A NO-REFERENCE QUALITY MEASURE FOR DIBR-BASED 3D VIDEOS

Mashhour Solh and Ghassan AlRegib
Georgia Institute of Technology
School of Electrical and Computer Engineering
{msolh,alregib}@gatech.edu

ABSTRACT
In this paper we present a no-reference objective quality measure for stereoscopic 3D videos generated by depth image-based rendering (DIBR). At first we will derive an ideal depth estimate for each pixel value. The ideal depth estimate will then be used to calculate three distortion measures: temporal outliers ($T_0$), temporal inconsistencies ($T_1$), and spatial outliers ($S_0$). The combination of the three measures constitute the proposed no-reference measure. Its performance is verified using subjective DMOS scores and compared to the full reference version of the proposed algorithm. The subjective results show that the proposed measure highly correlates with subjective scores and is close in performance to the full reference version of the measure.

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

1. INTRODUCTION

Three dimensional television (3D-TV) is broadly considered the future of television broadcasting that would bring a more life-like and visually immersive home entertainment experience. This trend is accompanied with a surge in the number of 3D movies that are being recorded in a stereoscopic format. While 3D content could be generated by capturing two video streams corresponding to left and right views, this approach is highly inefficient as it requires double the resources and it limits the resolution of the 3D video. As a solution, ATTEST (Advanced three-dimensional television system technologies) proposed a depth image-based rendering (DIBR) in which the two-views for 3D display can be generated from a single 2D image and a corresponding depth map [1]. DIBR has many advantages over two-views approach including high bandwidth efficiency, user interactivity, computational and cost efficiency, and 2D to 3D selectivity [1]. Moreover, DIBR eliminates photometric asymmetries in between the two views hence both of them are generated from the same original image. However, the perceived quality of DIBR-based stereoscopic 3D videos is error prone and can be affected by the following factors (Figure 1):

- 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,
- compression artifacts for the 2D video and depth map,
- transmission errors and streaming losses, and
- scaling and formatting algorithms in the 3D displays.

While 2D video quality is solely based on monocular color cues in one view, 3D video quality on the other hand is a combination of binocular and monocular cues. The depth illusion in stereoscopic video is constructed through binocular cues (horizontal disparity). Visual discomfort occurs whenever the depth through monocular cues (lighting, shading, motion parallax, texture gradient, blur, relative sizes ... etc.) mismatches the depth through binocular cues. Mismatches in depth cues such as unmatched color objects, conflicts between the blur in different depth planes and disparity, unmatched luminance and frame cancelation (near-edge cut-off for objects with front depth) can occur at any stage within the processing pipeline of 3D videos (Figure 1). These mismatches beside other factors such as excessive disparities, fast changing disparities, and geometric distortions constitute visual discomfort in stereoscopic videos [2]. 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 [3].

Existing stereoscopic video quality metrics are mostly based on measuring the 2D metrics for the left and the right image separately and then finding the combination of the values (e.g., sum and average) that would best predict the 3D video quality [4][5][6]. These methods assume that the perceived depth distortions are less significant than the perceived color distortions. These approaches have poor correlation with perceived 3D video quality and have been proved to be non-robust [7]. 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 swarm optimization [8]. 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 [9] 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 the 3D video quality, because the authors did not take into consideration the visual discomfort in analyzing depth information.

In our previous work, we have introduced a full reference objective quality measure for stereoscopic 3D videos generated by DIBR referred to as 3VQM [10]. 3VQM was based on evaluating elements of visual discomfort on an ideal depth estimate. We defined ideal depth as the per pixel depth that would generate a DIBR-based distortion-free 3D video. In this paper we introduce a no-reference 3D video quality measure for DIBR-based 3D video (NR-3VQM). First, we will show how to derive the ideal depth in a no-reference scenario. Then, we will present the three distortion measures that constitute NR-3VQM by quantitatively comparing the ideal depth estimate and the received depth maps in terms of visual discomfort. The proposed no-reference measure will also be evaluated against subjective scores.

The rest of the paper is organized as follows. The concept of ideal depth estimation and its derivation will be presented in Section 2. In Section 3 we will present the three distortion measures of visual discomfort in stereoscopic videos. The no-reference vision-based quality measure for 3D DIBR-based videos referred to as NR-3VQM will be introduced in Section 4. Finally, the experimental results and conclusion will be presented in Sections 5 and 6, respectively.

2. IDEAL DEPTH ESTIMATION

In DIBR, virtual views are generated by first projecting the pixels in the reference image to the world coordinates using depth map and camera information. The resulting pixels in the world coordinates are then sampled in the 2D plane from a different view-point to obtain a DIBR estimated image. This process is known as 3D wrapping [11].

Consider a reference camera $C_r$ and a virtual camera $C_v$, as shown in Figure 2. Where $F_r$ and $F_v$ are the focal lengths of the reference and the virtual cameras, respectively $^1$. $B$ is the baseline distance that separates the two cameras. $Z_c$ is the convergence distance of the two cameras. The horizontal coordinates vector $X_v$ of the virtual camera as a function of the horizontal coordinate vector $X_r$ of the reference camera is given by:

$$X_v = X_r + s \frac{F_v B}{Z_c} + h$$

$^1$ $F_r$ and $F_v$ will be assumed to be equal for the rest of this paper.
where \( s = -1 \) when the estimated view is to the left and \( s = +1 \) when the estimated view is to the right. \( \mathbf{Z} \) is a vector of the depth values at pixel location \((x_r, y_r)\), and \( h \) is the horizontal shift in the camera axis. During the 3D wrapping process errors in the depth map lead to errors in the relative pixel location. These displacements result in visual distortions in the form of temporal flickering and spatial noise. The 3D wrapping process leads to disocclusion, where some occluded areas in the reference image become visible in the virtual image. The process of removing disocclusion is referred to as hole filling. Several solutions to the disocclusion problem have been proposed in literature such as interpolation after depth map smoothing, extrapolation, background mirroring [1], asymmetric smoothing [12], distance dependent smoothing [13], edge-based smoothing [14], layered depth images (LDI) [15], and hierarchical hole filling (HHF) [16]. These hole filling algorithms may introduce artifacts around the disocclusion areas that correspond to regions where depth plane transition occurs.

In order to evaluate the perceived quality of the DIBR generated videos we need a reference depth for spatial and temporal analysis of the received depth and the rendered video. In [10] we have introduce the concept of ideal depth estimation to derive a valid depth reference for objective quality evaluation. A graphical illustration for ideal depth estimation for the no-reference scenario is defined as the per pixel depth that would generate a distortion-free virtual view assuming that same reference image and same DIBR parameters. The ideal depth is not given and must be estimated from the available colored images information.

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

The ideal depth estimation for the no-reference scenario is different from the full reference scenario. For full reference both the original depth map and the distortion-free view that needs to be estimated are given. The derivation for the no-reference ideal depth estimate is as follows:

1. Using the 3D wrapping equation in (1), we first express the horizontal coordinate \( \bar{X}_v \) vector of the synthesized virtual view as a function of the horizontal coordinate vector of the reference view \( \bar{X}_r \):

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

2. For the sake of derivation, let us first assume that we have obtained a distortion-free view that corresponds to the view we are trying to render by DIBR. Then the horizontal coordinate vector of that view \( \bar{X}_o \) can be expressed as a function of the horizontal coordinate vector \( \bar{X}_r \) of the the reference view:

\[
\bar{X}_o = \bar{X}_r + s \frac{F_v B}{Z_{IDEAL}} + h
\]

where \( Z_{IDEAL} \) is the ideal depth map vector to be estimated.

3. By subtracting (3) from (2) and then preforming direct substitution, the ideal depth vector \( Z_{IDEAL} \) can be expressed as:

\[
Z_{IDEAL} = \frac{s F_v B}{(X_o - X_v) + s \frac{F_v B}{Z}}
\]

4. In [3] the relation between the sum of squared differences (SSD) of the original video frame and its horizontal translations has been shown to be linear. Hence, for a small horizontal shift, \( \Delta \bar{X} \), the horizontal shift values for each pixel location can be estimated in terms of the intensity variations \( \Delta \bar{I} \approx \alpha \Delta \bar{X} \), where \( \alpha \) is a constant. As a result, the ideal depth vector can be estimated from the rendered virtual view intensity vector \( \bar{I}_v \), the distortion free view intensity vector \( \bar{I}_o \), the received depth map \( Z \) vector, focal length \( F_v \), and the baseline \( B \) as:

\[
\bar{Z}_{IDEAL} \approx \frac{s F_v B}{\alpha (I_o - I_v) + s \frac{F_v B}{Z}}
\]

5. Equation (5) is the full-reference ideal depth estimate. For the no-reference case \( I_o \) is not given. Hence, we cannot explicitly derive the ideal depth map. Instead, the difference \( (I_o - I_v) \) will be estimated as a function of intensity vector of the rendered virtual image \( I_v \) and the intensity vector \( I_r \) of the received reference image. If we assume this function to be \( f(\bar{I}_v, \bar{I}_r) \), then the ideal depth can be expressed as a function of \( f(\bar{I}_v, \bar{I}_r) \) as follows:

\[
\bar{Z}_{IDEAL} \approx \frac{s F_v B}{\alpha (f(\bar{I}_v, \bar{I}_r)) + s \frac{F_v B}{Z}}
\]
6. We define the function \( f(\bar{I}_v, \bar{I}_r) \) as the difference in intensity between each block in the reference view \( \bar{I}_r \) and the corresponding block in the rendered virtual view \( \bar{I}_v \) after applying a horizontal shift to the blocks of the reference view. \( \bar{Z}_{\text{IDEAL}} \) can then be calculated in an algorithmic manner as shown in Algorithm 1.

**Algorithm 1 Ideal depth approximation.**

\[
d \text{is variable initialized as the block size for } i = 1 \text{ to } \text{imagewidth} \text{ step } d \text{ do for } j = 1 \text{ to } \text{imageheight} \text{ step } d \text{ do } D = Z[i \text{ to } i + d, j \text{ to } j + d] \quad m = \text{mean}(D) \quad I_{\text{ref}} = I_v[i \text{ to } i + d, j + m \text{ to } j + d + m] \quad I_{\text{ver}} = I_v[i \text{ to } i + d, j \text{ to } j + d] \quad f[i \text{ to } i + d, j \text{ to } j + d] = I_{\text{ref}} - I_{\text{ver}} \quad \bar{Z}_{\text{IDEAL}}[i \text{ to } i + d, j \text{ to } j + d] = (s \frac{F_r}{B})/(\alpha(f) + s \frac{F_v}{D}) \quad \text{end for} \quad \text{end for}
\]

The choice of the block size \( d \) does affect the noise level in the estimated ideal depth. As we increase the block size, the horizontal shift applied to block in the reference view \( \bar{I}_r \) will less likely correspond to the right block in the rendered virtual view \( \bar{I}_v \). In the results section, we show that (Table 1) as the value of \( d \) increases the root means square error (RMSE) of the subjective results and the no-reference measure increase as well. Moreover, as the block size increases, the percentage of outliers increases. Nevertheless, from our experiments, a block size of \( d = 2 \) or \( d = 5 \) has a low RMSE, high correlation values and no outliers.

Now that we have derived the the ideal depth estimate, the next step is to calculate the distortion measures that would evaluate different elements of visual discomfort in the DIBR generated 3D video as a function of the estimated ideal depth map and the received depth map. These distortion measures capture the visual distortions due to the errors caused by bad pixels in the depth maps from stereo matching and/or compression as well as the errors caused by post processing of the synthesized colored video itself such as hole-filling and colored video compression.

### 3. DISTORTION MEASURES

The three distortion measures defined in [10] are going to be used in here as well. These measures are Temporal error outliers (TO), Spatial error outliers (SO) and Temporal inconsistency (TI). Both TO and SO are a function of \( \Delta Z \). Where \( Z \) is a matrix composed of the vectors \( \bar{Z} \)’s, defined as follows:

\[
\Delta Z = |Z_{\text{IDEAL}} - Z|
\]  

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 such error will be a slight increase or decrease in the perceived depth. This slight change has no effect on the perceived quality of the 3D video. Hence, a non-zero value of \( \Delta Z \) does not always mean a visual distortion. Visual distortions occur when an non-uniform error in depth plane causes a relocation of color pixel/blocks during the wrapping process to an alien position. The visual effect of this relocation on the synthesized view is spatially noticeable around texture areas in the form of inconsistent depth cues (unmatched object colors). Also, temporal inconsistency in \( \Delta Z \) 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 what preceded we defined the distortion measures as follows.

#### 3.1. Spatial Outliers (SO)

The spatial inconsistencies are measured through the spatial outliers (SO), defined as the standard deviation of \( \Delta Z \):

\[
SO = STD(\Delta Z)
\]  

The standard deviation in this case is meant to separate the spatially visible distortions caused by non-zero \( \Delta Z \) from the perceptually non-significant \( \Delta Z \)’s. The perceptually nonsignificant \( \Delta Z \)’s are spatially homogeneous and are caused by inaccuracies in the wrapping equation and inherited approximation in the camera modeling parameters.

#### 3.2. Temporal Outliers (TO)

Similarly, the temporal outlier (TO) is defined as the standard deviation of the change in \( \Delta Z \) for two consecutive frames:

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

The error introduced due to depth map noise and hole filling is random and temporally inconsistent. By taking the standard deviation, TO filters out the edginess in SO and will only keep the visible distortions due to depth map errors and hole filling.

#### 3.3. Temporal Inconsistencies (TI)

Temporal inconsistencies (TI) measures fast changing disparities which mainly lead to visual discomfort. Errors due to stereo matching, hole filling algorithms and depth compression lead to sudden changes in disparity. Their visual effect are usually observed in the form of flickering around smoothly textured areas and noise around highly structured regions. TI is defined as:

\[
TI = STD(Z_{t+1} - Z_t)
\]
4. NO-REFERENCE 3D VISION-BASED QUALITY MEASURE

Artifacts leading to visual discomfort in DIBR-based stereoscopic videos are captured by at least one of the three measures introduced above. Hence, we combined the three measures into one no-reference 3D vision-based quality measure for stereoscopic DIBR-based videos as follows:

\[
NR-3VQM = K \left( 1 - SO (SO \cap TO) \right)^a \left( 1 - TI \right)^b \left( 1 - TO \right)^c
\]

(11)

where SO, TO, and TI are normalized to the range 0 to 1 and \(a\), \(b\), and \(c\) are constants which were determined by running a few training sequences. \((SO \cap TO)\) is the logical intersection of SO and TO included in the equation to avoid accounting the outlier distortion more than once \(^2\). \(K\) is a constant for scaling where \(NR-3VQM\) ranges from 0 for lowest quality to \(K\) for highest quality. The overall quality measure is calculated as the mean of the values in the matrix \(NR-3VQM\).

5. SUBJECTIVE RESULTS

In order to validate our objective results we conducted subjective experiments using a Samsung 2233RZ display with the shutter glass solution from NVIDIA. The testing conditions were chosen to be consistent with the new requirements for subjective subjective video quality assessment methodologies for 3DDTV described in [17]. We recruited 20 volunteers who were mostly engineers with little to no previous experience of 3D video processing. Each volunteer was asked to assign each video sequence with a score indicating his/her assessment of the quality of that video. The quality was defined as the extent to which the distortions were visible and annoying. The raw scores for each subject were collected and processed to give Mean Opinion Scores (MOS) and a Difference Mean Opinion Score (DMOS) for each distorted videos. The tested videos were a total of 31 video sequences each of 30 seconds in length. The DMOS results for the video sequences were divided into two groups. For the 21 video sequences of first group we had both the reference distortion-free video the original depth (before processing) and hence the objective quality was measured using the full reference as in [10]. However, for the 10 video sequences of the second group we had no information regarding the original depth or the reference distortion-free videos. As a result, the objective quality of the second group was measured using the no-reference measure. Figure 4 shows the scatter plot for both the full reference and the no reference measures versus DMOS. The sequences were all DIBR synthesized videos for three different applications: depth map and colored video compression, depth estimation (stereo matching), and depth from 2D to 3D conversion using color information. In order for the values of the full reference measure to have meaningful representation as well as making it easier to compare against the MOS values we have set \(K = 5\) in (11). The constants \(a\), \(b\), and \(c\) were determined after a small training experiment conducted using three video sequences in which three different volunteers were asked to rate the synthesized videos. 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 so that we can make sure that will not bias our results. Consequently, the constants were set to the following values: \(a = 8\), \(b = 8\), and \(c = 6\).

Looking at the results in Figure 4, we notice that similar to the full-reference 3VQM all the objective ratings of the proposed no-reference measure are inside the outlier boundary defined by the quality ratings that are greater that two DMOS standard deviation away from the ideal rating. We also notice that almost more that 70% of the no reference measure values fall inside the one \(\sigma_{DMOS}\) boundary. Which means that the no reference measure is significantly consistent with the subjective scores and has no outliers. The validation scores for the full reference measure, the proposed no reference measure, and the combination of both measures are show in Table 1.

Now, we show results when we vary the block size, Table 1. The results in Table 1 show that the no-reference measure value for \(d = 2\) has lower RMSE and MAE values but slightly lower correlation values. This indicates that for \(d = 2\) no reference measure has a high accuracy but slightly less coherent than full reference. With \(d = 5\) the RMSE and MAE values are higher however both Pearson linear correlation coefficient (CC) and Spearman rank order correlation coefficient

\(^2\)For numerical values all nonzero values in the \(\cap\) are considered as 1’s
Table 1. Validation scores for the full reference, the no reference and the combination of both the full reference and the no reference measures. The validation criteria are: root mean squared error (RMS), Pearson linear correlation coefficient (CC), Spearman rank order correlation coefficient (ROCC), mean absolute error (MAE), Outlier Ratio (OR) and the standard deviation of the DMOS values $\sigma_{DMOS}$.

<table>
<thead>
<tr>
<th></th>
<th>RMSE</th>
<th>CC</th>
<th>ROCC</th>
<th>MAE</th>
<th>OR</th>
<th>$\sigma_{DMOS}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Reference ($d = 2$)</td>
<td>0.5870</td>
<td>0.8529</td>
<td>0.1180</td>
<td>0.5094</td>
<td>0</td>
<td>0.6652</td>
</tr>
<tr>
<td>No Reference ($d = 5$)</td>
<td>0.6384</td>
<td>0.8662</td>
<td>0.4445</td>
<td>0.5551</td>
<td>0</td>
<td>0.6652</td>
</tr>
<tr>
<td>No Reference ($d = 10$)</td>
<td>0.7139</td>
<td>0.8762</td>
<td>0.1180</td>
<td>0.6440</td>
<td>0</td>
<td>0.6652</td>
</tr>
<tr>
<td>No Reference ($d = 100$)</td>
<td>1.6857</td>
<td>0.9157</td>
<td>0.1003</td>
<td>1.6632</td>
<td>0.8</td>
<td>0.6652</td>
</tr>
<tr>
<td>Full Reference</td>
<td>0.6158</td>
<td>0.8942</td>
<td>0.7890</td>
<td>0.5173</td>
<td>0</td>
<td>1.0082</td>
</tr>
<tr>
<td>Full Reference And No Reference ($d = 5$)</td>
<td>0.6875</td>
<td>0.8728</td>
<td>0.7894</td>
<td>0.5967</td>
<td>0</td>
<td>0.7885</td>
</tr>
</tbody>
</table>

(ROCC) values improved. We can see that the no reference measure with $d = 5$ is more coherent and is closer in performance to the full reference. Outlier ratio is zero except at $d = 100$ which is due to the fact that the correlation is eliminated at large block size. For small block sizes the outlier ratio indicates a very consistent quality predictions for the no reference measure for small block size values. The validation scores in Table 1 for the combination of the full reference and the no reference measures show that ideal depth evaluation for visual discomfort yields a very accurate, coherent and consistent objective quality prediction for DIBR-based stereoscopic videos.

6. CONCLUSION

In this paper we introduced a no-reference objective video quality measure for DIBR-based stereoscopic 3D videos. We have also shown how to derive the ideal depth estimate for the no-reference scenario. The ideal depth estimate was then evaluated by the three distortion measures: Temporal error outliers (TO), Spatial error outliers (SO) and Temporal inconsistency (TI). The combination of the three indexes forming the proposed measure was verified against subjective rating and compared to the full reference version of the measure. The results have shown that the predictions of the no-reference measure highly correlates with subjective scores and is fairly close in performance to the full reference 3VQM.

7. ACKNOWLEDGEMENTS

We would like to thank Judit Martinez Bauza for her contributions to this work.

8. REFERENCES


