Optimum Image Quality Assessment for 3D Perception of Stereoscopic Image Generated from Upsampled Depth Map

Saeed Mahmoudpour and Manbae Kim
Dept. of Computer and Communications Engineering, Kangwon National University
Chunchon, Republic of Korea
E-mail: {saeed, manbae}@kangwon.ac.kr

Abstract – Depth map upsampling is an approach to increase the spatial resolution of depth maps obtained from ToF (Time of Flight) cameras. Since depth map quality directly affects 3D perception of stereoscopic image, applying different depth upsampling methods to a low resolution depth map causes a variety of perceptions of the stereoscopic images. In this paper, we investigate the relation between objective measurements and 3D subjective evaluation. For the former, diverse full-reference and no-reference quality assessment tools are applied to measure the quality of depth maps obtained. Subjective evaluation is carried out by DSCQS test on the stereoscopic images. We utilize several upsampling methods to achieve different depth map qualities of a scene as our experiment samples. Finally, the quantitative value of each measurement is compared with a subjective assessment using correlation coefficients. The experiment shows promising results that could help to select the most appropriate objective quality assessment tool(s) for stereoscopic image quality measurement.

Keywords: Depth map, Upsampling, Objective assessment, Correlation

1 Introduction

Following the recent technology advances in camera systems and computer vision, three-dimensional active cameras are capable to provide accurate distance information of a scene. The high speed Time of Flight (ToF) cameras extract reliable depth maps. However, the spatial resolution of depth maps is relatively low in comparison with original images. Therefore, diverse depth map upsampling approaches are provided to obtain high-resolution depth maps. Also, it is important to evaluate the upsampling quality in order to realize upsampling performance on 3D content quality.

Image quality assessment (IQA) helps us to evaluate the quality of the upsampled depth map using any assessment tools. The most frequently used method is PSNR (Peak Signal-to-Noise). As well, due to various sources of quality degradation and visual discomfort which degrade the end-user 3D experience and lack of accurate objective IQA tools, the subjective assessment is commonly used for 3D quality evaluation. Investigating the similarity between 2D quality evaluation and 3D perception, we will search for the most accurate IQA tool for 3D quality assessment. Using the proper automatic objective IQA tool will help to predict the quality of 3D image without using the expensive subjective test and even without watching the stereoscopic image.

Depth map quality has significant effect on 3D perception, therefore, first we have implemented seven upsampling methods and the depth map quality of each method is computed using different objective IQA tools. Then, the quality results are compared with a reliable subjective assessment using correlation.

The Pearson, Spearman and Kendall correlations are three proper approaches for similarity measurement between each objective assessment result and the subjective IQA. Comparing the objective and subjective scores, it will be inferred that which objective IQA tools show the most correspondence with human judgment and have superiority for 3D quality assessment. DIBR (depth image based rendering) or 2D+Depth is used to generate a stereoscopic image. Fig. 1 shows the overall framework that examines the relation between upsampling methods and 3D perception.

Fig. 1. The flow diagram for testing 3D quality of depth upsampling methods.
Since it is difficult to investigate all methods, seven approaches are selected to be used in this work. The bilinear upsampling (BLU) uses average weighted of four neighboring pixels for interpolation to achieve upsampled depth map. A similar method called bicubic upsampling (BCU) is based on sixteen neighboring pixels. The bilateral upsampling (BU) [1] is a prevalent approach combines a spatial filter and a range filter to preserve the edge regions in upsampling process. Another upsampling method based on the bilateral upsampling is joint bilateral upsampling (JBU) [2] which utilizes both a color data and its low-resolution depth map. The variance-based upsampling (VBU) [3] avoids the usage of the constant variance and uses a variance that is computed on each pixel block. The disadvantage of the JBU is that it is sensitive to homogeneous regions and the weighting function can be assigned a wrong variance in non-edge regions. To solve this problem, an adaptive bilateral upsampling method (ABU) [4] has been proposed, where a large weight is assigned to color image at edge pixels and a large weight is assigned to depth data at non-edge pixels. To overcome the limitation in reducing blur at low-gradient edge regions in prior methods, a distance transform-based bilateral upsampling (DTBU) [5] has been proposed.

This paper is organized as follows. In the next section, different IQA tools considered in this work are described. The correlation coefficients for similarity measurements are introduced in section 3 and the experimental results with correlation coefficients for similarity measurements are different IQA tools considered in this work are described. The has been proposed. Finally, we introduced in section 3 and the experimental results with correlation coefficients for similarity measurements are different IQA tools considered in this work are described. The

2 Objective Quality Assessment Tools

Full-reference image quality assessment (FR IQA) compares test and reference images, therefore, both ground-truth and upsampled depth map are needed. The no-reference/blind image quality assessment (NR IQA) refers to quality assessment of images by an algorithm where only the distorted image is accessible and no information about the reference image is available. In this paper, several FR IQA and NR IQA tools are used to evaluate the performance of different upsampling methods. The quality metrics are introduced as follows:

2.1 FR IQA tools

A. PSNR

PSNR is one of the most prevalent tools for image quality evaluation defined by (1):

$$PSNR = 10 \cdot \log\left[\frac{255^2}{\sum (D^b - D^a)^2}\right]$$

where $D^b$ and $D^a$ are ground-truth and upsampled depth maps respectively.

B. SSIM

A sophisticated tool for image quality evaluation is structural similarity index measure (SSIM) [6] that measures the similarity between two images and considered to be correlated with the quality perception of the human visual system (HVS). SSIM principle is based on the modeling any image distortion as a combination of luminance distortion, contrast distortion and loss of correlation. SSIM value for two images $f$ and $g$ is expressed by

$$SSIM = \frac{l(f,g)c(f,g)s(f,g)}{l(f,g)c(f,g)s(f,g)}$$

where $l(f,g)$, $c(f,g)$ and $s(f,g)$ are luminance, contrast and structure comparison functions respectively. $\sigma_l$ and $\sigma_g$ denote standard deviations, $\mu_l$ and $\mu_g$ are mean values and $\sigma_{lg}$ is covariance. $C_1$, $C_2$ and $C_3$ are positive constants added to avoid a null denominator. The SSIM is a value between 0 and 1 that higher value shows more similarity.

C. VIF

Visual Information Fidelity (VIF) [7] is a full-reference image quality metric that uses information theoretic criterion for image fidelity measurement. In an information-theoretic framework, the information that could ideally be extracted by the brain from the reference image and the loss of this information to the distortion are quantified in VIF method using natural scene statistics (NSS), HVS, and an image distortion (channel) model. The VIF is derived from a quantification of two mutual information quantities: the mutual information between the input and the output of the HVS channel when no distortion channel is present (called the reference image information) and the mutual information between the input of the distortion channel and the output of the HVS channel for the test image. Similar to SSIM, the assessment result is represented using a value between 0 and 1.

2.2 NR IQA tools

A. Sharpness Degree

Sharpness degree ($\theta$) is used to represent the extent of sharpness of the image and is defined by (2).

$$\text{Sharpness Degree} = \frac{1}{MN} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} G(x,y)$$

where

$$G(x,y) = \sqrt{(D(x,y) - D(x-1,y))^2 + (D(x,y) - D(x,y-1))^2}$$

B. Blur Metric

Another tool for measuring blur attempts to measure the spread of the edges. First, we apply an edge detector (e.g. a Sobel edge detector) to a grayscale image. We scan each row of the image for pixels corresponding to an edge location. The start and end positions of the edge are defined as the locations of the local extrema closest to the edge. The spread of the edge is then given by the distance between the end and start positions, and is identified as the local blur measure for this...
edge location. The global blur measure for the whole image is obtained by averaging the local depth values over all edges found [9].

\[
\text{Blur Metric} = \frac{\text{Sum of all edge widths}}{\text{No. of edges}} \tag{4}
\]

\section*{C. BIQI}

Blind image quality index (BIQI) [10] identifies the likeliest distortion in the image and then quantifies this distortion using a NSS-based approach. Given a distorted image, the algorithm first estimates the presence of a set of distortions in the image that consists of JPEG, JPEG2000, white noise (WN), Gaussian Blur (Blur) and Fast fading (FF). The amount or probability of each distortion in the image is denoted as \( p_i \) \( i=1,2,...,5 \). The method performs quality assessment in two stages. This first stage is a classification and the second stage attempts to evaluate the quality of the image along each of these distortions. The quality of the image is then expressed as a probability-weighted summation:

\[
\text{BIQI} = \sum_{i=4}^5 p_i \cdot q_i \tag{5}
\]

where \( q_i \) \( i=1,2,...,5 \) represents the quality scores from each of the five quality assessment algorithms (corresponding to the five distortions).

\section*{D. NIQE}

Natural image quality evaluator (NIQE) [11] is a completely blind image quality analyzer that only uses measurable deviations from statistical regularities observed in natural images, without training on human-rated distorted images. Unlike current general purpose no reference (NR) IQA algorithms which require knowledge about anticipated distortions in the form of training examples and corresponding human opinion scores, NIQE uses a quality aware collection of statistical features based on the simple and successful space domain, the NSS model. These features are derived from a corpus of natural, undistorted images.

The quality scores for both BIQI and NIQE are expressed by a value between 0 and 100 (0 represents the best and 100 the worst quality).

\section*{3 3D Subjective Quality Assessment}

The quality of stereoscopic images made by DIBR technique is evaluated using subjective quality experiment. We observed the stereoscopic images with a 3D monitor adopting DSCQS (Double Stimulus Continuous Quality Scale) subjective test [12]. In the first stage, original views were displayed to ten participants. Each participant watched an original stereoscopic image for 10 seconds and another stereoscopic image made by an upsampled depth map for the same period, and the effect of the 3D depth is evaluated. For each image data, similar viewing was carried out in order to examine the 3D perception. Depth perception was then subjectively judged on a scale of 1 (bad), 2 (poor), 3 (fair), 4 (good) and 5 (excellent) in terms of 3D perception.

\section*{4 Correlation Measurement Metrics}

In order to investigate the relationship between objective measurement tools and 3D subjective evaluation, three approaches are utilized.

\subsection*{4.1 Pearson Correlation}

A popular metric for measuring the association of two continuous variables is Pearson’s correlation coefficient. Pearson method shows the strength of relationship using a coefficient ranges from -1 to 1. Positive coefficient means simultaneous changes in two variables, negative coefficient implies inverse association and zero means no association between variables. Pearson’s coefficient (\( \rho \)) is defined by

\[
\rho = \frac{\text{Cov}(x,y)}{\sigma_x \sigma_y} \tag{6}
\]

where \( \text{Cov}(x,y) \) is the covariance between two groups and \( \sigma_x \) and \( \sigma_y \) denotes standard deviations.

\subsection*{4.2 Spearman Correlation}

This correlation coefficient is a rank-based version of the Pearson’s correlation coefficient. First, the samples of each group with \( n \) variables are ranked from 1 to \( n \) (value 1 shows the smallest and value \( n \) denotes the biggest sample). Then, Spearman’s correlation coefficient (\( \rho_s \)) can be calculated as

\[
\rho_s = \frac{\sum_{i=1}^n ((\text{rank}(x_i) - \text{rank}(x))((\text{rank}(y_i) - \text{rank}(y)))}{\sqrt{\sum_{i=1}^n ((\text{rank}(x_i) - \text{rank}(x))^2 \sum_{i=1}^n ((\text{rank}(y_i) - \text{rank}(y))^2} \tag{7}
\]

where \( \text{rank}(x_i) \) and \( \text{rank}(y_i) \) are the ranks of the observation in the sample. Similar to Pearson, Spearman’s coefficient varies from -1 to +1 and the absolute value of \( \rho_s \) indicating the strength of association.

\subsection*{4.3 Kendall’s Tau Correlation}

Kendall’s tau ranges from -1 to +1 and describes the strength of the relationship between the two variable similar to previous correlation coefficients. This metric is defined to measure how much two variables are correlated. This coefficient quantifies the difference between the number of concordant and discordant pairs. Any two pairs of ranks \( (x_i, y_i) \) and \( (x_j, y_j) \) are concordant if \( (x_i - x_j)(y_i - y_j) > 0 \) and discordant when \( (x_i - x_j)(y_i - y_j) < 0 \). Kendall’s tau coefficient is

\[
\tau = \frac{\sum_{i=1}^n \sum_{j=1}^n \text{sgn}(x_i - x_j)\text{sgn}(y_i - y_j)}{n(n-1)} \tag{8}
\]

where

\[
\text{sgn}(X) = \begin{cases} 
1 & \text{if } X > 0 \\
0 & \text{if } X = 0 \\
-1 & \text{if } X < 0 
\end{cases}
\]

\section*{5 Experimental Results}

The quality performance of the seven upsampling methods are evaluated using ten test depth maps from Middlebury
stereo dataset [13]. The test RGB images and related depth maps are shown in Fig. 2. In order to obtain low-resolution depth maps, we downsampled the original data and then, we made high-resolution depth maps.

Figs. 3–5 show the upsampled depth maps of the seven methods for aloe, cone, and bowling. The stereoscopic images are shown in Fig. 6.

Table 1. Average subjective measurement data of upsampled depth maps

<table>
<thead>
<tr>
<th>Method</th>
<th>Visual Fatigue</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLU</td>
<td>3.76</td>
</tr>
<tr>
<td>BCU</td>
<td>3.64</td>
</tr>
<tr>
<td>BU</td>
<td>3.8</td>
</tr>
<tr>
<td>JBU</td>
<td>3.84</td>
</tr>
<tr>
<td>VBU</td>
<td>4.03</td>
</tr>
<tr>
<td>ABU</td>
<td>4.46</td>
</tr>
<tr>
<td>DTBU</td>
<td>3.99</td>
</tr>
</tbody>
</table>

Table 2. Average objective measurement data of upsampled depth maps (PSNR unit: dB)

<table>
<thead>
<tr>
<th>Depth map</th>
<th>BLU</th>
<th>BCU</th>
<th>BU</th>
<th>JBU</th>
<th>VBU</th>
<th>ABU</th>
<th>DTBU</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSNR image</td>
<td>35.85</td>
<td>35.71</td>
<td>35.64</td>
<td>34.15</td>
<td>35.64</td>
<td>33.16</td>
<td>34.86</td>
</tr>
<tr>
<td>Edge PSNR</td>
<td>23.68</td>
<td>23.55</td>
<td>23.66</td>
<td>22.82</td>
<td>23.38</td>
<td>20.97</td>
<td>22.93</td>
</tr>
<tr>
<td>Non-edge PSNR</td>
<td>38.07</td>
<td>37.94</td>
<td>37.78</td>
<td>37.50</td>
<td>37.93</td>
<td>35.43</td>
<td>36.92</td>
</tr>
<tr>
<td>Sharpness</td>
<td>39.5</td>
<td>42.3</td>
<td>49.51</td>
<td>49.09</td>
<td>31.92</td>
<td>88.51</td>
<td>68.14</td>
</tr>
<tr>
<td>Blur</td>
<td>8.48</td>
<td>11.38</td>
<td>10.29</td>
<td>10.87</td>
<td>10.51</td>
<td>9.00</td>
<td>9.89</td>
</tr>
<tr>
<td>SSIM</td>
<td>0.976</td>
<td>0.955</td>
<td>0.975</td>
<td>0.956</td>
<td>0.971</td>
<td>0.962</td>
<td>0.972</td>
</tr>
<tr>
<td>Edge SSIM</td>
<td>0.957</td>
<td>0.915</td>
<td>0.955</td>
<td>0.944</td>
<td>0.952</td>
<td>0.942</td>
<td>0.9</td>
</tr>
<tr>
<td>VIF</td>
<td>0.518</td>
<td>0.539</td>
<td>0.424</td>
<td>0.422</td>
<td>0.478</td>
<td>0.398</td>
<td>0.438</td>
</tr>
<tr>
<td>BIIQ</td>
<td>57.8</td>
<td>66.34</td>
<td>63.11</td>
<td>32.81</td>
<td>41.94</td>
<td>29.15</td>
<td>72</td>
</tr>
<tr>
<td>NIQE</td>
<td>15.95</td>
<td>13.11</td>
<td>13.94</td>
<td>11.82</td>
<td>12.47</td>
<td>13.41</td>
<td>13.82</td>
</tr>
</tbody>
</table>

Table 1 and 2 represent the quality scores for each upsampling method measured from different objective quality metrics and subjective experiment (Visual Fatigue) respectively. The results are derived from averaging the quality scores of the collection of 16 images.

The 3D perception grades of each upsampling method in Table 1 are based on 3D visual fatigue. The visual fatigue values obtained by the subjective test are 3.76 (BLU), 3.64
(BCU), 3.89 (BU), 3.84 (JBF), 4.03 (VBU), 3.46 (ABU) and 3.99 (DTBU). Table 2 compares the objective measurement data of the seven upsampling methods.

The quality scores of upsampling depth maps obtained from each IQA metric is considered as a group of seven samples. All values are normalized by scaling between 0 and 1 and the similarity of samples distribution in each IQA group is compared with visual fatigue samples group using Pearson, Spearman and Kendall correlation coefficients. Table 3 shows the correlation results.

Before evaluating the strength of correlation using different correlation coefficients, it is worth mentioning that Pearson’s correlation coefficient takes into account both the number and degree of concordances and discordances, whereas Kendall’s tau correlation coefficient shows only the number of concordances and discordances. Spearman’s correlation is in between of the Pearson’s and Kendall’s, reflecting the degree of concordances and discordances on the rank scale. The disadvantage of Pearson is the sensitivity to outliers (an observation that is numerically distant from the rest of the data). In this case, Spearman and Kendall are less sensitive to outliers and preferable.

According to Table 3, edge PSNR shows higher value of correlation compare to common PSNR and non-edge PSNR. Also, Pearson coefficient is much higher than Spearman result that indicates the distribution is non-linear. In this case, Spearman and Kendall results are more reliable.

Sharpness degree and blur metric show negative and positive correlation values respectively. These two results confirm the fact that image with high spatial frequency (sharper) reveals much noticeable visual discomfort than that with low frequency [14].

SSIM uses luminance, contrast and structure features to measure quality. Similar to PSNR, Pearson coefficient is higher than two other correlation coefficients in this metric. SSIM has the highest Spearman value among other metrics. Thus, it is the most similar metric to visual fatigue in the case of samples order.

As mentioned earlier, Pearson is very sensitive to outliers and its value can be drastically influenced by a few extreme values. Negative value of Pearson with positive Spearman and Kendall coefficients for edge-SSIM show that Pearson can not be used due to outliers, therefore, Pearson correlation may severely underestimate the strength of a relationship between two variables. In this case, we should rely on Spearman results that reveals correlation more than edge-PSNR but less than SSIM.

VIF results are based on NSS, HVS, and an image distortion (channel) model in wavelet domain and shows a positive but low correlation to visual fatigue.

BIQI and NIQE are two NR IQA metrics that are expected to show lower correlation in comparison to FR IQA metrics. JPEG, JPEG2000 (JP2K), white noise (WN), Gaussian Blur (Blur) and Fast fading (FF) are five distortions that are considered in BIQI method for quality measurements. Similar to negative results of Sharpness degree, BIQI is not correlated with visual fatigue results.

NIQE metric delivers a positive correlation using Pearson coefficient. Also, Spearman and Kendall correlation results are comparative to some results derived from FR IQA methods. NIQE results are near to VIF, therefore, it can be inferred that NIQE is an acceptable quality assessment tool when there is no access to reference image. Fig. 4 shows correlation values for different quality metrics in column diagram mode.

<table>
<thead>
<tr>
<th></th>
<th>Pearson</th>
<th>Spearman</th>
<th>Kendall</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSNR</td>
<td>0.582</td>
<td>0.0357</td>
<td>0.0476</td>
</tr>
<tr>
<td>Edge-PSNR</td>
<td>0.608</td>
<td>0.1429</td>
<td>0.1429</td>
</tr>
<tr>
<td>Non-edge PSNR</td>
<td>0.554</td>
<td>0.0357</td>
<td>0.0476</td>
</tr>
<tr>
<td>Sharpness</td>
<td>-0.522</td>
<td>-0.3214</td>
<td>-0.1429</td>
</tr>
<tr>
<td>Blur Metric</td>
<td>0.273</td>
<td>0.1429</td>
<td>0.0476</td>
</tr>
<tr>
<td>SSIM</td>
<td>0.505</td>
<td>0.3571</td>
<td>0.1429</td>
</tr>
<tr>
<td>Edge-SSIM</td>
<td>-0.025</td>
<td>0.1441</td>
<td>0.0476</td>
</tr>
<tr>
<td>VIF</td>
<td>0.019</td>
<td>0.1071</td>
<td>0.1429</td>
</tr>
<tr>
<td>BIQI</td>
<td>-0.339</td>
<td>-0.3214</td>
<td>-0.2381</td>
</tr>
<tr>
<td>NIQE</td>
<td>0.132</td>
<td>0.1071</td>
<td>0.1429</td>
</tr>
</tbody>
</table>
6 Conclusions

In this paper, we investigated a method to find the superior quality assessment tools, which are commonly used in 2D images, for 3D quality assessment. Using correlation, we successfully achieved a reasonable relation between objective QA results and subjective assessment. Also, several upsampling results are provided to use as group samples in this work. As a result, PSNR and SSIM show the highest Pearson correlation coefficients. However, corresponding Spearman and Kendall for PSNR assessments are far from Pearson coefficients. The reason of this difference is outliers or highly skewed variables, therefore, the distribution is not linear and Pearson is unreliable.

According to Spearman and Kendall coefficients for Edge-PSNR, this metric is less correlated compare to SSIM but better than other metrics. Consequently, SSIM is the most concordant metric according to Spearman coefficient and also with high Pearson and Kendall coefficient values. Furthermore, Sharpness Degree and BIQI are not appropriate tools due to their negative correlation coefficients.

7 Acknowledgements

This work was supported by Small & Medium Business Administration (SMBA) under Technology Innovation Project (S2056930) and in part by the MKE (The Ministry of Knowledge Economy)/KEIT [10041082, System and Semiconductor Application Promotion Project].

8 References