STEREOSCOPIC VIDEO QUALITY ASSESSMENT BASED ON STEREO JUST-NOTICEABLE DIFFERENCE MODEL

Feng Qi¹, Tingting Jiang²,³, Xiaopeng Fan¹, Siwei Ma², Debin Zhao¹

1. School of Computer Science and Technology
   Harbin Institute of Technology
   Harbin, China
2. National Engineering Lab for Video Technology
3. Key Lab. of Machine Perception (MoE) School of EECS
   Peking University
   Beijing, China

ABSTRACT

In this paper, we propose a full reference Stereoscopic Video Quality Assessment (SVQA) algorithm based on the Stereo Just-Noticeable Difference (SJND) model. Firstly, SJND mimic the human binocular visual system characteristics from four factors, including: sensitivity of luminance contrast, spatial masking, temporal masking and binocular masking. Secondly, based on the SJND model, the full reference SVQA is developed, by capturing spatio-temporal distortions and binocular perceptions. Finally, experimental results have demonstrated that the proposed SVQA outperforms other four current evaluation methods and has a good consistency with the observers' subjective perception.

KEYWORDS—Video signal processing, Image quality

1. INTRODUCTION

As the middle of last century, stereoscopic videos are encountering the second upsurge which arouse by Hollywood’s 3D movies – Avatar, 2012 Titanic, etc. Different from its first prevalence, the noticeable development of 3D techniques including capturing, encoding and displaying provides more realistic experience and higher quality of stereoscopic videos. For stereoscopic video compression, the additional spatial and temporal statistical redundancy should mainly be removed. Therefore, 3DAV (3D Audio-Visual) group of Moving Picture Experts Group (MPEG) and Joint Video Team (JVT) of ITU-T Video Coding Experts Group (VCEG) develop a new standard for multiview video coding (MVC) [1]. Through the research of the Human Visual System (HVS) sensitivity to luminance contrast and spatial/temporal masking effects with the JND, 3D image/video coding has got higher compression efficiency and better perception quality. Although human binocular vision is an up-to-date sealed book in physiology and psychology, the JND method is still available to describe the perception redundancy quantitatively in the 3D IQA/VQA.

In the image/video quality assessment literature [2-9], JND models can be grouped into two categories: 1) transformation domain, such as DCT and wavelet domain JND[2-4], and 2) pixel-domain [5-9]. [2] proposed a DCT based JND model for monochrome pictures which combine spatial and temporal factors. A full reference VQA algorithm based on the Adaptive Block-size Transform JND model is proposed in [3]. Just like DCT-based JND, [4] defined JND model in DWT. Compared to the transformation domain models, the pixel domain JND can simplify calculation to spatially localized information, so it is more prevalently used in motion estimation and quality assessment. In [5][6], the spatial JND threshold could be modeled as a function of luminance contrast, spatial masking and temporal masking effect, respectively. [7] extended the JND model, where the Nonlinear Additively Masking Model (NAMM) is used to integrate the luminance masking and texture masking. [8] considered several factors including spatial contrast sensitivity function (CSF), luminance adaptation, and adaptive inter- and intra-band contrast masking. According to the non-uniform density of human photoreceptor cells on the retina, [9] proposed foveated JND based on the spatial-temporal JND. Although the above literatures have showed good performance in image processing systems, they are all based on the property of human monocular vision. However, HVS is a complicated system which is composed of two eyes. Three-dimensional IQA/VQA should take more account of the characteristic of stereoscopic images and human binocular vision. [10] Several works (e.g. [11][12]) illustrated that HVS compensates for the lack of high frequency components in one view if the other view is at sufficiently high quality. Binocular JND (BJND) model [13] is proposed to mimic the basic binocular vision properties in response to asymmetric noises in a pair of stereoscopic images by two psychophysical experiments. [14] derived a mathematical model to explain the just noticeable difference in depth. Based on the idea that human has different visual perception for the objects with different depths, [15] proposed a depth perception based joint JND model for stereoscopic images.
According to the studies of psychovision [5], there exists the inconsistency in sensitivity inherent of HVS, named as "perceptual redundancies". JND is based on this premise. However, for human binocular vision, one spatial point is projected into two different locations on both retinas. These differences, referred to as binocular disparity, provide information that the brain can use to calculate depth in the visual scene, providing a major means of depth perception. For each retina, account of the distribution densities of visual acuity cells are non-uniform. And human is only sensitive to the object close to the fixation point. The magnitude of disparity affects visual acuity, which is related to visual masking. Therefore, besides intra-view masking effects, there exist inter-view masking effects in human binocular vision. Four major factors have been validated to influence the distortion visibility threshold of stereoscopic videos by previous literatures [3-15]. They are:

(1) Luminance Contrast: As indicated by Weber’s law [7], human visual perception is sensitive to luminance contrast rather than absolute luminance value.
(2) Spatial Masking: The reduction in the visibility of the stimuli is caused by the increase in the spatial non-uniformity of the background luminance.
(3) Temporal Masking: Like the spatial masking characteristic of the HVS for images, temporal masking has similar peculiarity for videos.
(4) Binocular Masking: When dissimilar monocular stimuli are presented to corresponding retinal locations of the two eyes, one stimulus has an effect on the other stimulus.

Previous JND models only considered one or two kinds of above four factors. However, in the natural scene, human binocular vision system has synchronously various masking effects. In light of this, we propose a new binocular model to integrate these four masking effect factors specially, named SJND. And in SVQA, a full reference assessment algorithm (as shown in Fig.1) based on SJND is also proposed. Luminance contrast, spatial masking, temporal masking and binocular masking are combined to generate original and distorted SJND maps respectively, which are used to calculate a quality of stereoscopic video sequences. Firstly, for each view, the original and distorted videos are united to acquire original JND (TJND) maps. Secondly, for original and distorted videos, left and right views are integrated to obtain two inter-view JND (IJND) map sequences. Thirdly, both views’ TJND maps are combined with the original and distorted IJND maps to derive the original and distorted SJND maps respectively. Finally, the two kinds of SJND maps are computed by the proposed SVQA model to pool the final quality score.

The rest of this paper is organized as follows. The next section presents the SJND model. The third section presents the SVQA algorithm based on the SJND model. The experimental results and discussion are presented in Section 4. Finally, conclusion and future work are given in Section 5.

2. SJND MODEL

SJND mainly consists of temporal JND and inter-view JND. TJND introduces temporal property to the spatial JND model. IJND introduces binocular property to both views’ TJND models. The element of SJND is a classic spatial JND model [5], in which luminance contrast and spatial masking is the two factors that determine the JND threshold of the image. The perceptual model simplifies a very complex process for estimating the visibility threshold of JND, which is described as follow:

\[
JND(x, y) = \max\{f_1(bg(x, y), mg(x, y)), f_2(bg(x, y))\} \tag{1}
\]

where \( f_1(bg(x, y), mg(x, y)) \) and \( f_2(bg(x, y)) \) estimate the luminance contrast and spatial masking effect around the pixel at \((x, y)\), respectively.

\[
f_1(bg(x, y), mg(x, y)) = mg(x, y)\alpha(bg(x, y)) + \beta(bg(x, y)) \tag{2}
\]

through psychovisual experiments[5], the background-dependent function parameters \( \alpha \) and \( \beta \) are expressed as:

\[
\alpha(bg(x, y)) = bg(x, y) \cdot 0.0001 + 0.115 \tag{3}
\]

\[
\beta(bg(x, y)) = \lambda - bg(x, y) \cdot 0.01 \tag{4}
\]

The parameter \( \lambda \) affects the average amplitude of visibility threshold due to spatial masking effect.

\[
f_2(bg(x, y)) = \begin{cases} T_c \cdot (1 - bg(x, y)/127^{0.5}) + 3 & \text{for } bg(x, y) \leq 127 \\ \gamma \cdot (bg(x, y) - 127) + 3 & \text{for } bg(x, y) > 127 \end{cases} \tag{5}
\]

\( T_c \) denotes the visibility threshold when the background grey level is 0, and \( \gamma \) denote the slope of the linear function relating the background luminance to visibility threshold at background luminance level higher than 127.

Generally, as the curve shown in [7], the more luminance and texture difference of the inter-frame, the greater temporal
masking effect. We propose a new temporal JND model called TJND, which is determined by the inter-frame luminance difference, background luminance and texture difference. The resulting TJND is defined as:

\[
TJND(x, y, t) = \max \left\{ f_s(bg(x, y, t), \arg \max_{x, y} \frac{1}{v} \sum_{x, y} (P_x - P_{x+1}), f_s(bg(x, y, t)) \right\}
\]

where

\[
f_s(bg(x, y, t), \arg \max_{x, y} \frac{1}{v} \sum_{x, y} (P_x - P_{x+1}) = \arg \max_{x, y} \frac{1}{v} \sum_{x, y} (Q_x - Q_{x+1})
\]

The psychophysical experiment is designed as follows. The view distance to be three times the picture width and a 64×64 square has been located in the center of both left image and right image with constant gray level G. For the image at each possible gray level G, noise of fixed amplitude A has been randomly added or subtracted to each pixel in the square of right image. However the left image keeps the corresponding gray level G. Therefore, the amplitude of the pixel in the square area of right image was either G+A or G-A, bounded by the maximum and minimum luminance values. The amplitude of the noise has been adjusted from 0 and increased by 1 until the noise was becoming noticeable in the stereoscopic picture. The result is shown in Fig.2:

![Figure 2. the interocular spatial masking effect](image)

From Fig. 2, as a result, a=15, b=5.08, c=0.04 and d=5.08. TJND introduces temporal masking into the traditional spatial JND model, and IJND establishes a binocular masking model. Therefore, through combining the TJND and IJND, SJND is defined as:

\[
SJND(x, y, t) = [TJND(x, y, t)] + \mu [IJND(x, y, t)]
\]

where \(\mu, \eta\) denote the weight to adjust the balance of TJND and IJND. Similarly, for simplification, \(\mu, \eta\) are taken as 0.5 respectively.

### 3. SVQA ALGORITHM

In the proposed full reference SVQA algorithm, there exist three steps:

1. The SJND maps of the original and distorted stereoscopic videos are computed respectively by Eqn. (13).
2. Secondly, the quality maps of the original and distorted SJND maps are calculated by:

\[
q(x, y, t) = \frac{2 \cdot SJND_{orig}(x, y, t) \cdot SJND_{dist}(x, y, t) + \epsilon}{SJND_{orig}(x, y, t) + SJND_{dist}(x, y, t) + \epsilon}
\]

where \(SJND_{orig}(x, y, t)\) and \(SJND_{dist}(x, y, t)\) denote the SJND map at \((x, y)\) of the original and the distorted stereoscopic video respectively, \(\epsilon\) is a positive constant, here we take \(\epsilon = 0.1\).
Finally, the SVQA model pools the quality maps as a quality score:

\[
Q = \sum_{x,y,t} q(x, y, t)
\]  

(15)

The final score considers three kinds of quality of the encoded stereoscopic video, including spatial information distortion, temporal information distortion and binocular information distortion.

4. EXPERIMENTAL RESULTS

To the best of our knowledge, there is no public database for 3D VQA. Therefore, we choose a subjective experiment to evaluate the performance of the proposed model. This subjective experiment has been published in [18]. Fig.3 and Table.1 show the details of the subjective test setting.

<table>
<thead>
<tr>
<th>Stereo video</th>
<th>Poznan Street, Tsinghua Classroom, Balloons, Pantomime</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stereo video encoder</td>
<td>JMVM 2.1</td>
</tr>
<tr>
<td>Frame rate</td>
<td>25 fps</td>
</tr>
<tr>
<td>Frame number</td>
<td>250</td>
</tr>
<tr>
<td>Display Card</td>
<td>NVIDIA GeForce GTS 450</td>
</tr>
<tr>
<td>Display resolution</td>
<td>ViewSonic VX2268wsm</td>
</tr>
<tr>
<td>refresh rate</td>
<td>120 Hz</td>
</tr>
<tr>
<td>Glasses refresh rate</td>
<td>60 Hz</td>
</tr>
<tr>
<td>Subjective test standard</td>
<td>ITU-R BT.500-11</td>
</tr>
<tr>
<td>Test method</td>
<td>single-stimulus (SS)</td>
</tr>
<tr>
<td>observers</td>
<td>18</td>
</tr>
<tr>
<td>Age range</td>
<td>20-35</td>
</tr>
<tr>
<td>Viewing distance</td>
<td>1 m</td>
</tr>
<tr>
<td>Room illumination</td>
<td>Dark</td>
</tr>
</tbody>
</table>

It can be seen from Table.2 that the SJND model outperforms the other metrics in all performance criteria. Meanwhile, Fig.5 shows the proposed metric is in good consistency with the observers' subjective perception.

<table>
<thead>
<tr>
<th>Metrics</th>
<th>CC</th>
<th>SROCC</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>PQM [18]</td>
<td>0.8610</td>
<td>0.8935</td>
<td>0.5638</td>
</tr>
<tr>
<td>PHVS-3D [19]</td>
<td>0.7796</td>
<td>0.7832</td>
<td>0.6943</td>
</tr>
<tr>
<td>SFD [20]</td>
<td>0.6900</td>
<td>0.7049</td>
<td>0.8023</td>
</tr>
<tr>
<td>3D-STS [17]</td>
<td>0.9488</td>
<td>0.9398</td>
<td>0.3500</td>
</tr>
<tr>
<td>SJND model</td>
<td>0.9542</td>
<td>0.9585</td>
<td>0.1332</td>
</tr>
</tbody>
</table>

5. CONCLUSION

This paper proposes a stereoscopic video quality assessment based on a stereo JND model. Through the mimic of human binocular vision system, we suggest to use luminance contrast, spatial masking, temporal masking and binocular masking of stereoscopic video pair to evaluate its quality. Through testing on the previous subjective evaluation database, the experimental results have shown that the proposed model had good performance. Human binocular visual mechanism and stereoscopic video statistical characteristic need to be considered in the future work.

ACKNOWLEDGEMENTS

This work was supported in part by the National Science Foundations of China (61272386, 61103087, 61121002), in part by the Major State Basic Research Development Program of China (973 Program 2009CB320905).
6. REFERENCES