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

Feng Qi[1],Tingting Jiang[2,3], Xiaopeng Fan[1],Siwei Ma[2],Debin Zhao[1]

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, 2012Titanic, 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.

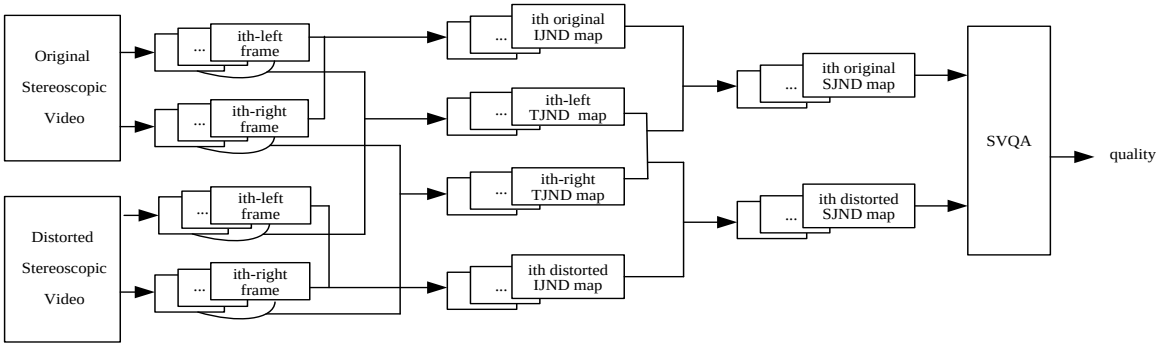

Figure 1. The framework for the proposed 3D VQA scheme based on JND profile.

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, on 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 temporal 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:

$$JND(x,y) = \max\{f_1(bg(x,y), mg(x,y)), f_2(bg(x,y))\} \quad (1)$$

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.

$$f_1(bg(x,y), mg(x,y)) = mg(x,y)\alpha(bg(x,y)) + \beta(bg(x,y)) \quad (2)$$

through psychovisual experiments[5], the background-dependent function parameters α and β are expressed as:

$$\alpha(bg(x,y)) = bg(x,y)\cdot 0.0001 + 0.115 \quad (3)$$

$$\beta(bg(x,y)) = \lambda - bg(x,y)\cdot 0.01 \quad (4)$$

The parameter λ affects the average amplitude of visibility threshold due to spatial masking effect.

$$f_2(bg(x,y)) = \begin{cases} T_0 \cdot (1-(bg(x,y)/127^{1/2})) + 3 \\ \qquad for \quad bg(x,y) \le 127 \\ \gamma \cdot (bg(x,y) - 127) + 3 \\ \qquad for \quad bg(x,y) > 127 \end{cases} \quad (5)$$

T_0 denotes the visibility threshold when the background grey level is 0, and γ 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:

$$TJND(x,y,t) = \max\{f_3(bg(x,y,t), mg(x,y,t)), f_4(bg(x,y,t))\} \quad (6)$$

where

$$f_3(bg(x,y,t), mg(x,y,t)) = \arg\max((P_t - P_{t-1}), \Delta\bar{P}) \quad (7)$$

$$f_4(bg(x,y,t)) = \arg\max((Q_t - Q_{t-1}), \Delta\bar{Q}) \quad (8)$$

P_t and Q_t denote $f_1(bg(x,y), mg(x,y))$ and $f_2(bg(x,y))$ of $JND(x,y)$ at pixel (x,y) in frame $t(t \geq 2)$, respectively. $\Delta\bar{P}$ and $\Delta\bar{Q}$ denote mean difference between the two adjacent frames of the whole video's P and Q. For each view of stereoscopic image, there are corresponding temporal JND which are denoted as $TJND_L(x,y,t)$, $TJND_R(x,y,t)$. According to the experimental results of the reference (right) and auxiliary (left) views from [11][12], TJND is defined as:

$$TJND(x,y,t) = \frac{3}{8}[TJND_L(x,y,t)] + \frac{5}{8}[TJND_R(x,y,t)] \quad (9)$$

According to the discovery that disparity sensitive neurons in the striate cortex of mammals are encoded to perceive stereopsis[16]. While this neural mechanisms are directly represented as binocular fusion and binocular rivalry. Binocular rivalry occurs when dissimilar monocular stimuli are presented to corresponding retinal locations of the two eyes, while in contrast binocular fusion needs similar monocular stimuli. Therefore, we divide each view image of stereoscopic video into occlusion pixels and non-occlusion pixels. Different with [15][17], we use the proposed IJND model instead of the traditional 2D pixel-based JND model to evaluate the interview masking.

For occlusion pixels, based on the distribution of these occlusion pixels focusing on the edge of foreground objects and the concept of HVS meeting Contrast Sensitivity Function (CSF), only the luminance contrast is adopted in their IJND model. Meanwhile, according to the temporal random properties and spatial equivalent properties of binocular rivalry, IJND is defined as:

$$\begin{aligned} IJND_O(x,y,t) = {} & p(t) \cdot f_{3L}(bg(x,y,t), mg(x,y,t)) \\ & + (1 - p(t)) \cdot f_{3R}(bg(x,y,t), mg(x,y,t)) \end{aligned} \quad (10)$$

where $p(t)$ is a random number less than 1 which varies about the time. f_{3L} and f_{3R} are luminance difference of the inter-frame from the left view and the right view. Here, in order to simplify the calculation, we let $p(t)$ is a sawtooth value between [0, 1].

For non-occlusion pixels, both views' temporal masking effects are combined to consider:

$$\begin{aligned} IJND_N(x,y,t) = \max\{ & f_3(bg(x,y,t), mg(x,y,t)), \\ & f_4(bg(x,y,t)), f_5(bg'(x,y,t))\} \end{aligned} \quad (11)$$

where $f_5(bg'(x,y,t))$ represents the left/right view's t -th frame spatial masking under the right/left view's spatial masking effect which is defined as Eqn.(12). A psychophysical experiment is conducted to parameterize the four parameters a, b, c and d.

$$f_5(bg'(x,y)) = \begin{cases} a \times (1 - (bg'(x,y)/127^{1/2})) + b, \\ \qquad bg'(x,y) \leq 127 \\ c \times (bg'(x,y) - 127) + d, \\ \qquad bg'(x,y) > 127 \end{cases} \quad (12)$$

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:

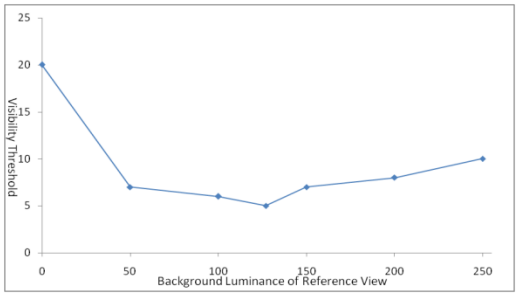

Figure 2. the interocular spatial masking effect

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:

$$SJND(x,y,t) = [TJND(x,y,t)]^{\mu} \cdot [IJND(x,y,t)]^{\eta} \quad (13)$$

where μ, η denote the weight to adjust the balance of $TJND$ and $IJND$. Similarly, for simplification, μ, η are taken as 0.5 respectively.

3. SVQA ALGORITHM

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

Firstly, the SJND maps of the original and distorted stereoscopic videos are computed respectively by Eqn. (13).

Secondly, the quality maps of the original and distorted SJND maps are calculated by:

$$q(x,y,t) = \frac{2 \cdot SJND_{ori}(x,y,t) \cdot SJND_{dis}(x,y,t) + \varepsilon}{SJND_{ori}^2(x,y,t) + SJND_{dis}^2(x,y,t) + \varepsilon} \quad (14)$$

where $SJND_{ori}(x,y,t)$ and $SJND_{dis}(x,y,t)$ denote t -th frame's SJND map value at (x,y) of the original and the distorted stereoscopic video respectively, ε is a positive constant, here we take $\varepsilon = 0.1$.

Finally, the SVQA model pools the quality maps as a quality score:

$$Q = \sum_{x,y,t} q(x,y,t) \tag{15}$$

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. EXPERMENTAL 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.

Figure 3. four sequences of thesubjective test

Table 1. Subjective test equipment and parameters.

Stereo video	Poznan Street, Tsinghua Classroom, Balloons, Pantomime
Stereo video encoder	JMVM 2.1
QP	0, 20, 30, 40, 50
Frame rate	25 fps
Frame number	250
Display Card	NVIDIA GeForce GTS 450
Display	ViewSonic VX2268wm
Display resolution	1680×1050
refresh rate	120 Hz
Glasses	Nvidia® 3D Vision shutter stereo glasses
Glasses refresh rate	60 Hz
Subjective test standard	ITU-R BT.500-11
Test method	single-stimulus (SS)
observers	18
Age range	20-35
Viewing distance	1 m
Room illumination	Dark

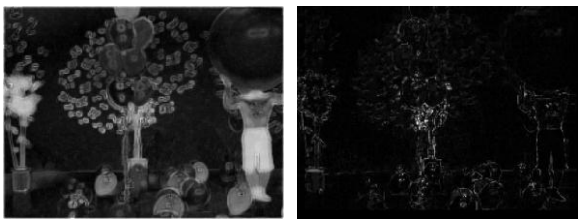

Figure 4. SJNDmap of Balloons

According to the definition of SJND, Fig. 4 shows the TJND and IJND map of the Balloons sequence (left and right view's QP are all equal to 30). In Fig. 4, left image is TJND map which is derived from the first two frames of the left view's sequence. Right image is IJND map which derived from the first frame of left and right views' sequences.

In order to evaluate the performance of the proposed algorithm, we use the nonlinear regression function to transform the proposed metric results, and compare them with our subjective scores in three evaluation criteria: 1) Correlation Coefficient (CC): accuracy of objective metrics; 2) Spearman Rank Order Correlation Coefficient (SROCC): monotonicity of objective metric; 3) Root Mean Square Error (RMSE): offset of objective metric. We choose four previously published 3D quality metrics to compare, including PQM [19], PHVS-3D [20], SFD [21] and 3D-STS[18].The performance comparison result is shown in Table 2.

Table 2. Performance comparison of SVQA metrics

Metrics	CC	SROCC	RMSE
PQM [18]	0.8610	0.8935	0.5638
PHVS-3D [19]	0.7796	0.7832	0.6943
SFD [20]	0.6900	0.7049	0.8023
3D-STS [17]	0.9488	0.9398	0.3500
SJND model	**0.9542**	**0.9585**	**0.1332**

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.

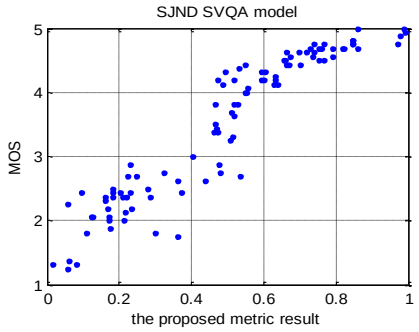

Figure 5. Scatter plot of DMOS versus the SJND model's values

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

[1] "Report on 3DAV exploration", ISO/IEC JTC/SC29/WG11, N5878, July 2003.

[2] Z. Wei, K.N. Ngan, "Spatio-temporal just noticeable distortion profile for greyscale image/video in DCT domain", IEEE Trans. Circuits Syst. Video Technol. vol. 19, no. 3 pp. 337-346,Mar. 2009.

[3] L. Ma, F. Zhang, S. Li, and K.N. Ngan,"Video Quality Assessment based on Adaptive Block-size Transform Just-Noticeable Difference model", in Proc. ICIP2010, pp.2501-2504.

[4] Z. Wang, A.C. Bovik, L. Lu, "Wavelet-based foveated image quality measurement for region of interest image coding", in Proc. ICIP2001, pp. 89-92.

[5] C.-H. Chou and Y.-C. Li, "A perceptually tuned sub-band image coder based on the measure of just-noticeable-distortion profile", IEEE Trans. Circuits Syst. Video Technol, vol. 5, no. 6, pp. 467–476, Dec. 1995.

[6] C.-H. Chou and C.-W. Chen, "A perceptually optimized 3-D sub-band codec for video communication over wireless channels", IEEE Trans. Circuits Syst. Video Technol, vol. 6, no. 2, pp. 143-156, Apr. 1996.

[7] X. Yang, W. Lin, Z. Lu, E. P. Ong, and S. Yao, "Motion-compensated residue preprocessing in video coding based on just-noticeable-distortion profile", IEEE Trans. Circuits Syst. Video Technol, vol. 15, no. 6, pp.742-752, Jun. 2005.

[8] X. Zhang, W. Lin, and P. Xue, "Just-noticeable difference estimation with pixels in images", J. Visual Commun. Image Represent, vol. 19, pp. 30-41, Jan. 2008.

[9] Z.Z. Chen and C. Guillemot, "Perceptually-Friendly H.264/AVC Video Coding Based on Foveated Just-Noticeable-Distortion Model", IEEE Trans. Circuits Syst. Video Technol, vol. 20, no. 6, pp. 806-819, Jun. 2010.

[10] Q.H. Thu, P.L. Callet, M. Barkowsky, "Video quality assessment: From 2-D to 3-D-Challenges and future trends," in Proc. ICIP2010, pp. 4025–4028.

[11] G. Saygh, C.G. Gurler, A.M. Tekalp, "Quality Assessment of Asymmetric Stereo Video Coding", in Proc. ICIP2010, pp.4009-4012.

[12] P. Aflaki, M.M. Hannuksela, J. Hakkinen, "Subjective Study on Compressed Asymmetric Stereoscopic Video",in Proc. ICIP2010, pp.4021-4024.

[13] Y. Zhao, Z. Chen, C. Zhu, Y.Tan, and L. Yu, "Binocular just noticeabledifference model for stereoscopic images", IEEE Signal Processing Letters, Vol. 18, No. 1, pp. 19-22, Jan. 2011.

[14] D. De. Silva, W. A. C. Fernando, S. T. Worrall, S. L. P. Yasakethu, and A. M. Kondoz, "Just noticeable difference in depth model for stereoscopic 3D displays", in Proc. ICME2010, pp. 1219-1224.

[15] X. Li, Y. Wang, D. Zhao, T. Jiang, and N. Zhang,"Joint just noticeable difference model based on depth perception for stereoscopic images", in Proc. VCIP2011, pp.1-4.

[16] D. J. Fleet, H. Wagner, D. J. Heeger. "Neural encoding of bincoular disparity: energy models, position shifts and phase shifts",J.Vision research, Vol. 36, No. 12, pp. 1839-1857, Jun. 1996.

[17] X. Wang, S. Kwong and Y. Zhang, "Considering binocularspatial sensitivity in stereoscopic image quality assessment",in Proc. VCIP2011, pp. 1-4.

[18] J. Han, T. Jiang, S. Ma, "Stereoscopic Video Quality Assessment Model Based on Spatial-Temporal Structural Information", in Proc. VCIP2012, pp. 119-125.

[19] P. Joveluro, H. Malekmohamadi, W. A. C. Fernando and A. M. Kondoz, "Perceptual video quality metric for 3D video quality assessment",in Proc. 3DTV-CON2010, pp. 1-4.

[20] L. Jin, A. Boev, A. Gotchev and K. Egiazarian, "3D-DCT based perceptual quality assessment of stereo video",in Proc. ICIP2011, pp. 2521-2524.

[21] F. Lu, H. Wang, X. Ji and G. Er, "Quality assessment of 3D asymmetric view coding using spatial frequency dominance model",in Proc. 3DTV-CON2009, pp.1-4