Quality of Experience Assessment for Stereoscopic Images

Feng Qi[1], Tingting Jiang[2,3], Siwei Ma[2,3], 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
fqi@jdl.ac.cn, ttjiang@pku.edu.cn, swma@jdl.ac.cn, dbzhao@jdl.ac.cn,

***Abstract*—Stereoscopic image quality assessment has been widely studied in last decades; however, the research on 3D quality of experience (QoE) is proposed recently. As a part of human stereo perception, 3D QoE plays an important role to stereoscopic image quality assessment. In this paper, an objective metric is proposed based on the hypothesis that binocular vision system is sensitive to the structure of low-level features and its discrepancy between the two view images of a stereoscopic image pair. Specifically, the correlation between the left and right views of a stereoscopic image pair could reflect the QoE. To represent the structure of low-level features, in each view of the stereoscopic image pair, the phase congruency (PC) and the saliency map are employed as the primary and secondary features to compose a feature map. To compute the correlation between the two views, a local matching function is suggested to weight the discrepancy between the two feature maps and generate a local quality. Then these local quality values are combined to derive a single quality score. The proposed metric is evaluated on one public subjective assessment database. The experimental results indicate that our metric exhibits good performance.**

I. INTRODUCTION

In the middle of last century, stereoscopic videos were at the height of their popularity, with realistic experience of stereoscopic feature films produced by Hollywood. However, unlike its initial success, broader acceptance of stereoscopic video has been hampered in a short time. One of the most important reasons was the safety and health issues such as visual discomfort and visual fatigue [1]. Therefore, study of stereoscopic image quality is significant to the development of 3D techniques, including capture, encoding, transmission and display. Different from monoscopic images, a stereoscopic image projected into our retinas are two slightly different images, because our eyes are separated about 6.5 cm. The disparity between the left and right retinal images is one important cue that constructs the sensation of depth.

From the viewpoint of human stereo perception, the quality of stereoscopic images is affected not only by the degree of distortion of the two images, but also the experience of binocular perception. The studies of stereoscopic image quality assessment have been studied for decades. Most of them focus on the correlation of stereoscopic image distortion and binocular perception, such as asymmetric assessment [2,3] and Just Noticeable Difference (JND) threshold measurement [4,5]. However, these assessment methods are mainly extension from 2D quality assessment; some methods which study safety and health issues related to stereo display are the beginning of 3D quality of experience (QoE). Reference [6] showed that disparity magnitude and disparity switch is more important in determining visual comfort. Although some people study stereo display issues for visual safety and health, the research of 3D QoE is still lacking. Especially, few of these above methods take into account the natural property between two views of a stereoscopic image pair. The natural property includes the statistical characteristic of stereoscopic image, such as global similarity and local discrepancy, which make it essential for QoE. [8] proposed a no-reference algorithm to assess the comfort associated with viewing stereo images and videos which is the first attempt to algorithmically assess the subjective QoE on a publicly available dataset [7].

As a mental and psychology act [9][10], the completely same features between stereoscopic image pair such as edge, flat and zero-crossing are regarded as one feature, while the slight discrepancy between stereoscopic image pair are fused to produce stereo sensation. In this higher cognition process, binocular fusion is the fundament of human apperception of stereoscopic images. Therefore, different from [8] which assumes that natural 3D images have certain 'natural' statistical properties, we believe that the process of human 3D experience should include binocular feature extraction and matching. Specifically, the phase congruency (PC) and saliency map are employed as two low-level features to form a perceptual feature map for each view of the stereoscopic image pair. The consequent correlation between the left and right feature maps could be used to evaluate 3D QoE. The proposed 3D QoE framework is shown in Figure. 1.

The rest of this paper is organized as follows. Section II describes the binocular feature extraction algorithm in detail. Section III represents the discrepancy calculation. Section IV reports the experimental results conducted on the database [7]. Finally the conclusion is drawn in Section V.

978-1-4673-0219-7/12/$31.00 ©2012 IEEE

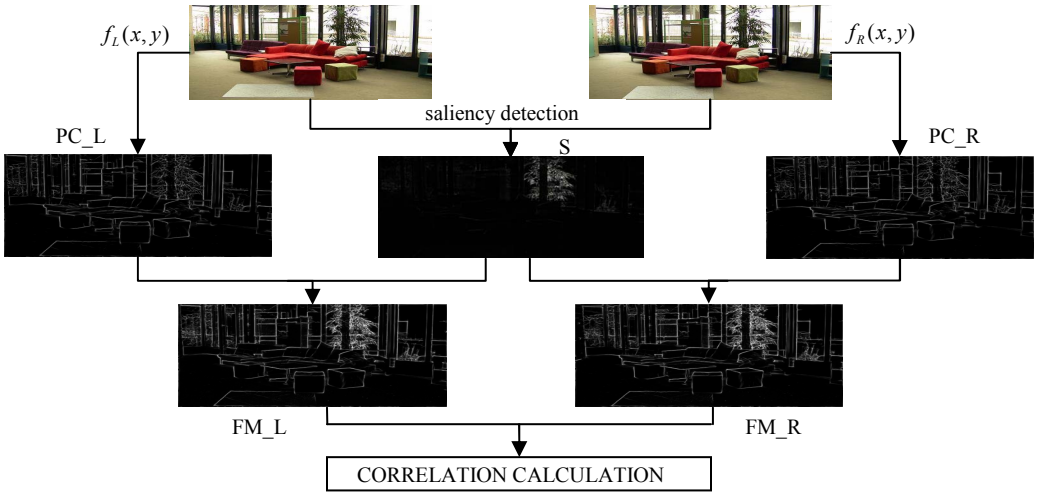

Figure 1. The process of 3D QoE. Firstly, given the left view of a stereoscopic image $f_L(x,y)$, the left PC map PC_L is computed. Similarly, the right PC_R is computed for $f_R(x,y)$. Secondly, the saliency map (S) is obtained through saliency detection of the stereoscopic image pair. Next, feature maps (FM_L, FM_R) of the left and right views are generated by combining PC_L and PC_R with SM, respectively. Finally, the correlation between the two views is used to evaluate 3D QoE.

II. BINOCULAR FEATURE EXTRACTION

According to the physiological and psychophysical evidences, human vision system (HVS) understands an image mainly based on its low-level features, such as edges and zero crossings [9]. In addition, due to the physical structure of the human binocular vision, these low-level features between a stereoscopic image pair are expected to follow structure congruency. Therefore, in order to mimic the human's perception of viewing stereoscopic images, we propose a binocular feature extraction model. The model generates feature maps for stereoscopic images by phase congruency calculation and saliency detection.

A. Phase Congruency

Under the definition of phase congruency (PC) in [11], PC can be expressed as a dimensionless quantity that keeps invariance to some changes in image, such as brightness or contrast. The PC theory provides a simple but biologically plausible model of how mammalian visual systems detect and identify features in an image. Hence, it provides an absolute measure of the significance of a local feature structure.

In this paper, we adopt the method developed by Kovesi Peter[12], which is widely used in literature. For a stereoscopic image, we compute the PC of each view's image respectively. According to [11], the PC of spreading function would preserve stability after being smoothed with Gaussian. Thus, using Gaussian as the spreading function, there exists a 2-D log-Gabor function which has the following transfer function:

$$G(\omega,\theta)=\exp\left(-\frac{(\log(\omega/\omega_0))^2}{2\sigma_r^2}\right)\cdot\exp\left(-\frac{(\theta-\theta_j)^2}{2\sigma_\theta^2}\right) \quad (1)$$

where ω and θ are the filter's frequency and angular respectively. ω_0 is the filter's center frequency, and σ_r^2 controls the filter's bandwidth. $\theta_j = j\pi/J$, $j=\{0,1,...,J-1\}$ is the orientation angle of the filter, J is the number of the orientations, and σ_θ determines the filter's angular bandwidth. From the 2-D image signal $f(x,y)$, a set of responses at each point (x,y) can be denoted by a quadrature pair include the even-symmetric filters M^e_{n,θ_j} and odd-symmetric filters M^o_{n,θ_j} on scale n and orientation θ_j [13]. They form a set of response vector:

$$[e_{n,\theta_j}(x,y),o_{n,\theta_j}(x,y)]=[f(x,y)*M^e_{n,\theta_j},f(x,y)*M^o_{n,\theta_j}] \quad (2)$$

The local amplitude on scale n is:

$$A_{n,\theta_j}(x,y)=\sum_j\sqrt{e^2_{n,\theta_j}(x,y)+o^2_{n,\theta_j}(x,y)} \quad (3)$$

and the local energy along orientation θ_j is:

$$E_{n,\theta_j}(x,y)=\sqrt{(\sum_n e_{n,\theta_j}(x,y))^2+(\sum_n o^2_{n,\theta_j}(x,y))^2}. \quad (4)$$

The 2-D image PC at (x,y) is defined as below:

$$PC[f(x,y)]=\frac{\sum_j E_{n,\theta_j}(x,y)}{\varepsilon+\sum_n A_{n,\theta_j}(x,y)}. \quad (5)$$

B. Binocular Saliency Detection

It is very important in stereoscopic image assessment to establish a complete binocular saliency model with stereo visual characteristics. According to the analyzing the log-spectrum of a given image [13], we propose a spectral difference approach to detect the binocular saliency. The spectral difference of the stereoscopic image in the log spectrum domain denotes its variation, which can be used to obtain the binocular saliency map.

Consider the left frame of stereoscopic image $f_L(x,y)$. Let $AF_L(u,v)$ and $PF_L(u,v)$ denote the amplitude and phase

spectrum of the Fourier Transform, respectively. The log spectrum $LS_L(u,v)$ can be computed as follows:

$$LS_L(u,v) = \log(AF_L(u,v)) . \quad (6)$$

Similarly, through the amplitude $AF_R(u,v)$ and phase spectrum $PF_R(u,v)$ of the Fourier Transform from the right frame, the log spectrum $LS_R(u,v)$ can be obtained. Therefore the spectral difference $D(u,v)$ can be denoted by:

$$D(u,v) = \left| LS_L(u,v) - LS_R(u,v) \right| . \quad (7)$$

In the model, the spectral difference contains the occlusion of the stereoscopic image, which should be paid more visual attention in a stereo scene. By Inverse Fourier Transform, the output image called the binocular saliency map $S(x,y)$ can be constructed in spatial domain.

$$S(x,y) = g(x,y) * \Im^{-1}[\exp(D(u,v) + \frac{1}{2}(PF_L(u,v) + PF_R(u,v)))]^2 , \quad (8)$$

where $\Im^{-1}$ denotes the Inverse Fourier Transform. $g(x,y)$ is a Gaussian filter to smooth the binocular saliency map for better visual effects.

From the viewpoint of visual attention, binocular saliency detection model here is used to detect the representative regions attracting our attention according to the stereo perception, and phase congruency model depicts another type of the representative regions attracting our attention according to the content. So combining PC and saliency detection is necessary to evaluate the quality of stereoscopic image. For simplification, here we let PC maps from left and right views plus the saliency map respectively, i.e.,

$$FM_L = PC_L + S , \quad FM_R = PC_R + S . \quad (9)$$

III. CORRELATION CALCULATION

From the physiology and psychology evidence [10], our brains combine both retina images to generate one perception image. According to the characteristic of stereoscopic image pair, there exists slight discrepancy between left and right images. However, human binocular vision system is sensitive to these differences. The highly similar content of stereoscopic image pairs makes discrepancy calculation difficult. To solve this problem, computing the correlation of the two feature maps from left and right views is chosen, because it simplifies the complicated scene and quantifies the discrepancy. Therefore, the correlation of the two generated feature maps above is used to evaluate the quality of stereo experience.

The proposed correlation calculation predicts the quality of a stereoscopic image by the following three steps.

Step 1: Compute the local quality map. The details are given in A.

Step 2: Pool the local quality score to the global quality, which will be described in B.

Step 3: Calculate the final quality score by adopting linar normalization. This process will be given in C.

A. Local matching function

According to the hypothesis that 3D QoE correlates with the feature and its discrepancy of the stereoscopic image, we adopt normalized cross-correlation method [15] to quantify the correlation between the two feature maps. The degree of correlation can be decomposed as the accumulation of the local similarity factor. Therefore, the local quality of two corresponding blocks centered at the location (u,v) in both views' feature maps can be calculated by a local matching function. The proposed function can be described as:

$$f(u,v) = \frac{\sum_{x=u}^{u+U+1}\sum_{y=v}^{v+V+1} DFM_L \cdot DFM_R}{\sqrt{\sum_{x=u}^{u+U+1}\sum_{y=v}^{v+V+1} DFM_L^2 \cdot \sum_{x=u}^{u+U+1}\sum_{y=v}^{v+V+1} DFM_R^2}}, \quad (10)$$

where

$$DFM_L = FM_L(x,y) - \overline{FM_L(u,v)}, DFM_R = FM_R(x,y) - \overline{FM_R(u,v)}. \quad (11)$$

$\overline{FM_L(u,v)}$ and $\overline{FM_R(u,v)}$ are the mean value of the block $U \times V$ from the left feature map and the right feature map with the center (u,v).

B. Global convergence

Global convergence pools the local quality as a score to represent a global binocular fusion quality. It is expressed as:

$$Q = \frac{\sum_{u=1}^{u=\lfloor M/U \rfloor}\sum_{v=1}^{v=\lfloor N/V \rfloor} f(u,v)}{\lfloor M/U \rfloor \times \lfloor N/V \rfloor} \quad (12)$$

where $M \times N$ is the size of the image. $\lfloor \bullet \rfloor$ is the rounding sign. Here, the 3D QoE score of the stereoscopic image is obtained.

C. Linear normalization

In order to express the final score in the range of [0,1], linear normalization is adopted. It can be described as:

$$P_i = \frac{(Q_i - \sum_{j=1}^{n} \frac{Q_j}{n}) + (Q_{\max} - Q_{\min}) + c}{2(Q_{\max} - Q_{\min}) + c} \quad (13)$$

where n is the total number of stereoscopic images, Q_i is the quality's value of the i-th stereoscopic image, $Q_{\max}$ and $Q_{\min}$ are the max and min value of the results respectively. c is a positive constant which is used to increase the stability of P_i.

IV. EXPERIMENTAL RESULTS

To the best of our knowledge, there is only one public database that can be used in 3D QoE research community. The dataset has recently been made public by researchers at EPFL [7]. It includes two parts, one is for images and the other is for

videos. Here, we only choose the image part. The EPFL 3D image database contains 10 scenes which has 100 stereoscopic images. Each scene includes 10 stereoscopic images with different camera distances and the image resolution is 1920× 1080 pixels. The camera distances vary in the range 10 – 60 cm. According to the subjective experiment of EPFL, 54 stereoscopic images are tested by 17 people. It contains 9 scenes as seen in Figure. 2. Each scene has 6 different camera distances. These subjective test scores are finally represented as 54 MOS of QoE.

Figure 2. Typical frames of stereoscopic images.

In this paper, we choose these 54 pictures as our test set, which can be downloaded at [16]. Through the binocular feature extraction and correlation calculation, the score of QoE can be obtained. For the 9 different scenes, we compute each scene's CC and then take the average across these 9 scenes.

The performance of the proposed model can be indicated by several performance metrics, including Pearson correlation coefficient (CC), Spearman Rank Order Correlation coefficient (SROCC). The evaluation results are summarized and compare to [8] in Table I. Because [8] does not provide its evaluation result of CC, here we only list its SROCC of principal component analysis (PCA) and forward feature selection (FFS) methods. Figure 3 shows that the proposed model is in good consistency with the observers' subjective perception. In addition, our method does not need training process while [8] requires to learn parameters for its model. Note that we are not sure how [8] computes the mean and standard deviation in details, because they might need different combinations of training and testing data while we don't have this process.

V. CONCLUSION

This paper proposes a novel approach for 3D QoE assessment. Through the mimic of human binocular vision system, we suggest to use binocular feature extraction and correlation calculation of a stereoscopic image pair to evaluate its quality. The proposed method is applied on the EPFL 3D image database. The experimental results show that the proposed model had good performance. More future work about feature selection mechanisms needs to be considered based on binocular human visual system.

VI. ACKNOWLEDGEMENTS

This work was supported in part by the National Science Foundations of China (60736043, 60821003, 60833013), in part by the Major State Basic Research Development Program of China (973 Program 2009CB320905).

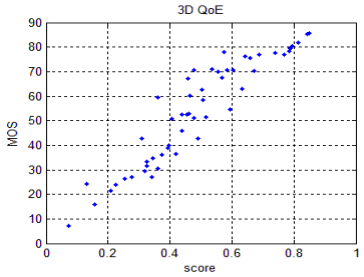

Figure 3. The Scatter plots of MOS versus our score. The 54 subjective MOS are listed in the y-axis and our scores are listed in the x-axis.

TABLE I. PERFORMANCE EVALUATION

Evaluation	SROCC		SROCC	CC
Method	*PCA*	*FFS*	*Our method*	*Our method*
Mean	0.79	0.86	0.96	0.93
Standard deviation	0.08	0.11	0.08	0.08

REFERENCES

[1] M. Lambooij, W. IJsselsteijn, M. Fortuin, and I. Heynderickx, "Visual discomfort and visual fatigue of stereoscopic displays: A review," Journal of Imaging Science and Technology, vol. 53, no. 3, pp. 030201/1–14, June 2009.

[2] G. Saygili, G. Gurler, A.M. Tekalp, "Quality assessment of asymmetric stereoscopic video coding," ICIP, pp:4009-4012, September 2010.

[3] F. Lu, H. Wang, X. Ji, G. Er, "Quality Assessment of 3D Asymmetric View Coding Using Spatial Frequency Dominance Model", Proc. IEEE 3DTV Conference, Potsdam, Germany, May 2009.

[4] D.V.S.X. De Silva, W.A.C. Fernando, G. Nur, E.Ekmekcioglu and S.T. Worrall, "3D video assessment with just noticeable difference in depth evaluation,"ICIP, pp. 4013-4016, September 2010.

[5] Y. Zhao, Z.Z. Chen, C. Zhu, Y.P. Tan, and L. Yu, "Binocular just-noticeable-difference model for stereoscopic images," IEEE Trans. Signal Processing, vol. 18, no. 1, pp.19-22, January 2011.

[6] F. Speranza, W. J. Tam, R. Renaud, and N. Hur, "Effect of disparity and motion on visual comfort of stereoscopic images," Proc. SPIE vol. 6055, no. 60550B, pp.94-103, January 2006.

[7] L. Goldmann, F. De Simone, and T. Ebrahimi, "A comprehensive database and subjective evaluation methodology for quality of experience in stereoscopic video," EI, 3DIP and Applications, 2010.

[8] Mittal,A., Moorthy,A.K., Ghosh,J., Bovik, A.C., "Algorithmic assessment of 3D quality of experience for images and videos," DSP/SPE, 2011 IEEE

[9] M.C.Morrone and D.C.Burr, "Feature detection in human vision: A phase-dependent energy model," Proc. R. Soc. Lond. B, vol. 235, no.1280, pp. 221–245, Dec. 1988.

[10] I. P. Howard and B. J. Rogers, "Binocular Vision and Stereopsis," Oxford University Press, ISBN 0-19-508476-4, 1995.

[11] M. C. Morrone, J. Ross, D. C. Burr, and R. Owens, "Mach bands are phase dependent,"Nature,vol.324,no.6049,pp.250–253,Nov.1986.

[12] P. Kovesi, "Image features from phase congruency," *Videre: J. Comp. Vis. Res.*, vol. 1, no. 3, pp. 1–26, 1999.

[13] L. Zhang, L. Zhang, X. Mou, andD. Zhang. "FSIM: a feature similarity index for image quality assessment", IEEE TIP, vol. 20, pp. 2378-2386, Aug. 2011.

[14] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. IEEE Conference on Computer Vision and Pattern Recognition, 2007

[15] J. Lewis. Fast normalized cross-correlation. In Proc. of Vision Interface, 1995

[16] http://mmspg.epfl.ch/3diqa