A Stereo Video Quality Assessment Method For Compression Distortion

Hong Zhu, Mei Yu, Yang Song, Gangyi Jiang
Faculty of Information Science and Engineering
Ningbo University
Ningbo, 315211 China
jianggangyi@126.com

***Abstract*—Based on feature analysis of video compression distortion, a new stereo video quality assessment (SVQA) method for compression distortion is proposed from temporal characteristics of video and binocular perception in this paper. First of all, significant pixels of monocular video are determined, and a new definition of region-of-interest (ROI) is given by using significant pixels and just noticeable distortion model. Secondly, taking ROIs and temporal perception characteristics of video and depth perception of stereo views into account, two quality metrics, left and right views' quality metric and depth perception metric, are designed. Finally, the two quality metrics are pooled to obtain final SVQA score. Experimental results with tested in the NAMA3DS1 - CoSpaD1 stereo video database show that the proposed method is quite efficient for evaluating stereo video quality, its linear correlation coefficient (LCC) is above 0.92, and its evaluating results are more consistent with subjective perception.**

Keywords—Stereo Video; Stereo Video Quality Assessment; region-of-interest; depth perception

I. INTRODUCTION

Recently, the rapid technological developments of three dimensional (3D) image/video processing have led to an explosion in consumer demand for 3D content [1-3]. 3D video and imaging technologies of creating the illusion of depth present new advances in video areas by offering viewers a stunning 3D experience. Stereo video, as an important expression form of 3D video, is captured at the same time from two different viewpoints to simulate the perspectives of the right and left human's eyes [4]. Stereo video quality plays an important role in assessing 3D video coding, transmission, and applications.

The signal fidelity measures refer to the traditional mean absolute error (MAE), mean square error (MSE), signal-to-noise ratio (SNR), PSNR (peak SNR), or one of their relatives [5]. Two other representative quality metrics are structural similarity index (SSIM) [6] and video quality measurement (VQM). Some researches on 3D/stereo video quality assessment have been reported. Classical methods are usage of 2D objective video quality metrics to evaluating 3D/stereo video quality [7-10]. Yasakethu et al. [7] had compared these methods with PSNR, SSIM and VQM, their results show that, by measuring left and right views separately, VQM can predict the overall image quality, and PSNR and SSIM results correlate better with depth perception of 3D video in comparison with VQM. Ozbek et al. used different weights to the PSNR scores of the left and right views [8]. Hewage et al. further analyzed more specific depth map based stereo video quality assessment in [9]. Although they are simple, well defined, with clear physical meanings and widely accepted, they can be a poor predictor of perceived visual quality, especially the noise is not additive [11]. Han et al. proposed a no reference objective video quality metric for 3D video quality using network packet loss rate and bit rate as input [12]. Jin et al. present a stereoscopic video quality assessment method based on 3D-DCT transform, in which similar blocks from left and right views of stereo video frames are found by block-matching, grouped into 3D stack and then analyzed by 3D-DCT [13].

In the previous researches, some important factors, such as binocular perception, temporal features of video, had not been considered sufficiently. In this paper, a compression distortion -oriented stereo video quality assessment method is proposed by combining with temporal features and binocular perception. Significant pixels of video are extracted, and region of interest are obtained by using significant pixels and just noticeable distortion (JND). Then, left and right views' quality metric and depth perception metric of stereo video are designed to obtain the final SVQA score. Experimental results show that the proposed method is quite efficient for evaluating stereo video quality, its linear correlation coefficient (LCC) is above 0.92, and its evaluating results are more consistent with subjective perception.

II. PROPOSED COMPRESSION DISTORTION ORIENTED STEREO VIDEO QUALITY ASSESSMENT METHOD

To take temporal features of video and binocular perception, a compression distortion -oriented stereo video quality assessment method is proposed, shown in Fig. 1. The proposed method consists of two metrics, that is, left-right views' quality metric and depth perception quality metric of stereo video. In the first metric, significant pixel is defined, ROI of image is extracted by using significant pixels and JND, perceptual weights is determined and temporal fusion strategy is given, and Left and right view's quality is computed. In the second metric, depth perception quality metric of stereo video

978-1-4673-9795-7/15 $31.00 © 2015 IEEE
DOI 10.1109/CSCI.2015.18

is computed by using disparity space image and three- dimensional discrete wavelet transform.

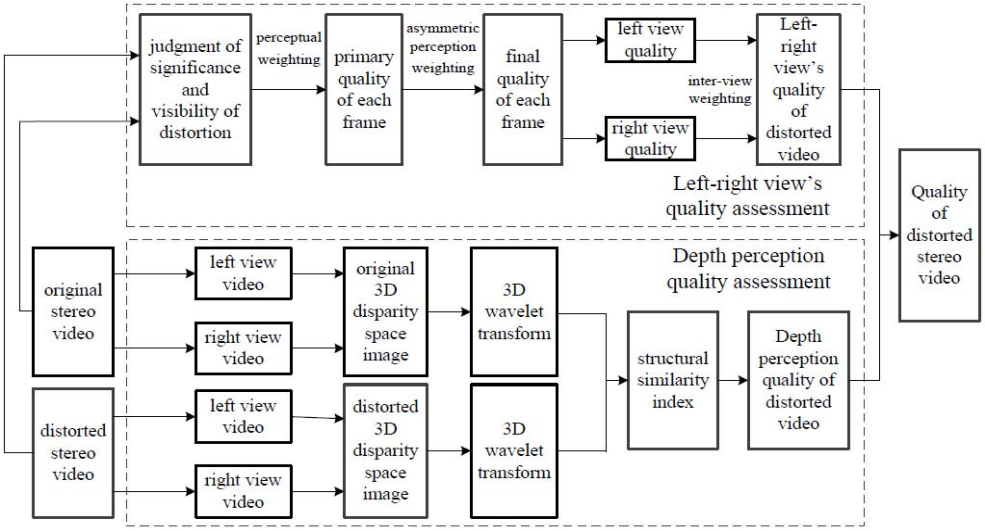


Fig. 1. Scatter plots between different methods and MOS.

A. Left-right views' quality metric of stereo video

1) Extraction of significant map and ROI

Let $\boldsymbol{Y}_{org}$ and $\boldsymbol{Y}_{dis}$ denote the original and distorted stereo videos, respectively, $\boldsymbol{Y}_{org}$ ={ $\boldsymbol{Y}^n_{org}$; n=0,1,…,} and $\boldsymbol{Y}_{dis}$={ $\boldsymbol{Y}^n_{dis}$; n=0, 1, … N} , n is an integer and N is the total number of stereo frame pairs of stereo video. $\boldsymbol{Y}^n_{org}$ and $\boldsymbol{Y}^n_{dis}$ denote the n-th stereo frame pairs of the original and distorted stereo videos, respectively. Here, $\boldsymbol{Y}^n_{org}$ ={ $\boldsymbol{L}^n_{org}$, $\boldsymbol{R}^n_{org}$ } and $\boldsymbol{Y}^n_{dis}$ ={ $\boldsymbol{L}^n_{dis}$, $\boldsymbol{R}^n_{dis}$ }, $\boldsymbol{L}^n_{org}$ and $\boldsymbol{R}^n_{org}$ are the left and right views of the n-th original stereo frame pair, similarly, $\boldsymbol{L}^n_{dis}$ and $\boldsymbol{R}^n_{dis}$ are the left and right views of the n-th distorted stereo frame pair.

Edge information is important feature in image or video. Here, horizontal gradient image of $\boldsymbol{L}^n_{org}$ and $\boldsymbol{L}^n_{dis}$ are computed by using Sobel operator, denoted as $\boldsymbol{G}^n_{Lh,org}$ and $\boldsymbol{G}^n_{Lh,dis}$. Similarly, the vertical gradient image of $\boldsymbol{L}^n_{org}$ and $\boldsymbol{L}^n_{dis}$ are obtained as $\boldsymbol{G}^n_{Lv,org}$ and $\boldsymbol{G}^n_{Lv,dis}$. Then, the gradient image, $\boldsymbol{G}^n_{L,org}$, of $\boldsymbol{L}^n_{org}$ is defined by

$$G^n_{L,org}(i,j)=\sqrt{\left(G^n_{Lh,org}(i,j)\right)^2+\left(G^n_{Lv,org}(i,j)\right)^2} \qquad (1)$$

where $\boldsymbol{G}^n_{Lh,org}$ (i,j) or $\boldsymbol{G}^n_{Lv,org}$ (i,j) is the value of $\boldsymbol{G}^n_{Lh,org}$ or $\boldsymbol{G}^n_{Lv,org}$ at the position of (i,j), respectively. In the same way, the gradient image, $\boldsymbol{G}^n_{L,dis}$, of $\boldsymbol{L}^n_{dis}$ is defined also.

If $\boldsymbol{G}^n_{L,org}$ (i,j) is smaller than a threshold T, the pixel at (i,j) is defined as a significant pixel in $\boldsymbol{G}^n_{L,org}$, thus, these significant pixels are constructed into a significant map of $\boldsymbol{G}^n_{L,org}$, denoted as $\boldsymbol{S}^n_{L,org}$. Likewise, the significant map of $\boldsymbol{G}^n_{L,dis}$ is defined as $\boldsymbol{S}^n_{L,dis}$. Let $\boldsymbol{X}^n$={$\boldsymbol{X}^n(i,j)$} be the absolute difference image between { $\boldsymbol{S}^n_{L,org}(i,j)$} and { $\boldsymbol{S}^n_{L,dis}$ (i,j)}, $\boldsymbol{X}^n(i,j)$= | $\boldsymbol{S}^n_{L,org}$ (i,j) - $\boldsymbol{S}^n_{L,dis}$ (i,j) |.

It is known that human's eye is sensitive to distorted regions. However, when the distortion is less than just noticeable distortion (JND) threshold around the pixel, it will not be perceived. JND simulates luminance contrast and spatial-temporal masking in human visual system (HVS) [14], human eyes cannot distinguish any changes below JND threshold around a pixel. Here, spatial JND model and the absolute difference image between the original and distorted images are used to determine if distorted significant pixel is visible or not. The region composed of significant pixels with visible distortion is defined as a region-of-interest (ROI). Let $J_S(i,j)$ be the spatial JND at (i,j), and $J_b(i,j)$ and $J_t(i,j)$ be the visibility thresholds at (i,j) for the two primary masking factors, background luminance adaptation and texture masking, respectively; and C denotes the overlapping effect in masking. A spatial JND model is computed by $J_s(i,j)=J_b(i,j)+J_t(i,j)-C\min\{J_b(i,j),J_t(i,j)\}$.

Larger JND value implies that human's eye can tolerate more image distortion. For a pixel at (i,j), if its $X^n(i,j)$ value is larger than its $J_S(i,j)$, the pixel at (i,j) is defined as a visible-distortion pixel (VDP). On the contrary, the pixel is an invisible-distortion pixel (IDP). This process is defined by the function $f()$ as follows

$$f(i,j)=\begin{cases}1 & if\ X(i,j)>J_s(i,j)\\ 0 & otherwise\end{cases} \qquad (2)$$

In (2), '1' or '0' implies that the pixel at (i,j) is a VDP or IDP, respectively. All VDPs in $\boldsymbol{L}^n_{org}$ construct a ROI map, denoted as $\boldsymbol{I}^n_{org}$, in the same way, the ROI map of $\boldsymbol{L}^n_{dis}$ is defined as $\boldsymbol{I}^n_{dis}$.

2) Determination of perceptual weights

It is known that the degree of human eye's visual interest to the ROI of image is approximately inverse-proportion to the ROI's area. When the ROI's area is quite small, human eye may be more sensitive to the distortion in ROI region. While the ROI's region is sufficiently large, human eye attention may move pay to non-ROI. Based on this phenomenon, perceptual weights are designed. Let ω_1 and ω_2 be perceptual weights of ROI and non-ROI regions, then, ω_1 is computed by $\omega_2=1-2\sqrt{S_1S_2}/(S_1+S_2)$ and $\omega_1=\omega_2+(S_1+S_2)(1-\omega_2)/S_1$, where S_1 and S_2 are the regions of ROI and non-ROI. Then, with normalization of the perceptual weights ω_1 and ω_2, the normalized perceptual weight, λ, is computed by $\lambda=1/(\omega_1+\omega_2)$, the perceptual quality of ROI with SSIM metric is calculated by

$$ssim_{rs}=\begin{cases}\lambda\times ssim_s & if\ f=1\\ ssim_s & otherwise\end{cases} \quad (3)$$

where $ssim_s$ denotes SSIM value of ROI in reference video frame, and $ssim_{rs}$ denotes the revised SSIM value. For a frame in a video, its quality score, denoted as $q(n)$, is computed by all pixels in the ROIs, and represented by $q(n)=\Sigma ssim_{rs}/M$, where M denotes the number of pixels in ROI.

3) Temporal fusion

Human visual system is thought to separate visual information into two temporal channels: a lowpass channel with an approximate cutoff around 10 Hz and a bandpass channel with a peak frequency of 15 Hz and an approximate bandwidth [15]. According to HVS, there exists in smooth effect for human visual perception on change of consecutive frames' qualities. In addition, there is another asymmetric visual perceptual feature on human being, for which human visual perception is more sensitive to quality reduction than quality improvement. Hence, an asymmetric perception model with lowpass part is used to pre-process frame quality of video and expressed by

$$q'(n)=\begin{cases}q(n-1)+a_{-}g(n) & if\ g(n)\le 0\\ q(n-1)+a_{+}g(n) & if\ g(n)>0\end{cases} \quad .(4)$$

where $q(n)$ or $q'(n)$ is the incipient or processed n-th frame quality of a monocular video, respectively. $g(n)$ = q(n)−q(n−1), and a- and a+ are coefficients, a-<a+.

If $G^n_{L,org}(i,j)$ is smaller than a threshold T, the pixel at (i,j) is defined as a significant pixel in $\boldsymbol{G}^n_{L,org}$, thus, these significant pixels are constructed into a significant map of $\boldsymbol{G}^n_{L,org}$, denoted as $\boldsymbol{S}^n_{L,org}$. Likewise, the significant map of $\boldsymbol{G}^n_{L,dis}$ is defined as $\boldsymbol{S}^n L,dis$. Let $\boldsymbol{X}^n=\{\boldsymbol{X}^n(i,j)\}$ be the absolute difference image between $\{\boldsymbol{S}^n_{L,org}(i,j)\}$ and $\{\boldsymbol{S}^n_{L,dis}(i,j)\}$, $\boldsymbol{X}^n(i,j)=|\boldsymbol{S}^n_{L,org}(i,j)-\boldsymbol{S}^n_{L,dis}(i,j)|$.

Eq. (4) embodies the asymmetry of human eye perception on video image quality. For the left view of stereo video, its quality, Q_L, is defined by $Q_L=\Sigma q'(n)/N_f$, where N_f denotes the number of frames in the left view of stereo video. Similarly, the right view quality, Q_R, of stereo video is computed in the same way.

4) Left and right view's quality metric

Let Q_{LR} be the fusion of left view's quality Q_L and right view's quality Q_R, w_{LR} be a weight, then Q_{LR} is expressed by

$$Q_{LR}=w_{LR}\times Q_L+(1-w_{LR})\times Q_R \quad .(5)$$

B. Left-right view's quality metric of stereo video

Due to the slight difference in the retina of left and right eyes, it will produce a perception of the binocular depth after the process of the brain, and it is also called stereo vision. Here, to simplify the calculation, disparity space image (DSI) [16] is used to de-scribe depth perception of stereo video. Three-dimensional wavelet transform (3D-DWT) is used to extract the mid- and low- frequency characteristics of DSI, which are used to depth perception evaluation of stereo image with the help of SSIM metric. Finally, a temporal pooling strategy is used to get the depth perception evaluation of stereo video. The proposed depth perception evaluation model is shown as a part of Fig. 1. Let $\boldsymbol{D}^n_{org}$ and $\boldsymbol{D}^n_{dis}$ denote DSIs of the n-th original and distorted stereo views, and computed by

$$D^n_{org}(i,j,d)=\left\|I^n_{L,org}(i,j)-I^n_{R,org}(i-d,j)\right\|^2 \quad (6)$$

where $\boldsymbol{D}^n_{org}(i,j,d)$ is a pixel' s value of $\boldsymbol{D}^n_{org}$ at (i,j) with the dis-parity of d, and $0\leqslant d\leqslant d_{max}$, d_{max} is the maximal search range for finding the optimal matching between the left and right views. In the same way, $\boldsymbol{D}^n_{dis}(i,j,d)$ can also be obtained.

Human visual perception has different sensitivity to different frequency components of an image, which is represented as the contrast sensitivity function (CSF), and used in depth perception evaluation. $\{\boldsymbol{D}^n_{org}(i,j,d)\}$ is decomposed by 3D-DWT into 8 sub-bands, including LLL, LLH, LHH, LHL, HLL, HHL, HLH and HHH. Based on human eye's sensitivity and computational load, the LLL sub-band of signal is mainly used and the LLL sub-band, $X^n_{org,LLL}=\{X^n_{org,LLL}(u,v,w)\}$, of $\boldsymbol{D}^n_{org}$ is expressed by

$$X^n_{org,LLL}(u,v,w)=DWT_{3D}\left(D^n_{org}(i,j,d)\right) \quad .(7)$$

Similarly, the LLL sub-band, $X^n_{dis,LLL}=\{X^n_{dis,LLL}(u,v,w)\}$, of $\boldsymbol{D}^n_{dis}$ is also obtained. Then, depth perception quality metric, $q_d(n)$, of each frame in stereo video is defined with $X^n_{dis,LLL}$ and represented by $q_D(n)=\Sigma ssim_{LLL}(u,v,w)/N_D$, where N_D is the number of pixels in $X^n_{dis,LLL}$.

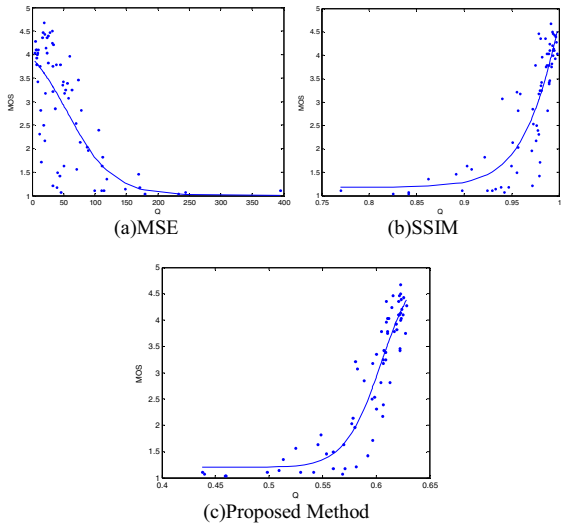
Finally, the depth perception quality metric Q_D is obtained by $Q_D=\Sigma q_D(n)/N_f$.

C. Total evaluation of stereo video quality

Final quality, Q, for distorted stereo video is given by fusing the two metric scores, the left-right views' quality Q_s and the depth perception quality Q_D. Let w be a weight, Q is computed by $Q=wQ_s+(1-w)Q_D$.

III. Prepare Experimental Results and Discussion

To verify the proposed method, the NAMA3DS1-CoSpaD1 stereo video [17] is used to test video quality assessment methods. In the database there are ten original 1920×1080 3D full HD stereo videos and corresponding one hundred distorted stereo videos at 25 frames per second. The distortions resulted from coding, transmission, and image processing including H.264/AVC, JPEG 2000, reduction of resolution, image sharpening, down sampling & sharpening. The mean opinion scores (MOS) of the videos range from 1 (bad) to 5 (excellent) are provided with 29 testers. Here, 70 stereo videos, only with the distortion of H.264/AVC and JPEG 2000, in the NAMA3DS1-CoSpaD1 database are used to compare different methods. Four statistical indexes are used to evaluate the performance in the experiments. They are the linear correlation coefficient (LCC), the Spearman's rank ordered correlation coefficient (SROCC), the root mean squared error (RMSE), and the outlier ratio (OR) between the predicted scores a and the mean opinion scores (MOS).



(a)MSE (b)SSIM

(c)Proposed Method

Fig. 2. Scatter plots between different methods and MOS.

TABLE I. Performance Comparison of Different Quality Assessment Methods

Distortion	Method	LCC	SROCC	OR	RMSE
H.264	MSE	0.6463	0.6288	0	0.8832
	SSIM	0.7843	0.7846	0	0.7181
	Proposed	**0.8743**	**0.8398**	0	**0.5618**
JPEG2000	MSE	0.8266	0.7974	0	0.7011
	SSIM	0.9399	0.9288	0	0.4255
	Proposed	**0.9548**	**0.9467**	0	**0.3703**
ALL	MSE	0.7297	0.6986	0	0.8278
	SSIM	0.8700	0.8723	0	0.5969
	Proposed	**0.9215**	**0.9101**	0	**0.4702**

Through a large number of experiments, the threshold T is set to 0.12, the weight C is 0.3, and a- and a+ are designed as a-=0.04 and a+=0.5, respectively. w is set as 0.75 for H.264 distortion and 0.70 for JPEG2000 distortion, respectively.

Fig. 2 shows scatter plots of different methods and MOS. It is clear that compared with other two methods, most points in the scatter plot locate closely to the fitted curve in the propose method. It means that the propose method performs more accurately in quality evaluation.

Table 1 shows performance comparison of different quality assessment methods, in which MSE denotes the mean square error metric and SSIM is structural similarity index. It is found that the proposed method gives better evaluation results than the two other methods.

IV. Conclusions

Considering the features of compression distorted stereo video, this paper presents a new stereo video quality assessment method with two metrics. Taking left-right views quality assessment into account, the concepts of significant pixels and just noticeable distortion are used to extract region of interest (ROI), and the ROIs are used to design the perceptual weight for quality assessment, an asymmetric perception model is designed to tempo-rally fuse qualities. Considering depth perception quality, disparity space image is generated and processed with three-dimensional wavelet transform for evaluating depth quality. The proposed method is tested with the common stereo video database. Experimental results show that the proposed method has well consistency with the subjective perceptual quality.

Rcknowledgements

This work was supported by Natural Science Foundation of China (U1301257, 61271270, 61311140261), and Natural Science Foundation of Zhejiang Province, China (LY15F010005).

References

[1] W Tam, F Speranza, S Yano, K Shimono, et al., "Stereoscopic 3D-TV: Visual comfort," IEEE Transactions on Broadcasting, vol. 57, no.2, pp. 335-346, 2011.

[2] F Shao, W Lin, G Jiang, M Yu, Q Dai, "Depth map coding for view synthesis based on distortion analyses," IEEE Journal on Emerging and Selected Topics in Circuits and Systems, vol. 4, no.1, pp. 106-117, 2014.

[3] T Kim, J Kang, S Lee, A C Bovik, "Multimodal Interactive Continuous Scoring of Subjective 3D Video Quality of Experience," IEEE Transactions on Multimedia, vol. 16, no.2, pp. 387 - 402, 2014.

[4] Y Chang, M Kim, "Binocular suppression-based stereoscopic video coding by joint rate control with KKT conditions for a hybrid video codec system," IEEE Transactions on Circuits and Systems for Video Technology, vol. 25, no.1, pp. 99 - 111, 2015.

[5] A Eskicioglu, P Fisher, "Image quality measures and their performance, IEEE Transactions on Communication," vol. 43, no.12, pp. 2959–2965, 1995.

[6] Z Wang, A C Bovik, H R Sheikh, E P Simoncelli, "Image quality assessment: From error visibility to structural similarity," IEEE Transactions on Image Processing, vol. 13, pp. 600-612, 2004.

[7] S L P Yasakethu, C T Hewage, et al., "Quality analysis for 3D video using 2D video quality models," IEEE Transactions on Consumer Electronics, vol. 54, no. 4, pp. 1969-1976, 2008.

[8] N Ozbek, A Tekalp, “Unequal inter-view rate allocation using scalable stereo video coding and an objective stereo video quality measure,” Int. Conf. Multimedia and Expo., pp. 1113-1116, 2008.

[9] C Hewage, S Worrall, et al., Quality evaluation of color plus depth mapbased stereoscopic video, IEEE. Journal Selected Topics Signal Processing, vol. 3, no. 2, pp 304-318, 2009.

[10] K Wang, K Brunnström, M Barkowsky, et al., “Stereoscopic 3D video coding quality evaluation with 2D objective metrics,” Proc. SPIE 8648, no.86481L, 2013.

[11] W Lin, C. J Kuo, “Perceptual visual quality metrics: A survey,” Journal Visual Communication Image Representation, vol. 22, pp. 297-312, 2011.

[12] Y Han, Z Yuan, G Muntean, “No Reference Objective Quality Metric for Stereoscopic 3D Video,” Int. Symp. on Broadband Multimedia Systems and Broadcasting, pp. 1-6, 2014.

[13] L Jin, A Boev, A Gotchev, K Egiazarian, “3D-DCT based perceptual quality assessment of stereo video,” Int. Conf. on Image Processing, pp. 2521-2524, 2011.

[14] X Yang, W Ling, Z Lu, E Ong, S Yao, “Just Noticeable Distortion Model and Its Applications in Video Coding,” Signal Processing: Image Communication, vol. 20, no.7, pp. 662–680, 2005.

[15] M Masry, S Hemami, Y Sermadevi, “A scalable wavelet-based video distortion metric and applications,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 16, no. 2, pp. 260-273, 2006.

[16] R Szeliski, D Scharstein, “Sampling the disparity space image,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 3, pp. 419-425, 2004.

[17] M Urvoy, M Barkowsky, R Cousseau, et al., “NAMA3DS1-COSPAD1: Subjective video quality assessment database on coding conditions introducing freely available high quality 3D stereoscopic sequences,” Int. Workshop on Quality of Multimedia Experience, pp. 109-114, 2012.

[18] W Wiseguy, D Argh, “An important paper”, IEEE Transactions on Signal Processing, vol. 99, pp. 9932–9948, Dec. 1988