3D-DCT BASED PERCEPTUAL QUALITY ASSESSMENT OF STEREO VIDEO

Lina Jin, Atanas Boev, Atanas Gotchev, Karen Egiazarian

Department of Signal Processing, Tampere University of Technology

ABSTRACT

In this paper, we present a novel stereoscopic video quality assessment method based on 3D-DCT transform. In our approach, similar blocks from left and right views of stereoscopic video frames are found by block-matching, grouped into 3D stack and then analyzed by 3D-DCT. Comparison between reference and distorted images are made in terms of MSE calculated within the 3D-DCT domain and modified to reflect the contrast sensitive function and luminance masking. We validate our quality assessment method using test videos annotated with results from subjective tests. The results show that the proposed algorithm outperforms current popular metrics over a wide range of distortion levels.

Index Terms— 3D video quality, depth perception, 3D-DCT, stereo-correspondence

1. INTRODUCTION

In the last decade, stereoscopic video technologies have burgeoned which led to an increasing interest in various 3D applications, for example 3DTV. The overall aim of a 3D video system is to deliver high quality video plus a natural sensation of depth. The 3D video delivery chain includes the stages of capture, encoding, transmission, possible post-processing at the receiver side and then display. Any of these stages may cause degradation of 3D visual quality and errors introduced at a certain step of the production process may propagate through the chain. Therefore, quality assessment is a key factor in the design and optimization of 3D video processing systems. Failing to evaluate the quality increases the risk of dissatisfied users who might experience effects attributed to poor stereo images, such as headache, eye pain or other simulator sickness symptoms.

The ultimate way to evaluate visual quality is to run subjective tests. However, subjective evaluation is time-consuming and expensive. The goal of objective stereoscopic quality assessment (QA) research is to design algorithms that can automatically assess the quality of 3D images or videos in a perceptually consistent manner. Although objective quality assessments of 2D image and video have been an active research topic for some decades, still very few efforts have been concentrated on 3D image and video quality evaluation.

For stereoscopic video quality metrics, a widely used approach is to apply 2D metric to evaluate the quality of each video channel separately, and then to calculate the overall 3D video quality as the mean of the two images. This approach might work for impairments equally affecting the left and right image; however, it would fail in many other cases. It is because this approach does not consider stereo perceptual information, such as rendered perception of depth, stereoscopic impairments, visual discomfort, relative size, motion, texture gradient, disparity, and temporal masking [1, 2]. To solve this problem, recently, some stereoscopic image and video quality metrics have been proposed by measuring the quality of disparity and cyclopean image separately and combing them in a compound measure. In [3], a monoscopic quality component and stereoscopic quality component for measuring stereoscopic image quality have been combined. The former component assesses the trivial monoscopic perceived distortions caused by blur, noise, contrast change etc; while the latter assesses the perceived degradation of binocular depth cues only. In [4], the popular 2D image quality metric called structural similarity index (SSIM) [5] has been applied for 3D images in the form of view plus depth, where information about depth has been added to the metric using a local or global approach. In [6], an overall quality metric has been suggested by combining image quality with disparity quality using a nonlinear function. In [7], a quality metric for color stereo images has been proposed based o the use of binocular energy contained in the left and right retinal images calculated by complex wavelet transform (CWT) and Bandelet transform. Other works have addressed the use of discrete cosine transform (DCT) as a component in 2D image and video quality metrics [8, 9, 10]. However, DCT has not been investigated for its implementation to stereoscopic image and video quality metric.

In this paper, we propose a full reference stereoscopic quality metric based on 3D-DCT, which takes into account some HVS properties, such as contrast sensitive function (CSF) and luminance masking In the proposed metric, 3D-DCT transform is used to analyze the perceptual similarity of blocks in stereo frames grouped using disparity correspondence and block-matching.

The rest of the paper is organized as follows. In Section 2 the proposed metric is described. The used set of test sequences and subjective tests are described in Section 3. Section 4 describes the experiment results. Final conclusions are presented in Section 5.

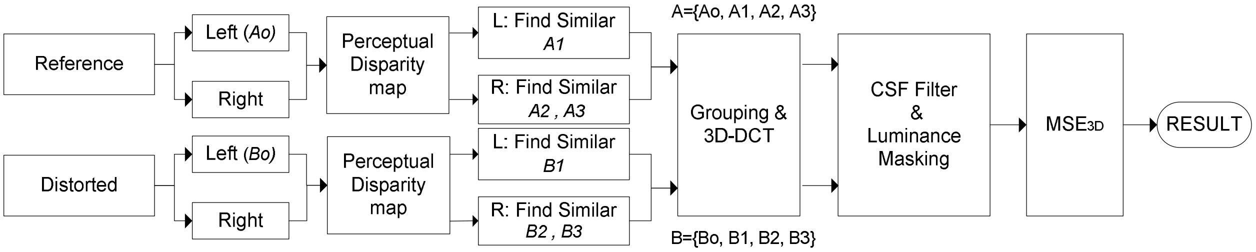

Figure 1. Flow chart of proposed model

2. PROPOSED ALGORITHM

The proposed stereo quality assessment scheme is given in Figure 1. We consider 3D video represented in the form of two channels – left and right – forming a stereo-pair of views. The depth perception is created by the slight different perspective of the two views manifested as disparity. The original (reference) video is distorted by some processing stage, e.g. compression and the level of distortions (or quality) between the reference and distorted videos has to be determined (full-reference metric). Both the reference and distorted videos are processed to find the disparity map between the left and right views. The assessment runs on blocks. For each reference block A_0 in the left reference view, the corresponding block B_0 is selected. In the reference video the most similar blocks in the left and right view, namely A_1, A_2, A_3 are found and stacked in a 3D structure, which then undergoes 3D-DCT. The same is done for the structure associated with the block B_0. Both 3D-DCT domain structures are then corrected with masking factors to account for the influence of the contrast sensitivity and luminance masking. Transform-domain mean square error is computed between the two set of coefficients to get a measure about the difference of three stereo blocks associated with A_0.

2.1 Finding block-disparity map

The disparity (or parallax) observed between the right and the left frame is generally inversely proportional to the distance to the object. Stereo matching is to search for a point in an image that corresponds to the point specified in the other image in terms of associated features. Stereo matching plays a key role in the structure-from-stereo algorithms, which aim at getting an image (a map) being indicative for the distance to the object. In our approach, we calculate a dense disparity map between the left and right frames using a color-weighted local search [11]. Considering rectified images, a window of size 9x9 from the left frame is run in horizontal direction to find similarity in terms of block matching, weighed by the color difference in a bilateral manner [11].

2.2 Block selection and 3D-DCT transform

DCT plays a key role in our approach. We rely on the capabilities of DCT to decorrelate data and achieve highly sparse representation [12]. In our approach, by the use of DCT we aim at modeling two processes taking place in the HVS. First, we model the binocular vision by combining together the left and right corresponding blocks [13]. Furthermore, we simultaneously model the saccades – the pseudo-random movements the eyes are performing while processing spatial information [14]. All those similar blocks are stacked together in a 3D structure to be jointly projected on the DCT basis. Thus, the resulting set of coefficients is expected to be informative about similarities across views and in spatial vicinity. Block-Matching (BM) is applied to find similar blocks around the reference block A_0 (B_0) and its stereoscopic correspondence within a search range region. Mean Squared Error (MSE) is used as dissimilarity measure

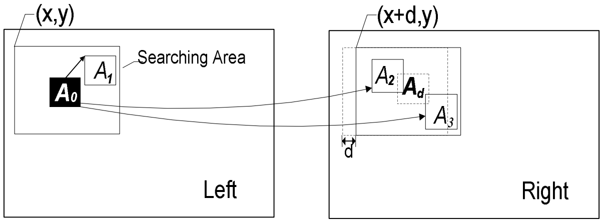

Figure 2. An example of block selection in stereoscopic image

$$MSE = \frac{1}{MN}\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}\left(S_{ij} - R_{ij}\right)^2 \qquad (1)$$

where $M \times N$ is the size of the macro block, S_{ij} and R_{ij} denote the pixel intensity of the searched block and reference block respectively. Using BM, one best matched block in the left view and two similar blocks in the right view are found and grouped together as shown in Figure 2. In the figure, A_0 is the reference block in the left-view frame, A_1 is the most similar block to A_0 in the same channel. A_d is the corresponding block to reference block A_0 in right channel found through the stereo-correspondence search, and A_2 and A_3 are the two best matching blocks to A_0 which are searched within search region around A_d in the right channel. Note that the similarity between A_0 and A_d has been found through another similarity mechanism an eventual A_d could be or could not be one of the selected blocks A_k, $k = 2$, 3. The four blocks in the reference stereoscopic image $A = \{A_0, A_1, A_2, A_3\}$ and the respective four blocks in the distorted stereoscopic image $B = \{B_0, B_1, B_2, B_3\}$ are grouped into two 3D arrays respectively.

A 3D-DCT transform is applied to these two 3D arrays, i.e. $A = \{A_0, A_1, A_2, A_3\}$ and $B = \{B_0, B_1, B_2, B_3\}$. In our setting, for nice symmetry and fast processing the size of the blocks is fixed to 4x4, thus fixing the size of the 3D structure to a cube of ridge length 4. However, any other block size is possible and also more blocks can be collected on the base of similarity, thus forming a 3D structure of bigger size. Our previous experiments with 2D images and videos have shown that finding one similar block is sufficient to account for this type of similarity around the reference block. The search region around the reference block and its stereo correspondence has been fixed to 22x22.

2.3 Modified MSE

In [9], a perceptually-driven metric for 2D image quality assessment has been suggested. It calculates an MSE between the 8x8 DCT coefficients of the reference and distorted block modified by coefficients reflecting the masking effect of the contrast sensitivity function [9]. The method has been further modified in [10] to take into account also between-coefficients contrast masking of DCT basis functions. While the original approach was developed for DCT of size 8x8 and using masking coefficients determined by the JPEG quantization table, we have modified it to work with DCT of size 4x4. Correspondingly, the coefficients of the top layer of the 4x4x4 3D-DCT, which concentrate most of the energy of the similar blocks are weighted with a down-sampled version of the masking table used in [10]. The lower layers are scaled down with coefficients determined by the energy distribution within the block. The modified version of PSNR is:

$$PHVS_{3D} = 10log(255^2/MSE_{3D}) \qquad (2)$$

$$MSE_{3D} = \frac{16}{I \times J} \sum_{i=1}^{I-3} \sum_{j=1}^{J-3} MSE\left(A_{ij}^D, B_{ij}^D\right) \cdot C_{4x4}^2 \cdot Mask \qquad (3)$$

where, I, J denote image size, A_{ij}^D is the first layer of 4x4x4 3D-DCT coefficients with indices i, j in reference image, B_{ij}^D is the one in distorted image, C_{4x4} is a correcting factor determined by the CSF and it is obtained by down-sampling the normalized quantization table of JPEG into 4-by-4 size and then squared [9], $Mask$ is to reduce the value of contrast masking in accordance to the model proposed in [10]. We have tested two versions of the formula: without masking effect, i.e. $Mask = 1$,denoted as 'PHVS-3D' and with masking effect calculated, denoted as 'PHVS-M-3D'. For the latter case, the table from [10] has been divided into four 2-by-2 blocks and then used in the formula from [10] with a normalizing factor $p = 2$ instead of 16. For efficient calculation, we have chosen that the reference blocks do not overlap.

3. TEST SEQUENCES

The performance of the proposed stereoscopic quality assessment model has been validated using the results from subjective tests [15]. Four multi-view video sequences have been used, namely Akko&Kayo, Champagne Tower, Pantomime, and Love Birds1. Different camera baselines have been selected to get stereo pairs with three different types of depth: no depth (2D), short baseline (3D) and wide baseline (3D). The sequences have been cut into 10 seconds and coded using the simulcast MPEG-4 standard encoder. Five different quantization parameters QP (=25, 30, 35, 40, 45) have been applied to each processed sequence. Thus, a total of 12 reference sequences and 60 distorted sequences have been obtained.

The video sequences have been used in experienced quality tests. The test group included 32 persons equally stratified by gender and age between 18 and 45. The visualization was done on an auto-stereoscopic display provided by NEC [15]. The tests collected the opinion in term of quality (11 point scale) and acceptance (binary scale). The overall ratings of stereoscopic videos have been ranked in terms of mean opinion score (MOS).

4. RESULTS

The results of the proposed approach are compared along with several state-of-art quality metrics: PSNR, MSSIM [16], SSIM [5], UQI [17], NRMSE [18], PSNR-HVS [9], PSNR-HVS-M [10], which are all 2D metrics and the 3D metric from [6]. For the latter, we have set SSIM to measure the image quality and UQI to measure the disparity quality [6]. All algorithms compare the luminance component only. The 2D metrics have been run on the left and right channels separately and the results have been averaged.

Figure 3 presents the results of fitting logistic curves on the dependence between MOS and the quality predicted by some of the compared measures. In Table I, Spearman, Pearson and Kendall correlations for each compared quality assessment are presented. Popular 2D quality metrics, such as MSSIM, SSIM and UQI for the given stereoscopic test set show lower correlation to MOS. Surprisingly, the method from [6] does not correlate to MOS well either, see Figure 3 d). The standard PSNR, NRMSE, along with PSNR-HVS and PSNR-HVS-M suit the visual perception remarkably better. Finally, it can be seen that the proposed metrics PHVS-M-3D and PHVS-3D, outperform the other considered metrics. Both Spearman correlation and Pearson correlation are above 0.9 for PHVS-M-3D and PHVS-3D. The performance is also confirmed by the well-behaving logistic curves shown in Figure 3 e) and f).

Table I. Spearman, Pearson, and Kendall correlations for considered metrics

Metric	Spearman Correlation	Pearson Correlation	Kendall Correlation
PSNR	0.8639	0.8431	0.6610
MSSIM[16]	0.6734	0.7246	0.4857
SSIM[5]	0.6232	0.7068	0.4437
UQI[17]	0.5113	0.5333	0.3726
NRMSE[18]	0.8431	0.8786	0.6327
PSNR-HVS[9]	0.8862	0.8850	0.7017
PSNR-HVS-M[10]	0.8594	0.8817	0.6644
Global comb.[6]	0.6307	0.6938	0.4416
PHVS-3D	**0.9168**	**0.9063**	**0.7436**
PHVS-M-3D	**0.9150**	**0.9232**	**0.7500**

5. CONCLUSION

In this paper, we proposed a novel full-reference stereoscopic quality metric based on 3D-DCT. We modeled the combined effects of binocular vision and saccades which are considered important for visual quality assessment through the decorrelation properties of DCT. In addition, the CSF and luminance masking were taken into account by applying proper masking. The approach is simple yet quite effective as demonstrated by comparing it against subjective tests. The experimental results have shown that our metrics outperform current state-of-the-art quality metrics. We have to note that our implementation does not take into account masking effects created by motion. Our experiments have shown that this masking plays minor role in estimating the quality. This observation has been confirmed by subjective tests on still images from the same database, which resulted in the same MOS as in the case of the respective videos. Another future improvement of the approach can be achieved by adding proper disparity-driven weighting to the other masking mechanisms. As the subjective experiments in [15] have shown, differences in depth presence (i.e. through different camera baseline) play rather marginal role in evaluation the overall quality compared with compression artifacts. Yet, they can and shall be modeled in a future version of the suggested approach.

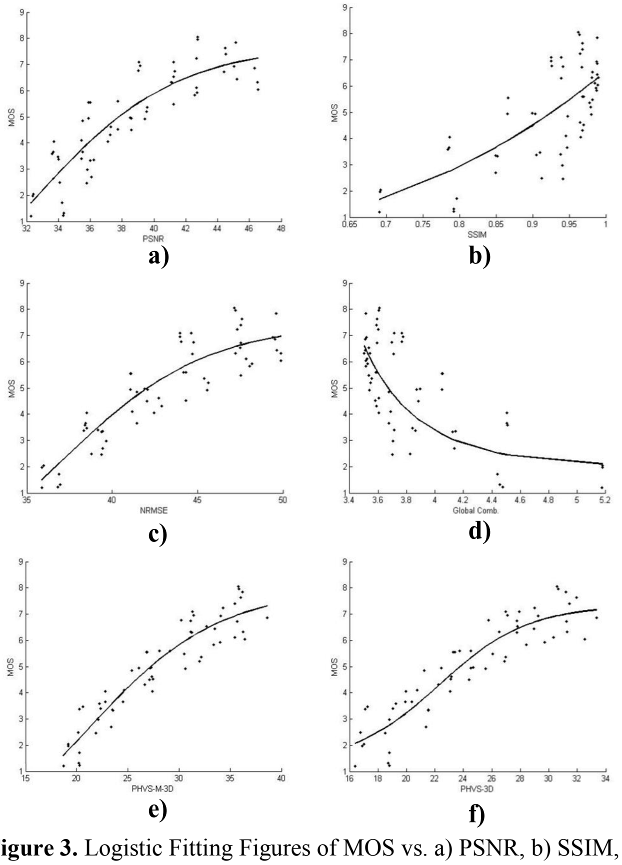

Figure 3. Logistic Fitting Figures of MOS vs. a) PSNR, b) SSIM, c) NRMSE, d) Global Comb. e) PHVS-M-3D, f) PHVS-3D

6. REFERENCES

[1] M. T. M. Lambooij, W. A. Ijsselsteijn, and I. Heynderickx, "Visual discomfort in stereoscopic displays: a review," in *Stereoscopic Displays and Virtual Reality Systems XIV of Proceedings of SPIE*, San Jose, USA, pp. 1-13, Jan. 2007.

[2] A. Boev, D. Hollosi, A. Gotchev, K. Egiazarian, "Classification and simulation of stereoscopoic artifacts in mobile 3DTV content," *Electronic Imaging Symposium 2009*, San Jose, USA, Jan. 2009.

[3] A. Boev, A. Gotchev, K. Egiazarian, A. Aksay, and G.B. Akar, "Towards compound stereo-video quality metric: a specific encoder-based framework," *IEEE Southwest Symposium on Image Analysis and Interpretation*, Denver, USA, pp. 218-222, June 2006

[4] A. Benoit, P. Le Callet, P. Campisi, and R. Cousseau, "Quality assessment of stereoscopic images," *EURASIP Journal on Image and Video Processing*, special issue on 3D Image and Video Processing, vol. 2008, Article ID 659024, 13 pages, 2008

[5] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, issue 4, pp. 600-612, April 2004.

[6] J. You, L. Xing, A. Perkis, and X. Wang, "Perceptual Quality Assessment for Stereoscopic Images Based on 2D Image Quality Metrics and Disparity Analysis," *Int. Workshop on Video Processing and Quality Metrics for Consumer Electronics -VPQM*, Scottsdale, AZ, USA, 2010

[7] R. Bensalma, and M.C. Larabi, "Towards A Perceptual Quality Metric For Color Stereo Images," *IEEE 17th International Conferences on Image Processing,* Hong Kong, Sep. 26-29, 2010

[8] A. B. Watson, "Toward a perceptual video quality metric," *Human Vision, Visual Processing, and Digital Display VIII*, 3299, pp. 139-147, 1998

[9] K. Egiazarian, J. Astola, N. Ponomarenko, V. Lukin, F. Battisti, and M. Carli, "New full-reference quality metrics based on HVS," *Int. Workshop on Video Proc. and Quality Metrics*, Scottsdale, USA, 2006

[10] N. Ponomarenko, F. Silvestri, K. Egiazarian, M. Carli, J. Astola, and V. Lukin, "On between-coefficient contrast masking of DCT basis functions," *Int. Workshop on Video Proc. and Quality Metrics*, USA, 2007

[11] S. Smirnov, A. Gotchev, and M. Hannuksela, "Comparative analysis of local binocular and trinocular depth estimation approaches", *Proc. of SPIE*, vol. 7724, 2010, pp. 77240H.

[12] K. Dabov, A. Foi, and K. Egiazarian, "Image restoration by sparse 3D transform-domain collaborative filtering," *Proc. SPIE Electronic Imaging '08*, no. 6812-07, San Jose, USA, Jan. 2008.

[13] Brian A. Wandell, "Foundations of Vision", Sinauer Associates, 1995.

[14] Martin J. Tovée, "An Introduction to the Visual System" 2nd ed. , Cambridge, 2008.

[15] S. Jumisko-Pyykkö, T. Haustola, A. Boev, and A. Gotchev, "Subjective Evaluation of mobile 3D content: depth range versus compression artifacts", *Proc. of SPIE*, vol. 7881A, 2011.

[16] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multi-scale structural similarity for image quality assessment," *IEEE Asilomar Conference on Signals, Systems and Computers*, Nov. 2003.

[17] Z. Wang, and A. Bovik, "A universal image quality index," *IEEE Signal Processing Letters*, vol. 9, pp. 81–84, March, 2002

[18] S. Baker, D. Scharstein, J. P. Lewis, S. Roth, M.J. Black, and R. Szeliski, "A database and evaluation methodology for optical flow," *IEEE Int'l Conf. on Computer Vision, Crete*, Greece, pp. 243-246, Dec. 2007