Binocular-Combination-Oriented Perceptual Rate-Distortion Optimization for Stereoscopic Video Coding

Yanwei Liu, *Member, IEEE,* Jinxia Liu, Antonios Argyriou, *Senior Member, IEEE* and Song Ci, *Senior Member, IEEE*

***Abstract*—In the chain of stereoscopic video processing, stereoscopic video coding and viewing are usually two independent stages. Conventional stereoscopic video coding puts emphasis on improving the coding efficiency by seeking the optimal trade-off between the coding bit-rate and the signal-based distortion, while neglecting the perceptual behaviors of binocular combination when stereoscopic video is viewed by human beings. In this paper, we propose to utilize binocular combination to optimize the stereoscopic video coding from the perspective of perceptual quality measurement. Specifically, we propose a novel binocular-combination-oriented measurement for visual distortion, and then derive the Lagrange multiplier for the binocular-combination-oriented rate-distortion optimization (RDO). Via extensive subjective tests, the results show that the proposed perceptual RDO can save more than 5% BD-rate over the traditional RDO in MV-HEVC (MultiView extension of HEVC) for stereoscopic video coding.**

***Keywords*—Binocular combination, Perceptual rate-distortion optimization, Stereoscopic video coding, Stereoscopic 3D video.**

I. Introduction

Advanced 3D display technology enables the human brain to experience depth beyond the conventional 2D plane, providing enriched experiences to viewers. As a result, there has an increasing number of consumer-oriented 3D video applications, such as digital cinema, home theatre, mobile 3D video, 3D games, and virtual/augmented reality. However, 3D video applications have not successfully entered into our daily lives due to their poor Quality of Experience (QoE) [1]. Therefore, enhancing QoE of 3D video applications is crucial for the success of all kinds of 3D video applications.

Generally, poor 3D QoE are caused by several factors impacting 3D perception. Unlike 2D video applications, several forms of distortion, as well as the neurological symptoms (such as visual fatigue and headache) induced by the ocular adjustment to 3D depth, may lead to quality degradation of 3D perception. In the full 3D video processing chain, 3D video coding is a key factor affecting 3D QoE, along with imprecise or improper 3D geometry incurred by the generation of 3D contents. Specifically, during 3D video compression, residual data quantization errors are inevitable and irreversible, which degrade the 3D video quality, evaluated by traditional video signal distortion measurement methods, such as mean squared error (MSE) or sum of squared errors (SSE). Although these measurement methods can sufficiently explain video signal quality degradation, they do not perform well on 3D video perceptual quality degradation [2][3][4]. Due to the intrinsic nature of 3D perception in human brain, a new perceptual quality metric for 3D video is necessary for quantifying 3D video compression performance.

In literature, the majority of research efforts have focused on quality assessment of compressed 3D videos. Several of them on 3D video quality assessment have proposed various quality metrics based on different 2D video quality evaluation schemes [5][6][7]. These quality metrics take considerations of the characteristics of human visual system (HVS) and 3D vision, such as binocular rivalry [6] and binocular suppression [7], focusing on modeling the coarse behavior of 3D perception [7] and evaluating the 3D video quality qualitatively. Another major drawback of those quality metrics is their high computational complexity. For 3D video compression, the computation requirements on quality metrics are even higher. The perceptual distortion introduced by the coding scheme must be quantitatively characterized at a low computational complexity. Hence, existing 3D video quality evaluation methods cannot be directly adopted by 3D video encoders.

Furthermore, in literature, existing stereoscopic 3D video coding frameworks only adopt inter-view prediction for exploiting the natural inter-view correlation without considering the human brain behavior on fusing two views in stereoscopic video perception. In this paper, we present a novel stereoscopic video coding scheme that is based on the binocular combination process in human brain. Therefore, it can be used for optimizing the perceptual 3D video quality under the given bit-rate constraint. To the best of our knowledge, there has no such binocular-combination-oriented perceptual rate-distortion optimization (RDO) for 3D video coding in literature.

The contributions of this paper are as follows.

1) A binocular-combination-oriented perceptual distortion metric is proposed to characterize 3D visual distortion in

This work was supported in part by Zhejiang Provincial Natural Science Foundation of China under Grant LY17F010001, Public welfare project of Zhejiang Province under Grant 2014C31072, National Natural Science Foundation of China under Grant 61472388.

Y. Liu is with the State Key Lab of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100193, China and also with the School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: liuyanwei@iie.ac.cn).

J. Liu is with the Zhejiang Wanli University, Ningbo 315100, China (e-mail: liujinxia@zjwu.edu.cn).

A. Argyriou is with the Department of Electrical and Computer Engineering, University of Thessaly, Greece (e-mail: anrgyr@uth.gr).

S. Ci is with the University of Nebraska-Lincoln, Omaha 680182, USA (e-mail: sci@unl.edu).

stereoscopic video coding, where the binocular 3D perceptual distortion is modeled as a linear fusion operation. To obtain the binocularly combined visual distortion, the gain-control combination of Difference of Gaussian (DoG) models in the frequency domain has been adopted. Hence, the binocular 3D perceptual distortion can be quantitatively characterized by the combination of the distortions from two views.

2) We propose a novel rate-distortion function by in cooperating the binocular combination characteristics of stereoscopic video coding. Specifically, based on the coding-induced 3D perceptual distortion measurement, the relationship between 3D perception distortion and coding bit-rate is derived based on the rate-distortion theory. The perception-based 3D RDO is subsequently implemented to optimize the perceptual stereoscopic 3D coding efficiency through optimal mode selection. The main advantage of the proposed 3D perceptual RDO over the traditional RDO is that the binocular combination in the viewing stage is integrated with the compression stage, making the coding optimization gear towards not only the signal fidelity but also the perceptual quality preservation under a given bit-rate constraint.

The rest of the paper is organized as follows. The related work is presented in Section II. The 3D perceptual RDO based on binocular combination is described in Section III. Experimental results are presented and discussed in Section IV. Finally, Section V concludes the paper.

II. Related work

The research work relevant to this paper can be classified into three categories: stereoscopic 3D video quality assessment, 3D video RDO, and binocular combination models.

A. Quality assessment for stereoscopic 3D video coding

A significant body of work exists in the area of stereoscopic 3D video quality assessment. The first works on stereoscopic 3D video quality evaluation were based on traditional 2D video quality evaluation tools. These works process two-view videos (or video plus depth) independently and then average the evaluation scores to assess the total stereoscopic 3D video quality [8][9]. Though this type of approach can evaluate the stereoscopic 3D video quality to a certain extent, it neglects the natural multiview geometry in 3D video which means that the evaluation accuracy is limited. Newer works proposed the quality assessment of the cyclopean view through the binocular rivalry mechanism based on psychological 3D perception [5][6][10][11]. Since the binocular suppression (a special case of binocular rivalry) behavior also exists in binocular vision, the authors in [5] proposed an adaptation of the 3D spatial sensitivity by binocular fusion and suppression for performing quality assessment. Another class of schemes for stereoscopic 3D image quality evaluation utilizes the pooling of different classes of quality indexes to calculate the total quality measurement [7][12][13][14]. Beyond the pooling of multiple quality metrics, a quality assessment approach based on binocular integration has been recently proposed [15]. This study is the first to introduce the binocular combination principle into a stereoscopic 3D video quality assessment model.

The previous approaches focused on general 3D video/image quality assessment metrics. They considered different types of artifacts and distortions in several stages of the stereoscopic 3D video processing and focused on building efficient and objective 3D quality metrics. However, these metrics cannot be directly adopted for measuring the stereoscopic 3D video distortion by the video encoder. A stereoscopic 3D video quality metric must be accurate so that it can optimize the video coding efficiency. Since video is naturally a digital signal, a quality metric based on perceptual signal distortion is ideal for measuring 3D video quality degradations under different bit-rate constraints. This is one of the main goals of this paper.

B. RDO in 3D video coding

Recently, significant progress has been made in 3D video coding (3DVC) [16]. Research efforts in 3D/multiview video coding are typically aimed at removing the signal redundancy among the 3D video data. Several coding tools have been developed to improve 3D video coding efficiency [17]. RDO [18] is an important tool for improving the video coding efficiency. During video coding, the selection of different encoding modes can lead to different video fidelities and different amount of produced bits. To find the optimal encoding mode, a Lagrange-based optimization method is predominantly used to identify the optimal trade-off between the amount of encoding bits and the video distortion.

In the field of 2D video coding, the literature is rich with seminal papers that accurately model the rate-distortion function for different video coding standards [19][20], that is later used for RDO. In recent years, perceptual RDO technologies have gradually emerged. The goal now is to optimize the perceptual video coding performance by removing redundancy based on perceptual metrics [21][22].

Furthermore, for 3D video coding, several RDO-based mode selection approaches have been proposed based on different distortion metrics [23][24][25][26]. In [25], the authors proposed a rate control scheme based on binocular suppression for stereoscopic video coding. This work only partially considered the binocular vision characteristics of stereoscopic video for performing the rate control step. In [26], the authors proposed a mode selection scheme for 3D video coding based on single-view HVS-based perceptual distortion measurement, while they did not touch upon the issue of binocular vision.

The previous approaches used the signal MSE or the binocular suppression characteristic to assess the 3D video quality. Clearly, this does not accurately reflect the fusion process of the two views in the human brain. The main problem is that the 3D perception is generated by the human binocular visual system. Hence, the binocular vision characteristics must be introduced into the stereoscopic 3D video distortion model.

C. Binocular combination model

During the formation of the 3D perception in the human brain, one of the most important processes, besides the generation of the depth perception, is the generation of the cyclopean perception. A fundamental issue in vision science, that has been the focus of many psychophysical and physiological investigations [27][28][29], is how the HVS combines information from the two eyes to form a single cyclopean representation of the external world. Studies in the past have found that the binocular combination mechanism exhibits several psychological properties, such as the cyclopean perception [30] and Fechner's paradox [29]. The cyclopean perception means that people will perceive the single view of a unified visual scene from the combination of the images projected to the two eyes. It is one of the basic functions of binocular combination. Fechner's paradox refers to the fact that a bright light presented to one eye may actually appear less bright when a dim light is shone into the other eye [29]. These two properties highlight the theoretical foundations for modeling the binocular combination process.

Based on the understanding of binocular combination mechanism in the human brain, a large number of computational models of binocular combination have been proposed. In [31], a weighting form of binocular combination of two views was proposed. To capture more accurately the cyclopean perception, the vector summation model was proposed in [32], which characterized the binocular brightness perception as the sum of two orthogonal vectors (with appropriate normalizations). The nature of neural cell processing in the HVS was considered by the authors in [33] that computed the neural responses from each eye (e.g. excitation and inhibition) for binocular combination. However, the accuracy of interpreting cyclopean perception in this model depends on the adopted neural response description. Based on gain-control theory, Ding and Sperling [34] proposed to measure the appearance of a cyclopean image resulting from binocular combination of two view signals by using the energy responses of the left and right images. The model can explain well the cyclopean perception and Fechner's paradox.

III. Binocular-Combination-Oriented Perceptual Stereoscopic 3D RDO

In this section, we introduce the proposed RDO scheme for stereoscopic video coding that is based on binocular combination. In the proposed scheme, the traditional RDO based on signal distortion is extended to RDO based on visual distortion for stereoscopic video coding. Specifically, we first derive a new 3D perception distortion model in terms of the binocular combination mechanism, and then according to the successive encoding relationship of the left and right views we derive a new 3D perceptual rate-distortion function for right view encoding, in addition to the traditional rate-distortion function for left view encoding.

For the stereoscopic 3D video applications, stereoscopic video compression and human brain binocular combination can be understood as a single process that leads to the perception of the 3D scene signal, as illustrated in Fig. 1. Specifically, the 3D scene signal I_{3D} can be expressed as two view signals namely I_L and I_R. When people view the raw stereoscopic video, the raw signals of the two views are combined into one 3D signal by the human brain with a pair of raw combination coefficients ξ_L and ξ_R. Thus, during the 3D signal capturing stage, the raw 3D signal can be seen as the combined form of two separate raw signals $\xi_L \cdot I_L$ and $\xi_R \cdot I_R$. During the compression stage, I_L and I_R are encoded by the encoders Enc_L and Enc_R. Although the current MV-HEVC or AVC-MV (H.264/AVC MultiView profile) encoder simultaneously encodes the two views with inter-view prediction, the two views are indeed successively encoded by the same encoder and this is equivalent to two views being encoded by two encoders. The compressed video signals of the two views are distributed to the client and then successively decoded by the decoders Dec_L and Dec_R. Finally, the decoded signals $\hat{I}_L$ and $\hat{I}_R$ are combined into one decoded 3D signal $\hat{I}_{3D}$ with another pair of binocular combination coefficients $\hat{\xi}_L$ and $\hat{\xi}_R$ while being viewed by humans.

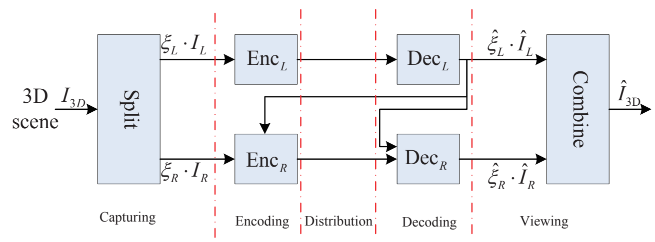


Fig. 1. The stereoscopic 3D visual signal compression and human brain combination process.

A. Binocular combination distortion

The objective measurement of perceptual 3D video quality remains a challenging problem. The issue is the complicated interaction among the neural signals of the two eyes that are involved in the visual cognition process of 3D scene in the human brain. This visual cognition mechanism is currently not understood to the desired degree. Thus, it is challenging to efficiently measure the 3D perception in an objective way. In recent years, with the advancement in vision science, the community has reached a better understanding of the binocular combination process that occurs in the human brain. Currently, we can use the binocular combination process to approximate the early stages of the formation of the 3D perception.

As mentioned earlier, several computational models of binocular brightness combination have been proposed [29], such as the eye weighting model, vector summation model, neutrally inspired models, and gain-control theory models. Research in the past has shown that among these models, the early stage of psychological binocular combination can be approximated with sufficient accuracy with the help of gain-control theory [34][35]. The binocularly combined cyclopean 3D video signal I_{3D} can be obtained as

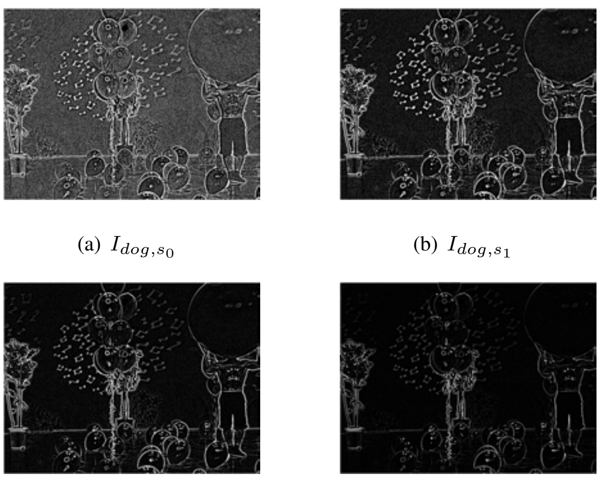
$$\begin{aligned} I_{3D} &= f_c(I_L, I_R) \\ &= \left(\frac{1+\varepsilon_L(I_L)}{1+\varepsilon_L(I_L)+\varepsilon_R(I_R)}\right) I_L + \left(\frac{1+\varepsilon_R(I_R)}{1+\varepsilon_L(I_L)+\varepsilon_R(I_R)}\right) I_R \\ &= \xi_L \cdot I_L + \xi_R \cdot I_R, \end{aligned} \quad (1)$$

where $\varepsilon_L(I_L)$ and $\varepsilon_R(I_R)$ denote the total energies in different frequency bands for the left view video signal I_L and the right view video signal I_R, respectively, ξ_L and ξ_R denote the different combination coefficients for the raw video signals of the left and right view, respectively, and $f_c(\cdot)$ denotes the binocular combination processing.

One important result that allows us to proceed further, is that the visual response of an image in HVS can be characterized via Difference-of-Gaussian models [36] [15] [37]. The physiological study in [36] indicated that the optimal number of models (the number of DoG spatial frequency bands) is 4. The DoG model for an image I can be expressed as

$$\begin{aligned} & I_{dog,\, s}(x,y) \\ & = (G(s) - G(ks)) * I \\ & = (\tfrac{1}{2\pi s^2} e^{-(x^2+y^2)/(2s^2)} - \tfrac{1}{2\pi k^2 s^2} e^{-(x^2+y^2)/(2k^2 s^2)}) * I \end{aligned} \tag{2}$$

where s is the standard deviation of Gaussian response $G(\cdot)$, and k is the space constant with typical value equal to 1.6. Fig. 2 shows the DoG decomposition results of one given image in the Balloons video sequence. The different frequency band components in Fig. 2 indicate different human visual responses for an image.



(a) I_{dog,s_0} (b) I_{dog,s_1}

(c) I_{dog,s_2} (d) I_{dog,s_3}

Fig. 2. The different spatial frequency band components I_{dog,s_0}, I_{dog,s_1}, I_{dog,s_2}, I_{dog,s_3} of an given image in Balloons clip, and $s_0 = 0$, $s_1 = 1$, $s_2 = k \cdot s_1$, $s_3 = k \cdot s_2$ denote the standard deviations of Gaussian response, respectively

The binocular combination of two views in human brain can be approximated as the weighted combination of the DoG spatial frequency bands for the left and right views. The HVS response for each view in a stereoscopic pair, can be decomposed to two independent DoG spatial frequency bands. Then the energy for each DoG band can be computed, and also the total energy $\varepsilon_L(I_L)$ for all bands for the left and right views. Consequently, the binocular combination coefficients ξ_L and ξ_R, can be computed as described in Eq. (1). The total energy over all DoG bands for one complete image is

$$\varepsilon(I) = \textstyle\sum_i \varepsilon(b_i) = \sum_i \sum\limits_{p \in b_i} p^2, \tag{3}$$

where $\varepsilon(b_i)$ denotes the energy of the i^{th} band, and p is the value of the pixel in band b_i of image I. With Eq. (3), $\varepsilon_L(I_L)$ and $\varepsilon_R(I_R)$ for the left and right views can be derived.

After the cyclopean view processing for the raw and compressed stereoscopic images, the 3D visual distortion that makes use of binocular combination is computed as

$$\begin{aligned} D_c &= E\{I_{3D} - \hat{I}_{3D}\} \\ &= E\{(f_c(I_L, I_R) - f_c(\hat{I}_L, \hat{I}_R))^2\} \\ &= E\{(\xi_L \cdot I_L + \xi_R \cdot I_R - \hat{\xi}_L \cdot \hat{I}_L - \hat{\xi}_R \cdot \hat{I}_R)^2\} \\ &= E\{(\xi_L \cdot I_L + \xi_R \cdot I_R - \xi_L \cdot \hat{I}_L + \xi_L \cdot \hat{I}_L \\ &\quad -\hat{\xi}_L \cdot \hat{I}_L - \hat{\xi}_R \cdot \hat{I}_R + \xi_R \cdot \hat{I}_R - \xi_R \cdot \hat{I}_R)^2\} \\ &= E\{(\xi_L \cdot (I_L - \hat{I}_L) + \xi_R \cdot (I_R - \hat{I}_R) \\ &\quad +\hat{I}_L \cdot (\xi_L - \hat{\xi}_L) + \hat{I}_R \cdot (\xi_R - \hat{\xi}_R))^2\}, \end{aligned} \tag{4}$$

where $E\{\cdot\}$ denotes the expectation operator. Due to the negligible effect of compression on the combination coefficients, the coefficients for the compressed and raw stereoscopic images are assumed to the same [15]. We have $\xi_L \approx \hat{\xi}_L$ and $\xi_R \approx \hat{\xi}_R$. Furthermore, we assume that the quantization noise for the left view and right views are mutually uncorrelated so that $E\{(I_L - \hat{I}_L) \cdot (I_R - \hat{I}_R)\}$ becomes zero. Hence,

$$\begin{aligned} D_c &= E\{(\xi_L \cdot (I_L - \hat{I}_L) + \xi_R \cdot (I_R - \hat{I}_R))^2\} \\ &= (\xi_L)^2 \cdot E\{(I_L - \hat{I}_L)^2\} + (\xi_R)^2 \cdot E\{(I_R - \hat{I}_R)^2\} \\ &\quad +2\xi_L \xi_R \cdot E\{(I_L - \hat{I}_L) \cdot (I_R - \hat{I}_R)\} \\ &= (\xi_L)^2 \cdot D_L + (\xi_R)^2 \cdot D_R \\ &\quad +2\xi_L \xi_R \cdot E\{(I_L - \hat{I}_L) \cdot (I_R - \hat{I}_R)\} \\ &\approx (\xi_L)^2 \cdot D_L + (\xi_R)^2 \cdot D_R, \end{aligned} \tag{5}$$

where D_L and D_R are the distortion values for the left and right views.

B. Stereoscopic 3D visual RDO

It is well known that in the high bit-rate regime of video encoding, the relationship between the distortion and the quantization level can be approximated as [38]

$$D = \frac{q^2}{12}, \tag{6}$$

where q is the quantization step size. The binocular combination distortion can be described as

$$D_c = \frac{(\xi_L)^2 \cdot q_L^2}{12} + \frac{(\xi_R)^2 \cdot q_R^2}{12}, \tag{7}$$

where q_L and q_R are the quantization step sizes for left and right views, respectively.

For 2D video encoding, there is a trade-off between distortion and bit-rate. RDO attempts to find the optimal operating point between these two conflicting objectives. For stereoscopic video encoding, that adds the element of binocular combination in human brain, a perceptual RDO is necessary to determine the optimal video encoding mode. We note in Fig. 1 that stereoscopic video encoding is executed in two stages. The first stage is the left view video encoding that uses traditional RDO. It satisfies

$$J_L = D_L + \lambda_L R_L \tag{8}$$

with

$$R_L = a_L \cdot \log_2(\frac{b_L}{q_L^2/12}) \text{ and } D_L = \frac{q_L^2}{12}, \tag{9}$$

where R_L is the left view encoding bit-rate, a_L and b_L are constants, and λ_L is the Lagrange multiplier for the left view encoding. Then

$$\begin{aligned} \lambda_L &= -\frac{dD_L}{dR_L} \\ &= -\frac{dD_L/dq_L}{dR_L/dq_L} \\ &= -\frac{2}{12}q_L \Big/ \frac{-2a_L}{q_L \ln 2} \\ &= \frac{\ln 2 \cdot q_L^2}{12 \cdot a_L} \end{aligned}, \tag{10}$$

We notice in Eq. (10) that the Lagrange multiplier for RDO is dominated by the quantization level. In the reference encoding models of H.264/AVC (JM) and H.265/HEVC (HM), the Lagrange multiplier is determined by the quantization parameter first and then refined in terms of the video coding structure. Based on the relationship between quantization parameter and the quantization step size in MV-HEVC, Eq. (10) can be empirically modified as [43]

$$\lambda_L = c \cdot 2^{(qp-12)/3}, \tag{11}$$

where c is a constant related to the coding structure, and qp denotes the coding quantization parameter (QP).

Similarly, during the second stage of stereoscopic video coding for the right view, the distortion must be computed to control the right view coding bit-rate. The RDO for the right view can be characterized as

$$J_R = D_c + \lambda_R R_R, \tag{12}$$

where J_R denotes the Lagrange cost and R_R is the right view encoding bit-rate. Fig. 1 illustrates that the decoded left view is used as the inter-view prediction reference for the encoding of the right view. Thus, the Lagrange multipliers for the left and right views are different. The total perceptual distortion corresponds to the two encoding channels for left and right views. For the left view, the traditional RDO is used and so λ_L is determined by the quantization parameter of the left view. However, for the right view the video encoding mode is selected based on the perceptual RDO. Thus, the obtained stereoscopic 3D visual quality is optimal under the total bit-rate constraint of the two views.

For the right view, when considering the human brain's binocular combination at the encoder, the perceptual RDO in Eq. (12) can be further expanded as

$$J_R = \frac{(\xi_L)^2 \cdot q_L^2}{12} + \frac{(\xi_R)^2 \cdot q_R^2}{12} + \lambda_R R_R \tag{13}$$

Similarly with before, the derivation for right view encoding proceeds as follows:

$$\frac{dR_R}{dq_R} = \frac{-2a_R}{q_R \ln 2}, \tag{14}$$

where a_R is a constant. And also

$$\frac{dD_c}{dq_R} = 2\frac{(\xi_R)^2 \cdot q_R}{12}, \tag{15}$$

Thus, the Lagrange parameter for the right view can be formulated as

$$\begin{aligned} \lambda_R &= -\frac{dD_c}{dR_R} \\ &= -\frac{dD_c/dq_R}{dR_R/dq_R} \\ &= \ln 2 \cdot \left[\frac{(\xi_R)^2 \cdot q_R^2}{12 \cdot a_R}\right] \\ &= (\xi_R)^2 \cdot \lambda'_R, \end{aligned} \tag{16}$$

where λ'_R is the Lagrange parameter for traditional SSE-based RDO for the right view and in MV-HEVC it is also obtained as Eq. (11).

IV. Experimental Results

Experimental study has been carried out based on the MV-HEVC reference software, following the common test requirement of the multiview texture-based MV-HEVC in 3DV core experiments [40]. We have implemented the proposed binocular-combination-based 3D RDO for MV-HEVC. The coding parameters used in experiments are shown in TABLE I. In the proposed RDO, the luminance distortion is used for computing the binocular combination distortion, and the chroma distortion is not considered. In our implementation, λ'_R in Eq. (16) is derived from the coding structure of the HM (HEVC Model) reference software [43]. To evaluate the performance of the proposed encoding scheme, the SSE-based RDO in MV-HEVC is chosen as the baseline scheme referred as the *traditional approach*.

TABLE I. Experimental test conditions

Encoding Profile	multiview-main
Intra Period	24
GOP size	8
QP	20, 25, 30, 35
MaxCUWidth	64
MaxCUHeight	64
Coding structure	Hierarchical B
Other parameters	Default setting

Extensive subjective tests were conducted to verify the accuracy of the proposed distortion model. The subjective tests adopted the Double Stimulus Impairment Scale (DSIS) method in [39], in which five grades of the mean opinion score (MOS) were used (0 for the lowest quality and 5 for the highest one). All tests were performed on a 27 inch NVIDIA 3D vision based 3D monitor (ASUS VG278, 144Hz) with seven test sequences from the 3D video coding experiments in MPEG [40] (the specific sequences are illustrated in Fig. 3). For the test sequences, the selected views are listed in TABLE II. A total of 21 subjects participated in the tests, consisting of 14 males and 7 females. The viewers sat in front of the screen with comfortable distance and the field of view was about 30^o. The specific test environment and procedures followed the suggestion of ITU-R BT.2021 [41].

A. Accuracy of binocularly combined distortion measurement

Four DoG models of an given image were used to perform the binocular combination in all experiments. To restrict the

(a)Balloons,1024×768 (b)Kendo, 1024×768 (c)Newspaper, 1024×768

(d) GTFly,1920×1088 (e) PoznanStreet, 1920×1088 (f) UndoDancer, 1920×1088 (g) Shark, 1920×1088

Fig. 3. Test sequences of stereoscopic video.

TABLE II. THE SELECTED LEFT AND RIGHT VIEWS FOR DIFFERENT SEQUENCES

Sequences	Left	Right
Balloons	1	2
Kendo	0	1
Newspaper	2	3
GTFly	1	2
PoznanStreet	3	4
UndoDancer	1	2
Shark	1	2

range of distortion values, the MSE distortion is transformed into Peak Signal to Noise Ratio (PSNR) by using

$$\text{BC-PSNR} = 10 \cdot \log_{10}\left(\frac{255}{\text{BC-distortion}}\right), \tag{17}$$

where BC-distortion denotes the binocularly combined distortion. The values of the binocularly combined distortion in terms of BC-PSNR are further transformed into the subjective MOS values by a symmetrical logistic function [39]. Fig. 4 shows the correlations between the proposed distortion measurements and the measured MOS values for all stereoscopic test videos. It can be observed from Fig. 4 that the proposed distortion measurement model correlates well with the MOS value obtained from the actual subjective evaluation.

The combination coefficients for the Balloons and Shark videos are shown in Fig. 5. Since spatial frequency features for different 3D sequences are different, the gains of binocular frequency combination for different sequences are also different. For video sequences with similar energy distributions in the spatial frequency domain of the two views, the binocular combination tends to obtain the same combination coefficients for the two views. Since motion in temporal domain imposes an impact on image content, the binocular combination coefficient changes with the increasing frame number.

The accuracy of the proposed distortion measurement model is summarized in TABLE III for different quantization levels and test sequences. In TABLE III, the consistency evaluation metrics are: the Pearson correlation coefficient (PCC), Spearman rank correlation coefficient (SRCC), and root mean squared error (RMSE). A low RMSE, high PCC and a high SRCC each suggest high accuracy. To reduce the computational complexity of the proposed distortion measurement model, in our experiments we did not combine multiple frequency bands with multiple pairs of coefficients. Instead, only one pair of coefficients was used over all bands. Results in TABLE III show that the proposed distortion measurement model can characterize the 3D visual quality degradation of compressed stereoscopic video with sufficient accuracy. Specifically, both PCC and SRCC have the average values higher than 0.91, and the RMSE value is lower than 0.03.

In binocular combination, only spatial frequency of luminance component was used, and phase and contrast of the stereoscopic images were not considered as part of the binocular combination process. This may lower the accuracy of the proposed distortion measurement. Although the results of the proposed distortion measurement model do not exactly follow the actually measured MOS values, they still reflect the perceptual quality variations among different compression levels in a quantitative way, meaning that the proposed model can be used for encoding mode selection in the RDO of the stereoscopic video encoder.

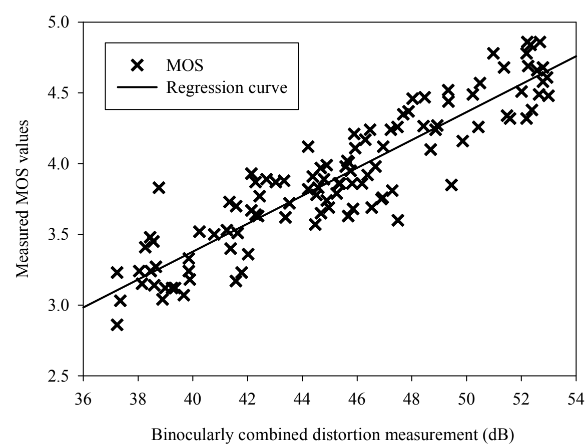


Fig. 4. Correlation between the proposed distortion measurements and the measured MOS values (in horizontal axis, the distortion is characterized by BC-PSNR value).

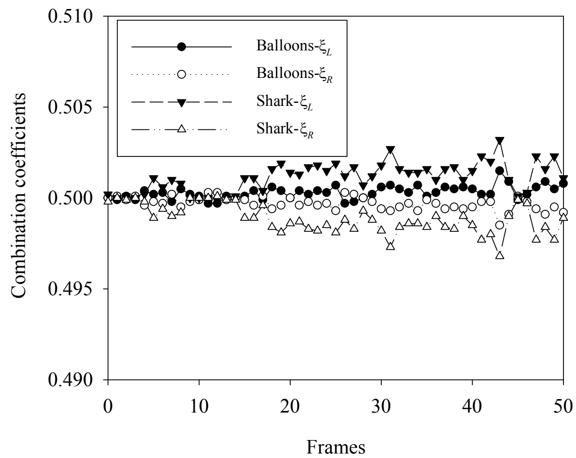


Fig. 5. Binocular combination coefficients for different sequences.

TABLE III. ACCURACY OF THE PROPOSED DISTORTION MEASUREMENT MODEL

Sequences	Binocular combination measurement		
	RMSE	PCC	SRCC
Balloons	0.022	0.9284	0.9236
Kendo	0.033	0.9134	0.9126
Newspaper	0.012	0.9319	0.9322
GTFly	0.036	0.9021	0.9102
PoznanStreet	0.024	0.9223	0.9265
UndoDancer	0.041	0.9043	0.8945
Shark	0.037	0.9156	0.8924
Average	0.029	0.9176	0.9131

B. Binocular perceptual rate-distortion performance

In this subsection we will validate whether the proposed perceptual RDO performs as designed. The symmetric stereoscopic encoding of the Balloons sequence with QP=30 was used to analyze the encoding mode differences among the proposed perceptual RDO and the traditional RDO. Fig. 6 illustrates the subjective quality comparison between the stereoscopic anaglyph (red-cyan) pictures of the proposed RDO and the traditional RDO for the tenth frame of the Balloons sequence (and their coding partition (CU) type comparison). We notice that the proposed approach can achieve a superior subjective 3D quality than the traditional RDO, especially in the areas near certain balloons covered by the dash line rectangle, the boundaries of some balloons in the depth dimension of the perceived cyclopean view are clearer under the proposed RDO scheme due to the fact that a higher number of sub-blocks are selected in those areas in the proposed encoding approach. When comparing results between Fig. 6(a) and Fig. 6(b), the boundary of that balloon pointed by arrow in Fig. 6(a) is clearer than that of Fig. 6(b), indicating that the proposed RDO is able to optimize the perceptual stereoscopic video coding performance by using cyclopean view perception.

In traditional RDO in MV-HEVC, binocular combination is not considered during the stereoscopic video encoding stage. Specifically, the distortion measurement at the encoder adopts the single view video distortion for performing RDO mode selection. In our scheme, the binocular combination distortion of two views is used to assist the right view video coding mode selection during the stereoscopic video coding. Hence, the video coding result correlates with the HVS 3D perception better than the traditional approach. Binocular combination can well explain the process of the cyclopean view formation in human brain. Thus, for the proposed RDO, the video coding modes are selected towards more consistent binocular combination during the formation of the cyclopean view in human brain. Under the traditional RDO, since the encoding qualities of specific areas in two views are not consistent, a two-view image cannot perfectly match the cyclopean view perception of the 3D world (such as the boundary of the balloon to which the arrow points in Fig. 6).

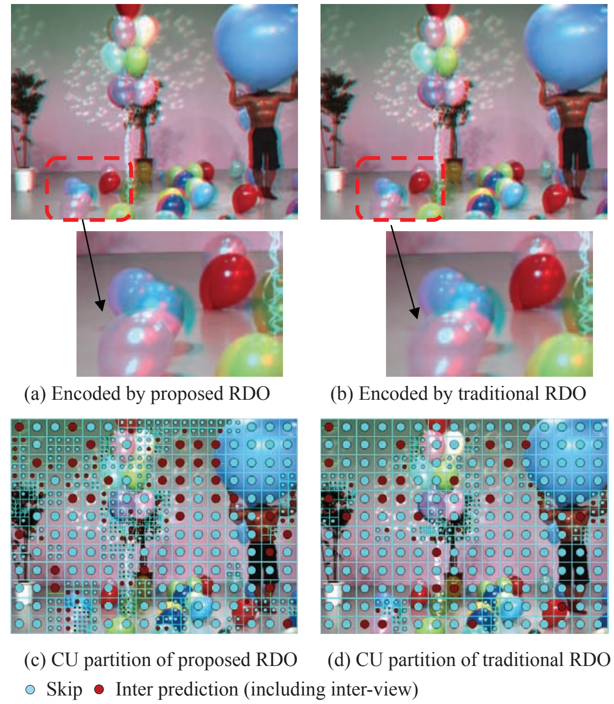


Fig. 6. The subjective stereoscopic anaglyph (red-cyan) quality comparison between the proposed RDO and the traditional RDO for the tenth frame of Balloons sequence, and their coding type comparison.

To verify the statistical performance of the proposed perceptual stereoscopic RDO, both symmetric and asymmetric coding for stereoscopic video were conducted. Since the effectiveness of the BC-PSNR metric has been verified, the coding performance was assessed by both BC-PSNR and subjective evaluation.

Symmetric Stereoscopic Encoding: The symmetric encoding of stereoscopic video can deliver almost the same quality for the two views. Accordingly, the combination of binocular brightness can achieve the cyclopean view with a similar quality. Fig. 7 shows the BC-PSNR performances under different bit-rates for the Balloons and Shark videos for both the proposed perceptual RDO and the traditional RDO. It can be seen from Fig. 7 that the proposed binocular combination-

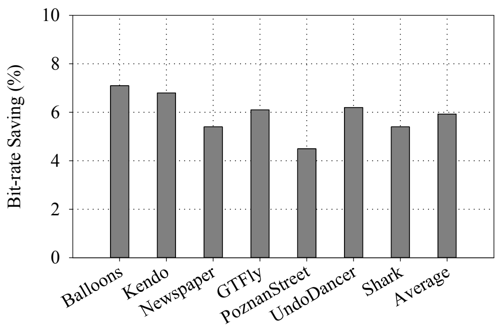


Fig. 11. BD-rate saving results over the traditional RDO.

based RDO can improve the 3D perceptual encoding efficiency. At low bit-rate, the perceptual coding efficiency is improved slightly more than that at high bit-rate, indicating that the quantization-induced distortion at low bit-rate is larger than that at high bit-rate. Thus, the binocularly combined distortion may impose a more significant impact on the coding mode selection of the right view in the proposed perceptual RDO. Fig. 8 illustrates the MOS values over different bit-rates for the Balloons and Shark videos. The subjective test results in Fig. 8 lead to the same conclusions as that of Fig. 7.

Asymmetric Stereoscopic Encoding: In tests of asymmetric stereoscopic encoding, QP gap between two views was set to 8. Fig. 9 shows that the binocularly combined distortion measurement in terms of BC-PSNR under different bit-rates for Balloons and Shark with asymmetric encoding. In Fig. 9, when compared with symmetric encoding, the binocular coding performances for low and high bit-rates are improved with similar increments, because that the binocular combination can efficiently regulate the right view encoding mode to suppress the asymmetric quantization-induced 3D quality degradation for both low and high bit-rate coding. The subjective results in Fig. 10 are consistent with Fig. 9 in terms of the perceptual 3D quality improvements.

The overall performance enhancement of perceptual coding in terms of the BD-rate savings [42] for the symmetric and asymmetric encodings under all test sequences has been summarized in Fig. 11. The average BD-rate saving for all sequences is up to 5.93%. In our current experiments binocular combination is performed over all DoG models and only the overall energies of the two views are involved in computational reduction. Hence, the combination accuracy can be further refined by one-to-one combination over two-view DoG models with higher computational complexity.

C. Encoding complexity

In this work, simulations were performed on a personal computer (Linux CentOS) with Intel Xeon 2.10GHz processor and 12GB random access memory. For the proposed binocular-combination-oriented RDO, the DoG computation is introduced for each pair of stereoscopic video images. The additional DoG computation involves pixel-based filtering computation and it does not increase the computational complexity. Compared with traditional RDO, the increment of computation time is defined as,

$$\Delta T = \frac{T_p - T_o}{T_o} \times 100\%, \tag{18}$$

where T_p and T_o denote the encoding times for the proposed and traditional solutions, respectively. TABLE IV shows the coding time increment for different sequences. It can be observed from TABLE IV that the complexity of the proposed binocular-combination-oriented RDO only increased on average less than 1% over traditional encoding.

V. Conclusion

In this paper, we proposed a binocular perception distortion model by taking considerations of the distortion combination from two views inspired by human brain behaviors. Based on the proposed distortion model, we developed a low-complexity perceptual RDO scheme for enhancing the stereoscopic video encoding efficiency. The proposed scheme jointly optimizes both encoding and viewing processes of stereoscopic video with the proposed 3D perceptual optimization metric for higher encoding efficiency. Subjective tests show that based on binocular combination the proposed perceptual RDO can reduce the BD-rate by up to 5.93% over the traditional RDO in MV-HEVC.

Acknowledgement

The authors would like to thank Prof. Lingfen Sun at Plymouth University for her valuable comments.

References

[1] T. Kim, J. Kang, S. Lee, and 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, Feb. 2014.

[2] M. Lambooij, W. Ijsselsteijn, D. G. Bouwhuis, and I. Heynderickx, “Evaluation of stereoscopic images: Beyond 2D quality”, IEEE Trans. Broadcast., vol. 57, no. 2, pp. 432-444, Jun. 2011.

[3] Z. Wang and A. Bovik, “Mean squared error: Love it or leave it? A new look at signal fidelity measures”, IEEE Signal Process. Mag., vol. 26, no. 1, pp. 98117, Jan. 2009.

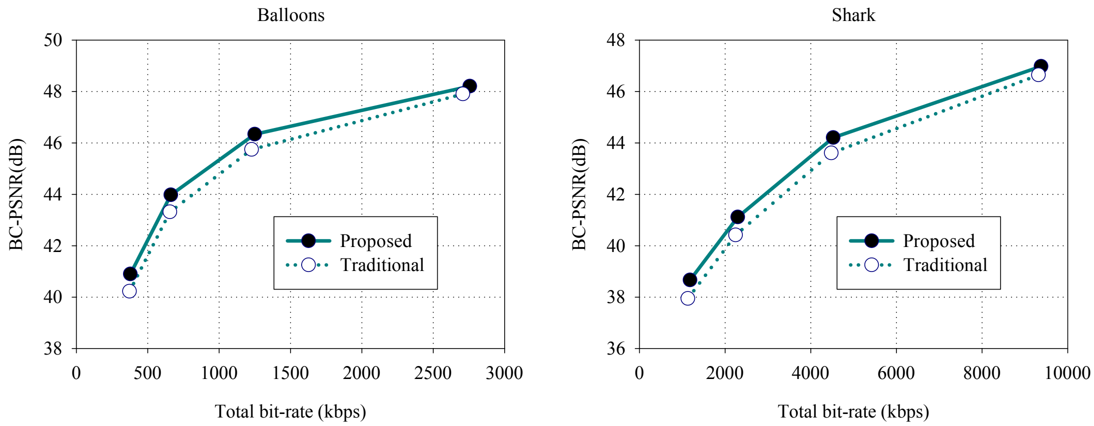
[4] Y. Liu, S. Ci, H. Tang, Y. Ye, and J. Liu, “QoE-oriented 3D Video Transcoding for mobile streaming”, ACM Transactions on Multimedia Computing, Communications and Applications, vol.8, no.3s, article 42, 20 pages, 2012.

[5] F. Shao, W. Lin, S. Gu, G. Jiang, and T. Srikanthan, “Perceptual full reference quality assessment of stereoscopic images by considering binocular visual characteristics”, IEEE Trans. Image Process., vol. 22, no. 5, pp. 1940-1953, May 2013.

[6] M. J. Chen, C. C. Su, D. L. Kwon, L. K. Cormack, and A. C. Bovik, “Full-Reference Quality Assessment of Stereopairs Accounting for Rivalry”, Signal Processing: Image Commun., vol. 28, no. 9, pp. 1143-1155, Oct. 2013.

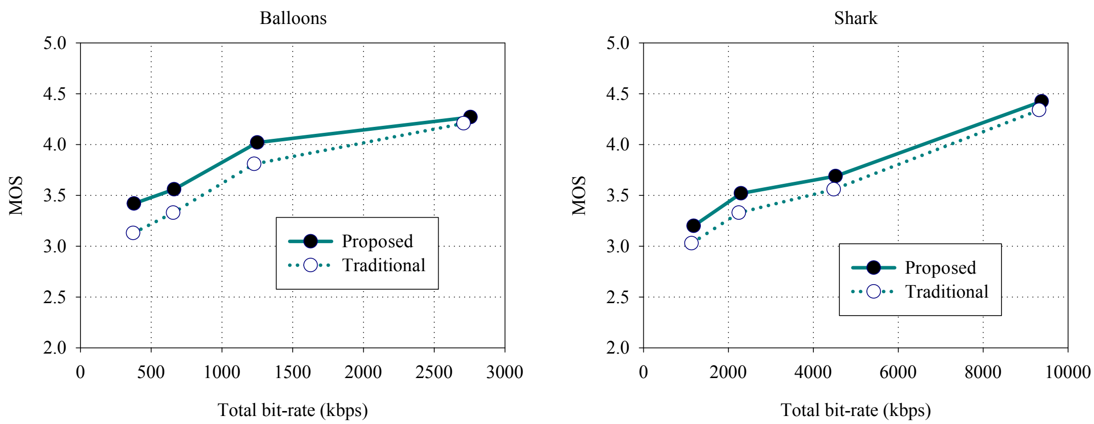
[7] K. Lee and S. Lee, “3D Perception Based Quality Pooling: Stereopsis”, Binocular Rivalry, and Binocular Suppression, IEEE Journal of Selected Topics in Signal Processing, vol. 9, no. 3, pp. 533-545, April 2015.

[8] S. L. P. Yasakethu, C. T. E. R. Hewage, W. A. C. Fernando, and A. M. Kondoz, “Quality analysis for 3D video using 2D video quality models”, IEEE Trans. Consumer Electron., vol. 54, no. 4, pp. 1969-1976, Nov. 2008.



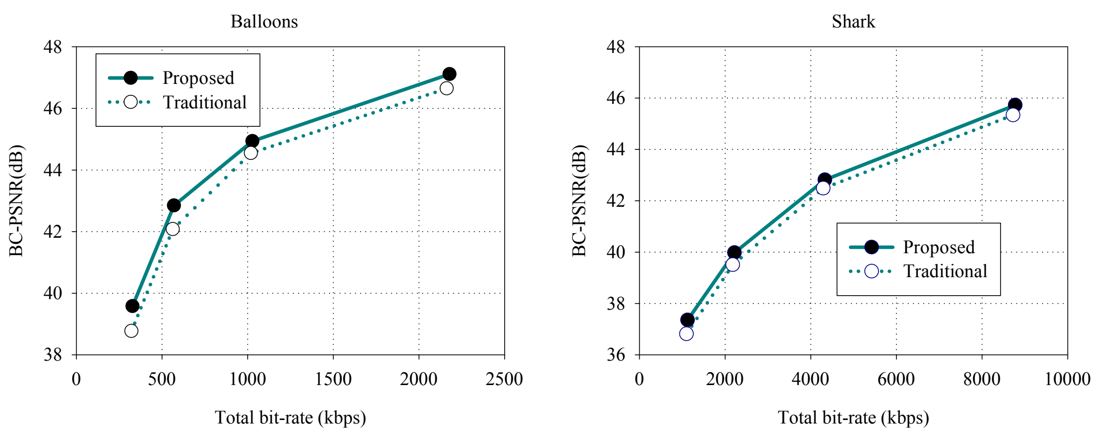
(a) (b)

Fig. 7. The BC-PSNR performances under different bit-rates for Balloons and Shark.



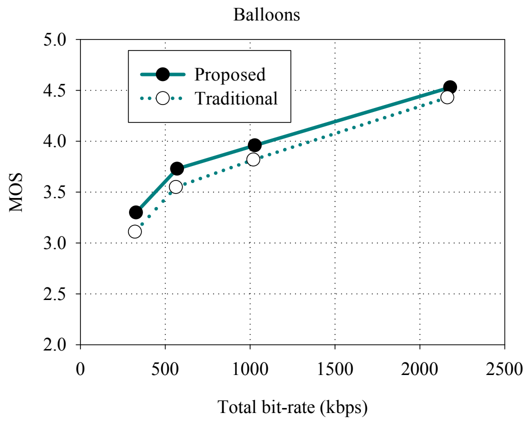
(a) (b)

Fig. 8. The subjective test results under different bit-rates for Balloons and Shark.

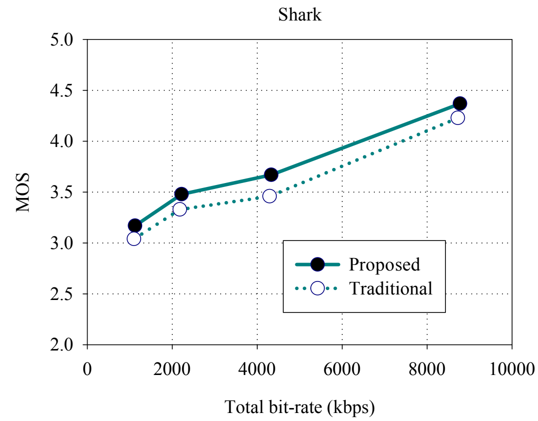


(a) (b)

Fig. 9. The BC-PSNR performances under different bit-rates for asymmetric encoding.



(a)



(b)

Fig. 10. The subjective test results under different bit-rates for Balloons and Shark for asymmetric encoding.

TABLE IV. ENCODING TIME INCREMENT (%)

Sequences	Symmetric encoding				Asymmetric encoding			
	20	25	30	35	20	25	30	35
Balloons	1.38	1.27	1.21	1.12	1.22	1.11	1.23	1.09
Kendo	1.25	1.34	1.41	1.35	1.56	1.67	1.44	1.65
Newspaper	1.48	1.72	1.37	1.31	1.29	1.36	1.21	1.22
GTFly	0.31	0.3	0.27	0.35	0.34	0.31	0.32	0.27
PoznanStreet	0.32	0.35	0.31	0.30	0.34	0.33	0.29	0.30
UndoDancer	0.27	0.34	0.31	0.32	0.33	0.30	0.27	0.29
Shark	0.26	0.44	0.36	0.31	0.30	0.32	0.39	0.32
Average	0.75	0.83	0.75	0.72	0.77	0.77	0.74	0.73

[9] A. Boev, A. Gotchev, K. Egiazarian, A. Aksay, and G. B. Akar, “Toward compound stereo-video quality metric: A specific encoder-based framework”, in Proc. IEEE Southwest Symp. Image Anal. Interpret., Jun. 2006, pp. 218-222.

[10] M.-J. Chen, L.K. Cormack and A.C. Bovik, “No-reference quality assessment of natural stereopairs”, IEEE Transactions on Image Processing, vol. 22, no. 9, pp. 3379-3391, Sep. 2013.

[11] L. Jin, A. Boev, K. Egiazarian and A. Gotchev, “Quantifying the importance of cyclopean view and binocular rivalry-related features for objective quality assessment of mobile 3D video”, EURASIP Journal on Image and Video Processing, 2014:6.

[12] J. Park, K. Seshadrinathan, S. Lee, and A. C. Bovik, “VQPooling: Video quality pooling adaptive to perceptual distortion severity”, IEEE Trans. Image Process., vol. 22, no. 4, pp. 610-620, Feb. 2013.

[13] A. K. Moorthy and A. C. Bovik, “Visual importance pooling for image quality assessment”, IEEE J. Sel. Topics Signal Process., vol. 3, no. 2, pp. 193-201, Apr. 2009.

[14] V. De Silva, H. K. Arachchi, E. Ekmekcioglu, A. Kondoz, “Toward an Impairment Metric for Stereoscopic Video: A Full-Reference Video Quality Metric to Assess Compressed Stereoscopic Video”, IEEE Transactions on Image Processing, vol. 22, no.9, pp. 3392-3404, Sep. 2013.

[15] Y. Lin, J. Wu, “Quality assessment of stereoscopic 3D image compression by binocular integration behaviors”, IEEE Transactions on Image Processing, vol. 23, no. 4, pp. 1527-1542, April 2014.

[16] A. Vetro, T. Wiegand, and G. J. Sullivan, “Overview of the Stereo and Multiview Video Coding Extensions of the H.264/AVC Standard”, Proc. of the IEEE, vol. 99, no. 4, 626-642, April 2011.

[17] G. Tech, Y. Chen, K. Müller, et al., “Overview of the Multiview and 3D Extensions of High Efficiency Video Coding”, IEEE Transactions on Circuits and Systems for Video Technology, vol. 26. no. 1, pp. 35-49, January 2016.

[18] G. J. Sullivan and T. Wiegand, “Rate-distortion optimization for video compression”, IEEE Signal Process. Mag., vol. 15, no. 6, pp. 74-90, Nov. 1998.

[19] E. H. Yang and X. Yu, “Rate distortion optimization for H.264 inter-frame video coding: A general framework and algorithms”, IEEE Trans. Image Process., vol. 16, no. 7, pp. 1774-1784, Jul. 2007.

[20] S. Ma, S. Wang, S. Wang, L. Zhao, Q. Yu, and W. Gao, “Low complexity rate distortion optimization for HEVC”, IEEE Data Compression Conference (DCC), 2013, pp. 73-82.

[21] S. Wang, A. Rehman, Z. Wang, S. Ma and W. Gao, “SSIM-Motivated Rate-Distortion Optimization for Video Coding”, IEEE Transactions on Circuits and Systems for Video Technology, vol. 22, no.4, pp. 516-529, April 2012.

[22] H. Wu, W. Lin, K. N. Ngan, “Rate-perceptual-distortion optimization (RpDO) based picture coding issues and challenges”, 19th International Conference on Digital Signal Processing, 2014.

[23] H. Iravani, M. Hashemi, S. Shirmohammadi, M. Gabbouj, “A new rate distortion model for multi-view/3D video coding”, IEEE International Conference on Multimedia and Expo workshop 2013.

[24] W. Kim, A. Ortega, P. Lai, D. Tian, “Depth Map Coding Optimization Using Rendered View Distortion for 3D Video Coding”, IEEE Transactions on Image Processing, vol. 24, no. 11, pp. 3534-3545, 2015.

[25] 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, January 2015.

[26] S. Valizadeh, P. Nasiopoulos, R. Ward, “Perceptual distortion measurement in the coding unit mode selection for 3D-HEVC”, The International Conference on Consumer Electronics, pp.347-350, 2016.

[27] P. A. Anderson, J. A. Movshon, “Binocular combination of contrast signals”, Vision Research, vol. 29, no. 9, pp. 1115-1132, 1989.

[28] D. H. Baker, S. A. Wallis, M. A. Georgeson, and T. S. Meese, “Nonlinearities in the binocular combination of luminance and contrast”, Vision Research, vol. 56, pp. 1-9, Mar. 2012.

[29] S. Grossberg and F. Kelly, “Neural dynamics of binocular brightness perception”, Vis. Res., vol. 39, no. 22, pp. 3796-3816, 1999.

[30] B. Julesz, Foundations of Cyclopean Perception. Univ. Chicago Press, 1971.

[31] G.R. Engel, “The visual processes underlying binocular brightness summation”, Vision Research, vol. 7, no. 9, .pp. 753-767, Sep. 1967.

[32] D. W. Curtis, S. J. Rule, “Binocular processing of brightness information: a vector-sum model”, Journal Journal of Experimental Psychology Human Perception & Performance, vol. 4, no. 1, pp. 132-143, March 1978.

[33] A. I. Cogan, “Human binocular interaction: towards a neural model”, Vision Research, vol .27, no. 12, pp. 2125-2139, 1987.

[34] J. Ding and G. Sperling, “A gain-control theory of binocular combination”, Proc. Nat. Acad. Sci. United States Amer., vol. 103, no. 4, pp. 1141-1146, 2006.

[35] C.-B. Huang, J. Zhou, Z.-L. Lu, Y. Zhou, “Deficient binocular combination reveals mechanisms of anisometropic amblyopia: Signal attenuation and interocular inhibition”, Journal of Vision, vol. 11, no. 6, pp. 1-17, 2011.

[36] H. R. Wilson and J. R. Bergen, “A four mechanism model for threshold spatial vision”, Vis. Res., vol. 19, no. 1, pp. 19-32, 1979.

[37] G. Mather, Foundations of Sensation and Perception. Oxon, U.K., Psychology Press, 2008.

[38] T. Wiegand, B. Girod, “Lagrangian multiplier selection in hybrid video coder control”, IEEE International conference on Image Processing, pp. 542-545, Oct. 2001.

[39] International Telecommunication Union (ITU), “Methodology for the Subjective Assessment of the Quality of Television Pictures”, ITU Recommendation BT.500-13, Jan. 2012.

[40] D. Rusanovskyy, K. Müller, and A. Vetro, “Common test conditions of 3DV core experiments”, JCT3V-E1100, Vienna, Austria, Aug. 2013.

[41] International Telecommunication Union (ITU), “Subjective Methods for the Assessment of Stereoscopic 3DTV Systems”, ITU-R BT.2021, May 2012.

[42] G. Bjontegard, “Calculation of Average PSNR Differences Between RD Curves”, document Rec. VCEG-M33, ITU-T Q.6/SG16, Apr. 2001.

[43] I.-K. Kim, K. D. McCann, K. Sugimoto, B. Bross, W.-J. Han and G. J. Sullivan, “High Efficiency Video Coding (HEVC) Test Model 13 (HM13) Encoder Description”, Document: JCTVC-O1002, Joint Collaborative Team on Video Coding (JCT-VC) of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, 15th Meeting: Geneva, CH, 23 Oct. - 1 Nov., 2013.

Yanwei Liu received the B.S. degree in applied geophysics from Jianghan Petroleum University, China, in 1998, the M.S. degree in computer science from China Petroleum University (Beijing) in 2004 and the Ph.D. degree in computer science from Institute of Computing Technology, Chinese Academy of Sciences in 2010. Currently, he is an Associate Professor with the Institute of Information Engineering, Chinese Academy of Sciences. His research interests include digital image/video processing, multiview and 3D video processing, multimedia communication and wireless video communication. He has published over 50 scientific papers. He serves on the TPCs of several international conferences in the area of multimedia, communications, and networking.

Jinxia Liu received the B.S. degree and M.S. degree in physics from Harbin Normal University, China, in 1994, and 2005, respectively. In 2005, she joined the Zhejiang Wanli University as an Associate Professor. Currently, she is a Professor with Zhejiang Wanli University. Her research interests include laser imaging, digital image/video processing, multiview and 3D video coding, and wireless communication.

Antonios Argyriou (S’00-M’07-SM’15) received the Diploma degree in electrical and computer engineering from Democritus University of Thrace, Xanthi, Greece, and the M.S. and Ph.D. degrees in electrical and computer engineering (as a Fulbright Scholar) from Georgia Institute of Technology, Atlanta, GA, USA, in 2001, 2003, and 2005, respectively. Currently, he is an Assistant Professor with the Department of Electrical and Computer Engineering, University of Thessaly, Volos, Greece. From 2007 to 2010, he was a Senior Research Scientist with Philips Research, Eindhoven, The Netherlands. From 2004 to 2005, he was a Senior Engineer with Soft.Networks, Atlanta, GA, USA. He currently serves on the Editorial Board of the Journal of Communications. His research interests include wireless communication systems and networks, and video delivery. He has also served as Guest Editor for the IEEE TRANSACTIONS ON MULTIMEDIA Special Issue on Quality-Driven Cross-Layer Design, and he was also a Lead Guest Editor for the Journal of Communications, Special Issue on Network Coding and Applications. He serves on the TPCs of several international conferences and workshops in the area of communications, networking, and statistical signal processing.

Song Ci (S’98–M’02–SM’06)received the B.S. degree from the Shandong University of Technology (now Shandong University), Jinan, China, in 1992, the M.S. degree from the Chinese Academy of Sciences, Beijing, China, in 1998, and the Ph.D. degree from the University of Nebraska-Lincoln, Lincoln, NE, USA, in 2002, all in electrical engineering. He is currently an Associate Professor with the Department of Electrical and Computer Engineering at the University of Nebraska-Lincoln. His current research interests include dynamic complex system modeling and optimization, content-aware quality-driven cross-layer optimized multimedia over wireless, Energy Internet and green computing. Dr. Ci serves as Editor, Guest Editor or Associate Editor in editorial boards of many journals, including the IEEE TRANSACTIONS ON CSVT, the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, and the IEEE ACCESS. He also serves as the TPC member for numerous conferences. He is a senior member of IEEE and a member of ACM and AAAS.