J. Vis. Commun. Image R. 38 (2016) 246–255

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

Binocular perception based reduced-reference stereo video quality assessment method ☆

Mei Yu [a,b], Kaihui Zheng [a], Gangyi Jiang [a,b,*], Feng Shao [a], Zongju Peng [a]

[a] Faculty of Information Science and Engineering, Ningbo University, Ningbo, China
[b] National Key Lab of Software New Technology, Nanjing University, Nanjing, China

ARTICLE INFO

Article history:
Received 4 September 2015
Revised 3 January 2016
Accepted 7 March 2016
Available online 9 March 2016

Keywords:
Stereo video quality assessment
Binocular vision
Temporal characteristics
Reduced-reference frame
Motion intensity

ABSTRACT

A new reduced-reference (RR) stereo video quality assessment method is proposed in this paper by considering temporal characteristics of video and binocular perception in human visual system (HVS). Firstly, motion intensity is utilized to extract RR frames for the purpose of temporal characteristics in stereo video. Secondly, according to internal generative mechanism of HVS, fusion and rivalry in the process of binocular perception is modeled, and the RR frames are divided into binocular fusion portion and binocular rivalry portion. Then, RR frame quality indicators are computed for these two portions. Finally, the RR frame quality indicators of the original and distorted frames are compared. A temporal pooling strategy is utilized on these quality indicators to obtain final stereo video quality score, where the motion intensity is used for toning the pooling parameters. Experimental results show that the proposed method has better performances when compared to other state-of-the-art quality assessment methods.

1. Introduction

Three dimensional (3D) video becomes popular in digital multimedia in last decade [1,2]. These 3D contents can bring users stereo perception and immersive viewing experiences, and they have been displayed not only in theaters, but also on television, portable or other kinds of consumer electronic devices. Thus, for the purpose of higher quality of 3D video content generation and processing, stereo video quality assessment (SVQA) becomes more and more important [3–5].

In this new paradigm of image quality assessment, some works have been done in different approaches. For example, full-reference (FR) methods in classical manner with complete reference information from the original content were proposed [6,7]. However, the disadvantage of FR methods is that the reference images/videos are not usually accessible in most practical situations. As a solution, no-reference (NR) methods were proposed without the original reference [8]. But many existing NR methods may be only effective for some specific image/video contents or specific types of distortions. Reduced-reference (RR) method is an alternative of the above two kinds of methods, depending on partial information of the original contents [9]. Compared with the FR method, the RR method greatly reduces the reference data to be transmitted; and taking advantage of the reference data, the RR method ensures the evaluation effectively and universally in comparison with the NR method.

Different from viewing single viewpoint images/videos, human visual system (HVS) can perceive the difference between two retinal images to create a mental image with depth perception, which is the result of two binocular interactions, i.e., binocular fusion and suppression. Thus the perceptual quality of stereo image/video with two views is also different from the one obtained by using simple weighting of the two views' two dimensional (2D) quality in stereo image/video. The additional dimension, depth information which is generated from the left and right views in stereo image/video, needs to be carefully considered. Some SVQA methods were proposed through stereo image quality assessment (SIQA) models. Seo et al. proposed a FR objective SVQA metric by using blocking artifacts, blurring in edge regions and video quality difference between two views [10]. The first two terms describe the distortion in compressed video, while the third is used to represent 3D effect of stereoscopic video. Jin et al. proposed a FR SIQA model for mobile 3D video [11], which presents three quality components to measure the image quality, including the cyclopean view, binocular rivalry, and the scene geometry, respectively; the final 3D image quality is assessed through a machine learning

☆ This paper has been recommended for acceptance by Wiesi Lin.

* Corresponding author at: Faculty of Information Science and Engineering, Ningbo University, Ningbo, China.
E-mail address: jianggangyi@126.com (G. Jiang).

http://dx.doi.org/10.1016/j.jvcir.2016.03.010
1047-3203/

approach. Hewage et al. proposed a RR 3D image quality metric by extracting edges and contours of depth map [12]. Malekmohamadi et al. proposed a RR quality metric for 3D video by extracting side information from edge properties and gray level co-occurrence matrices from color and depth sections [13]. Taking into account the spatial information in video quality assessment (VQA), Ma et al. proposed a RR VQA model [14], which not only considers spatial statistical characteristics of video, but also temporal statistical characteristics. Soundararajan et al. proposed RR VQA models by utilizing spatial and temporal entropic differences [9], in which a Gaussian scale mixture model is used to measure the amount of spatial and temporal information differences. These RR models may improve the evaluation accuracy by utilizing temporal information.

In this paper, considering demands of practical applications and human visual perception on spatial and temporal characteristics of stereo video, a new RR-SVQA method is proposed, which consists of three parts. Firstly, according to the analysis of temporal characteristics of video, motion intensity is defined and used to extract the RR frame pairs from stereo video so as to reduce the amount of data to be processed in quality assessment while keep the temporal characteristic of video, and then these RR frames are decomposed into binocular fusion portion (BFP) and binocular rivalry portion (BRP). Secondly, the BFP is used to construct the cyclopean view, and then the generalized Gaussian distribution (GGD) features are extracted from the cyclopean view and the BRP. Finally, the quality indicators of BFP and BRP of stereo video are obtained. These quality indicators are pooled in spatial and temporal domains to get the final stereo video quality score. The proposed method is tested on the NAMA3DS1-COSPAD1 stereo video database, and the corresponding experimental results show that the proposed method has more excellent performance and better consistency with human visual perception compared with other start-of-art methods.

The remainder of this paper is organized as follows. In Section 2, the proposed RR SVQA method is described in detail. In Section 3, the experimental results are given and the performance indices are discussed. Finally, Section 4 concludes the paper.

2. The proposed RR-SVQA method

Considering human visual perception of temporal variability and binocular perception characteristics of stereo video, a new RR SVQA method is proposed, and its framework is shown as Fig. 1. The proposed method mainly consists of three parts, including (1) selection of RR frame for reflecting temporal variability of video, (2) extracting visual perception RR features in BFP and BRP of stereo video, and finally (3) quality indicator calculation and pooling for stereo video.

Since there are strong correlations among successive frames of stereo video, the RR frames are defined and selected from the original and distorted stereo videos to represent the videos, so as to decrease the amount of data to be processed in the following quality assessment while describe the temporal variability of the videos. In addition, according to binocular perception characteristics, the RR frames of stereo view are decomposed into two portions, that is, BFP and BRP, which resulted from binocular fusion and binocular rivalry of binocular vision. Then, multi-channel decomposition is applied on these portions and the GGD model [15] is used to normalize the coefficients to gain the RR features. Then, at the sender, by compression coding, some RR features in original video are sent to the receiver as the side information of H.264 standard compatible video stream. Finally, at the receiver, the quality of these two portions is computed and pooled as the final stereo video quality score.

2.1. Extraction of the RR frame

For a video, when a series of high correlation consecutive images are played continuously, if the correlation coefficient between the images is quite close or equal to 1, it can be understood as almost still image. But if the correlation between the frames before and after the current frame is weak, perceptual difference between the frames will be large, which implies that the played video is significantly varying. Thus, inter-frame difference can be used to represent the motion intensity of a video. The inter-frame motion is an important feature to describe temporal variability of video. In SVQA research, the existing methods pay little attention to temporal characteristics of video. However, it has been known that the temporal, spatial, and quantization variations will impact the perceived quality of video [16]. When the motion intensity in video is not strong, video's spatial features show higher significance. With the movement being stronger, the temporal features will play a more important role in the quality of video. Therefore, we put forward the following scheme for extracting the RR frame.

2.1.1. Motion intensity of video

As one of the most important temporal characteristics of video, motion intensity of video will be defined to describe temporal variability of video and considered in designing a new RR SVQA method. For videos with different motion intensity, the redundancy between the frames before and after the current frame is also different. If motion in video is slight, there are more temporal redundancies in the video; on the contrary, there are less temporal redundancies. Hence, the different RR frame selection strategy should be executed on video with different motion intensity. Considering HVS being sensitive to motion intensity, motion intensity

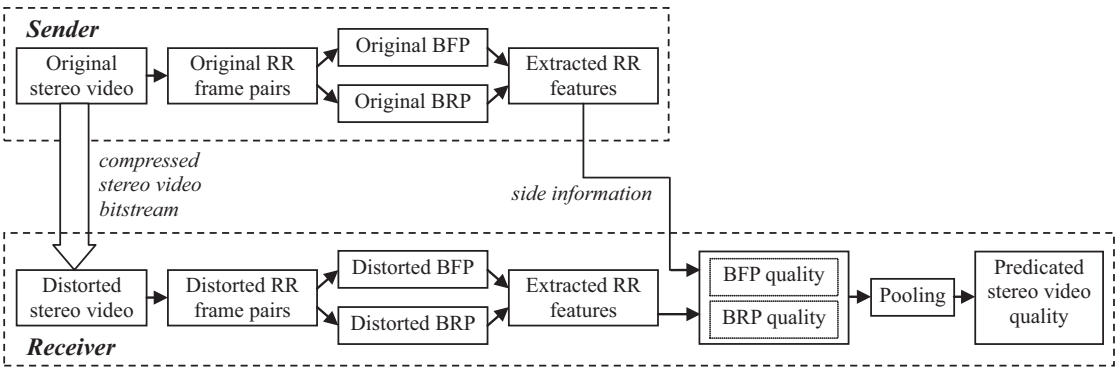

Fig. 1. Framework of the proposed RR SVQA method.

of stereo video is divided into three levels, strong, general and weak, respectively. Let T denote motion intensity of a group-of-pictures (GOP) in video, which is computed by

$$T = \begin{cases} 2, & \text{if } v_i > v_{mean} \text{ and } m_i > m_{mean} \\ 0, & \text{if } v_i < v_{mean} \text{ and } m_i < m_{mean} \\ 1, & \text{otherwise} \end{cases} \quad (1)$$

where v_i is the average motion vector of the i-th GOP and v_{mean} is the mean of the whole video's motion vectors. m_i is the variance of the i-th GOP's motion vectors and m_{mean} is the variance of the whole video's motion vectors. These four parameters are computed in a way of scalar type. For example, the x-component of m_i, denoted as m_i^x, is computed by

$$m_i^x = \frac{1}{L}\sum_{j=1}^{L}(mv_j^x - v_{mean}^x)^2 \quad (2)$$

where mv_j^x is the x-component of the average motion vector of the j-th frame in the i-th GOP, v_{mean}^x is the x-component of the v_{mean}, and L is the number of frames in a GOP, here, L is empirically set to 16. The y-component of m_i, denoted as m_i^y, can be computed similarly.

In Eq. (1), $T = 2$ implies that the motion intensity of video is strong, $T = 1$ means the general level of motion intensity, while $T = 0$ means that the motion intensity is weak.

Using 'Barrier gate' video in NAMA3DS1-COSPAD1 stereo video database [17] as an example, some test results are shown in Fig. 2. Fig. 2(a) is the first frame of left view in 'Barrier gate' video, and Fig. 2(b) and (c) shows the average motion vectors and motion intensities of each GOP in the left view of 'Barrier gate' video. In Fig. 2(b) and (c), the horizontal axis is frame number and the vertical axis is the average motion vector or motion intensity, respectively. 'Barrier gate' video describes a car coming from far away until it is out of the gate and the motion intensity changes from weak to strong. From Fig. 2(b), the variation of motion vector reflects such situation clearly. For interval from the 200-th to 250-th frames, there is a valley because the car stops and is waiting until the door is open. In Fig. 2(c), the levels of motion intensity are classified efficiently.

2.1.2. RR frame selection scheme

The RR frame is defined as one frame extracted from GOP in video, which can represent primary characteristics of the GOP. The high correlation between the frames before and after the current frame results in massive redundancy. Thus, a GOP with similar frames can be completely substituted by one frame in the GOP to reduce redundancy and keep characteristics of the GOP. So, how to select one frame in a GOP as the RR frame is a key step for the proposed RR-SVQA method. Four test videos, 'Barrier gate', 'Basket', 'Hall' and 'Phone', from the NAMA3DS1-COSPAD1 stereo video database are used to determine the RR frame selecting strategy. Fig. 3 shows the left views of the four test videos.

Let S_R denote selecting rate of the RR frame in a GOP of video. The test steps to determine the RR frame are designed as follows

(1) In GOPs of a video, different RR frame selection schemes with different selecting rates, S_R, are tested. For GOP with the length of L, S_R can be set to 1, 1/2, 1/4, . . ., or 1/L, respectively, in which 1 means that each frame is the RR frame, while 1/2 means that half of the L frames are the RR frame. Similarly, 1/4 means that selecting one frame from every 4 frames as the RR frame, and so on.
(2) Compute the objective quality of the selected RR frames and fuse them as the objective evaluation quality.
(3) Then, calculate correlation coefficients (CC) between objective evaluation quality and subjective mean opinion score (MOS) of these videos. Here, two classical objective quality metrics, peak signal to noise ratio (PSNR) and structural similarity index (SSIM) [18], are applied.
(4) The mean and variance about CC with respect to different RR frame selecting rates are computed.
(5) Finally, the RR frame selecting rates are determined with respect to different motion intensity. If L is 16, when $T = 0$ and $T = 2$, the RR frame selecting rate S_R 1/16 is applied; while for $T = 1$, 1/4 is a suitable selecting rate of the RR frame.

As an example, four videos distorted with H.264 compression, named 'Barrier gate', 'Basket', 'Hall' and 'Phone call', are used to test the proposed RR frame selection scheme. The mean and variance about CC with respect to different RR frame selecting rates corresponding to different levels of motion intensity are shown in

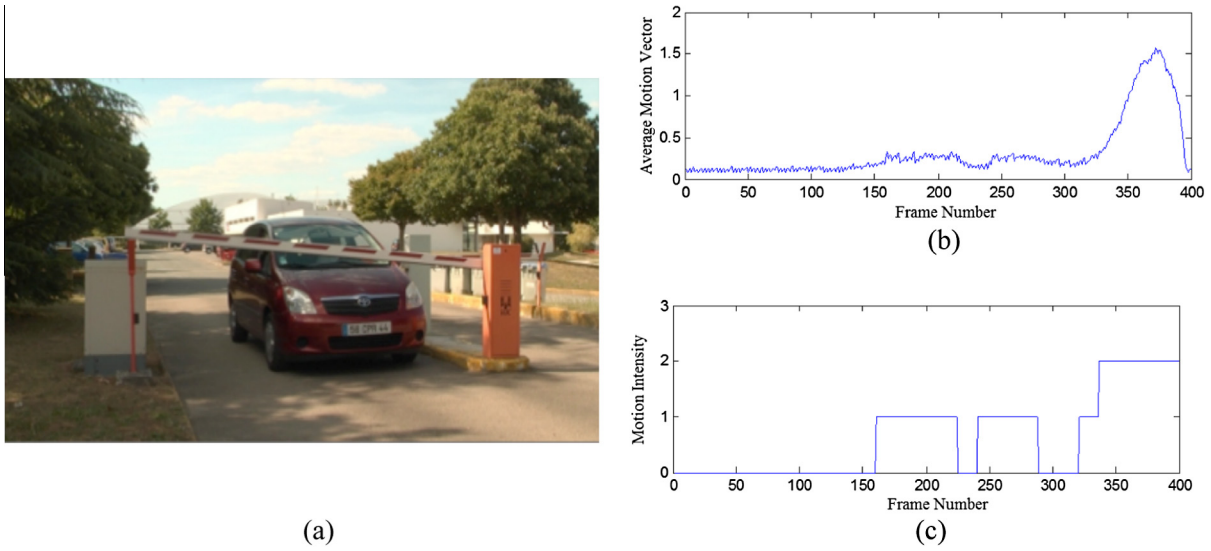

Fig. 2. Motion intensities of 'Barrier gate' video. (a) Left view of 'Barrier gate' video. (b) Average motion vectors of left view of 'Barrier gate'. (c) Motion intensities of left view of 'Barrier gate'.

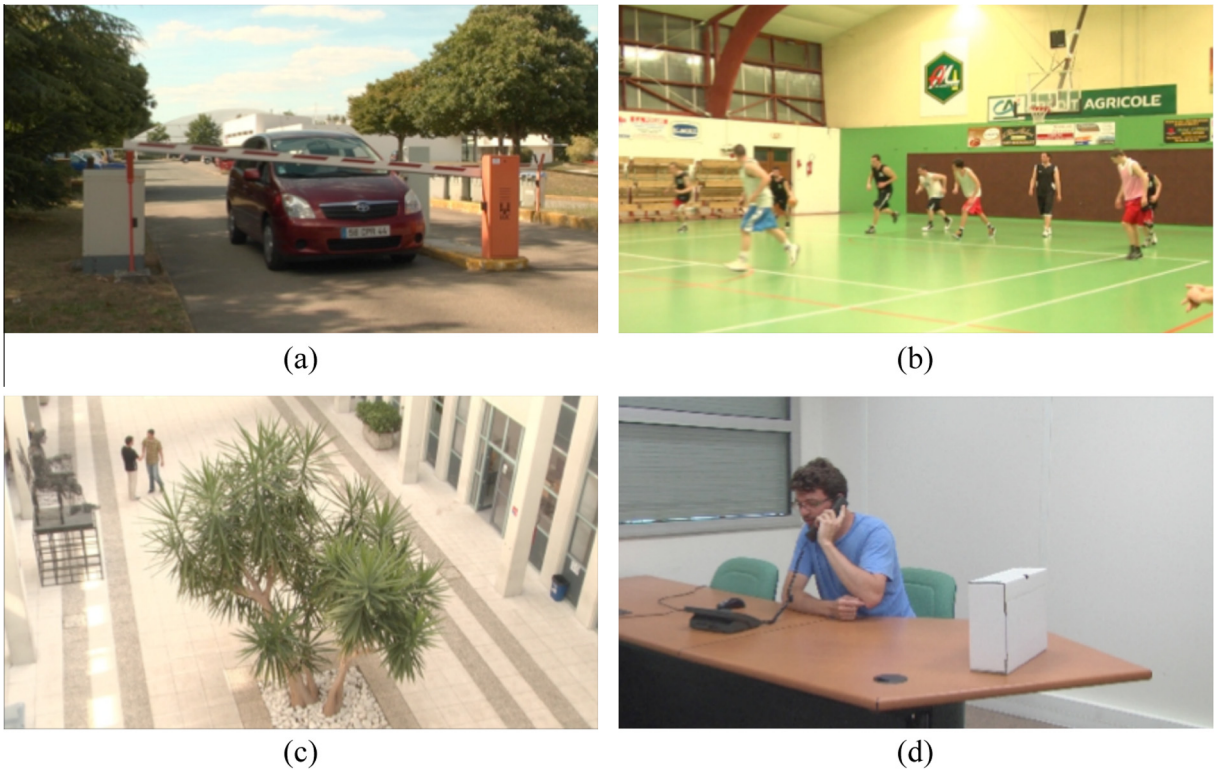

Fig. 3. Four test videos in the NAMA3DS1-COSPAD1 stereo video database [17]. (a) Left view of 'Barrier gate' video. (b) Left view of 'Basket' video. (c) Left view of 'Hall' video. (d) Left view of 'Phone call' video.

Fig. 4. Since the mean about CC with respect to different RR frame selecting rates indicates the consistency between the objective evaluation quality and subjective MOS, thus, from Fig. 4 (a) and (c), it is clear that video clips with different levels of motion intensity at different selecting rates make different contributions to quality assessment. On the other hand, the variance about CC with respect to different S_Rs describes stability of the RR frame selection. For video clip with $T = 0$, either mean or variance about CC with respect to different S_Rs are very stable, no matter for PSNR or for SSIM. The fold lines of the variance about CC of $T = 0$ in Fig. 4 (b) and (d) are close to 0. Thus for GOPs with weak motion intensity, the selecting rate 1/16 is enough for maintaining accuracy of quality assessment. When $T = 1$, if S_R is not less than 1/4, the mean and variance are relatively stable. But when S_R is less than or equal to 1/8, there is an obvious mutation. In Fig. 4(a) and (c), the mean about CC with respect to selecting rates 1/8 and 1/16 drops sharply, which means that evaluation accuracy will be impacted seriously if selecting rate 1/8 or 1/16 is utilized in this case. When $T = 2$, the motion intensity is the relatively most. The mean and variance about the CC are also stable, similar with the situation of $T = 0$, but the mean about CC is lower than that of $T = 0$, as well as that of $T = 1$ at large selecting rates. This means that the contribution of video clips with $T = 2$ are relatively smaller to quality of the video, compared with video clips with the other two motion intensity.

2.2. Binocular visual perception model

The study on binocular visual perception shows that visual information is handled by two visual pathways at the same time, cyclopean view and binocular view, respectively [11]. The binocular visual perception model with cyclopean view and binocular view is depicted in Fig. 5.

Theoretical studies on internal generative mechanism (IGM) [19] have shown that the brain actively predicts the visual sensation and avoids the residual uncertainty. Based on the IGM theory, Wu et al. proposed a model to decompose image into predicted portion and disorderly portion [20]. The distortions on the predicted portion will damage the primary visual information, such as blur the edge or destroy the structure. Chen et al. put forward that binocular rivalry occurs when the two eyes view mismatched images at the same retinal location [11]. If the stimuli received by human's two eyes are sufficiently different from each other, it will causes match failures or to otherwise affect stereo perception. However, the distortion on the disorderly portion is somewhat content independent, which do not disturb the inference of the primary visual information and mainly arouse uncomfortable perception. In stereo video, the uncomfortable perception will cause the binocular rivalry. In the processing of predicting sensory information, the binocular fusion proceeds easily, frequently and naturally. However, brain tries to avoid residual uncertainty information which causes chaos in the human eye perception. Hence, for binocular perception the predicted portion is primary visual region and it is the portion to fuse into a cyclopean view easily. The disorderly portion cannot be fused well, and binocular rivalry often occurs on this portion. It is the portion to present the binocular view. The predicted portion and disorderly portion of frames are obtained by the IGM model. Here, the predicted portion is defined as BFP and the disorderly portion as BRP in Fig. 1.

2.2.1. Cyclopean view model of binocular fusion

As a binocular fusion case, cyclopean view model is used to depict binocular perception. The usual cyclopean view is usually created by matching the left and right views directly. But these models will cause the loss of texture detail information.

Here, we present a new generation scheme of cyclopean view which obtains the cyclopean view by matching the BFPs of the left

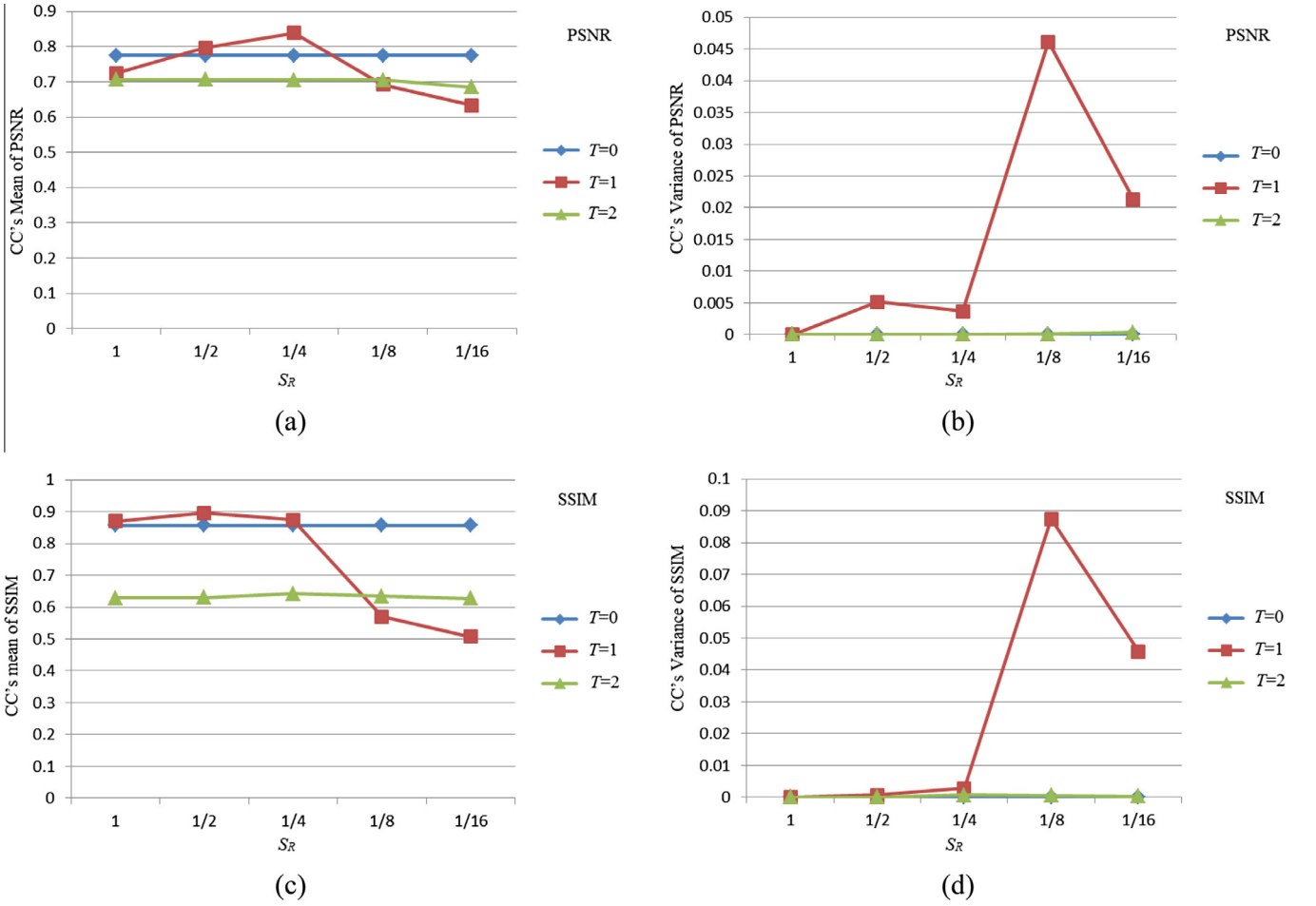

Fig. 4. Mean and variance about CC of the different RR frame selection schemes. (a) CC's mean of PSNR values in different levels of motion intensity. (b) CC's variance of PSNR values in different levels of motion intensity. (c) CC's mean of SSIM values in different levels of motion intensity. (d) CC's variance of SSIM values in different levels of motion intensity.

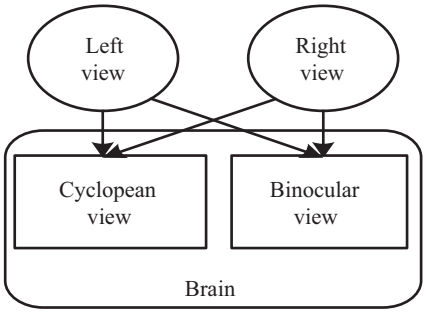

Fig. 5. Binocular visual perception model with cyclopean view and binocular view.

and right views. Through IGM decomposition, stereo views are divided into BFP and BRP. As mentioned above, the BFP is the main body of edge and structure in views, which is easy to match and reduce the loss of the details.

The proposed cyclopean view C is defined by

$$C = \omega_l E_l + \omega_r E_r \tag{3}$$

where E_l and E_r are the matching blocks at the BFPs of stereo views, and ω_l and ω_r are weights for the matched left and right blocks, respectively. For a symmetrically distorted stereo views, both of ω_l and ω_r are set to 0.5.

2.2.2. Extracting RR feature and computing stereo video quality

Take into account the multi-channel characteristics of HVS, the RR frames of the original and distorted cyclopean view are decomposed with discrete wavelet transform (DWT), since sub-band coefficients of DWT can be easily collected statistically and features of RR frames are also easy to be extracted.

By estimating a few of parameters, the GGD can describe a mass of statistic data accurately. It is widely used to describe coefficients of DWT and discrete cosine transform (DCT). The GGD model is expressed as follows

$$P(x) = \frac{\beta}{2\alpha\tau(1/\beta)} \exp\left(-\left(\frac{|x|}{\alpha}\right)^{\beta}\right) \tag{4}$$

where α is the scale parameter and β is the shape parameter in the GGD model.

Through analyzing DWT high-frequency sub-band coefficients, it can be found that the statistical characteristics of sub-band coefficients conform to the GGD. Here, α and β of the cyclopean view about BFP and BRP are obtained as the RR quality indicators. Let $D_\alpha(\alpha^n_{org}, \alpha^n_{dis})$ denote the quality indicator of α about the n-th DWT sub-band, which can be computed by

$$D_\alpha(\alpha^n_{org}, \alpha^n_{dis}) = \left|\frac{\alpha^n_{org} - \alpha^n_{dis}}{\alpha^n_{org} + \alpha^n_{dis}}\right| \tag{5}$$

where α^n_{org} or α^n_{dis} denotes α of the n-th DWT sub-band in the original or distorted view. Similarly, let $D_\beta(\beta^n_{org}, \beta^n_{dis})$ denote the quality indicators of β about the n-th DWT sub-band, which can be calculated by

$$D_\beta(\beta^n_{org}, \beta^n_{dis}) = \left|\frac{\beta^n_{org} - \beta^n_{dis}}{\beta^n_{org} + \beta^n_{dis}}\right| \tag{6}$$

where β^n_{org} or β^n_{dis} denote β value of the n-th DWT sub-band in the original or distorted view, respectively.

Then, we compare the difference about α and β indicators of the original and distorted views. Let q^n_{cyc} denote the cyclopean view quality indicators of BFP, and it is computed by

$$q^n_{cyc} = \delta \cdot D_\alpha(\alpha^n_{org}, \alpha^n_{dis}) + \gamma \cdot D_\beta(\beta^n_{org}, \beta^n_{dis}) \tag{7}$$

where δ and γ are the weights of quality indicators, and $\delta + \gamma = 1$.

Different DWT sub-bands make different contribution to HVS. The contrast sensitive function (CSF) [21] is used for combining DWT sub-bands, so the cyclopean view quality Q_{cyc} can be gained by combining n DWT sub-band quality indicators $\{q^n_{cyc}\}$. In the case of 4 levels of DWT, as shown in Fig. 6, the indicators q^n_{cyc} with respect to the four DWT levels can efficiently reflect the change of different DMOS. In the figure, horizontal axis denotes DMOS value of distorted view and the vertical axis denotes the quality indicator q^n_{cyc}. It can be found that in each level, the larger the distortion (the larger the DMOS) is, the greater the quality indicator is.

In the BRP, the quality indicators of BRPs in left and right views are denoted as $q^n_{dis,l}$ and $q^n_{dis,r}$, respectively, which can be obtained in the same way as q^n_{cyc}. The BRP qualities of the left and right views can be calculated from the sub-band quality indicators by using CSF. Let $Q_{riv,l}$ and $Q_{riv,r}$ denote BRP qualities of the left and right views, respectively, and let Q_{riv} denote BRP quality of stereo views, which can be computed by

$$Q_{riv} = wQ_{riv,l} + (1 - w)Q_{riv,r} \tag{8}$$

where w is a weight. For symmetric stereo view pair, w is set to 0.5.

The cyclopean view quality Q_{cyc}, and BRP quality Q_{riv}, are pooled as the stereo RR frame quality Q_f by

$$Q_f = f(Q_{cyc}, Q_{riv}) \tag{9}$$

After taking into account that it is difficult to combine Q_{cyc} and Q_{riv} being exactly coincident with HVS, support vector regression technique [22] is used to establish the function $f()$ as a weighting function. The cyclopean view quality Q_{cyc} and BFP quality have the same meaning here.

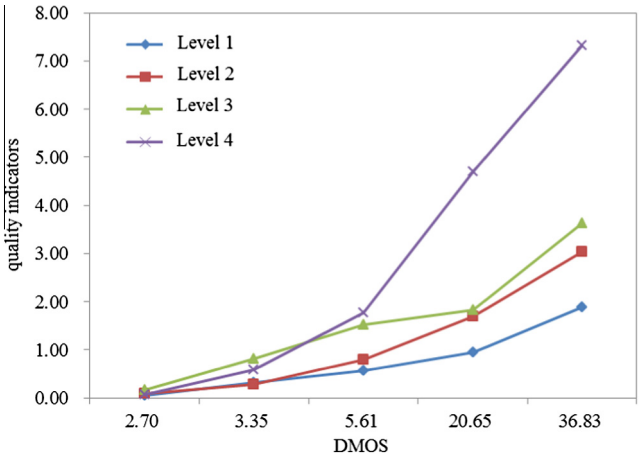

Fig. 6. The relationship between DMOS and sub-bands quality indicators.

Finally, we propose a temporal pooling strategy that fuses the stereo RR frame quality Q_f by using motion intensity to obtain the final stereo video quality score Q, which can be computed by

$$Q = \frac{1}{N}\sum_{i=1}^{N} Q_f(i) \cdot mv_i^\eta \tag{10}$$

where mv_i is the average motion vector of the i-th RR frame, N is the total number of RR frames of stereo video, and η is a regulatory factor.

3. Experimental results and discussions

To test the performances of the proposed RR-SVQA method, the NAMA3DS1-COSPAD1 stereo video database [17] is used, in which 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, downsampling and sharpening. The provided mean opinion scores (MOS) of the videos range from 1 (bad) to 5 (excellent). Among the distortion videos, the videos distorted by reduction of resolution, image sharpening, downsampling and sharpening are abandoned because few amount of them are able to obtain statistically significant results by training. Additionally, JPEG 2000 is not commonly used in video coding. Hence, we finally choose the 30 H.264 distortion videos as the test sequences.

Three criteria are used for performance comparison of objective quality assessment methods [23], including Pearson linear correlation coefficient (LCC), Spearman rank correlation coefficient (SROCC) and root mean squared error (RMSE). The values of LCC and SROCC range from 0 to 1. The better the performance of the method is, the SROCC and LCC values are more close to 1 and RMSE values more close to 0.

3.1. Decision of parameters

Here, the "db1" wavelet is used to decompose BFP and BRP into 4 levels. The values of δ and γ are obtained by training on the Ningbo University's stereo image database [24], which includes 12 original stereo images and 312 distorted stereo images with five types of distortions, including JPEG, JPEG2000 (JP2 K), Gaussian blur (Gblur), white noise (WN) and H.264 distortions. Through testing different values of δ and γ for getting the most effective correlation coefficient between the stereo image quality and the DMOS, α and β is determined as 0.1 and 0.9, respectively. This also shows that the standard derivation parameter has more contribution than mean parameter for this scheme. Since the test stereo video database is of symmetrical distortion, the weights ω_l, ω_r and w are all set to be 0.5. The motion vector regulatory factor η is set to 0.21.

3.2. Performance comparison and analysis

The motion intensity and RR frame extraction strategy of original video in the NAMA3DS1-COSPAD1 database are shown in Fig. 7. Every original video in the NAMA3DS1-COSPAD1 database is divided into three levels of motion intensity, strong, general and weak, respectively. In each subfigure, the upper part is the varied average motion vector of each frame in video. The middle part is the motion intensity level of the video. At the bottom part of subfigure, every peak represents a RR frame at the corresponding position of video. Although average motion vector of each stereo video computed by Eq. (1) is different, almost all videos have three motion intensities according to Eq. (1), as shown in Fig. 7. It is

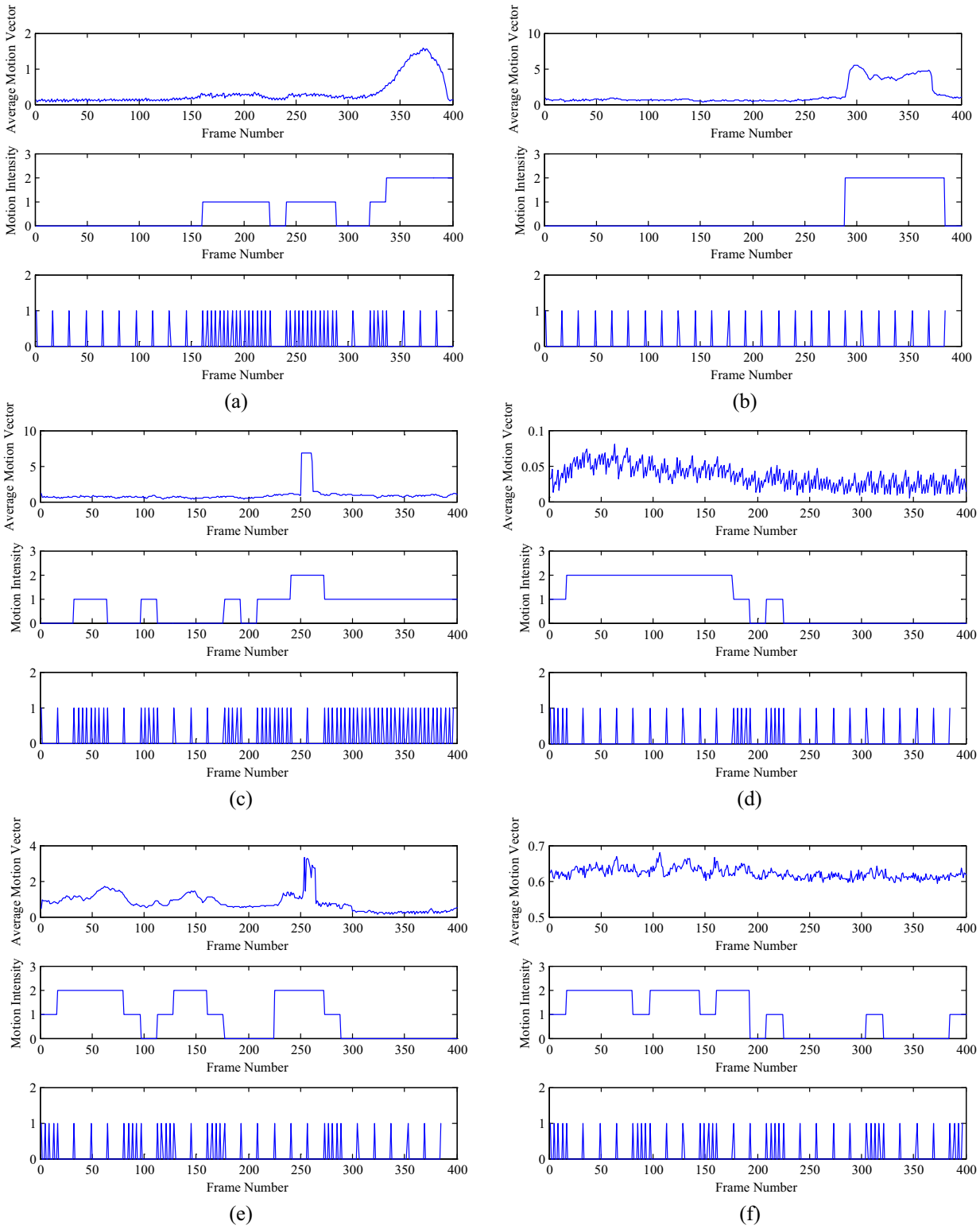

Fig. 7. Selecting rates, S_R, of ten original videos in the NAMA3DS1-COSPAD1 database. (a) 'Barrier gate'. (b) 'Basket'. (c) 'Boxers'. (d) 'Hall'. (e) 'Lab'. (f) 'News report'. (g) 'Phone call'. (h) 'Soccer'. (i) 'Tree branches'. (j) 'Umbella'.

because that the motion intensity level indicates a relative intensity. This strategy is more accord with human visual temporal characteristics of CSF. After RR frame extracting, the redundancy of video information is greatly reduced and the computational efficiency is significantly improved in the follow-up process.

In [25], existing state of the art 2D objective image and video quality metrics are tested on the 3D stereo video NAMA3DS1-COSPAD1 database to study the benchmark performance. The test results of these metrics are presented in Table 1. In the table, "VQM" [26], "RVQM" [27] and "PARMENIA" [28] are 2D video qual-

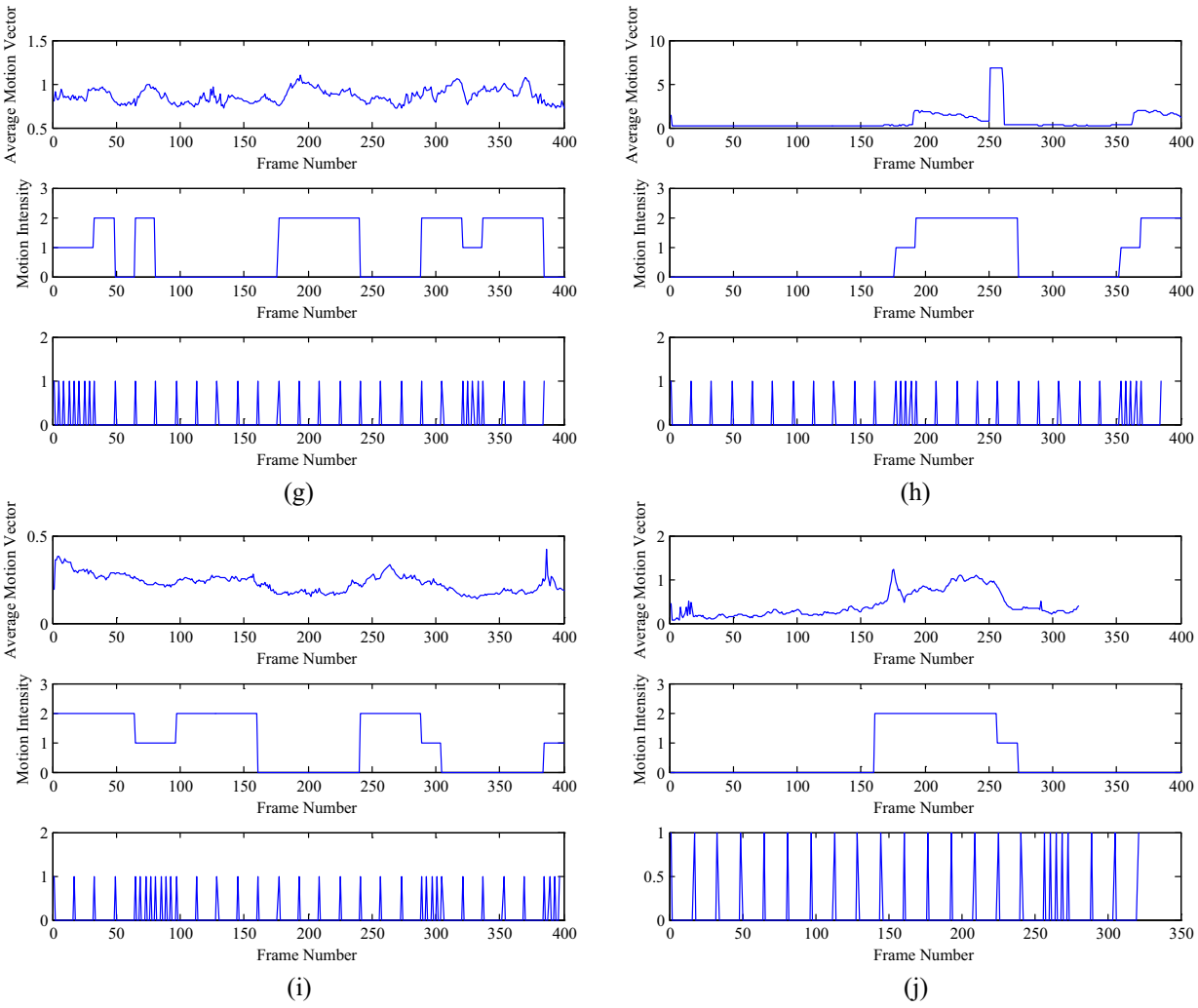

Fig. 7 (*continued*)

Table 1
2D objective image and video quality assessment performance of SROCC on H.264.

Metrics	SROCC	Metrics	SROCC
MS-SSIM	0.6112	IW-PSNR	0.8347
NQM	0.6676	VQM [26]	0.8804
PSNR	0.5064	VQM_VFD	0.8906
SSIM	0.7257	IQM2	0.8033
VIF	0.8726	RVQM [27]	0.6012
VSNR	0.7041	PARMENIA [28]	0.6464
IW-SSIM	0.8396	Proposed	0.9243

ity metrics, while the others are image quality metrics. More detailed information of these metrics can be found in [25]. "Proposed" in Table 1 denotes the proposed method in this paper.

As shown in Table 1, compared with all other metrics, the proposed method shows its better correlation with the MOS of test videos distorted with H.264 distortion. The SROCC index of the proposed method is up to 0.9243. It is clear that existing 2D image or video quality metrics are congenitally deficient to apply on stereo image or video. By contrast, the proposed method considers more binocular vision especially about binocular fusion and binocular rivalry. The proposed method extends the RR features from BFP and BRP of stereo video, which is shown to be necessary and effective.

Table 2
Performance of PSNR, SSIM on H.264.

	LCC	SROCC	RMSE
PSNR	0.5733	0.5200	0.9483
SSIM	0.7257	0.6973	0.8238
Proposed	0.9487	0.9243	0.3602

Table 3
Performances of T-PSNR and T-SSIM on H.264.

H.264	LCC	SROCC	RMSE
T-PSNR	0.7906	0.7876	0.7087
T-SSIM	0.8486	0.8047	0.6122
Proposed	0.9487	0.9243	0.3602

To validate the effect of the proposed RR-SVQA method, it is compared with the classical PSNR and SSIM on three criteria, that is, LCC, SROCC and RMSE. PSNR and SSIM are directly applied to evaluate stereo images. The weighted average is adopted between two viewpoints and temporal pooling is used to obtain the final stereo video quality. The test results are shown in Table 2. Because of the different test environments, such as the test parameters and test software platform, the SROCC index of PSNR and SSIM is slightly different from that in Table 1.

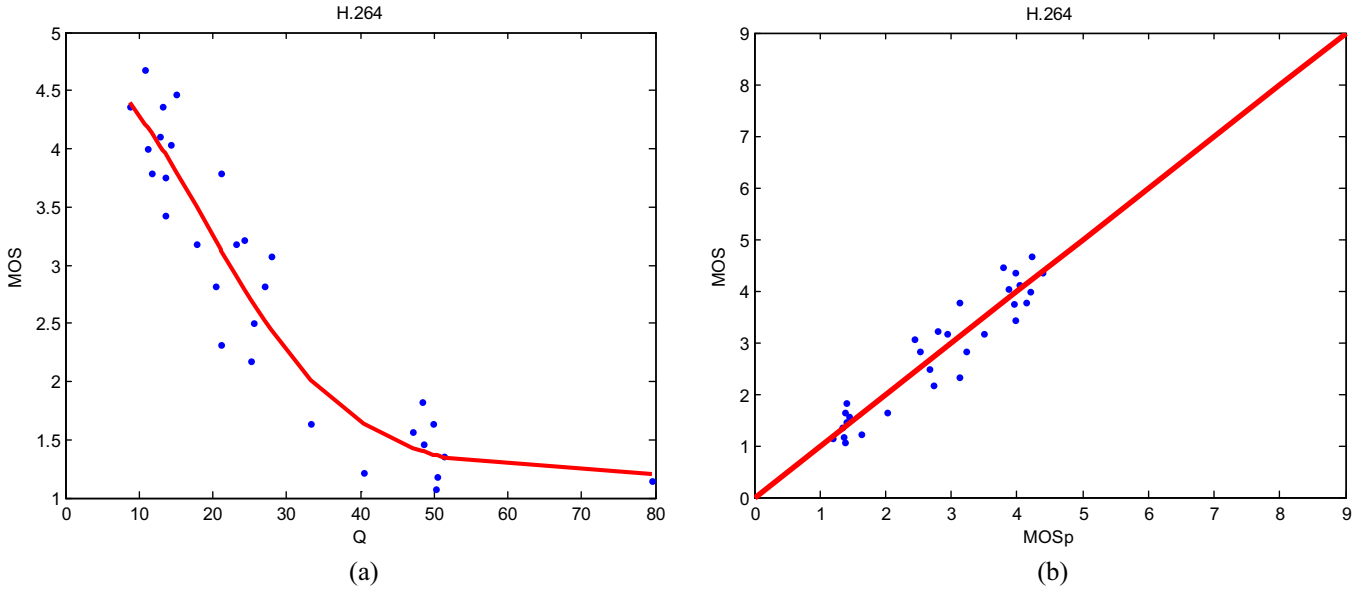

Fig. 8. Scatter plots of subjective scores versus the objective SVQA scores. (a) Q versus MOS. (b) MOS_p versus MOS.

Compared with the standard 2D quality assessment methods of PSNR and SSIM, the proposed method achieves better performance on the three criteria. The results imply that 2D quality assessment methods cannot reflect human perception accurately when viewing 3D scenes. The usual cyclopean view models [11] match the left and right views directly to get the cyclopean view, but they will cause the loss of detail texture information. The proposed method considers binocular characteristics of BFP as well as BRP. It makes a contribution to retain detail texture information in BRP and obtain the cyclopean view more easily and reasonably in BFP. In addition, according to the statistical characteristics of wavelet coefficients, the effective RR quality indicators are obtained to make the proposed method more accurate. Hence, the proposed method can represent human visual properties effectively.

In order to validate effect of the RR frame selection and temporal pooling strategy, the PSNR or SSIM of the RR frames are weighted by using the proposed temporal pooling strategy. The test results denoted as T-PSNR and T-SSIM are shown in Table 3.

Compared with the results in Table 2, the performances of T-PSNR and T-SSIM are significantly improved according to the three criteria, that is, LCC, SROCC and RMSE. The LCC of T-PSNR is improved from 0.5733 to 0.7906 and SROCC reaches 0.7876. T-SSIM also has a prominent improvement. LCC and SROCC of T-SSIM are 0.8486 and 0.8047, which are lift 0.1229 and 0.1074 compared with SSIM, respectively.

In addition, stereo video quality Q versus corresponding MOS value is shown in Fig. 8(a). The objective quality assessment scores of the proposed method have good consistency with the MOS scores. According to the fitting curve in Fig 8(b), the MOS_p is computed and the results show that the MOS_p accords well with the subjective scores MOS. It means that the proposed RR frame selection and temporal pooling strategy are effective, so that the corresponding temporal pooling image quality assessment methods can achieve better performance. The proposed RR frame extraction scheme can reduce a massive of redundancy and ensure the assessment performance of the proposed RR-SVQA method. For uniform distributed distortion on the temporal domain, the proposed method can extract RR frames effectively. The discrepancy between previous and succeeding frames is the essential difference between image and video and represented as motion vector. Motion vector is used to weight the RR frames in the proposed method, since utilizing temporal motion information rationally can effectively reflect temporal characteristics of video. The experimental results also show that the motion intensity of video has a great influence on human visual perception and should not be ignored.

4. Conclusion

In this paper, we propose a RR stereo video quality assessment method by modeling the binocular perception effect. Compared with the full-reference quality assessment method, the proposed method is more feasibility for stereo video quality assessment because not entire original video is required for the assessment. The RR frame extraction scheme can reduce redundancy and ensure the performance of the proposed RR SVQA method. In addition, considering binocular perceptual theory, a RR frame is divided into binocular fusion portion (BFP) and binocular rivalry portion (BRP). Even though the BFP and BRP are roughly divided, the experimental results show that the proposed method is highly consistent with the subjective perception. Furthermore, the experimental results also show that the RR features obtained from the two portions can measure the quality of stereo video and represent human visual properties effectively. For uniform distributed distortion on temporal domain, the RR frames are easy to be extracted, but for unstable temporal distortion, the RR temporal strategy may be less effective.

Acknowledgements

This work was supported by the Natural Science Foundation of China under Grant Nos. U1301257, 61271270, 61271021 and 61311140262, the National High-tech R&D Program of China under Grant No. 2015AA015901, and the Natural Science Foundation of Zhejiang Province in China under Grant Nos. LY15F010005 and LY16F010002. It is also sponsored by K.C. Wong Magna Fund in Ningbo University.

References

[1] Y. Chen, M.M. Hannuksela, T. Suzuki, S. Hattori, Overview of the MVC + D 3D video coding standard, J. Vis. Commun. Image Represent. 25 (4) (2014) 679–688.

[2] F. Shao, W. Lin, G. Jiang, M. Yu, Q. Dai, Depth map coding for view synthesis based on distortion analyses, IEEE J. Emerg. Select. Topics Circ. Syst. 4 (1) (2014) 106–117.
[3] S. Ryu, K. Sohn, No-reference quality assessment for stereoscopic images based on binocular quality perception, IEEE Trans. Circuits Syst. Video Technol. 24 (4) (2014) 591–602.
[4] W. Zhou, G. Jiang, M. Yu, F. Shao, Z. Peng, Reduced-reference stereoscopic image quality assessment based on view and disparity zero-watermarks, Signal Process. Image Commun. 29 (1) (2014) 167–176.
[5] N. Yun, Z. Feng, J. Yang, J. Lei, The objective quality assessment of stereo image, Neurocomputing 120 (10) (2013) 121–129.
[6] J.Y. Lin, C. Wu, I. Katsavounidis, Z. Li, A Aaron, C.C. J. Kuo, EVQA: an ensemble-learning-based video quality assessment index, in: 2015 IEEE International Conference on Multimedia & Expo Workshops, Torino, Italy, June 2015, pp. 1–6.
[7] F. Shao, W. Lin, S. Gu, G. Jiang, T. Srikanthan, Perceptual full-reference quality assessment of stereoscopic images by considering binocular visual characteristics, IEEE Trans. Image Process. 22 (5) (2013) 1940–1953.
[8] T. Zhu, L. Karam, A no-reference objective image quality metric based on perceptually weighted local noise, EURASIP J. Image Video Process. 2014 (1) (2014) 1–8.
[9] R. Soundararajan, A.C. Bovik, Video quality assessment by reduced reference spatio-temporal entropic differencing, IEEE Trans. Circuits Syst. Video Technol. 23 (4) (2013) 684–694.
[10] J. Seo, X. Liu, D. Kim, K. Sohn, An objective video quality metric for compressed stereoscopic video, Circ. Syst. Signal Process. 31 (3) (2012) 1089–1107.
[11] L. Jin, A. Boev, K. Egiazarian, A. Gotchev, Quantifying the importance of cyclopean view and binocular rivalry-related features for objective quality assessment of mobile 3D video, EURASIP J. Image Video Process. 2014 (2) (2014) 1–18.
[12] C.T.E.R. Hewage, M.G. Martini, Reduced-reference quality metric for 3D depth map transmission, in: 3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON), Tampere, Finland, June 2010, pp. 1–4.
[13] H. Malekmohamadi, A. Fernando, A. Kondoz, A new reduced reference metric for color plus depth 3D video, J. Vis. Commun. Image Represent. 25 (3) (2014) 534–541.
[14] L. Ma, K.N. Ngan, L. Xu. Reduced reference video quality assessment based on spatial HVS mutual masking and temporal motion estimation, in: IEEE International Conference on Multimedia and Expo (ICME'2013), California, USA, July 2013, pp. 1–6.
[15] K. Sharifi, A. Leon-Garcia, Estimation of shape parameter for generalized Gaussian distributions in subband decompositions of video, IEEE Trans. Circuits Syst. Video Technol. 5 (1) (1995) 52–56.
[16] A. Rossholm, M. Shahid, B. Lovstrom, Analysis of the impact of temporal, spatial, and quantization variations on perceptual video quality, in: Network Operations and Management Symposium (NOMS'2014), Krakow, Poland, May 2014, pp: 1–5.
[17] M. Urvoy, M. Barkowsky, R. Cousseau, Y. Koudota, NAMA3DS1-COSPAD1: Subjective video quality assessment database on coding conditions introducing freely available high quality 3D stereoscopic sequences, in: International Workshop on Quality of Multimedia Experience (QoMEX), Melbourne, Australia, July 2012, pp. 109–114.
[18] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.
[19] D.C. Knill, A. Pouget, The Bayesian brain: the role of uncertainty in neural coding and computation, Trends Neurosci. 27 (12) (2004) 712–719.
[20] J. Wu, W. Lin, G. Shi, A. Liu, Perceptual quality metric with internal generative mechanism, IEEE Trans. Image Process. 22 (1) (2013) 43–54.
[21] J. Mannos, D. Sakrison, The effects of a visual fidelity criterion on the encoding of images, IEEE Trans. Inf. Theory 20 (4) (1974) 525–536.
[22] C.C. Chang, C.J. Lin, LIBSVM: a library for support vector machines, ACM Trans. Intell. Syst. Technol. (TIST) 2 (3) (2011) 389–396.
[23] K. Brunnström, D. Hands, F. Speranza, A. Webster, VQEG validation and ITU standardization of objective perceptual video quality metrics, IEEE Trans. Sig. Process. Magaz. 26 (26) (2009) 96–101.
[24] J. Zhou, G. Jiang, X. Mao, M. Yu, Subjective quality analyses of stereoscopic images in 3DTV system, in: Visual Communication and Image Processing, Taiwan, November 2011, pp, 1–4.
[25] E. Dumic, S. Grgic, D.J. Bermejo, L.S. Cruz, Benchmark of state of the art objective measures for 3D stereoscopic video quality assessment on the Nantes database, in: International Symposium on ELMAR, Zadar, Croatia, September 2014, pp. 1–4.
[26] M.H. Pinson, S. Wolf, A new standardized method for objectively measuring video quality, IEEE Trans. Broadcast. 50 (3) (2004) 312–322.
[27] E. Dumic, S. Grgic, Reduced video quality measure based on 3D steerable wavelet transform and modified structural similarity index, in: International Symposium ELMAR, Zadar, Croatia, September 2013, pp. 65–69.
[28] D. Jiménez, High Definition Video Quality Assessment Metric Built Upon Full Reference Ratios, Ph.D. Thesis <http://oa.upm.es/14712/>.