Reduced-Reference Video Quality Assessment of Compressed Video Sequences

Lin Ma, *Student Member, IEEE,* Songnan Li, *Student Member, IEEE,* and King Ngi Ngan, *Fellow, IEEE*

***Abstract*—In this paper, a novel reduced-reference (RR) video quality assessment (VQA) is proposed by exploiting the spatial information loss and the temporal statistical characteristics of the interframe histogram. From the spatial perspective, an energy variation descriptor (EVD) is proposed to measure the energy change of each individual encoded frame, which results from the quantization process. Besides depicting the energy change, EVD can further simulate the texture masking property of the human visual system (HVS). From the temporal perspective, the generalized Gaussian density (GGD) function is employed to capture the natural statistics of the interframe histogram distribution. The city-block distance (CBD) is used to calculate the histogram distance between the original video sequence and the encoded one. For simplicity, the difference image between adjacent frames is employed to characterize the temporal interframe relationship. By combining the spatial EVD together with the temporal CBD, an efficient RR VQA is developed. Evaluation on the subjective quality video database demonstrates that the proposed method outperforms the representative RR video quality metric and the full-reference VQAs, such as peak signal-to-noise ratio and structure similarity index in matching subjective ratings. This means that the proposed metric is more consistent with the HVS perception. Furthermore, as only a small number of RR features are extracted for representing the original video sequence (each frame requires only one parameter for describing EVD and three parameters for recording GGD), the RR features can be embedded into the video sequences or transmitted through the ancillary data channel, which can be used in the video quality monitoring system.**

***Index Terms*—Energy variation descriptor (EVD), generalized Gaussian density (GGD), human visual system (HVS), reduced-reference (RR), video quality assessment (VQA).**

I. Introduction

IMAGE OR VIDEO visual quality measurement is becoming more and more important, especially in the transmission of multimedia content over the Internet. It is very useful for various image or video processing applications, such as communication, compression, displaying, registration, restoration, enhancement, quality monitoring, and so on [1]. The human eyes are the ultimate receivers of the processed images or videos. Accordingly, the most reliable way for evaluating the image or video visual quality is the subjective testing method. However, the subjective testing method [2], [3] requires many observers to participate in the experiments and provide their personal opinions of the image or video quality. It is very time-consuming and expensive, and so cannot be employed for practical image or video applications. Consequently, the image or video quality metrics [4] that can automatically assess the visual quality are desired.

Based on the availability of the reference image or video, the quality metrics [4] can be categorized into three classes: full-reference (FR) [5]–[12], no-reference (NR) [13]–[19], and reduced-reference (RR) [20]–[45]. The FR quality metrics require the whole information of the reference image or video to evaluate the visual quality of the distorted one. These metrics can be utilized in image or video compression, watermarking, and so on. The mean square error (MSE) and the corresponding peak signal-to-noise ratio (PSNR) are the most widely used FR metrics, because of their simple formulations, easy optimizations, and clear physical meanings. However, MSE or PSNR does not consider any perceptual properties of the human visual system (HVS). The effectiveness of MSE or PSNR is doubted for evaluating the perceptual quality [5]–[46], as their predictions correlate poorly with the subjective ratings of the viewers. In order to handle the drawbacks, the structure similarity (SSIM) index [5], [6] is proposed to depict the structural distortions rather than the pixel absolute differences. Recently, Chandler *et al.* [7] derived an image quality assessment (IQA) to depict the visual signal-to-noise ratio (VSNR) in the wavelet domain. Ma *et al.* [9] proposed to incorporate the orientation sensitivity and conspicuity of the HVS into SSIM to derive a more accurate IQA. And the quality metric [11] based on the just-noticeable difference (JND) in discrete cosine transform (DCT) domain was developed and employed for perceptual video coding. Furthermore, Zhang *et al*. [10] considered the contrast sensitivity function and texture masking effect of the HVS to develop a simple image quality metric, which demonstrates its effectiveness for perceptual image compression. In [12], the authors proposed a motion-compensation-based approach to assess the temporal quality of the video sequences.

However, in many real-world applications, the original image or video is not available, e.g., image or video denoising, restoration, and super-resolution. In order to evaluate the qualities of these processed images or videos, the NR metrics

Manuscript received July 30, 2011; revised December 20, 2011; accepted February 17, 2012. Date of publication June 1, 2012; date of current version September 28, 2012. This work was supported in part by a grant from the Research Grants Council of Hong Kong, under Project 415712. This paper was recommended by Associate Editor F. Lavagetto.

The authors are with the Department of Electronic Engineering, Chinese University of Hong Kong, Shatin, Hong Kong (e-mail: lma@ee.cuhk.edu.hk; snli@ee.cuhk.edu.hk; knngan@ee.cuhk.edu.hk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2012.2202049

[13] are thus developed, which can be utilized for controlling the processing stages. However, it is an extremely difficult task. Many metrics take the behaviors of specific distortions into consideration. In [14], the wavelet statistical model was employed to capture the distortion introduced by JPEG 2000. Liang *et al*. [15] combined the sharpness, blurring, and ringing measurements together to evaluate the visual quality of the JPEG 2000 coded image. Brandao *et al*. [16] proposed an NR IQA based on the DCT domain statistics to evaluate the quality of JPEG coded image. Afterwards, the NR IQA in [16] is extended to NR video quality assessment (VQA) [17] for evaluating the visual quality of the H.264/AVC coded videos. Furthermore, Ferzli *et al*. [18] did the psychophysical experiment to test the blur tolerance ability of the HVS, which is denoted as just-noticeable blur. In [19], the authors proposed an NR VQA for assessing the perceptual quality of the frame freezing impairments. All of these NR metrics try to depict the visual qualities of video sequences contaminated by a certain specific distortion. When some new distortions emerge, the performances of these metrics may degrade. Therefore, in order to provide a compromise between FR and NR metrics, RR methods have been developed for quality assessment by employing partial information of the corresponding reference image or video. With a limited number of features extracted from the reference image or video, the RR metric can efficiently evaluate the visual quality of the distorted one. As the extracted features require only a small number of bits for representation, they can be coded and transmitted together with the images or videos. Consequently, it is practical for quality monitoring during the image or video transmission and communication, which can finally lead to a better quality of user experience.

This paper deals with the RR quality assessment by a tradeoff between the quality prediction accuracy and the amount of extracted RR features. The proposed method tries to extract several features from both the spatial and temporal perspectives, which are sensitive to the perceptual quality. With evaluation on the subjective quality video database, the proposed RR VQA can outperform the representative RR metrics, such as video quality metric (VQM) [32], J.246 [34], and the RR video metric in [35], and even the FR quality metrics PSNR and SSIM [5], [6]. Furthermore, the RR feature extraction and comparison are of low complexity, which can be easily computed from the distorted video sequences. Therefore, the proposed RR VQA can be employed in the video quality monitoring systems.

The remainder of this paper is organized as follows. After a brief description of the existing RR image or video quality metrics in Section II, the detailed algorithm is introduced in Section III. Section IV demonstrates the performance comparisons. Finally, the conclusion is given in Section V.

II. Related Works

The RR quality metrics aim to monitor the video perceptual quality during the transmission and communication processes. Therefore, many approaches [21]–[26] try to model the distortions of the encoded video sequences, such as the MPEG-2 compressed videos, in the quality monitoring system. For example, Wolf *et al*. [21], [22] extracted a set of spatial and temporal features that are very sensitive to the distortions introduced in the standard video compression framework. Le Callet *et al*. [23] depicted the blur, blocking, and temporal artifacts of the MPEG-2 coded sequences by some representative features. By accounting for differences between these features, the degradation level of the coded videos can be estimated. Yang [24] employed the ratio information of DCT coefficients to measure the perceptual qualities of MPEG-2 coded sequences. In [25], the artifacts of the AVC/H.264 coded video sequence, such as blur and blocking, are depicted and measured by the objective features. They are combined together into a single measurement for the overall video quality. Furthermore, Tagliasacchi *et al*. [26] approximated the SSIM value of the video corrupted by channel errors through employing coding tools provided by the distributed source coding theory.

Furthermore, in order to provide a more accurate performance, the HVS properties [27]–[36] have been considered during the feature extraction. Le Callet *et al*. [27] employed a neural network to train and evaluate the perceptual qualities of video sequences, based on the perception related features of the video frames. In [28] and [29], the authors extracted perceptual features motivated from the computational models of the low level vision. These features are utilized as the reduced descriptors to represent the visual quality. Tao *et al*. [30] incorporated the merits of the contourlet transform, the contrast sensitivity function, and Weber's law of JND to derive an RR IQA. Engelke *et al*. [31] designed an RR IQA for wireless imaging by accounting for different structural information that is observed in the distortion model of wireless link. Then, the structural information from the viewing area is trained for the HVS. In [32], the authors extracted several HVS related features to indicate the spatial information losses, edge information changes, contrast information, and color impairments. By combining these different components with different weights, the final video perceptual quality index is obtained. These HVS related features are compressed for video quality monitoring [33]. It is demonstrated that a compression ratio of more than $30\times$ can be achieved with only a small error introduced in the final quality values. Moreover, as the HVS is sensitive to the degradation around the edges, the RR video quality metric proposed in [34] mainly measures the edge degradations. The edge degradation is computed by measuring the MSE of the edge pixels. Therefore, this method is named as edge PSNR. In [35], the authors employed discriminative local harmonic strength with motion consideration to evaluate the distorted video quality. The gradient information of each frame is employed for harmonic and discriminative analysis. Furthermore, the authors [36] derived the RR quality metric for 3-D videos. The edge information of depth maps and information from the corresponding color image in the areas in the proximity of edges are extracted for the RR quality metric, which can be utilized for 3-D video compression and transmission.

Recently, the natural image or video statistics modeling [20], [37]–[45] has also been considered for developing an efficient RR quality metric. Li *et al*. [39] employed the divisive normalization to depict the coefficient distributions

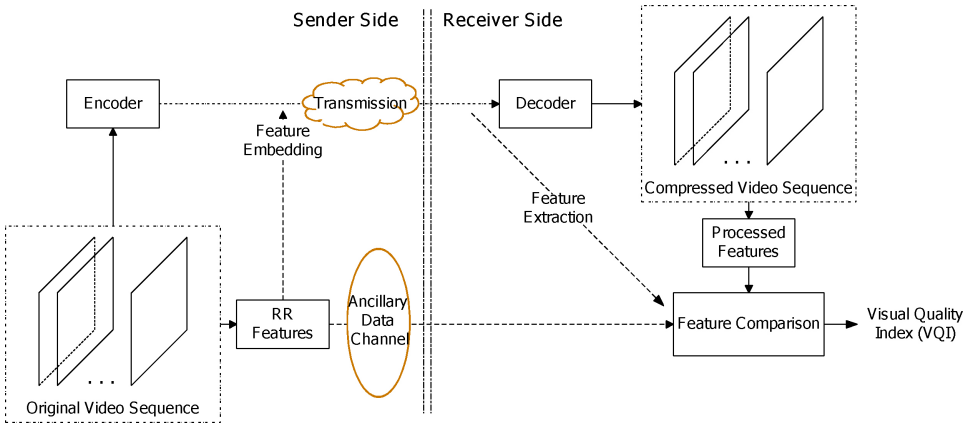

Fig. 1. General framework of the RR VQA system.

in the wavelet domain, where the divisive normalization transformation can accurately depict the coefficient distribution. And a training process is introduced to learn several parameters for the RR IQA. Wang *et al*. [37], [38] proposed a wavelet-domain natural image statistic metric, which tries to model the marginal probability distribution of the wavelet coefficients of a natural image by the generalized Gaussian density (GGD) function. Then, the Kullback–Leibler distance (KLD) is used to depict the distribution difference, which represents the perceptual quality of the distorted image. However, the KLD is asymmetric [40]. As demonstrated in [41], it is not suitable for IQA, because the visual quality distance from one image to another should be identical no matter how it is measured. In order to handle the problem, Ma *et al*. [41] proposed an RR IQA by using the reorganized DCT-based image representation, where the coefficient distribution in the DCT domain has been modeled. In [42], the differences between entropies of the wavelet coefficients of the reference and distorted images are employed to measure the image information change. And Redi *et al*. [43] proposed using the color distribution for evaluating the perceived image quality. Furthermore, Zeng *et al*. [44], [45] extended the RR IQA to VQA, by modeling the video's natural temporal statistics. In [44], the temporal motion smoothness of a video sequence is proposed to examine the temporal variations of local phase structures in the complex wavelet transform domain. In [45], both intraframe and interframe RR features are calculated based on the statistical modeling of natural videos. Together with a robust video watermarking approach, a quality-aware video system is developed. It has been demonstrated that these two RR VQAs present good measurements of the individual distortion level. However, these metrics are not evaluated over the subjective quality video database, which leads to a deficiency of the evaluation results.

In this paper, with inspiration of all the aspects, an efficient RR VQA for compressed video sequences is proposed. First, from the spatial perspective, an energy variation descriptor (EVD) is proposed to measure the energy change of each distorted frame. The proposed EVD can also be utilized to simulate the texture masking property of the HVS. For the temporal distortion, GGD is employed to model the histogram distribution of the interframe difference. The city-block distance (CBD) is used to calculate the histogram difference between the original video and the distorted one. Finally, the perceptual quality index is derived by combining the spatial EVD together with temporal CBD.

III. Proposed RR Video Quality Metric

The general framework of the RR VQA system is illustrated in Fig. 1. On the sender side, the RR features that are sensitive to the HVS perception are first extracted from the original video sequence. Then, the original video is encoded and transmitted to the receiver side. The corresponding RR features can be embedded into the coded bit-streams or transmitted through an ancillary data channel to the receiver side. After decoding, the processed features can be calculated from the compressed video sequence. By comparing the processed features with the ones of the original video sequence, the visual quality index of the compressed video can be generated.

In order to develop an efficient RR VQA, several challenges need to be considered. On the sender side, the extracted features need to be sensitive to a variety of video coding distortions, not only from the spatial perspective but also from the temporal perspective. Also, these features have to be relevant to the HVS perception of the video quality. The second important issue is the computational complexity of the RR feature calculation. If the complexity is too high, the receiver cannot easily compute the processed features from the compressed video. Consequently, it cannot practically monitor the visual quality of the distorted video. Therefore, the feature computation process should be efficient. Another important factor to consider is that the RR feature selection should consider not only the prediction accuracy of the quality metric, but also the data rate of the RR features. For a higher data rate, one may include more information of the reference video. Thus a good performance can be obtained, but this on the other hand will introduce a heavy burden to the RR feature transmission. Actually, the FR VQA is one extreme case of RR VQA, with the data rate being the whole reference video. With a smaller data rate, little information of the reference image or video is available, resulting in a poor quality prediction accuracy. As such, we can regard the NR VQA as another extreme case of RR VQA, with no information from the

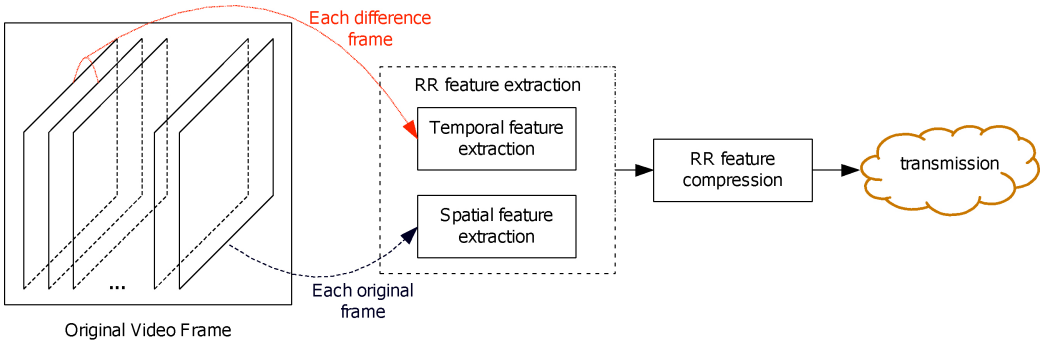

Fig. 2. RR feature extraction on the sender side.

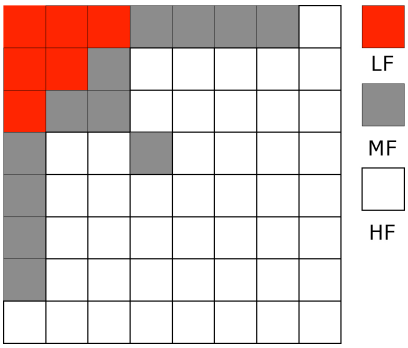

Fig. 3. DCT subband categorization based on different frequencies.

reference video. How to balance the data rate and performance is the key point for RR feature selection.

The framework of extracting the RR features on the sender side is illustrated in Fig. 2. For each original video frame, the RR feature representing the distortions from the spatial perspective is calculated. As the difference frame can depict the temporal relationship between adjacent frames, the temporal features are extracted from each difference frame. After the feature extraction, the compression process is performed to represent the RR features in limited bits, which can be easily transmitted to the receiver side for visual quality analysis. The following sections will introduce the detailed information of feature extraction from both spatial and temporal perspectives.

A. RR Feature Extraction From the Spatial Perspective

The distortion of the video sequence encoded by MPEG-2 and H.264 is introduced during the quantization process, which quantizes the DCT coefficients of the spatial blocks into different levels. It can help to efficiently reduce bit-rates for representing the video sequence. However, the quantization process results in the useful information loss. Intuitively, the larger the quantization step, the more the information loss is, and the worse is the perceptual quality of the encoded video. Therefore, the information loss has a certain implicit relationship with the video perceptual quality. In this paper, we propose an EVD to represent the spatial information loss.

For each block-based DCT (take an 8×8 DCT as an example), the DCT subbands can be categorized into different frequency bands, namely, high-frequency (HF), medium-frequency (MF), and low-frequency (LF). In JND estimation [47], [48], the authors employed the energies of different subbands to indicate different block types. Based on these different types, the visual texture masking property is described. The frequency categorization of DCT subbands is illustrated in Fig. 3. Let L, M, and H represent the sums of the absolute DCT coefficient values in the LF, MF, and HF groups, respectively. It should be noted that the quantization matrix is not uniformly distributed. The higher the DCT frequency, the larger the quantization parameter is. The reason is that the HVS is more sensitive to the LF components, which should be preserved during the quantization process. Therefore, it is not reasonable to record the absolute values of L, M, and H, which cannot effectively depict information loss. In this paper, the corresponding frequency ratio EVD is proposed to depict the HVS-related information loss, which is defined as

$$EVD = \frac{(M + H)}{L}. \tag{1}$$

The above definition is for each 8×8 DCT block. You can sum all the L, M, and H values over all the blocks to get the EVD value for a whole image or frame. From the definition, we can see that the EVD depicts the frequency energy proportion of the original video frame. When the distortion is introduced, specifically in the quantization process, the energies of MF and HF components will change more significantly than the LF ones. Thus, the EVD can accurately depict the changes and effectively capture the information losses. Furthermore, the larger the value of EVD, the more energy the MF and HF components possess. It means that the DCT block is more likely to contain texture information. For the plain block, the energy mostly concentrates in the LF components. For the edge block, there will be only a small number of DCT coefficients in the HF group. Consequently, the texture block will present higher EVD. As discussed in the JND models [47], [48], the texture block can tolerate more distortions than the plain and edge block, which is interpreted as the texture masking property of the HVS. Therefore, the proposed EVD can be employed to simulate the texture masking property for the derivation of the final video quality metric.

B. RR Feature Extraction From the Temporal Perspective

The temporal RR feature extraction strategy is illustrated in Fig. 4. First, the temporal relationship between adjacent frames

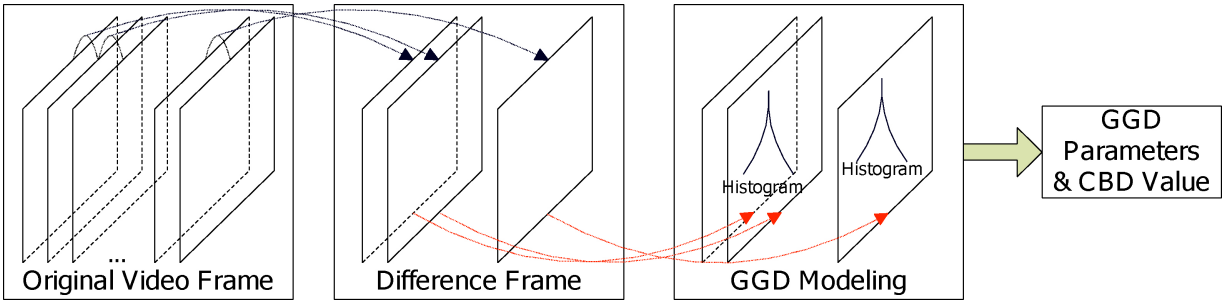

Fig. 4. Proposed method for temporal RR feature extraction.

needs to be depicted. The block-based motion estimation [49], [50] and optical flow [51] are employed to explore the motion information between the corresponding blocks or pixels of adjacent frames. However, although they can provide much more accurate information for describing the motion, the computational complexity is too high for practical implementations, especially on the receiver side. Therefore, in this paper, we simply employ the difference image for characterizing the temporal relationship between adjacent frames

$$D(i) = I(i) - I(i-1), \quad i \in \{2, 3, \ldots, N\} \tag{2}$$

where $I(i)$ is the ith original video frame, $D(i)$ is the corresponding difference frame, and N is the total frame number of the video sequence. This simple scheme has been proved to be effective for detecting the visual saliency map of the natural video sequences [52], [53]. Since luminance is more important than chrominance for our visual system, only the luminance information is considered to compute the difference frame.

In order to illustrate the statistical property of the difference image, several original video sequences, such as PA, PR, RB, and TR, are selected from the LIVE video quality database [54] for demonstration, as illustrated in Fig. 5. In order to provide a better visualization, the difference image has been reconstructed by $128 + (Pixel\_\text{value})$. It can be observed that the pixel values of the difference image mostly concentrate around zero, which generates a highly kurtotic distribution (with a sharp peak at zero and a fat-tail distribution). As demonstrated in [37] and [38], the histogram distribution of the wavelet coefficient is highly kurtotic. And this highly kurtotic distribution can be well fitted by GGD function. Furthermore, the coefficient distribution of the reorganized DCT subband, presenting highly kurtotic, can also be modeled by GGD [41]. Therefore, in this paper, GGD is employed to model the histogram distribution of the difference image. The probability density function (PDF) of GGD is defined as follows:

$$p_{\alpha,\beta}(x) = \frac{\beta}{2\alpha\Gamma(\frac{1}{\beta})} \exp\left\{-\left(\frac{|x|}{\alpha}\right)^{\beta}\right\} \tag{3}$$

where $\beta > 0$ and α are two parameters of the GGD function. Γ is the Gamma function given by

$$\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t}\, dt. \tag{4}$$

Here, α models the width of the PDF peak (standard deviation), while β is inversely proportional to the decreasing rate of the peak. α and β are also referred to as the scale and shape parameters, respectively. The GGD model can accurately model the histogram distribution, as demonstrated in Fig. 5, where the actual histogram distribution and the fitted GGD curve overlap with each other. Furthermore, it can be observed that the GGD model can work effectively with different types of video sequences. For example, the PA video sequence is captured by a static camera, which results in a great proportion of the pixel values around zero, whereas the PR video sequence is captured by a moving camera; hence, the pixel value distribution is much flatter. On the other hand, the RB video sequence is rich of dynamic texture information, and the TR video sequence is captured with a camera zooming effect.

By considering the maximum-likelihood estimation and assuming $\beta > 0$, we can obtain the approximated $\hat{\alpha}$ [55] according to

$$\hat{\alpha} = \left(\frac{\beta}{L}\sum_{i=1}^{L}|x_i|^{\beta}\right)^{\frac{1}{\beta}} \tag{5}$$

where x_i is the pixel sample from the corresponding difference image and L denotes the total number of the pixels. From (5), it can be observed that the estimated $\hat{\alpha}$ is related to the energy of the difference image in the β-norm. The difference energy can somewhat reflect the temporal changes between adjacent frames. That is the reason why we introduce GGD to model the histogram distribution of the difference image, because of not only the modeling accuracy but also its ability to indicate the energy of frame difference. As demonstrated in [23] and [27], the energy of the frame difference is useful to measure the temporal content for VQA. Furthermore, in order to improve the modeling accuracy, another parameter besides (α, β) is introduced, which is named CBD [41]

$$d_{\text{CBD}}(p, p_{\alpha,\beta}) = \sum_{i=1}^{h_L} |p(i) - p_{\alpha,\beta}(i)| \tag{6}$$

where $p(i)$ is the actual histogram of the difference image, $p_{\alpha,\beta}(i)$ is the fitted GGD curve, and h_L is the total number of the histogram bins. Compared with KLD, CBD is symmetrical, which makes it more reasonable for evaluating the histogram distance [41].

For each video frame, one parameter EVD is recorded to depict the spatial information loss, and three GGD parameters $\{\alpha, \beta, d_{\text{CBD}}(p, p_{\alpha,\beta})\}$ are extracted from each difference image

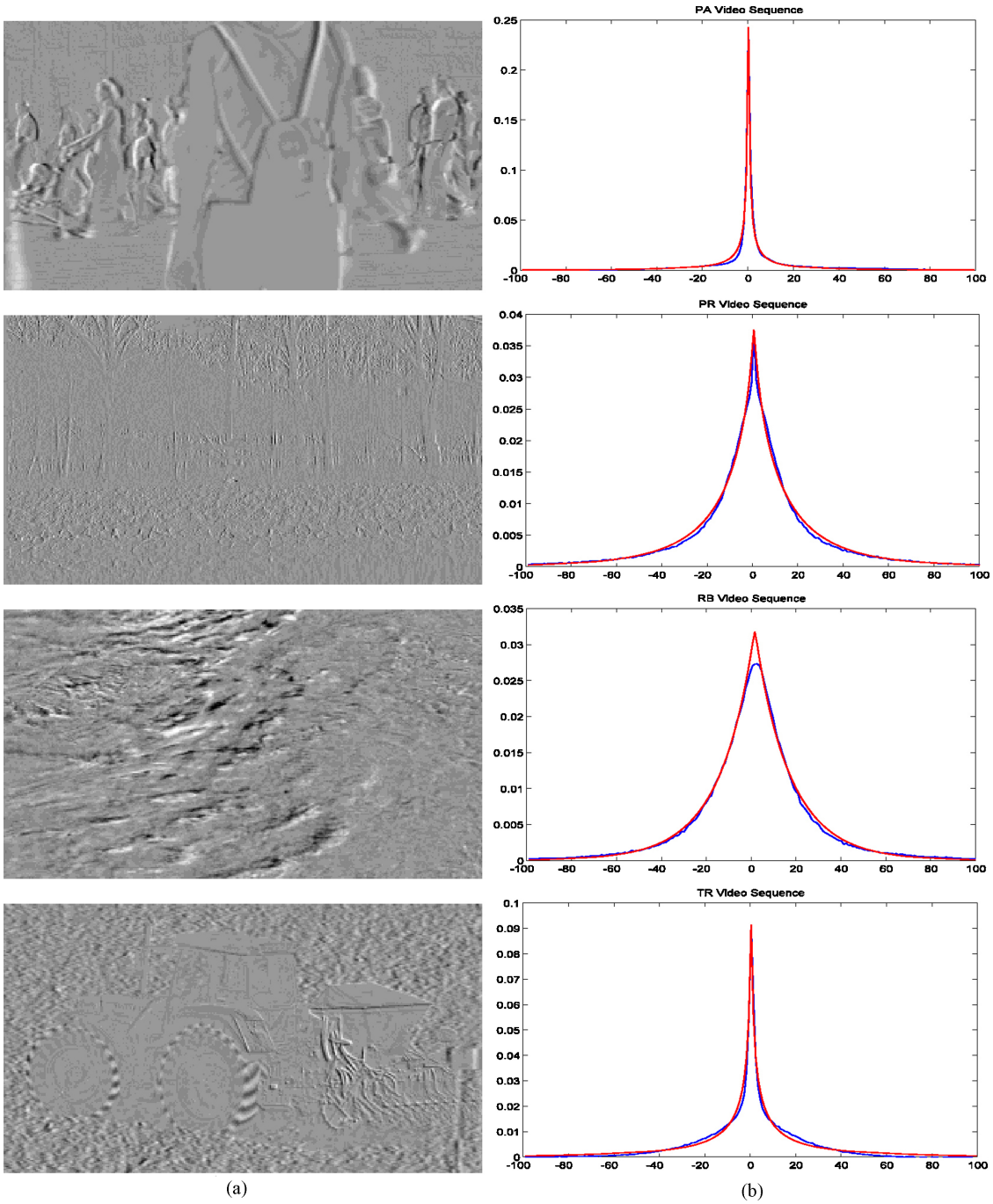

Fig. 5. (a) Eleventh difference image of the original video sequence. (b) Its corresponding histogram (blue line), and the fitted GGD curve (red line). From top to bottom: the PA, PR, RB, and TR video sequences from the LIVE video quality database [54].

for describing the temporal information. Therefore, there will be four parameters per frame in total to be recorded and transmitted to the receiver side for the quality assessment. For the EVD parameter, it is quantized into 8-bit precision for transmission. For the three GGD parameters, the same as in [37], β and $d_{\mathrm{CBD}}(p, p_{\alpha,\beta})$ are quantized into 8-bit precision, and α is represented using the 11-bit floating point, with 8 bits for mantissa and 3 bits for exponent. The quantization steps are set uniformly to represent the corresponding parameters in a limited number of bits. Therefore, for each frame, only $8 + 8 + 8 + 8 + 3 = 35$ bits are required to represent the RR features. As the data rate is very small, the features can be easily transmitted through an ancillary data channel. Furthermore, they can also be embedded into the same video signal with a robust watermarking scheme [45].

C. Visual Quality Analysis on the Receiver Side

On the receiver side, as shown in Fig. 1, we need to evaluate the visual quality of the compressed video sequence based on the RR features of the original video. The framework of the visual quality analysis on the receiver side is illustrated in

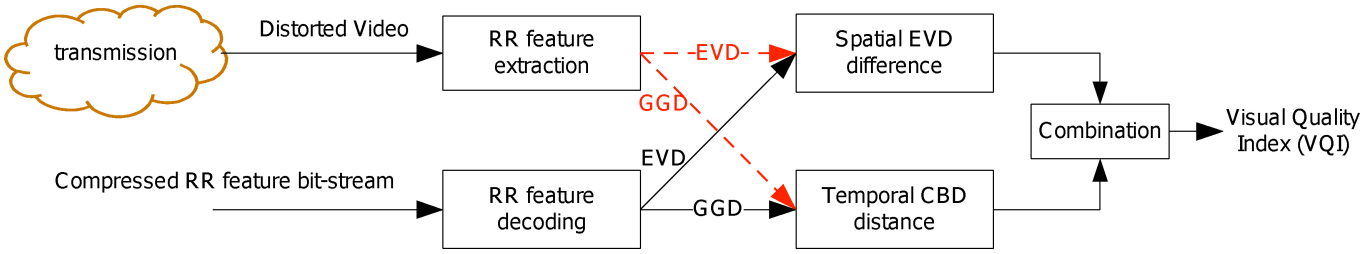

Fig. 6. Framework of visual quality analysis on the receiver side.

Fig. 6. First, the feature calculation procedure is performed on the distorted sequence to obtain the processed features, which consist of the spatial EVD and temporal GGD. The original RR features are decoded from the transmitted bit-streams. By comparing the original features with the processed ones, the spatial EVD difference and temporal CBD distance are obtained. By combining the two distances together, the visual quality score of each frame is generated. The final video quality index (VQI) of the corresponding video is obtained by temporally pooling the frame-level scores together.

For the spatial EVD, as the compression process will discard more HF and MF components than the LF ones, the degradation of EVD can effectively represent the information loss caused by the compression

$$EL = |EVD_{\text{ori}} - EVD_{\text{pro}}| \tag{7}$$

where EVD_{ori} is the original feature and EVD_{pro} is calculated from the compressed video sequence. For the coded video sequences, the compression artifacts are superposed onto the original video sequence, which is regarded as the masker signal. Therefore, the original sequence is utilized to mask the compression artifacts, which are introduced by the quantization process. As discussed before, a larger EVD value indicates more texture information. Consequently, more distortion can be masked by a larger EVD. Therefore, the extracted EVD can be utilized to simulate the HVS texture masking property. The information loss in (7) is weighted by the original feature EVD_{ori}

$$EL_V = \frac{EL}{EVD_{\text{ori}}} = \frac{|EVD_{\text{ori}} - EVD_{\text{pro}}|}{EVD_{\text{ori}}} \tag{8}$$

where EL_V is the final HVS-related features for depicting the spatial information loss.

For the temporal difference image, the CBD is employed to measure the difference between the reference video and the distorted one

$$d_{\text{CBD}}(p, p_d) = \sum_{i=1}^{h_L} |p(i) - p_d(i)| \tag{9}$$

where p depicts the difference image histogram of the original video and p_d is the distorted one. However, as the original video is unavailable on the receiver side, the fitted GGD curve is employed to approximate the distance

$$d_{\text{CBD}}(p, p_d) \triangleq |d_{\text{CBD}}(p_{\alpha,\beta}, p_d) - d_{\text{CBD}}(p, p_{\alpha,\beta})| \tag{10}$$

where $d_{\text{CBD}}(p, p_{\alpha\beta})$ is the third parameter introduced on the sender side. On the receiver side, only $d_{\text{CBD}}(p_{\alpha,\beta}, p_d)$ needs to be calculated. Their difference will be recorded to represent the statistical feature distance from the temporal perspective. As in [37] and [38], the logarithm process is employed to scale the temporal CBD distance as $\log_{10}(1+d_{\text{CBD}}(p, p_d)/c)$, where c is utilized to scale the CBD distance to avoid the variation being too small, and it is set as 0.001 for simplicity.

After obtaining the spatial EL_v value and temporal $\log_{10}(1+d_{\text{CBD}}(p, p_d)/c)$ value, how to combine them together remains a problem. In [45], the authors employed the averaging process to combine the spatial and temporal values together. However, it is not suitable for our obtained spatial and temporal values, because their magnitudes are quite different. In order to make the spatial EL_v value and temporal $\log_{10}(1 + d_{\text{CBD}}(p, p_d)/c)$ value contribute equally to the final quality score Q_s for each frame, the simple multiplication process is employed

$$Q_s = EL_V \times \log_{10}\left(1 + \frac{d_{\text{CBD}}(p, p_d)}{c}\right). \tag{11}$$

Based on the frame-level quality score Q_s, the VQI for depicting the perceptual quality of the entire compressed video is obtained by temporally pooling the Q_s scores together. In our implementation, the averaging process is employed to generate the final VQI

$$\text{VQI} = \sum_{i=1}^{N} Q_s(i)/N \tag{12}$$

where N is the total number of the video frames. According to the definition of VQI, the smaller the VQI, the better visual quality the compressed video sequence is. And the VQI of the original sequence is 0 according to its definition.

IV. Performance Evaluation

In this section, different VQAs are compared to demonstrate the effectiveness of the proposed RR VQA for evaluating the video perceptual quality. First, similar to [44] and [45], the consistency between the quality index generated by our proposed method and the distortion level is evaluated. Subsequently, the effectiveness of the proposed RR VQA is evaluated based on the LIVE video quality database [54], compared with the other VQAs. Finally, each component of the proposed algorithm is evaluated separately to demonstrate their corresponding contributions.

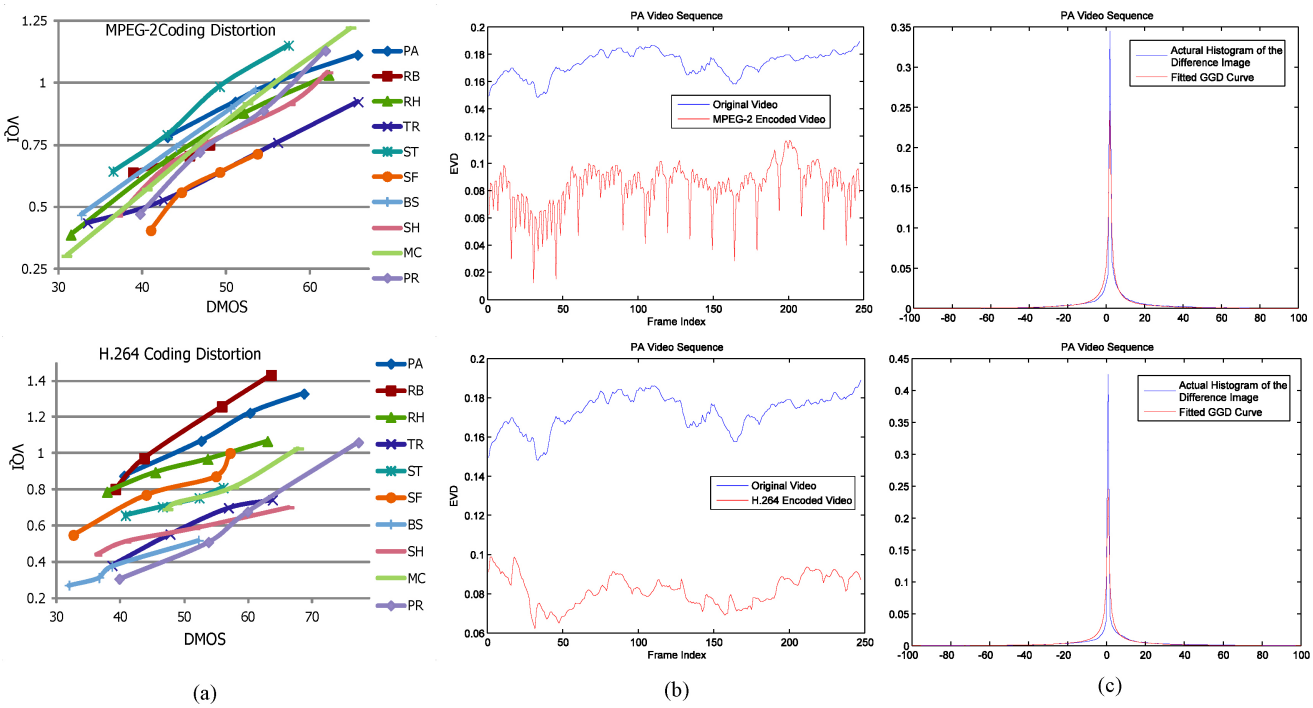

Fig. 7. Consistency evaluation of the proposed RR VQA over MPEG-2 and H.264 coded video sequences. (a) Proposed distortion measure VQI versus the DMOS value of each distorted video sequence. (b) Spatial EVD value of the PA video sequence (with the largest VQI value). (c) Temporal histogram of the eleventh difference image of the distorted video PA (with the largest VQI value) and the fitted GGD curve.

A. Consistency Test of the Proposed VQA Over Compressed Video Sequences

We first test the consistency of our proposed RR VQA on the coding artifacts, specifically, the MPEG-2 compression and H.264 compression. The LIVE video quality database contains the coded video sequences and their corresponding differential mean opinion score (DMOS) values. The consistency results of our proposed RR VQA on the coded video sequences are illustrated in Fig. 7. It can be observed that the relationship between the VQI and DMOS values is monotonic for a given source video, specifically the VQI value is monotonically increasing with the DMOS value for a given source video. The larger the VQI, the worse visual quality is the compressed video sequence, which possesses a larger DMOS value. For all the original video sequences, the relationship between the VQI and DMOS values is approximately linear for both MPEG-2 and H.264 coded video sequences. For each original video sequence, if a new MPEG-2 or H.264 coded video sequence is introduced, we can utilize the slope information that can be derived from Fig. 7, and its corresponding VQI value to predict its DMOS value with high accuracy. Consequently, the true perceptual quality of the coded sequence is obtained. In the following section, we will further evaluate the proposed RR VQA metric in the standardized way by measuring the relationship of the obtained VQI values and the provided subjective DMOS values.

The middle column of Fig. 7 shows the EVD values of the original and distorted videos, respectively. The MPEG-2 and H.264 compression will change the EVD value of each frame. During the compression process, more HF and MF components have been discarded than the LF components, which results in a smaller value of $M + H$ in (1). Therefore, a smaller EVD value of each frame is obtained, compared to the original value. The histogram of the difference image indicating the temporal information is illustrated in the right column of Fig. 7. Compared to the fitted GGD curve, the histogram distribution has been changed. As MPEG-2 and H.264 introduce more zero coefficients during the quantization process, a sharper and narrower distribution can be obtained from the distorted video sequence. Moreover, the actual histogram of the difference frame and the fitted GGD curve appear very close, while the EVD curves of the original and coded videos are quite different. The EVD is believed to affect the final VQI value more. As the compression distortion increases, although the LF component starts to be affected, the HF and MF components are quantized even more severely. Therefore, by computing EVD in (1), its value becomes even smaller. However, the quantization step of each frame is usually the same during the compression process. The temporal CBD changes will not be as significant as the spatial EVD variations. That is the reason why the spatial EVD contributes more to the final quality score than the temporal EVD, as to be demonstrated in the following section.

As shown in Fig. 7, the MPEG-2 coded sequences with the same perceptual quality (same DMOS value) demonstrate similar VQI values for different original sequences. On the contrary, the VQI values of the H.264 coded sequences with the same perceptual quality (same DMOS value) appear diversely. It means that the performance of the proposed RR VQA on MPEG-2 coded video sequences is more robust than that of H.264 coded video sequences. The reason is that the

EVD is calculated based on the 8×8 DCT. The energy variation of MPEG-2 can be accurately depicted, as the transform and quantization are performed based on an 8×8 block. For H.264, different block-size based intra prediction, inter motion estimation, and DCT result in an inaccurate energy variation calculation. The final VQI values of H.264 coded video sequences of the same perceptual quality will be different.

B. Consistency Test of the Proposed RR VQA Over Video Sequences With Simulated Distortions

Furthermore, as in [44] and [45], the consistency property of the proposed RR VQA is evaluated over the simulated video sequences with five distortion types at different distortion levels. These five distortions include: 1) Gaussian noise contamination, where the mean value is set as 0 and the distortion level is defined as the variance; 2) Gaussian blur distortion, where the filter is fixed as a 7×7 window, and the corresponding distortion level is determined by the standard deviation; 3) line jittering, where each line in a frame is shifted horizontally by a random number uniformly distributed between $[-S, S]$, and S defines the line jittering level; 4) frame jittering, where the whole frame is shifted together by a random number uniformly distributed between $[-S, S]$, and S defines the frame jittering level; and 5) frame dropping, which is simulated by discarding every 1 of N frames and repeating the previous frame to fill the empty frame, and $10 - N$ defines the distortion level. As claimed in [44], all these distortion types are associated with certain real-world scenarios. For example, frame jittering is often caused by irregular camera movement; line jittering often occurs when two fields of interlaced video signals are not synchronized.

Fig. 8 illustrates the consistency evaluation results over different distortion types of different levels. As we do not have the DMOS values of the video sequences contaminated by the aforementioned five distortions, the corresponding distortion level is utilized to indicate its perceptual quality. For each distortion type, the higher the distortion level, intuitively the worse is the perceptual quality of the processed video. Similar with MPEG-2 and H.264 coded video sequences, the relationship between the distortion level and the VQI is monotonic. Specifically, the VQI value is monotonically increasing with the distortion level for a given source video. Therefore, from this aspect, we can conclude that the proposed VQI is sensitive to the levels of different distortions. It demonstrates a consistent relationship with the distortion level of different distortion types. The spatial EVD and the temporal CBD information of the PA video sequence (at the largest distortion level) are illustrated. For Gaussian noise contamination and line-jittering distortion, the EVD value of the distorted video is larger than that of the original video. It means that the HF and MF components increase more than the LF components. For the Gaussian noise contamination, the Gaussian noise dominates the histogram distribution of the difference image. It demonstrates a much flatter distribution, compared to the fitted GGD curve. For the line-jittering and frame-jittering distortion, as the temporal relationship still exists, the histogram distribution of the difference image appears to be similar with the GGD fitted curve. However, the pixel values of the difference image will increase due to the jittering distortion. Therefore, there will not be so many zero values, which results in a smaller peak value as shown in Fig. 8. For Gaussian blur distortion, more HF and MF components are discarded compared with LF component. The EVD value decreases after the Gaussian blur process. And a sharper and narrower histogram distribution is obtained as more zero pixel values appear due to the filtering process. For the frame dropping distortion, the spatial EVD varies slightly, because of the close temporal relationship between adjacent frames. However, the pixel values of the difference image are all zero, as the previous frame is simply copied to fill the empty frame.

C. Performance Evaluation of the Proposed RR VQA on Compressed Video Sequences

In order to provide a more convincing result of the proposed RR VQA, we test the proposed method on the LIVE video quality database [54]. The performance can be evaluated by depicting the relationship of the obtained VQI values and the provided subjective ratings, specifically the DMOS value of each distorted video. The DMOS value is obtained by subjective viewing tests where many observers participated and provided their opinions on the visual quality of each distorted video. Therefore, it can be regarded as the ground truth for evaluating the metric performances. As suggested by video quality experts group HDTV test [56] and that in [57], we follow their evaluation procedure to evaluate the performance of the proposed metric. Let x_j represent the visual quality index of the jth distorted image obtained from the corresponding VQA. The five-parameter $\{\beta_1, \beta_2, \beta_3, \beta_4, \beta_5\}$ monotonic logistic function is employed to map x_j into V_j

$$V_j = \beta_1 \times \left(0.5 - \frac{1}{1 + e^{\beta_2 \times (x_j - \beta_3)}}\right) + \beta_4 \times x_j + \beta_5. \quad (13)$$

The corresponding five parameters are determined by minimizing the sum of squared differences between the mapped objecting score V_j and the subjective DMOS value. In order to evaluate the performances, three statistical measurements are employed. The linear correlation coefficient (LCC) measures the prediction accuracy. The Spearman rank-order correlation coefficient (SROCC) provides an evaluation of the prediction monotonicity. The root mean square prediction error (RMSE) is introduced for evaluating the error during the fitting process. According to the definitions, larger values of LCC and SROCC mean that the objective and subjective scores correlate better, that is to say, a better performance of the VQA. And the smaller RMSE values indicate smaller errors between the two scores, therefore a better performance.

We compare the performance of our proposed RR VQA with the representative RR video quality metric VQM [32], Yang's metric [24], Gunawan *et al.*'s metric [35], and J.246 [34], as well as several FR metrics: PSNR, SSIM [6], MSSIM [59], VSNR [7], VIF [58], and V-JND [11]. The corresponding results together with the reference type and RR data rates are illustrated in Table I. As PSNR, SSIM, MSSIM, VSNR, VIF, Yang's metric, and J.246 only provide frame-level quality

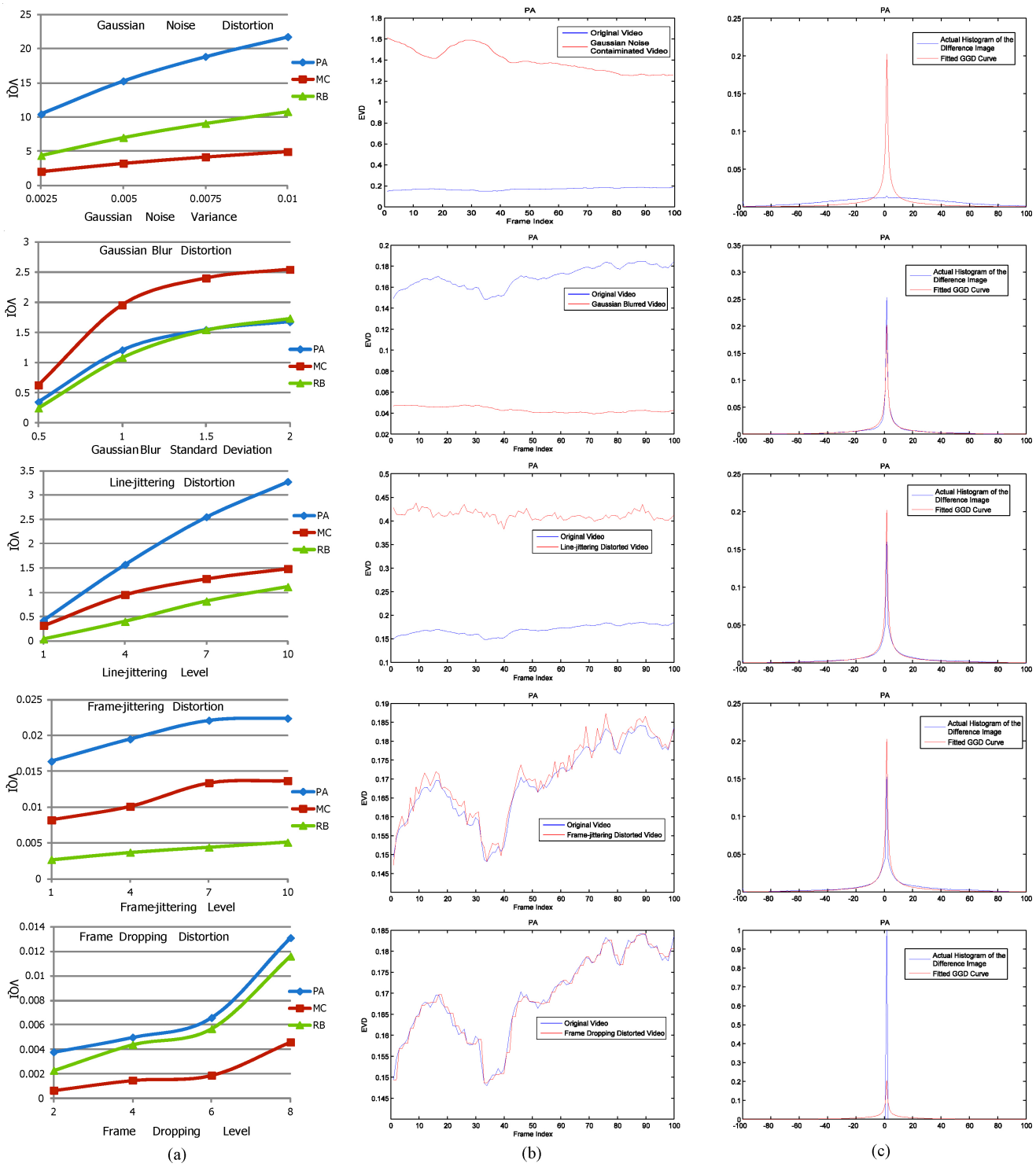

Fig. 8. Consistency evaluation of the proposed RR VQA over different distortions of different levels. (a) Proposed distortion measure VQI versus the distortion level. (b) Spatial EVD value of the PA video sequence at the largest distortion level. (c) Temporal histogram of the tenth difference image of the distorted video PA at the largest distortion level and the fitted GGD curve.

TABLE I
PERFORMANCES OF DIFFERENT VQAS OVER THE LIVE VIDEO QUALITY DATABASE (MPEG-2 AND H.264 ENCODED VIDEOS)

	LCC	SROCC	RMSE	Reference Type	Data Rate (25 f/s)
PSNR	0.4488	0.4157	9.188	FR	–
SSIM [6]	0.5946	0.5969	8.267	FR	–
MSSIM [59]	0.6671	0.6944	7.717	FR	–
VSNR [7]	0.3097	0.3041	9.777	FR	–
VIF [58]	0.6447	0.6350	7.860	FR	–
V-JND [11]	0.7558	0.7233	6.733	FR	–
J.246 [34]	0.5036	0.4460	8.883	RR	10 kb/s
Yang's metric [24]	0.5654	0.5366	8.484	RR	0.2 kb/s
Gunawan *et al.*'s metric [35]	0.4557	0.4082	9.152	RR	64 kb/s
VQM [32]	0.7003	0.6790	7.340	RR	150 kb/s
Proposed metric	0.7567	0.7486	6.722	RR	0.875 kb/s

scores, the final quality index of the video sequence is generated by averaging their outputs for each frame. For PSNR, SSIM, MSSIM, VSNR, VIF, and V-JND are FR metrics, the whole original frame should be available for quality analysis. Therefore, the RR data rates are regarded as the whole original video sequence. As for the RR VQAs, in order to ensure a fair comparison, the RR data rate is calculated based on video sequences of 25 f/s. For J.246, the locations and edge pixel values need to be encoded. As shown in [34], 14 extracted edge pixels per frame will result in the data rate as about 10 kb/s. For Yang's metric, the only one extracted ratio parameter can be quantized in 8-bit precision. The data rate (about 0.2 kb/s) is relatively small. For the Gunawan *et al.*'s metric, as shown in [35], the bit rate of the RR data is 64 kb/s. For VQM, the compression method has been researched in [33], which ensures a more than $30\times$ compression ratio compared to the original VQM features. The bit rate of the RR feature is about 150 kb/s. For the proposed method, as only 35 bits for each frame are required to encode all the features, $35 \times 25 = 875$ b/s are required to represent the features. Compared with the other RR VQAs except Yang's metric, the RR data rate of the proposed metric is much smaller. However, the performance of the proposed metric is better. Furthermore, if only the spatial EVD is employed for constructing the RR metric, the RR data rate will be the same as Yang's metric. The performance is better, as to be illustrated in the following section.

From Table I, it can be observed that the FR PSNR performs poorly, because it is not related to the HVS perception. Also, the VSNR performs badly, which can be attributed to two reasons. The first is that VSNR analyzes the HVS perception of the distortion in the wavelet domain. But the MPEG-2 and H.264 compression schemes introduce the distortions during the quantization process in DCT domain. The second is that VSNR is an image quality metric designed to capture the spatial distortions. For VQA, the temporal information is very important and needs to be accounted for. This is also the reason why SSIM, MSSIM and VIF perform successfully in image quality evaluation, but not so well on VQA. Although V-JND is a frame-based video quality metric, the JND model developed for each frame considers both the spatial and temporal HVS properties. Therefore, the V-JND method can accurately model the distortion of the video sequence. Hence, it provides a very good performance. However, the V-JND model is an FR VQA, which requires the whole video sequence for quality analysis. Yang's metric employs the DCT coefficient ratio to measure the video quality. Although a smaller RR data rate is required, Yang's metric only depicts the DCT coefficient distortion from the spatial aspect. The temporal information is not considered. For J.246, only the edge pixels in spatial domain are extracted for quality comparison. For Gunawan *et al.*'s metric, the harmonic and discriminative analysis is employed to depict the blocking and blur artifacts in the spatial domain. And the temporal motion information is employed to finally correct the quality values. From Table I, it can be observed that the performances of these metrics are not good enough, with SROCC values smaller than 0.6. The reason is that the temporal information is not accurately modeled. For VQA, the temporal distortion is very important and needs to be considered for developing an effective video quality metric. The RR VQM [32] is derived by recording several features which depict the spatial information losses, edge information changes, contrast information, and the color impairments. However, the feature extraction process is of high complexity. And the RR data rate after compression is still very large. As for our proposed method, it outperforms the VQM and the other FR quality metrics. It means that the proposed metric can effectively depict the perceptual quality of the compressed videos. Furthermore, the RR data rate is very small compared with the other RR VQAs, which will not introduce heavy burden for transmitting the RR features from the sender to the receiver side. The scatter-plots of different VQAs over the LIVE video quality database are illustrated in Fig. 9. It can be observed that for our proposed method the sample points scatter more closely around the fitted line. It means that the values predicted by the proposed method correlate better with the subjective ratings, specifically the DMOS values, demonstrating a better performance.

Moreover, for the proposed RR VQA and VQM [32], the triangles representing MPEG-2 coded videos scatter more closely to the fitted line, while several star points indicating H.264 coded videos are under or over estimated. As mentioned before, such scattering may be attributed to the fact that the features are calculated based on fixed block size, specifically EVD from the 8×8 DCT for the proposed RR VQA and the quality-related features from $(8 \times 8) \times 0.2\,second$ S-T region for the VQM. By considering the fixed 8×8 block in the spatial domain, the distortion of MPEG-2 can be accurately depicted, as the transform and quantization are performed based on 8×8 block. However, for H.264, different block-size based intra prediction, inter motion estimation, and DCT result in an inaccurate energy variation calculation. Therefore, the DMOS values correlate worse with the quality values of H.264 coded videos than that of MPEG-2 coded videos. In the future, we will consider the information of the H.264 coded video, specifically the transform and quantization block size. Then the EVD calculation can be extended to different block sizes for accurately capturing the energy variation, which is believed to be able to improve the performance of our proposed RR VQA.

As illustrated in Table I and Fig. 9, the effectiveness of our proposed RR VQA has been clearly demonstrated compared with the other RR metrics or even FR metrics in terms of both performance and required RR data rate. Therefore, as in [33], we may consider incorporating the proposed RR VQA into the video quality monitoring system, where the computational complexity of feature extraction and comparison needs to be evaluated. The spatial EVD of the proposed RR VQA only requires several addition and division processes after DCT, which can be calculated during the DCT process of the video encoding and the decoding procedure. For the temporal GGD modeling and CBD calculation, the processing complexity on the sender side is different from that on the receiver side. On the sender side, as shown in Fig. 4, the difference image is first obtained. Then, the histogram depicting the pixel value distribution is modeled by the GGD. Finally, the CBD distance as shown in (6) is calculated to indicate the modeling error. We implement the temporal feature extraction in MATLAB. During our implementation, we do not perform any optimization. A speed test is performed on our PC with a 3.0 GHz Quad-core CPU and 6.0 GB memory. For each difference frame, it only requires 0.7 s on average for obtaining the temporal features. On the receiver side, we only construct the histogram of the difference image and compare it with the fitted GGD. The distance shown in (10) is approximated. As we need not perform the fitting process, the computation of the temporal information is faster. The speed test is performed on the same PC, which indicates that only 0.14 s per difference frame on average is needed for the temporal quality analysis. If further optimization is employed, it is believed that the quality analysis on the receiver side can perform even faster, which can be incorporated into the video quality monitoring system.

D. Performance Evaluation of the Proposed RR VQA on Video Sequences Containing Transmission Distortions

As the transmission errors over wireless channel and IP network are more realistic for the video quality monitoring, the proposed RR VQA and other representative RR quality metrics are evaluated on the LIVE video sequences containing transmission distortions. These distortions are simulated transmission of H.264 compressed bit-streams through error-prone IP networks and wireless channel. The performance of these RR video quality metrics are illustrated in Table II. It can be observed that the proposed RR VQA can outperform J.246 [34], Yang's metric [24], and Gunawan *et al.*'s metric [35]. However, it performs worse than VQM [32]. For Yang's metric, it only employs the ratio between the parent coefficient (second DCT coefficient) and the child coefficient (the third and fourth DCT coefficient) to measure the video perceptual quality. Therefore, the distortion introduced by compression can be depicted, as the quantization process will change the ratio of DCT coefficients. However, the transmission distortion is not related to the ratio of DCT coefficients. Consequently, the perceptual qualities of the video sequences containing transmission distortion cannot be accurately depicted, which results in a bad performance of Yang's metric on these video sequences, as illustrated in Table II. For J.246 and Gunawan *et al.*'s metric, the performances are not as good as the proposed one, although they required more RR data rates for representing the original video sequence. On the other hand, VQM extracts many features for quality analysis, which are related to the specific distortions, such as blur, edge shifting, chroma spreading, color impairments, and so on. Most of these features can help to depict the distortions introduced during the video transmission. Therefore, the VQM performs well on these distorted video sequences. However, considering the data rates of different RR VQAs shown in Table I, the RR data rate of VQM after compression is 150 kb/s, which is about 170 times the proposed VQA (0.875 kb/s). It will introduce a heavy burden for the RR feature transmission.

TABLE II
PERFORMANCES OF DIFFERENT VQAS OVER THE LIVE VIDEO QUALITY DATABASE (IP AND WIRELESS DISTORTION)

		Proposed Method	Gunawan *et al.*'s Metric [35]	J.246 [34]	VQM [32]	Yang's Metric [24]
IP distortion	LCC	0.6000	0.4602	0.3168	0.6553	0.2684
	SROCC	0.5582	0.3766	0.3437	0.6383	0.1462
	RMSE	7.488	8.873	8.873	7.071	9.017
Wireless distortion	LCC	0.5546	0.4684	0.5061	0.7416	0.0842
	SROCC	0.5386	0.4638	0.4051	0.7220	0.1041
	RMSE	8.586	9.117	8.900	6.922	10.282

For the proposed RR VQA, although it can outperform the other RR metrics except VQM employing very low RR data rates, the performance is still not good enough. This can be attributed to two reasons. First, the transmission distortions over wireless channel and IP network are simulated from the H.264 compressed bit-streams. As discussed in Section IV-C, the proposed spatial EVD is calculated based on the fixed block size, specifically from the 8×8 DCT. However, for H.264, different block-size based intra prediction, inter motion estimation, and DCT are utilized, which result in an inaccurate energy variation calculation. Therefore, the transmission distortions simulated from H.264 compressed bit-streams cannot be accurately depicted. Second, in the RR VQA, the properties of the transmission errors, such as the error patterns, are not considered. If some features related with these errors are further incorporated, the RR VQA can more accurately depict the perceptual qualities of these degraded video sequences. In this paper, the authors only focus on the RR VQA for the compressed video sequences. In the future, as the RR data rate of the proposed metric is relatively small, we will consider incorporating more RR features to better handle the transmission errors.

E. Performance Analysis of Each Component

In this section, we evaluate the corresponding contribution for each component of our proposed metric in (11). To this end, we derive three different metrics to generate the frame-level quality score. The first one is the spatial EVD distance, which means $Q_s = EL$, as in (7). The second one is the weighted spatial EVD distance, which means $Q_s = EL_V$, as in (8). The third is the temporal CBD distance defined as

$$Q_s = \log_{10}\left(1 + \frac{d_{\mathrm{CBD}}(p, p_d)}{c}\right) \tag{14}$$

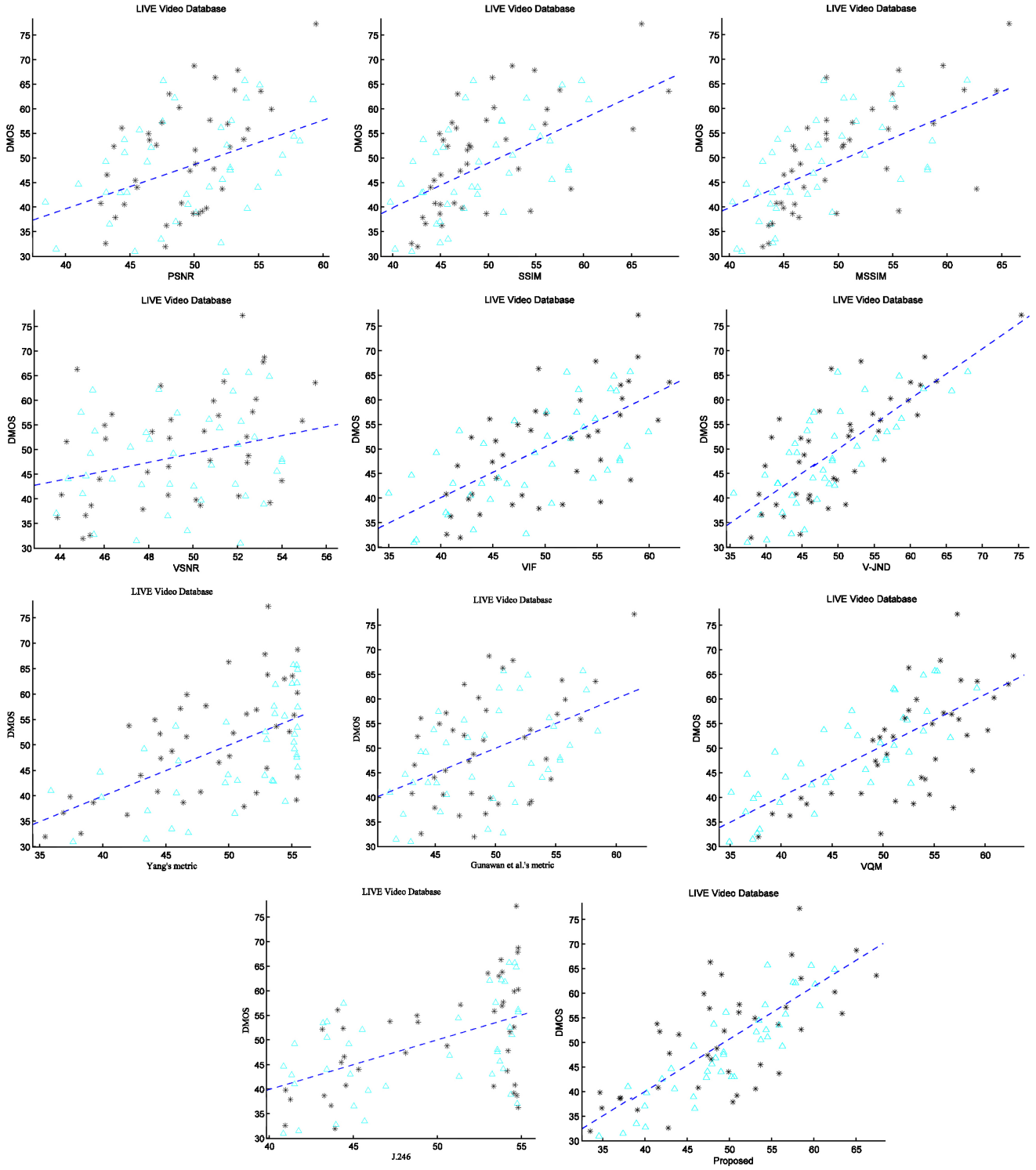

Fig. 9. Scatter plots of the DMOS values versus model predictions on the LIVE video quality database. Each sample point represents one test video. The star indicates H.264 encoded video sequence, while the triangle indicates the MPEG-2 compressed video sequence. First row from left to right: PSNR, SSIM, and MSSIM. Second row from left to right: VSNR, VIF, and V-JND. Third row from left to right: Yang's metric, Gunawan *et al.*'s metric, and VQM. Fourth row from left to right: J.246, and the proposed method.

TABLE III
PERFORMANCES OF DIFFERENT COMPONENTS OF THE PROPOSED RR VQA OVER THE LIVE VIDEO QUALITY DATABASE (MPEG-2 AND H.264 ENCODED VIDEOS)

	Spatial EVD Distance as in (7)	Weighted Spatial EVD Distance as in (8)	Weighted Spatial EVD Distance as in (15)	Temporal CBD Distance as in (14)
LCC	0.3986	0.5965	0.5958	0.4135
SROCC	0.3475	0.5992	0.5717	0.3950
RMSE	9.430	8.253	8.258	9.362

where c is also set as 0.001. Their corresponding performances are illustrated in Table III.

It can be observed that all of these three components are necessary for the proposed RR VQA. The spatial EVD distance as in (7) performs the worst. The reason is that it only considers the absolute difference of corresponding DCT coefficients, which captures the information loss during the quantization process. Therefore, it does not correlate well with the HVS perception. Furthermore, the distance in (7) is performed in the spatial domain. It does not consider the temporal information, which is critical to VQA. The HVS related weighting strategy of the EVD distance is tested as formulated in (8). As discussed before, the EVD of the original frame can represent its texture characteristic. The higher the EVD value, the more texture information it may contain. And the more texture information, the more distortion it can mask. Therefore, the EVD value is employed to simulate the texture masking property of the HVS as shown in (8). Compared with (7), the performance is significantly improved. It means that the EVD can accurately model the texture masking property of the HVS. Actually, Yang's metric [24] also employed the ratio of DCT coefficient to measure the video quality in spatial domain. It employs the ratio between the parent coefficient (second DCT coefficient) and the child coefficient (the third and fourth DCT coefficient). However, it does not consider the texture masking effect of the HVS. Therefore, the performance, as illustrated in Table I, is not as good as that of (8).

For the coded video sequences, the artifacts in the processed video are superposed onto the original video sequence, which is regarded as the masker signal. Therefore, as shown in (8), we employed EVD_{ori} to mask the compression artifacts, which are introduced by quantization process. However, as discussed in [60], for the content of video sequence and the compression artifacts, one's presence will affect the visibility of the other. It is believed that the coded video sequence lacking of detailed information can also mask the artifacts. Therefore, we also evaluated the weighted spatial EVD distance, where EVD_{pro} is employed for simulating the HVS texture masking property

$$EL'_V = \frac{EL}{EVD_{\text{pro}}} = \frac{|EVD_{\text{ori}} - EVD_{\text{pro}}|}{EVD_{\text{pro}}}. \tag{15}$$

The corresponding performance is shown in Table III. It can be observed that EL'_V performs better than the spatial EVD distance formulated in (7). It means that EVD_{pro} can also simulate the texture masking property of HVS. However, EVD_{ori} as the masker signal can generate a better performance. Therefore, we only consider employing the original video signal to simulate the HVS texture masking effect in this paper. It means that EVD_{ori} is employed to weight the spatial EVD distance as in (8). In the future, we will research on how to accurately model the HVS texture masking effect by considering both the original and processed video signal.

The temporal CBD distance is evaluated as expressed in (14). The temporal CBD distance depicts the temporal statistical characteristic. It has been demonstrated to be related to the HVS perception, as shown in [44] and [45]. The distortions in the video will result in the statistical characteristic changes. By accurately capturing these changes, the corresponding perceptual quality can be described. Comparing the performances in Table III with those in Table I, it is clear that the spatial distance or the temporal distance alone cannot outperform the integrated one. It means that only the spatial or temporal distortion alone is not sufficient to depict the perceptual quality of the video sequence. An effective RR VQA needs to accurately capture not only the spatial distortion but also the temporal one. This is the main reason why our RR VQA outperforms the other quality metrics, such as PSNR, SSIM, VSNR, Yang's metric, J.246, and VQM.

V. CONCLUSION

In this paper, an effective RR VQA was proposed by depicting the distortions from both the spatial and temporal perspectives. The EVD captured the information loss of each individual frame, which was also employed to simulate the texture masking property of the HVS. The GGD function and CBD distance were utilized to describe the temporal statistical characteristics. With evaluation on the subjective quality video database, the proposed RR VQA outperformed the representative RR metrics, and also the FR metrics. Due to its simplicity and efficiency in terms of feature representation, the proposed metric may be considered for incorporation into the video quality monitoring system.

REFERENCES

[1] Z. Wang, H. R. Sheihk, and A. C. Bovik, "Objective video quality assessment," in *The Handbook of Video Databases: Design and Application*, B. Furht and O. Marqure, Eds. Boca Raton, FL: CRC Press, Sep. 2003, pp. 1041–1078.

[2] G. Zhai, J. Cai, W. Lin, X. Yang, W. Zhang, and M. Etoh, "Cross-dimensional perceptual quality assessment for low bit-rate videos," *IEEE Trans. Multimedia*, vol. 10, no. 7, pp. 1316–1324, Nov. 2008.

[3] A. K. Moorthy, K. Seshadrinathan, R. Soundararajan, and A. C. Bovik, "Wireless video quality assessment: A study of subjective scores and objective algorithms," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 4, pp. 587–599, Apr. 2010.

[4] Z. Wang and A. C. Bovik, *Modern Image Quality Assessment*. New York: Morgan and Claypool, 2006.

[5] Z. Wang, A. C. Bovik, H. R. Sheihk, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[6] Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal Process.: Image Commun.*, vol. 19, no. 2, pp. 121–132, Feb. 2004.

[7] D. M. Chandler and S. S. Hemami, "VSNR: A wavelet-based visual signal-to-noise ratio for natural images," *IEEE Trans. Image Process.*, vol. 16, no. 9, pp. 2284–2298, Sep. 2007.

[8] K.-C. Yang, C. C. Guest, K. El-Maleh, and P. K. Das, "Perceptual temporal quality metric for compressed video," *IEEE Trans. Multimedia*, vol. 9, no. 7, pp. 1528–1535, Nov. 2007.

[9] L. Ma, S. Li, and K. N. Ngan, "Visual horizontal effect for image quality assessment," *IEEE Signal Process. Lett.*, vol. 17, no. 7, pp. 627–630, Jul. 2010.

[10] F. Zhang, L. Ma, S. Li, and K. N. Ngan, "Practical image quality metric applied to image coding," *IEEE Trans. Multimedia*, vol. 13, no. 4, pp. 615–624, Aug. 2001.

[11] L. Ma, K. N. Ngan, F. Zhang, and S. Li, "Adaptive block-size transform based just-noticeable difference model for images/videos," *Signal Process. Image Commun.*, vol. 26, no. 3, pp. 162–174, Mar. 2011.

[12] A. K. Moorthy and A. C. Bovik, "Efficient video quality assessment along temporal trajectories," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 11, pp. 1653–1658, Nov. 2010.

[13] S. S. Hemami and A. R. Reibman, "No-reference image and video quality estimation: Applications and human-motivated design," *Signal Process. Image Commun.*, vol. 25, no. 7, pp. 469–481, Aug. 2010.

[14] H. R. Sheikh, A. C. Bovik, and L. Cormack, "No-reference quality assessment using nature scene statistics: JPEG 2000," *IEEE Trans. Image Process.*, vol. 14, no. 11, pp. 1918–1927, Nov. 2005.

[15] L. Liang, S. Wang, J. Chen, S. Ma, D. Zhao, and W. Gao, "No-reference perceptual image quality metric using gradient profiles for JPEG 2000," *Signal Process. Image Commun.*, vol. 25, no. 7, pp. 502–516, Aug. 2010.

[16] T. Brandao and M. P. Queluz, "No-reference image quality assessment based on DCT domain statistics," *Signal Process.*, vol. 88, no. 4, pp. 822–833, Apr. 2008.

[17] T. Brandao and M. P. Queluz, "No-reference quality assessment of H.264/AVC encoded video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 11, pp. 1437–1447, Nov. 2010.

[18] R. Ferzli and L. J. Karam, "A no-reference objective image sharpness metric based on the notion of just noticeable blur (JNB)," *IEEE Trans. Image Process.*, vol. 18, no. 4, pp. 445–448, Apr. 2009.

[19] Q. Huynh-Thu and M. Ghanbari, "No-reference temporal quality metric for video impaired by frame freezing artifacts," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2009, pp. 2221–2224.

[20] Z. Wang and A. C. Bovik, "Reduced and no-reference image quality assessment: The natural scene statistic model approach," *IEEE Signal Process. Mag.*, vol. 28, no. 6, pp. 29–40, Nov. 2011.

[21] S. Wolf and M. H. Pinson, "Low bandwidth reduced reference video quality monitoring system," in *Proc. Int. Workshop Video Process. Quality Metrics Consumer Electron.*, Jan. 2005, pp. 76–79.

[22] S. Wolf and M. H. Pinson, "Spatio-temporal distortion metrics for in-service quality monitoring of any digital video system," *Proc. SPIE*, vol. 3845, pp. 266–277, Sep. 1999.

[23] P. Le Callet, C. V. Gaudin, and D. Barba, "Continuous quality assessment of MPEG2 video with reduced reference," in *Proc. Int. Workshop Video Process. Quality Metrics Consumer Electron.*, Jan. 2005, pp. 11–16.

[24] S. Yang, "Reduced reference MPEG-2 picture quality measure based on ratio of DCT coefficients," *Electron. Lett.*, vol. 47, no. 6, pp. 382–383, Mar. 2011.

[25] T. Oelbaum and K. Diepold, "Building a reduced reference video quality metric with very low overhead using multivariate data analysis," *J. Syst. Cybern. Informatics*, vol. 6, no. 5, pp. 81–86, 2008.

[26] M. Tagliasacchi, G. Valenzise, M. Naccari, and S. Tubaro, "A reduced-reference structural similarity approximation for videos corrupted by channel errors," *Multimedia Tools Appl.*, vol. 48, no. 3, pp. 471–492, Jul. 2010.

[27] P. Le Callet, C. V. Gaudin, and D. Barba, "A convolutional neural network approach for objective video quality assessment," *IEEE Trans. Neural Netw.*, vol. 17, no. 5, pp. 1316–1327, May 2006.

[28] M. Carnec, P. Le Callet, and D. Barba, "An image quality assessment method based on perception of structural information," in *Proc. IEEE Int. Conf. Image Process.*, vol. 3. Sep. 2003, pp. 185–188.

[29] M. Carnec, P. Le Callet, and D. Barba, "Visual features for image quality assessment with reduced reference," in *Proc. IEEE Int. Conf. Image Process.*, vol. 1. Sep. 2005, pp. 421–424.

[30] D. Tao, X. Li, W. Lu, and X. Gao, "Reduced-reference IQA in contourlet domain," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 39, no. 6, pp. 1623–1627, Dec. 2009.

[31] U. Engelke, M. Kusuma, H. J. Zepernick, and M. Caldera, "Reduced-reference metric design for objective perceptual quality assessment in wireless imaging," *Signal Process. Image Commun.*, vol. 24, no. 7, pp. 525–547, Aug. 2009.

[32] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Trans. Broadcasting*, vol. 50, no. 3, pp. 312–322, Sep. 2004.

[33] M. Makar, Y.-C. Lin, A. F. de Araujo, and B. Girod, "Compression of VQM features for low bit-rate video quality monitoring," in *Proc. IEEE Workshop Multimedia Signal Process.*, Oct. 2011.

[34] *Perceptual Visual Quality Measurement Techniques for Multimedia Services Over Digital Cable Television Networks in the Presence of a Reduced Bandwidth Reference*, ITU-T Rec. J.246, Aug. 2008 [Online]. Available: http://www.itu.int/rec/T-REC-J.246/en

[35] I. P. Gunawan and M. Ghanbari, "Reduced-reference video quality assessment using discriminative local harmonic strength with motion consideration," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 1, pp. 71–83, Jan. 2010.

[36] C. T. E. R. Hewage and M. G. Martini, "Reduced-reference quality assessment for 3-D video compression and transmission," *IEEE Trans. Consumer Electron.*, vol. 57, no. 3, pp. 1185–1193, Aug. 2011.

[37] Z. Wang, G. Wu, H. R. Sheikh, E. P. Simoncelli, E. Yang, and A. C. Bovik, "Quality-aware images," *IEEE Trans. Image Process.*, vol. 15, no. 6, pp. 1680–1689, Jun. 2006.

[38] Z. Wang and E. P. Simoncelli, "Reduced-reference image quality assessment using a wavelet-domain natural image statistic model," in *Proc. SPIE Hum. Vision Electron. Imaging*, Jan. 2005, pp. 149–159.

[39] Q. Li and Z. Wang, "Reduced-reference image quality assessment using divisive normalization-based image representation," *IEEE J. Select. Top. Signal Process.*, vol. 3, no. 2, pp. 201–211, Apr. 2009.

[40] T. M. Cover and J. A. Thomas, *Element of Information Theory*. New York, Wiley, 1991.

[41] L. Ma, S. Li, F. Zhang, and K. N. Ngan, "Reduced-reference image quality assessment using reorganized DCT-based image representation," *IEEE Trans. Multimedia*, vol. 13, no. 4, pp. 824–829, Aug. 2011.

[42] R. Soundararajan and A. C. Bovik, "RRED indices: Reduced reference entropic differencing framework for image quality assessment," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, May 2011, pp. 1149–1152.

[43] J. A. Redi, P. Gastaldo, I. Heynderickx, and R. Zunino, "Color distribution information for reduced-reference assessment of perceived image quality," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 12, pp. 1757–1769, Dec. 2010.

[44] K. Zeng and Z. Wang, "Temporal motion smoothness measurement for reduced-reference video quality assessment," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Mar. 2010, pp. 1010–1013.

[45] K. Zeng and Z. Wang, "Quality-aware video based on robust embedding of intra and interframe reduced-reference features," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2010, pp. 3229–3232.

[46] B. Girod, "What's wrong with mean-squared error," in *Digital Images and Human Vision*. Cambridge, MA: MIT Press, 1993, pp. 207–220.

[47] X. Zhang, W. Lin, and P. Xue, "Improved estimation for just-noticeable visual distortion," *Signal Process.*, vol. 85, no. 4, pp. 795–808, Apr. 2005.

[48] X. Zhang, W. Lin, and P. Xue, "Just-noticeable difference estimation with pixels in images," *J. Visual Commun. Image Representation*, vol, 19, no. 1, pp. 30–41, Jan. 2008.

[49] J. R. Jain and A. K. Jain, "Displacement measurement and its application in interframe image coding," *IEEE Trans. Commun.*, vol. 29, no. 12, pp. 1799–1808, Dec. 1981.

[50] Y. C. Lin and S. C. Tai, "Fast full-search block-matching algorithm for motion-compensated video compression," *IEEE Trans. Commun.*, vol. 45, no. 5, pp. 527–531, May 1997.

[51] B. K. P. Horn and B. G. Schunck, "Determining optical flow," *Artif. Intell.*, vol. 17, nos. 1–3, pp. 185–203, Aug. 1981.

[52] C. Guo, Q. Ma, and L. Zhang, "Spatio-temporal saliency detection using phase spectrum of quaternion Fourier transform," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 2908–2915.

[53] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its application in image and video compression," *IEEE Trans. Image Process.*, vol. 19, no. 1, pp. 185–198, Jan. 2010.

[54] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack. (2010). *LIVE Video Quality Database* [Online]. Available: http://live.ece.utexas.edu/research/quality/live_video.html

[55] M. N. Do and M. Vetterli, "Wavelet-based texture retrieval using generalized Gaussian density and Kullback-Leibler distance," *IEEE Trans. Image Process.*, vol. 11, no. 2, pp. 146–158, Feb. 2002.

[56] VQEG. (2000). *Final Report From the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment* [Online]. Available: http://www.vqeg.org

[57] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, Nov. 2006.

[58] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006.

[59] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multi-scale structural similarity for image quality assessment," in *Proc. IEEE Asilomar Conf. Signals Syst. Comput.*, Nov. 2003, pp. 1398–1402.

[60] S. Li, F. Zhang, L. Ma, and K. N. Ngan, "Image quality assessment by separately evaluating detail losses and additive impairments," *IEEE Trans. Multimedia*, vol. 13, no. 5, pp. 935–949, Oct. 2011.

Lin Ma (S'09) received the B.E. and M.E. degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 2006 and 2008, respectively. He is currently pursuing the Ph.D. degree with the Department of Electronic Engineering, Chinese University of Hong Kong (CUHK), Shatin, Hong Kong.

He was a Research Intern with Microsoft Research Asia, Beijing, China, from October 2007 to March 2008. He was a Research Assistant with the Department of Electronic Engineering, CUHK, from November 2008 to July 2009. He was a Visiting Student with the School of Computer Engineering, Nanyang Technological University, Singapore, from July 2011 to September 2011. His current research interests include image or video quality assessment, superresolution, restoration, and compression.

Mr. Ma received the Best Paper Award in the Pacific-Rim Conference on Multimedia in 2008. He received the Microsoft Research Asia Fellowship in 2011.

Songnan Li (S'08) received the B.S. and M.S. degrees from the Department of Computer Science, Harbin Institute of Technology, Harbin, China, in 2004 and 2006, respectively. He is currently pursuing the Ph.D. degree with the Department of Electronic Engineering, Chinese University of Hong Kong, Shatin, Hong Kong.

His current research interests include visual quality assessment, video deinterlacing, video compression, and code optimization.

King Ngi Ngan (M'79–SM'91–F'00) received the Ph.D. degree in electrical engineering from Loughborough University, Loughborough, U.K.

He is currently a Chair Professor with the Department of Electronic Engineering, Chinese University of Hong Kong, Shatin, Hong Kong. He was previously a Full Professor with Nanyang Technological University, Singapore, and with the University of Western Australia, Perth, Australia. He holds honorary and visiting professorships with numerous universities in China, Australia, and South East Asia. He has published extensively, including three authored books, six edited volumes, over 300 refereed technical papers, and has edited nine special issues in journals. He holds ten patents in image or video coding and communications.

Dr. Ngan has served as an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the *Journal on Visual Communications and Image Representation*, the *EURASIP Journal of Signal Processing: Image Communication*, and the *Journal of Applied Signal Processing*. He has chaired a number of prestigious international conferences on video signal processing and communications, and has served on the advisory and technical committees of numerous professional organizations. He co-chaired the IEEE International Conference on Image Processing, Hong Kong, in September 2010. He is a fellow of IET, U.K., and IEAust, Australia, and was an IEEE Distinguished Lecturer from 2006 to 2007.