REDUCED REFERENCE VIDEO QUALITY ASSESSMENT BASED ON SPATIAL HVS MUTUAL MASKING AND TEMPORAL MOTION ESTIMATION

Lin Ma[†‡], King N. Ngan[†], and Long Xu[#]
[†]Department of Electronic Engineering, the Chinese University of Hong Kong, Hong Kong
[‡]Lenovo Corporate Research Hong Kong Branch, Hong Kong
[#]School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing, China

ABSTRACT

In this paper, an effective reduced reference (RR) video quality assessment (VQA) is proposed by depicting both the spatial and temporal statistical characteristics of the video signals. For each video frame, spatial information change (SIC) is employed to depict the energy variation. A novel mutual masking strategy based on the extracted SIC is proposed to accurately simulate the human visual system (HVS) texture masking property. For adjacent video frames, the temporal relationship is depicted by block-based motion estimation (BME). The generalized Gaussian density (GGD) function is employed to depict the histogram natural statistic of the residual frame after BME. The city-block distance (CBD) is used to measure the distance between histograms of the original and distorted video sequence. By pooling the measurements from both spatial and temporal perspectives, an efficient RR VQA is constructed. With the evaluations on the public video quality database, the proposed RR VQA demonstrated to be more effective than the representative RR VQAs and even the full-reference (FR) VQAs, such as peak signal-to-noise ratio (PSNR) and structure similarity index (SSIM) in matching the subjective ratings. Furthermore, the proposed RR VQA demonstrated to be much more effective and efficient, requiring only a very small number of bits for the RR feature representation.

Index Terms— Reduced reference, quality assessment, motion estimation, mutual masking, human visual system (HVS)

1. INTRODUCTION

Nowadays, more and more information blooms in the internet and is presented to consumers in the form of visual signals, especially images and videos, as the intuitive and faithful depiction of things in life and work. The images and videos are affected by a wide variety of distortions during acquisition, compression, storage, processing, transmission, and reproduction processes, which result in the perceptual quality degradation [1]. As a result, perceptual quality assessment plays a very important role in today's visual signal processing and communication systems.

The subjective quality assessment methods are believed to be the most reliable way for evaluating the perceptual quality of the visual signal. However, subjective quality assessment suffers from drawbacks as following. First, it is time-consuming, laborious and expensive, and requires many human subjects and repeated viewing sessions. Second, it is not feasible for on-line or practical signal manipulations (such as visual signal transmission). Third, even in the cases where human assessment is possible (such as manufacturing assembly lines) and cost is not a problem, it depends upon the assessor's physical conditions, emotion states, personal experience and preference, and so on. Therefore, in order to tackle these drawbacks, many researchers carried on subjective testing processes by recruiting many observers into the subjective assessment. The subjective rating scores after processing are believed to accurately indicate the true perceptual quality of the visual signal. In this manner, several public image/video databases are constructed, which consist of the distorted visual signals as well as the corresponding subjective ratings. The representative image databases include LIVE [2], IRCCyN/IVC [3], TID [4], the retargeting [5], and so on. The representative video databases include LIVE [6], IVP [7], and so on. Based on these databases, researchers can efficiently evaluate the performances of quality metrics and thereby develop more effective ones to automatically measure the perceptual quality of the visual signal.

Objective quality assessment can be roughly classified into three categories: full-reference (FR), reduced-reference (RR), and no-reference (NR). FR quality metrics require the full information of the image/video to evaluate the corresponding perceptual quality. Therefore, FR quality metrics are mostly employed for visual signal compression, watermarking, and so on, where the original visual signal is fully provided. The most famous and widely utilized quality metric is mean square error (MSE) and the related peak signal-to-noise ratio (PSNR). They are appealing for their simple formulation, easy computation and optimization [10]. However, it has been well recognized that MSE/PSNR does not correlate closely with human visual system (HVS) perception [9]. Another representative FR quality metric is structural similarity index (SSIM) [10], as well as its relatives [11] [12]. SSIM is derived by capturing the information loss

The work described in this paper was partially supported by a grant from the Research Grants Council of the Hong Kong SAR, China (Project CUHK415712), and and in part by the National Natural Science Foundation of China under Grant 61202242.

of image structures, which are believed to be more sensitive to HVS. However, for practical applications, the original image/video is always unavailable for quality analysis. Therefore, NR quality metrics are demanded [13]-[15]. Most of the NR quality metrics in previous literatures focus on evaluating images degraded by specific distortions, such as JPEG compression, JPEG 2000, blurring, and so on. Recently, in [14] [15], machine learning methods are employed to fuse the quality metrics designed for specific distortions together. However, as the reference image/video is unavailable, the performances of NR quality metrics can still not be ensured.

RR quality metrics are the compromise between FR and RR quality metrics, which require part information or extract several representative features from the original visual signal. Based on the extracted features, the perceptual quality of the distorted visual signal can be evaluated. The RR features can be easily encoded and represented in limited bits, which can be embedded into the visual signal or transmitted to the client side for quality monitoring. Based on different ways of the approaches, RR quality metrics can be roughly categorized into three classes [16] [17]. The first approach is based on modeling the distortions. The quality metrics [18] [19] are mostly developed for the videos degraded by specific distortions, such as MPEG-2 compression. The second approach is developed on modeling HVS. The quality metrics [20] [21] [26] are developed by considering the HVS properties. For example, in [21], several HVS related features are extracted to indicate the spatial information losses, edge information changes, contrast information, and color impairments. Therefore, an effective RR quality metric named as VQM is developed, which has been adopted as a North American standard by ANSI. The third type of approaches is based on modeling natural visual signal statistics. The underline essential of these metrics [16] [17] [22]-[25] is that most real-world distortions will disturb the visual signal statistics. The variations of the statistics can be used to quantify the degradation level of the image/video. For example, generalized Gaussian density (GGD) is employed to depict the wavelet coefficient distribution in [22]. In [23] [25], GGD is employed to depict the coefficient distribution in reorganized DCT (RDCT) domain, which can ensure better performance. These three types of approaches are different but related. Therefore, in order to develop a more effective RR quality metric, the visual signal statistics and the HVS properties need to be considered together.

In this paper, a novel RR video quality assessment (VQA) is proposed by depicting the distortions from both spatial and temporal perspectives. Spatial information change (SIC) is utilized to depict the energy variation of each video frame. Moreover, SIC is researched to model the HVS texture masking property in a mutual manner. Block-based motion estimation (BME) is employed to depict the temporal relationship between adjacent video frames. The histogram of the prediction residual is characterized by GGD to depict the natural statistics. CBD is employed to measure the histogram distribution differences between the original and distorted video sequence. Finally, by pooling the spatial and temporal measurements together, an effective RR VQA is developed.

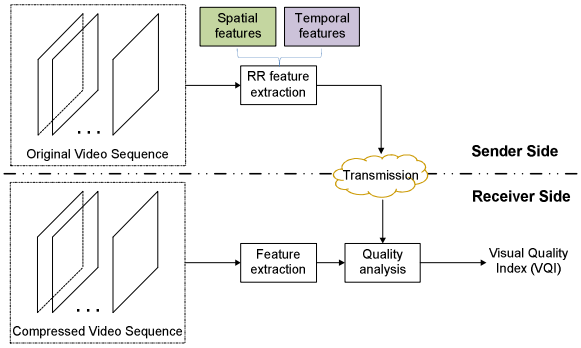

Figure. 1. General framework of RR VQA

The rest of the paper is organized as follows. Section 2 will introduce the proposed RR VQA in detailed. The experimental results in Section 3 will demonstrate its effectiveness. Finally, Section 4 will conclude the paper.

2. PROPOSED REDUCED REFERENCE VIDEO QUALITY ASSESSMENT

The general framework of the RR VQA is illustrated in Figure. 1. At the sender side, RR features which sensitively correlate with the video distortions are extracted. With efficient representation, RR features are transmitted to the receiver side. By performing the quality analysis, the visual quality index (VQI) of each distorted video sequence is obtained to indicate its perceptual quality.

The key components for constructing an effective and efficient RR VQA are the processes of RR feature extraction, RR feature representation for transmission, and quality analysis based on the RR features, respectively. First, at the sender side, the RR features need to correlate with HVS perception, which can effectively represent the level of the introduced distortion. Second, RR features need to be efficiently coded and represented. The amount of the total RR features should not be too large to introduce heavy overhead burden for transmission. Third, at the receiver side, how to compare the features from the original and distorted video sequence needs to be further researched. Based on the feature comparison, the perceptual quality of the distorted sequence can be analyzed. In this paper, an RR VQA is proposed to handle the aforementioned three challenges, the detailed information of which is addressed in the following.

2.1. Reduced reference feature extraction

As discussed before, the extracted RR features should not only be sensitive to the HVS perception, but also represent the degradation level of the distorted video sequence. Therefore, in order to accurately capture the distortion occurred in both spatial and temporal domain, the proposed RR VQA

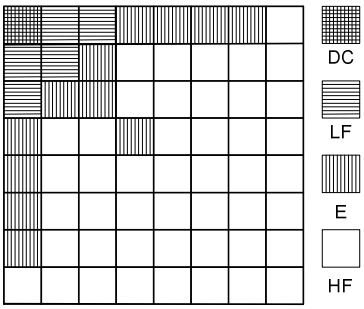

Figure. 2. DCT coefficient classification

will extract RR feature by considering the statistical information of each single frame (spatial) and adjacent frames (temporal).

For compressed video sequences, distortion will be introduced to the discrete cosine transform (DCT) coefficients, because of the quantization process. In order to accurately depict the quantization distortions, an RR feature SIC is employed to capture the distortions from the spatial perspective. After performing DCT, the DCT coefficients can be grouped into four categories, namely DC coefficient, low frequency (LF), edge (E), and high frequency (HF), as shown in Figure. 2. According the experimental results in [27], the coefficient values of LF, E, and HF can help to determine the presence of edges or good approximation of the block texture energy. Therefore, based on the values of the DCT coefficients in the corresponding areas, the 8×8 block can be categorized into three types, namely plain, edge, and texture. The HVS sensitivity to error is generally highest in plain regions and decreases in the order of plain, edge, and texture. In this manner, the HVS texture masking property can be modeled, which have been widely employed for developing HVS just noticeable difference (JND) model [28].

In this paper, in order to capture the information changes due to the compression, SIC is utilized by accounting for DC, LF, E, and HF DCT coefficients defined as following:

$$SIC = \frac{\sum|E|+\sum|HF|}{\sum|DC|+\sum|LF|}. \quad (1)$$

The quantization process will introduce different distortions to different DCT coefficients. Higher quantization will be assigned to higher frequency DCT coefficients, such as E and HF coefficients. Therefore, SIC can help to capture the distortions introduced by the compression process. Moreover, as discussed in [27], different DCT coefficients in different areas can represent the edge or texture information. which can thus model the HVS texture masking property. Therefore, SIC can further model the HVS texture masking property to quantify the perceptual distortion.

From the temporal perspective, in order to extract the perceptual related RR features, temporal relationships between adjacent frames need to be depicted. In [20] [24], the adjacent frame difference is employed to measure the temporal content. However, the object motion information cannot be accurately depicted to model the HVS temporal property [29]. In this paper, BME is employed to depict the temporal relationship:

$$PE(i) = BME\big(I(i), I(i-1)\big),\ i \in \{2,3,\dots,N\}, \quad (2)$$

where $PE(i)$ indicates the prediction residual after performing BME, which compensates the inaccurate BME.

In order to further illustrate the statistical property of the prediction residual, several video frames are chosen from LIVE video database [6], as illustrated in Figure. 3. The pixel values of $PE(11)$ are scaled by $128 + PE(11)$ for better visualization. Its histogram is illustrated as the blue line. It can be observed that the histogram distribution is of highly kurtosis (with a sharp peak at zero and a fat-tail distribution). As demonstrated in [22]-[25], the highly kurtotic distribution can be modeled by generalized Gaussian density (GGD) function defined as following:

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

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

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

The GGD fitting curve of each $PE(11)$ for the original video sequences is illustrated in Figure. 3 (red line). It can be observed that the two curves present nearly the same distribution, which means that GGD can accurately depict the histogram distribution. Furthermore, in order to improve the GGD modeling accuracy, another parameter named as city-block distance (CBD) is employed to measure the difference between the actual histogram p and the GGD distribution $p_{\alpha,\beta}$, which is defined as:

$$d_{CBD}\big(p, p_{\alpha,\beta}\big) = \sum_{i=1}^{h_L}\big|p(i) - p_{\alpha,\beta}(i)\big|, \quad (5)$$

where h_L is the histogram bin number.

With the GGD modeling, the histogram statistical property of each prediction residual frame can be accurately depicted by three parameters $\{\alpha, \beta, d_{CBD}\big(p, p_{\alpha,\beta}\big)\}$, which are employed as the temporal RR features.

2.2. Reduced reference feature representation

For each video frame, one RR feature SIC and three RR features related to GGD and CBD, are extracted for quality analysis from the spatial and temporal perspectives, respectively. How to encode and represent these RR features more efficiently is another key challenge for RR VQA design. In this paper, SIC parameter is quantized into 8-bit precision for transmission. For the 3 GGD parameters, same as in [22] [24], β and $d_{CBD}\big(p, p_{\alpha,\beta}\big)$ are quantized into 8-bit precision, and α is represented using 11-bit floating point, with 8 bits for mantissa and 3 bits for exponent. In this manner, there will be $8 + 8 + 8 + 8 + 3 = 35$ bits required to encode and represent the extracted RR features for each frame. If the video sequence is of 25 fps, the RR data rate will be 0.875 kbps. Compared with other representative RR VQAs [21] [30] [31], the data rate is very small, and therefore it can be easily transmitted to the receiver side for quality analysis.

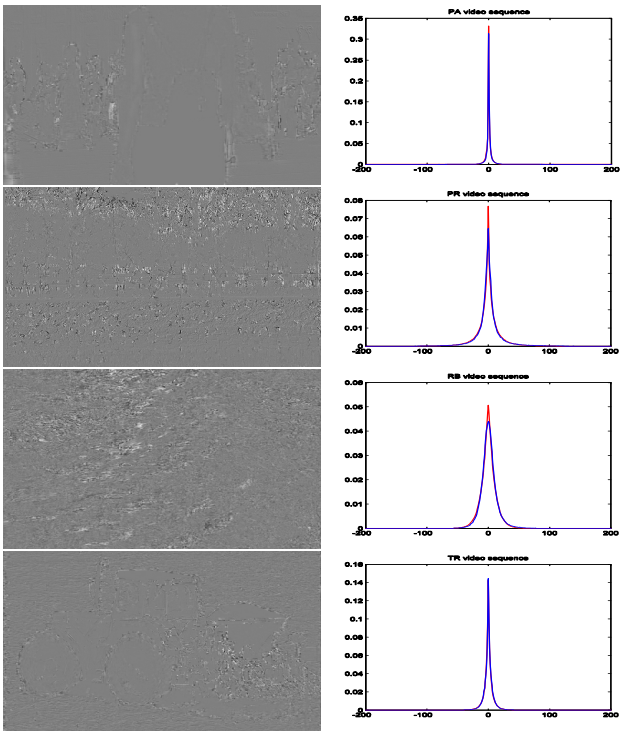

Figure. 3. Left: $\boldsymbol{PE(11)}$ obtained from each original video sequence; right: the corresponding histogram (blue line), and the fitted GGD curve (red line).

2.3. Quality analysis

At the receiver side, the transmitted RR features of the original sequence need to be firstly decoded. Then by comparing the RR feature difference between the original and the distorted video sequences, the perceptual quality of the distorted video sequence can be analyzed. As the RR features are extracted from both spatial and temporal perspectives, the RR feature comparison should also be performed from both perspectives. For the spatial SIC, not only the difference is computed, but also a mutual masking property is considered to depict the spatial distortion:

$$SD = |SIC_{dis} - SIC_{ori}| \tag{6}$$

where SD captures the spatial distortions related to the spatial energy variation, which is caused by quantization process. As discussed before, quantization process will discard more E and HF information than the DC and LF information. Therefore, according to Eq. (6), the larger the SD value, the more the quantization distortion is introduced. However, as aforementioned, different spatial contents will present different HVS masking properties. The HVS sensitivity to error is generally highest in plain regions and decreases in the order of plain, edge, and texture. Therefore, in this paper, in order to model the HVS masking property, a novel mutual masking strategy is employed:

$$SD_v = \begin{cases} 1 - \frac{SIC_{ori}}{SIC_{dis}}, & SIC_{ori} < SIC_{dis} \\ 1 - \frac{SIC_{dis}}{SIC_{ori}}, & SIC_{ori} \geq SIC_{dis} \end{cases}. \tag{7}$$

where SD_v is the final HVS-related features to depict the spatial energy variation. SD in the denominator is employed to scale SD_v into the range [0,1]. When a frame containing texture information is smoothed by the distortions, such as the compression, the detailed texture information cannot be perceived by the HVS. Therefore, no visual masking effect should occur. Also if a smooth frame is distorted to be highly textured by the distortion, such as ringing and blocking artifacts, only the noise can be perceived from the degraded frame. In this case, there should be no visual masking effect either. This phenomenon is named as the mutual masking [33]. In [32], the mutual masking effect is determined by the minimum value of the thresholds calculated from the original and distorted image. In this paper, Eq. (7) is employed to model the mutual masking effect of the HVS perception, where the smaller value of SIC_{ori} and SIC_{dis} is employed to model the masking effect. In this way, only the image is highly textured in both the reference and distorted images (large SIC_{ori} and SIC_{dis} values) can produce a significant masking effect.

For the temporal distance, what needs to be computed is the histogram distance TD between the original and distorted video:

$$TD = \sum_{i=1}^{h_L} |p(i) - p_d(i)|, \tag{8}$$

where the histogram p is constructed from the original prediction residual. The histogram p_d is obtained by revisiting same motion vectors derived for the original sequence. Actually, for compressed video sequences, the decoder will generate the same motion vectors for reconstruction. In this way, p_d can be easily constructed without introducing much overhead computation. However, the histogram p is not available at the receiver side. Therefore, the GGD model introduced at the sender side is employed to approximate the temporal distance TD_v:

$$TD_v = \sum_{i=1}^{h_L} |p(i) - p_{\alpha,\beta}(i)| - \sum_{i=1}^{h_L} |p_{\alpha,\beta}(i) - p_d(i)|. \tag{9}$$

$\sum_{i=1}^{h_L} |p_{\alpha,\beta}(i) - p_d(i)|$ can be easily computed at the receiver side, while $\sum_{i=1}^{h_L} |p(i) - p_{\alpha,\beta}(i)|$ can be decoded from the transmitted RR features.

As the feature distances from the spatial and temporal perspectives have been computed, how to combine these values together needs to be further considered. In this paper, in order to balance the contributions of the spatial and temporal features, a simply multiplication process is employed:

$$Q_f = SD_v \times log_{10}\left(1 + \frac{TD_v}{c}\right). \tag{10}$$

Q_f is the quality score for each video frame. By combining Q_f values of the video frames, the final video quality index (VQI) for each sequence can be generated. It is well known that HVS is more sensitive to bad experiences of the visual signals. It means that the HVS perception will be very easily biased by the frames of the worst qualities during viewing a video sequence. Therefore, in this paper, only the top K worst video frames rather than all the frames are employed to generate the VQI. For simplicity, K is experimentally set as 17% of the total number of video frames per second.

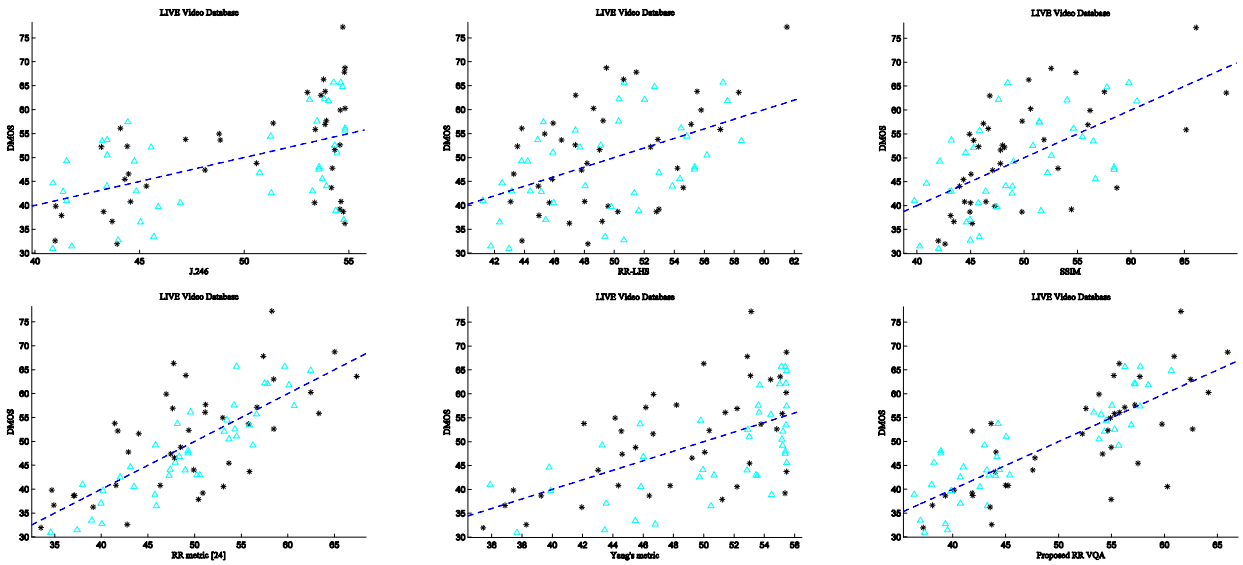

Figure. 4. Scatter plots of the DMOS values versus model predictions on the LIVE video quality database. The star indicates H.264 encoded video sequence, while the triangle indicates the MPEG-2 compressed video sequence. Top row from left to right: J.246, RR-LHS, and SSIM; bottom row from left to right: RR metric [24], Yang's RR VQA, and proposed RR VQA.

3. EXPERIMENTAL RESULTS

In this section, the effectiveness of the proposed RR VQA is demonstrated by comparisons with other representative VQAs. LIVE video database [6] is employed to illustrate the performances of different VQAs. The performances are compared by computing the statistical relationships between the subjective rating values, specifically the difference mean opinion score (DMOS) values, and the VQA outputs. The procedure introduced in [8] [34] is followed to evaluate the VQA's performance. First, a logarithmic function is employed to fit the objective and subjective scores through a nonlinear mapping process. Subsequently, the linear correlation coefficients (LCC) which provides an evaluation of the prediction accuracy, the Spearman rank-order correlation coefficients (SROCC) which measures the prediction monotonicity, root mean square prediction error (RMSE) of the fitting procedure are utilized to measure the VQA's effectiveness. Larger LCC and SROCC values mean that the objective and subjective scores correlate better, which generates a better performance. On the other hand, the smaller RMSE value provides a better performance.

The performance of the proposed RR VQA is compared with other VQAs, including the FR VQA PSNR and SSIM [10], and recently developed RR VQAs J.246 [30], Yang's RR VQA [35], RR-LHS [31], and VQM [21]. Detailed information is illustrated in Table I. It can be observed that PSNR performs the worst, even though it is the FR VQA and requires the whole video sequence for quality evaluation. The reason is that PSNR only compares the pixel-level absolute difference, which does not account for any HVS and signal statistical properties. The corresponding perceptual quality cannot be accurately depicted. SSIM has been demonstrated to be very effective for evaluating image perceptual quality. However, for the video sequences, the performance is not good enough, compared with other VQAs. There may be two reasons. First, the video sequence is encoded based 8×8 DCT block. SSIM computes the mean, variance, and co-variance values based on overlapped blocks. The distortions by the quantization process cannot be clearly indicated. Second, the temporal distortions are not considered in SSIM, which is proved to be crucial for evaluating the perceptual quality of the video sequence.

Table I. Performances of different VQAs over LIVE video database (MPEG-2 and H.264 encoded sequences)

	LCC	SROCC	RMSE	Reference type	Data rate (25fps)
PSNR	0.4488	0.4157	9.188	FR	-
SSIM [10]	0.5946	0.5969	8.267	FR	-
J.246 [30]	0.5036	0.4460	8.883	RR	10 kbps
Yang's RR VQA [35]	0.5654	0.5366	8.484	RR	0.2 kbps
RR-LHS [31]	0.4557	0.4082	9.152	RR	64 kbps
VQM [21]	0.7003	0.6790	7.340	RR	150 kbps
RR metric [24]	0.7567	0.7486	6.722	RR	0.875 kbps
Proposed RR VQA	**0.7945**	**0.7856**	**6.244**	**RR**	**0.875 kbps**

For recently developed RR VQAs, Yang's RR VQA employs the DCT coefficient ratio to measure the video quality, which results in a very small RR data rate (only 0.2 kbps). However, the performance is not pleasing, as the temporal information is not considered. For J.246, the edge pixels are extracted for quality comparison, which generates higher RR data rate while the performance is not good enough. For RR-LHS, the harmonic and discriminative analysis is employed to depict the blocking and blur artifacts. The temporal motion information is employed to finally correct the quality values. The poor performance is due to that the temporal information is not accurately modeled. The VQM [21] is derived by recording several features which depict the spatial and temporal distortions. The per-

formance has been significantly improved. However, the RR data rate is of huge burden for transmission (about 150 kbps after feature compression process). The RR metric [24] well balanced the performance and the RR data rate, with higher LCC and SROCC values and a smaller RR data rate.

For the proposed method, it can be observed that the proposed method outperforms the FR and RR VQAs. The large improvement attributes to the following reasons. First, the SIC feature can accurately depict the distortion introduced by the compression process. Second, a mutual masking strategy is employed to model the HVS texture masking property. In this way, the HVS property can be accurately modeled and employed for quality analysis. Third, BME is employed to depict temporal relationships between adjacent frames. With this consideration, motion prediction residual can be more accurately modeled by GGD, which generates an accurate temporal statistical property modeling. The scatter-plots of different VQAs over the LIVE video quality database are illustrated in Figure. 4. It can be observed that for the 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.

4. CONCLUSION

In this paper, a novel RR VQA is proposed by depicting the spatial and temporal distortions, respectively. From the spatial perspective, the extracted RR feature SIC not only measures the spatial distortions but also simulates the HVS texture masking property in a mutual way. From the temporal perspective, BME helps to accurately model the temporal relationships, resulting in more effective temporal statistical property characterization. Experimental results demonstrate the proposed RR VQA can produce better performances with a lower RR data rate.

5. REFERENCES

[1] Z. Wang, *et al.*, "Objective video quality assessment," *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] H. R. Sheikh, *et al.*, "LIVE image quality assessment database release 2," http://live.ece.utexas.edu/research/quality.

[3] P. Le Callet, *et al.*, "Subjective quality assessment IRCCyN/IVC database," http://www.irccyn.ec-nantes.fr/ivcdb/.

[4] N. Ponomarenko, *et al.*, "TID2008 - a database for evaluation of full-reference visual quality assessment metrics," *Advances of Modern Radioelectronics*, vol. 10, pp. 30-45, 2009.

[5] L. Ma, *et al.*, "Image retargeting quality assessment: a study of subjective scores and objective metrics," *JSTSP*, vol. 6, no. 6, pp. 626-639, Oct. 2012.

[6] K. Seshadrinathan, *et al.*, "LIVE video quality database," http://live.ece.utexas.edu/research/quality/livevideo.html.

[7] F. Zhang, *et al.*, "IVP subjective quality video database," http://ivp.ee.cuhk.edu.hk/research/database/subjective/.

[8] VQEG Final Report from the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment 2000 [Online]. Available: http://www.vqeg.org.

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

[10] Z. Wang, *et al.*, "Image quality assessment: from error visibility to structural similarity," *TIP*, vol. 13, no. 4, pp. 600-612, Apr. 2004.

[11] L. Ma, *et al.*, "Visual horizontal effect for image quality assessment," *SPL*, vol. 17, no. 7, pp. 627-630, Jul. 2010.

[12] L. Zhang, *et al.*, "FSIM: a feature similarity index for image quality assessment," *TIP*, vol. 20, no. 8, pp. 2378-2386, 2011.

[13] S. S. Hemami, *et al.*, "No-reference image and video quality estimation: applications and human-motivated design," *SPIC*, vol. 25, no. 7, pp. 469-481, Aug. 2010.

[14] M. A. Saad, *et al.*, "A perceptual DCT statistics based blind image quality metric," *SPL*, vol. 17, pp. 583-586, 2010.

[15] A. K. Moorthy, *et. al*, "Blind image quality assessment: from natural scene statistics to perceptual quality," *TIP*, pp. 3350-3364, 2011.

[16] Q. Li, *et al.*, "Reduced-reference image quality assessment using divisive normalization-based image representation", *JSTSP*, vol. 3, no. 2, pp. 202-211, Apr. 2009.

[17] Z. Wang, *et al.*, "Reduced-and no-reference image quality assessment: the natural scene statistic model approach," *SPM*, vol. 28, Nov. 2011.

[18] S. Wolf, *et al.*, "Low bandwidth reduced reference video quality monitoring system," *QoMEX*, Jan. 2005.

[19] M. Tagliasacchi, *et. al*, "A reduced-reference structural similarity approximation for videos corrupted by channel errors," *Multimedia Tools Appl.*, vol. 48, no. 3, pp. 471-492, Jul. 2010.

[20] P. Le Callet, *et al.*, "A convolutional neural network approach for objective video quality assessment," *TNN*, vol. 17, no. 5, pp. 1316-1327, May. 2006.

[21] M. H. Pinson, *et al.*, "A new standardized method for objectively measuring video quality'" *IEEE Trans. Broadcasting*, vol. 50, pp. 312-322, Sept. 2004.

[22] Z. Wang, *et al.*, "Quality-aware images," *TIP*, vol. 15, no. 6, pp. 1680-1689, Jun. 2006.

[23] L. Ma, *et al.*, "Reduced-reference image quality assessment using reorganized DCT-based image representation," *TMM*, vol. 13, no. 4, pp. 824-829, Aug. 2011.

[24] L. Ma, *et al.*, "Reduced-reference video quality assessment of compressed video sequences," *TCSVT*, vol. 22, no. 10, pp. 1441-1456, Oct. 2012.

[25] L. Ma, *et al.*, "Reduced-reference image quality assessment in reorganized DCT domain," *SPIC*, accepted.

[26] J. A. Redi, *et al.*, "Color distribution information for reduced-reference assessment of perceived image quality," *TCSVT*, vol. 20, pp. 1757-1769, Dec. 2010.

[27] Henry H. Y. Tong, *et al.*, "A perceptual model for JPEG applications based on block classification, texture masking, and luminance masking", *ICIP*, 1998.

[28] L. Ma, *et al.*, "Adaptive block-size transform based just-noticeable difference model for images/videos", *SPIC*， vol. 26, no. 3, pp. 162-174, Mar. 2011.

[29] L. Ma, *et al.*, "Motion trajectory based visual saliency for video quality assessment", *ICIP*, 2011.

[30] ITU-T Recommendation J.246, "Perceptual visual quality measurement techniques for multimedia services over digital cable television networks in the presence of a reduced bandwidth reference", Aug. 2008. Available [Online]: http://www.itu.int/rec/T-REC-J.246/en.

[31] I. P. Gunawan, *et al.*, "Reduced-reference video quality assessment using discriminative local harmonic strength with motion consideration", *TCSVT*, vol. 18, no. 1, pp. 71-83, Jan. 2010.

[32] A. P. Bradley, "A wavelet visible difference predictor", *TIP* vol. 8, no. 5, pp. 717-730, May 1999.

[33] S. Daly, "The visible difference predictor: an algorithm for the assessment of image fidelity," in *Digital Images and Human Vision*, A. B. Watson, Ed. Cambridge, MA: MIT Press, pp. 179-206, 1993.

[34] H. R. Sheikh *et al.*, "A statistical evaluation of recent full reference image quality assessment algorithms," *TIP*, vol. 15, no. 11, pp. 3441–3452, Nov. 2006.

[35] 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.