A Novel Objective No-Reference Metric for Digital Video Quality Assessment

Fuzheng Yang, Shuai Wan, Yilin Chang, and Hong Ren Wu

***Abstract*—A novel objective no-reference metric is proposed for video quality assessment of digitally coded videos containing natural scenes. Taking account of the temporal dependency between adjacent images of the videos and characteristics of the human visual system, the spatial distortion of an image is predicted using the differences between the corresponding translational regions of high spatial complexity in two adjacent images, which are weighted according to temporal activities of the video. The overall video quality is measured by pooling the spatial distortions of all images in the video. Experiments using reconstructed video sequences indicate that the objective scores obtained by the proposed metric agree well with the subjective assessment scores.**

***Index Terms*—Digital video quality assessment (VQA), human visual system (HVS), no-reference (NR) metrics.**

I. Introduction

ALONG with the rapid development of wireless and internet protocol (IP) video technologies, there has been growing emphasis placed on real-time assessment of digital video quality for various visual communications services [1], [2]. Full-reference objective video quality assessment (VQA) methods have been widely used, such as the peak signal-to-noise ratio (PSNR), which requires both the reference and the compressed/transmitted videos. However, in the absence of the original or reference videos, they are not practical for picture quality assessment in real-time video transmission applications. In these cases, an objective no-reference (NR) metric for VQA is required, which does not require the reference video. Research on NR metrics has recently attracted a great deal of attention, and the video quality experts group (VQEG) is working toward the standardization of NR metrics.

It is very difficult to design an objective NR video quality metric, which is mainly due to our limited understanding of the human visual system (HVS) and cognitive aspects of the brain. Only a few methods have been proposed for objective NR quality assessment. An effective approach proposed is to extract certain features (e.g., blurring and blocking artifacts) that reflect the video quality to a certain degree [3]–[6]. In [3] and [4], the magnitude of extracted blocking artifacts was used to assess the video quality. In [6], an NR blurring distortion metric was presented for video. An NR MPEG-2 video quality rating method was proposed in [7], which predicted the PSNR based on the quantization parameters obtained from the video stream. However, the applications of those methods have inherent limitations.

In this contribution, an objective NR metric is proposed for VQA. The spatial distortion of each image in a video is predicted using the differences between the corresponding regions of two adjacent frames in the video sequence. The spatial distortion is then weighted according to temporal activities of the video. The video quality is measured by pooling the spatial distortions of all images in the sequence. Section II describes the proposed video quality metric. Experimental results are presented in Section III, which is followed by concluding remarks.

II. NR Video Quality Metric

It is well known that human observers are able to assess the quality of distorted images without reference images. In the assessment process, a comparison is naturally made between the distorted images and the information and knowledge about natural images that have been acquired by observers (i.e., assessors) and stored in their brain. In other words, any assessment by a human observer without given reference material (in this case, reference images) is not really an NR process, where the observer does use information or relevant knowledge "prestored" in his/her memory. Therefore, it is necessary to find out some available reference information in order to design an objective NR assessment metric. In this letter, the previous image is used as the reference of the current image in the video based on the temporal dependency between adjacent images.

Furthermore, for the HVS to perceive and to comprehend a video sequence, the video content is supposed to be consistent in several successive images. A rapid change of the video content between adjacent images may lead to poor subjective picture quality, i.e., when the change is too sharp, the video content is not perceivable to the HVS anymore. Equivalently, abrupt changes induced by coding distortions can impair the subjective video quality. Therefore, the differences between adjacent images can be utilized to evaluate the video quality. Since the spatial contrast sensitivity of the HVS is low where the video content is with rapid speed, temporal activities of video must be considered in VQA. Our experiments have also shown that the differences between the corresponding translational regions of high spatial complexity in adjacent images reflect the perceptual distortion of the regions well and can be utilized in the NR video quality assessment.

Manuscript received December 19, 2004; revised April 20, 2005. This work was supported in part by the foundation of HuaWei Technology Limited Cooperation under Grant YJCB2003017MU. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Dipti Prasad Mukherjee.

F. Yang, S. Wan, and Y. Chang are with National Key Laboratory of ISN, Xidian University, Xi'an 710071, China (e-mail: y_fuzheng@yahoo.com.cn; shwan@mail.xidian.edu.cn; ylchang@xidian.edu.cn).

H. R. Wu is with School of Electrical and Computer Engineering, Royal Melbourne Institute of Technology, Melbourne, Victoria 3001, Australia (e-mail: henry.wu@rmit.edu.au).

Digital Object Identifier 10.1109/LSP.2005.855553

In the proposed method, the motion vector (MV) field of each image in the sequence is estimated to determine the translational regions. The spatial distortion of each image is then predicted using the differences between the translational regions of high spatial complexity in adjacent reconstructed images. Finally, the quality of the entire video is obtained by pooling the spatial distortions of all images in the sequence. The detailed method for estimating the spatial distortion is presented as follows.

A. MV Estimation

In this letter, the block matching algorithm is used for MV estimation. Since the noise in the reconstructed images influences the precision of motion search, the distorted images are smoothed with the two-dimensional Gaussian filter for noise removal before the MV estimation

$$f'_n(x,y) = f_n(x,y) * G(x,y), \quad 0 < x \le W, \quad 0 < y \le H \tag{1}$$

where $f_n(x,y)$ and $f'_n(x,y)$, respectively, represent pixels in the nth image in the video sequence and its filtered counterpart, and W stands for the width of the image and H the height. $G(x,y)$ is the two-dimensional Gaussian filter, defined by

$$G(x,y) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) \tag{2}$$

where σ is the standard deviation of the filter. The filtered image f'_n is used for both the motion estimation and the spatial distortion computation in Section II-C.

Let (a,b) be the geometrical center of the $(2N_1+1)\times(2N_1+1)$-pixels Block B in the nth frame. Search the MV of Block B in the $(n-1)$th frame using block matching algorithm, and the obtained MV is considered as the MV of the pixel at (a,b). Let $mv_{x,n}(x,y)$ and $mv_{y,n}(x,y)$ be the horizontal and the vertical components of the motion vector at (x,y) in the nth frame, respectively. By MV estimation of every pixel, the motion vector field $F = \{(mv_{x,n}(x,y), mv_{y,n}(x,y)) | 1 \le x \le W, 1 \le y \le H\}$ of the nth frame can be obtained.

B. Determination of Translational Regions of High Spatial Complexity

The regions with translational motion in the images can be determined by the consistency in motion vectors of neighboring pixels according to the obtained motion vector field. For the region R of $(2N_2+1)\times(2N_2+1)$-pixels centered at (c,d), the consistency of the motion vectors is estimated by the variance $\sigma^2_{mv,n}(c,d)$ of the MV values.

For the region of $(2N_3+1)\times(2N_3+1)$-pixels centered at (c,d), the spatial complexity is estimated by its variance $\sigma^2_{f,n}(c,d)$ of the pixel values.

Given thresholds T_1 and T_2, the set I_n for the nth frame can be defined as follows:

$$I_n = \{(x,y) | (\sigma^2_{mv,n}(x,y) < T_1) \cap (\sigma^2_{f,n}(x,y) > T_2) \cap \left(\overline{mv_{x,n}(x,y)} \ne 0\right) \cap \left(\overline{mv_{y,n}(x,y)} \ne 0\right)\} \tag{3}$$

where $\overline{mv_{x,n}(x,y)}$ and $\overline{mv_{y,n}(x,y)}$ represent, respectively, the mean value of the horizontal and the vertical components of the motion vectors in the region R.

All the pixels that satisfy (3) are considered to belong to translational regions of high spatial complexity.

C. Spatial Distortion of Image

For a pixel $(a,b) \in I_n$, according to its motion vector $(mv_{x,n}(a,b), mv_{y,n}(a,b))$, find the region in the $(n-1)$th frame that corresponds to the region of $(2N_4+1)\times(2N_4+1)$-pixels centered at (a,b) in the nth frame. Then compute the following two sums of squared differences, respectively:

$$D_n(a,b) = \sum_{k=-N_4}^{N_4}\sum_{l=-N_4}^{N_4}[f_n(a+k,b+l) - f_{n-1}(a+k+mv_{x,n}(a,b) \; b+l+mv_{y,n}(a,b))]^2 \tag{4}$$

$$D_n(a,b) = \sum_{k=-N_4}^{N_4}\sum_{l=-N_4}^{N_4}[f'_n(a+k,b+l) - f'_{n-1}(a+k+mv_{x,n}(a,b) \; b+l+mv_{y,n}(a,b))]^2 \tag{5}$$

where f'_n is the filtered image obtained by (1). Compute $D_n(x,y)$ and $D'_n(x,y)$ $((x,y) \in I_n)$, and they are averaged to obtain $\overline{D}_n$ and $\overline{D'}_n$, respectively.

Filters are used in some video compression and postprocessing algorithms to improve the quality of compressed video. Different filtering algorithms may have different influence on the reconstructed video, thus affecting $\overline{D}_n$. The employed Gaussian filter can smooth the distorted video and impair the above-mentioned effect. Therefore, we determine the spatial distortion of the image mainly using $\overline{D'}_n$. The distortion is modified by the value of $\overline{D}_n - \overline{D'}_n$, which reflects the influence of the filters used in video compression or postprocessing. Using $\overline{D}_n$ and $\overline{D'}_n$, we define the spatial distortion of the image d_n as

$$d_n = \overline{D'}_n \cdot \left(\alpha - \frac{(\overline{D}_n - \overline{D'}_n)}{\overline{D'}_n}\right) \tag{6}$$

where α is a parameter determined by experiments.

D. Influence of Temporal Activity

Since the human visual system can tolerate the distortions in fast-moving regions to a considerable extent, different weightings are applied to the measured spatial distortions of the image according to temporal activities of the video, which are computed as the mean value of the motion vectors in the image. The temporal activity of the nth frame A_n is defined as

$$A_n = \overline{|mv_{x,n}(x,y)|} + \overline{|mv_{y,n}(x,y)|} \quad 1 \le x \le W, \; 1 \le y \le H. \tag{7}$$

The weighted spatial distortion is then defined as

$$d'_n = \frac{d_n}{\left(\beta + \frac{(\max(A_n,\gamma))^2}{\delta}\right)} \tag{8}$$

where d'_n is the final spatial distortion of the nth frame, and β, γ, and δ are constants determined by experiments.

The video quality metric is obtained by pooling and averaging the weighted spatial distortions calculated using (8) of the images over the entire video sequence.

III. Experimental Results

In our experiments, 144 (i.e., 9×16) distorted video sequences have been used, which are from the VQEG Phase I FR-TV test, consisting of source sequences (SRCs) 2–10 undergone hypothetical reference circuits (HRCs) 1–16 [8]. The single stimulus continuous quality evaluation (SSCQE) [9] is used in the subjective test. The guidelines specified by the VQEG in [10] are followed for quality tests, involving 25 nonexpert viewers. The viewers (12 males and 13 females) have evaluated the video quality in real time using a slider device and a continuous grading scale marked with "Excellent," "Good," "Fair," "Poor," and "Bad." The subjective scores are quantized on a scale of [0...100]. The analysis of subjective test results shows that the standard deviation of our subjective scores is 3.6, which may be caused by varying delays in response time by different viewers.

In order to compare our SSCQE test data with the subjective data published by the VQEG using the double stimulus continuous quality scale (DSCQS) [8], the continuous quality scores are averaged to obtain a quality score by each viewer for each sequence. The mean opinion score (MOS) S_i of the ith video sequence is obtained by averaging its subjective scores by different viewers (as shown in Fig. 2), where the standard deviation is 2.5. The comparison between the DSCQS scores and SSCQE scores shows that they correlate well for the sequences from the same reference video. However, for the sequences from different reference videos, there are different offsets between their DSCQS and SSCQE scores, which are dependent on the video content.

The mean opinion scores are then linearly scaled to a nominal range of [0, 1], where "0" and "1" represent the best and the worst ratings, respectively. The normalized (or scaled) MOS $\overline{S}_i$ is defined as

$$\overline{S}_i = \frac{S_i - S_{\text{best}}}{S_{\text{worst}} - S_{\text{best}}} \tag{9}$$

where S_{best} and S_{worst} are the best rating and the worst rating of all the scores, respectively.

The adjacent images tend to have similar spatial distortions in a sequence compressed with the same algorithm. To reduce the computational complexity, M neighboring frames are grouped together in natural order, and one image is randomly selected from each group. The spatial distortions of the selected images are computed using the proposed method. We define the quality of the ith video sequence as $S_{p,i}$, and its value is obtained by calculating the mean value of the spatial distortions of the selected images from the sequence. The proposed quality assessment method is used to evaluate all of the distorted sequences. The model parameters are determined by the training experiments. The distorted video sequences from SRC18-SRC21 under HRC1-HRC16 in the VQEG Phase I FR-TV test [8] are used for training purposes. The parameters are adjusted so that the objective scores obtained using the proposed method coincide with the SSCQE scores. The resulting values are $N_1 = 8$, $N_2 = N_3 = N_4 = 4$, $M = 5$, $T_1 = 5$, $T_2 = 500$, $\sigma = 0.8$, $\alpha = 2.5$, $\beta = 2.5$, $\gamma = 5$, and $\delta = 30$. Thresholds T_1 and T_2 are selected to determine the appropriate translational regions of high spatial complexity. They are trained using the distorted sequences from SRC18-SRC21 under HRC4. α and σ are selected to ensure that the proposed method is applicable to video sequences compressed with different algorithms. They are trained utilizing the distorted sequences from SRC21 under HRC1-HRC16. β, γ, and δ are adjusted to adapt to video sequences with different motion activities. They are mainly determined by the distorted sequences from SRC18-SRC21 under the HRC6.

TABLE I
Performance Comparison of Objective Data and Subjective Data

SRC Exclusion Sets	RMSE	PCC	SCC	OR
SRC8 and SRC3	0.1218	0.8496	0.7968	0.6429
SRC8	0.1860	0.7673	0.7409	0.6719
None	0.2196	0.6512	0.6430	0.8531

The following linear polynomial fit is used for our objective quality measurements to fit the scaled subjective scores [10]:

$$\overline{S_{p,i}} = p_1 \cdot S_{p,i} + p_2 \tag{10}$$

where $\overline{S_{p,i}}$ is the scaled objective score, and p_1 and p_2 are the fitting parameters. In our experiments, $p_1 = 0.003\,397$ and $p_2 = -0.065\,45$.

A number of performance evaluation metrics suggested by the VQEG are used to evaluate the performance of the proposed video quality metric (VQM) [10]. The root-mean-squared error (RMSE), Pearson correlation coefficient (PCC), Spearman rank order correlation coefficient (SCC), and outlier ratio (OR) are computed between the fitted objective data and the corresponding subjective data [10]. The results are shown in Table I, where the SRC exclusion sets denote the excluded SRC subsets. From Table I, we observe that the objective scores obtained through the proposed method are consistent with the subjective assessment for video sequences containing natural scenes (SRC8 and SRC3 excluded in our experiments). However, the effectiveness of the proposed VQM degrades significantly when applied to SRC8 (Sequence Horizontal scrolling 2) and SRC3 (Sequence Harp). The reasons lie in that SRC8 contains an unnatural scene; in SRC3, zooming decreases the accuracy of motion vector estimation between adjacent frames.

For the distorted sequences except those from SRC8 and SRC3, the scatter plot of the objective scores obtained by the proposed method and the scaled subjective scores is shown in Fig. 1, and the scaled objective scores with the scaled subjective scores are shown in Fig. 2, where the order of the distorted sequences is from SRC2 under HRC1-HRC16 to SRC10 under HRC1-HRC16. We observe that the two traces in Fig. 2 are quite close. In addition, because the temporal activity is considered, satisfactory results can be obtained for both the

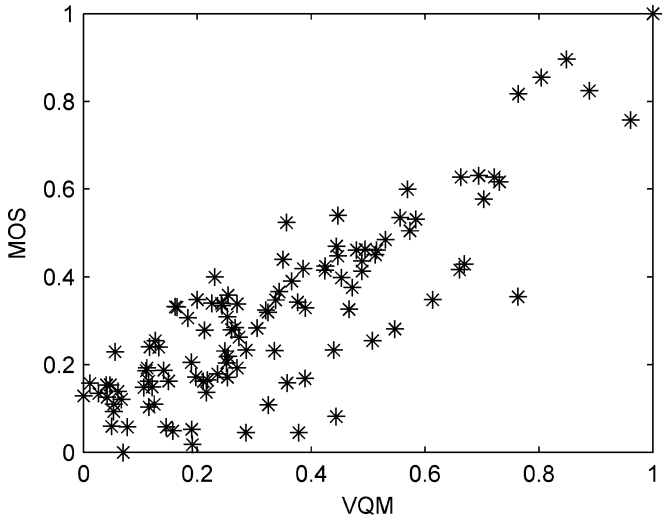

Fig. 1. Scatter plot of MOS versus objective prediction.

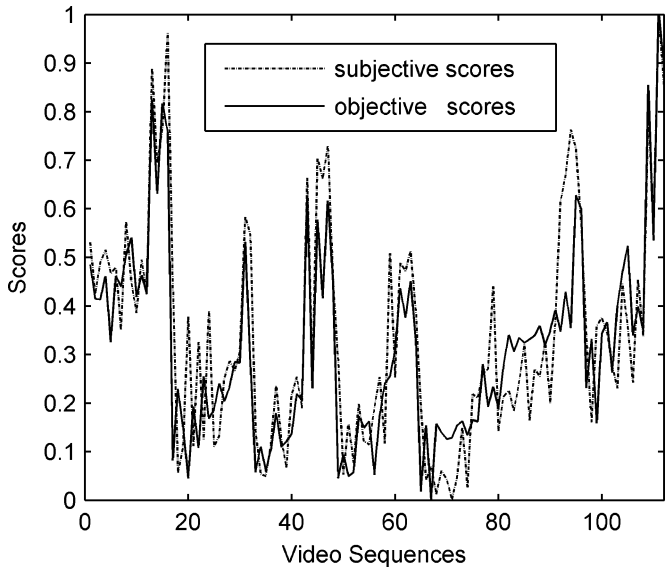

Fig. 2. Scaled subjective scores and objective scores.

low-motion sequences (e.g., SRC10) and the high-motion ones (e.g., SRC9).

It should be noted that when the video content of adjacent images in a sequence changes too much, for instance, during a scene change, few translational regions can be found. Under such circumstances, the calculated quality of the whole video is mainly the quality of those parts with steady contents. The fast-changing segments of the video and their effects are omitted in computation of the metric, which is reasonable when taking into account the temporal masking characteristic of the HVS.

The MV estimation constitutes most of the computational burden in the proposed quality metric. Therefore, the computational complexity can be significantly reduced by the following methods. The MVs of adjacent pixels are generally similar so that the MV of the pixel at (a, b) can be considered as the MVs of its surrounding pixels. In addition, we can first calculate the spatial complexity of all regions of an image and only perform the MV estimation for pixels that are classified as of high spatial complexity.

IV. Conclusion

We proposed a novel objective NR metric for digital video quality assessment based on the temporal dependency of video images, taking into account characteristics of the HVS. This method does not require any reference and is independent of the video compression algorithm. Therefore, the proposed VQM is suitable for the NR video assessment. Experimental results indicate that the objective scores obtained by the proposed method correlate well with the subjective assessment scores for reconstructed natural videos. Since computational complexity can be an issue when using NR models, we have also discussed how to improve the computation efficiency for the proposed method. Future studies include error detection techniques based on temporal dependency of video images and picture quality evaluation for video communications over error-prone networks.

References

[1] S. Winkler and F. Dufaux, "Video quality evaluation for mobile applications," in *Proc. SPIE/IS&T VCIP*, vol. 5150, Lugano, Switzerland, Jul. 8–11, 2003, pp. 593–603.

[2] S. Winkler and R. Campos, "Video quality evaluation for Internet streaming applications," in *Proc. SPIE/IS&T Human Vision Electronic Imaging*, vol. 5007, Santa Clara, CA, Jan. 20–24, 2003, pp. 104–115.

[3] H. R. Wu and M. Yuen, "A generalized block-edge impairment metric for video coding," *IEEE Signal Process. Lett.*, vol. 4, no. 11, pp. 317–320, Nov. 1997.

[4] Z. Wang, A. C. Bovik, and B. L. Evans, "Blind measurement of blocking artifacts in images," in *Proc. ICIP*, vol. 3, Sep. 2000, pp. 981–984.

[5] J. Caviedes and F. Oberti, "No-reference quality metric for degraded and enhanced video," in *Proc. SPIE VCIP*, vol. 5150, Lugano, Switzerland, Jul. 2003, pp. 621–632.

[6] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi, "A no-reference perceptual blur metric," in *Proc. ICIP*, vol. 3, Rochester, NY, Sep. 22–25, 2002, pp. 57–60.

[7] M. Knee. The Picture Appraisal Rating (PAR) a single-ended picture quality measure for MPEG-2. presented at *Proc. Int. Broadcast. Conv.*. [Online]. Available: http://www.snellwilcox.com/products/mosalina/content/downloads/parpaper.pdf

[8] VQEG Report. (2000, Mar.) Final Report From the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment. [Online]. Available: http://www-ext.crc.ca/vqeg/frames.html

[9] ITU-T Recommendation BT.500-10, Methodology for the Subjective Assessment of the Quality of Television Pictures, 2000.

[10] VQEG Report. (2004, Jun.) RRNR-TV Group Test Plan, Draft Version 1.7. [Online]. Available: http://www-ext.crc.ca/vqeg/frames.html