Journal ohejiang University SCIENCE A
ISSN 1673-565X (Print); ISSN 1862-1775 (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

A video structural similarity quality metric based on a joint spatial-temporal visual attention model

Hua ZHANG, Xiang TIAN, Yao-wu CHEN‡

(*Institute of Advanced Digital Technology and Instrumentation, Zhejiang University, Hangzhou 310027, China*)

E-mail: emma_zhanghua@zju.edu.cn; xiang.t@163.com; cyw@mail.bme.zju.edu.cn

Received Jan. 14, 2009; Revision accepted Apr. 23, 2009; Crosschecked Oct. 18, 2009

Abstract: Objective video quality assessment plays a very important role in multimedia signal processing. Several extensions of the structural similarity (SSIM) index could not predict the quality of the video sequence effectively. In this paper we propose a structural similarity quality metric for videos based on a spatial-temporal visual attention model. This model acquires the motion attended region and the distortion attended region by computing the motion features and the distortion contrast. It mimics the visual attention shifting between the two attended regions and takes the burst of error into account by introducing the non-linear weighting functions to give a much higher weighting factor to the extremely damaged frames. The proposed metric based on the model renders the final object quality rating of the whole video sequence and is validated using the 50 Hz video sequences of Video Quality Experts Group Phase I test database.

Key words: Quality assessment, Structural similarity (SSIM) index, Attended region, Visual attention shift
doi:10.1631/jzus.A0920035 **Document code:** A **CLC number:** TN919.8

INTRODUCTION

Digital videos are subject to a wide variety of distortions during acquisition, compression, restoration, transmission, etc. Any of the distortions may result in a degradation of visual quality. For applications in which videos are ultimately to be viewed by human beings, the most reliable way of assessing the quality of the videos is subjective evaluation. The difference mean opinion score (DMOS), a subjective quality measurement obtained from a number of human observers, has been regarded for many years as the most reliable form of quality measurement (Sheikh *et al*., 2006). However, subjective evaluation is usually inconvenient, time-consuming and expensive. With the rapid increase in popularity of digital video applications, automatic evaluation of video quality through objective video quality metrics is becoming increasingly important.

Objective video quality metrics can be classified into full-reference (FR), no-reference (NR) and reduced-reference (RR) according to the availability of an original video sequence (Wang *et al*., 2003). The FR quality metrics are widely embedded into the video coding system to optimize the parameter settings and employed to benchmark the different video compression algorithms (Martinez-Rach *et al*., 2006). This paper focuses on FR video quality assessment.

The very powerful FR image quality metric named as the structural similarity (SSIM) index has been proposed in Wang *et al*.(2004a) based on the degradation of structural information. Furthermore, a simple and preliminary extension of the SSIM index was proposed for video quality assessment in Wang *et al*.(2004b) using a simple frame-by-frame implementation of the SSIM image quality metric. Wang and Simoncelli (2005) reported a complex wavelet SSIM (CWSSIM) index implemented in the wavelet domain. However, these metrics could not account for the temporal distortions in videos as a result of their utilizing only relative spatial information and their

‡ Corresponding author

ignorance of temporal information, which is the main difference from still to video images.

Recently, extensive research in different video processing domains has focused on the detection of attended regions according to the characteristics of the human visual system (HVS) (Chen *et al*., 2007; Tang, 2007) and the traditional SSIM index has been improved by the visual attention model. In Lu *et al*. (2005), the perceptual quality significance map (PQSM) was formed to adjust the weights of the SSIM index. Seshadrinathan and Bovik (2007) and Wang and Li (2007) both incorporated the motion model as spatial-temporal weighting factors into video quality assessments. Brooks *et al*.(2008) proposed a perceptually weighted multi-scale variant of CWSSIM. However, not only the motion but also other occurrences attract the observer's visual attention. For example, the visual attention may shift from a moving object to a suddenly appearing object, which may be a serious distortion caused by packet losses or bit errors during the transmission.

In this paper, we propose a perceptual structural similarity metric based on a spatial-temporal visual attention model for FR video quality assessment. The framework of this metric is shown in Fig.1. First, three coefficient maps named as the motion intensity, the motion spatial coherence (defined as the entropy of the motion directions), and the contrast coherence of the motion intensity are jointly considered to produce the motion attended region. Secondly, the distortion attended region is detected by the distortion contrast coherence coefficient map, which includes the explicit distortion in the frame. Thirdly, an attention shift mechanism based on the two attended regions is applied to integrate the four coefficient maps into the visual attention saliency map. The frame-level quality is calculated based on this saliency map and the local structural similarity (which is measured by the SSIM index). Finally, the overall quality of the video sequence is weighted frame by frame. We take the burst of error into account by introducing non-linear weighting functions to give a much higher weighting factor to the severely damaged frames.

STRUCTURAL SIMILARITY INDEX MAP

As described in Wang *et al*.(2004b), the SSIM video image quality assessment is a simple frame-by-frame implementation of the SSIM still image quality assessment. Hence, its basic philosophy is identical to that of the SSIM still image quality assessment. The motivation behind the structural similarity approach for measuring image quality is that the HVS is designed to detect the structural similarity instead of the error visibility (Wang *et al*., 2004a). The corresponding structural similarity index measure $\boldsymbol{S}(\boldsymbol{x}, \boldsymbol{y})$ is (Wang *et al*., 2004a; 2004b)

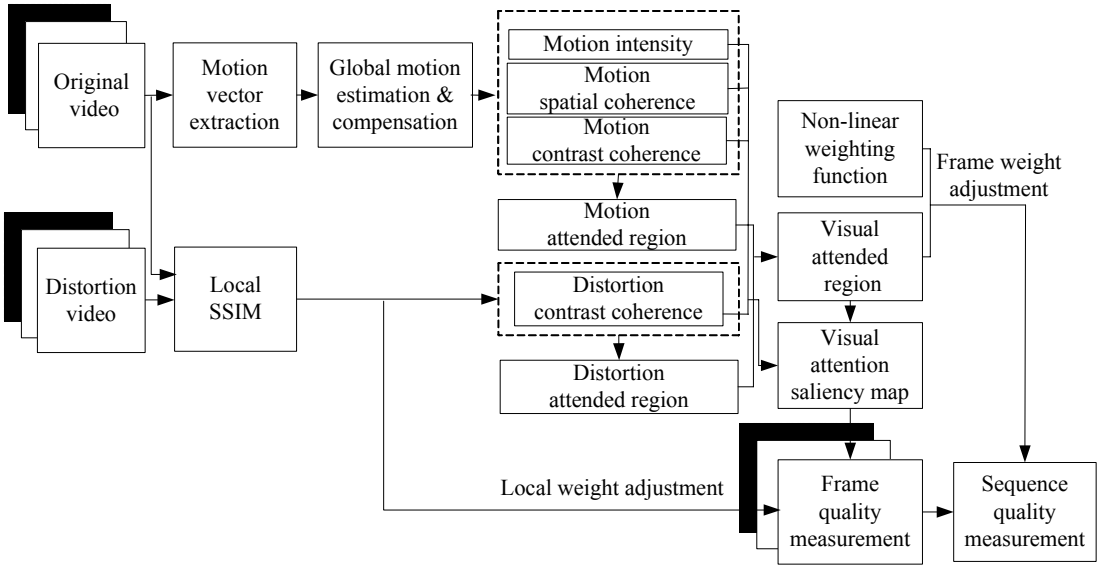

Fig.1 Framework of the proposed video structural similarity metric based on a spatial-temporal visual attention model

$$\boldsymbol{S}(\boldsymbol{x},\boldsymbol{y})=\frac{(2\mu_x\mu_y+C_1)(2\sigma_{xy}+C_2)}{(\mu_x^2+\mu_y^2+C_1)(\sigma_x^2+\sigma_y^2+C_2)}, \quad (1)$$

where $\boldsymbol{x}$ and $\boldsymbol{y}$ are the two images to be compared, and μ_x, μ_y, σ_x, σ_y, and σ_{xy} are the mean of $\boldsymbol{x}$, the mean of $\boldsymbol{y}$, the variance of $\boldsymbol{x}$, the variance of $\boldsymbol{y}$, and the covariance of $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively. C_1 and C_2 are two constants.

The SSIM index is applied for quality assessment of images using a sliding window approach. The window size is fixed to $N\times N$ (N=11) (Wang *et al*., 2004a). The SSIM is calculated within the sliding window, which moves pixel by pixel from the top-left to the bottom-right corner of the image. This results in an SSIM index map of the image, also considered as the quality map of the distorted image being evaluated. The overall quality value is defined as the average of the quality map.

PROPOSED VIDEO STRUCTURAL SIMILARITY METRIC

One strategy for the HVS processing incoming information serially is to pay almost all the visual attention to the object-selective region (Grill-Spector and Malach, 2004; Itti, 2005). It means if the visual attention was focused on some object in the scene, the remaining details in the environment would be unnoticed and, if some other emergent objects were more attractive, the visual attention would be shifted from the original object to the new scene. Motion and contrast are two salient features that attract visual attention in HVS. Motion always stimulates the visual attention (Grill-Spector and Malach, 2004). Many of the current visual attention models are based on motion features (Tang, 2007). Contrast, such as color contrast and luminance contrast, also attracts visual attention (Aziz and Mertsching, 2008). Distortion contrast is even more attractive, because the quality itself is a highly concerned feature during the video quality assessment; an observer's attention tends to be fixing on a region with a large distortion contrast, which reflects a relatively low quality. Thus, the distortion contrast should be regarded as another feature for visual attention models. Moreover, in a burst-of-error situation where most of the frames in a video sequence have high quality and only a few frames are of extremely low quality, a human observer tends to give a quality score lower than the average of all the frames, and the whole quality of the video is greatly reduced.

Motion attended region

Visual attention, which consists of static and dynamic attention, can be guided by the stimulus-driven (bottom-up) mechanism. Itti and Baldi (2005) showed that the temporal surprise is 20 times the spatial surprise. In other words, the dynamic changes are stronger predictors of human saccades compared with static features. As is well known, the dynamic attention is always caused by the motion of objects in a video sequence. In turn, the frame quality is mostly determined by the quality of the motion attended region. Therefore, the detection of the motion attended region is the crucial process in the proposed attention model.

The motion attended region is built based on motion features in the scene, such as motion intensity, motion spatial coherence and motion contrast coherence (Fig.1). Thus, the motion vectors extracted by the MATLAB optical flow block become the basis of this process. However, the video sequences are always captured with the camera panning or zooming, making the motion features not to be computed correctly by the original motion vectors. Thus, the global motion estimation and motion compensation are indispensable. Zheng (2008) proposed a fast global motion estimation method based on symmetrical elimination and the difference of motion vectors. A geometric global motion estimation model with four parameters is defined as follows:

$$f(\boldsymbol{z}\,|\,\boldsymbol{A},\boldsymbol{T})=\boldsymbol{A}\boldsymbol{z}+\boldsymbol{T}=\begin{bmatrix}a_1 & -a_2\\ a_2 & a_1\end{bmatrix}\begin{bmatrix}i\\ j\end{bmatrix}+\begin{bmatrix}t_{\mathrm{h}}\\ t_{\mathrm{v}}\end{bmatrix}, \quad (2)$$

where a_1 and a_2 are two parameters used for zooming and rotating, respectively, and t_{h} and t_{v} are used for horizontal and vertical shifting, respectively. After the compensation of this global motion model, the motion vectors are ready for computing the motion features.

Assume that a frame of a video sequence is divided into $K\times L$ blocks each with $M\times M$ (M=8), and let the motion vector of the block indexed (i, j) in the nth frame be $(\mathrm{MV_h}(i, j), \mathrm{MV_v}(i, j))$. The motion intensity of the nth frame becomes

$$I_n(i,j)=\sqrt{\mathrm{MV_h}(i,j)^2+\mathrm{MV_v}(i,j)^2}. \quad (3)$$

The global mean of motion intensities (M_n), the global variance of motion intensities (V_n), and the global mean of motion vectors (D_n) of the nth frame are as follows:

$$M_n=\frac{1}{K\times L}\sum_{i=1}^{K}\sum_{j=1}^{L}I_n(i,j), \quad (4)$$

$$V_n=\frac{1}{K\times L}\sum_{i=1}^{K}\sum_{j=1}^{L}(I_n(i,j)-M_n)^2, \quad (5)$$

$$D_n=\sqrt{\left(\sum_{i=1}^{K}\sum_{j=1}^{L}\mathrm{MV_h}(i,j)\right)^2+\left(\sum_{i=1}^{K}\sum_{j=1}^{L}\mathrm{MV_v}(i,j)\right)^2}. \quad (6)$$

Following Tang (2007), the motion intensity $M_\mathrm{i}(i, j)$ is computed inside a spatial window that contains $W\times W$ (W=5 experimentally) blocks and centers on the (i, j)th block in the nth frame:

$$M_\mathrm{i}(i,j)=\frac{1}{W\times W}\sum_{n=j-(W-1)/2}^{j+(W-1)/2}\sum_{m=i-(W-1)/2}^{i+(W-1)/2}I_n(m,n). \quad (7)$$

The motion spatial coherence $M_\mathrm{s}(i, j)$ is defined as the entropy of the directions of motion vectors inside the spatial window:

$$M_\mathrm{s}(i,j)=-\sum_{k=1}^{n_\mathrm{s}}p(k)\lg p(k), \quad (8)$$

where $p(k)$ is the probability of occurrence of the kth bin and n_s is the number of histogram bins. The motion spatial coherence measures the consistency of magnitudes of motion vectors in a region.

The motion contrast coherence $M_\mathrm{c}(i, j)$ measures the contrast of motion intensities within the neighborhood of the (i, j)th block:

$$M_\mathrm{c}(i,j)=\begin{cases}|1-I_\mathrm{lmin}|/I_\mathrm{lmax}, & \text{if } I_\mathrm{lmin}\neq 0,\\ I_\mathrm{lmax}/I_\mathrm{max}, & \text{if } I_\mathrm{lmin}=0 \text{ and } I_\mathrm{max}\neq 0,\\ 0, & \text{otherwise,}\end{cases} \quad (9)$$

where I_lmin and I_lmax are the minimum and maximum motion intensities in the spatial window, respectively, and I_max is the maximum motion intensity in the nth frame. Then, $M_\mathrm{i}(i, j)$, $M_\mathrm{s}(i, j)$ and $M_\mathrm{c}(i, j)$ are normalized in the range of [0, 1], and three coefficient maps are produced for easy computation.

The motion attended region $R_\mathrm{m}(i, j)$ is acquired by the following equations:

$$R_\mathrm{m}(i,j)=\begin{cases}1, & \text{if } M_n<M_1^* \text{ and } V_n<V_1^*,\\ 0, & \text{if } M_n>M_2^* \text{ or } D_n<D^* \text{ or } V_n<V_2^*,\\ C_n(i,j), & \text{otherwise,}\end{cases} \quad (10)$$

$$C_n(i,j)=\begin{cases}S_n'(i,j)\,||\,C_n'(i,j), & \text{if } M_\mathrm{i}(i,j)>I^*,\\ S_n'(i,j)\,\&\,C_n'(i,j), & \text{otherwise,}\end{cases} \quad (11)$$

$$S_n'(i,j)=\begin{cases}1, & \text{if } M_\mathrm{s}(i,j)>S^*,\\ 0, & \text{otherwise,}\end{cases} \quad (12a)$$

$$C_n'(i,j)=\begin{cases}1, & \text{if } M_\mathrm{c}(i,j)>C^* \text{ and } M_\mathrm{i}(i,j)>I^*,\\ 0, & \text{otherwise.}\end{cases} \quad (12b)$$

If M_n and V_n are small enough (M_1^* and V_1^* are both set to 0.015 experimentally), there is a small motion in the scene; in such a case, the whole frame would be taken as a motion attended region and every weighting coefficient would be the most salient. If M_n is very large (M_2^* is set to 5.50 experimentally) or D_n, V_n are smaller than the thresholds (D^* and V_2^* are both set to 0.01 experimentally), all blocks in the scene would be taken as unattended ones and the quality of this frame would not contribute to the final quality of the video sequence. Otherwise, the motion attended region is identified by $M_\mathrm{i}(i, j)$, $M_\mathrm{s}(i, j)$ and $M_\mathrm{c}(i, j)$ (S^*, C^*, and I^* are set to 0.95, 0.93, and 0.93, respectively).

Distortion attended region

Visual attention can also be guided by the goal-driven (top-down) mechanism; it means that information can feed back from high-level cortical regions in the parietal and prefrontal cortex to early processing stations (Grill-Spector and Malach, 2004). When the observer views a video sequence in which most of the frames have high qualities, the sudden appearance of an apparent distortion would annoy the observer. Since the quality of the video sequence is affected by this emotion of the observer, the detection of the distortion attended region, which has a relatively low quality compared with its neighbors, is another important process in the proposed attention

model. If the scene has a coherent quality, there will be no distortion attended region, even though the entire scene is terribly damaged.

The local structural similarity $\boldsymbol{S}(\boldsymbol{x}, \boldsymbol{y})$ is measured using the SSIM index (Eq.(1)). The structural similarity $S_n(i, j)$ of the block indexed (i, j) is the average of $\boldsymbol{S}(\boldsymbol{x}, \boldsymbol{y})$ in the block. The distortion contrast coherence $D_c(i, j)$ measures the contrast of structural similarities within the neighborhood of the (i, j)th block:

$$D_c(i,j)=\begin{cases}\dfrac{S_{l\max}-S_{l\min}}{S_{l\max}+S_{l\min}}, & \text{if } S_{l\min}\neq 0,\\ \dfrac{S_{l\max}}{S_{\max}}, & \text{if } S_{l\min}=0 \text{ and } S_{\max}\neq 0,\\ 0, & \text{otherwise,}\end{cases} \quad (13)$$

where $S_{l\min}$ and $S_{l\max}$ are the minimum and maximum structural similarities in the spatial window (which contains $W\times W$ blocks and centers on the (i, j)th block, W=5 experimentally), respectively, and $S_{\max}$ is the maximum structural similarity in the nth frame. $D_c(i, j)$ is normalized in the range of [0, 1] (the closer $D_c(i, j)$ is to 1, the more explicit the contrast of the quality becomes), and a distortion contrast coherence coefficient map is produced. The region with $D_c(i, j)$ larger than D_c^* (D_c^* is set to 0.3 experimentally), implying a relatively low quality, is the distortion attended region:

$$R_d(i,j)=\begin{cases}1, & \text{if } D_c(i,j)>D_c^*,\\ 0, & \text{otherwise.}\end{cases} \quad (14)$$

Object quality of the video

The visual attention saliency map $P_n(i, j)$ of the nth frame is integrated by the above four coefficient maps, $M_i(i, j)$, $M_s(i, j)$, $M_c(i, j)$ and $D_c(i, j)$, through an attention shift mechanism based on the area of the distortion attended region $R_d(i, j)$. Normally, the quality of the nth frame is measured in the motion attended region. But when the area of the distortion attended region is relatively large to a degree, the artifact will make the observer shift his or her attention from the motion attended region to the distortion attended region. The observer is always more concerned with this annoying information, and the quality of the nth frame is measured in the distortion region.

The visual attention saliency map $P_n(i, j)$ is defined as

$$P_n(i,j)=\begin{cases}M_i(i,j)\times(\alpha+M_s(i,j))+\beta\times M_c(i,j), & \\ \qquad\qquad \text{if } A_n=A_m, & \\ \gamma\times D_c(i,j), \quad \text{otherwise.}\end{cases} \quad (15)$$

where

$$A_n=\begin{cases}A_m, & \text{if AREA}(R_d(i,j))<A^*,\\ A_d, & \text{otherwise,}\end{cases} \quad (16)$$

α, β and γ are set to 1, 2 and 4 respectively, and A_n is the attribute of the visual attended region for the nth frame. When the area of the distortion attended region is smaller than A^* (A^* is set to 16 experimentally), A_n is set to the attribute of motion (A_m); otherwise, it is set to the attribute of distortion (A_d).

The structural similarity $S_n(i, j)$ of the block is combined into a quality of the nth frame using

$$Q_n=\frac{\sum_{j=1}^{L}\sum_{i=1}^{K}(S_n(i,j)\times P_n(i,j))}{\sum_{j=1}^{L}\sum_{i=1}^{K}P_n(i,j)}. \quad (17)$$

The object quality of the video sequence is the average of all the N frames:

$$Q_N=\frac{1}{N}\sum_{n=1}^{N}Q_n. \quad (18)$$

However, every frame-level quality offers a different contribution to the final sequence-level quality. The qualities of the damaged frames impact the final quality more than others. Hence, the object quality of the video sequence with a total of N frames predicted by the proposed video structural similarity metric with the weighting adjustment is gained by

$$Q_N=\frac{\sum_{n=1}^{N}\lambda_n Q_n}{\sum_{n=1}^{N}\lambda_n}, \quad \lambda_n=\begin{cases}F_m(f_n), & \text{if } A_n=A_m,\\ F_d(f_n), & \text{otherwise,}\end{cases} \quad (19)$$

where λ_n, the weighting factor of the nth frame, is decided by two non-linear weighting functions $F_m(f_n)$ and $F_d(f_n)$. These two functions are for the frames

with the motion attended regions and the distortion attended regions respectively (Fig.2). λ_n for the frame with the distortion attended regions is much larger than for the frame with the motion attended regions, reflecting the impact of the few extremely damaged frames.

$$F_{\mathrm{m}}(f_n)=\begin{cases}a_1+b_1\lg(1+c_1\sqrt{f_n}), & 0\le f_n<40,\\ 1.0, & f_n\ge 40,\end{cases} \quad (20a)$$

$$F_{\mathrm{d}}(f_n)=\begin{cases}a_2+b_2\lg(1+c_2 f_n^2), & 0\le f_n<15,\\ 10.0, & f_n\ge 15.\end{cases} \quad (20b)$$

The parameters are decided by experiments: a_1=0.5, b_1=0.4, c_1=0.4, a_2=1, b_2=2, c_2=0.5. f_n is defined as a frame index that increases from 0 and, returns to 0 when the attended region is shifting between the motion attended region and the distortion attended region:

$$f_n=\begin{cases}f_{n-1}+1, & \text{if } A_n=A_{n-1},\\ 0, & \text{otherwise.}\end{cases} \quad (21)$$

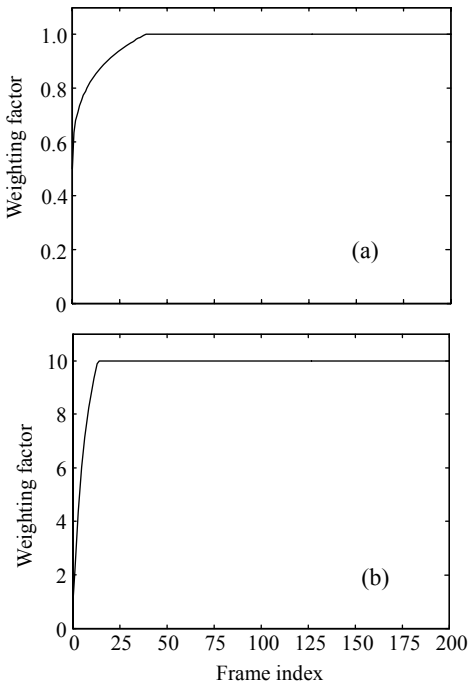

Fig.2 The non-linear weighting functions
(a) $F_{\mathrm{m}}(f_n)$ for the frame with motion attended regions; (b) $F_{\mathrm{d}}(f_n)$ for the frame with distortion attended regions

EXPERIMENTS AND DISCUSSION

To validate the proposed video structural similarity metric for the videos, a different set of video sequences which are the Video Quality Experts Group (VQEG) Phase I 50 Hz datasets (VQEG, 2000), was used afterwards. There are 10 sources (SRC 1~10) and 16 hypothetical reference circuits (HRC 1~16) to produce 160 decoded video sequences with associated DMOSs. For the validation set, the results of the motion based classification are listed in Table 1. SRCs 1 (Tree), 2 (Barcelona) and 4 (Moving graphic) have a low motion content while SRCs 5 (Canoa), 7 (Fries) and 9 (Rugby) comprise very high motion scenes. The remaining SRC sequences, i.e., SRCs 3 (Harp), 6 (F1 car), 8 (Horizontal scrolling) and 10 (Mobile and Calendar), contain moderate motions.

Table 1 The motion classification of the VQEG's Phase I 50 Hz datasets

SRC	M_n	D_n	V_n	Motion
1	0.00	0.00	0.00	Low
2	1.18	0.72	0.70	Low
3	1.25	0.41	0.90	Moderate
4	0.07	0.04	0.37	Low
5	3.80	1.31	13.01	High
6	1.75	0.77	8.70	Moderate
7	3.08	0.94	17.85	High
8	2.06	1.51	25.09	Moderate
9	3.56	1.23	20.71	High
10	1.50	0.61	1.21	Moderate

The motion attended regions of SRCs 4 (Moving graphic), 9 (Rugby) and 10 (Mobile and Calendar) are shown in Fig.3, representing the different motion classification. In Fig.3a, the 183rd frame of SRC 4, which does not have any motion, is all classified as attended, whereas in Fig.3b, the 107th frame of SRC 9 has very strong motions such that the observer will pay no attention to it. In Fig.3c, the ball and the train are attractive at the beginning of the video, and they are moving in the same direction. This repeated action makes the observer fatigue. Thus, the characters on the calendar which are moving up and down take over the observer's attention. When the ball meets the rolling toy, the observer will turn his or her attention back on the ball again.

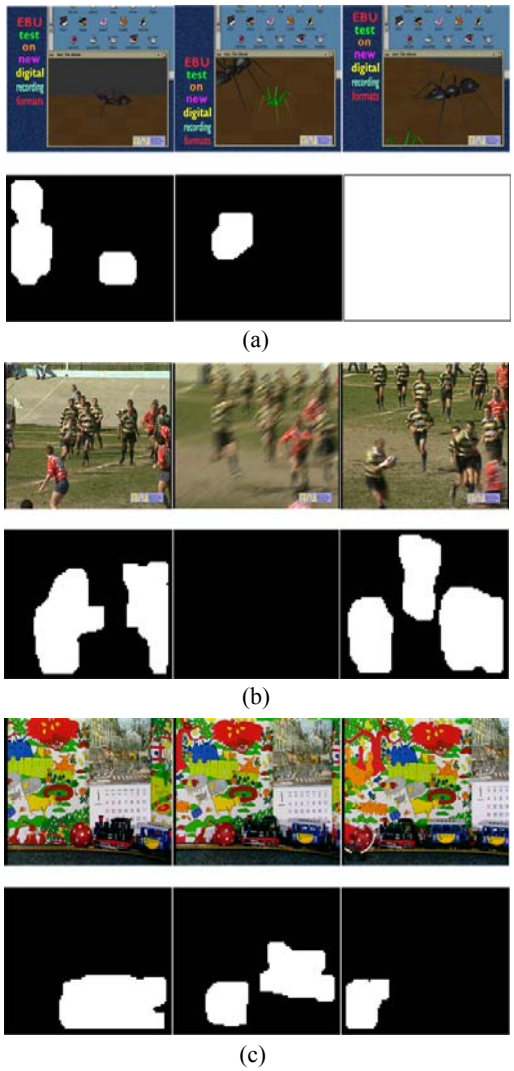

Fig.3 Results of the motion attended region

(a) The 38th, 142nd, 183rd frames of SRC 4; (b) The 41st, 107th, 131st frames of SRC 9; (c) The 28th, 99th, 196th frames of SRC 10. The white area is the motion attended region and the black area is unnoticed

The results of the visual attention model on SRC 8 (Horizontal scrolling) HRC 11 are illustrated in Fig.4. There are a few extremely damaged frames in SRC 8 HRC 11, such as the 64th frame with a slice of artifacts at the top (Fig.4a). In this frame, the scrolling characters and the waving background should be looked on (Fig.4b), but the observer shifts his or her attention to the distortion attended region shown in Fig.4c. Therein the visual saliency map is decided by the distortion contrast coherence. In Fig.4d, the region with relatively low quality displays the final visual saliency map.

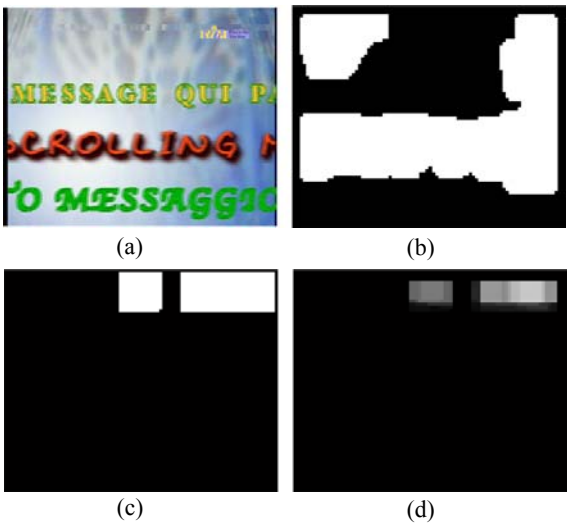

Fig.4 Results of the visual attention model

(a) The 64th frame of SRC 8 HRC 11; (b) The motion attended region (the white area is the motion attended region and the black area is unnoticed); (c) The distortion attended region (the white area is the distortion attended region and the black area is unnoticed); (d) The visual attention saliency map (the lighter, the more salience)

Followed by the performance evaluation procedures employed in the VQEG Phase I test (VQEG, 2000), a multi-parameter logistic function was used to estimate the predicted $\mathrm{DMOS_p}$ from the output Q_N of Eq.(19):

$$\mathrm{DMOS_p} = \beta_1 \mathrm{logistic}(\beta_2, Q_N - \beta_3) + \beta_4 Q_N + \beta_5, \quad (22)$$

where

$$\mathrm{logistic}(\tau, x) = \frac{1}{2} - \frac{1}{1+\mathrm{e}^{\tau x}}.$$

To search for the parameters, the line fitting was conducted. The video database with 160 decoded videos was split into two disjoint sets of videos for fitting and prediction. The fitting set videos were not used in the prediction set. Eighty videos randomly chosen from the database were used for fitting, and the remaining 80 videos were used for prediction. The first 20 frames of each video sequence did not participate in prediction because their qualities are not stable. Figs.5a and 5b show the scatter plots and the fitting lines comparisons between the traditional SSIM index and the proposed video structural similarity metric, respectively. Each sample point represents one test video sequence.

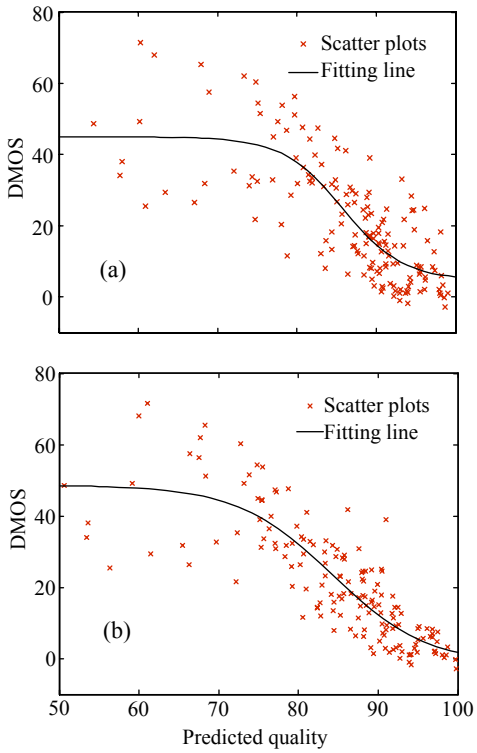

Fig.5 The scatter plot comparison between (a) SSIM index and (b) the proposed video structural similarity metric

Three parameters (VQEG, 2000) were employed to quantitatively measure the performance of the proposed metric.

● Pearson linear correlation coefficient between $DMOS_p$ and DMOS, M_1, which provides an evaluation of the prediction accuracy.

● Spearman rank order correlation coefficient between $DMOS_p$ and DMOS, M_2, which is considered as a measure of prediction monotonicity.

● Outlier ratio of outlier sequences to the total number of testing sequences, M_3.

The higher the M_1 and M_2, and the lower the M_3, the better the match between $DMOS_p$ and DMOS. In Table 2, the proposed metric with the average of frames by Eq.(18) provided reasonably good results compared with the SSIM index (Wang *et al*., 2004b), and the proposed metric with weighting adjustment by Eq.(19) performed the best as a result of its dealing with the burst of error using the non-linear weighting functions.

Some predicted qualities of the video sequences with the burst of error (Table 3) were produced by HRC11 and HRC12. The Q_N was calculated by Eqs.(18) and (19) respectively. The $DMOS_p$ was obtained from the fitting line (Eq.(22)). The proposed metric with weighting adjustment produced $DMOS_p$ more consistent than the average.

Table 3 Some predicted qualities of the video sequences with the burst of error

Sequence	DMOS	Proposed_avg		Proposed_w	
		Q_N	$DMOS_p$	Q_N	$DMOS_p$
SRC 2 HRC 11	32.78	82.87	26.22	81.50	29.26
SRC 3 HRC 12	23.29	86.85	18.52	85.23	21.51
SRC 7 HRC 11	29.43	84.87	22.27	80.62	30.98
SRC 8 HRC 11	28.29	92.66	9.87	85.51	20.92
SRC 9 HRC 12	41.08	83.17	25.63	81.07	30.11
SRC 10 HRC 11	29.84	85.05	21.92	83.27	25.63

Fig.6 shows the predicted quality of SRC 8 HRC 11. From the 21st to 220th frame, the quality of each frame predicted by Eq.(18) is presented in Fig.6a (100 stands for the highest quality, and 0 stands for the lowest quality). The quality was higher than 90 at first and then fell down to 75; this could be the result of the artifacts caused by block spatial shifting. Since about the 70th frame, most of the frames had a high quality except the extremely damaged frames from 169th to 179th. When an observer views this video sequence, his or her opinion may change along with time. The proposed metric mimics this phenomenon. Fig.6b shows the predicted quality acquired by Eq.(19) from 0 to N s. The video sequence lasted nearly 9 s, and the final quality was about 85, but the average of 200 frames was about 92, which is too high to reflect the real quality of this video.

Table 2 Performance comparison on the VQEG's Phase I 50 Hz datasets

Video quality assessment metric	M_1	M_2	M_3	Comment
SSIM_avg	0.788	0.774	0.654	Without weighting adjustment, the average of 200 frames
SSIM_w	0.812	0.804	0.631	With weighting adjustment
Proposed_avg	0.817	0.812	0.619	Without weighting adjustment, the average of 200 frames
Proposed_w	0.845	0.859	0.580	With weighting adjustment

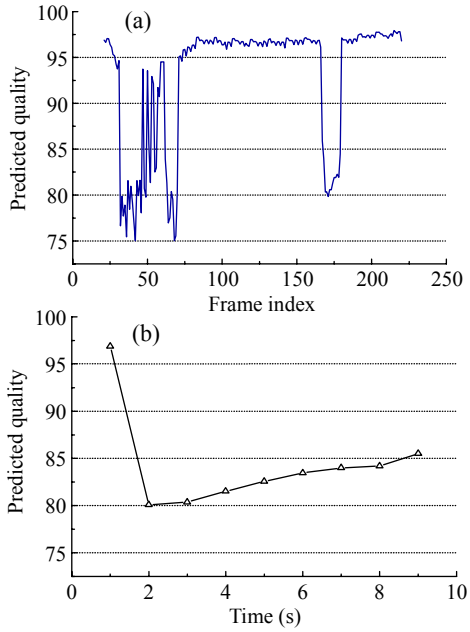

Fig.6 The predicted quality of SRC 8 HRC 11
(a) Predicted quality of the *n*th frame; (b) Predicted quality of a video sequence from 0 to *N* s

CONCLUSION

A structural similarity quality metric for videos based on a spatial-temporal visual attention model is proposed. One feature of the model, especially for video assessment, is that it is based not only on the motion features but also on the distortion contrast; the model mimics the visual attention shifting between the motion attended region and the distortion attended region. The other feature is about its inclusion of a non-linear weighting function to deal with the burst of error by offering a much higher weighting factor to the extremely damaged frames. Experiments on VQEG Phase I 50 Hz database illustrated that, the objective qualities predicted by the proposed metric are correlated with the subjective qualities.

In order to improve the proposed metric, the attention shift mechanism and the form of the weighting factor need further investigation to make the objective video quality prediction become more consistent with the subjective result of the observation.

References

Aziz, M.Z., Mertsching, B., 2008. Fast and robust generation of feature maps for region-based visual attention. *IEEE Trans. Image Process.*, **17**(5):633-644. [doi:10.1109/TIP.2008.919365]

Brooks, A.C., Zhao, X.N., Pappas, T.N., 2008. Structural similarity quality metrics in a coding context: exploring the space of realistic distortions. *IEEE Trans. Image Process.*, **17**(8):1261-1273. [doi:10.1109/TIP.2008.926161]

Chen, Q.Q., Chen, Z.B., Gu, X.D., Wang, C., 2007. Attention-based adaptive intra refresh for error-prone video transmission. *IEEE Commun. Mag.*, **45**(1):52-60. [doi:10.1109/MCOM.2007.284538]

Grill-Spector, K., Malach, R., 2004. The human visual cortex. *Ann. Rev. Neurosci*., **27**:649-677. [doi:10.1146/annurev.neuro.27.070203.144220]

Itti, L., 2005. Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. *Vis. Cogn.*, **12**(6):1093-1123.

Itti, L., Baldi, P., 2005. A Principled Approach to Detecting Surprising Events in Video. IEEE Int. Conf. on Computer Vision and Pattern Recognition, p.631-637. [doi:10.1109/CVPR.2005.40]

Lu, Z.K., Liu, W.S., Yang, X.K., Ong, E.P., Yao, S.S., 2005. Modeling visual attention's modulatory aftereffects on visual sensitivity and quality evaluation. *IEEE Trans. Image Process.*, **14**(11):1928-1942. [doi:10.1109/TIP.2005.854478]

Martinez-Rach, M., Lopez, O., Pinol, P., Malumbres, M.P., Oliver, J., 2006. A Study of Objective Quality Assessment Metrics for Video Codec Design and Evaluation. Eighth IEEE Int. Symp. on Multimedia, p.517-524. [doi:10.1109/ISM.2006.15]

Seshadrinathan, K., Bovik, A.C., 2007. A Structural Similarity Metric for Video Based on Motion Models. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, p.I-869-I-872. [doi:10.1109/ICASSP.2007.366046]

Sheikh, H.R., Sabir, M.F., Bovik, A.C., 2006. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Trans. Image Process.*, **15**(11):3440-3451. [doi:10.1109/TIP.2006.881959]

Tang, C.W., 2007. Spatiotemporal visual considerations for video coding. *IEEE Trans. Multim.*, **9**(2):231-238. [doi:10.1109/TMM.2006.886328]

VQEG, 2000. Final Report from the Video Quality Expert Group on the Validation of Objective Models of Video Quality Assessment. Video Quality Expert Group. Available from http://www.vqeg.org [Accessed on Aug. 22, 2008].

Wang, Z., Li, Q., 2007. Video quality assessment using a statistical model of human visual speed perception. *J. Opt. Soc. Am. A*, **24**:B61-B69. [doi:10.1364/JOSAA.24.000B61]

Wang, Z., Simoncelli, E.P., 2005. Translation Insensitive Image Similarity in Complex Wavelet Domain. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, p.573-576.

Wang, Z., Sheikh, H.R., Bovik, A.C., 2003. Objective Video Quality Assessment. *In*: Furht, B., Marques, O. (Eds.), The Handbook of Video Databases: Design and Applications. CRC Press, Florida, USA, p.1041-1078.

Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., 2004a. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, **13**(4):600-612. [doi:10.1109/TIP.2003.819861]

Wang, Z., Lu, L., Bovik, A.C., 2004b. Video quality assessment based on structural distortion measurement. *Signal Process.: Image Commun.*, **19**(2):121-132.

Zheng, Y.Y., 2008. Research on H.264 Region-of-Interest Coding Based on Visual Perception. PhD Thesis, Zhejiang University, Hangzhou, China (in Chinese).