Journal of Information Hiding and Multimedia Signal Processing ©2016 ISSN 2073-4212
Ubiquitous International Volume **7**, Number **4**, July **2016**

Saliency Region and Motion Characteristics Combined Video Quality Assessment

Ying Zhou[1], Yongsheng Liang[1], Wei Liu[1],Cheng He[1],Yue Zhu[2]

[1]Shenzhen Key Laboratory of Visual Media Processing and Transmission
Shenzhen Institute of Information Technology
Shenzhen 518172, PRChina
zhouying722@163.com
[2]CNNC Lanzhou Uranium EnrichmentCo., Ltd.
Lanzhou 730060, PRChina

Received October, 2015; revised May, 2016

Abstract. *In order to accurately evaluate the video quality and make it consistent with the subjective evaluation result, a saliency region and motion characteristics combined video quality assessment is proposed in this paper. This method is based on traditional SSIM and improves its assessment performance. According to the relationship between human visual attention and video's region of saliency, frame saliency is weighted by spatial saliency and temporal motion characteristics, in which the spatial saliency is extracted by spectrum analysis and the temporal one got by motion visual attention model. This saliency is weighted by SSIM index for single frame quality assessment. Finally, based on HVS characteristics and motion characteristics, the entire video quality SMW-SSIM can be got by weighting single-frame quality. Experiment results prove that this method is faster, more convenient and accurate than other SSIM-based assessments.*
Keywords: Human Visual System(HVS); Region of saliency(ROS); Spatio-temporal saliency; Motion characteristics; SSIM.

1. **Introduction.** With the development of video codec, transmission technology and broadcast end, user's demand for video quality has increasing dramatically. For IP network video streaming applications, due to the limits of acquisition apparatus, compression method, transmission medium and broadcasting terminal, the video quality tends to be reduced when capturing, compressing, transmitting or broadcasting video streaming. In order to timely feedback the decline of video quality to the coding, decoding and transmission, it is necessary to evaluate accurately the video quality. Then the user's visual perception can be improved through modifying algorithm and transmission mechanism.

According to structural similarity theory, Zhou Wang et al. [1] proposed Structural Similarity Index Measurement(SSIM). The distortion of image's brightness, contrast, and structure were calculated and averaged to get the entire distortion. However, this method is only accurate for the single image without considering the dynamic information of the video. When the image is severely blurred, the evaluation result is unsatisfactory. In recent years, many scholars have proposed a series of improved algorithms based on SSIM.

Later Wang proposed the Multi-Scale Structural Similarity Index Measurement [2] (MS-SSIM) and the Information Weighted Structural Similarity Index Measurement [3] (IW-SSIM). Literature [4] proposed the Edge-based Structural Similarity for Image Quality

Assessment(ESSIM) to improve the distortion analysis of fuzzy image. Literature [5] proposed the Motion and Edge Information Based Structural Similarity Index Measurement(MESSIM) to weigh the scene structure and motion information according to HVS characteristics. Literature [6] proposed the HVS Based Structural Similarity Index Measurement(HSSIM),which considered HVS sensitivity to different frequency and different regions. The brightness, texture and space location were introduced to adjust the macroblock's weight. The weight of single frame was calculated by the motion vector. The Visual Saliency and Distortion Weighting Based Video Quality Assessment was proposed in literature [7]. Literature [8] averaged density, color and motion vector to obtain the saliency map and weighted SSIM. In literature [9], author first extracted video motion characteristics by multi-scale method to obtain temporal saliency map, then weighted density, color and contrast to obtain spatial saliency map. The final saliency by Gaussian filter was used to weigh SSIM.

Currently, most VQAs are still based on calculating the image quality frame by frame, and then taking the average or the weighted average of all frames as the score of the whole video quality. However, the most significant characteristics distinguishing video from image is video's dynamic nature. When viewing a video sequence, the human eyes can obviously feel the direction and consistency of the moving object. Such VQAs mentioned above ignored the most important video motion characteristics.

When viewing video, the human eyes will quickly transfer to the region of interest (ROS),which mostly contains rich textures, more detailed information, or more severe movement. The background is often ignored by human eyes. If the same error distortion occurs in ROS, the corresponding error propagation distortion is more severe than that occurred in the non-ROS. As time goes on, the human visual attention on moving target will gradually decrease, and then transfer to new emerging objects or the lower significant regions.

Secondly, HVS has different response to videos with different motion characteristics. The movement in the same video scene is more intense, human visual perception of the distortion is worse. For fast movement, it is difficult to distinguish the details, and therefore the distortion occurred in the fast moving video frame is hard to detect. When the scene change occurs, human eyes are almost imperceptible for the distortion in the next few frames, therefore, the distortion in scene changing frames should be given very little weight.

With the above consideration, we propose a Saliency and Motion Characteristics Weighted SSIM(SMW-SSIM), which is based on the dynamic motion description and saliency region extraction combining with video motion characteristics. The subjective human visual perception for ROS and moving objects should be fully considered when evaluate the video quality. The core idea is to weighed accumulate the magnitude of motion and the influence of distortion in ROS. If the distortion occurred in the same frame's ROS, the greater weight is given; while the distortion occurred in the frame with violent motion, the smaller weight is given. The framework of this VQA is shown in Figure 1.

2. **Description of Saliency and Motion Characteristics Weighted SSIM(SMW-SSIM).**

2.1. **Saliency Calculation Integrating Video Motion Characteristic.**

When viewing the video, the human eyes are sensitive to moving objects or objects with the great contrast to surroundings. According to these human perceptual characteristics, the spatial and temporal saliency which can reflect the human visual characteristics are extracted,

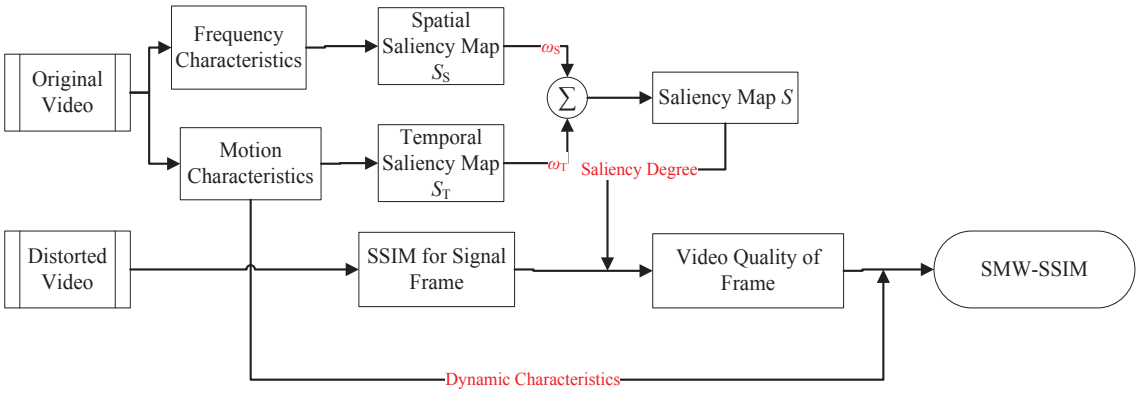


FIGURE 1. The block diagram of video quality assessment

then ROS is weighted accumulated by fusing the spatial and temporal saliency map dynamically.

According to the statistics invariance of image, the spatial characteristics can be maintained still in frequency domain. So the spatial feature of image can be characterized by the energy spectrum in the frequency domain. This method avoids feature selection in the traditional spatial saliency map extraction. The amplitude spectrum can represent the quantity of information changes, and phase spectrum can indicate the location of information changes. In this paper, the spectral Residual (SR) method is used [10] to extract spatial saliency map. The most of image's redundant information is removed by analyzing and processing logarithmic spectrum in the frequency domain [11]. And then through the inverse Fourier transform, the spatial saliency map is extracted. The algorithm details are shown as follows:

$$A(f) = \log\left|\hat{F}(f)\right| * g(f) \tag{1}$$

$$I(A(f)) = A(f) * l(f,k) \tag{2}$$

$$R(f) = A(f) - I(A(f)) \tag{3}$$

$$S_S(f) = g(f) * \hat{F}^{-1}[\exp(R(f) + P(f))]^2 \tag{4}$$

Assuming the input image is f, then $\hat{F}(f)$ is its Fourier transform; $A(f)$ is its logarithmic amplitude spectrum; $h_n(f)$ is the mean filter by $n * n$; $R(f)$ is its redundancy of spectrum; $P(f)$ is its phase spectrum; $g(f)$ is its Gaussian filter; $S_S(f)$ is output spatial saliency map. Among them, $l(f,k) = \frac{1}{\sqrt{2\pi}2^k\sigma} * \exp(-\frac{u^2+v^2}{(2^k\sigma)^2})$ is the Gaussian kernel function; $k = 1, 2...K$ for the scale parameter; $K = \log_2 \min\{H, W\}$ is determined by the image size; H, W is image's height and width respectively.

The quality of ROS extraction is determined by the scale parameter k, if k is too small, the redundant information cannot be effectively removed; while if k is too large, only the boundary information of ROS is significant. The optimal scale parameter k must meet the following equation:

$$k_e = \arg\min(-\sum_{i=1}^{n} S_S(k)_i \log(S_S(k)_i)) \tag{5}$$

Wherein, $S_S(k)_i$ is saliency map with scale parameter k.

The video contains motion information, which is different from the image. When observing video, due to the visual attention mechanism, HVS is more likely attracted by moving object and scene, and has different subjective perception for video with different motion characteristics. According to statistics, when the moving object's intensity is larger or

direction of movement is complicated, the human eyes' attention to this part will increase rapidly. By correlation experiments, three temporal characteristics are selected,which can best reflect motion characteristics of ROS and be calculated easily: motion density, motion intensity and motion direction.

Firstly, the method mentioned in literature [12] is used to extract the moving target, then the target's motion vector map is filtered with Gaussian filter. Finally according to motion density, intensity and direction three feature channels, the temporal saliency map of current frame is obtained by Visual Attention Model(VAM).

With time going on, the human eye's attention to moving object will change [12]. When moving target appears continually, the human eye's sensitivity will decrease. At this time, human eyes will turn to focus on the newly emerging object or ROS with lower saliency then before [13]. Therefore, the final temporal saliency map $S_T(i)'$ is modified as following:

$$S_T(i)' = norm\{S_T(i) + \int_{t_i-\Delta t}^{t_i} [(S_T(i) - S_T(i-\Delta i)) * g]dt\} \tag{6}$$

Wherein, g is a Gaussian function, $(i-\Delta i)$is another frame with interval Δi, $S_T(i)'$ is the corrected temporal saliency map.

According to human visual perception characteristics, spatio-temporal saliency map $S(n)$ is integrated adaptively for each frame of a video, which is shown as the following formula:

$$S(n) = \frac{\sum_{i=1}^{N} \omega_S \times S_S(i) + \sum_{i=1}^{N} \omega_T \times S_T(i)}{\sum_{i=1}^{n} \omega_S + \sum_{i=1}^{n} \omega_T} \tag{7}$$

Wherein, $S_S(i)$ denotes the spatial saliency map of the i-th frame, $S_T(i)$ represents the temporal saliency map,S is weighted spatio-temporal saliency map. ω_S, ω_T are weights.

According to the human eye's concentric feature and the relationship between ROS distribution and human eye's attention [14], this paper assumes that the image center position is (x, y), any pixel position of ROS is (x_i, y_i), then:

$$\omega_S = \|(x_i, y_i) - (x, y)\| * \frac{1}{2\pi\sigma} \exp(-\frac{(x_i - x)^2 + (y_i - y)^2}{2\sigma^2}) \tag{8}$$

Wherein, ω_S is the Gaussian weighted Euclidean distance. If ROS is closer to the center, distribution is more concentrated, so the human eyes are more sensitive and the saliency weight is higher.

ω_T is adjusted dynamically according to the video motion characteristics, considering the motion spatial distribution ω_1, motion intensity ω_2 and motion complexity ω_3 three factors [11] simultaneously, so that

$$\omega_T = \omega_1 \times \omega_2 \times \omega_3 \tag{9}$$

$$\omega_1 = \frac{N_O}{N(s)} \tag{10}$$

Wherein, N_O denotes the number of non-zero macroblocks of target's motion vector, $N(s)$ is the number of macroblocks of frame, then ω_1 represents the spatial distribution of motion, which is larger, the motion distribution is more concentrated.

$$\omega_2 = \frac{\sum_{i=1}^{N_{kB}} (\|v_x\| + \|v_y\|)}{\sum_{i=1}^{N(s)} (\|v_x\| + \|v_y\|)} \tag{11}$$

Wherein, v_x and v_y denote the horizontal and vertical coordinates of the target's motion vector; ω_2 denotes motion energy, which is greater, the motion information is more abundant.

$$\omega_3 = \frac{-\left[\sum_{i=1}^{m} \frac{N(s_i)}{N(s)} \times \log\left(\frac{N(s_i)}{N(s)}\right)\right]}{\lg(36)} \tag{12}$$

Wherein, s_i denotes the non-empty dimension of each direction histogram of target motion vector; $N(s_i)$ denotes the number of macro blocks with non-zero motion vectors in each dimension; $i \leq 36$, ω_3 is macroblock motion vector's distribution in each dimension, which is greater, the motion direction is more complicated.

2.2. **Video Motion Characteristics Weighting Factor.** In the video sequence, $\vec{v}(i, j, n)$ denotes the motion vector of the n-th frame at $(i\ ,\ j)$, $v = \|\vec{v}\|_2$ indicates its strength. Motion vector field of a frame is comprised of three motion vectors: the absolute motion $\vec{v}_a$, background motion $\vec{v}_b$ and relative motion $\vec{v}_r$. Three motion vectors meet:

$$\vec{v}_r = \vec{v}_a - \vec{v}_b \tag{13}$$

Wherein, $\vec{v}_r$ can be calculated by the global motion estimation mentioned in the literature [12]. Then the motion intensity of the n-th frame M_v is weighted by motion vector intensity v_r and v_b, which is expressed as the following equation:

$$M_v = (1 - \omega_b)v_r + \omega_b v_b \tag{14}$$

Wherein, ω_b is composed of motion spatial distribution ω_1, motion intensity ω_2 and motion complexity ω_3, which is shown as the following formula:

$$\omega_b = \omega_1 \times \omega_2 \times \omega_3 \tag{15}$$

The human eyes' response to the visual stimuli is non-linear. The video motion is more dramatic, the human eyes are more likely ignore its distortion. So the weight based on video motion characteristics can be shown as following:

$$\omega_M = \log(\frac{M_{\max}}{M_v^{\alpha}}) \tag{16}$$

Wherein, $M_{\max}$ is the maximum of weighted sum of motion vector intensity, α is a constant.

2.3. **Saliency and Motion Characteristics Weighted Video Quality Assessment.** Supposing the video sequence consisting of $n = 1, 2, \ldots, N$. (i, j) represents the position coordinates of corresponding pixel, (H, W) represents the height and width of video frame, $SSIM_{ij}(n)$ represents the structural similarity index on the position (i, j) of the n-th frame.

SSIM proposed by Zhou Wang [1] is used to calculate $SSIM_{ij}(n)$ between distortion and original video frame by frame. This index is weighted by each frame's saliency map mentioned in this paper. The improved Frame Quality Metric(FQM) can be calculated as the following formula:

$$FQM(n) = \frac{\sum_{i=1}^{H}\sum_{j=1}^{W}[SSIM_{ij}(n) \cdot S(n)]}{\sum_{i=1}^{H}\sum_{j=1}^{W} S(n)} \tag{17}$$

According to HVS characteristics, ROS and scene with dramatic movement can obviously attract human eyes' attention. While the single-frame quality considering ROS, the entire

VQA should consider motion characteristics. The single-frame quality is weighted by motion severity index and the entire video quality SMW-SSIM can be got as shown:

$$SMW - SSIM = \frac{\sum_{n=1}^{N} \omega_M \cdot FQM(n)}{\sum_{n=1}^{N} \omega_M} \tag{18}$$

3. **Simulation Experiments and Results Analysis.** In order to prove SMW-SSIM is closer to the human eyes' subjective assessment, the public database LIVE VQA is used for experiment [15]. The database consists of ten kinds of video sequences. Each kind of video sequence comprises of an original video, fifteen corresponding distortion sequences, four distortion videos through wireless transmission, three through IP transmission, four compressed videos by H.264, four by MPEG-2. The entire database contains150 test sequences.

According to the report provided by VQEG [16], the relationship between objective and subjective VQA is non-linear, this mapping relationship can be regressed using the following non-linear equation:

$$f(x) = \alpha_1 + \frac{\alpha_2 - \alpha_1}{1 + e^{-\frac{x-\alpha_3}{\alpha_4}}} \tag{19}$$

Wherein, x represents the objective assessment index, $f(x)$ is the processed result. In this paper, the nlinfit-function in Matlab is adopted as the non-linear least square optimization.

For comparing SMW-SSIM with other VQAs and proving that this index is closer to the subjective evaluation, the Spearman Rank Order Correlation Coefficient (SROCC), Linear Correlation Coefficient (LCC) and Outlier Ratio (OR) three indicators are chosen to judge the performance of various evaluation methods. The larger SROCC and LCC and the smaller OR, the objective assessment result is better.

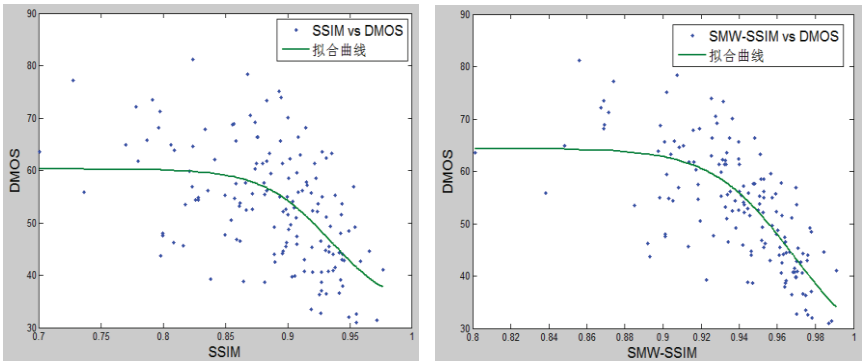
MS-SSIM can also be weighted by the method in this paper for each frame. The improved SMW-MSSSIM is accumulated by weighing every fame according to the motion characteristics. The comparing result of each VQA is shown in Table 1:

Table 1. The quality evaluation results of LIVE VQA database

VQA	Evaluation index		
	LCC	SROCC	OR
PSNR	0.3984	0.3492	99.63
VSNR	0.6842	0.6719	62.69
MOVIE	0.8067	0.7795	40.94
VQM	0.7162	0.6981	56.9
VIF	0.5670	0.5594	80.37
SSIM	0.5378	0.5163	86.71
MS-SSIM	0.7243	0.7146	57.66
SMW-SSIM	**0.7301**	**0.7182**	**56.65**
SMW-MSSSIM	**0.7372**	**0.731**	**55.3874**

As can be seen from Table 1, compared with the original SSIM and MS-SSIM algorithm, SMW-SSIM and SMW-MSSSIM index have higher accuracy, monotonicity and lower out ratio, which proves the superiority of this method.

For comparing the VQAs more intuitively, the experimental scatter plots are introduced. The closer objective predictive value to subjective evaluation, the more linear fitted curve is. The scatter plots of subjective/objective scores is shown in Figure 2:



(a) (b)

FIGURE 2. Scatter plots comparison of different VQA on LIVE VQA database:(a) SSIM; (b)SMW-SSIM

As can be seen from Figure 2, the scatter plots of method in this paper is more similar to a straight line, which indicates that SMW-SSIM is more effective and closer to the human eyes' subjective evaluation.

4. **Conclusion.** Compared with other video quality assessments(VQA) such as MOVIE (Motion-based Video Integrity Evaluation), PSNR(Peak Signal-to-Noise Ratio), VSNR (Video Signal-to-Noise Ratio), VQM(Video Quality Metric), VIF(Visual Information Fidelity) and so on, SSIM has been widely spread due to simple calculation and good accuracy. So this paper proposes a full-reference objective VQA based on SSIM and improves its performance. HVS focusing on ROS and dynamic motion characteristics are two important features distinguishing video from image. They should be considered when evaluating video quality. The saliency map of each frame's is weighted by spatial and temporal saliency, then according to dynamic motion characteristics, entire VQA is accumulated by weighting SSIM index of each frame. Experiments with LIVE VQA standard database show that this assessment result is closer to the human eyes' subjective judgment than other conventional SSIM index. SMW-SSIM is easier to apply and more accurate.

Acknowledgments. This work is supported by National Natural Science Foundation of China 61172165 and Natural Science Foundation of Guangdong Province S2011010006113. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] Z. Wang, A. C. Bovik, H. R. Sheikh Image Quality Assessment: From Error Visibility to Structural Similarity *IEEE Tran. on Image Proc.*, vol.13, no.4, pp.600-612. 2004

[2] Z. Wang, E. Simoncelli, A. Bovik, Multi-scale Structural Similarity for Image Quality Assessment, *Proc. of the 37th IEEE Asilomar Conference on Signals, Systems and Computers*, 2003.

[3] Z. Wang, Q. Li Information Content Weighting for Perceptual Image Quality Assessment *IEEE Trans. on Image Processing*, vol. 20 no. 5 pp.1185-1198, 2011.

[4] G. Chen, C. Yang, L. Po, Edge-based Structural Similarity for Image Quality Assessment, *ICASSP*, 2006.

[5] G. Lu, J. Li, G. Chen, Video Quality Assessment Measurement Based on Motion Information and Structural Distortion, *Computer Simulation*, vol.27, no.6, pp. 262-267, 2010

[6] Z. Wang, Y. Liao, B. Wang, A Novel HVS-based SSIM on Video Quality Assessment, *Communications Technology*, vol.43, no.2, pp. 77-81, 2010.

[7] L. Zhu, L. Su, Q. Huang, et al, Visual Saliency and Distortion Weighting Based Video Quality Assessment *Pacific-rim Conference on Multimedia*, 2012.

[8] X. Gao, N. Liu, W. Lu, et al, Spatio-temporal Salience Based Video Quality Assessment, *IEEE International Conference on Systems Man and Cybernetics*, 2010.

[9] B. Fu, Z. Lu, X. Wen, et al, Visual Attention Modeling for Video Quality Assessment With Structural Similarity *16th International Symposium on Wireless Personal Multimedia Communications*, 2013.

[10] X. Hou, L. Zhang, Saliency Detection: A Spectral Residual Approach, *IEEE Conference on Computer Vision and Pattern Recognition* 2007.

[11] Y. Zhou, J. Zhang, Y. Liang, W. Liu, Video Motion Characteristics based Spatial-Temporal Salient Region Extraction Method, *Journal of Information Hiding and Multimedia Signal Processing*, vol. 6, no. 2, pp. 225-233, 2015.

[12] Y. Zhou, W. Liu, J. Zhang, Content-Aware Based Sorting Approach of Scalable Video Bit Stream, *Journal of Signal Processing* vol.29, no.8, pp. 1012-1018, 2013.

[13] X. Chen, J. Zhang, W. Liu, New Scalable ROI Algorithm Based on Visual Attention, *Journal of Shandong University*, vol. 43, no. 1, pp. 1-8, 2013.

[14] V. Gopalakrishnan, Y Hu, D. rajan, Salient Region Detection by Modeling Distributions of Color and Orientation *IEEE transactions on multimedia*, vol.2, no.5, pp.892-905, 2011.

[15] Z. Wang, Q. Li, Video Quality Assessment Using a Statistical Model of Human Visual Speed Perception, *Journal of The Optical Society Of America*, vol.24, no.12, pp. 61-69, 2007.

[16] Final report from the video quality experts group on the validation of objective quality metrics for video quality assessment.