



Low-Complexity Video Quality Assessment Based on Spatio-Temporal Structure

Yaqi Lu, Mei Yu, and Gangyi Jiang^(✉)

Ningbo University, Ningbo 315211, China
jianggangyi@126.com

Abstract. Low-complexity is as important as prediction accuracy for video quality assessment (VQA) metrics to be practically deployable. In this paper, we develop an effective and efficient full-reference VQA algorithm, called Spatio-temporal Structural-based Video Quality Metric (SSVQM). To be more specific, spatio-temporal structural information is sensitive to both spatial distortions and temporal distortions. We calculate spatio-temporal structure based local quality according to spatio-temporal gradient characteristics and chrominance information. Then, these local quality scores are integrated to yield an overall video quality via a spatio-temporal pooling strategy simulating three most important global temporal effects of the human visual system, i.e. the smooth effect, the asymmetric tracking effect. Experiments on VQA databases LIVE and CSIQ demonstrate that our SSVQM achieves highly competitive prediction accuracy and delivers very low computational complexity.

Keywords: Video quality · Low-complexity · Spatio-temporal structure

1 Introduction

Video quality assessment (VQA) plays a central role in various video-related applications. It is often employed in quality control of various video services and performance comparison of various video processing algorithms. Objective VQA models [1, 2] have been designed to evaluate video quality in line with the human visual system (HVS). Considering accurate and stable, full-reference VQA (FR-VQA) has been widely used in video processing and coding [2]. FR-VQA intends to estimate video quality quickly and precisely, in which peak signal-to-noise ratio (PSNR) has been the dominant quantitative indicator in image and video processing. However, PSNR does not always correlate well with HVS [3]. The drawback will slow down the development of video processing technology, especially the further promoting of video compression efficiency. In image quality assessment (IQA), based on the assumption that the HVS is highly adapted for extracting structural information from the scene, Structural SIMilarity (SSIM) index [4] was proposed to bring IQA from pixel-based stage to structure-based stage. Then, some state-of-the-art FR-IQA algorithms were developed, such as Feature SIMilarity (FSIM) index [5] and Gradient Magnitude Similarity Deviation [6], etc.

A straightforward approach to VQA is to use these state-of-the-art IQAs for an independent frame and then average all the frame level scores to obtain a composite score. However, due to the lack of temporal information, this approach did not worked well. Other methods deal with motion information either by incorporating motion characteristics into the frame weighting to account for their effects on the spatial distortion, or by modeling the temporal distortion and spatial distortion independently. Video Quality Metric (VQM) [7], MOTion-based Video Integrity Evaluation (MOVIE) [8], Spatio-Temporal MAD (STMAD) [9], spatial and spatiotemporal slices based ViS3 [10] and [11, 12] are the state-of-the-art FR-VQA metrics. Although these approaches improved performance compared to PSNR, a number of shortcomings, including high implementation and computational complexity, long latency and difficulty to integrate, still need to be resolved.

This paper proposes an efficient FR-VQA algorithm, called Spatio-temporal Structural-based Video Quality Metric (SSVQM). We extend the assumption of SSIM to VQA, i.e. HVS is highly adapted for extracting spatio-temporal structural information from dynamic scene. The whole video sequence is regarded as a pixel volume, and local quality is calculated according to spatio-temporal gradient and chrominance characteristics, which are both sensitive to spatial distortions and temporal distortions. Then, local quality scores are integrated to yield an overall video quality via a spatio-temporal pooling strategy simulating three most important global temporal effect.

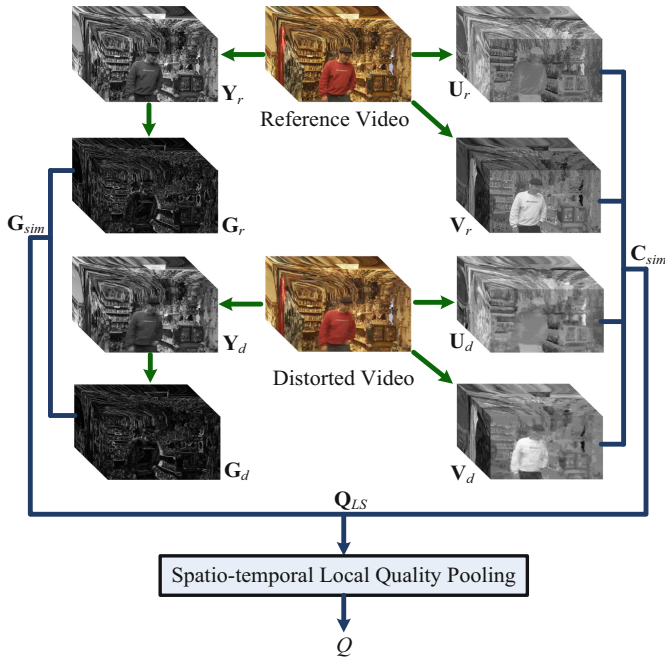


Fig. 1. Framework of the proposed SSVQM.

2 The Proposed SSVQM

In Fig. 1, the framework of the proposed SSVQM mainly consists of two phases: spatio-temporal structure based local quality calculating and spatio-temporal pooling. We use 3D Prewitt operator to calculate spatio-temporal gradients of reference and distorted video sequences.

We compare the video gradients detected by 2D Prewitt operator used in [7] and 3D Prewitt operator used in our SSVQM and the results are shown in Fig. 2. It can be clear that 2D Prewitt operator only detects spatial gradients, whereas 3D Prewitt operator can detect both temporal and spatial gradients. The rough edge represents motion, and the thin edge represents static, which can be used to measure temporal distortion.

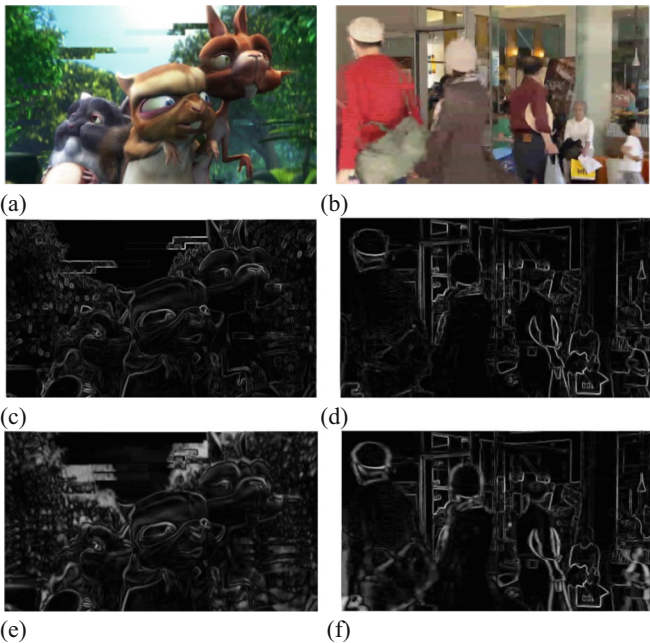


Fig. 2. Comparison of gradient calculated by 2D Prewitt operators and 3D Prewitt operators. (a) and (b) are frames from distorted video, (c) and (d) are gradients detected by 2D Prewitt operators, (e) and (f) are gradients detected by 3D Prewitt operators.

The proposed SSVQM estimates video quality of the reference and distorted videos in the YUV color space. The luminance component of the reference and distorted video are denoted as \mathbf{Y}_r and \mathbf{Y}_d , respectively. The spatial resolution of each video is $W \times H$, F is the total frames evaluated. Assume the luminance component of each video as a 3-D matrix with size of $H \times W \times F$. The original video gradient amplitude and distorted video gradient magnitude, denoted as \mathbf{G}_r and \mathbf{G}_d , are computed via.

$$\mathbf{G}_r = \sqrt{(\mathbf{Y}_r \otimes \mathbf{F}_x)^2 + (\mathbf{Y}_r \otimes \mathbf{F}_y)^2 + (\mathbf{Y}_r \otimes \mathbf{F}_z)^2} \quad (1)$$

$$\mathbf{G}_d = \sqrt{(\mathbf{Y}_d \otimes \mathbf{F}_x)^2 + (\mathbf{Y}_d \otimes \mathbf{F}_y)^2 + (\mathbf{Y}_d \otimes \mathbf{F}_z)^2} \quad (2)$$

where the symbol \otimes denotes the convolution operation, and \mathbf{F}_x , \mathbf{F}_y and \mathbf{F}_z are three templates of 3D Prewitt operator along direction of x -axis, y -axis and t -axis, respectively. The gradient similarity of reference and distorted video sequence denoted as \mathbf{G}_{sim} , is computed via

$$\mathbf{G}_{sim} = \frac{2\mathbf{G}_r \cdot \mathbf{G}_d + c_1}{\mathbf{G}_r^2 + \mathbf{G}_d^2 + c_2} \quad (3)$$

where c_1 is a positive constant to increase stability and fine-tune the feature sensitivity to distortions.

Distortions of chrominance information will also seriously affect the video quality. Generally, the spatial size of two chrominance components both are only 1/4 of the luminance component in widely used YUV420 format. In addition, chrominance components adopt the same encoding scheme as luminance component when video is encoded. Both factors make chrominance components an easy source of degeneration, and therefore chrominance information is an important factor to be considered in the VQA algorithm.

We calculate the chrominance similarity using two chrominance components in YUV format video as [5]. Two chrominance components of the original and distorted video are denoted as \mathbf{U}_r , \mathbf{V}_r and \mathbf{U}_d , \mathbf{V}_d respectively, and the chrominance similarity denoted as \mathbf{C}_{sim} , is computed via

$$\mathbf{C}_{sim} = \frac{2\mathbf{U}_r \cdot \mathbf{U}_d + c_2}{\mathbf{U}_r^2 + \mathbf{U}_d^2 + c_2} \cdot \frac{2\mathbf{V}_r \cdot \mathbf{V}_d + c_3}{\mathbf{V}_r^2 + \mathbf{V}_d^2 + c_3} \quad (4)$$

where c_2 and c_3 are positive constants. Here, we set $c_1 = 90$, $c_2 = c_3 = 300$, and fix them for all databases so that SSVQM can be conveniently used. \mathbf{G}_{sim} and \mathbf{C}_{sim} can then be combined to get the spatio-temporal structure based local quality, denoted as \mathbf{Q}_{LS} .

$$\mathbf{Q}_{LS} = \mathbf{G}_{sim} \cdot (\mathbf{C}_{sim})^\lambda \quad (5)$$

where $\lambda > 0$ is the parameter for adjusting the importance of two chromatic components. Here, λ is set to 3 in order to emphasize the effect of severely distorted in chrominance components.

In this paper, each video is regarded as a pixel volume at the local quality calculating stage ignoring global temporal effects of HVS. At the spatio-temporal pooling stage, the frame-level quality is obtained via an efficient spatial pooling method, then these frame-level quality scores are integrated to yield an overall video quality via a temporal pooling method simulating three most important global temporal effects.

Spatial pooling: The simplest and widely used pooling strategy is average pooling. Here, we adopt a more efficient standard deviation pooling [7] method to calculate frame-level quality. Each pix-level quality in \mathbf{Q}_{LS} is denoted as $Q_{LS}(x, y, t)$, where $x \in \{1, \dots, H\}$, $y \in \{1, \dots, W\}$, $z \in \{1, \dots, F\}$, and frame-level quality can be computed via.

$$Q_{mean}(t) = \frac{1}{H \times W} \sum_{x=1}^H \sum_{y=1}^W Q_{LS}(x, y, t) \quad (6)$$

$$Q_{frame}(t) = \sqrt{\frac{1}{H \times W} \sum_{x=1}^H \sum_{y=1}^W (Q_{LS}(x, y, t) - Q_{mean}(t))^2} \quad (7)$$

Note that a smaller value indicate higher quality, which is opposite to the average pooling method.

Temporal pooling: In this paper, we mainly consider three most important global temporal effects of HVS: (1) the smooth effect [13], *i.e.* the subjective ratings of the whole video sequence typically demonstrate far less variation than the frame-level quality scores. This effect can be simulated by low-pass filtering frame-level quality scores; (2) the asymmetric tracking effect [14], *i.e.* HVS is more sensitive to frame-level quality degradation than improvement. This effect can be simulated using different weights for quality rising and quality declining; (3) the recency effect [15], *i.e.* subjects tend to put a higher weigh on what they have seen most recently. This effect can be simulated using a time-related weight. The implementation in [16] accurately describes the smooth effect and the asymmetric tracking effect. We improved the model by integrating logarithmic function weights to simulate the recency effect. The final video quality denoted as Q , is computed via

$$Q_{LP} = \begin{cases} Q_{LP}(t-1) + \alpha \cdot \Delta Q(t), & \text{if } \Delta Q(t) \leq 0 \\ Q_{LP}(t-1) + \beta \cdot \Delta Q(t), & \text{if } \Delta Q(t) > 0 \end{cases} \quad (8)$$

$$Q = \frac{1}{F} \sum_{t=1}^F (Q_{LP}(t) \cdot \ln(r \cdot t + 1)) \quad (9)$$

where $\Delta Q = Q_{frame}(t) - Q_{LP}(t-1)$, $Q_{LP}(1) = Q_{frame}(1)$, α and β are asymmetric weighs, and γ is a positive constant for adjusting time-related weighs. Unlike [11], we just rough-tune these parameters and fix them as follow: $\alpha = 0.03$, $\beta = 0.2$ and $\gamma = 1000$. It is worth noting that the performance can be improved if we independently fine-tune these parameters for each database, but generalization capability of our pooling method will be slightly affected.

3 Experimental Results and Discussion

Two publicly VQA databases are used in experiments:

- (1) *LIVE* [17]: LIVE VQA database contains 10 reference videos with four different distortions: MPEG-2 compression, H.264/AVC compression, simulated transmission of H.264-compressed bit-streams through error-prone IP networks and error-prone wireless networks, a total of 150 distorted videos. All videos are in YUV420 format with a resolution of 768×432 pixels, two kinds of frame rates: 25 and 50 fps, and about 10 s in duration.
- (2) *CSIQ* [18]: CSIQ VQA database contains 12 reference videos with six different distortions: MJPEG compression, H.264/AVC compression, H.265/HEVC compression, wavelet compression, packet-loss simulated transmission of H.264-compressed bit-streams through wireless networks and additive white Gaussian noise, a total of 216 distorted videos. All videos are in YUV420 format with a resolution of 832×480 pixels, 10 s in duration and five kinds of frame rates: 24, 25, 30, 50 and 60 fps.

Two performance indicators are used in our experiments: Linear Correlation Coefficient (LCC) and Spearman Rank Order Correlation Coefficient (SROCC). LCC measures the prediction accuracy and SROCC measures the prediction monotonicity. We apply a four-parameter logistic transform before computing LCC as recommended by video quality experts group (VQEG) [19]. It is worth noting that VQA on the appropriate scale will be most effective and efficient. Instead of up-sampling two chrominance components for YUV420 format video, luminance component is filtered by a 2×2 average filter in spatial domain first, and then down-sampled by a factor of 2. In this way, not only the size of luminance component and two chrominance components match, but also the complexity of SSVQM is reduced without jeopardizing the prediction accuracy.

The proposed FR-VQA metric, SSVQM, compared with eight well-known quality metrics, including three FR-IQA metrics: PSNR, SSIM, FSIM, four FR-VQA metrics: VQM, MOVIE, STMAD, ViS₃, and one RR-VQA metric STRRED [20] on the LIVE and CSIQ VQA databases. Here, three FR-IQA metrics are extended to video. The results are listed in Table 1 and clearly SSVQM is superior to other metrics on the LIVE database. On the CSIQ database, SSVQM and ViS₃ provide similar SROCC, and both are superior to the other metrics.

We also compare the complexity of our SSVQM with other eight metrics on a video sequence with 250 frames from LIVE VQA database using a Lenovo desktop (Intel Core i5-4590 CPU @3.30 GHz, 8G RAM, Windows 7 64-bit, MATLAB® R2014b). Except MOVIE implemented using C++, all other metrics are implemented using MATLAB, and all source codes were obtained from the original authors. Our SSVQM is also implemented using MATLAB without any optimization. The results are listed in Table 2. The complexity of SSVQM is very low, only 5 times of PSNR. The complexity of ViS₃, STMAD and MOVIE are 28 times, 70 times and 380 times of our SSVQM, respectively.

Table 1. LCC and SROCC on LIVE and CSIQ VQA databases

Metrics	LIVE		CSIQ	
	LCC	SROCC	LCC	SROCC
PSNR	0.5397	0.5231	0.5663	0.5798
SSIM	0.5068	0.5233	0.6389	0.5811
FSIM	0.6823	0.7318	0.7211	0.7503
VQM	0.7708	0.7563	0.7697	0.7893
MOVIE	0.8116	0.7890	0.7884	0.8061
STRRED	0.8038	0.8007	0.7894	0.8129
STMAD	0.8303	0.8251	0.7237	0.7355
ViS ₃	0.8336	0.8168	0.8222	0.8325
SSVQM	0.8632	0.8475	0.8099	0.8302

Table 2. Running time of SSVQM and other eight metrics

Metrics	Running time (s)
PSNR	1.3
SSIM	6.3
SSVQM	6.5
VQM	26
FSIM	53
STRRED	97
ViS ₃	182
STMAD	455
MOVIE	2470

4 Conclusions

In this paper, an effective and efficient full-reference video quality assessment algorithm has been proposed, called Spatio-temporal Structural-based Video Quality Metric (SSVQM). Spatio-temporal structure based local quality is calculated according to spatio-temporal gradient characteristics and chrominance information. Then, these local quality scores are integrated to yield an overall video quality via a spatio-temporal pooling strategy simulating three most important global temporal effects of human visual system, i.e. the smooth effect, asymmetric tracking effect. Experiments on LIVE and CSIQ VQA databases demonstrate that the proposed SSVQM can achieve highly competitive prediction accuracy and delivers very low computational complexity. In future work, we will focus on low-complexity video quality assessment of high-definition videos and stereoscopic videos.

Acknowledgments. The work was partly supported by the Natural Science Foundation of China (61671258,61871247).

References

1. Fan, Q., Luo, W., Xia, Y., Li, G., He, D.: Metrics and methods of video quality assessment: a brief review, *Multimedia Tools and Applications* (2017)
2. He, M., Jiang, G., Yu, M., Song, Y., et al.: Video quality assessment method motivated by human visual perception. *J. Electron. Imaging* **25**(6), 061613 (2016)
3. Huynh-Thu, Q., Ghanbari, M.: Scope of validity of PSNR in image/video quality assessment. *Electron. Lett.* **44**(13), 800–801 (2008)
4. Wang, Z., Bovik, A.C., Sheikh, H.R., et al.: Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* **13**(4), 600–612 (2004)
5. Zhang, L., Zhang, L., Mou, X., et al.: FSIM: a feature similarity index for image quality assessment. *IEEE Trans. Image Process.* **20**(8), 2378–2386 (2011)
6. Xue, W., Zhang, L., Mou, X., et al.: Gradient magnitude similarity deviation: a highly efficient perceptual image quality index. *IEEE Trans. Image Process.* **23**(2), 684–695 (2014)
7. Pinson, M.H., Wolf, S.: A new standardized method for objectively measuring video quality. *IEEE Trans. Broadcast.* **50**(3), 312–322 (2004)
8. Seshadrinathan, K., Bovik, A.C.: Motion tuned spatio-temporal quality assessment of natural videos. *IEEE Trans. Image Process.* **19**, 335–350 (2010)
9. Vu, P., Vu, C., Chandler, D.: A spatiotemporal most-apparent-distortion model for video quality assessment, *International Conference on Image Processing (ICIP)* (2011)
10. Vu, P., Chandler, D.: ViS3: an algorithm for video quality assessment via analysis of spatial and spatiotemporal slices. *J. Electron. Imaging* **23**(1), 013016 (2014)
11. Li, S., Ma, L., Ngan, K.N.: Full-reference video quality assessment by decoupling detail losses and additive impairments. *IEEE Trans. Circuits Syst. Video Technol.* **22**(7), 1100–1112 (2012)
12. Loh, W., Bong, D.B.L.: An error-based video quality assessment method with temporal information. *Multimedia Tools Appl.* **77**(23), 30791–30814 (2018)
13. Tan, K.T., Ghanbari, M., Pearson, D.E.: An objective measurement tool for MPEG video quality. *Sig. Process.* **70**(3), 279–294 (1998)
14. Horita, Y., Miyata, T., Gunawan, I.P., et al.: Evaluation model considering static-temporal quality degradation and human memory for SSCQE video quality, *Visual Communications and Image Processing*, pp. 1601–1611 (2003)
15. Hands, D.S., Avons, S.E.: Recency and duration neglect in subjective assessment of television picture quality. *Appl. Cogn. Psychol.* **15**(6), 639–657 (2001)
16. Masry, M., Hemami, S.S., Sermadevi, Y.: A scalable wavelet-based video distortion metric and applications. *IEEE Trans. Circuits Syst. Video Technol.* **16**(2), 260–273 (2006)
17. Seshadrinathan, K., Soundararajan, R., Bovik, A.C., et al.: Study of subjective and objective quality assessment of video. *IEEE Trans. Image Process.* **19**(6), 1427–1441 (2010)
18. Laboratory of Computational Perception & Image Quality, Oklahoma State University: CSIQ video database (2013). <http://vision.okstate.edu/csiq/>
19. VQEG, Final Report from the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment, Phase II, August 2003. <http://www.vqeg.org>
20. Soundararajan, R., Bovik, A.C.: Video quality assessment by reduced reference spatio-temporal entropic differencing. *IEEE Trans. Circuits Syst. Video Technol.* **23**(4), 684–694 (2013)