SPATIO-TEMPORAL SSIM INDEX FOR VIDEO QUALITY ASSESSMENT

Yue Wang[1]*,Tingting Jiang*[2]*, Siwei Ma*[2]*, Wen Gao*[2]

[1]Graduate University of Chinese Academy of Sciences, Beijing, China
[2] National Engineering Lab for Video Technology, Key Lab of Machine Perception(MoE),
School of EECS, Peking University，Beijing, China
wangyue@jdl.ac.cn; {ttjiang,swma,wgao}@pku.edu.cn

ABSTRACT

An ideal objective metric for video quality assessment (VQA) should achieve consistency between video distortion prediction and psychological perception of human visual system (HVS), and is important in many video processing applications. In general, both spatial distortion and temporal distortion should be carefully considered in the designing of VQA metrics. In this paper, we propose a novel spatio-temporal structural information based video quality metric. Motivated by the fact that pixels in natural videos are highly structured in both spatial domain and temporal domain, we propose to perform structural similarity evaluation in x-y, x-t and y-t dimensions respectively and pooled them adaptively based on local spatio-temporal activities. Experimental results on LIVE database show that such a conceptually simple and computationally efficient algorithm is competitive with state-of-the-art VQA metrics, and is very robust to various types of video distortions.

Index Terms—video quality assessment (VQA), human visual system (HVS), SSIM, temporal distortion

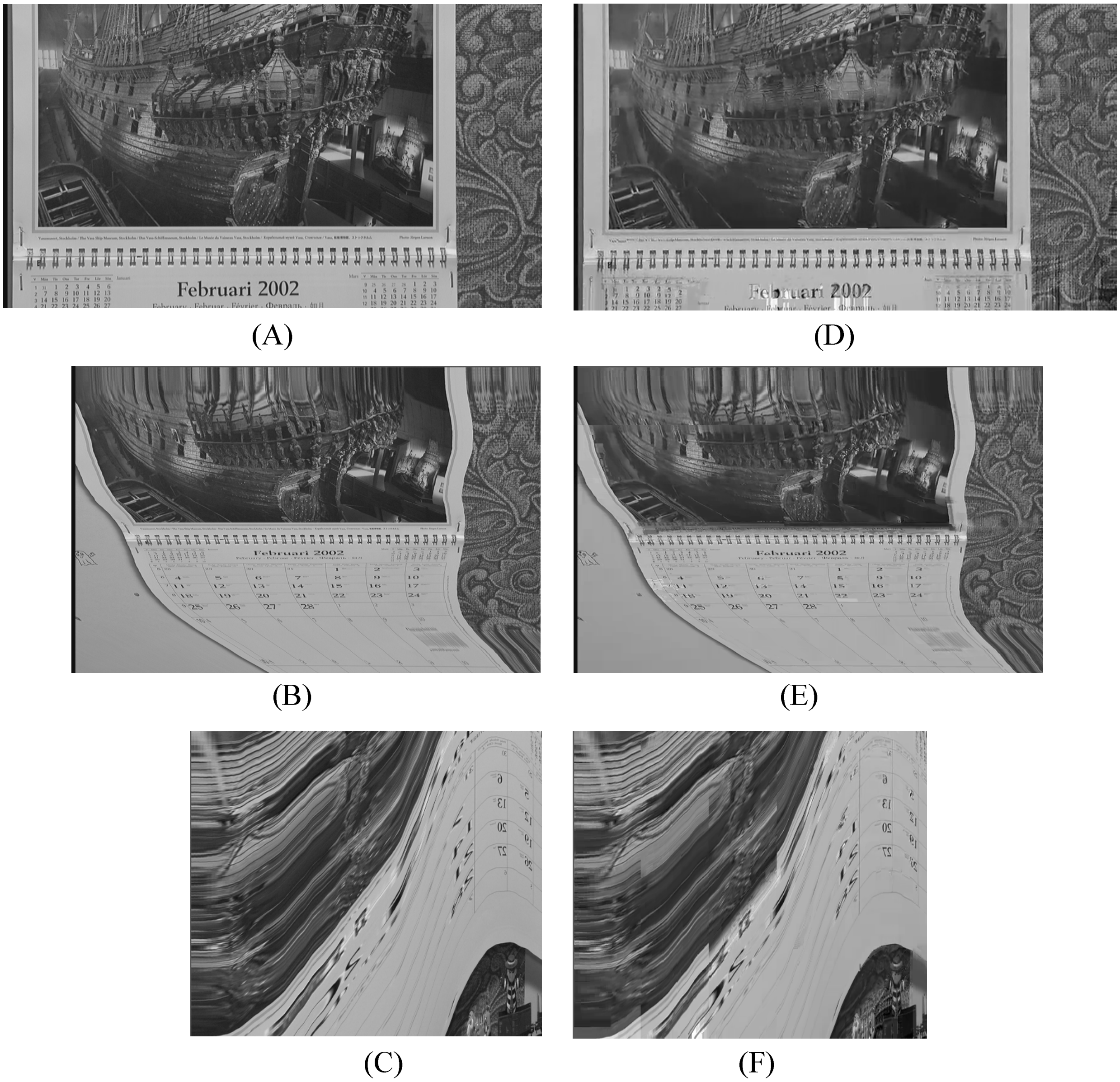
1. INTRODUCTION

Quality assessment of digital video resources is very important in the systems of video acquisition, compression, transmission and storage. The most straightforward way to evaluate the quality of a video is to use the quality scales directly rated by human observers. However, such subjective evaluations are quite time-consuming and expensive, and could not be applied in real-time scenarios and automatic systems. Therefore, there has been an increasing demand for objective quality metrics that are in agreement with the HVS' judgment. This work focuses on the full-reference (FR) objective video quality metrics, in which both the reference and distorted videos could be accessed.

The pixel-based FR metrics such as mean-squared-error (MSE) and related peak signal to noise ratio (PSNR) have been the dominant quantitative performance metrics in the field of signal processing for several decades since they are simple to calculate and have clear physical meanings in terms of Shannon information theory. However, it has been well acknowledged that these pixel-based signal fidelity metrics do not always correlate well with the HVS' perception [1]. Natural images are not random collections of pixels, but have strong statistical dependencies between the pixels. For the perception of the HVS, the visual information obtained from natural videos is not reflected in the individual pixels, but in the dependency between the pixels. Under the assumption that HVS is highly adapted to extract structural information from the viewing field, a new philosophy of SSIM for image quality assessment (IQA) was proposed by Wang *et al.* [2] and has drawn lots of attention from IQA researchers.

Different from the quality assessment for single images, motion information and temporal distortion should be carefully considered for video quality assessment (VQA). To extend the successful SSIM metric from images to videos, existing methods could be summarized to three categories. The first category is to calculate SSIM index on each individual frame and then summate all the frame scores to obtain a composite score. For example, to pool frame SSIM indexes to a video quality score, a motion-weighting model [3] is proposed to account for the fact that the accuracy of visual perception is significantly reduced when the speed of motion is large, and in [4], an alternate weighting scheme based on human perception of motion information is utilized. Although motion characteristic was more or less explored in these temporal weighting schemes, however, temporal distortions were not yet taken into account [5]. To account for the temporal distortion, the second category [6] proposed to perform motion estimation firstly and then calculate SSIM index between motion compensated blocks. In the third category [10], structural similarity analysis is performed in 3D video trunks using different filter directions, and the pooling weights for the different directions are tuned according to motion information.

Besides the structural information based methods, temporal distortion was modeled in different ways by other state of the art works. Video Quality Metric (VQM) proposed by NTIA [7] is a popular VQA metric and is included in the Recommendation ITU-T J.144 [8] as a normative FR VQA model. This metric extracts seven



(A) (D)

(B) (E)

(C) (F)

Fig.1. observing reference video and distorted video in three dimensions. (A),(B) and (C): reference video in x-y, x-t and y-t dimensions. (D),(E) and (F): distorted video in x-y, x-t and y-t dimensions

features from spatio-temporal blocks to compute the video distortion. Frame differences are embedded into one feature to account for the interaction between motion and spatial distortion. In [9], temporal distortion is defined as temporal evolution of the spatial distortion in a spatio-temporal “tube”, since the perception of spatial distortions over time can be largely modified by their temporal changes. Seshadrinathan and Bovik proposed a Motion-based Video Integrity Evaluation (MOVIE) in [5], where they defined the temporal distortion as the differences between the filter responses along computed motion trajectories.

In this paper, we propose a new methodology to account for temporal distortion. We evaluate SSIM for patches in x-t and y-t dimensions as the same way in x-y dimensions and SSIM indexes calculated from 3 dimensions are averaged to obtain a final video quality score.

Local quality scores are integrated to an overall score by a binary weighting scheme, in which the weighting factors are adapted to local spatio-temporal activities. This metric has a clear physical meaning in accordance with visual perceptions of the HVS and is quite computationally efficient. Experimental results demonstrate that the proposed metric’s performance is competitive with other state-of-the-art VQA metrics.

The remainder of the paper is organized as follows. In Section 2, we focus on the details of the proposed video quality metric. Simulation results are presented in Section 3. Finally, Section 4 concludes this paper.

2. PROPOSED VQA METRIC

As we know, the videos don't only provide us the spatial information about the natural scenes, but also the motion related temporal information. Local consistence of motion leads to highly structured feature along the temporal axis. Therefore, we extend the insights of structural similarity to spatio-temporal case.

2.1 Spatial SSIM

The SSIM metric evaluates image quality by using some low level structural information such as mean, variance, and covariance of intensity values of pixels in local patches. Let x and y denote two non-negative signals that have been aligned with each other (e.g., two image patches extracted from the same spatial location from two images being compared, respectively), and let μ_x, μ_y, σ_x^2, σ_y^2 and σ_{xy} represent the mean of x, the mean of y, the variance of x, the variance of y, and the covariance of x and y, respectively. The SSIM index between x and y is defined as [2]:

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

C_1 and C_2 are constants given by:

$$C_1=(K_1L)^2$$
$$C_2=(K_2L)^2$$

where L is the dynamic range of the pixel values (for 8 bits/pixel gray scale images, L = 255), and K_1, $K_2$are two small constants.

The SSIM index lie in the range of [-1, 1], and a higher value indicates a better quality.

Besides, gradient based SSIM index was proposed in [11] [12], which demonstrates to be more precise in predicting the quality of badly blurred images.

Gradient based SSIM is calculated as:

$$SSIM(x,y)=\frac{(2\mu_x\mu_y+C_1)(2\mu_{GxGy}+C_2)}{(\mu_x{}^2+\mu_y{}^2+C_1)(\mu^2{}_{Gx}+\mu^2{}_{Gy}+C_2)} \qquad (2)$$

where Gx and Gy are the gradient magnitude values in x and y directions respectively.

2.2 Temporal SSIM

Natural videos could be regarded as pixels arranged along two spatial and one temporal axis. Pixels have strong dependencies not only in spatial domain, but also in temporal domain. As illustrated in Fig.1, observing the pixel volume from different views, we could see that "images" in x-t dimension and y-t dimension exhibit highly structured characteristics as same as in x-y dimension. The distorted video also exhibits structural distortions such as blocky or blurring artifacts in x-t and y-t dimensions.

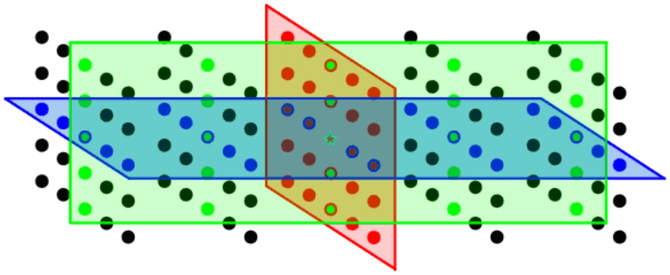


Fig.2. Illustration for calculating SSIM indexes in 3 dimensions

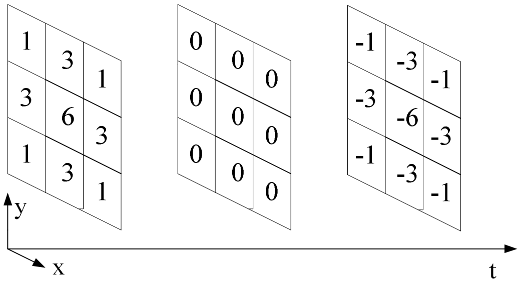


Fig.3. Sobel kernel for t direction

Accordingly, we propose to evaluate structural similarity for the patches in x-t and y-t dimensions as the same way in x-y dimension.

2.3 Pooling method

Pooling of the SSIM indexes are processed by 2 steps.

Since SSIM is evaluated in an overlapping fashion, there will be 3 patches in x-y, x-t and y-t dimensions centered on each pixel, which is illustrated in Fig.2. In the first step, for each pixel, we integrate the SSIM indexes calculated from 3 dimensions to one index by simply averaging:

$$SSIM_{xyt}=(SSIM_{xy}+SSIM_{xt}+SSIM_{yt})/3 \qquad (3)$$

In the second step, we need to pool all the local scores to a final video quality score. Recently, visual attention has been widely investigated in VQA studies [13]. Since human attention is not allocated equally to all regions in the visual field, but focused on certain regions known as salient regions [14], it is believed that distortion in these salient regions plays a crucial rule in the HVS' judgment on the overall quality of the video. Such an attention related mechanism is incorporated into our algorithm. We take a simple but efficient saliency detection method. Considering that HVS is sensitive to edges, motion regions and abruptly emerged artifacts, which usually have high spatial-temporal activities, we judge a pixel is a salient pixel if its spatio-temporal gradient magnitude is above a certain threshold ε in either original video or distorted video. To be specific, we utilize the 3D Sobel kernels to calculate the spatial and temporal gradients. Fig.3 shows the kernel for calculating the gradients along time direction. The kernels for x and y directions could be obtained by rotating this kernel by 90 degree along y axis and x axis respectively.

After the salient pixels have been selected, the local scores of the patches are pooled by a binary weighting scheme: we assign weighting factor of 1 to the patches centered on the salient pixels and 0 to the patches centered on the non-salient pixels. This means we can perform saliency detection firstly and then perform SSIM calculation only on the patches centered on the salient pixels. Discarding the non-salient pixels not only improves the accuracy of quality evaluation, but also saves up much computing time, which is quite important for a real-time VQA metric.

3. PERFORMANCE EVALUATION

3.1 Simulation Conditions

To evaluate the performance of the proposed VQA metric, we test it alongside other state-of-the-art VQA metrics on the LIVE Video Quality Database [15]. The LIVE Video Quality Database consists of 10 reference videos with 15 distortions each, to give a total of 150 distorted videos. Subjective scores (DMOS) were recorded. Test sequences in LIVE Video Quality Database are distorted by four different distortion processes - MPEG-2 compression, H.264 compression, and simulated transmission of H.264 compressed bitstreams through error-prone IP networks and through error-prone wireless networks.

For comparison, these same set of videos were evaluated by the following VQA metrics.

PSNR: the classic pixel-based VQA metric which is always used as baseline for performance evaluation of the VQA algorithms.

VQM: a widely used VQA metric proposed by NTIA [7], which was recommended by ITU J.144

SW-SSIM: frame based SSIM with motion associated weighting [4]

MOVIE: the representative optical flow based VQA metric proposed in [5]

MC-SSIM: motion compensated SSIM proposed in [6]

Additionally, we evaluate the database using the Tektronix PQA500 Picture Quality Analyzer, which is a leading video quality assessment product for industry application. Two indicators, namely PQR and DMOS exported by PQA500 are used for comparison.

As for performance criteria, Pearson correlation coefficient (CC) and Spearman rank order correlation coefficient (SROCC) are used as performance indicator. For the indicator of Pearson correlation coefficient, a nonlinear mapping between the objective scores and subjective quality ratings was used according to VQEG recommendations [16]. In this work, the mapping function chosen for regression for each of the metrics is a 4-parameter logistic function:

$$f(x)=\frac{\tau_1-\tau_2}{1+\exp(-\frac{x-\tau_3}{\tau_4})}+\tau_2 \tag{4}$$

3.2 Results and Discussions

Table 1 shows the overall results. It's impressive that the proposed metric with gradient SSIM [12] significantly outperforms other metrics on the LIVE database according to the two indicators, and is competitive with the MOVIE index. The PSNR performs especially poorly on this database. One important reason for the significant performance gap is that this database contains many spatio-temporally localized distortions that conventional metrics are incapable to account for. For example, packet loss will cause severe temporal distortions such as abrupt change or violent fluctuations along the motion trajectories. Compared to the blurring and blocky artifacts introduced by compression, the HVS is usually more intolerable to this type of distortion. However, the metrics based on spatial distortion could not accurately capture and characterize these local temporal distortions. Table 2 shows the results on each kind of distortion in LIVE database, which demonstrate that the proposed metric is rather robust to various types of video distortions.

Table 3 shows the performance comparison between SSIM indexes calculated in each dimension and integrated SSIM index (ε=1000). We could see that the performances of individual SSIM indexes are poorer than the integrated SSIM index, which powerfully demonstrates the importance of evaluating and integrating both spatial and temporal distortion.

Another advantage of the proposed algorithm is its low computational complexity. It can easily be shown that the time complexity of Sobel gradients computation used for saliency detection is $O(MN_{Sobel}^3)$, where M is the number of pixels in one frame and N_{Sobel} is the kernel size of the Sobel operator, which is 3 in our work. The time complexity of SSIM index calculation is $O(M'N_W^2)$, where M' is the number of selected saliency pixels in one frame, and N_W is the size of the window used for local structural information calculation, which is set as 7 in our implementation. As a result, the time complexity of the proposed algorithm is $O(MN_{Sobel}^3 + M'N_W^2)$.

In comparison, we will analyze the computational complexity of other optical flow based VQA algorithms. Generally, the most time-consuming part of these algorithms is the calculation of optical flow. Therefore, here we only take the complexity of optical flow calculation as a comparison. As we know, there are quite a lot of algorithms for optical flow calculation and algorithms with higher

Table 1. Performance Comparsion

Methods	Spearman CC	Pearson CC
MOVIE	0.786	0.810
VQM	0.702	0.723
SW-SSIM	0.585	0.596
MC-SSIM	0.679	0.698
PSNR	0.368	0.404
PQR(by PQA500)	0.695	0.712
DMOS(by PQA500)	0.695	0.711
Prop(standard SSIM)	0.717	0.726
Prop(gradient SSIM)	0.784	0.792

Table 2. Pearson CC scores of VQA metrics on each kind of distortion in LIVE database

Methods	Wireless	IP	H.264	MPEG2
PSNR	0.468	0.411	0.439	0.386
SW-SSIM	0.587	0.559	0.721	0.627
VQM	0.733	0.648	0.646	0.786
PQR(by PQA500)	0.646	0.730	0.746	0.646
DMOS(by PQA500)	0.643	0.730	0.743	0.645
Prop(standard SSIM)	0.693	0.681	0.796	0.621
Prop(gradient SSIM)	0.822	0.804	0.860	0.747

Table 3. Performance comparison between SSIM indexes in each dimensions and integrated SSIM index (Spearman CC).

	x-y	x-t	y-t	integrated
Standard SSIM	0.672	0.696	0.699	0.717
Gradient SSIM	0.612	0.759	0.759	0.784

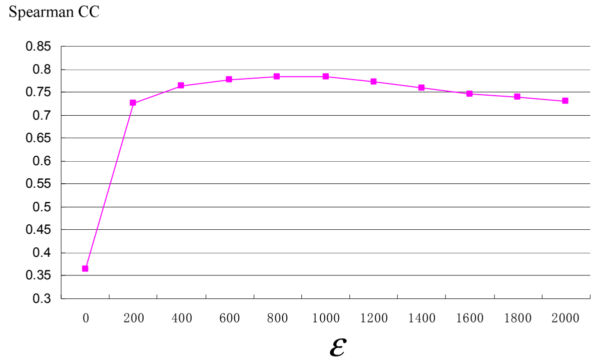


Fig.4. Performance of the metric versus different thresholds

accuracy usually needs higher computing cost. The MOVIE index [5] and the method in [10] utilized the phase based method [17] proposed by Fleet and Jepson, which is one of the most accurate method but with the highest computing cost. Time complexity of this method is $O(MV_{max}^3)$ [18], where $V_{max}$is the maximum motion velocity with the magnitude of dozens. The motion based SSIM [6] adopts the simplest block-matching based motion estimation method, whose complexity is $O(MV_{max}^2)$.

However, this method could not generate dense motion field and is unable to handle complex motion such as rotating and deforming. Another famous method is the gradient based LK method [19], which has a good tradeoff between performance and complexity. Its complexity is $O(M(n^2N_p+N_p^3))$ [20], where n is the size of a spatial template and N_p is the number of warping parameters. Usually n is set as 5, and 6-parameter affine model is used for warping. We can conclude that our algorithm has superiority in terms of computational complexity compared to other optical flow based VQA algorithms.

There is only one parameter that affect the performance and computational complexity of the algorithm: the saliency threshold ε. An appropriate threshold value can not only extract the HVS-sensitive regions of the video, but also maintain the computational complexity at a proper level. Fig.4 illustrates the performance versus different thresholds in terms of Spearman CC coefficient. We can see the metric (with gradient SSIM) gives the best performance when ε is set around 1000. When ε=0, all pixels are included in the evaluation, and the performance is quite poor. The accuracy of the metric will be also decrease as the threshold is larger than 1000, since some important video details may be missed. When ε=1000, the number of selected salient pixels is averagely reduced to 1/10 of original pixels in LIVE database. As a result, computing cost is largely cut down.

4. CONCLUSION

In this paper, we propose a novel structural information based video quality metric, in which spatial and temporal structural similarity are evaluated respectively and then pooled to a final video quality score. Experimental results on the LIVE database show that the proposed metric outperforms conventional quality metrics such as PSNR, VQM, VSSIM and performs competitively with MOVIE metric. Moreover, the experimental results also show that the proposed metric is rather robust to various types of video distortions and the performance is not parameter-dependent.

5. ACKNOWLEDGMENT

This work was supported in part by Major State Basic Research Development Program of China (973 Program, 2009CB320903), National Science Foundation (61121002, 61103088) and National High-tech R&D Program of China(863 Program, SS2012AA010805).

Table 4
Summary of time complexity of the proposed algorithm and algorithms based on optical flow

Algorithm	Proposed VQA algorithm	Optical flow algorithms		
		Phase based method	Block matching method	LK method
time complexity	$O(MN_{Sobel}^{3} +M'N_{W}^{2}))$	$O(MV_{max}^{3})$	$O(MV_{max}^{2})$	$O(M(n^{2}N_{p}+ N_{p}^{3}))$

6. REFERENCES

[1] A. M. Eskicioglu and P.S. Fisher, "Image quality measures and their performance," IEEE Trans. Communications, vol. 43, no. 12, pp. 2959–2965, Dec.1995.

[2] Z. Wang, A.C. Bovik, H.R. Sheikh and E. Simoncelli, "IQA: From error visibility to structural similarity", IEEE Trans. Image Process, vol. 13, no. 4, pp. 600-612, Apr. 2004.

[3] Z. Wang, L.G. Lu, and A.C. Bovik, "Video quality assessment based on structural distortion measurement," Signal Processing: Image Communication, vol. 19, no. 2, pp. 121–132, Jan. 2004.Z.

[4] Wang and Q. Li, "Video quality assessment using a statistical model of human visual speed perception," Journal of the Optical Society of America A, vol. 24, no 12, pp. B61–B69, 2007.

[5] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," IEEE Trans. Image Process., vol. 19, no. 2, pp. 335–350, Feb. 2010.

[6] Moorthy, A.K and Bovik, A.C. "Efficient Video Quality Assessment Along Temporal Trajectories," IEEE Transactions on Circuits and Systems for Video Technology, vol. 20, no. 11, pp. 1653-1658, Nov. 2010.

[7] M. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," IEEE Trans. on Broadcasting, vol. 50, no. 3, pp. 312–322, Sept. 2004.

[8] ITU-T Recommendation J.144, "Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference," Recommendations of the ITU, Telecommunication Standardization Sector

[9] A. Ninassi, O. L. Meur, P. L. Callet, and D. Barba, "Considering temporal variations of spatial visual distortions in video quality assessment," IEEE J. Sel. Top. Signal Process., vol. 3, no. 2, pp. 253–265, Apr. 2009.

[10] Moorthy, A.K and Bovik, A.C. ``Efficient Motion Weighted Spatio-Temporal Video SSIM Index''. SPIE Proceedings Human Vision and Electronic Imaging. January 2010

[11] Guan-Hao Chen, Chun-Ling Yang, Sheng-Li Xie," Gradient-based Structural Similarity for Image Quality Assessment", ICIP2006, 8-11, October, 2006, Atlanta, Georgia, U.S.A, pp: 2929-2932.

[12] Ming-Jun Chen, Alan C Bovik, "Fast structural similarity index algorithm", ICASSP 2010, 14-19 March 2010, Dallas, TX, U.S.A, pp: 994 - 997

[13] Lu Z., Lin W., Yang X., et al., "Modeling visual attention's modulatory aftereffects on visual sensitivity and quality evaluation, " IEEE Trans. Image Processing, vol. 14, no. 11, pp. 1928-1942, Nov. 2005.

[14] Itti L. and Koch C. "Computational modelling of visual attention", Nat. Rev. Neurosci., vol. 3, no. 2, pp. 194-203, Mar. 2001.

[15] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," IEEE Trans. Image Process., vol. 19, no. 6, pp. 1427–1441, Jun.2010.

[16] VQEG: The Video Quality Experts Group, http://www.its.bldrdoc.gov/vqeg

[17] Fleet, D.J. and Jepson, A.L., "Computation of Component Image Velocity from Local Phase Information", International Journal of Computer Vision, vol. 5, no.1, pp. 77-104, 1990.

[18] Hongche Liu, Tsai-Hong Hong, Martin Herman and Rama Chellappa. "Accuracy vs. Efficiency Trade-offs in Optical Flow Algorithms", In Proceedings of ECCV (2). pp.174~183 1996

[19] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In Proceedings of the International Joint Conference on Artificial Intelligence,pp.674–679, 1981.

[20] Simon Baker and Iain Matthews, "Lucas-Kanade 20 Years On: A Unifying Framework", International Journal of Computer Vision, Volume 56, Number 3, pp. 221-255