Spatio-temporal Salience Based Video Quality Assessment

Xinbo Gao [1], *Senior Member, IEEE*, Ni Liu[1], Wen Lu[1], Dacheng Tao[2], *Member, IEEE*, Xuelong Li[3], *Senior Member, IEEE*

1 VIPS Lab and Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education of China, School of Electronic Engineering, Xidian University, Xi'an 710071, China

2 School of Computer Engineering, Nanyang Technological University, Singapore 639798

3 Center for **OPT**ical **IM**agery **A**nalysis and **L**earning (OPTIMAL), State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, Shaanxi, P. R. China.

{xbgao, NLiu, luwen}@mail.xidian.edu.cn, dacheng.tao@gmail.com, xuelong_li@opt.ac.cn

***Abstract*— It is important to design an effective and efficient objective metric of the video quality in video processing areas. The most reliable way is subjective evaluation, thus the most reasonable objective metric should adequately consider characteristics of the *human visual system* (HVS). *Visual attention* (VA) is one of the essential visual phenomena of HVS, the realization of which relies on the saliency of visual field. Moreover, the saliency of visual field has a great influence on recognition, memorization and subjective evaluation of the image. This paper explores the saliency of visual field for objective quality assessment of videos. The proposed method first uses the VA model to obtain visual saliency map of the distorted video, including color, intensity and motion. Then the salient map is used to weight a structural similarity map between the original and distorted videos to get the final value of the video quality. Experimental results prove that the proposed method achieves a good correlation with subjective valuation.**

***Keywords*—Video quality assessment, human visual system, visual attention, spatio-temporal salience, motion**

I. INTRODUCTION

The digital videos play an important role in many practical applications, such as video conference, video on demand, and video surveillance, etc. However, the video data are often subject to various kinds of distortions during acquisition, compression, transmission, and reproduction such as mosquito noise, blocking artifacts, color bleeding, false contouring and pocket losses. So there is demand for a video system to control and quantify the degree of the video quality degradations. Hence, an effective video quality metric [1] is of great urgency for the aim.

Owning to human eyes are the ultimate receivers of a video sequence, the best way to evaluate the quality of a video sequence is by subjective experiment. However, it is time-consuming, cumbersome, and cannot be implemented in real-time systems. To resolve this dilemma, researchers have developed a series of objective quality assessment methods, which aim to evaluate the distortions of video sequences automatically and accurately.

PSNR is the most widely metric but it doesn't provide a satisfactory reflection of human visual perception. Recently, a great deal of effort has been made to the development of image and video quality assessment that takes advantage of the known characteristics of *human visual system* (HVS) [2][3]. According to the properties and cognitive mechanism of the human visual system, the approaches can be classified into two groups, namely top-down approach and bottom-up approach. 1) The top-down approach is primarily based on the extraction and analysis of certain features or artifacts in the video. Wang et al. [4] proposed video structural similarity (V-SSIM) index, the mean, variance and covariance of small image patch are computed to get a distortion map and motion estimation is used for a weighting of the SSIM index of each frame in the video. In Wolf's *video quality metric* (VQM) [5], a number of features measuring the amount of activity are computed from the spatial luminance gradient. The features extracted from test and reference videos are compared by a masking process. 2) The bottom-up approach simulates the function of each component of the HVS and combines them together. It tries to incorporate aspects of human vision relevant to image quality, such as color perception, contrast sensitivity and pattern masking, using models and data from psychophysical experiments. Winkler [6] proposed a method called *perceptual distortion metric* (PDM), which employs the steerable pyramid transform to emulate the multi-channel mechanism. Watson proposed the *digital video quality* (DVQ) [7], which is implemented to transform the signals into different spatial and temporal frequency in the DCT domain.

From the above discussion, we can see that various characteristics of HVS have been utilized in existing video quality metrics. In recent years, *visual attention* (VA) model for image quality assessment has been a research hotspot. Lots of literatures and psychophysics experiments have revealed that the realization of VA relies on the saliency of visual field. Moreover, the saliency of visual field has a great influence on recognition, memorization and subjective evaluation of the image. In [8], VA models find the salient regions and the resulting map is used to weight the distortion. The aim of the method is to evaluate the compression artifacts in coded images. In [9], the authors investigate two weighting methods, weighting the SSIM quality map with visual importance from a visual fixation predictor and assigning larger weighting value to pixels with lower SSIM scores. In contrast, work in VA related with video quality is less investigated. [10] proposes an embedded reference-free video quality metric. It

978-1-4244-6587-3/10/$25.00 É2010 IEEE

first estimates the perceptually important areas. For the salient areas detected, a distortion measure is then computed using a specialized metric. It outperforms the standard peak signal to noise ratio in evaluating the perceived video quality.

By intuition, visual attention aiming at the video is more complex than that of the image. Though similar to image, key elements such as color contrast, object size will attract the viewers' attention. The difference is that videos are comprised of numerous image frames. When videos are being played, the frames in a video stream into viewers' eyes one by one and viewers are natural to pay attention to the advent of new video contents. So, the image salient model cannot be applied directly to video quality assessment because image-oriented models of VA only consider spatial features while neglect temporal features which are important to video quality assessment, such as motion features. For applying visual attention to video quality assessment, there is need to find a salient model which can capture adequately and effectively the salient parts in a video. The Kr model is [11] saliency detection model of videos and incorporates color and motion apart from intensity. In this paper, we first employ Kr model to compute the salient areas of videos and then build interactions between SSIM index and visual salience to finish the complicated task of quality assessment of blocky videos.

The rest of this paper is organized as follows: in Section II, the theories of salience-based visual attention are provided. Section III presents the proposed metric. Experimental results are presented in Section IV, and Section V is the conclusion.

II. The Salience-based Visual Attention

Visual attention is an appearance that we often pay attention to one or some scene due to the demand for behavior purposes or local scene clues as the surrounding environment observed, so that some certain spots or area are selected as the representation of the scenery. James [12] originally proposed the idea of attention. The paradigm of computational visual attention has been widely investigated during the last two decades, and numerous computational models of visual attention have been suggested in computer vision. One of the first computational architecture for visual attention was proposed by Koch and Ullman [13]. Their idea is the salience-based visual attention and salience map. *Gaze attentive fixation finding engine* (GAFFE) [9] is a fixation finding algorithm for visual attention. It selects the center of an image as the first fixation, then foveates the image around this point. The foveate image is filtered to create a fixation map. But this model has some limitations. It starts searching fixations from the center of an image while the center of an image is not always the fixation; every fixation point is treated equally while it is not so in practical situation. Itti *et al.* develop a trainable model of bottom-up, saliency-based selective visual attention [14]. The first step is fast and parallel pre-attentive extraction of visual features (for orientation, intensity and color). The features are computed using linear filtering and center-surround structures. And then the spatial distribution of each saliency map is modeled with a dynamical neural network, in order to select locations of attention in the images. Itti's model has been widely used in the computer vision community including experimental proof of object detection and video compression and so on.

These models belong to image-oriented ones which only consider spatial information. These image-oriented salient models have been proved useful on image quality assessment. Such as shown in [9], by visual importance weighting of the computed fixations, better agreement with subjective scores can be produced for image quality metrics. However, it is well known that videos are more complicated than images and there is need to consider temporal features and motion information. So image-oriented salient model cannot be directly used to design video quality metrics and true video-oriented salient models which describe both spatial and temporal features are required to bring more benefits to video quality assessment.

Recently, the concept of saliency has recently begun to be extended to spatiotemporal counterpart. A dense spatiotemporal salient model [11] proposed by *Konstantinos Rapantzikos* (Kr) is one of the state of the art video-oriented salient models. The Kr model uses a multi-scale volumetric representation of the video and involves spatiotemporal operations at the voxel level. Saliency is computed by a global minimization process constrained by pure volumetric constraints, each of them being related to an informative visual aspect, namely spatial proximity, scale and feature similarity (intensity, color, motion). Points are selected as the extrema of the saliency response and prove to balance well between density and informativeness. Then there is an introduction to the video-oriented salient model which plays a positive role in improving the performance of video quality assessment algorithm.

Let V be a video volume including a set of consequent frames. $q=(x,y,t)$ is an individual space-time point and become the equivalent to a voxel in the volume. Let $V(q)$ be the value of volume V at q. Main steps as follows:

1) V is decomposed into three conspicuity volumes $C_i(i=1,2,3)$ corresponding to three different features, namely intensity, color and motion;

2) Each conspicuity volume is decomposed into multiple scales j. A set of $C=\{C_{i,j}\}$ are created with $i=1,2,3$, and $j=1,...L$ is representing a Gaussian volume pyramid;

3) Minimize an energy function E composed of a data term E_d and a smoothness term E_s (the details of the two terms can be found in [11]. Finally, a set of modified conspicuity multiscale volumes $\overline{C}=\{\overline{C}_{i,j}\}$ are obtained;

$$E(C)=\lambda_d.E_d(C)+\lambda_s.E_s(C) \quad . (1)$$

4) Saliency is computed as the average of all volumes across features:

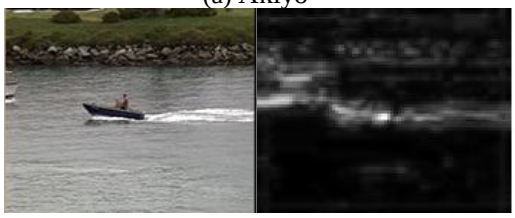
$$S=\{S_j\}=\frac{1}{3}.\sum_{i=1}^{3}\overline{C}_{i,j} \quad , (2)$$

where $j=1,...L$.

Figure 1 shows how the Kr model can capture the salience in a video in relevant with viewers' feelings. For Akiyo, most people pay more attention to the movements of eyes and

mouth of the woman. Fewer people will see the clothes she wears and even no people give an eye to the background. For Coastguard, the moving ship is the fixation according to the viewers. The salient maps in Figure 2 where more white means more evident are derived from Kr model. It can be found that it corresponds well with views' subjective feelings.

(a) Akiyo



(b) Coastguard

Fig. 1. Left to right: frame image and corresponding salient map of Akiyo and Coastguard

III. THE PROPOSED METHOD

Based on above analysis, how to assess video quality by salient model is the key in this paper. Obviously, if some parts in the frame are more salient, human eye will pay more attention to them, and then the salient areas may further deserve to be used to assess video quality. Therefore, salient areas have a louder voice when deciding the overall video quality. Based on above motivation, a spatio-temporal salience-based video quality assessment is proposed. The framework of salience-based video quality assessment can be shown in Figure 2.

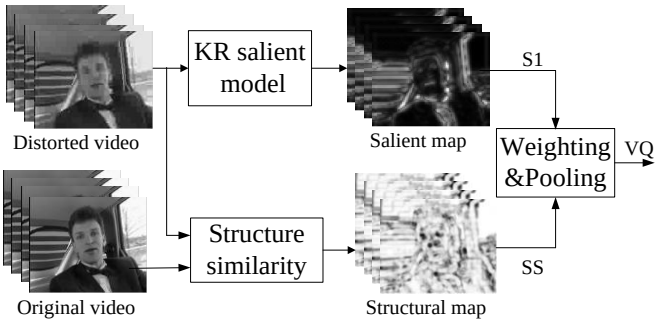


Fig. 2. The framework of salience-based video quality assessment

A. Structural map

Let V_o and V_d be the original and distorted video with a dimension of $M*N*F$. g and f represents one frame of V_o and V_d. To compute structural map SS_f of each distorted video frames, g and f are divided into a lot of small image patches namely x and y, then compute the SSIM index between x and y as following:

$$SSIM(x,y)=\frac{(2\mu_x\mu_y+C_1)(2\sigma_{xy}+C_2)}{(\mu_x^2+\mu_y^2+C_1)(\sigma_x^2+\sigma_y^2+C_2)} \quad , (3)$$

μ_x, μ_y and σ_x, σ_y is means and variances of x and y; σ_{xy} is the covariance between x and y; C_1, C_2 are constants. All SSIM indexes of small image patches in one frame comprise the structural map SS_f ($f=1...F$).

B. Salient map

The salience of distorted video V_d as section 2 described is computed. And the first scale is chosen as final salient map. In (2), let $j=1$, then

$$S1=\frac{1}{3}.\sum_{i=1}^{3}\overline{C}_{i,1} \quad , (4)$$

$S1$ also has F frames and each frame can be noted by $S1_f$.

C. Weighting and pooling

Then use the saliency map as the weighting factor to the structural similarity map of each video frames. The overall distortion of a video is the average of the saliency-weighted similarity map.

The quality value of each frame VQ_f is:

$$VQ_f=\sum_{m=1}^{M}\sum_{n=1}^{N}\frac{S1_f.*SS_f}{\sum_{m=1}^{M}\sum_{n=1}^{N}S1_f(m,n)} \quad . (5)$$

Finally, the quality value of the whole video VQ is computed as:

$$VQ=\sum_{f=1}^{F}VQ_f \quad . (6)$$

IV. EXPERIMENT RESULT

In this section, we design experiments to evaluate the performance of the proposed method. These experiments are based on the set of video test created by *video/image processing system labs* (VIPSL) [15] in Xidian University.

A. Database and Subjective Experiment

A total of 50 QCIF video sequences were generated from 10 reference sequences. They were subjected to H.264 video coding, with different bit rates (200kb/s to 16 Mb/s). Each of the video sequences consists of 300 frames. Figure 3 shows the original videos for testing.

Fig. 3. the original videos for testing

The subjective video quality test has been carried out for the evaluation of video sequences, which was performed by 20 subjects. Furthermore, we applied a series of metric during the subjective test, the experiment was organized in two phases: 1) the observers are asked to sort the videos from the test set according to their visual quality through a pair-wise comparison. To facilitate this process, a platform is designed as shown in Figure 4. In the platform, the original video is shown in the bottom and the two distorted ones to be compared are displayed on the top, and then the observer chooses the video that has less distortion with respect to the reference video. The observer has to choose until the platform shows ending, which means all videos have been sorted well; 2) the observers are asked to rate the annoyance of impairments in the test set using a continuous scale [0,100], where '100' corresponds to the highest quality and '0' to the worst quality.

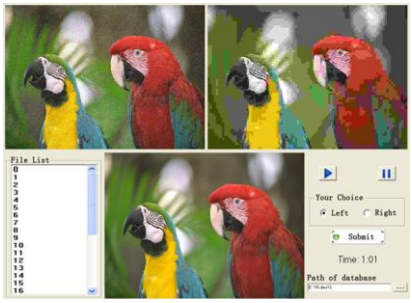


Fig. 4. The platform for subjective quality test.

Figure 5 shows the mean opinion scores of all the distorted videos. It can be seen that the quality of videos is widely dispersed. So the database can ensure the validity of evaluating quality methods.

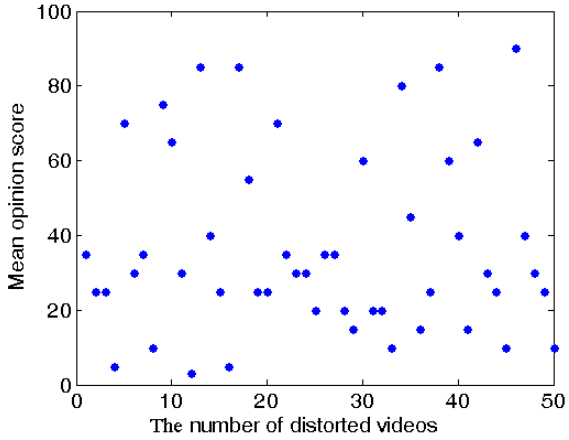
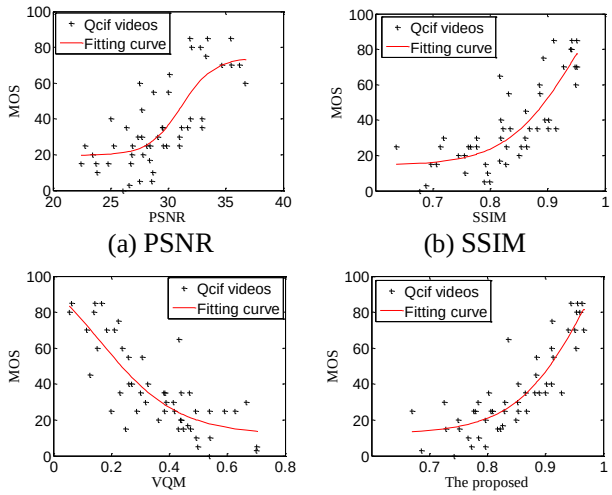


Fig. 5. The MOS for all distorted videos

B. *Consistency Experiment*

For the sake of examining the proposed method in the aspect of consistency with subjective perception, four measurements are used as evaluation criteria to test the consistency against subjective perception[16]: the Pearson linear *correlation coefficient* (CC) which presents the accuracy of objective method, the spearman *rank order correlation coefficient* (ROCC) which presents the monotonicity of objective method, the *mean absolute error* (MAE) and the *outlier ratio* (OR).

The performance of the proposed method is compared with some classical VQA methods shown in Table 1. From Table 1 the proposed method has greater improvement than the existing methods, so the objective assessment results have good consistency with subjective perception. Figure 6 presents the scatter plots of MOS versus the predicted score by PSNR [1], SSIM [4], VQM [5], and the proposed after the nonlinear regression.



(a) PSNR (b) SSIM

(c) VQM (d) the proposed

Fig. 6. Scatter plots of MOS versus different VQA methods

Table 1. The performance of PSNR, SSIM, VQM, and the proposed method.

MODEL	CC	ROCC	MAE	OR
PSNR	0.7628	0.6979	12.2196	0.6400
SSIM	0.8397	0.8071	10.4938	0.7000
VQM	0.8521	0.7616	10.4692	0.7000
Proposed	**0.8789**	**0.8596**	**9.0190**	**0.5800**

V. CONCLUSION

This paper describes a video quality assessment based on spatio-temporal salience. It employs the Kr model based on spatio-temporal salience to obtain the salient map of videos, and then computes the SSIM map of original video and distorted video. At last, salient map is used to weight SSIM map to acquire the final quality result. The experimental results show that the proposed method performs better than that of other VQA models. It has good consistency with subjective perception and can well reflect the visual quality of videos. However, the characteristics of visual attention deserve to be further investigated in the future to improve and simplify the video quality assessment model in both theory and practice.

ACKNOWLEDGMENT

This research was supported by the National Natural Science Foundation of China (60771068, 60702061, 60832005,

60902082), supported by the Ph.D. Programs Foundation of Ministry of Education of China (No. 20090203110002), the Natural Science Basic Research Plane in Shaanxi Province of China (2009JM8004), supported by the Natural Science Foundation in Shaanxi Province of China (2010JQ8022), the Open Project Program of the National Laboratory of Pattern Recognition (NLPR) in China and the National Laboratory of Automatic Target Recognition, Shenzhen University, China.

REFERENCES

[1] Wang, Z., Sheikh, H. R. and Bovik, A. C., "Objective video quality assessment," in The Handbook of Video Databases: Design and Applications, B. Furht and O. Marques, eds., CRC Press, Florida, pp. 1041-1078, 2003.

[2] Gao, X. B., Lu, W., Li, X. L. and Tao, D. C., "Image quality assessment based on multiscale geometric analysis," *IEEE Trans. Image Processing*, vol. 18, no. 7, pp. 1409-1423, 2009.

[3] Lu, W., Zeng, K., Tao, D. C., Yuan, Y. and Gao, X. B., "No-reference image quality assessment in contourlet domain," *Neurocomputing*, vol. 73, no. 4-6, pp. 784-794, 2010.

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

[5] Wolf, S. and Pinson, M. H., "Spatial-temporal distortion metrics for in-service quality monitoring of any digital video system," in *Proc. SPIE*, vol. 3845, pp. 266-277, 1999.

[6] Winkler, S., "A perceptual distortion metric for digital color video," in *Proc. SPIE*, vol. 3644, pp. 175-184, 1999.

[7] Watson, A. B., Hu, J. and McGowan, J. F., "DVQ: A digital video quality metric based on human vision," *Journal of Electronic Imaging*, vol. 10, no. 1, pp.20-29, 2001.

[8] Osberger, W., Maeder, A. J., and Mclean, D., "A computational model of the human visual system for image quality assessment," in *Proc. Digital Image Computing: Techniques and Applications*, pp. 337-342, 1997.

[9] Moorthy, A., K., and Bovik, A., C., "Visual Importance Pooling for Image Quality Assessment," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 2, pp. 193-201, 2009.

[10] Oprea, C., Prinog, I., Paleologu, C. and Udrea, M., "Perceptual Video Quality Assessment Based on Salient Region Detection," in *Fifth Advanced International Conf. on Telecommunications*, pp. 232-236, 2009.

[11] Rapantzikos, K., Avrithis, Y., and Kollias, S., "Dense saliency-based spatiotemporal feature points for action recognition," in *Conf. on Computer Vision and Pattern Recognition*, 2009.

[12] James, W., The Principles of Psychology. Cambridge, MA: Harvard University Press, 1890.

[13] Koch, C. and Ullman, S., "Shifts in selective visual attention: towards the underlying neural circuitry", *Human Neurobiology*, vol. 4, no. 4, pp. 219-227, 1985.

[14] Itti, L., Koch, C., and Niebur, E., "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254-1259, 1998.

[15] Video & Image Processing System lab (VIPSL), School of Electronic Engineering, Xidian University, Xi'an, China. Available: http://see.xidian.edu.cn/vipsl/index.html.

[16] Final report from the Video Quality Experts Group (VQEG) on the Validation of Objective Models of Video Quality Assessment, Phase II VQEG, 2003. [Online]. Available: http://www.vqeg.org/.