MOTION TRAJECTORY BASED VISUAL SALIENCY FOR VIDEO QUALITY ASSESSMENT

Lin Ma, Songnan Li, King N. Ngan
Department of Electronic Engineering, The Chinese University of Hong Kong
Email: {lma, snli, knngan}@ee.cuhk.edu.hk

ABSTRACT

In this paper, we propose a novel visual saliency detection method for video sequences by considering the object motion trajectories. Firstly, each frame of the video sequence is described in a new Quaternion Representation (QR), which comprises the spatial image content and the temporal motion characteristics. Based on the QR, Quaternion Fourier Transform (QFT) is employed to construct the visual saliency of the video sequence. Finally, the detected visual saliency map is incorporated with several video quality metrics. Compared with other visual saliency models, the proposed method can improve the performances of video quality metrics. It further confirms that the proposed visual saliency model can accurately depict the Human Vision System (HVS) properties.

Index Terms—Motion Trajectory, Visual Saliency, Video Quality Assessment (VQA), Human Visual System (HVS)

1. INTRODUCTION

Video quality assessment has been recently investigated due to its important role in video applications, such as video quality monitoring, streaming, transmission over the Internet, and video compression. The straightforward way for assessing the video quality is to conduct a large scale subjective study where a group of observers are asked to provide their personal opinions of the video on a particular scale. The observers' Mean Opinion Score (MOS) can be regarded as the true subjective quality values of the video sequences. However, the subjective experiment is very time-consuming and expensive. Therefore, the video quality metrics that can automatically evaluate the video perceptual quality are greatly demanded.

Recently, besides the Mean Square Error (MSE) and Peak Signal-to-Noise Ratio (PSNR), which have been verified to correlate poorly with the visual quality [1], many image/video quality metrics have been researched and developed [2]-[10]. Wang et al. develops the most popular Image Quality Assessment (IQA) Structural SIMilarity (SSIM) [2] that captures the structure information loss to depict the perceptual quality of the distorted image. Visual Information Fidelity (VIF) [3] employs the mutual information between the original image and the distorted one to evaluate the subjective quality. As SSIM and VIF fail to describe the temporal distortions of videos, several Video Quality Assessment (VQA) methods [4]-[8] have been developed. Video Quality Model (VQM) [6], adopted by the American national standards institute, analyzes 3D spatio-temporal blocks of the video sequence to extract the salient features for generating the video quality index. MOtion-based Video Integrity Evaluation (MOVIE) [5] is proposed to track perceptually relevant distortions along motion trajectories, and augment the measurement of spatial artifacts.

Since the Human Visual System (HVS) is the ultimate receiver of the images/videos, it is very important and advantageous to incorporate HVS properties into the IQA/VQA. In [9] [11]-[13], the HVS perceptual characteristics are depicted by the Just Noticeable Distortion (JND) model, which considers the visual contrast sensitivity, luminance adaptation, and video temporal characteristics. The JND has demonstrated good performances while incorporating with video quality metrics [14]. In [10], the authors model the HVS orientation properties by the horizontal effect, demonstrating that the HVS orientation preference can help improve the IQA performances. Among the HVS properties, the visual saliency is straightforward and extremely important for IQA, as revealed in [10]. Nowadays, many computational models [15]-[17] have been proposed to simulate human's visual attention. Itti et al. propose a bottom-up model and build a system named Neuromorphic Vision C++ Toolkit [17]. Hou et al. propose a Spectral Residual (SR) approach [16], which is proved to be useful for IQA [10]. However, SR only considers the spatial information for images. Guo et al. propose Phase Spectrum (PS) [15] for detecting the video saliency. Its temporal information is simply modeled by the frame differences. As claimed and verified in [7], the performances of VQAs can be improved by considering the distortions along the temporal trajectories. Therefore, we propose to incorporate the motion trajectory for efficiently detecting the visual saliency of video sequences. A Quaternion Representation (QR) for each frame is constructed, which comprises the spatial image content, the motion trajectories, and the temporal residuals. Based on the QR, the Quaternion Fourier Transform (QFT) is employed to construct the visual saliency. Finally, the visual saliency is incorporated with several video quality metrics for evaluating its efficiency.

This work was partially supported by a grant from the Chinese University of Hong Kong under the Focused Investment Scheme (Project 1903003).

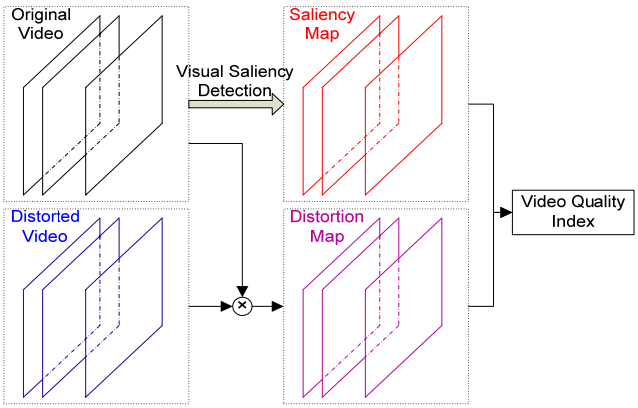

Figure 1. The VQA framework based on the proposed visual saliency

The rest of the paper is organized as follows. In Section 2, the proposed visual saliency model and its application on VQAs are introduced. Experimental results are demonstrated in Section 3. Finally, Section 4 concludes the paper.

2. THE PROPOSED MOTION TRAJACTORY BASED VISUAL SALIENCY FOR VQA

As illustrated in **Figure 1**, the proposed visual saliency model is applied on the original video sequences by considering both the image spatial content and the temporal motion trajectory. The distortion map is obtained by performing different VQAs, e.g. MSE, SSIM, on the original and distorted videos. Finally, by incorporating the saliency map with the distortion map, the video quality index of the distorted video is generated.

2.1 New Quaternion Representation (QR) for Each Frame

In order to apply the proposed visual saliency model, each frame of the original video sequence needs to be represented as a quaternion image [18]. It consists of four components, each of which captures the useful information from one certain aspect. As we only perform VQAs on the luminance part of the distorted videos, the chroma information is not accounted to construct the quaternion image. Define the video sequence as $V(t)$, t=1, 2, ..., N, where N is the total frame number. $l(t)$ denotes the luminance part of $V(t)$.

The Overlapped Block-based Motion Estimation (OBME) scheme is employed to depict the temporal motion trajectory. After the OBME, 3 temporal components of each frame are obtained. $MV_x(i,j)$ and $MV_y(i,j)$ denote the horizontal and vertical motion vector of the block centered at (i,j)-*th* pixel, respectively. $PE(i,j)$ indicates the corresponding motion prediction error. Together with the luminance $l(t)$, we have obtained the four components of the quaternion image. $l(t)$ represents the spatial image content. $MV_x(t)$ and $MV_y(t)$ describe the motion trajectory. $PE(t)$ depicts the temporal residual information, which compensates the inaccurate OBME. Each frame can be represented as the new quaternion image $q_i(t)$ [18] according to:

$$q_i(t) = l(t) + PE(t)\mu_1 + MV_x(t)\mu_2 + MV_y(t)\mu_3,$$

where

$$\begin{aligned} &\mu_i^2 = -1, \quad i = 1,2,3 \\ &\mu_1 \perp \mu_2, \mu_2 \perp \mu_3, \mu_3 \perp \mu_1. \\ &\mu_3 = \mu_1\mu_2 \end{aligned} \tag{1}$$

We can further represent $q_i(t)$ in a symplectic form:

$$\begin{aligned} q_i(t) &= f_1(t) + f_2(t)\mu_2 \\ f_1(t) &= l(t) + PE(t)\mu_1 \\ f_2(t) &= MV_x(t) + MV_y(t)\mu_1 \end{aligned} \quad . \tag{2}$$

In [15], the quaternion image comprises one intensity channel, two color channels, and one motion channel. However, the motion channel is simply described by the adjacent frame difference. On the contrary, our new quaternion image consists of one luminance channel, two motion vector channels depicting the temporal trajectory, and one temporal residual channel. With the consideration of the temporal trajectory, the visual saliency map can be faithfully reconstructed, which will benefit the VQAs.

2.2 From QR to the Saliency Map

As clarified in [15], only the phase spectrum is sufficient to represent the saliency information of each frame. Given an image $I(x,y)$,

$$\begin{aligned} f(x,y) &= F(I(x,y)) \\ p(x,y) &= P(f(x,y)) \\ SA(x,y) &= g(x,y) * \left\| F^{-1}(e^{i \cdot p(x,y)}) \right\|^2 \end{aligned}, \tag{3}$$

where F and F^{-1} denote the Fourier transform and inverse Fourier transform, respectively. $P(f)$ represents the phase spectrum of the image. $g(x,y)$ is a Gaussian filter. After the process in (3), the saliency map $SA(x,y)$ of $I(x,y)$ is generated.

For a quaternion image, the Quaternion Fourier Transform (QFT) [18] is employed to generate the visual saliency map. The QFT of a quaternion image $q(n,m)$ can be expressed as:

$$\begin{aligned} Q(u,v) &= F_1(u,v) + F_2(u,v)\mu_2 \\ F_i(u,v) &= \frac{1}{\sqrt{MN}} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} e^{-\mu_1 2\pi((mv/M)+(nu/N))} f_i(n,m) \end{aligned}, \tag{4}$$

where (n,m) and (u,v) are the locations of each pixel in time and frequency domain. N and M are the image height and width. $f_i, i \in \{1,2\}$ is obtained from (2).

The inverse QFT is defined as:

$$f_i(n,m) = \frac{1}{\sqrt{MN}} \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} e^{-\mu_1 2\pi((mv/M)+(nu/N))} F_i(u,v). \tag{5}$$

By applying (4), the frequency response $Q_i(t)$ of $q_i(t)$ can be obtained in the polar form as:

$$Q_i(t) = \left\| Q_i(t) \right\| e^{\mu \cdot p_i(t)}, \tag{6}$$

where $p_i(t)$ is the phase spectrum of $Q_i(t)$ and μ is a unit pure quaternion.

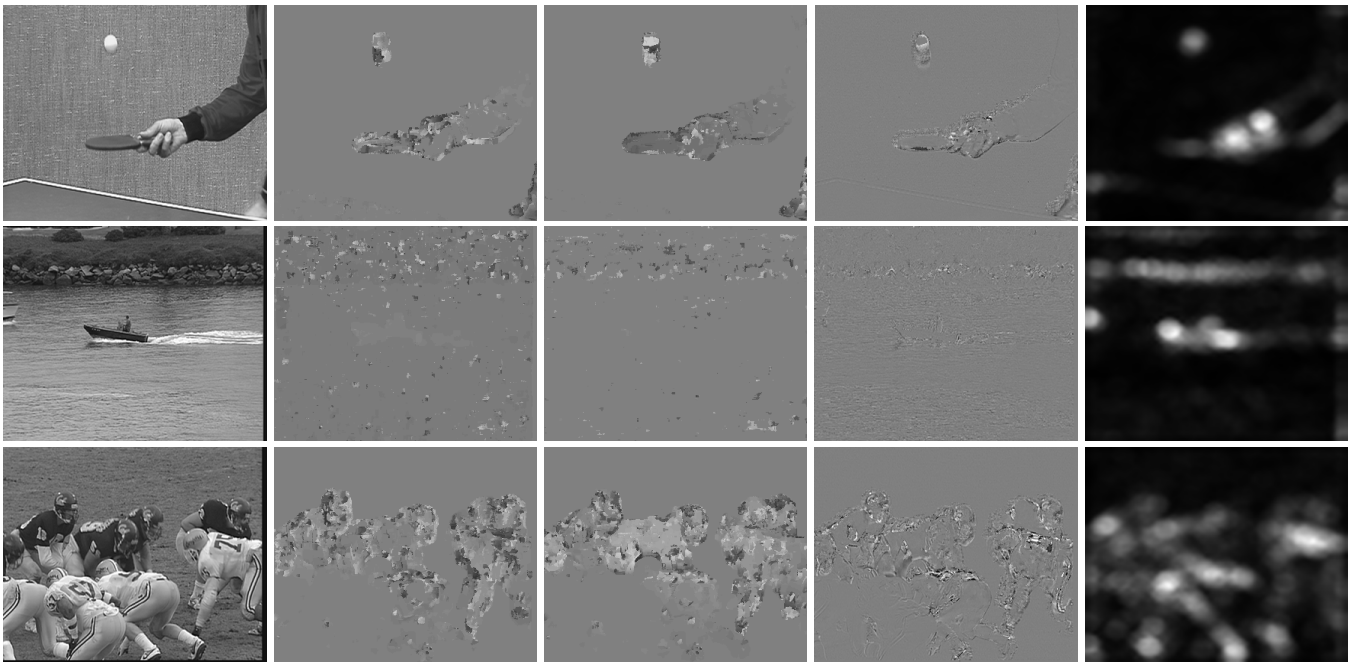

Figure 2. The Quaternion Representation (QR) of each frame and the visual saliency map. From left to right: luminance $l(t)$, horizontal motion vector $MV_x(t)$, vertical motion vector $MV_y(t)$, motion prediction error $PE(t)$, and the visual saliency map.

As shown in (3), only the phase spectrum is sufficient to construct the visual saliency map. Therefore, $\|Q_i(t)\|$ is set as 1. Then by applying the inverse QFT in (5), the reconstructed quaternion image q_i' is generated. Finally, the visual saliency map is constructed by the Gaussian filtering:

$$SA(t) = g * \left\| q_i'(t) \right\|^2 . \quad (7)$$

2.3 Incorporating Visual Saliency with VQAs

In this section, several VQAs, such as MSE, SSIM [2], MSSIM [19], and VABT-JND [14], incorporate with the detected visual saliency for improving their performances. For MSE and VABT-JND, the visual saliency map is employed to weight the calculated differences:

$$sDiff(t) = \left| Diff(t) \right| \cdot SA(t) , \quad (8)$$

where $Diff(t)$ denotes the differences between the original frame $O(t)$ and the distorted frame $D(t)$. For MSE, $Diff(t)=O(t)-D(t)$. As to VABT-JND, $Diff(t)$ indicates the difference after the JND masking process [14]:

$$Diff(t) = \left\{ O(t) - D(t) \right\} / T(t) , \quad (9)$$

where $T(t)$ is the ABT-JND threshold. The quality index for each frame is obtained by summing $sDiff(t)$ together:

$$Index(t) = 10 \log_{10} \left[mean\left(sDiff^2(t) \right) \right], \quad (10)$$

where $Index(t)$ is the perceptual quality index of each distorted frame.

As to SSIM, the visual saliency pooling strategy is performed over the structural distortion map, as demonstrated in [10]:

$$Index(t) = \sum SA(t) \cdot sM(t) / \sum SA(t) , \quad (11)$$

where $sM(t)$ is the structural distortion map obtained by applying SSIM:

$$sM(t) = SSIM(O(t), D(t)) . \quad (12)$$

MSSIM tries to apply SSIM over different scales of the image and sum the quality indexes together to evaluate the image quality. Adapting to this scheme, we down-sample the visual saliency map to different scales. By saliency pooling over different scales, the quality index of each frame is generated.

For each VQA, the quality index of each frame has been generated by considering the visual saliency map. Then the indexes are finally averaged to yield the Video Quality Index (VQI):

$$VQI = \sum_{t=1}^{N} Index(t) \Big/ N . \quad (13)$$

where N is the total frame number of the video sequence.

3. EXPERIMENTAL RESULTS

In this section, firstly we provide the processing results during the visual saliency detection, which is illustrated in **Figure 2**. As we have discussed in Section 2.1, each frame will be represented as a quaternion image, comprising luminance $l(t)$, horizontal and vertical motion vector $MV_x(t)$ and $MV_y(t)$, and prediction error $PE(t)$. For better visualization, MV is rescaled by $5 \times MV+128$; PE is rescaled by $PE+128$. It can be observed that the entire object generates nearly the same motion information, such as the ball, the boat, and the players in the video sequences. After performing the OBME, the prediction error is obtained. By incorporating the motion trajectory information (depicted by the motion vectors) and the temporal residual information, the visual saliency map for the corresponding frame is constructed using QFT, as shown in **Figure 2**. It can be observed that the visual saliency can significantly detect the motion object (highlighted

white) in the saliency map. By considering the accurate visual saliency map, the VQA performances can be significantly improved.

Table I. Performance comparisons between different VQAs

VQA Methods	LCC	SROCC	RMSE
MSE	0.5398	0.5234	9.241
SSIM	0.4999	0.5247	9.507
MSSIM	0.6754	0.7329	8.095
VABT-JND	0.7627	0.7372	7.099
SR-MSE	0.6164	0.6104	8.644
SR-SSIM	0.6215	0.6012	8.600
SR-MSSIM	0.7472	0.7360	7.296
SR-VABT-JND	0.7623	0.7322	7.105
PS-MSE	0.6230	0.6191	8.588
PS-SSIM	0.6051	0.5909	8.740
PS-MSSIM	0.7371	0.7245	7.419
PS-VABT-JND	0.7685	0.7338	7.023
VS-MSE	**0.6295**	**0.6268**	**8.531**
VS-SSIM	**0.6308**	**0.6187**	**8.518**
VS-MSSIM	**0.7583**	**0.7468**	**7.157**
VS-VABT-JND	**0.7768**	**0.7484**	**6.913**

We incorporate the detected saliency map with MSE, SSIM [2], MSSIM [19], and VABT-JND [14]. All of these VQAs are tested on the LIVE video database [22], which contains 150 distorted videos (obtained from 10 raw reference videos of natural scenes). The distorted videos are created using four different commonly encountered distortion types, including MPEG-2 compression, H.264 compression, wireless distortions, and IP distortions. We follow the performance evaluation procedure employed in the Video Quality Experts Group (VQEG) HDTV test [20] and that in [21]. A five parameter monotonic logistic function is employed for nonlinearly regression. After the nonlinearly mapping, the Pearson Linear Correlation Coefficients (LCC), Spearman Rank-Order Correlation Coefficients (SROCC), and Root Mean Square prediction Error (RMSE) are employed to evaluate different IQA performances. According to the definitions, larger LCC and SROCC values mean that the objective and subjective scores correlate better, and smaller RMSE indicates a better performance. The performances of VQAs incorporating different visual saliency models are shown in **Table I**, where SR denotes the saliency model in [16]; PS is the saliency model in [15]; VS is our proposed method. It can be observed that all the saliency weighted metrics can outperform the non-weighted metrics. It means that the visual saliency is important to HVS and helpful for the VQAs. Furthermore, VS weighted VQAs outperform the other saliency weighted VQAs. The reason is that the proposed method considers the motion trajectory, which is useful to improve the VQA performances, as demonstrated in [7]. However, the saliency weighted methods still perform inferiorly to MOVIE [5]. The reason is that MOVIE has employed complex HVS model for depicting the temporal and spatial distortions, compared to the proposed saliency weighting method. Another observation is that the improvement of VABT-JND is not so significant, compared with the other metrics. The reason is that ABT-based JND has considered some HVS properties, such as contrast masking, which has somehow modeled the HVS saliency property.

4. CONCLUSION

In this paper, we propose a new quaternion representation for each frame of the video sequence. Then the quaternion image is employed to generate the corresponding visual saliency map. By incorporating the visual saliency map with different VQAs, the metric performances can be significantly improved, which further confirms that the proposed method can accurately model the HVS saliency property.

5. REFERENCES

[1] B. Girod, "What's Wrong with Mean Squared Error", *Digital Images and Human Vision*, MIT Press, 1993, pp. 207-220.

[2] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity", *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600-612, 2004.

[3] H. R. Sheikh, and A. C. Bovik, "Image Information and Visual Quality", *IEEE Trans. Image Process.*, vol. 15, pp. 430-444, Feb. 2006.

[4] Z. Wang, L. Lu, and A. C. Bovik, "Video Quality Assessment Based on Structural Distortion Measurement", *Signal Processing. Image Communication*, vol. 19, no. 2, pp. 121-132, Feb. 2004.

[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] M. H. Pinson, and S. Wolf, "A New Standardized Method for Objectively Measuring Video Quality", *IEEE Trans. Broadcast.*, vol. 50, no. 3, pp. 312-322, Sep. 2004.

[7] A. K. Moorthy, and A. C. Bovik, "Efficient Video Quality Assessment Along Temporal Trajectories", *IEEE Trans. Circuits Syst. Technol.*, vol. 20, no. 11, pp. 1653-1658, Nov. 2010.

[8] M. Masry, S. S. Hemami, and Y. Sermadevi, "A Scalable Wavelet-Based Video Distortion Metric and Application", *IEEE Trans. Circuits Syst. Technol.*, vol. 16, no. 2, pp. 260-273, 2006.

[9] W. Lin, L. Dong, and P. Xue, "Visual Distortion Gauge Based on Discrimination of Noticeable Contrast Changes", *IEEE Trans. Circuits Syst. Technol.*, vol. 15, pp. 900-909, Jul. 2005.

[10] L. Ma, et al., "Visual Horizontal Effect for Image Quality Assessment", *IEEE Signal Process. Letters*, vol. 17, pp. 627-630, Jul. 2010.

[11] Z. Wei, and K. N. Ngan, "Spatio-Temporal Just Noticeable Distortion Profile for Grey Scale Images/Videos in DCT Domain", *IEEE Trans. Circuits Syst. Technol.*, vol. 19, no. 3, pp. 337-346, Mar. 2009.

[12] Z. Lu, et al., "Modeling Visual Attention's Modulatory Aftereffects on Visual Sensitivity and Quality Evaluation", *IEEE Trans. Image Process.*, vol. 14, no. 11, pp. 1928-1942, 2005.

[13] X. Yang, et al., "Motion-Compensated Residue Pre-processing in Video Coding Based on Just-Noticeable-Distortion Profile", *IEEE Trans. Circuits Syst. Technol.*, vol. 15, pp. 742-750, 2005.

[14] L. Ma, et al., "Video Quality Assessment Based on Adaptive Block-Size Transform Just-Noticeable Difference Model", *Proc. ICIP* 2010.

[15] C. Guo, et al., "Spatio-temporal Saliency Detection Using Phase Spectrum of Quaternion Fourier Transform", *Proc. CVPR,* 2008.

[16] X. Hou, and L. Zhang, "Saliency Detection: A Spectral Residual Approach", *Proc. CVPR,* 2007.

[17] L. Itti, C. Koch, E. Niebur, et al. "A Model of Saliency Based Visual Attention for Rapid Scene Analysis", *IEEE Trans. Pattern Anal. Mach. Intelli.* vol. 20, pp. 1254-1259, 1998.

[18] T. Ell, et al., "Hypercomplex Fourier Transforms for Color Images", *IEEE Trans. Image Process.*, vol. 16, no. 1, pp. 22-35, Jan. 2007.

[19] Z. Wang, et al., "Multiscale Structure Similarity for Image Quality Assessment", *IEEE Asilomar Conf. Signals, System and Computers*, 2003.

[20] VQEG, Final report from the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment. [Online] Available: http://ww.vqeg.org.

[21] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A Statistical Evaluation of Recent Full Reference Image Quality Assessment Algorithms", *IEEE Trans. Image Process.*, vol. 15, pp. 3441-3452, Jan. 2006.

[22] K. Seshadrinathan, R.Soundararajan, A. C. Bovik, and L. K. Cormack, "A Study of Subjective and Objective Quality Assessment of Video", *IEEE Trans. Image Process.*, vol. 19, pp. 1427-1441, 2010.