PERCEPTUAL-QUALITY SIGNIFICANCE MAP (PQSM) AND ITS APPLICATION ON VIDEO QUALITY DISTORTION METRICS

Zhongkang Lu, Weisi Lin, Eeping Ong, Susu Yao and X. K. Yang[*]

Laboratories for Information Technology,
Agency for Science, Technology and Research,
21 Heng Mui Keng Terrace, Singapore 119613
Email: {zklu, wslin, epong, ssyao, xkyang}@lit.a-star.edu.sg

ABSTRACT

This paper presents a new and general concept, PQSM (Perceptual Quality Significance Map), to be used in measuring the visual distortion. It makes use of the mechanism that the HVS (Human Visual System) pays more attention to certain areas of visual signal due to one or more of the following factors: salient features in image/video, cues from domain knowledge, and association of other media (e.g., speech or audio). PQSM is a 3D/4D array whose elements represent the relative perceptual-quality significance levels for the corresponding pixels/regions for images or video. Due to its generality, PQSM can be incorporated into any visual distortion metrics: to improve effectiveness or/and efficiency of perceptual metrics; or even to enhance a PSNR-based metric. A three-stage PQSM generation method is also proposed in this paper, with an implementation of motion, luminance, skin-color and face mapping. Experimental results show the scheme can significantly improve the performance of current image/video distortion metrics.

1. INTRODUCTION

Visual distortion metrics play an important role on broadcasting quality monitoring, compression controlling and gauging many image enhancement processes. There are generally two classes of quality or distortion assessment approaches. The first class is based on mathematically defined measures, such as the widely used MSE (mean square error), PSNR (peak noise to signal ratio), etc. Recently, Z. Wang *et al* [1] presented a new method that combines three factors (i.e., loss of correlation, luminance distortion and contrast distortion) to measure distortion of image. The second class is to measure the distortion by simulating the human visual system (HVS) characteristics [2, 3]. The advantage of the first class approaches is mathematically simple and low computational complexity, and therefore they have been widely adopted. The second class of approaches aims at perception results closer to human vision and hence can lead to better methodology in visual assessment and information processing (especially compression and coding). However, because of the incomplete understanding on the HVS and the lag in incorporating the physiological/psychological findings, the performance of the second class of metrics still cannot reach our expectation [4].

Since the aim of visual quality evaluation is to make the machine perceive as the human being does, it is always desirable for new discovery and observation regarding human perception to be incorporated. In this paper, we will explore possibilities when the selectivity on visual contents exhibited in human vision is considered.

Human vision has excellent selectivity on what one sees in a scene, as the result of evolutional process. There have been physiological and psychological evidence that the HVS does not pay equal attention to all visual information it is exposed to and simply focuses on certain areas on image. Visual attention is an important mechanism in the HVS that is believed to cause eye (fovea) movement after some transient. It has been explored by a large number of bioneuroscientists and psychologists for more than one hundred years, and refers to deciding the saliency among objects in visual signal [5, 6].

Computationally, visual attention can be defined as *a set of strategies that attempts to reduce the computational cost of the search processes inherent in visual perception* [6]. The mechanism can be modelled by *bottom-up*, image feature based stimuli or *top-down*, task/knowledge based cues.

For a *bottom-up* process, levels of saliency are decided by visual features/stimuli (e.g., illumination, color, motion, shape) mainly from visual data themselves [5] to form a *saliency map* [7]. The compound effect of various stimuli can be a nonlinear sum of all stimuli in different domains, and the additivity has been studied by subjective tests and modelled mathematically for pairwise combination of orientation, motion, luminance and color contrast [8]. For the *top-down*, or task/knowledge based process, significance for

[*] Thanks to Prof. T. Kanade, Dr. T. Moriyama and Dr. H. A. Rowley for providing face detection packages.

pixels/regions can be defined with domain/prior knowledge or/and indication from other media (like speech and audio), together with proper processing of visual data.

In this paper, a general framework will be proposed for assigning a visual importance measure for each pixel or region in a frame of video based upon the HVS' selectivity on visual contents, in order to form the Perceptual Quality Significance Map (PQSM) that would guide a perceptual distortion evaluation process for closer match with that of the HVS. In Section 2 of this paper, the concept of PQSM and its possible inclusion in various perceptual metrics is discussed. The experimental results for incorporating PQSM with a few exemplary quality metrics are presented in Section 3 to demonstrate the effectiveness of the scheme. Section 4 gives concluding remarks and possible future work.

2. PQSM AND ITS INCLUSION IN METRICS

PQSM is a 3D/4D array whose elements represent the relative perceptual-quality significance levels for the corresponding pixels/regions for images or video. The general block-diagram of a three-stage computational model of PQSM estimation is shown in Figure 1. In the first Stage, visually significant stimuli/features are extracted from input video video. The second Stage is the decision maker to integrate the extracted features based on prior knowledge to generate a gross PQSM. Post-processing is to enhance the gross PQSM with Gaussian Smoothing to remove impulse noise caused by errors in feature extraction.

Feature extraction and stimulus integration will be discussed in more details in the following subsections since they are critical in PQSM generation.

2.1. Feature Extraction

Figure 2 shows the current implementation of the first stage based upon extraction of luminance, motion and skin color in video.

Motion feature extraction includes two steps: global motion estimation, and motion mapping. Relative motion vectors are obtained by Heuer's 4-parameter global motion estimation method [9] combined with Lucas-Kanade's optical flow technique [10]. Two motion vectors can be obtained for an object in a video: *relative motion* vector and *absolute motion* vector. The former is the object motion vector against background or other objects in the scene while the latter is the object motion against the view frame. They have different effect on attention and perceptual quality, as shown in Table 1.

High *absolute motion* usually attracts the HVS's attention, but volition of the human brain may not be fast enough [11] to figure out the detailed information (such as exact shape and color changes) of the object. In the example mentioned in Introduction, human eyes will follow the movement of the car, but do not pay attention to the distortion of the car too much. Only when the *absolute motion* is low and *relative motion* is high (e.g., camera is following the object), the HVS cares more about the quality of the object. It can be seen that attention and perceptual quality requirement is not always the same with motion stimuli.

Table 2 gives an example on how the concept in Table 1 is linked to numerical range of RM ν_r and AM ν_a that are detected via Lucas-Kanade's algorithm (inclusive of a temporal-spatial smoothing pre-processing) for video of framerate at 25/30 Hz.

A statistical model is adopted to detect regions with skin color on $Cb - Cr$ domain in video. Based on the skin-color model, Rowley's face detection algorithm [12] is used to locate faces in image. The result of skin color detection is used to re-correct the face detection result because Rowley's algorithm is based on gray-scale image. A simple luminance mapping is also used since too bright or dark regions of video do not draw as much attention as other regions.

2.2. Stimulus Integration

Nothdurft's research results suggest that saliency effects in different stimulus dimensions add but do not add linearly [8]. In his model, the saliency effect with two dimensions can be expressed as:

$$s_{12} = s_1 + s_2 - s^*_{12} \tag{1}$$

and

$$s^*_{12} = min(c_{12} \cdot s_1, c_{21} \cdot s_2) \tag{2}$$

where s_1 and s_2 are the salience effects of two individual stimuli; s_{12} is their integrated effect; s^*_{12} is their coupling effect; c_{12} and c_{21} represent the cross-dimensional activation rates, or coupling factors between s_1 and s_2.

In this paper, Equation 1 is extended to three or more stimuli:

$$s_{all} = \sum_{i}^{N} s_i - \sum_{i}^{N} f(c_{ip} \cdot s_p, c_{pi} \cdot s_i) \tag{3}$$

where $p = \arg\max_i(s_i)$, $i = 1, \cdots, N$, and $f(\cdot,\cdot)$ is an appropriate nonlinear function. In this paper, $f() = min()$ is chosen. It worth noting that only coupling between the main and other stimuli are considered. Compare with other integration method like neural network and fuzzy rules, the advantages of this method are its simpleness and training-free.

The coupling factors are assigned based upon the conclusion from Nothdurft [8] as well as our experiments. Assume $i = 1, 2, 3$ and 4 for luminance, motion, skin color and face detection, respectively. Coupling factors used in this paper are: $c_{12} = c_{21} = 0.5$, $c_{13} = c_{31} = 0.5$, $c_{14} =$

$c_{41} = 0.5$, $c_{23} = c_{32} = 0.1$, $c_{24} = c_{42} = 0.1$ and $c_{34} = c_{43} = 1.0$.

Figure 3 shows a PQSM-modified perceptual video quality model when the PQSM weighting is inserted in the generalized block-diagram for Winkler's metric [2]. Channel decomposition is performed by spatio-spatial filterbands [2] or temporal filters plus Discrete Cosine Transform [3]. PQSM weighting is done by multiplying both the original signal and distorted signal at each spatio-spatial channel with the corresponding scale.

In the metric proposed by Z. Wang *et al* [1] for images, a distortion image $q(i,j)$ of size $(W-7) \times (H-7)$ is formed by evaluating loss of correlation, luminance distortion and contrast distortion. Therefore the PQSM-modified distortion measure is

$$q^{PQSM} = \frac{1}{(W-7)(H-7)} \sum_{i}^{W-7} \sum_{j}^{H-7} m_{i,j} \cdot q(i,j) \qquad (4)$$

where $m_{i,j}$ donates the PQSM component. Following the notation in Figure 3, the PQSM-modified MSE measure is

$$mse^{PQSM} = \frac{1}{WH} \sum_{i}^{W} \sum_{j}^{H} m_{i,j} \cdot \big(o(i,j) - d(i,j)\big)^2 \qquad (5)$$

and PQSM-modified PSNR measure is

$$psnr^{PQSM} = 10lg \frac{255}{mse^{PQSM}} \qquad (6)$$

3. EXPERIMENTAL RESULTS

Experiments is executed to compare the performance between current VQMs and PSQM-modified VQMs. PSNR, Wang's [1] and Winkler's metrics [2] are used. Two VQEG testing video sequences are used for experiments, they are 'harp' and 'waterfall'.

Evaluation uses VQEG's method [13]. The experimental results is listed in table 3. We can see that PQSM-modified VQM bring a performance improvement, except Wang's VQM on 'Autumn_leaves'. The reason is that: 1) Spearman correlation value on 'waterfall' is very high; 2) Subjective rating exist variations.

4. CONCLUSION AND FUTURE WORK

In this paper, a three-step model for PQSM estimation is presented. The model includes the last development of visual attention in bioneuro-science and psychological studies. Its application on measuring perceptual distortion of compressed video sequences are then discussed. PQSM can be incorporated into any visual distortion metrics (both objective-error or perception based ones). Experimental results prove that the PQSM (derived from luminance, motion, skin-color and face detection in current phase) as an independent module in visual quality metrics improves the performance of such metrics.

Table 1. Relationship among relative motion (RM), absolute motion (AM), attention and perceptual quality significance level (PQSL).

RM	AM	Attention	PQSL
low	low	low	low
high	low	high	high
low	high	low	low
high	high	high	medium

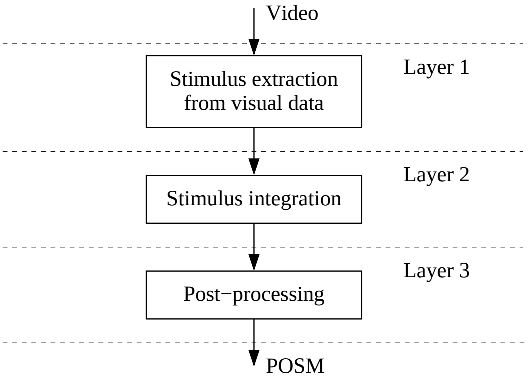

Fig. 1. General Block-diagram of computational model of pre-selective visual attention.

Embedded criteria can be also developed with PQSM for visual compression and coding. Other applications may be found in video indexing and retrieval, data hiding, and Video Surveillance.

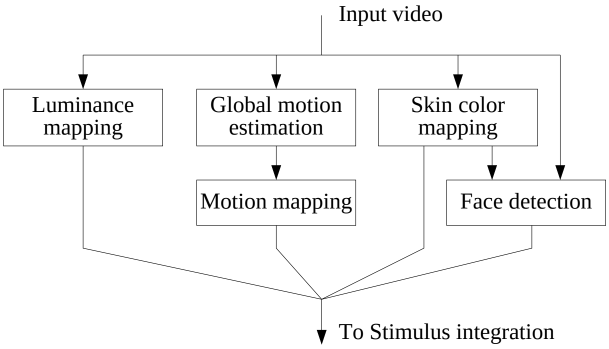

Fig. 2. Current Implementation of Feature Extraction.

Table 2. Motion mapping table.

Motion value	$\nu_r \in [0.0, 1.0)$	[1.0,3.0)	[3.0,5.0)	[5.0,7.0)	[7.0,9.0)	[9.0,∞)
$\nu_a \in [0.0, 1.0)$	0.0	0.4	0.7	0.9	0.9	1.0
[1.0,3.0)	0.0	0.4	0.7	0.9	0.9	1.0
[3.0,5.0)	0.0	0.4	0.7	0.9	0.9	1.0
[5.0,7.0)	0.0	0.3	0.6	0.8	0.8	0.9
[7.0,9.0)	0.0	0.3	0.5	0.7	0.7	0.7
[9.0,∞)	0.0	0.3	0.5	0.5	0.5	0.5

Table 3. Comparison between current VQMs and PQSM-modified VQMs (expressed as V-VQM) by Spearman correlation.

VQM	PSNR	P-PSNR	Wang's	P-Wang's	Winkler's	P-Winkler's
'Harp'	0.8118	0.85	0.6706	0.6853	0.6912	0.7412
'Autumn_leaves'	0.1324	0.5441	0.9324	0.9265	0.8235	0.8647

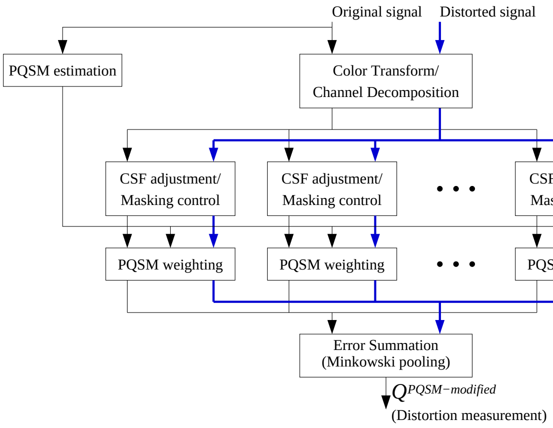
Original signal
Distorted signal
PQSM estimation
Color Transform/ Channel Decomposition
CSF adjustment/ Masking control
CSF adjustment/ Masking control
CSF adjustment/ Masking control
PQSM weighting
PQSM weighting
PQSM weighting
Error Summation (Minkowski pooling)
$Q^{PQSM-modified}$
(Distortion measurement)

Fig. 3. PQSM-modified Perceptual Model for Visual Quality Measurement (Winkler [2], Watson [3]).

5. REFERENCES

[1] Zhou Wang and Alan C. Bovik, "A universal image quality index," *IEEE Signal Processing Letters*, vol. 9, no. 3, pp. 81–84, March 2002.

[2] S. Winkler, *Vision models and quality metrics for image processing applications*, Ph.D. thesis, Ecole Polytecnique Federale De Lausanne (EPFL), Swiss Federal Institute of Technology, Thesis No. 2313, Lausanne, Switzerland, December 2000.

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

[4] Alan C. Bovik Zhou Wang and Ligang Lu, "Why is image quality assessment so difficult?," *Proc. IEEE International Conference on Acoustics, Speech, & Signal Processing 2002.*, May 2002.

[5] L. Itti, "Visual attention," in *The Handbook of Brain Theory and Neural Networks, 2nd Ed.*, M. A. Arbib, Ed. MIT Press, in press.

[6] J. K. Tsotsos, "Motion understanding: Task-directed attention and representations that like perception with action," *International Journal of Computer Vision*, vol. 45, no. 3, pp. 265–280, December 2001.

[7] L. Itti, "Real-time high-performance attention focusing in outdoors color video streams," in *Proc. SPIE Human Vision and Electronic Imaging IV (HVEI'02), San Jose, CA*, in press.

[8] H.-C. Nothdurft, "Salience from feature contrast: additivity across dimensions," *Vision Research*, vol. 40, no. 10-12, pp. 1183–1201, June 2000.

[9] J. Heuer and A. Kaup, "Global motion estimation in image sequences using robust motion vector field segmentation," in *Proceedings ACM Multimedia 99*, November 1999.

[10] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. DARPA IU Workshop*, 1981.

[11] J. M. Wolfe, G. A. Alvaraz, and T. S. Horowitz, "Attention is fast but volition is slow," *Nature*, vol. 406, no. 691, 2000.

[12] H. A. Rowley, "Neural network-based face detection," *Ieee Transactions On Pattern Analysis And Machine Intelligence*, vol. 20, no. 1, pp. 23–38, Jan. 1998.

[13] VQEG (Video Quality Experts Group), "Vqeg subjective test plan," Tech. Rep., www.vqeg.org, Feb. 1999.