VIDEO QUALITY ASSESSMENT BY INCORPORATING A MOTION PERCEPTION MODEL

Qiang Li and Zhou Wang

Dept. of Electrical Engineering, Univ. of Texas at Arlington, Arlington, TX 76019, USA
Emails: qiangli@uta.edu, zhouwang@ieee.org

ABSTRACT

Motion is one of the most important types of information contained in natural video, but direct use of motion information in the design of video quality assessment algorithms has not been deeply investigated. Here we propose to incorporate a recent motion perception model in an information theoretic framework. This allows us to estimate both the motion information content and the perceptual uncertainty in video signals. Improved video quality assessment algorithms are obtained by incorporating the model as spatiotemporal weighting factors, where the weight increases with the information content and decreases with the perceptual uncertainty. The proposed approach is validated using the Video Quality Experts Group Phase I test dataset.

Index Terms— video quality assessment, motion perception, information content, perceptual uncertainty, visual attention

1. INTRODUCTION

The capability of representing motion is one of the key features that distinguish a natural video sequence from a stack of independent still image frames. If we believe that the main purpose of vision is to extract useful information from the visual scene, then the perception of motion information would play an important role in the perception of natural video.

Nevertheless, in the literature of video quality assessment (VQA), motion information has typically been employed indirectly. The most frequently used method is temporal filtering [1, 2], where linear filters or filter banks are applied along the temporal direction (or along the spatial and the temporal directions simultaneously), and the filtered signals are normalized to reflect the variation of human visual sensitivity as a function of temporal frequency. Advanced models may also include the temporal masking effects [2] or statistics of the temporal filter coefficients [3]. Since motion in the visual scene may cause variations in signal intensity along the temporal direction, temporal filtering can, to some extent, capture motion. However, motion may not be the sole reason for temporal signal intensity variations, and the speed of motion cannot be directly related to the strength of temporal filter responses. Moreover, many visual experiments that measure temporal visual sensitivities were done with flickering patterns [1], which do not involve any physical motion.

Only a small number of existing VQA algorithms detect motion and use motion information directly. In [4], a heuristic weighting model was combined with the structural similarity (SSIM) based quality assessment method [5] to account for the fact that the accuracy of visual perception is significantly reduced when the speed of motion is extremely large. In [6], a set of heuristic fuzzy rules were proposed that use both absolute and relative motion information to describe visual attention and motion suppression. It was shown that these rules are effective in improving the performance of VQA algorithms.

In this paper, we propose to directly incorporate motion information in an information theoretic framework. Our approach is based on the following assumptions and observations. First, we believe that the human visual system (HVS) is an optimal information extractor (subject to certain physical constraints such as power consumption), as widely hypothesized in computational vision science [7]. As a result, the areas in the visual scene that contain more information should be more likely to attract visual attention. Such *information content* can be quantified using statistical information theory, provided that a statistical model about the information source is available. Indeed, information content-based method has shown to be useful in still image quality assessment [8].

Second, as in [3, 8], we model visual perception as an information communication process, where the information source (the video signal) passes through an error-prone communication channel (the HVS). The key difference here is that the noise level in the communication channel is not fixed. This is based on the observation that the HVS does not perceive all the information content with the same degree of certainty. For example, when the global motion in a video sequence is very large (or the head/camera motion is very large), the HVS cannot identify the objects presented in the video with the same accuracy as in still images, i.e., the video signal is perceived with higher uncertainty. Again, such *perceptual uncertainty* can be quantified based on information theory, by relating the channel distortion model with the speed of motion. In particular, recent psychophysical studies on the speed of motion [9] suggest that the internal noise of visual speed perception is approximately proportional to the true stimulus speed and inverse-proportional to the stimulus contrast.

IEEE Inter. Conf. Image Proc., San Antonio, TX, Sept. 16-19, 2007

2. METHOD

Our method is largely inspired by a recent paper by Stocker and Simincelli [9] about the visual speed perception. Based on a Bayesian optimal observer hypothesis, the authors measured the prior and the likelihood probability distributions of speed perception simultaneously from a set of carefully designed psychovisual experiments. These provide us with the essential ingredients in the computation of both the motion information content and the perceptual uncertainty.

2.1. Information Content

The motion information in a video sequence can be represented as a 3-D field of motion vectors, where each spatial location (x, y) and time instance t is associated with a motion vector $\vec{v}(x, y, t) = [v_x(x, y, t) \;\; v_y(x, y, t)]^T$. For notational convenience, in the rest of the paper, we often drop the space and time indices and write a motion vector as $\vec{v}$. For a given video sequence, we consider three types of motion fields – absolute motion, background motion, and relative motion. The absolute motion $\vec{v}_a$ is estimated as the absolute pixel movement at each spatial location between two adjacent video frames. The background motion $\vec{v}_g$ is global, often caused by the movement of the image acquisition system. The relative local motion $\vec{v}_r$ is defined as the vector difference between the absolute and the global motion, i.e.,

$$\vec{v}_r = \vec{v}_a - \vec{v}_g \, . \tag{1}$$

The speed of motion can be computed as the length of the motion vector, which, for convenience, we denote as $v = \|\vec{v}\|_2$. Thus, v_g, v_a and v_r represent the speed of the background motion, the absolute motion, and the relative motion, respectively.

It is believed that object motion is associated with visual attention and can be used for predicting visual fixations [10]. This is intuitively sensible because statistically, most of the objects in the visual scene are static (or close to static) relative to the background. As a result, an object with significant motion relative to the background would be a strong surprisal to the visual system. If the HVS is an optimal information extractor, as discussed in Section 1, then it should pay attention to such a surprising event. This intuitive idea can be converted into a mathematical measure if the prior distribution about the speed of motion is known. Early work on Bayesian speed perception has assumed Gaussian distribution for the speed prior [11], but the psychovisual study in [9] suggests that the distribution has a much longer tail than Gaussian and approximately constitutes a straight line in the logarithmic speed domain. Thus, we use a power-law function to describe it:

$$p(v_r) = \frac{\tau}{v_r^{\alpha}} \, , \tag{2}$$

where τ and α are two positive constants. Since the power-law function does not sum to a finite number, this is not a strictly valid probability density function and can only be used when v_r is away from 0. For any observed motion v_r, we can then estimate the information content associated with it by computing its self-information or surprisal as

$$I = -\log p(v_r) = \alpha \log v_r + \beta, \tag{3}$$

where $\beta = -\log \tau$ is a constant.

2.2. Perception Uncertainty

If we model visual perception as an information communication process, then the amount of information that can be received (perceived) at the receiver end will largely depend on the noise in the distortion channel (the HVS). In [9], the internal noise probability distribution is modeled as a likelihood function of perceived motion for a given true stimulus motion. It was found that a log-normal distribution can provide a good description of the likelihood function:

$$p(m|v_s) = \frac{1}{\sqrt{2\pi}\sigma m} \exp\left[\frac{-(\log m - \log v_s)^2}{2\sigma^2}\right] \, , \tag{4}$$

where v_s and m are the speed of the true stimulus motion and the perceived motion, respectively. Furthermore, the experimental results in [9] suggest that in the log-speed domain, the width parameter σ in the log-normal distribution is roughly constant for any stimulus speed v_s and inversely dependent on the stimulus contrast c. Note that the width here is represented in the log-domain, and thus it indeed scales linearly with v_s in the linear speed domain. Mathematically, we model it as

$$\sigma = \frac{\lambda}{c^{\gamma}} \, , \tag{5}$$

where λ and γ are both positive constants.

For a given video sequence, assume that the underlying stimulus speed v_s is the speed of the global motion v_g, we can measure the motion perceptual uncertainty using the entropy of the likelihood function, which can be computed as

$$\begin{aligned} U &= -\int_{-\infty}^{\infty} p(m|v_g) \log p(m|v_g) dm \\ &= \frac{1}{2} + \frac{1}{2}\log(2\pi\sigma^2) + \log v_g \\ &= \log v_g - \gamma \log c + \delta \, , \end{aligned} \tag{6}$$

where $\delta = \frac{1}{2} + \frac{1}{2}\log(2\pi) + \log\lambda$ is a constant. This perceptual uncertainty measure is consistent with our intuition. On the one hand, it increases with the global motion of the video frame, suggesting that when the global motion is very large, the HVS cannot extract the structural information about the objects presented in the video with the same accuracy as in still images. On the other hand, it decreases with the increase of the stimulus contrast, suggesting that higher contrast objects are perceived with higher accuracy.

2.3. VQA Based on Motion Perception

We measure the motion information content and the perceptual uncertainty at every spatial location and time instance (x, y, t) in the video sequence. Based on the efficient coding hypothesis about the HVS, the importance of a visual event should increase with the information content, and decrease with the perceptual uncertainty. Therefore, we define the following spatiotemporal importance function at every (x, y, t)

$$w = I - U = (\alpha \log v_r + \beta) - (\log v_g - \gamma \log c + \delta) \,. \quad (7)$$

This importance function alone cannot serve as a full VQA algorithm. However, it can be incorporated into a local image distortion/quality measure as a weighting function. The local image distortion/quality measure must provide a 3D quality map of the video sequence being evaluated. Let $q(x, y, t)$ be the quality/distortion map given by the local quality/distortion metric, the final VQA score is computed as

$$Q = \frac{\sum_t \sum_x \sum_y w(x,y,t)\, q(x,y,t)}{\sum_t \sum_x \sum_y w(x,y,t)} \,. \quad (8)$$

3. IMPLEMENTATION

We use a multi-layer optical flow algorithm [12] with a five-level pyramid decomposition for motion estimation. The background motion is obtained by a maximum likelihood estimation to identify the peak of the motion vector histogram on the 2D grid [13]. The relative motion vector $\vec{v}_r$ are then computed using Eq. (1). The local contrast is estimated as the ratio between the local standard deviation normalized by the local mean:

$$c' = \frac{\sigma_p}{\mu_p + \mu_0} \,, \quad (9)$$

where σ_p and μ_p are the standard deviation and the mean computed within a local patch respectively, and μ_0 is small constant. In order to take into account the contrast response saturation effect [14], we pass the contrast computation through a pointwise nonlinear function given by

$$c = 1 - e^{-(c'/\theta)^\rho} \,, \quad (10)$$

where ρ and θ are two constants that control the slope and the position of the function, respectively.

A practical issue in the implementation of the algorithm is that the global motion v_g, the local relative motion v_r, and the local contrast c may be close to zero. This could result in unstable evaluation of the weighting function Eq. (7). As in [9], instead of computing $\log v_r$, $\log v_g$ and $\log c$, we replace them with $\log(1 + v_r/v_0)$, $\log(1 + v_g/v_0)$ and $\log(1 + c/c_0)$, respectively, where v_0 and c_0 are both small positive constants. Furthermore, to avoid negative weighting, we threshold it at 0. Therefore, Eq. (7) becomes

$$w = \max \left\{ 0, \left[\alpha \log \left(1 + \frac{v_r}{v_0}\right) + \beta \right] - \left[\log \left(1 + \frac{v_g}{v_0}\right) - \gamma \log \left(1 + \frac{c}{c_0}\right) + \delta \right] \right\} \quad (11)$$

Assuming a 32 pixels/degree of viewing distance and $v_0 = 0.3$ degree/sec, we obtain $v_0 = 0.384$ pixles/frame for 25 frames/sec video and $v_0 = 0.32$ pixles/frame for 30 frames/sec video, respectively. The other parameters are hand-picked and we find that the following parameters give reasonable results and use them in all the experiments reported later in this paper: $\alpha = 0.2$, $\beta = 0.09$, $\gamma = 2.5$, $\delta = 2.25$, $\mu_0 = 6$, $\theta = 0.05$, $\rho = 2$, and $c_0 = 0.7$.

4. TEST

We test the proposed method using the video quality experts group (VQEG) Phase I dataset (available at www.vqeg.org), which contains 20 SDTV reference video sequences. The reference data set includes ten 50Hz (25 frames/sec) and ten 60Hz (30 frames/sec) video sequences. Every reference video sequence has 16 distorted versions. This results in a total of 320 distorted video sequences. The subject score for each sequence is given by the mean opinion score (MOS) from the ratings given by multiple human subjects. The difference of MOS (DMOS) score is then calculated for each distorted video sequence by subtracting its MOS by the MOS of its corresponding reference video sequence.

The proposed model is incorporated as weighting factors with two types of local distortion/quality maps − the squared error map and the SSIM index map. If no weighting is added, then averaging the squared error map and the SSIM index map results in the standard mean squared error (MSE), and the mean SSIM measure, respectively. One can also convert the MSE to a peak signal-to-noise ratio (PSNR) measure. With the local weights added, we can compute a weighted version of MSE/PSNR and a weighted version of SSIM.

We use the Spearman rank order correlation coefficient (ROCC) between the subject and objective scores to evaluate the performance of the VQA algorithms. Table 1 shows the ROCC results of three datasets − the 50Hz dataset, the 60Hz dataset, and all data combined. The results suggest that the proposed weighting method is quite effective. It gives clear and consistent improvement to all test datasets based on two completely different types of image distortion/quality maps. Figs. 1 (a), (b), (c), (d) show the scatter plots of the subjective/objective comparisons on all VQEG test video sequences for PSNR, PSNR with proposed weighting, SSIM, and SSIM with proposed weighting, respectively. These scatter plots confirm the ROCC results shown in Table 1. It can be seen that after applying the proposed weighting method, the clusters of sample points (each associated with a video sequence) become much tighter, which implies better consistency between subjective and objective quality evaluations.

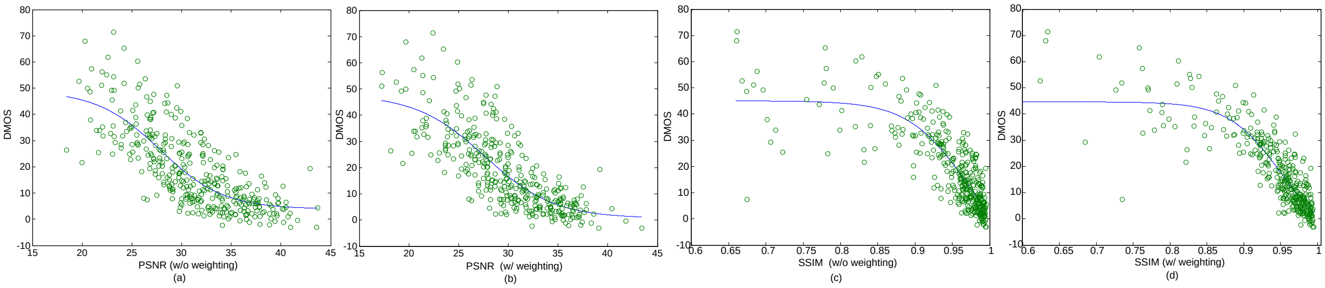

Fig. 1. Scatter plots of subjective/objective scores on VQEG Phase I test database. Each sample point represents the subjective/objective scores of one test video sequence. (a) PSNR; (b) PSNR with proposed weighting method; (c) SSIM; (D) SSIM with proposed weighting method. All SSIM values were raised to the 8th power for better visualization.

Table 1. ROCC results of VQA algorithms. PSNR(w): PSNR with proposed weighting; SSIM(w): SSIM with proposed weighting.

dataset	PSNR	PSNR(w)	SSIM	SSIM(w)
50hz	0.8152	0.8278	0.8301	0.8948
60hz	0.7112	0.7303	0.7680	0.7985
All	0.7818	0.8048	0.8127	0.8621

5. CONCLUSION

We propose a new method to incorporate motion information in VQA. Our tests with the VQEG Phase I dataset show that the weighting function computed based on our model is effective and consistent in improving VQA algorithms. A distinctive feature of our approach, as compared to the heuristic methods proposed in [4, 6], is that the use of motion information is well justified from the recent findings in psychophysical studies of human motion perception [9] and is formulated using an information theoretic framework. A underlying assumption we are making in this paper is that the information content and perceptual uncertainty of the video signal is proportional to the information content and perceptual uncertainty of speed perception. In the future, other types of information content (such as the local structure content measured in [8]) and perceptual uncertainty models may also be included into the same framework.

6. REFERENCES

[1] T. N. Pappas, R. J. Safranek, and J. Chen, "Perceptual criteria for image quality evaluation," in *The Handbook of Image and Video Processing*, A.Bovik, Ed. Academic Press, 2nd Ed., 2005.

[2] Z. Wang, H. R. Sheikh, and A. C. Bovik, "Objective video quality assessment," in *The Handbook of Video Databases: Design and Applications*, B. Furht and O. Marques, Eds. CRC Press, 2003.

[3] H. R. Sheikh and A. C. Bovik, "A visual information fidelity approach to video quality assessement," *The First International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, Jan 2005.

[4] Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," in *IEEE Trans. Signal Processing*, Feb. 2004, vol. 19, pp. 121–132.

[5] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structual similarity," *IEEE Trans. Image Processing*, vol. 13, pp. 600–612, Apr 2004.

[6] Z. K. Lu, W. Lin, X. K. Yang, E. P. Ong, and S. S. Yao, "Modeling visual attention's modulatory afterreffects on visual sensitivity and quality evaluation," *IEEE Trans. Image Processing*, vol. 14, pp. 1928–1942, 2005.

[7] E. P. Simoncelli and B. Olshausen, "Natural image statistics and neural representation," *Annual Review of Neuroscience*, vol. 24, pp. 1193–1216, May 2001.

[8] Z. Wang and X. L. Shang, "Spatial pooling strategies for perceptual image quality assessment," in *Proc. IEEE Int. Conf. Image Proc.*, Oct. 2006.

[9] A. A. Stocker and E. P. Simoncelli, "Noise characteristics and prior expectations in human visual speed perception," *Nature Neuroscience*, vol. 9, pp. 578–585, 2006.

[10] Z. Wang, L. Lu, and A. C. Bovik, "Foveation scalable video coding with automatic fixation selection," *IEEE Trans. Image Processing*, vol. 12, no. 2, pp. 243–254, Feb. 2003.

[11] E. P. Simoncelli, E. H. Adelson, and D. J. Heeger, "Probability distributions of optical flow," in *IEEE Inter. Conf. Computer Vision and Pattern Recognition*, June 3-6 1991, pp. 310–315.

[12] M. J. Black and P. Anandan, "The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields," *Computer Vision and Image Understanding*, vol. 63, no. 7, pp. 75–104, 1996.

[13] T. Vlachos, "Simple method for estimation of global motion paramenters using spares translational miton vector fields," *Electronics letters*, vol. 34, pp. 90–91, 1998.

[14] D. J. Heeger and P. C. Teo, "A model of perceptual image fidelity," in *Proc. IEEE Int. Conf. Image Proc.*, 1995, pp. 343–345.