IEEE International Conference on Acoustics, Speech & Signal Processing (ICASSP10), Dallas, TX, Mar. 2010

TEMPORAL MOTION SMOOTHNESS MEASUREMENT FOR REDUCED-REFERENCE VIDEO QUALITY ASSESSMENT

Kai Zeng and Zhou Wang

Dept. of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ontario, Canada
kzeng@engmail.uwaterloo.ca, zhouwang@ieee.org

ABSTRACT

Reduced-reference (RR) video quality measures aim to predict the perceptual quality of distorted video signals using only partial information about the reference video. Existing RR video quality assessment models are mostly designed and/or trained for specific applications such as lossy compression, where the detectable distortion types are often fixed and limited. Here we propose a novel approach that measures temporal motion smoothness of a video sequence by examining the temporal variations of local phase structures in the complex wavelet transform domain. We show that the proposed measure can detect a wide range of well-known practical distortions, including noise contamination, blurring, line or frame jittering, and frame dropping. In addition, the proposed algorithm does not require a costly motion estimation process and has a low RR data rate, making it much easier to be adopted in real-world visual communication applications.

Index Terms— perceptual image quality, temporal motion smoothness, reduced-reference video quality assessment, complex wavelet transform, local phase correlation, video jittering, video frame dropping

1. INTRODUCTION

Objective video quality assessment models typically require the access to the reference video that is assumed to have perfect quality. In practical visual communication applications, such full-reference (FR) methods may not be applicable because the reference video are unavailable [1]. On the other hand, no-reference (NR) video quality assessment is extremely difficult, especially when the types of distortions between senders and receivers are unknown [1]. Reduced-reference (RR) video quality measures provide a solution that lies between FR and NR models. They are designed to evaluate the visual quality of the distorted video with only partial information about the reference video.

The most challenging task in the design of RR video quality measures is to find appropriate RR features that 1) provide an efficient summary of the reference video; 2) are sensitive to the targeted types of video distortions; 3) are relevant to the perceptual characteristics of the human visual system; and 4) have relatively low data rate (so that they do not add too much burden to the visual communication systems that need to transmit the RR features) [2]. Most existing RR video quality models are developed and trained for specific applications such as lossy compression [1]. This makes the design task easier because the distortion types are known and fixed. Meanwhile, it also significantly limits their application scope.

In this paper, we are interested in discovering novel RR features that can capture video quality degradations caused by a wide variety of practical distortions, especially those related to motion (because the capability of representing motion is probably the most critical feature that distinguishes video from still images). In particular, we develop a novel method to quantify the temporal motion smoothness [3] of video sequences, which is affected by many types of distortions commonly encountered in real-world video acquisition, communication and processing systems, including noise contamination, blurring, frame jittering and frame dropping.

2. TEMPORAL MOTION SMOOTHNESS

Let $f(x)$ be a given real static signal, where x is the index of spatial position. When $f(x)$ represents an image, x is a 2-D vector. For simplicity, in the derivations below, we assume x to be one dimensional. However, the results can be easily generalized to two and higher dimensions. A time varying image sequence can be created from the static image $f(x)$ with rigid motion and constant variations of average intensity:

$$h(x,t) = f(x+u(t)) + b(t)\,. \tag{1}$$

Here $b(t)$ is real and accounts for the time-varying background luminance changes, and $u(t)$ indicates how the image positions move spatially as a function of time. We call the motion N-th order smooth if the $(N+1)$-th and higher order derivatives of $u(t)$ with respect to t are all zeros [3].

Now consider a family of symmetric complex wavelets whose "mother wavelets" can be written as a modulation of a low-pass filter $w(x)$= $g(x)\,e^{j\omega_c x}$, where ω_c is the center frequency of the modulated band-pass filter, and $g(x)$ is a slowly varying and symmetric function. The family of wavelets are dilated/contracted and translated versions of the mother wavelet: $w_{s,p}(x) = \frac{1}{\sqrt{s}}\, w\left(\frac{x-p}{s}\right)$, where $s \in R^+$ is the scale factor, and $p \in R$ is the translation factor. Using the convolution theorem and the scaling and modulation properties of the Fourier transform, we can compute the complex wavelet transform of $f(x)$ as

$$\begin{aligned} F(s,p) &= \int_{-\infty}^{\infty} f(x)\, w^*_{s,p}(x)\, dx \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega)\sqrt{s}\, G(s\,\omega - \omega_c)\, e^{j\omega p}\, d\omega\,, \end{aligned} \tag{2}$$

where $F(\omega)$ and $G(\omega)$ are the Fourier transforms of $f(x)$ and $g(x)$, respectively. Applying such a complex wavelet transform to both sides of Eq. (1) at time instance t, we have

$$\begin{aligned} H(s,p,t) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega)\sqrt{s}\, G(s\,\omega - \omega_c)\, e^{j\omega(p+u(t))}\, d\omega \\ &\approx F(s,p)\, e^{j(\omega_c/s)u(t)}\,. \end{aligned} \tag{3}$$

Here $b(t)$ is eliminated because of the bandpass nature of the wavelet filters. The approximation is valid when the envelope $g(t)$ is slowly varying and the motion $u(t)$ is small. A more convenient way to understand Eq. (3) is to take a logarithm on both sides, which gives

$$\log H(s,p,t) \approx \log F(s,p) + j(\omega_c/s)u(t)\,. \tag{4}$$

The key point here is that at a given scale s and a given spatial position p, the imaginary part of the logarithm of the complex wavelet coefficient changes linearly with $u(t)$. In other words, the local phase structures over time can be fully characterized by the movement function $u(t)$.

In order to relate temporal motion smoothness with the time-varying complex wavelet transform relationship, we must examine the complex wavelet coefficients at multiple time instances. A convenient choice is to start from a time instance t_0 and sample the sequence at consecutive time steps $t_0 + n\Delta t$ for $n = 0, 1, ..., N$. We define the N-th order temporal correlation function as [3]

$$L_N(s,p) = \sum_{n=0}^{N} (-1)^{n+N} \binom{N}{n} \log H(s,p,t_0 + n\Delta t)\,. \tag{5}$$

When the motion is (N-1)-th order smooth, i.e., $u^{(N)}(t_0) = 0$, then it can be derived that $L_N(s,p) \approx 0$ [3]. It needs to be kept in mind that this approximation is achieved based on the ideal formulation of Eq. (1) and the ideal assumption of (N-1)-th order temporal motion smoothness. Real natural image sequences are expected to deviate from these assumptions. However, by looking at the statistics of the imaginary part of $L_N(s,p)$, one may be able to quantify such deviation and use it as an indicator of temporal motion smoothness.

As a counterpart of the temporal correlation function $L_N(s,p)$, we can also define a temporal energy function

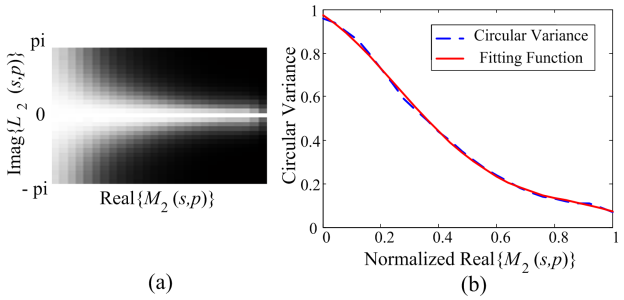
$$M_N(s,p) = \sum_{n=0}^{N} \binom{N}{n} \log H(s,p,t_0 + n\Delta t)\,, \tag{6}$$

which is useful for us to observe the strength of temporal motion smoothness as a function of local energy. An example of the imaginary part of $L_N(s,p)$ conditioned on the real part of $M_N(s,p)$ is shown in Fig. 1(a), where each column in the 2-D histogram is normalized to one. The conditional histogram shows strong temporal motion smoothness (when the values of $imag\{L_2(s,p)\}$ are close to zero), and such a statistical regularity becomes stronger with the increase of local signal strength (as the width of the column in the 2D histogram becomes narrower). This is not surprising because small magnitude coefficients typically come from the smooth background regions in an image and are easily disturbed by background noise.

3. RR VIDEO QUALITY ASSESSMENT

A full RR video quality assessment system consists of three modules: 1) RR feature extraction at the sender side; 2) Transmission of RR features from the sender to the receiver (may through an auxiliary channel [1] or through the same channel as video transmission [2, 4]); 3) Feature extraction and quality evaluation of the distorted video at the receiver side. This section focuses on the first and the third modules.

At the sender side, the given reference video sequence is first divided into groups of pictures (GOPs), each containing three consecutive frames. For each GOP, all three frames were decomposed using the complex version [5] of the steerable pyramid [6], an overcomplete wavelet transform that avoids aliasing in subbands. The



(a) (b)

Fig. 1. (a) Conditional histogram of $imag\{L_2(s,p)\}$ versus $real\{M_2(s,p)\}$ of a natural video sequence; (b) Variation of circular variance and the best fourth order polynomial fitting.

second order temporal correlation and temporal energy functions $L_2(s,p)$ and $M_2(s,p)$ are then computed for each subband. Instead of using the marginal histogram of $imag\{L_2(s,p)\}$ to quantify temporal motion smoothness (as in [3]), here we extract RR features based on the conditional histogram of $imag\{L_2(s,p)\}$ versus $real\{M_2(s,p)\}$. The reason behind this choice is that temporal motion smoothness is much stronger at high energy coefficients (as can be seen in Fig. 1(a)), but marginal histogram of $imag\{L_2(s,p)\}$ cannot distinguish such differences and takes all coefficients into equal account. Furthermore, the trend of how temporal motion smoothness varies with the increase of local signal energy provides additional information that can help characterize the reference video. Specifically, we use the circular variance (CV) [7] of each column in the conditional histogram to quantify the spread of the angle variables. For each column, the circular variance is computed as

$$CV = 1 - \frac{\left|\sum_{i=1}^{M} h_i e^{j\theta_i}\right|}{\sum_{i=1}^{M} h_i}\,, \tag{7}$$

where M is the total number of histogram bins, and h_i and θ_i are the height and center angle of the i-th histogram bin, respectively. The column CV values computed based on the conditional histogram of Fig. 1(a) are shown in Fig. 1(b) as a dashed curve, which provides an adequate description about the variation trend of temporal motion smoothness. Depending on the application environment, transmitting the CV curve as the RR features to the receiver may not be a realistic solution because it requires a fairly large RR data rate. To overcome this problem, we use a parametric model to describe the CV curve and only send the model parameters to the receiver. In particular, we find that a fourth order polynomial can very well approximate a typical CV curve, as demonstrated by the solid fitting curve in Fig. 1(b). Consequently, only 5 parameters (that uniquely define the fourth order polynomial) are employed as RR features and are transmitted to the receiver.

At the receiver side, the distorted video sequence is processed the same way as at the sender side, i. e., GOP division and complex wavelet signal decomposition, followed by the computation of the conditional histogram and the CV curve. Meanwhile, the received RR features (polynomial parameters) are used to reconstruct the model CV curve. Finally, we quantify the overall video quality distortion as

$$D = \left\{ \frac{1}{K} \sum_{k=1}^{K} [CV(k) - CV_{model}(k)]^2 \right\}^{1/2}\,, \tag{8}$$

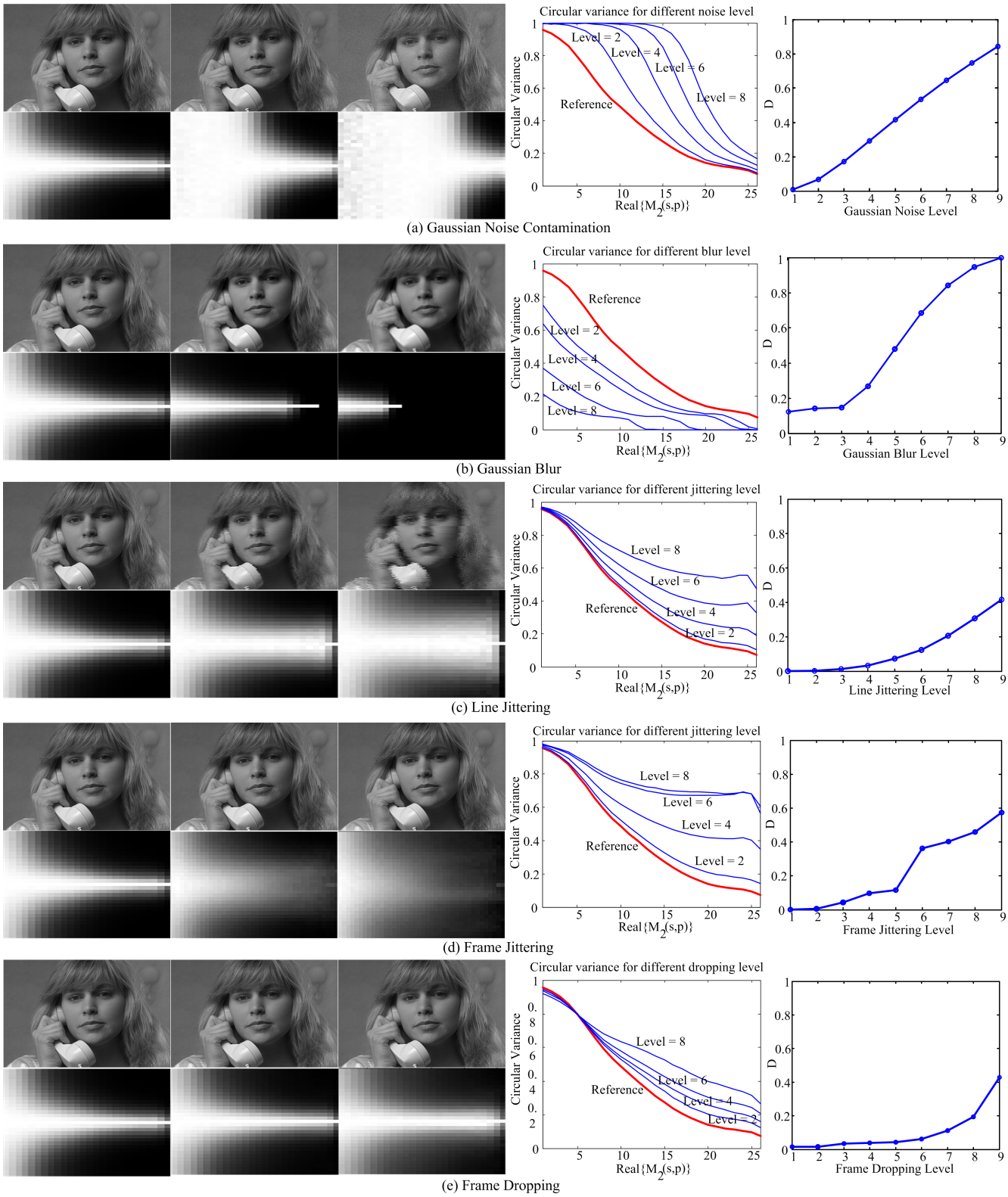


Fig. 2. Left: conditional histograms of $Imag\{L_2(s,p)\}$ versus $Real\{M_2(s,p)\}$ of different distortion types at low, middle and high distortion levels; Middle: circular variance as a function of $Real\{M_2(s,p)\}$ for the reference video sequence and distorted sequences at different distortion levels; Right: proposed distortion measure as a function of distortion level. From top to bottom: (a) Gaussian noise contamination; (b) Gaussian blur; (c) Line jittering; (d) Frame jittering; (e) Frame dropping.

where K is the total number of columns in the conditional histogram, and $CV(k)$ and $CV_{model}(k)$ are the CV values of the k-th column of the distorted CV curve and the model CV curve, respectively. Because the CV values are bounded between 0 and 1, this distortion measure is also bounded by the same range.

4. EXPERIMENTS

To the best of our knowledge, no existing RR VQA method was designed to evaluate general distortions (most of them are application specific, for example, for testing video compression only). There is also a lack of subject-rated video databases that cover general distortions. Therefore, we test the proposed algorithm using simulated video with five distortion types at different distortion levels. These include 1) Gaussian noise contamination, where the distortion level is defined as the standard deviation of the noise; 2) Gaussian blur, where the standard deviation of the Gaussian filter size defines the distrotion level; 3) Line jittering, where each line in a frame is shifted horizontally by a random number uniformly distributed between $[-S, S]$, and S defines the jittering level; 4) frame jittering, where the whole frame is shifted together by a random number uniformly distributed between $[-S, S]$; and 5) frame dropping, which is simulated by discarding every 1 of N frames and repeating the previous frame to fill the empty frame, and $12-N$ defines the distortion level. All distortion types are associated with certain real-world scenarios. For example, line jittering occurs when two fields of interlaced video signals are not synchronized; frame jittering is often caused by irregular camera movement such as hand shaking; and frame dropping usually happens when the bandwidth of a real-time communication channel drops and some frames have to be discarded to reduce the bit rate of the video signal being transmitted.

Figure 2 shows the results of the experiment. First, it is interesting to observe that different distortions lead to different changes to the conditional histogram of $imag\{L_2(s,p)\}$ versus $real\{M_2(s,p)\}$. For example, noise contamination and jittering cause the histogram to spread, but Gaussian blur results in shrinkage of the histogram (as the energy reduces, especially at high frequencies). The observed changes are well captured by the departure of the CV curves of the distorted video sequence from the reference CV curves. Specifically, for each distortion type, the CV curve moves away from the reference CV curve with the increase of distortion level. This is further confirmed by computing the overall distortion measure D, which is monotonically increasing with the distortion level. From this experiment, we observe that the same objective distortion measure D works consistently for each individual type of distortion. This demonstrates the potential of the proposed method for general-purpose RR video quality assessment, which is different from most approaches in the literature where ad-hoc features tuned to specific distortion types (such as blocking and ringing artifacts) are often used. Another interesting observation is regarding the frame jittering and frame dropping distortions. Notice that with these two types of distortions, the quality of each individual frame remains high quality, and thus frame-by-frame quality assessment approaches would give high quality scores to the image sequences undergoing these distortions, but the proposed method can capture them quite effectively without any specific change to the algorithm.

5. CONCLUSIONS

We propose a complex wavelet transform domain temporal motion smoothness measure and demonstrate its potential for general-purpose RR video quality assessment. There are several useful properties of the proposed algorithm: 1) it is applicable to a wide range of practical distortion types; 2) it captures relevant motion characteristics without explicit motion estimation, which often involves a complicated search procedure [8] or requires solving adaptive equations at every spatial location [9]; 3) it has a low RR data rate (only 5 scalar features). All these properties make it an attractive approach in real-world visual communication applications. For example, it can be directly adopted in a quality-aware video system [4].

The current work may be extended in many ways. First, higher-order temporal correlation functions may be employed to characterize the smoothness of higher-order motion (such as acceleration). Second, appropriate adjustments are needed to accommodate the cases of scene changes and very large motion (which may be solved by adopting a multi-scale, coarse-to-fine strategy). Third, temporal motion smoothness is only one aspect that affects perceived video quality, other RR features (such as intra-frame statistical features [2]) may be incorporated under a unified framework to provide a full solution to the problem of RR video quality assessment.

6. ACKNOWLEDGMENT

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada in the form of Discovery and Strategic Grants, and in part by Ontario Ministry of Research & Innovation in the form of an Early Researcher Award, which are gratefully acknowledged.

7. REFERENCES

[1] Zhou Wang and A. C. Bovik, *Modern Image Quality Assessment*, Morgan & Claypool Publishers, Mar. 2006.

[2] Zhou Wang, Guixing Wu, Hamid R. Sheikh, Eero P. Simoncelli, En-Hui Yang, and Alan C. Bovik, "Quality-aware images," *IEEE Trans. Image Processing*, vol. 15, no. 6, pp. 1680–1689, June 2006.

[3] Z. Wang and Q. Li, "Statistics of natural image sequences: Temporal motion smoothness by local phase correlations," in *Human Vision and Electronic Imaging IX, Proc. SPIE*, Jan. 2009, vol. 7240.

[4] B. Hiremath, Q. Li, and Z. Wang, "Quality-aware video," in *Proc. IEEE Int. Conf. Image Proc.*, San Antonio, TX, Sept. 2007.

[5] J. Portilla and E. P. Simoncelli, "A parametric texture model based on joint statistics of complex wavelet coefficients," *Inter. J. Computer Vision*, vol. 40, no. 1, pp. 49–71, Dec. 2000.

[6] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger, "Shiftable multi-scale transforms," *IEEE Trans. Information Theory*, vol. 38, pp. 587–607, 1992.

[7] N. I. Fisher, *Statistical analysis of circular data*, Cambridge University Press, New York, 2000.

[8] F. Dufaux and F. Moscheni, "Motion estimation techniques for digital TV: a review and a newcontribution," *Proceedings of the IEEE*, vol. 83, no. 6, pp. 858–876, June 1995.

[9] S. S. Beauchemin and J. L. Barron, "The computation of optical flow," *ACM Computing Surveys*, vol. 27, no. 3, pp. 433–467, Sept. 1995.