IEEE Inter. Conf. Image Processing, Hong Kong, China, Sept. 26-29, 2010.

QUALITY-AWARE VIDEO BASED ON ROBUST EMBEDDING OF INTRA- AND INTER-FRAME REDUCED-REFERENCE FEATURES

Kai Zeng and Zhou Wang

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

ABSTRACT

With the rapid development of network visual communications, there is an urgent need of effective and efficient video quality assessment (VQA) methods for quality control and resource allocation purposes. In this paper, a spatial and temporal reduced-reference (RR) VQA measure is combined with a robust video watermarking approach, leading to a quality-aware video (QAV) system. At the sender side, both intra- and inter-frame RR features are calculated from the original video based on statistical models of natural video. This is followed by error control coding to improve robustness. The encoded features are then embedded invisibly into the same video signal using a robust angle quantization index modulation based watermarking method in 3D discrete cosine transform domain. At the receiver side, the RR features are extracted and decoded from the distorted video and employed to predict the perceptual degradation of the video signal. Experimental results demonstrate the applicability of the proposed approach to a wide range of distortion types and levels.

Index Terms— video quality assessment, quality-aware video, natural video statistics, video watermarking, temporal motion smoothness, angle quantization index modulation

1. INTRODUCTION

Objective video quality assessment (VQA) metrics play an essential role in network visual communication systems for the evaluation, control, and improvement of the perceptual quality of video content. Although recent full-reference (FR) VQA measures have achieved notable success in predicting perceived image/video quality [1], they are not applicable in visual communication scenarios because full access to the original video is not available. Reduced-reference (RR) VQA measures provide a practically useful solution, which evaluate video quality with only partial information about the original video in the form of a set of RR features extracted from the original video at the sender side [1]. One difficulty in the deployment RR-VQA approaches is that they require the RR features to be transmitted to the receiver through a lossless ancillary channel [1], which is often hard to provide in real-world application environment. This motivated the ideas of quality-aware image (QAI) [2] and quality-aware video (QAV) [3], where the extracted RR features are embedded into the original image/video signal as invisible messages and transmitted to the receiver together with the image/video content.

In this paper, we propose a novel QAV system based on spatial and temporal RR-VQA and robust video watermarking. The general framework is depicted in Fig. 1. At the sender side, the extracted RR features include intra-frame features based on a statistical model of the marginal distribution of wavelet coefficients [2], and inter-frame RR features calculated by temporal motion smoothness measurement computed in the complex wavelet transform domain [4]. An error control encoding scheme, which consists of cyclic redundancy check (CRC) for error detection and low-density parity-check (LDPC) for error correction [5], is employed to improve the robustness in the subsequent transmission of the RR features. This is followed by embedding the encoded RR features into the same video signal invisibly using a robust angle quantization index modulation (AQIM) [6] based video watermarking approach in 3D discrete cosine transform (3D-DCT) domain. The resulting video is called a QAV, which is transmitted to the receiver through a lossy communication channel. At the receiver side, after a distorted version of the QAV is received, the same feature extraction process as at the sender side is applied to the distorted video. Meanwhile, the hidden messages are extracted, followed by error control decoding to recover the RR features. The error control code has the capability to identify errors. If it is found that the RR features are not fully recovered correctly, then the system reports an error message, indicating a failure in assessing the video quality. Otherwise, the recovered RR features, together with the corresponding features extracted from the distorted video, are employed by an RR-VQA algorithm, which evaluates the perceptual quality degradation of the distorted QAV.

2. RR VQA METHOD

2.1. Intra-frame feature extraction and distortion measure

Let $p(x)$ and $q(x)$ denote the probability density functions of the wavelet coefficients in the same subband of the same frame in the reference and distorted images, respectively. The Kullback-Leibler distance (KLD) between them is

$$d(p||q) = \int p(x) \log \frac{p(x)}{q(x)} dx \,. \tag{1}$$

$q(x)$ can be easily calculated from the distorted frame at the receiver. $p(x)$ needs to be transmitted from the sender. To do that efficiently, it is useful to summarize it using a 2-parameter generalized Gaussian density model that provides a good approximation [2]

$$p_m(x) = \frac{\beta}{2\alpha\Gamma(1/\beta)} e^{-(|x|/\alpha)^\beta} \,, \tag{2}$$

where $\Gamma(a)$ is the Gamma function. The model approximation error is computed as the KLD between $p_m(x)$ and $p(x)$:

$$d(p_m||p) = \int p_m(x) \log \frac{p_m(x)}{p(x)} dx \,. \tag{3}$$

In the end, only three RR parameters, α, β and $d(p_m||q)$, are extracted from each subband. At the receiver side, the intra-frame distortion is computed as an estimate of $d(p||q)$ given by

$$D_{\text{intra}} = \hat{d}(p||q) = d(p_m||q) - d(p_m||p) \,. \tag{4}$$

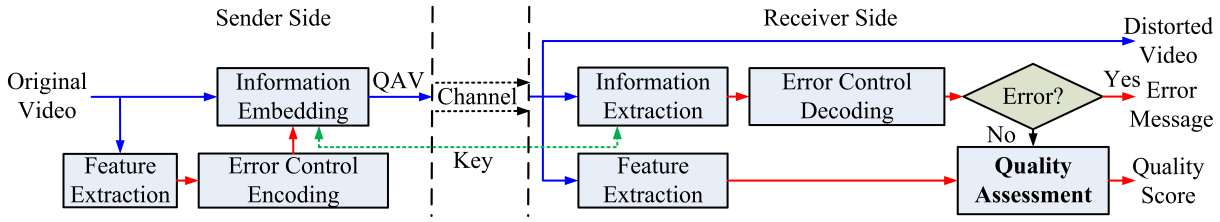

Fig. 1. Framework of the proposed QAV system.

2.2. Inter-frame feature extraction and distortion measure

The inter-frame features are extracted from 2D complex wavelet transforms applied on a frame-by-frame basis. 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. Let $f(x)$ be a real signal, where x is the spatial position index. Using Fourier transform properties, we can compute the complex wavelet transform of $f(x)$ as

$$F(s,p) = \tfrac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega)\sqrt{s}\, G(s\,\omega - \omega_c)\, e^{j\omega p}\, d\omega \qquad (5)$$

where $F(\omega)$ and $G(\omega)$ are the Fourier transforms of $f(x)$ and $g(x)$, respectively. A time varying image sequence can be created from $f(x)$ with rigid motion and constant variations of average intensity:

$$h(x,t) = f(x + u(t)) + b(t)\,, \qquad (6)$$

where $u(t)$ and $b(t)$ indicate image position and background luminance changes as a function of time. Applying complex wavelet transform to both sides of Eq. (6) at time instance t, we have

$$\begin{aligned} H(s,p,t) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega)\sqrt{s}\, G(s\,\omega - \omega_c)\, e^{j\omega(p+u(t))}\, d\omega \\ &\approx F(s,p)\, e^{j(\omega_c/s)u(t)}\,. \end{aligned} \qquad (7)$$

We can then define an N-th order temporal correlation function and energy function as

$$\begin{aligned} L_N(s,p) &= \sum_{n=0}^{N} (-1)^{n+N} \binom{N}{n} \log H(s,p,t_0 + n\Delta t)\,, \\ M_N(s,p) &= \sum_{n=0}^{N} \binom{N}{n} \log H(s,p,t_0 + n\Delta t)\,. \end{aligned} \qquad (8)$$

The strength of temporal motion smoothness can be characterized by the circular variance (CV) curve of the conditional distribution of the imaginary part of $L_2(s,p)\}$ versus the real part of $M_2(s,p)$. We found that the CV curve can be well fitted using a 4-th order polynomial, and therefore the 5 fitting parameters used to describe the polynomial are employed as the RR features for each complex wavelet subband.

At the receiver side, the CV curve of the distorted video is calculated and compared with that of the model CV curve reconstructed from the RR features. This leads to an inter-frame distortion measure

$$D_{\text{inter}} = \left\{ \frac{1}{N} \sum_{n=1}^{N} [\text{CV}(n) - \text{CV}_{\text{model}}(n)]^2 \right\}^{1/2}, \qquad (9)$$

where N is the number of samples in CV curve, and $\text{CV}(n)$ and $\text{CV}_{\text{model}}(n)$ are the n-th sample computed from the distorted video and the model CV curve, respectively. Finally, the overall distortion is computed as the average of intra- and inter-frame distortions:

$$D = \frac{1}{2}(D_{\text{intra}} + D_{\text{inter}})\,. \qquad (10)$$

3. ROBUST INFORMATION EMBEDDING

Robustness of information embedding is a critical issue to the success of QAV systems. To achieve it, the scalar RR features are first quantized to 7-bit representations, resulting in a binary RR bitstream. The bitstream is then expanded by a 16-bit CRC code for error detection, and then encoded using a binary LDPC code for error correction [5]. The column number of the sparse parity-check matrix of LDPC encoder was designed to be twice of the row number, so that it can correct up to 1 bit of error out of every 2 bits.

The error control coded bitstream is embedded invisibly into the original video using a watermarking scheme. Our method is based on an AQIM approach, which was shown to be extremely robust to contrast scaling attacks [6]. The novelty of our scheme is to apply it to pairs of coefficients in 3D-DCT domain, so that it is not only robust to scaling, but also to blur and other types of attacks. An example is illustrated in Fig. 2, where 1 bit of information is embedded into the plane composed of 2 3D-DCT coefficients. The plane is divided into R_0 and R_1 regions, corresponding to 0 and 1, respectively. The division is based on angular values and the angular quantization step is $\Delta = \pi/4$. Let a and b be the values of a pair coefficients, and $\angle c$ be the angle of the complex number $c = a + jb$. Then the AQIM

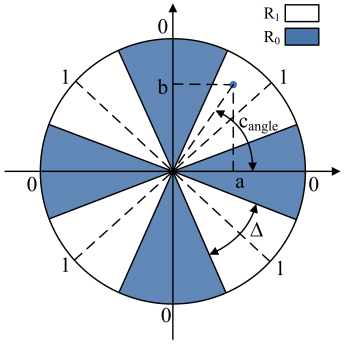

Fig. 2. Illustration of AQIM for $\Delta = \pi/4$.

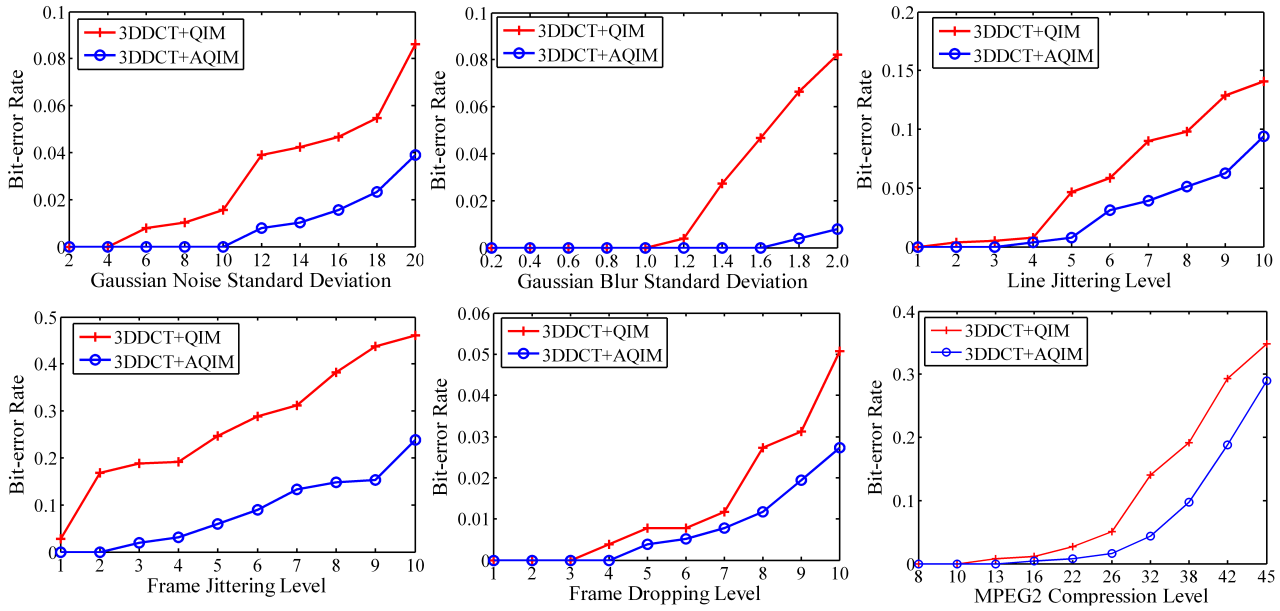

Fig. 3. Robustness test of information embedding schemes.

embedding scheme is given by an angular quantization operation

$$\begin{aligned} \angle c_{\text{new}} &= Q(\angle c + d(m)) - d(m) \equiv Q^m(\angle c)\,, \\ c_{\text{new}} &= |c| \exp(j \angle c_{\text{new}})\,, \end{aligned} \tag{11}$$

where m is the bit being embedded, Q is an angular quantization operator as exemplified by Fig. 2, c_{new} is the complex coefficient pair after embedding, and $d(m)$ is a dithering operator defined as

$$d(m) = \begin{cases} -\Delta/4, & \text{if } m = 0 \\ \Delta/4, & \text{if } m = 1\,. \end{cases} \tag{12}$$

At the receiver side, after a distorted version (denoted as c_d) of the embedded complex coefficient pair c_{new} is received, the embedded bit can be estimated using a minimum angular distance criterion:

$$\hat{m}(\angle c_d) = \underset{m \in \{0,1\}}{\operatorname{argmin}} \| \angle c_d - Q^m(\angle c_d) \|\,. \tag{13}$$

3D-DCT often leads to strong energy concentration when applied to natural video signals. As a result, the coefficients corresponding to low spatial and temporal frequencies have much higher energy than that of the high frequency ones. To maximize robustness, we choose the low frequency coefficients for AQIM embedding that are much less sensitive to typical distortions such as compression and noise contamination. Since both 3D-DCT and contrast scaling are linear operators, 3D-DCT domain AQIM is automatically robust to contrast scaling attack because the angular value in Fig. 2 is invariant to scaling. In addition, the coefficients selected for embedding are paired so that two coefficients that form a pair correspond to the same spatial and temporal frequencies (though may be different in orientation). This is critical to make the AQIM scheme robust to blur attack, because blur causes the two coefficients to scale down by the same ratio, such that the angular value in Fig. 2 remains unchanged. The value of Δ is tuned to achieve a compromise between robustness and imperceptibility of information embedding. The locations of the selected 3D-DCT coefficients are shared between the sender and receiver as the embedding key, as illustrated in Fig. 1.

4. IMPLEMENTATION AND EXPERIMENT

In our implementation, every 30 consecutive frames form a group of picture (GOP), where each frame is decomposed using a complex version [7] of a two-orientation steerable pyramid transform [8]. The subband statistics are carried out on the two orientation subbands at the finest scale by accumulating the coefficients of all frames in the GOP. These include the marginal statistics of real coefficients for intra-frame features and the statistics of the temporal correlation function conditioned on the energy function for inter-frame features. The intra- and inter-frame RR features are then extracted using the methods described in Section 2. This results in 8 features for each subband (3 intra- and 5 inter-frame features) and a total of 16 scalar features for both subbands. They are converted to 116 bits after 7-bit quantizations, and 256 bits after CRC and LDPC coding. The resulting encoded RR bitstream is then embedded into a 3D-DCT transform of the GOP using the method described in Section 3.

We simulated six types of distortions to test the proposed QAV system, which include 1) Gaussian noise contamination, where the distortion level is defined as the standard deviation of noise; 2) Gaussian blur, where the standard deviation of the blur filter defines the distortion level; 3) line jittering, simulated by shifting each line horizontally by a random number uniformly distributed between $[-S, S]$, and S defines the jittering level; 4) frame jittering, which is similar to line jittering except that the whole frame shifts together; 5) frame dropping, simulated by discarding every 1 out of N frames (empty frames are filled by repeating their previous frame) and 12-N defines the distortion level; and 6) MPEG2 compression, where the compression ratio defines the distortion level. All distortion types are observed in real-world scenarios. For example, frame dropping occurs when the bandwidth of a real-time communication channel drops; and frame jittering is often caused by irregular camera movement such as hand shaking.

Figure 3 shows the test results for the robustness of information embedding, where the bit-error rates are calculated without LDPC

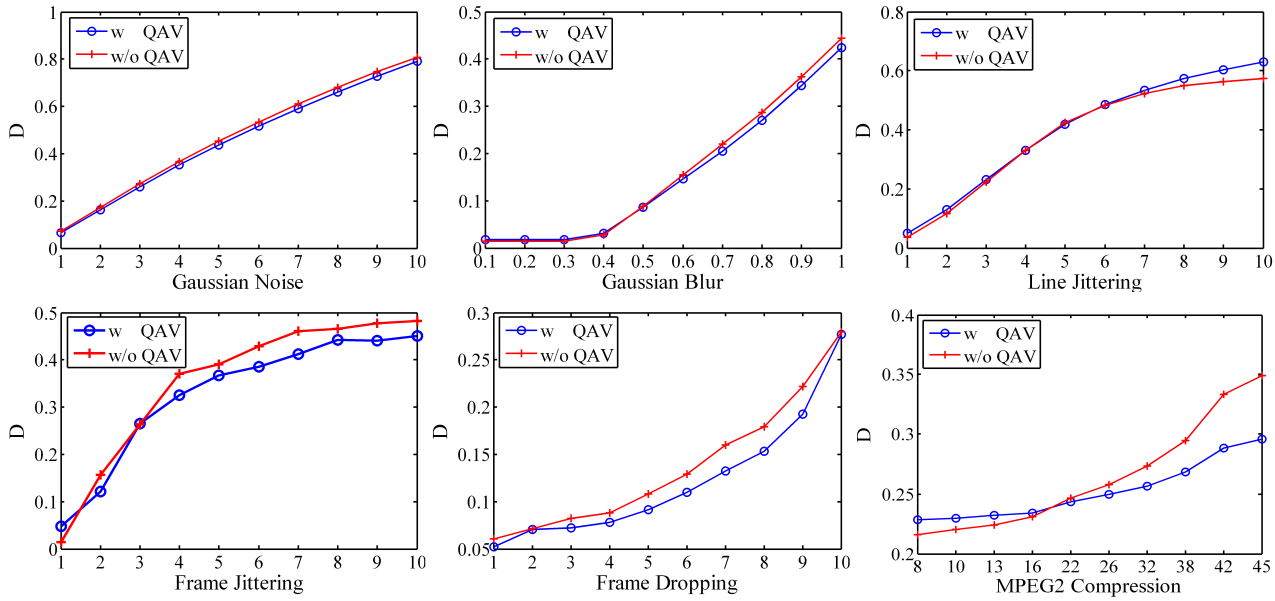

Fig. 4. RR VQA consistency with and without QAV information embedding.

correction, which can further improve the robustness. Compared with the traditional "3DDCT+QIM" method, "3DDCT+AQIM" leads to consistent improvement for all distortion types. As expected, the improvement is the most significant for blur distortions. Since information embedding alters the original video signal and thus its statistics, it is important to verify that such alteration does not have significant impact on the performance of the VQA algorithm. A comparison between the RR-VQA evaluation results with and without QAV information embedding is shown in Fig. 4 for six types of distortions. It appears that the differences are generally small relative to the distortion measures. This may be explained by the fact that the VQA algorithm mostly relies on the variations of the statistics of the fine scale coefficients, while information embedding mainly affects relatively lower frequencies of the video content.

5. CONCLUSION

We propose a QAV system that incorporates state-of-the-art RR-VQA algorithms with a novel robust information data hiding approach. Such a QAV system has a number of attractive properties: It provides the useful functionality of "quality-awareness" without affecting the conventional use of the video content; It avoids the necessity of an ancillary channel in the deployment of RR-VQA schemes; It allows the video content to be converted and distributed using any existing or user-defined formats, provided the embedded messages are not corrupted during lossy format conversion; It also provides an opportunity at the receiver side to partially "repair" the distorted video signal using the embedded RR features. Future work includes improving the performance of both the accuracy of RR-VQA and the robustness of information embedding, and providing meaningful video quality evaluations in the case that the RR features cannot be fully recovered (for example, by relating decoding error rate with perceived video quality).

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] Z. Wang and A. C. Bovik, *Modern Image Quality Assessment*, Morgan & Claypool Publishers, March 2006.

[2] Z. Wang, G. Wu, H. R. Sheikh, E. P. Simoncelli, E.-H. Yang, and A. C. Bovik, "Quality-aware images," *IEEE Trans. Image Processing*, vol. 15, no. 6, pp. 1680–1689, June 2006.

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

[4] K. Zeng and Z. Wang, "Temporal motion smoothness measurement for reduced-reference video quality assessment," in *IEEE Inter. Conf. Acoustics, Speech & Signal Proc.,*, Mar. 2010.

[5] Todd K. Moon, *Error Correction Coding: Mathematical Methods and Algorithms*, Wiley-Interscience, 2005.

[6] F. Ourique, V. Licks, R. Jordan, and F. Perez-Gonzalez, "Angle qim: a novel watermark embedding scheme robust against amplitude scaling distortions," in *IEEE Inter. Conf. Acoustics, Speech & Signal Proc.*, Mar. 2005, vol. 2, pp. 797–800.

[7] 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.

[8] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger, "Shiftable multi-scale transforms," *IEEE Trans. Info. Theory*, vol. 38, no. 2, pp. 587–607, 1992.