# QUALITY-AWARE VIDEO

*Basavaraj Hiremath, Qiang Li and Zhou Wang*

Dept. of Electrical Engineering, The Univ. of Texas at Arlington, Arlington, TX 76019, USA
basavaraj.hiremath@uta.edu, qxl7033@exchange.uta.edu, zhouwang@ieee.org

**ABSTRACT**

Recent development in network visual communications has emphasized on the need of objective, reliable and easy-to-use video quality assessment (VQA) systems. This paper introduces a novel idea of *quality-aware video* (QAV), in which extracted features about the original video sequence are invisibly embedded into the same video data. When such a QAV sequence is distributed over an error-prone network, a network user who receives it can decode the hidden messages and use them to evaluate the quality degradations between the original and the received video sequences. Our first implementation of QAV employs 1) a novel reduced-reference VQA method based on a statistical model of natural video, and 2) a 3D discrete cosine transform-based data hiding algorithm. The proposed approach does not assume any prior knowledge about image distortions, and the simulation results demonstrate its potentials to be generalized for different types and degrees of image distortions.

***Index Terms***— video quality assessment, quality aware video, reduced reference, data hiding, natural video statistics

## 1. INTRODUCTION

Video quality assessment (VQA) techniques are essential for network video communication systems to maintain, control, and enhance the quality of video data distributed over the network. Perhaps the most reliable VQA method is subjective testing, but the subjective nature makes it practically difficult to implement and limits its usage to test rooms. In the past decade, there has been an increasing need of objective VQA systems that are self-dependent, reliable, widely applicable to various scenarios, and consistent with subjective quality evaluations.

Depending on whether the original, perfect-quality video sequence is available as a reference to the quality assessment system, a VQA method may be classified as full-reference (FR), no-reference (NR), or reduced-reference (RR) [1]. FR methods are simply not applicable in our intended application because the network users would not have access to the original video data. On the other hand, all the existing NR methods, which do not assume any knowledge about the reference video, are limited to specific types of image distortions [1], e.g., blocking artifacts created in block-based video compression. However, knowledge of the distortions that arise between the original and distorted video is in general not available to VQA systems. Therefore, it is desirable to design VQA techniques that do not assume any particular image distortion types. RR methods provide a tradeoff between NR and FR methods. It does not require full access to the original video, but only needs partial information, in the form of a set of extracted RR features. Moreover, general-purpose RR quality assessment method has shown to be feasible for still image quality assessment [2].

In [2], we proposed the concept of quality-aware image, in which certain extracted features of the original image are embedded into the same image as invisible hidden messages. When a distorted version of such an image is received, users can decode the hidden messages and use an RR quality assessment method to evaluate the quality of the distorted image. In this paper, we extend the idea to *quality-aware video* (QAV). The transition from image to video, to some extent, makes the information embedding task easier because of the increase of the data volume (and thus the data embedding capacity). But it also casts new challenges to the quality assessment task because general-purpose RR VQA is a rarely explored research direction. In the literature, data hiding techniques have been employed for VQA purposes. In [3], [4] and [5], a pseudo-random bit sequence or a watermark image is hidden inside the video being transmitted. The bit error rate or the degradation of the watermark image measured at the receiver side is then used as an indication of the quality degradation of the host video. Strictly speaking, these methods are not VQA methods because no extracted features about either the reference or the distorted video are actually used in the quality evaluation process. Instead, the distortion processes that occur in the distortion channel are gauged. The fundamental difference of QAV is that the video quality degradations are evaluated by the variations of the video's own features (but not the watermarks), leading to better quality prediction accuracy.

The advantages of QAV are multifold. First, it uses an RR method that makes the VQA task feasible (as compared to FR and NR methods). Second, it does not affect the conventional usage of the video data because the data hiding process causes only invisible changes to the video. Third, it does not require a separate data channel to transmit the RR information. Fourth, it allows the video to be stored, converted and distributed using any existing or user-defined formats without losing the functionality of "quality-awareness", provided the hidden messages are not corrupted during lossy format conversion. This is an important advantage in comparison with the methods that include the RR features in image headers. Finally, it provides the users with a chance to partially "repair" the received distorted video by making use of the embedded RR features about the original video.

## 2. RR VQA BASED ON NATURAL VIDEO STATISTICS

At the core of a QAV system is an RR VQA algorithm. Here we propose an RR VQA method based on a statistical model of *temporal motion smoothness* in the complex wavelet transform domain.

Let $f(x)$ be a real static signal (e.g., a still image frame), where $x$ is the index of spatial position. For simplicity, in the derivations below, we assume $x$ to be one dimensional. However, the results can be easily generalized to higher dimensions. A time varying image sequence can be created from the static image $f(x)$ with rigid motion and constant variations of contrast and average intensity:

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

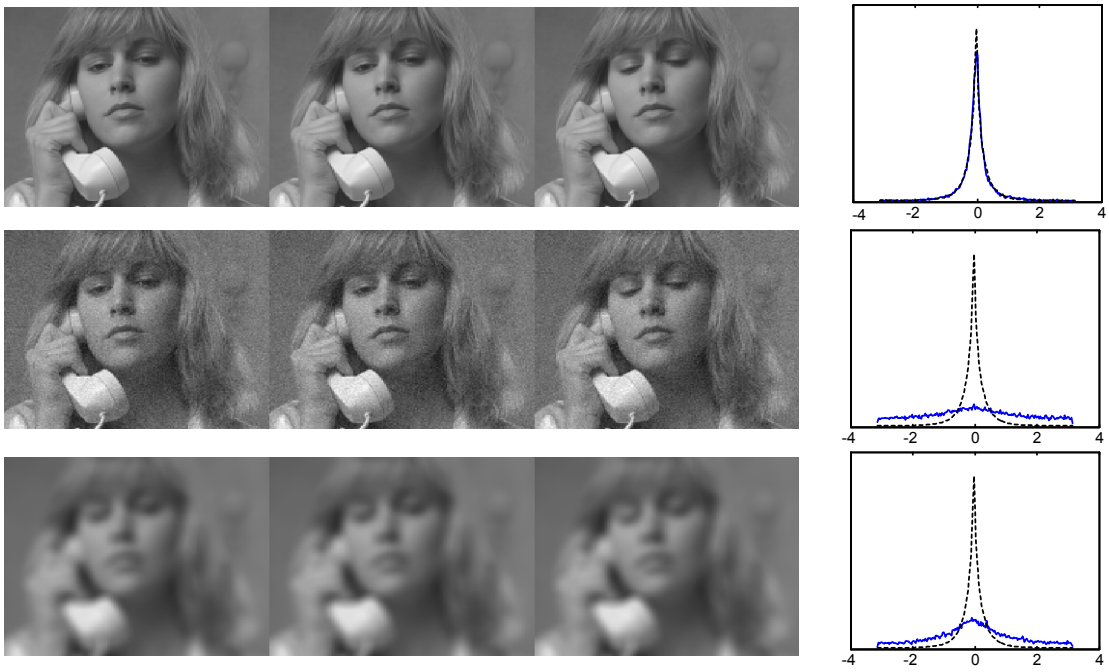


**Fig. 1**. Consecutive frames of an original (top) and two distorted (mid and bottom) video sequences and the histograms (solid curves) of the imaginary parts of $L_2(s,p)$. For comparison, the fitting model [Eq. (6)] of the original sequence are also shown as dashed curves.

Here $u(t)$ indicates how the image positions move spatially as a function of time. $a(t)$ and $b(t)$ are both real and account for the time-varying contrast and luminance changes, respectively.

Now consider a family of symmetric complex wavelets whose "mother wavelets" can be written as a modulation of a band-pass filter $w(x)$= $g(x)\,e^{j\omega_c x}$, where $\omega_c$ 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:

$$w_{s,p}(x) = \frac{1}{\sqrt{s}}\, w\left(\frac{x-p}{s}\right) = \frac{1}{\sqrt{s}}\, g\left(\frac{x-p}{s}\right)\, e^{j\omega_c(x-p)/s}\,, \quad (2)$$

where $s \in R^+$ is the scale factor, and $p \in R$ is the translation factor. We can then compute a continuous complex wavelet transform of $f(x)$ as

$$F(s,p) = \int_{-\infty}^{\infty} f(x)\, w^*_{s,p}(x)\, dx\,. \quad (3)$$

Applying such a complex wavelet transform to both sides of Eq. (1) at a given time instance $t$, we can derive that

$$H(s,p,t) \approx F(s,p) a(t)\, e^{j(\omega_c/s)u(t)}\,. \quad (4)$$

Here $b(t)$ is eliminated because of the bandpass nature of the wavelet filters. The approximation is valid when the movement $u(t)$ is small compared to the width of the slowly varying window $g(x)$. The key observation from Eq. (4) is that the phase change of $H(s,p,t)$ is approximately a linear function of $u(t)$. We call $u(t)$ $N$-th order smooth if its ($N$+1)-th and higher order derivatives with respect to $t$ are all zeros. For instance, zero-order smooth motion implies no motion [$u(t)$ is a constant over time], first-order smooth motion corresponds to constant speed [$u'(t)$ is a constant], and second-order smooth motion leads to constant acceleration, and so on. Furthermore, if we observe $H(s,p,t)$ at consecutive time steps $t_0 + n\Delta t$ for $n = 0, 1, ..., N$, we find that the following temporal correlation function is useful to test the ($N$-1)-th order temporal motion smoothness:

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

In particular, it can be shown that when the motion is ($N$-1)-th order smooth, the imaginary part of $L_N(s,p)$ is zero. Of course, this is achieved based on the ideal formulation of Eq. (1) and the ideal assumption about temporal motion smoothness. Real natural images are expected to depart 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 departure and use it as an indication of the strength of temporal motion smoothness.

Given a video sequence, we divide it into groups of pictures (GOPs) and then decompose each image frame independently into subbands using the complex version [6] of the steerable pyramid [7]. By aligning the subbands at the same orientation and scale but across different frames, we obtain a discrete version (both in space and time) of the function $H(s,p,t)$. We then compute $\mathcal{L}_2(s,p)$ for all the coefficients within the subbands. A sample histogram of the imaginary part of $\mathcal{L}_2(s,p)$ is shown in Fig. 1. A high peak at zero is observed, demonstrating a strong prior of temporal motion smoothness. The histogram of each GOP can be well fitted with a four-parameter function given by

$$P(\theta) = \frac{1}{Z}\left\{\exp\left[-\left(\frac{|\sin[(\theta-\theta_0)/2]|}{\alpha}\right)^{\beta}\right] + C\right\}, \quad (6)$$

where $\theta \in [-\pi, \pi]$, $Z$ is a normalization constant, and the four parameters $\theta_0$, $\alpha$, $\beta$ and $C$ controls the center position, width, peakedness and the baseline of the function, respectively. In Fig. 1, it is also observed that image distortions, such as noise contamination and blur, can significantly affect temporal motion smoothness.

For each subband in each GOP, we create the histogram of the imaginary part of $L_2(s,p)$ and fit it using the model given by Eq. (6). The fitting process is optimized to minimize the Kullback-Leibler distance (KLD) [8] between the model and the observed distributions (denoted as $p_m$ and $p$, respectively). The four fitting parameters, together with the KLD between the fitting model and the true distribution [denoted as $d(p_m \parallel p)$], are included as RR features. In addition, as in [2], we use a generalized Gaussian model (GGD) to fit the marginal distribution of real wavelet coefficients, and three more parameters are created for each subband. Detailed descriptions about the GGD model and the parameters can be found in [2]. To reduce the RR data rate, two subbands are involved in our implementation. This results in a total of 16 scalar features (8 for each subband) for each GOP. These features are embedded into the video sequence using the data hiding method described in Section 3.

At the receiver side, the same GOP and wavelet decompositions are applied to the received video and the distributions of the imaginary part of $L_2(s,p)$ are estimated (denoted as $q$). Meanwhile, the embedded RR features are extracted (using the method described in Section 3) to recreate the fitting model distribution. We can then compute the KLD between the model and the distorted video distributions and denote it as $d(p_m \parallel q)$. Finally, the KLD between the original and the distorted video is estimated using

$$\hat{d}(p \parallel q) = d(p_m \parallel q) - d(p_m \parallel p)\,, \tag{7}$$

which can be easily shown to be a close approximation of $d(p \parallel q)$. In addition, as in [2], we compute the probability distortion measure between the marginal distributions of the real wavelet coefficients (see [2] for details). These distortion measures are computed and summed for all subbands and averaged over all GOPs to create the final distortion measure of the video sequence.

## 3. DATA HIDING

The 16 scalar RR features for each GOP (as described in Section 2) are encoded using a 7 bit binary representation to obtain a total of 112 bits of information. To improve robustness, these bits are further encoded using BCH(15, 2, 7) code, which can correct 2 errors for every 7 bits. This results in a total of 240 bits for each GOP. A GOP of 8 frames are grouped together into a 3D volume and a 3D discrete cosine transform (3D-DCT) is applied globally to the whole volume. Because of the energy compaction property of DCT, most of the signal energy is concentrated in the low-frequency DCT components. We select a subset of DCT coefficients and embed every information bit into one DCT coefficient using a quantization index modulation (QIM) method [9], which allows for blind decoding (i.e., the original video is not needed in decoding the hidden messages). The information embedding operation for a single bit can be written as

$$c_q = Q(c + d(m)) - d(m) \equiv Q^m(c), \tag{8}$$

where $c_q$ is the marked coefficient, $Q(\cdot)$ is the base quantization operator with step size $\triangle$, and $d(m)$ is a dithering operator given by

$$d(m) = \begin{cases} -\Delta/4, & \text{if } m = 0 \\ \Delta/4, & \text{if } m = 1 \end{cases}. \tag{9}$$

The value of $\triangle$ is tuned such that the embedding of information is imperceptible. The locations of the DCT coefficients with information embedded are regarded as the embedding key that are shared between the transmitting and the receiving ends.

At the receiver side, 3D-DCT is applied and an embedded bit is extracted from the distorted DCT coefficient $c_d$ using the minimum distance criterion:

$$\hat{m}(c_d) = \arg\min_{m \in \{0,1\}} \|c_d - Q^m(c_d)\|\,. \tag{10}$$

The extracted bits are then decoded with BCH decoding and dequantized to create the decoded RR feature parameters, which are channelized to the VQA system, resulting in a distortion measure.

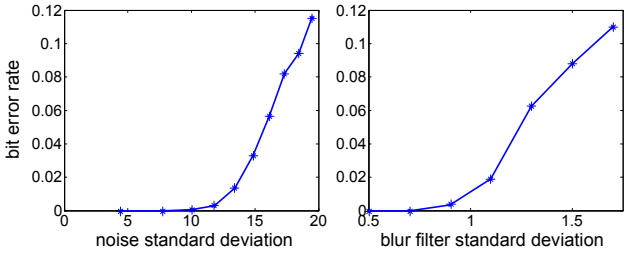


**Fig. 2**. Robustness test of the data hiding algorithm.

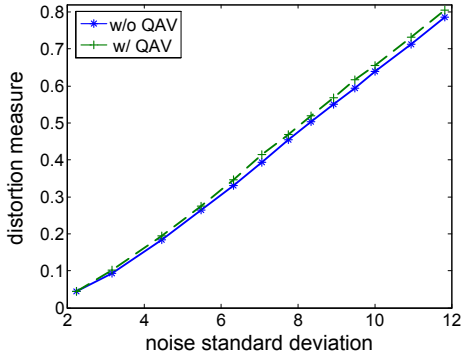


**Fig. 3**. Distortion measure for white Gaussian noise contamination with (solid curve) and without (dashed curve) QAV encoding.

## 4. SIMULATION RESULTS

We first test the robustness of the data hiding algorithm. Fig. 2 shows the detecting bit error rate (BER) under white Gaussian noise contamination and Gaussian blur, where we define the image distortion levels using the noise standard deviation and the standard deviation of the blur filter, respectively. The BER values are given without BCH coding, which can further improve the robustness.

The quality assessment algorithm is tested in two cases. In the first case, the distortions are applied directly to the original video sequences. In the second case, the original sequences are converted into QAV sequences (by feature extraction and data embedding) before the distortions are applied. It is important to verify that the distortion/quality measurement does not differ significantly for these two cases. This is because the data embedding process (though only

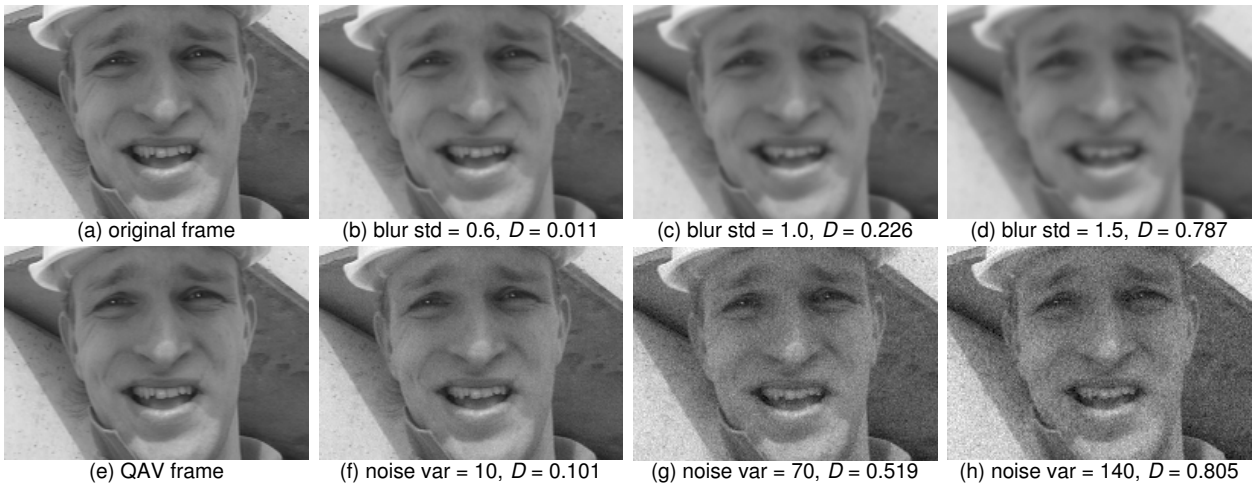


**Fig. 5**. An original video frame (a) encoded to a QAV frame (e) and passes through different levels of Gaussian blur distortion [(b)-(d)] and white Gaussian noise contamination [(f)-(h)]. The proposed distortion measures are given by $D$.

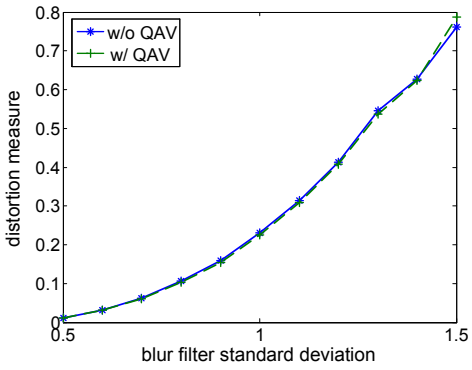


**Fig. 4**. Distortion measure for Gaussian blur distortion with (solid curve) and without (dashed curve) QAV encoding.

causes invisible changes to the video) may potentially change the statistical features of the video, while the quality assessment algorithm may rely on certain image statistics to evaluate video quality. Figs. 3 and 4 provide the results of the proposed VQA measure for both cases for white Gaussian noise and Gaussian blur distortions, respectively. The results are not only consistent with the degree of image distortions. They also show that the QAV encoding does not significantly affect the quality assessment performance because the two curves appear to be very close to each other in both figures. Fig.5 shows a sample QAV frame and the distortion measurement results for different types and degrees of image distortions.

## 5. CONCLUSION

We introduce the concept of QAV and provide a first implementation, which includes a novel RR VQA method and a 3D-DCT based data hiding algorithm. It is worth mentioning that the proposed approach does not assume any knowledge about the image distortions occurred between the original and the distorted video sequences. The simulation results lead us to believe that it has the potentials to be generalized for a wide range of image distortion types and levels. Future work includes testing the system with a larger variety of image distortion types and designing new algorithms to partially "repair" distorted QAV using the RR features.

## 6. REFERENCES

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

[2] Z. Wang, G. Wu, H. R. Sheikh, E. P. Simoncelli, E. Yang, and A. C. Bovik, "Quality-aware images," *IEEE Transactions on Image Processing*, vol. 15, pp. 1680–1689, June 2006.

[3] O. Sugimoto, R. Kawada, M. Wada, and S. Matsumoto, "Objective measurement scheme for perceived picture quality degradation caused by MPEG encoding without any reference pictures," *Proc. SPIE*, vol. 4310, pp. 932–939, 2001.

[4] M. C. Q. Farias, S. K. Mitra, M. Carli, and A. Neri, "A comparison between an objective quality measure and the mean annoyance values of watermarked videos," in *Proc. IEEE Int. Conf. Image Proc.*, Rochester, Sept. 2002.

[5] P. Campisi, M. Carli, G. Giunta, and A. Neri, "Blind quality assessment system for multimedia communications using tracing watermarking," *IEEE Trans. Signal Processing*, vol. 51, no. 4, pp. 996–1002, Apr. 2003.

[6] J. Portilla and E. P. Simoncelli, "A parametric texture model based on joint statistics of complex wavelet coefficients," *Int'l Journal of Computer Vision*, vol. 40, no. 1, pp. 49–71, December 2000.

[7] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger, "Shiftable multi-scale transforms," *IEEE Trans Information Theory*, vol. 38, no. 2, pp. 587–607, March 1992, Special Issue on Wavelets.

[8] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley-Interscience, New York, 1991.

[9] B. Chen and G. W. Wornell, "Quanitization index modulation: A class of provably good methods for digital watermarking and information embedding," *IEEE Transactions on Information Theory*, vol. 47, pp. 1423–1443, May 2001.