A REDUCED-REFERENCE VIDEO STRUCTURAL SIMILARITY METRIC BASED ON NO-REFERENCE ESTIMATION OF CHANNEL-INDUCED DISTORTION

A. Albonico, G. Valenzise, M. Naccari, M. Tagliasacchi, S. Tubaro

Dipartimento di Elettronica e Informazione - Politecnico di Milano
P.za Leonardo da Vinci 32, 20133 Milano, Italy
{valenzise,naccari,tagliasa,tubaro}@elet.polimi.it

ABSTRACT

The reduced-reference (RR) approximation of a full-reference (FR) video quality assessment method is a convenient way to build evaluation metrics which are both intrinsically well correlated with human judgments and feasible to implement in a network scenario, without the need to explore the perceptual significance of new video features through mean opinion score tests. In this paper, we propose a RR approximation of the video structural similarity index (VSSIM), a FR metric which is known to be well descriptive of the video quality perceived by users. We focus on the visual degradation produced by channel transmission errors: first, at the encoder, a small set of salient structural video features is assembled and transmitted through the RR channel to the end-user; then, at the decoder the feature vector is combined with a fine-granularity, no-reference estimate of the channel-induced distortion to produce the VSSIM approximation. By uniformly quantizing the feature vector and compressing it using a context-adaptive, variable length encoder, we show that good correlation coefficients with ground-truth VSSIM ($\rho = 0.85$) may be achieved spending, respectively, less than 12 and 27 kbps for a video sequence with CIF or SD resolution.

Index Terms— Video signal processing, Video coding

1. INTRODUCTION

The estimation of the perceived quality of video sequences is a crucial task when visual contents are transmitted over communication networks, where annoying artifacts in the received video stream may be introduced due to channel errors or jitter. In the last decades, a great deal of effort has been made to develop objective video quality assessment techniques [1] which resemble perceptual judgments given by human observers more accurately than the traditional PSNR, whose correlation with Mean Opinion Score (MOS) tests is notoriously poor [2]. Most of these techniques – known as *full reference* (FR) methods because, to be computed, they require the complete availability of the original (reference) signal at the decoder – have been recently tested by the video quality experts group (VQEG) [3] in their full reference television (FRTV) phase 2 tests [4]. In practice, FR methods are hardly implementable by a decoder in a network scenario, since the end-users do not have access to the original frames at their terminals. Therefore, in the literature two alternative solutions have been proposed to estimate the quality at the decoder: no-reference (NR) methods and reduced-reference (RR) methods.

The presented work was developed within VISNET II, a European Network of Excellence (http://www.visnet-noe.org), funded under the European Commission IST FP6 programme.

In NR methods, the end-user tries to infer the distortion of the received frames from just the reconstructed video available at the output of the decoder or from the transmitted bitstream itself, without any sort of access to the original video. These techniques can be easily integrated into existing broadcasting systems, but generally lack in estimation accuracy. Specific NR methods have been proposed to estimate the distortion introduced by video coding [5, 6] or to take in consideration the effects of channel losses [7]. We have recently proposed a NR system [8] for H.264/AVC video sequences which models the effect of temporal concealment to estimate the channel-induced distortion at the decoder; this method has been subsequently expanded in [9] and, since it acts as a substratum for the proposed quality assessment system, it will be briefly summarized in Section 2.2.

In contrast with NR methods, RR techniques can achieve a more accurate distortion estimation by using some feature vector extracted from the original bitstream that is made available at the decoder side through an ancillary, low bit-rate, noiseless data channel. RR methods can be either designed independently with respect to pre-existing FR methods [10, 11] or as an approximation of some FR metrics as in [12]. The methods developed using the first approach are specifically conceived to target the peculiarities of the RR scenario, but need to be matched against MOS tests for their effectiveness to be validated. On the other hand, for RR techniques which approximate FR methods it is sufficient to exhibit good correlation between RR scores and FR ground-truth data, since the correlation between FR scores and MOS is supposed to have already been assessed elsewhere.

In this paper, we propose a RR method which takes inspiration from the second approach described above in the fact that it approximates a popular objective perceptual quality assessment method, the video structural similarity index (VSSIM) [13], whose correlation with MOS has been exhaustively proved. In order to do this, we collect a few significant features from each video frame and transmit them to the decoder, where they are used to estimate the VSSIM by leveraging the side information provided by a NR distortion estimator. To keep the size of the RR information small, different techniques have been used in the literature, such as non-linear quantization [12] or distributed source coding [14]. In the proposed system we uniformly quantize the DCT-transformed DPCM residuals, and encode them with a variable-length, context-adaptive entropy coder, as the CABAC module of H.264/AVC [15].

The rest of this paper is organized as follows: Section 2 describes the building blocks of the proposed RR quality assessment system; Section 3 illustrates the performance of our method in terms of correlation with the VSSIM and rate spent for the RR information; finally, Section 4 gives some concluding remarks.

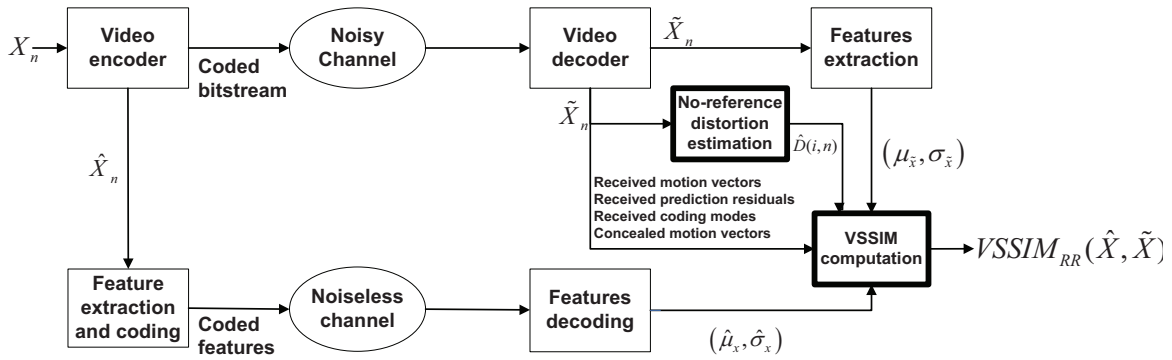

Fig. 1. The building blocks of the proposed RR quality assessment system.

2. SYSTEM DESCRIPTION

The proposed video quality assessment system is illustrated in Figure 1. The original video frame X_n is encoded and transmitted through a noisy channel that drops the coded packets according to a given packet loss rate (PLR). During the encoding process, the error-free reconstructed frame $\hat{X}_n$ is fed into the *features extraction and coding* module which computes, for the luminance component of each macroblock i, belonging to frame n, its average ($\mu_x(i,n)$) and standard deviation ($\sigma_x(i,n)$). The ensemble of $\mu_x(i,n)$ and $\sigma_x(i,n)$ represents the RR feature vector which is encoded and transmitted through the noiseless RR channel. At the receiver terminal, the transmitted video is decoded and fed, together with side information extracted during the decoding process, into the *no-reference distortion estimation* module [8]. The output of this module consists in an estimate $\hat{D}_n^i$ of the mean square error (MSE) between the error-free reconstructed macroblock and its decoded counterpart. From the decoded frame $\tilde{X}_n$, the features $\mu_{\tilde{x}}(i,n)$ and $\sigma_{\tilde{x}}(i,n)$ are extracted and, together with the ones received through the RR channel, are used in the *VSSIM computation* module to calculate a RR approximation of the VSSIM at the frame level. Finally, with the additional information provided by the motion vectors, the frame-level VSSIM's are aggregated to compute the approximate metric at the sequence level.

2.1. VSSIM approximation

The structural similarity index (SSIM) was initially proposed for still images [16] and successively extended to video sequences with the name of video SSIM or VSSIM [13]. The core idea of the structural similarity metric is to compare the reference and distorted video signals from a structural distortion point of view by computing a product of independent terms, namely the luminance, the contrast and the similarity between two images. This leads to the following SSIM formula between signals $\mathbf{x}$ and $\tilde{\mathbf{x}}$:

$$\mathrm{SSIM}(\mathbf{x},\tilde{\mathbf{x}}) = \frac{(2\mu_x\mu_{\tilde{x}} + C_1)(2\sigma_{x\tilde{x}} + C_2)}{(\mu_x^2 + \mu_{\tilde{x}}^2 + C_1)(\sigma_x^2 + \sigma_{\tilde{x}}^2 + C_2)}, \qquad (1)$$

where μ_x and $\mu_{\tilde{x}}$ are, respectively, the mean values of $\mathbf{x}$ and $\tilde{\mathbf{x}}$, σ_x^2 and $\sigma_{\tilde{x}}^2$ represent their variances and $\sigma_{x\tilde{x}}$ is the covariance between the two signals. The two constants C_1 and C_2 are added to avoid possible division by zero, and are selected as in [13]. In the full-reference VSSIM, the signals $\mathbf{x}$ and $\tilde{\mathbf{x}}$ are the luminance values of 8×8 sliding windows, moved pixel by pixel on, respectively, the error-free frame $\hat{X}_n$ and the decoded frame $\tilde{X}_n$. Equation (1) is the basic ingredient to calculate the video SSIM, which is obtained by aggregating the SSIM either at the frame or sequence level, taking into account that dark regions of a frame do not attract fixation and errors in fast moving scenes are less annoying than errors in a still or slowly moving background. Clearly, (1) can be computed only when both the original signal $\mathbf{x}$ and its degraded version $\tilde{\mathbf{x}}$ are available at the decoder. With respect to this full-reference statement of the VSSIM metric, the proposed RR quality assessment algorithm differs in the following points:

1) The feature vector, consisting of μ and σ, is computed over a disjoint grid of 16×16 macroblocks. We have found in our experiments (see Figure 2) that the VSSIM computed in this way closely approximates the one that would have been obtained by implementing it as described in [13].

2) At the receiver side, the original, error-free frame $\hat{X}_n$ is not available for comparison and thus the values μ_x and σ_x are unknown. To produce the RR approximation we first transform μ_x and σ_x for each frame in the DCT domain, to exploit the spatial correlation between features; the transformed coefficients are passed through a DPCM encoder and the prediction residuals are uniformly quantized and entropy-coded with a context-adaptive, arithmetic encoder. At the decoder, the reconstructed features are $\hat{\mu}_x$ and $\hat{\sigma}_x$.

3) In the RR approximation, also the covariance term $\sigma_{x\tilde{x}}$ cannot be calculated. However, by the covariance definition and after a little algebra we can write:

$$\hat{\sigma}_{x\tilde{x}}(i,n) = \frac{1}{2}\big[(\hat{\sigma}_x(i,n))^2 + (\sigma_{\tilde{x}}(i,n))^2 + (\hat{\mu}_x(i,n) - \mu_{\tilde{x}}(i,n))^2 - \hat{D}(i,n)\big], \qquad (2)$$

where the term $\hat{D}(i,n)$ denotes the estimate of the channel induced distortion provided by the *No-reference distortion estimation* algorithm proposed in [8].

4) The quantization of μ_x and σ_x, as well as the estimation error in $\hat{D}(i,n)$, have the effect of introducing a bias in the estimated VSSIM with respect to the original metric. Through experimental tests we have verified that this bias depends more on the quantization step used for the means rather than on the one used for standard deviations; furthermore, it comes out that this bias is independent from the tested sequence or from the specific degradation pattern of the received video. To overcome this bias, we use have estimated a single look-up table containing the estimated biases for each quantization step size used to encode μ_x and σ_x.

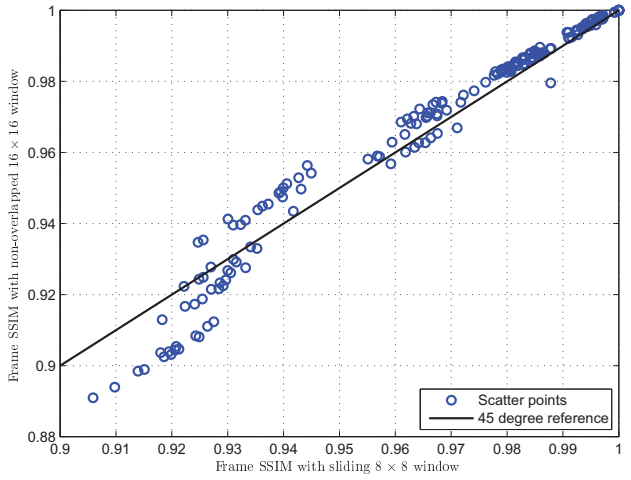

Fig. 2. The VSSIM estimated over 16×16 disjoint blocks vs. the original full-reference metric.

2.2. No-reference distortion estimation

In [8] we proposed a NR quality monitoring algorithm to estimate the distortion induced by channel losses. The estimation process is specifically designed for motion-compensated predictive (MCP) video codecs as those belonging to the MPEG-x and H.26x families. The algorithm provides, for each macroblock i of frame n, an estimate of the channel induced distortion $\hat{D}(i,n)$ assuming the MSE as distortion measure. The estimation process explicitly accounts for the distortion induced by temporal error concealment, the lack of both motion vectors and prediction residuals as well as the error propagation due to the predictive nature of modern video codecs. This technique has been improved and further extended in [9] in order to account for the distortion induced by spatial concealment (i.e. the concealment typically used in intra coded slices [17]). The overall algorithm can be easily integrated in any MCP decoder compliant with the adopted standard. Finally the estimated $\hat{D}(i,n)$ is used to compute the RR VSSIM approximation as described in Section 2.1.

3. RESULTS

The performance of the proposed RR VSSIM approximation has been tested by simulating the errors produced by the transmission of a video sequence over a broadcasting network. Before discussing the results obtained, we briefly describe the source coding and transmission conditions adopted in the paper.

3.1. Source and transmission conditions

In our experiments a CIF and a 625-SD (standard definition) video sequences, respectively *Soccer* and *Mobile & Calendar*, have been coded with the H.264/AVC video coding standard with the Baseline profile and the reference software in [17]. The 300 frames of the *Soccer* sequence have a frame rate of 30 Hz, with an intra frame every 15 frames, and are coded at 256 kbps. Conversely, the 220 frames of the *Mobile & Calendar* sequence are coded at 25 fps, with an intra frame every 15 frames, spending a bit-rate of 4 Mbps. Every coded frame is divided into slices where each coded slice contains a horizontal row of macroblocks and corresponds to a transmitted packet. The

QP(μ) \ QP(σ)	25	30	35	40	45	51
20	0.98	0.97	0.97	0.96	0.97	0.94
25	0.97	0.97	0.97	0.95	0.95	0.94
30	0.94	0.94	0.94	0.93	0.91	0.88
35	0.91	0.88	0.86	0.83	0.83	0.83
40	0.9	0.87	0.86	0.81	0.78	0.77
45	0.9	0.86	0.85	0.8	0.77	0.75
51	0.88	0.84	0.84	0.75	0.71	0.69

Table 1. Linear correlation coefficients for the *Soccer* sequence.

QP(μ) \ QP(σ)	25	30	35	40	45	51
20	0.97	0.95	0.92	0.91	0.91	0.90
25	0.96	0.94	0.91	0.91	0.89	0.88
30	0.94	0.94	0.90	0.86	0.83	0.82
35	0.94	0.94	0.89	0.86	0.83	0.82
40	0.93	0.93	0.89	0.85	0.82	0.82
45	0.93	0.92	0.88	0.84	0.81	0.81
51	0.92	0.92	0.83	0.79	0.78	0.77

Table 2. Linear correlation coefficients for the *Mobile* sequence.

packets are then packetized according to the real-time transfer protocol (RTP). The simulated error-prone channel drops coded packets according to a packet loss rate (PLR) equal to 2.5%, with error patterns generated according to a two state Gilbert's model [18] with average burst length of 3.1 packets. To encode the feature vector as described in Section 2.1 we quantize DCT-transformed DPCM residuals of the features. The step-size Δ of the dead-zone quantizer follows an exponential law through a QP parameter as in the H.264/AVC standard:

$$\Delta = 2^{\frac{QP-4}{6}}. \quad (3)$$

As for the entropy coding of the quantized transformed residuals, we use the context-adaptive binary arithmetic coder (CABAC) module of the H.264/AVC standard [15].

3.2. Experimental results and discussion

In order to measure the accuracy of the VSSIM estimation, we have measured the Pearson's correlation coefficient ρ between the full-reference metric and its reduced-reference approximation. Tables 1 and 2 show the correlation coefficients for the two tested video sequences, when different quantization parameters $QP(\mu)$ and $QP(\sigma)$ are used, respectively, to determine the quantization step size used in the encoding of μ_x and σ_x as in (3). For both sequences, as QP increases the correlation with the FR metric degrades. Furthermore, the quantization of the elements of the feature vector produces different effects on the correlation, as the VSSIM approximation is more tolerant to quantization noise in the standard deviations rather than in the means: thus, it is convenient to quantize more heavily the values σ_x in order to obtain an equivalent ρ at a lower RR rate.

Tables 3 and 4 report the rates spent for encoding the feature vector for each couple $(QP(\mu), QP(\sigma))$. Passing from the CIF sequence to the the SD video, the rates increase by a factor between 2.5 and 3, which is less than the ratio between the number of features for the SD resolution and the number of features of the CIF sequence. In addition, if we match Tables 3-4 with the previous ones (Tables

QP(μ) \ QP(σ)	25	30	35	40	45	51
20	47.24	39.26	33.12	29.2	27.67	27.23
25	39.41	31.43	25.29	21.37	19.84	19.4
30	33.13	25.15	19.01	15.09	13.56	13.12
35	29.08	21.1	14.96	11.04	9.51	9.07
40	26.85	18.87	12.73	8.81	7.28	6.84
45	25.91	17.93	11.79	7.87	6.34	5.9
51	25.31	17.33	11.19	7.27	5.74	5.3

Table 3. Rates needed to encode the RR information for the *Soccer* sequence.

QP(μ) \ QP(σ)	25	30	35	40	45	51
20	132.79	115.81	101.04	89.64	81.72	78.49
25	104.12	87.14	72.37	60.97	53.05	49.82
30	82.15	65.17	50.4	39	31.08	27.85
35	73.08	56.1	41.33	29.93	22.01	18.78
40	69.66	52.68	37.91	26.51	18.59	15.36
45	68.21	51.23	36.46	25.06	17.14	13.91
51	67.59	50.61	35.84	24.44	16.52	13.29

Table 4. Rates needed to encode the RR information for the *Mobile* sequence.

1-2), we can notice that good correlation coefficients ($\rho \geq 0.85$) may be attained spending a rate that is less than 12 kbps for the CIF sequence and no more than 27 kbps for the SD sequence.

4. CONCLUSIONS

In this paper we have proposed a reduced-reference approximation of the video structural similarity index, which leverages a no-reference, fine-granularity distortion estimation computed at the decoder to produce an accurate, objective estimation of the perceived quality of a video sequence. We have focussed to channel-induced degradations, but our system is intrinsically extensible to other kinds of errors by further expanding the no-reference module at the decoder. A possible development of the system will be the exploration of alternative coding paradigms for the feature vector, as distributed source coding, to further reduce the bandwidth required for the RR channel.

5. REFERENCES

[1] Z. Wang, H.R. Sheikh, and A.C. Bovik, *The Handbook of Video Databases: Design and Applications*, chapter 41: Objective video quality assessment, pp. 1041–1078, CRC Press, 2003.

[2] B. Girod, "What's wrong with mean-squared error?," *MIT Press Cambridge, MA, USA*, 1993.

[3] "The Video Quality Expert Group web site," http://www.its.bldrdoc.gov/vqeg.

[4] P. Corriveau and A. Webster, "Final report from the video quality experts group on the validation of objective models of video quality assessment, phase II," Tech. Rep., Video Quality Expert Group, July 2003.

[5] Q. Li and Z. Wang, "A no-reference perceptual blockiness metric," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Las Vegas, NV, USA, April 2008.

[6] I. P. Gunawan and M. Ghanbari, "Reduced-reference video quality assessment using discriminative local harmonic strength with motion consideration," vol. 18, no. 1, pp. 71–83, January 2008.

[7] A. R. Reibman, V. A. Vaishmpayan, and Y. Sermadevi, "Quality monitoring of video over a packet network," *IEEE Trans. Multimedia*, vol. 6, no. 2, pp. 327–334, April 2004.

[8] M. Naccari, M. Tagliasacchi, F. Pereira, and S. Tubaro, "No-reference modeling of the channel induced distortion at the decoder for H.264/AVC video coding," in *Proc. IEEE Int. Conf. Image Processing*, San Diego, CA, USA, Oct 2008.

[9] M. Naccari, M. Tagliasacchi, and S. Tubaro, "No-reference video quality monitoring for H.264/AVC coded video," *IEEE Trans. Multimedia (submitted)*, 2008, available at http://home.dei.polimi.it/naccari/nrtransMMnaccariV1.pdf.

[10] T. Yamada, Y. Miyamoto, M. Serizawa, and H. Harasaki, "Reduced-reference based video quality-metrics using representative-luminance values," in *Third International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, Scottsdale, AZ, USA, 2007.

[11] Z. Wang and E.P. Simoncelli, "Reduced-reference image quality assessment using a wavelet-domain natural image statistic model," *Human Vision and Electronic Imaging X Conference, San Jose, CA, January*, pp. 17–20, 2005.

[12] M. Pinson and S. Wolf, "Low bandwidth reduced reference video quality monitoring system," in *First International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, Scottsdale, AZ, USA, 2005.

[13] Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion meausure," *Signal Processing: Image Communication*, vol. 19, no. 2, pp. 121–132, 2004.

[14] G. Valenzise, M. Naccari, M. Tagliasacchi, and S. Tubaro, "Reduced-reference estimation of channel-induced video distortion using distributed source coding," in *Proc. ACM Int. Conf. on Multimedia*, Vancouver, Canada, 2008.

[15] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 560–576, July 2003.

[16] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structual similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, April 2004.

[17] Joint Video Team (JVT), "H.264/AVC reference software version JM12.3," downloadable at http://iphome.hhi.de/suehring/tml/download/.

[18] E. N. Gilbert, "Capacity of a burst-noise channel," *Bell System Technical Journal*, vol. 39, pp. 1253–1266, 1960.