Reduced-Reference Estimation of Channel-Induced Video Distortion using Distributed Source Coding*

G. Valenzise, M. Naccari, M. Tagliasacchi, S. Tubaro
{valenzise, naccari, tagliasa, tubaro}@elet.polimi.it
Dipartimento di Elettronica e Informazione, Politecnico di Milano
P.za Leonardo da Vinci 32, 20133 Milano, Italy

ABSTRACT

Channel-induced distortion estimation is an important aspect in the delivery of video contents over IP networks: the QoS requirements of both content providers and content users conflict with the intrinsic best-effort nature of packet-switched networks, which may introduce annoying artifacts in the received streams due to channel errors or jitter. In this paper we propose a Reduced-Reference video quality assessment method, based on objective quality metrics, which enables distortion estimation at the macroblock level. The content provider transmits a small feature vector for each frame, starting from random projections computed for each macroblock. In order to reduce the bit rate of the transmitted feature vector, we encode it using Distributed Source Coding (DSC) tools. The content user decodes the feature vector using the received sequence as side information. Additionally, the end-user may take advantage of some prior information about the support of the errors in the frame in such a way that the required bit length of the transmitted feature vector is further reduced. In our experiments, using 4 random projections, the use of DSC enables a bit saving of 70% w.r.t. scalar quantization and transmission of the original feature vector; when also the a priori error map is available at the decoder, the average length of the transmitted partial reference can be further reduced by another 5% of average.

Categories and Subject Descriptors

I.4.2 [**Image Processing and Computer Vision**]: Compression (Coding)

General Terms

MEASUREMENT, PERFORMANCE

*This work has been partially sponsored by UE under VISNET II Network of Excellence.

ACM Multimedia '08 Vancouver, BC, Canada

Keywords

Distortion Estimation, Quality Assessment, DSC

1. INTRODUCTION

The use of IP networks for the delivery of multimedia contents is gaining an increasing popularity as a means for broadcasting media files from a content provider to many end-users. In the case of video contents, for instance, packet-switched networks are used to distribute video programs in IPTV applications. Typically, these kind of networks provide only best-effort services, i.e. there is no guarantee that the content will be delivered without errors to the final user. In practice, this implies that the video sequence received by the end-user may contain artifacts due to packet losses or jitter, in addition to the distortion introduced by possible quantization of the original bit-stream at the content producer side. On the other hand, the content provider and the users generally stipulate a Service Level Agreement (SLA) that fixes an expected perceived quality at the end-user terminal: the provider imposes a price to the customers for assuring the agreed Quality of Service (QoS), and pays a penalty if the SLA is unfulfilled.

For this reason, it is fundamental in IP networks to assess the visual quality of distributed video contents. In the case of video quality assessment, a commonly used objective metric is the PSNR, which is based on the Mean Square Error (MSE) between the original and the received frames. Although it is known that PSNR could be poorly correlated with the actual perceived quality, measured through Mean Opinion Score (MOS) tests, recent studies [3] have shown that applying a region-of-interest weighting to the MSE computed at a local scale (e.g. macroblock) may considerably improve the fidelity of the estimation. However, since the end-users do not have access to the original frames at their terminals, it is difficult to directly compute the PSNR at the receiver. In the literature, the techniques for estimating the distortion at the end-user side fall in two main categories: No-Reference (NR) and Reduced-Reference (RR) methods [8].

In NR methods, the end-user does not have any information about the original video stream, and tries to infer the distortion of the received frames from the reconstructed video available at the output of the decoder or from the transmitted bit-stream itself. These techniques can be easily integrated into existing broadcasting systems, but generally lack in estimation accuracy. The NR method in [5] evaluates the distortion introduced by video coding by automatically and perceptually quantifying blocking artifacts

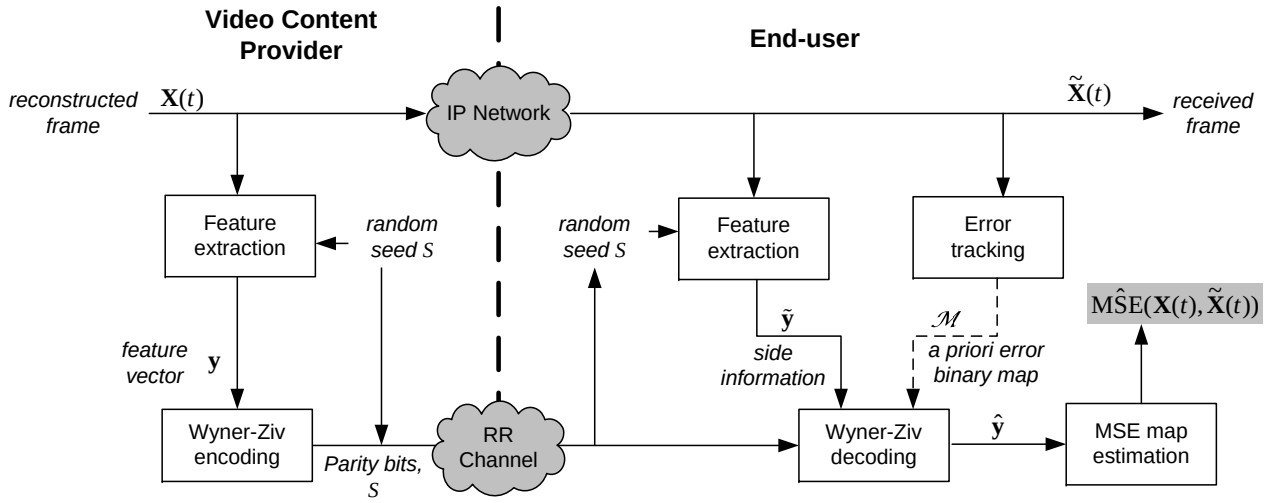

Figure 1: Block diagram of the proposed tampering localization scheme

of the DCT coded macroblocks. When also channel losses have to be considered, NR techniques extract detailed local information regarding the spatial impact and the temporal extent of the packet loss, as in the work of Reibman et. al. [6]. If some error concealment technique is available at the decoder, the distortion can be determined from the macroblocks for which the concealment is judged to have been ineffective, as in [11].

On the other hand, RR methods can achieve a more accurate distortion estimation than NR approaches without assuming the complete availability of the reference signal, by using some feature vector extracted from the original bitstream that is made available at the decoder side through an ancillary, low bit-rate data channel. The RR method in [4] uses local harmonic strength descriptors to evaluate video compression distortion. In order to keep the size of the transmitted feature vector as small as possible, Chono et. al. [2] employ Distributed Source Coding (DSC) tools to efficiently encode the partial reference data, in the scenario of image delivery with distortion due to compression. Some portion of the whitened spectrum of the original image is transmitted as the feature vector, according to a predetermined admissible image quality; at the end-user terminal, the decoder reconstructs the feature vector using the received image as side information and estimates the PSNR from the feature measurements.

In this work, we propose an objective RR video quality assessment scheme that uses distributed source coding to efficiently transmit a small hash to the end-users, taking inspiration from the PSNR estimation technique for compressed images described in [2]. With respect to [2], we focus on channel-induced distortion estimation. In our scheme, outlined in Section 2, the content provider generates a small hash for each frame, starting from a few random projections computed from each macroblock; the hash is then encoded using DSC tools in order to reduce its payload. Moreover, if a coarse knowledge of the error support in a frame is available at the decoder (e.g. by means of an error-tracking module such as the one devised in [1]), an a priori map of the degraded macroblocks may be used to aid hash decoding, reducing th bit-rate. We show in Section 3 the results of our experiments on different video sequences for different numbers of random projections, whether using or not some a priori map.

2. SYSTEM OVERVIEW

Figure 1 illustrates the architecture of the proposed quality assessment system. The content provider encodes the video program and broadcasts it over an IP network to the content users. The frames $\tilde{\mathbf{X}}(t)$ reconstructed at the decoder differ from the reconstructed frames $\mathbf{X}(t)$ at the encoder side because of channel-induced errors (e.g. lost slices) and drift caused by motion propagation. For each reconstructed frame $\mathbf{X}(t)$, the content provider extracts a feature vector $\mathbf{y}$ consisting of a number of pseudo-random projections for each macroblock. The feature vector is given in input to the Wyner-Ziv encoder, which is based on Low-Density Parity Check (LDPC) codes [7]; the parity bits produced by the encoder are sent as part of the reduced reference information through the RR channel, together with the pseudo-random seed S. At the end-user side, the decoder extracts a feature vector $\tilde{\mathbf{y}}$ with the same procedure carried out at the content provider side, using the same random seed S received through the RR channel; this vector is used as side information to reconstruct an approximation of $\mathbf{y}$, obtaining an estimate $\hat{\mathbf{y}}$ of the feature vector of the frame reconstructed at the encoder. If some error tracking module is available at the decoder, an a priori map $\mathcal{M}$ can be built in order to aid Wyner-Ziv decoding. Finally, with the reconstructed $\hat{\mathbf{y}}$, the end-user can estimate the MSE between $\mathbf{X}(t)$ and $\tilde{\mathbf{X}}(t)$, with a macroblock granularity.

2.1 Feature extraction and MSE map estimation

Feature extraction consists in computing $m \geq 1$ pseudo-random projections for each macroblock and collecting them in a feature vector $\mathbf{y} \in \mathbb{R}^M$, $M = m \cdot N$, where N is the number of macroblocks in a frame. Thus, each entry of $\mathbf{y}$ is a dot product between one rasterized macroblock $\mathbf{x}^{(k)}$ and a random vector $\mathbf{a}$:

$$y_i = \mathbf{a}^T \mathbf{x}^{(k)}, \tag{1}$$

where we set $\|\mathbf{a}\| = 1$ in such a way that the macroblock energy is conserved. An estimate $\hat{\text{MSE}}_k$ of the Mean Square Error for a macroblock k can be computed from the feature vector $\tilde{\mathbf{y}}$ extracted from the decoded frame and from the reconstructed vector of random projections $\hat{\mathbf{y}}$ as:

$$\hat{\text{MSE}}_k = \frac{1}{m} \sum_{i=mk}^{m(k+1)-1} (y_i - \tilde{y}_i)^2 , \quad (2)$$

while the estimated MSE at the frame level is the average MSE of the macroblocks:

$$\hat{\text{MSE}}(\mathbf{X}(t), \tilde{\mathbf{X}}(t)) = \frac{1}{N} \sum_{k=1}^{N} \hat{\text{MSE}}_k . \quad (3)$$

2.2 Wyner-Ziv feature vector coding

Instead of quantizing and transmitting over the RR channel the feature vector $\mathbf{y}$, we make use of Distributed Source Coding principles [10] to reduce the rate requirements of the hash. From experimental observations, we observe that $\mathbf{y}$ and $\tilde{\mathbf{y}}$ are well correlated: the higher is the correlation between the source and the side information (i.e. the smaller the ratio between the energy of the error and that of the original frame), the larger is the coding gain enabled by the application of DSC tools. To implement the WZ codec, we use LDPC Accumulate codes [7]. The number of bitplanes passed as input to the LDPC encoder for the feature vector $\mathbf{y}$ corresponds to the number of bits that would have been needed if a simple uniform scalar quantizer had been designed in such a way that the signal-to-quantization noise ratio (SQNR) is fixed to 30 dB. The LDPC returns a syndrome for each bitplane.

The LDPC decoder iteratively reconstructs an estimate $\hat{\mathbf{y}}$ of the original feature vector $\mathbf{y}$, starting from the most significant bitplane and conditionally decoding the subsequent bitplanes. In our setting, we use an ideal rate allocation which requires exactly the minimum necessary number of bits for decoding the feature vector of a frame through a feedback channel; realistically, a 10% to 20% increase in the required bit rate should be taken in consideration if further modeling of the source statistics is introduced to remove feedback.

2.3 Error tracking and prior error map incorporation

To conditionally decode feature vector bitplanes, information about the correlation between the source and the side information is taken into account, by considering the variance σ_z^2 of the correlation noise $\mathbf{z} = \mathbf{y} - \tilde{\mathbf{y}}$. This information is spatially non-stationary, i.e. it may change from macroblock to macroblock, and can be integrated with some prior knowledge of the spatial support of the errors in the received frame to reduce the bit rate demand at the end-user terminal. In our implementation, we have computed a simple binary error map $\mathcal{M}$ by performing error tracking at the decoder, in a similar fashion to [1]. For each frame $\tilde{\mathbf{X}}(t)$, the macroblocks marked as corrupted are the ones that belong to a lost packet (reported by the transmission protocol), plus the macroblocks predicted from regions of the previous frames $\tilde{\mathbf{X}}(t-1), \tilde{\mathbf{X}}(t-2), \ldots$ that have already been labeled as erroneous. The basic idea is to exploit this prior map to assign two possible values of σ_z^2 to the projections of each macroblock: a "high" value, corresponding to the average variance of the corrupted macroblocks, and a "low" value for the uncorrupted macroblocks. Since our simplified error tracking scheme does not account for errors due to intra-prediction or to de-blocking filters, the "low" σ_z^2 level is not set exactly to zero, but to a small value estimated from experiments in such a way that it targets the average distortion of the macroblocks not significantly affected by channel errors.

3. EXPERIMENTAL RESULTS

The proposed objective quality assessment system has been evaluated using three test 4CIF video sequences (*Harp & piano, Mobile & Calendar* and *Rugby*). Each sequence contains 220 frames at a spatial resolution of 704×576 pixels, at a frame rate of 30 Hz. Each sequence is encoded using H.264/AVC main profile at a bit-rate of 3 Mbps, and packetized using Real-time Transport Protocol (RTP). The errors on the transmission channels are simulated according to the patterns in [9], with a Packet Loss Rate (PLR) of 3%. In order to extract the feature vector, we divide each frame in macroblocks of size 32×32, and compute m random projections as in (1).

For each sequence, we evaluate the performance of the MSE map estimation computing the average correlation coefficient ρ_{MB} over all frames between the estimated MSE at the macroblock level $\hat{\text{MSE}}$ (2) and the actual MSE. We also compute an average correlation coefficient ρ_{F} over the whole sequence at the frame level, where the estimated MSE for each frame is computed as in (3).

Figure 2 shows the two correlation coefficients for the three test video sequences, for $m = 1, 2, 4, 8$. As the number of extracted projections increases, the estimation of the MSE improves, at the cost of a larger rate requirement for the hash. It has to be noted that ρ_{MB} is in general much smaller than ρ_{F} for a given m. This is mainly due to the fact that a few large errors in the macroblock-level MSE estimation can considerably influence the resulting correlation coefficient; conversely, in the case of ρ_{F} the effect of outliers is alleviated by the frame-level averaging. From Figure 2 one can conclude that a good degree of correlation ($\rho_{\text{MB}} \geq 0.8$) may be obtained for $m \geq 4$.

Clearly, using a larger m to estimate the MSE reduces the estimation variance but calls for more bits for encoding the feature vector. By sweeping different values of m, it is possible to draw rate-correlation curves as the ones shown in Figure 3, for the sequence *Mobile*. In this graph, we have plotted the correlation coefficients ρ_{MB} and ρ_{F} for different rates, corresponding to $m = 1, 2, 4, 8$. The no Wyner-Ziv (NO-WZ) curve represents the bit rate needed for encoding the feature vector with a uniform scalar quantizer, using the same number of bits as the bitplanes used in WZ coding. Also, we report the rate spent using the binary prior map $\mathcal{M}$ described in Section 2.3 (WZ+prior), where the σ_z^2 of the uncorrupted blocks is set to 0.016, while the variance of the macroblocks with errors is $\sigma_z^2 = 1600$. Finally, for comparison, we show the rate spent when an ideal, real-valued prior map of the actual distortion is provided by an oracle (WZ+oracle). The use of WZ coding enables a bit saving w.r.t. the NO-WZ case of about 70%; if the binary prior map is introduced, the bit-rate is further reduced by another 5%, i.e. half the reduction due to the map produced by the oracle, which can be quantified in approximatively 10%. Similar results are obtained for the other test sequences; for

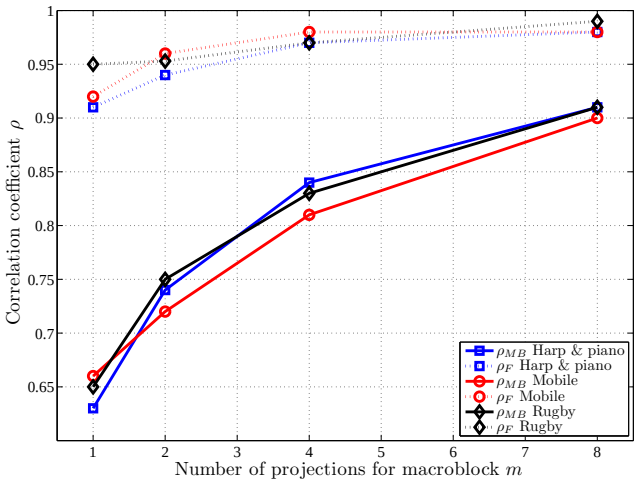

Figure 2: Correlation coefficient between the estimated and the actual MSE.

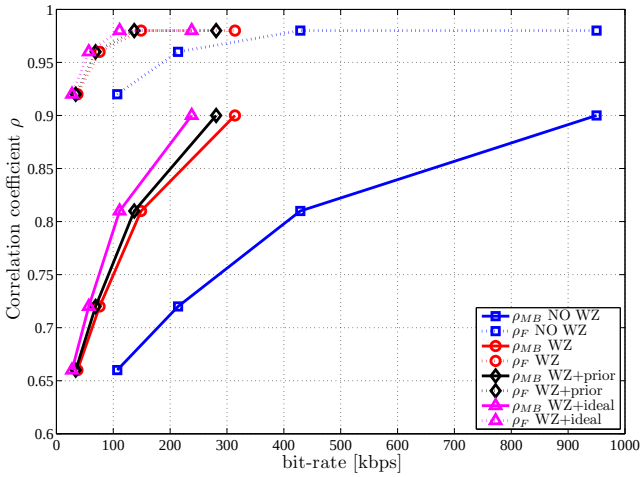

Figure 3: Correlation coefficient at different bit-rates for the hash, for the sequence *Mobile & Calendar*

convenience, Table 1 summarizes the rates spent without WZ coding and the percentage gains introduced by DSC and prior maps, for $m = 4$. It can be observed that as the sequence complexity increases (e.g. in *Rugby* there is much more motion than in *Harp & piano*), more bits have to be transmitted on the RR channel in order to obtain the same quality assessment performance, due to the higher energy of correlation noise produced by drift.

4. CONCLUSIONS

In this paper we presented an RR video quality assessment system to evaluate channel-induced distortion. The key point of the proposed method is to produce a small feature vector at the encoder, which does not contain the full information of the original frame but that can be still used to estimate a macroblock-level MSE map. The size of the feature vector must be traded off with the bit-rate availability, and to keep the hash size small we employ DSC principles and prior information about the corrupted macroblocks. At the decoder, the estimated MSE map can be used for further perceptual processing, such as region-of-interest weighting.

	R NO-WZ [kbps]	ΔR WZ [%]	ΔR WZ+PRIOR [%]	ΔR WZ+ORACLE [%]
Harp & piano	427	71.42	75.14	82.28
Mobile & calendar	429	65.27	68.07	74.10
Rugby	428.33	61.52	67.84	70.37

Table 1: Bit-Rate for the RR hash

5. REFERENCES

[1] R. Bernardini, M. Naccari, R. Rinaldo, M. Tagliasacchi, S. Tubaro, and P. Zontone. Rate allocation for robust video streaming based on distributed video coding. *To appear in Signal Processing: Image Communication*, 2008.

[2] K. Chono, Y. C. Lin, D. Varodayan, Y. Miyamoto, and B. Girod. Reduced-reference image quality estimation using distributed source coding. In *IEEE International Conference on Multimedia and Expo*, Hannover, Germany, June 2008.

[3] U. Engelke, V. Nguyen, and H. Zepernick. Regional attention to structural degradation for perceptual image quality metric design. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Las Vegas, NV, USA, April 2008.

[4] I. P. Gunawan and M. Ghanbari. Reduced-reference video quality assessment using discriminative local harmonic strength with motion consideration. *IEEE Trans. Circuits Syst. Video Technol.*, 18(1):71–83, January 2008.

[5] Q. Li and Z. Wang. A no-reference perceptual blockiness metric. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Las Vegas, NV, USA, April 2008.

[6] A. R. Reibman, V. A. Vaishmpayan, and Y. Sermadevi. Quality monitoring of video over a packet network. *IEEE Trans. Multimedia*, 6(2):327–334, April 2004.

[7] D. Varodayan, A. Aaron, and B. Girod. Rate-adaptive codes for distributed source coding. *Signal Processing*, 86(11):3123–3130, 2006.

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

[9] S. Wenger. Error patterns for internet experiments. Technical report, Joint Video Team (JVT), October 1999.

[10] A. Wyner and J. Ziv. The rate-distortion function for source coding with side information at the decoder. *IEEE Transactions on Information Theory*, 22(1):1–10, 1976.

[11] T. Yamada, Y. Miyamoto, and M. Serizawa. No-reference video quality estimation based on error-concealment effectiveness. *Packet Video 2007*, pages 288–293, November 2007.