NO-REFERENCE MODELING AND ESTIMATION OF THE CHANNEL INDUCED DISTORTION AT THE DECODER FOR VIDEO CODING APPLICATIONS[1]

Matteo Naccari, Marco Tagliasacchi*, Fernando Pereira‡, Stefano Tubaro**

*Dipartimento di Elettronica e Informazione, Politecnico di Milano
({naccari,tagliasa,tubaro}@elet.polimi.it)
‡Instituto Superior Técnico - Instituto de Telecomunicações
fp@lx.it.pt

ABSTRACT

This paper proposes a technique for estimating, at the decoder side, the distortion induced by the transmission over an error-prone channel, when error-free reconstructed frames are not available as a reference. The proposed estimate considers explicitly the temporal error concealment algorithm adopted at the decoder. In the evaluation of the induced distortion, we model the effects of the absence of motion vectors and prediction residuals in the decoding process. In addition, we take into account the error propagation along successive frames. Experimental results conducted over real video sequences coded with the state-of-art standard H.264/AVC validate the proposed model. In fact, the distortion estimated when no reference is available is strongly correlated both at the frame and group of pictures level with the actual distortion. This technique represents an effective no-reference video quality monitoring tool that can be embedded in any H.264/AVC compliant decoder.

Index Terms— Channel induced distortion, video quality evaluation, error concealment, temporal distortion propagation

1. INTRODUCTION

Broadband access in IP networks makes feasible the delivery of video content for services as IP television (IPTV). In this scenario, evaluating the quality of the received videos becomes a challenging activity due to the packet losses and jitter phenomena that afflict such networks. Video quality assessment at the receiver allows end users to benefit from scalable billing contracts based on the actual received video quality. The effort of the video quality expert group (VQEG) witnesses this need [1], as demonstrated by the standardization of objective metrics for the evaluation of the received quality.

Modern video coders enable efficient compression of the video data by exploiting both spatial and temporal redundancies. In recent standards, including H.264/AVC, each frame of a video sequence is partitioned into nonoverlapping regions called macroblocks (MBs). Each macroblock can be coded exploiting either spatial redundancy (intra-frame coding) or temporal redundancy between consecutive frames (inter-frame coding). Coded data relative to macroblocks are gathered into slices, then packetized and transmitted through a noisy channel that drops packets at a given packet loss rate (PLR). At the receiver side, macroblocks belonging to lost packets cannot be decoded. Therefore, the decoder tries to recover the missing pixel values running an error concealment algorithm. Despite the concealment, the original macroblocks cannot be recovered and a channel-induced distortion is inevitably introduced. In addition, due to the predictive nature of the video encoding process, errors propagate along time affecting also correctly received macroblocks.

In the literature, the problem of estimating the channel-induced distortion at the encoder side has been addressed in [2, 3]. A statistical characterization of the channel is given, since the actual error pattern is unknown at the encoder. Conversely, the estimation of the distortion at the decoder side is simplified by the deterministic knowledge of actual loss pattern. At the same time, the lack of error-free reconstructed frames introduces the issue of no-reference distortion estimation. This fact has motivated the investigation of methodologies that try to estimate video quality degradation connected to packet losses based on the analysis of the received bitstream and networks parameters. The work in [4] provides an estimate of the received quality from the perspective of a network service provider. The proposed method parses the bitstream at different levels in order to extract information that is fed into a model, whose parameters are estimated on training data. Each level of parsing adds further computational complexity, thus the estimation accuracy is tuned on the basis of the available resources. The method proposed in [5] provides an accurate estimation based on both the video codec type, the network characteristics and the bitstream parameters (bitrate, enabled error resiliency tools, etc.). Online quality assessment requirements make the estimation of model parameters unfeasible, therefore the authors introduce a new metric (relative PSNR (rPSNR)) to accommodate the real time constraints. Finally, the recent work in [6] proposes a polynomial fitting between the measured mean square error (MSE) and the number of macroblocks in a frame for which the error concealment is ineffective.

The end-to-end distortion between the original video data and the reconstructed sequence at the decoder is due both to lossy coding (quantization) and channel losses. In [7] it has been shown that these two contributions can be considered uncorrelated. The interested reader can refer to [8, 9] for algorithms that estimate the distortion induced by lossy coding at the decoder side. Conversely, this paper complements the aforementioned works, presenting a model that estimates the channel-induced distortion introduced by packet losses in bitstreams coded with the state-of-art H.264/AVC video coding standard [10, 11]. Unlike the blind approach in [6], we explicitly model the distortion taking into account the concealment strategy adopted at the decoder side, the lack of motion vectors and prediction residuals. In addition, we consider temporal distortion propagation due to the predictive nature of modern video coders.

The rest of this paper is organized as follows. Section 2 de-

[1]THE WORK PRESENTED WAS DEVELOPED WITHIN VISNET II, A EUROPEAN NETWORK OF EXCELLENCE (HTTP://WWW.VISNET-NOE.ORG), FUNDED UNDER THE EUROPEAN COMMISSION IST FP6 PROGRAMME.

Fig. 1. *No-reference channel-induced distortion estimation.*

tails the model used to describe the effect of packet losses on the reconstructed sequence. Section 3 presents the testing scenarios and the experimental results obtained to validate the proposed model. Finally, Section 4 concludes the paper and discusses the ongoing work.

2. CHANNEL INDUCED DISTORTION ESTIMATION

This section introduces a model aimed at estimating the distortion induced by packet losses in the reconstructed video sequence at the decoder, without having access to error-free reconstructed reference frames (see Figure XX). Hereafter, the term distortion refers explicitly to the channel induced distortion only (i.e. neglecting the distortion due to quantization). The rest of this section is organized as follows: Section 2.1 introduces the notation and presents the overall model that takes into accounts the channel induced distortion for both correctly received and lost macroblocks. For this latter case, we address the estimation of the distortion due to the lack of motion vectors (Section 2.2), prediction residuals (Section 2.3) and temporal distortion propagation (Section 2.4). Finally, Section 2.5 discusses temporal attenuation of the distortion propagation due to spatial filtering.

2.1. Distortion estimation model

The loss of a packet implies that all the macroblocks that belongs to it cannot be correctly decoded. In particular, it is impossible to decode a macroblock due to the lack of coding modes (intra or inter coding, macroblock partitioning), motion vectors and prediction residuals. Errors might arise also when a packet is correctly received, due to temporal distortion propagation. Furthermore, the state-of-art video coding standard H.264/AVC exploits spatial redundancy between neighboring macroblocks performing intra prediction. Thus, the distortion not only propagates along the temporal dimension but also in the spatial one. In this paper, we do not explicitly take into account the spatial propagation of the distortion, and this is topic is left for future work.

In order to formally describe the proposed model for channel induced distortion estimation, we adopt the following notation:

- $\hat{D}_n^i$: estimated distortion introduced by the i-th macroblock in frame n.
- $\hat{D}_{n,(MV|PR|DP)}^i$: estimated distortion introduced by the i-th macroblock in frame n, due the lack of either motion vectors (MV), prediction residuals (PR) or distortion propagation (DP).
- $e(x, y, n)$: prediction residual transmitted for frame n at the pixel location (x, y).
- $\mathbf{v}^{i,k} = (v_x^{i,k}, v_y^{i,k})$: motion vector assigned to the k-th macroblock partition in the i-th macroblock ($k = 1, \ldots, K_i$, where K_i denotes the number of partitions in macroblock i).
- $\tilde{\mathbf{v}}^i = (\tilde{v}_x^i, \tilde{v}_y^i)$: motion vector used by the concealment algorithm for the i-th macroblock.
- $\bar{\mathbf{v}}^i = (\bar{v}_x^i, \bar{v}_y^i)$: motion vector of the i-th macroblock obtained by averaging the motion vectors of the K_i partitions.

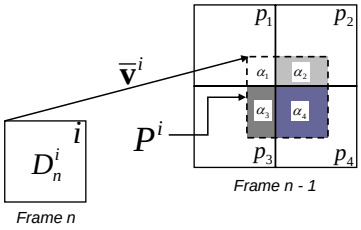

Fig. 2. *Temporal distortion propagation when the predictor P^i overlaps with four macroblocks in the reference frame $(n-1)$*

In the following, we will always use the mean square error as distortion metric. Therefore, the distortion affecting frame X_n due to packet losses is given by:

$$MSE(\hat{X}, \tilde{X}) = \sum_{x,y} (\hat{X}_n(x,y) - \tilde{X}_n(x,y))^2 = \frac{1}{M} \sum_{i=1}^{M} D_n^i \quad (1)$$

where M denotes the number of macroblocks in a frame and $\hat{X}_n$, $\tilde{X}_n$ are, respectively, the reconstructed frame at the encoder and decoder side.

In order to estimate the D_n^i terms in (1), the proposed model considers two separate cases, depending on whether a macroblock i has been lost or correctly received:

1) ***Lost MB***: Lost data consist of coding modes, motion vectors and prediction residuals. Since the coding mode is unknown, we apply temporal error concealment to all macroblocks. The temporal concealment implemented in the H.264/AVC decoder reference software [12] replaces the missing pixels with the ones pointed by the concealed motion vector $\tilde{\mathbf{v}}^i$. This motion vector is chosen among the motion vectors of the 8×8 blocks surrounding the lost macroblock i, minimizing the sum of absolute differences (SAD) along the boundaries of i. We assume that the contributions to the channel distortion due to the lack of motion vectors, prediction residuals and temporal propagation as uncorrelated. Thus we can write:

$$\hat{D}_n^i\{lost\} = \hat{D}_{n,MV}^i + \hat{D}_{n,PR}^i + \hat{D}_{n,DP}^i \quad (2)$$

The estimation of the terms on the right hand side of equation (2) will be detailed below.

2) ***Correctly received MB***: Channel-induced distortion propagates from previous frames due to the motion-compensation loop performed at the decoder. In particular the predictor $\mathbf{P}^i$ of the i-th macroblock can overlap with up to four macroblocks in the reference frame (see Figure 2). The H.264/AVC video coding standard enables the assignment of different motion vectors to each macroblock partition. For the sake of simplicity, here we consider the predictor $\mathbf{P}^i$ as referenced by the average motion vector $\bar{\mathbf{v}}^i$.

Let N_0 ($1 \leq N_0 \leq 4$) be the number of blocks in the reference frame overlapping with the predictor $\mathbf{P}^i$ and γ_p the number of pixels in $\mathbf{P}^i$ that overlap with the p-th macroblock ($1 \leq p \leq N_0$). The temporal distortion propagation is modeled as follows:

$$hatD_n^i\{received\} = \sum_{p=1}^{N_0} \alpha_p \cdot \hat{D}_{n-1}^p, \quad \text{with} \quad \alpha_p = \frac{\gamma_p}{256} \quad (3)$$

2.2. Motion Vector Distortion

In case of translational motion, the difference between the predictor provided by the concealment algorithm $\tilde{\mathbf{P}}^i$ and the one correspond-

ing to the motion vector estimated at the encoder, $\mathbf{P}^i$, consists into a spatial shift related to the difference between $\tilde{\mathbf{v}}^i$ and $\bar{\mathbf{v}}^i$ [13]:

$$\tilde{\mathbf{P}}^i(x,y) = \mathbf{P}^i(x-\delta_x, y-\delta_y) \tag{4}$$

where $\delta_x = \bar{v}_x^i - \tilde{v}_x^i$ and $\delta_y = \bar{v}_y^i - \tilde{v}_y^i$.

In the frequency domain, equation (4) corresponds to a phase rotation: $\tilde{\mathbf{P}}^i(\vec{\omega}) = \mathbf{P}^i(\vec{\omega}) \cdot e^{-j\vec{\omega}\vec{\delta}}$, with $\vec{\omega} = (\omega_x, \omega_y)$ and $\vec{\delta} = (\delta_x, \delta_y)$. From the Parseval's theorem, the distortion due to the lack of motion vectors, $D_{n,MV}^i$, is given by:

$$\hat{D}_{n,MV}^i = \frac{1}{(2\pi)^2} \cdot \int_{\vec{\omega}} \phi(\vec{\omega}) \left| 1 - e^{-j\vec{\omega}\vec{\delta}} \right|^2 d\vec{\omega} \tag{5}$$

where the term $\phi(\vec{\omega})$ denotes the power spectral density of $\mathbf{P}^i(\vec{\omega})$. From equation (5) we notice that the estimation of $D_{n,MV}^i$ involves two unknown quantities: $\phi(\vec{\omega})$ and $\vec{\delta}$. The first one is obtained computing the periodogram spectral estimate of $\tilde{\mathbf{P}}^i$. The second one implies the knowledge of the original average motion vector $\bar{\mathbf{v}}^i$. A reasonable estimate $\hat{\mathbf{v}}^i$ of $\bar{\mathbf{v}}^i$ is given as the average of all the motion vectors belonging to the 8×8 blocks surrounding the macroblock i. Therefore, substituting $\bar{\mathbf{v}}^i$ with $\hat{\mathbf{v}}^i$ in $\vec{\delta}$ the distortion in Equation (5) is readily computed.

2.3. Prediction Residuals Distortion

The distortion due to the absence of prediction residuals is estimated as follows:

$$\hat{D}_{n,PR}^i = \frac{1}{256} \sum_{x=1}^{16} \sum_{y=1}^{16} \hat{e}(x,y,n)^2 \tag{6}$$

Since the packet which the macroblock i belongs to has been dropped, the prediction residuals $e(x,y,n)$ are not available. In order to provide an estimate $\hat{e}(x,y,n)$, it should be noted that the energy of the prediction residuals is higher in zones where occlusions occur. Furthermore temporal correlation exists between occlusions in neighboring frames. Therefore, it is likely that the residuals in the region pointed by the concealed motion vector constitute a good estimate:

$$\hat{e}(x,y,n) = e(x+\tilde{v}_x, y+\tilde{v}_y, n-1) \tag{7}$$

2.4. Distortion Propagation

The quantity $\hat{D}_{n,DP}^i$ accounts for the temporal distortion that propagates from previous frames due to motion-compensated temporal concealment. Thus, $\hat{D}_{n,DP}^i$ is modeled as in equation (3) using $\tilde{\mathbf{v}}^i$ instead of $\bar{\mathbf{v}}^i$.

2.5. Deblocking Filter and Quarter Pixel Motion Interpolation Issues

The work in [3] provides a thorough analysis about the effect of error propagation due to motion compensation and the temporal decay of the distortion due to the spatial filtering. Spatial filtering performed in the H.264/AVC decoder is due to two reasons: the deblocking filter and the quarter pixel motion interpolation. These contributions have been modeled by two attenuation coefficients θ_1 and θ_2 ($0 \leq \theta_i \leq 1$) whose values have been derived experimentally. Therefore, the overall estimate introduced in Section 2.1 is modified as follows:

$$\hat{D}_n^i\{received\} = \theta_1 \cdot \left(\sum_{l=1}^{N_O} \alpha_l \cdot \hat{D}_{n-1}^l \right) \tag{8}$$

Sequence	Format	Frame rate [Hz]	GOP length	Bitrate [kbps]	# of frames
Foreman	QCIF	10	10	64	100
Coastguard	QCIF	10	10	64	100
Soccer	CIF	30	15	1200	300

Table 1. *Test video sequences.*

$$\hat{D}_n^i\{lost\} = \theta_2 \cdot \left(\hat{D}_{n,DP}^i \right) + \hat{D}_{n,MV}^i + \hat{D}_{n,PR}^i \tag{9}$$

3. MODEL VALIDATION

To validate the channel distortion estimate proposed in this paper, experimental simulations have been conducted over real video sequences.

3.1. Test Scenarios and Coding Conditions

We consider two target application scenarios: conversational services and video streaming over the internet. As for the conversational scenario, the *Foreman* and *Coastguard* video sequences are encoded with the parameters listed in the first two rows of Table 1. As for the streaming scenario, the *Soccer* video sequence has been selected with parameters indicated in the last row of Table 1. All the video sequences have been coded with the H.264/AVC standard using the reference software version JM12.3 [14], baseline profile.

3.2. Channel Simulation Parameters

In the target video coding applications mentioned above the transmission takes place over an IP network. Bitrates and packet sizes have been chosen according to the guidelines given in [15]. In both scenarios, packetization follows the real-time transfer protocol (RTP), with each packet corresponding to a coded slice. The packet size is, respectively, 64 bytes for the conversational scenario and 256 bytes for the video streaming one. We have set the target bitrates equal to 64 kbps for the conversational scenario and 1200 kbps for video streaming. In order to simulate IP networks, packet loss rates (PLR) of 3,5,10 and 20% have been tested, where the loss patterns have been generated according to [16].

3.3. Experimental Results

In order to validate the accuracy of the proposed model, the correlation coefficient between the actual distortion and the one provided by the model has been measured over 30 channel simulations, with $\theta_1 = 0.1$ and $\theta_2 = 0.2$. The correlation coefficient has been computed at different levels of granularity: macroblock, frame and group of pictures (GOP) level. As an example, Figure 3 shows the relationship between the estimated and the actual distortion at the frame level for the *Foreman* sequence, corresponding to one channel realization at PLR equal to 3%. Figure 4 depicts the temporal tracks of the distortion. We notice the effect of distortion propagation, which is stopped when I frames are decoded (in our simulations I frames are assumed to be error-free). Experimental results for all of the tested sequences are listed in Table 2. We notice relatively high values of the correlation coefficient (> 0.7) both at the frame and at the GOP level, which proves that the proposed channel-induced distortion estimation algorithm can be effectively used in applications for real-time monitoring of the video quality at the decoder side. Lower values of the correlation coefficient are obtained at the macroblock level. This behavior can be justified by the fact that some of

the model assumptions do not hold exactly for real video sequences: 1) the uncorrelation of the $\hat{D}^i_{n,MV}$, $\hat{D}^i_{n,PR}$ and $\hat{D}^i_{n,DP}$ terms in equation (2); 2) the translation motion hypothesis in equation (4); 3) the original average motion vector estimate ($\hat{\mathbf{v}}^i$) might fail for those macroblocks with poor motion correlation with their neighbors.

We want also to emphasize that the bulk of the good performance both at frame and GOP level comes from the explicit modeling of the distortion propagation expressed by equation (3). As a matter of fact, the second column relative to each sequence in Table 2, shows the correlation coefficients where the distortion has been estimated neglecting the temporal propagation ($\theta_1 = 0$ and $D^i_{n,DP} \equiv 0$ in equation (2)). We notice that in this case the correlation coefficient is significantly smaller.

4. CONCLUSIONS

This paper introduces a novel algorithm aimed at estimating the channel-induced distortion at the decoder side. The proposed model addresses explicitly both the effect of temporal error concealment and distortion propagation. The accuracy of the proposed model enables to accurately monitor the video quality at the receiver side at the frame and GOP granularity. Ongoing work is moving along the extension of the proposed model to include intra-coded slices, considering spatial distortion propagation due to intra prediction. This would enable optimal spatial/temporal error concealment selection at the decoder side.

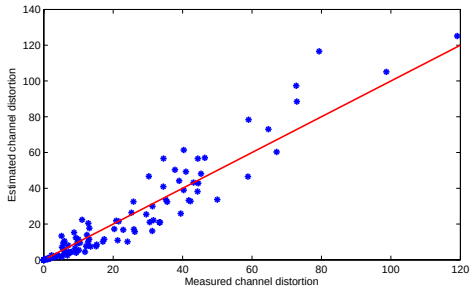

Fig. 3. Correlation between the measured channel distortion and the estimated one at the frame level for the *Foreman* video sequence with PLR = 3%

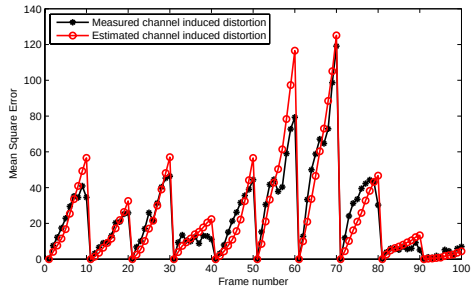

Fig. 4. Temporal evolution track of the channel distortion estimate for the *Foreman* video sequence with PLR = 3%

5. REFERENCES

[1] ITU-T, *Recommendation J.143: User requirements for objective perceptual video quality measurements in digital cable television*, 2000.

Sequence		Foreman		Coastguard		Soccer	
PLR [%]	Level	DP	No DP	DP	No DP	DP	No DP
3	MB	0.66	0.43	0.56	0.4	0.58	0.42
	Frame	0.95	0.66	0.93	0.77	0.72	0.65
	GOP	0.97	0.92	0.93	0.82	0.89	0.73
5	MB	0.62	0.38	0.47	0.38	0.59	0.41
	Frame	0.93	0.62	0.94	0.65	0.72	0.66
	GOP	0.99	0.89	0.93	0.8	0.80	0.71
10	MB	0.5	0.33	0.4	0.34	0.54	0.36
	Frame	0.89	0.53	0.86	0.62	0.73	0.59
	GOP	0.98	0.81	0.9	0.64	0.79	0.69
20	MB	0.4	0.24	0.39	0.35	0.43	0.32
	Frame	0.82	0.43	0.7	0.4	0.71	0.58
	GOP	0.95	0.59	0.81	0.54	0.74	0.65

Table 2. *Correlation coefficient between the measured channel distortion and the estimated one for each video sequence.*

[2] R. Zhang, S. L. Regunathan, and K. Rose, "Video coding with optimal inter/intra-mode switching for packet loss resilience," vol. 18, no. 6, pp. 966–976, June 2000.

[3] N. Färber, K. Stuhlmuller, and B. Girod, "Analysis of error propagation in hybrid video coding with application to error resilience," in *ICIP*, 1999.

[4] A. R. Reibman and V. Vaishampayan, "Low complexity quality monitoring of MPEG-2 video in a network," in *ICIP*, 2003.

[5] S. T., J. Apostolopoulos, and R. Guérin, "Real-time monitoring of video quality in IP networks," in *International Workshop on Network and Operating System Support for Digital Audio and Video*, 2005.

[6] T. Yamada, Y. Miyamoto, and M. Serizawa, "No-reference video quality estimation based on error-concealment effectiveness," in *Packet Video*, 2007.

[7] Z. He, J. Cai, and C. W. Chen, "Joint source channel rate-distortion analysis for adaptive mode selection and rate control in wireless video coding," *IEEE Transactions on Circuit and Systems for Video Technology*, vol. 12, no. 6, pp. 511–523, June 2002.

[8] T. Brandao and M. P. Queluz, "Blind PSNR estimation of video sequences using quantized DCT coefficient data," in *PCS*, 2007.

[9] A. Ichigaya, M. Kurozumi, N. Hara, Y. Nishida, and E. Nakasu, "A method of estimating coding PSNR using quantized DCT coefficients," *IEEE Transactions on Circuit and Systems for Video Technology*, vol. 16, no. 2, pp. 251–259, 2006.

[10] ITU-T, *Information Technology - Coding of audio-visual objects - Part 10: advanced video coding*, 2003, Final Draft International Standard, ISO-IEC FDIS 14 496-10.

[11] T. Wiegand, G. J. Sullivan, G. Bjøntegaard, and A. Luthra, "Overview of the H.264/AVC Video Coding Standard," *IEEE Transactions on Circuit and Systems for Video Technology*, vol. 13, pp. 560–576, 2003.

[12] G. J. Sullivan, T. Wiegand, and K.-P. Lim, "Joint Model Reference Encoding Methods and Decoding Concealment Methods," Tech. Rep. JVT-I049, Joint Video Team (JVT), 2003.

[13] A. Secker and D. Taubman, "Highly Scalable Video Compression With Scalable Motion Coding," *IEEE Transactions on Image Processing*, vol. 13, no. 8, pp. 1029–1041, 2004.

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

[15] S. Wenger, "H.264/AVC Over IP," *IEEE Transactions on Circuit and Systems for Video Technology*, vol. 13, pp. 645–656, 2003.

[16] S. Wenger, "Error patterns for Internet experiments," Tech. Rep. JVT-Q15-I-16r1, Joint Video Team (JVT), 1999.