ANALYSIS-BY-SYNTHESIS DISTORTION COMPUTATION FOR RATE-DISTORTION OPTIMIZED MULTIMEDIA STREAMING

*Enrico Masala, Juan Carlos De Martin**

Dipartimento di Automatica e Informatica / *IEIIT-CNR
Politecnico di Torino
Corso Duca degli Abruzzi, 24 — I-10129 Torino, Italy
E-mail: [masala|demartin]@polito.it

ABSTRACT

This paper presents an analysis-by-synthesis technique to evaluate the perceptual importance of multimedia packets for rate-distortion optimized streaming. The proposed technique, instead of relying on a priori information, computes the distortion that would be caused by the loss of each single packet, including the effects of error propagation and receiver-side error concealment. A rate-distortion optimized streaming algorithm is presented to compare the perceptual performance obtained using content-adaptive analysis-by-synthesis distortion values versus distortion values obtained using a priori knowledge of the statistical importance of the elements of the compressed multimedia bitstream. Simulations with video test sequences compressed with the MPEG-2 coding standard show that the proposed technique delivers substantial and consistent PSNR gains (1.2–2.8 dB) with respect to ideal frame-type-driven a priori distortion evaluation for a wide range of channel conditions. Compared to distortion-agnostic streaming techniques such as SoftARQ, the gain is even more pronounced.

1. INTRODUCTION

Multimedia applications are one of the most appealing services in the communications industry today, with growth forecast to be very strong for most wireline and wireless scenarios. Multimedia traffic is thus expected to quickly become the dominant kind of network traffic, raising the interest on more and more advanced multimedia communications techniques.

Several challenges, however, have to be met, due to the strict constraints on bandwidth and delay variations associated to good audio-visual quality. Many methods have been proposed so far to enhance the quality of real-time multimedia communications in general, and for streaming in particular. For a survey of state-of-the-art video streaming technologies, see, e.g., [1].

Given the non-uniform perceptual importance of multimedia signals, several techniques have investigated the application of different levels of protection to the various parts of the compressed bitstream. Most Unequal Error Protection techniques rely on *a priori* classification of the importance of the data to be protected, typically based on statistical studies of the perceptual importance of the elements of the compressed bitstream. For motion-compensated video, one common approach is to assign different levels of importance to packets based on the type of frame (I, P or B) to which they belong [2], or, in case of scalable video coding, to the various layers of the compressed bitstream.

The instantaneous characteristics of the video signal, however, may be exploited at a finer level of granularity. A measure of perceptual importance can, in fact, be determined for each individual packet. Based on this information, the transmitter may decide, in conjunction with other information such as the current state of network, which is the best transmission policy at any given time instant.

A rate-distortion (R-D) optimized approach has recently been proposed to enhance the performance of multimedia transmissions. The technique presented in [3] optimizes the transmission policy taking into account for each packet, among the other factors, its impact on quality in case of loss. In a number of studies the distortion values associated to each data unit are assumed to be known. The determination of such values, however, is not always a trivial task. Moreover, the accuracy level of the distortion values may have a significant effect on the overall performance of multimedia communications systems.

We present a new *analysis-by-synthesis* technique to compute the distortion associated to each packet, based on an accurate simulation of the decoding stage, including concealment. This method has already been successfully used to improve multimedia transmission over DiffServ IP networks [4] [5]. Distortion values can be also used for transmissions protection using forward error correction schemes, as shown in [6]. Hence the importance of an accurate algorithm to compute the distortion associated to each transmission unit.

To document the performance gains delivered by the analysis-by-synthesis distortion computation technique, detailed comparisons with traditional *a priori* distortion estimation methods are presented using the same streaming optimization algorithm and the same transmission channel.

This paper is organized as follows. Section 2 examines the analysis-by-synthesis technique. Section 3 describes the rate-distortion optimized selection algorithm used for the simulations. Detailed comparisons with other techniques are reported in Section 4. Finally, conclusions are presented in Section 5.

2. ANALYSIS-BY-SYNTHESIS DISTORTION COMPUTATION

The analysis-by-synthesis distortion computation method consists in analyzing the compressed video with a packet-level granularity. Figure 1 shows the block diagram of the proposed technique.

This work was supported in part by CERCOM, the Center for Wireless Multimedia Communications, Torino, Italy, http://www.cercom.polito.it.

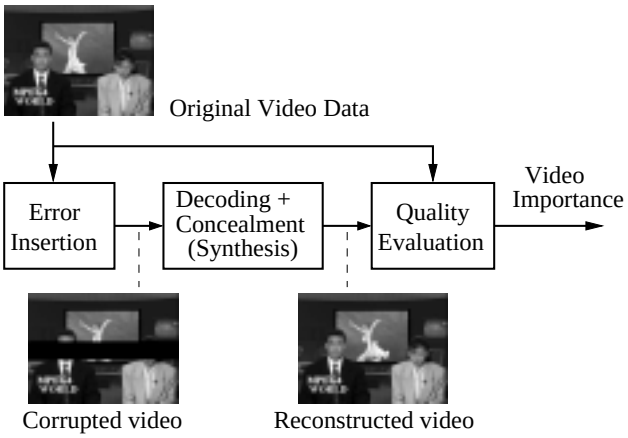

Fig. 1. Block diagram of the analysis-by-synthesis distortion computation technique.

The algorithm is activated for each new packet. The original compressed bitstream is corrupted by simulating its loss. Then, the complete decoder behavior, including concealment, is simulated (synthesis stage). The final step is the distortion computation of the corrupted video with respect to the non-corrupted one. The technique can be also used to compute the distortion at the macroblock level, as in [4], to make the encoding process adaptive, further enhancing the overall performance. Full simulation of the decoder behavior takes into account the effects of the concealment technique, often neglected for simplicity reasons in many studies. Moreover, the distortion computation is precise because it exactly reproduces the operations of the decoder, taking into account both spatial and temporal error propagation. The simulation of the decoder also makes the method independent of any particular video coding algorithm.

Due to the inter-dependencies usually present between data units, the simulation of the loss of an isolated data unit is not completely realistic, particularly for high packet loss rates. Every possible combination of events should ideally be considered, weighted by its probability, and its distortion computed by the analysis-by-synthesis technique, obtaining the expected distortion value. For simplicity, we assume all preceding data units have been correctly received and decoded. In [6] a similar approach is considered, noting that the assumption is equivalent to compute the first order approximation of the expected distortion, interpreted as a function of the loss events of each packet.

The distortion values generated by the analysis-by-synthesis technique can be often obtained as a by-product of the encoding process. Several coding methods, in fact, already reconstructs the data simulating the decoder operations, since it is needed for motion-compensated encoding. If no predictive coding is used, the distortion estimation simply consists in computing a quality measure (e.g. the MSE) between the pixels represented by the current packet and those, typically belonging to the preceding frame, that would be used for concealment in case of loss. With predictive coding, the distortion caused in the subsequent frames must be evaluated until it becomes negligible, i.e., at the beginning of the next Group of Pictures (GOP) for MPEG video. In the latter case, the complexity of the proposed approach could be high, but it is suitable for stored-video scenarios that allow an heavy precomputation.

Figure 2 shows the distortion values computed by the proposed

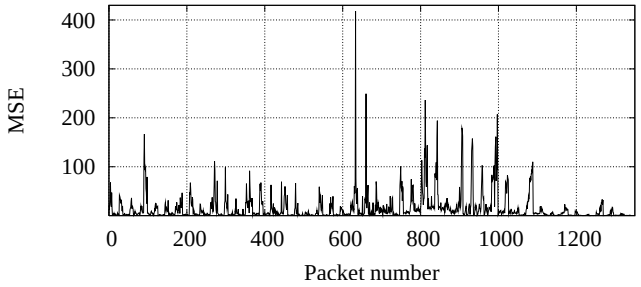

Fig. 2. Analysis-by-synthesis-evaluated distortion as a function of packet number; *Foreman* sequence, 128 kbit/s, 15 fps.

analysis-by-synthesis technique for an example test sequence, *Foreman*, as a function of packet number. Distortion varies widely and quite irregularly, since it depends on the actual video content and on how well the decoder can replace it at any given time. The distortion values associated to the first I-type frame of the sequence are not reported in Figure 2, due to their magnitude; in this case, in fact, the concealment technique assigns a fixed value to the missing pixels, resulting in very high distortion. An MPEG-2 [7] encoder producing a slice per row of macroblocks was used. Each slice was placed into an IP packet, one slice per packet. Nine slices per frame are present (QCIF size), and a 10-frames per GOP coding scheme (IPBBPBBPBB in display order) with closed GOP is used. A simple temporal concealment technique that replaces the missing areas with the pixels in the same position in the previous available frame was implemented. Table 1 shows the mean distortion value for each type of frame for the *Foreman* and *News* sequences. The number next to the frame type is its position in the GOP. The first P-type frames (P1 in the table) are associated to the highest distortion values. The reason lies in the distance of the frame that is used for concealment. While I- and B-type frames, in fact, refer to the frame immediately preceding them, P-type frames reference three-frames away in the past, resulting in generally poor concealment. The distortion due to B-type frames is low, since there is no error propagation.

3. R-D OPTIMIZED STREAMING ALGORITHM

In streaming scenarios, a certain delay occurs between the time the packet arrives at the receiver and the time the packet is played. One or more retransmissions of the generic packet l are, therefore, possible, until a last opportunity, called deadline time t_{D_l}, which

Table 1. Average MSE distortion for each frame type in a GOP; test sequences *Foreman* and *News*, QCIF, 128kbit/s, 15fps.

Frame type	*Foreman* (MSE)	*News* (MSE)
I	21.36	11.59
P1	55.96	12.82
B1	4.13	1.32
B2	4.52	0.70
P2	39.49	14.44
B3	4.49	0.65
B4	3.77	2.45
P3	17.46	3.65
B5	4.61	0.69
B6	4.35	1.13

depends on the playback time of the packet. Given the distortion information associated to each packet, it is possible to use it to optimize the retransmission policy in a rate-distortion sense. We refer to the framework presented in [3], with some modifications. The optimization process takes into account the packet size B_l, the deadline time t_{D_l} and the distortion value d_l, obtained using the analysis-by-synthesis technique. The distortion value d_l is the MSE computed with respect to the correctly decoded pixels, and it includes the effect of the distortion caused by the loss in the frame to which the packet belongs, as well as the distortion caused by the erroneous reconstruction of subsequent frames up to the beginning of the next GOP. The areas of the frame that are not interested by errors give a contribution to the MSE equal to zero, since it is computed with respect to the correctly decoded pixels and not to the original ones.

For each packet l, given a transmission policy π_l, it is possible to estimate the expected loss probability $\epsilon(\pi_l)$ and expected cost $\rho(\pi_l)$ in terms of bytes transmitted per byte of original data. $\epsilon(\pi_l)$ is null if the packet has been acknowledged, otherwise it can be computed constructing the trellis of the future events, evaluating all the branches with their associated probability, as in [3]. This method requires the ability to compute the probability that a packet is subject to a given delay. In our experiments we assumed an exponential distribution of the form

$$p_F(t) = (1 - p_{loss,F}) \cdot a_F e^{-a_F(t - FTT_{min})}, \quad (1)$$

where $p_{loss,F}$ is the packet loss probability of the forward channel and FTT_{min} is the minimum time a packet needs to go from sender to receiver. Analogous considerations hold for the backward channel. The packet loss events are modeled as infinite delay events. We also assume that the receiver sends an acknowledgement as soon as a packet arrives. The receiver starts to playback the received video packets $\Delta T + W$ seconds after the first packet has been sent; we assume $\Delta T = \overline{FTT}$ and W, called transmission window, to be equal to the size, expressed in seconds, of the playout buffer at the receiver.

Given a Lagrangian multiplier λ', it is possible to find the optimal transmission policy π_l^* of a packet as

$$\pi_l^* = \arg\min_{\pi_l} \epsilon(\pi_l) + \lambda' \rho(\pi_l). \quad (2)$$

At each time instant t, L packets form the transmission buffer T_B. A packet belongs to T_B if:

1. the sender has not already received an acknowledgement for that packet;
2. $t_{D_l} > t + FTT_{min}$;
3. $t_{D_l} < t + \Delta T + W$.

The last expression limits the number of packets in T_B. Moreover, the transmission window W can be used to influence the tradeoff between the complexity of the packet selection algorithm, that depends on the number of considered packets, and the performances of the proposed method.

The transmission policy used by the sender at a given time t can be expressed as

$$\pi = (\pi_1, ..., \pi_l, ..., \pi_L), l \in T_B. \quad (3)$$

Let the expected rate be defined as

$$R(\pi) = \sum_{l \in T_B} B_l \rho_l. \quad (4)$$

A good estimation of the distortion at the decoder can be expressed by

$$D(\pi) = \sum_{l \in T_B} d_l \epsilon_l. \quad (5)$$

As explained above, d_l is computed with respect to the currently decoded pixel, thus it considers only the distortion due to the loss and not the source coding distortion. Assuming that the distorted areas of the frames do not overlap in case of multiple losses, it is possible to obtain the total distortion as the sum of the distortion contributions of the various packets. The results will confirm that such non-overlapping assumption is good and the sum does lead to a useful estimation of the total distortion.

The policy π^* that minimizes

$$J(\pi) = D(\pi) + \lambda R(\pi) \quad (6)$$

will be used by the transmitter to decide which packets to send at a given time t. The minimization of (6) requires to find the optimal policy for each packet in T_B. To solve the problem, note that (6) can be expressed as

$$J(\pi) = \sum_{l \in T_B} d_l \epsilon_l + \lambda \sum_{l \in T_B} B_l \rho_l = \sum_{l \in T_B} (d_l \epsilon_l + \lambda B_l \rho_l). \quad (7)$$

The problem of finding the optimal policy π^* can be solved separately for each packet, defining $\lambda_l' = \lambda B_l / d_l$ and finding the π_l^* given by (2). The policy selection process is repeated every T seconds, after having discarded from T_B all packets that have been acknowledged in the meantime. To achieve constant-bit-rate transmission, the λ parameter is adjusted, at each transmission opportunity, to meet as close as possible a given rate constraint.

4. RESULTS

Although the analysis-by-synthesis distortion computation technique has been applied to constant bit-rate video sequences coded using the MPEG-2 coding standard [7], the method can be straightforwardly adapted to any video codec. Every T seconds the packet scheduler evaluates which packets are to be sent. The standard RTP/UPD/IP protocol stack is used; packets are reordered at the receiver using the RTP sequence number. Table 2 shows the parameters used in all simulations, unless otherwise noted. The retransmission bandwidth is the amount of bandwidth allocated for packet retransmission and is expressed as a given percentage of the source coding rate.

The results compare the performance of the rate-distortion packet selection algorithm using —all other aspects equal— two different sets of distortions: distortion values computed using the *analysis-by-synthesis* algorithm, and distortion values computed using an ideal implementation of the classic *a priori* estimation approach. For further comparison, the performance of a well-known scheduling algorithm not based on distortion, *SoftARQ* [8], is also reported. SoftARQ sends the packets in bitstream order, associating a timeout to each of them. At every retransmission opportunity, if an acknowledgement has not yet been received for a packet whose timeout has expired, but whose deadline has not, it schedules that packet for retransmission. The packets to be retransmitted are ordered and sent according to a priority given by their deadline.

The *a priori* distortion evaluation approach consists in giving higher priority to those compressed-bitstream elements that are considered to be the most perceptually important on a statistical basis. The effects of errors on each bitstream element over a

Table 2. Parameters used for all simulations, unless otherwise noted.

Parameter	Value
Test sequence size	QCIF (176 × 144)
Coding rate	128 kbit/s, 15 fps
T	50 ms
Transmission window, W	500 ms
a_F, a_B	20
$p_{loss,F}, p_{loss,B}$	0.20
FTT_{min}, BTT_{min}	50 ms
Playout buffer	500 ms
Retransmission bandwidth	20%
Timeout (SoftARQ only)	149 ms
# of channel realizations	50

large database of representative material are observed, resulting in a sensitivity ranking. For motion-compensated video, a typical a priori technique consists in giving the highest priority to I-frames, a lower priority to P-frames and no priority to B-frames. In the case of our simulations, the a priori scheme is based on the distortion values of Table 1, i.e., each frame type in a GOP receives a level of priority proportional to the actual average distortion that its loss would cause. This a priori scheme is, therefore, ideal, an upper bound for all possible a priori methods, since the values, instead of being averages based on statistics of the expected plurality of input video sequences, are the values, computed by the analysis-by-synthesis technique, for the specific sequences used in the simulations.

Table 3 shows the PSNR values when the analysis-by-synthesis, the a priori and the softARQ methods are applied to the well-known *Foreman*, *News* and *Mobile* sequences. The PSNR is computed with respect to the original uncompressed material. The analysis-by-synthesis method shows significant performance gains, ranging from 1.3 dB to 3.1 dB depending on the sequence, with respect to the other methods. In particular, the ideal a priori technique is outperformed by 2.8 dB (News), 1.73 dB (Foreman) and 0.77 dB (Mobile). The a priori method delivers indeed only slightly better performance with respect to the non-distortion-based SoftARQ method. Adapting the transmission policy to the characteristics of the input video signal thus delivers significant gains with respect to the statistically–based *a priori* approach. The highest gains are obtained for the *News* sequence: in that case, the analysis-by-synthesis algorithm identifies the low impact of the packets representing the static background image, and sometimes does not even transmit them, thus saving bandwidth for other, more critical parts of the sequence.

Figure 3 shows the PSNR performance of the three techniques as a function of packet loss rate. Analysis-by-synthesis distortion computation consistently outperforms the other methods over a

Table 3. PSNR performances of three different streaming techniques; *News* & *Foreman* test sequences: QCIF, 128 kbit/s; *Mobile*: CIF, 600 kbit/s; 15 fps.

Sequence	SoftARQ	RD opt. A priori	RD opt. Analysis-by-Synthesis
News	29.51 dB	29.85 dB	32.65 dB
Foreman	25.10 dB	25.76 dB	27.49 dB
Mobile	20.91 dB	21.48 dB	22.25 dB

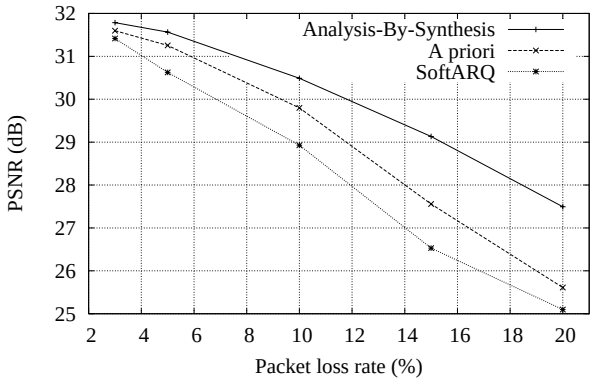

Fig. 3. PSNR as a function of packet loss rate; *Foreman* sequence.

wide range of channel conditions.

5. CONCLUSIONS

An analysis-by-synthesis technique to evaluate the importance of individual video packets has been presented. The analysis-by-synthesis method computes the importance as the distortion due to the loss of the video packet, also taking into account the distortion in subsequent frames due to the predictive coding. The technique is independent of video coding algorithm. The values produced by the analysis-by-synthesis method have been used as the input of a rate-distortion optimized packet scheduler in a streaming scenario. Simulations showed that the proposed technique consistently outperforms classical a priori distortion evaluation, leading to higher PSNR values over a wide range of channel conditions.

6. REFERENCES

[1] "Issue on streaming technologies," *IEEE Transactions on Multimedia*, vol. 3, no. 1, March 2001.

[2] S. H. Kang and A. Zakhor, "Packet Scheduling Algorithm for Wireless Video Streaming," in *Proc. Packet Video Workshop*, Apr. 2002.

[3] P.A. Chou and Z. Miao, "Rate-Distortion Optimized Streaming of Packetized Media," *submitted to IEEE Trans. on Multimedia*, Feb. 2001.

[4] E. Masala, D. Quaglia, and J.C. De Martin, "Adaptive Picture Slicing for Distortion-Based Classification of Video Packets," in *Proc. IEEE Workshop on Multimedia Signal Processing*, Oct. 2001, pp. 111–116.

[5] F. De Vito, L. Farinetti, and J.C. De Martin, "Perceptual Classification of MPEG Video for Differentiated-Services Communications," in *Proc. IEEE Int. Conf. on Multimedia & Expo*, Aug. 2002, vol. 1, pp. 141–144.

[6] R. Zhang, S. L. Regunathan, and K. Rose, "Optimized Video Streaming over Lossy Networks with Real-Time Estimation of End-to-End Distortion," in *Proc. IEEE Int. Conf. on Multimedia & Expo*, Aug. 2002, vol. 1, pp. 861–864.

[7] ISO/IEC, "MPEG-2 generic coding of moving pictures and associated audio information," *ISO/IEC 13818*, 1996.

[8] M. Podolsky, S. McCanne, and M. Vetterli, "Soft ARQ for Layered Streaming Media," in *Tech. Rep. UCB/CSD-98-1024, University of California, Computer Science Division, Berkeley*, November 1998.