Model–based Distortion Estimation For Perceptual Classification of Video Packets

Fabio De Vito*, Davide Quaglia* and 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: [fabio.devito|davide.quaglia|demartin]@polito.it

Abstract— **In video communications over IP networks, quality of service (QoS) guarantees must be introduced to limit the effect of packet losses. In particular, end–to–end QoS can be improved if packets are protected according to the distortion that would be introduced at the receiver by their loss. In the traditional Analysis–by–Synthesis (AbS) approach, each packet is assumed lost, error concealment applied, the sequence decoded and the resulting overall distortion computed. This process produces reliable distortion estimates, but is computationally demanding. In this work we present a hybrid approach: the distortion introduced in the current frame is evaluated with the AbS method, while the distortion in future frames is estimated by means of a statistical error–propagation model. Results obtained on eight, widely different H.264 sequences show that the proposed model successfully estimates overall distortion with very low complexity. Network simulations also show that model–based packet classification, when used for video transmission over DiffServ networks, delivers PSNR results which are consistently within 0.1 dB compared to the AbS technique.**

I. Introduction

VIDEO communications over IP networks, both wireline and wireless, are at the focus of an extraordinary deal of attention. However, for this appealing class of applications to succeed, quality of service (QoS) guarantees must be introduced to limit the effect of packet losses due to congestion, bit errors, and late delivery. Several approaches have been proposed for the robust transmission of video packets; they range from source coding solutions (e.g., resynch marker and intra/inter coding mode selection) to the addition of redundancy (e.g., forward error correction and retransmissions).

One promising QoS solution is the Differentiated Services (DiffServ) architecture [1] according to which routers apply different forwarding behaviors to packets depending on the value of the type–of–service field in the IP header. Using a DiffServ network architecture, video packets are assigned to classes, thus obtaining a different treatment by the network; in particular, packets assigned to the most privileged class will be lost with a very small probability, while packets belonging to the lowest priority class will experience the traditional best–effort service.

For multimedia transmission over DiffServ networks, a key problem is how to optimally assign speech, audio, and video packets to the available DiffServ classes. One way to perform the assignment is to proceed on a packet-by-packet basis, to exploit the highly non-uniform perceptual importance of compressed multimedia data. Working on the perceptual importance of each individual data unit has been shown, in recent years, to deliver better performance than relying on the average error sensitivity of each bitstream element, as done in traditional data partitioning (see, e.g., [2]).

The perceptual importance of a video packet can be expressed as the distortion that would be introduced at the receiver by its loss, taking into account the effects of both error concealment and error propagation due to temporal prediction. Such distortion can be computed following an Analysis–by–Synthesis (AbS) approach [3] [4]: each packet is assumed lost, error concealment applied, the sequence decoded and the resulting overall distortion computed. This process produces reliable distortion estimates, but, for motion–compensated video, is computationally demanding.

Other, less CPU intensive, methods for the estimation of the impact of packet loss on distortion have been proposed. The Recursive Optimal Per-pixel Estimate (ROPE) [5] [6] recursively computes the pixel–level distortion of decoder frame reconstruction due to quantization, error propagation, and error concealment; this approach requires, however, a considerable amount of memory and its complexity is increased by the presence of B–frames. In the First–Order Distortion Estimate (FODE) [7] all possible loss patterns in the GOP are considered and the estimation formula is approximated through a first–order Taylor expansion. In [8] [9] the overall loss impact of each packet is estimated through the distortion introduced in the current frame, the average size of motion vectors, and the number of intra–coded macroblocks in the packet. In [10] the same task is performed by considering the distortion introduced in the current frame and the product of the number of intra–coded macroblocks in the subsequent frames. In [11] the loss impact of each packet is estimated from the distortion introduced in the current frame and the distance from the next key frame. These techniques, however, do not take into account the presence of B-frames.

In this work a hybrid approach is presented: the distortion introduced in the current frame is evaluated with the AbS

0-7803-8578-0/04/$20.00 ©2004 IEEE

method, while the distortion in future frames is estimated by means of an *error propagation model*. The relationship —as observed in a test video database— between the distortion introduced in the current frame and the distortion introduced in future frames is first analyzed. Then, a statistical model that predicts distortion propagation based on current–frame distortion is presented. This model can be applied to GOP structures derived from the periodic repetition of I-, P-, and B–frames even with sub–pixel motion compensation and in–loop decoding filtering. Results obtained on several, widely different, H.264 test sequences show that the proposed model successfully estimates overall distortion. Network simulations also show that model–based packet classification, when used for video transmission over DiffServ networks, delivers PSNR results which are consistently within 0.1 dB compared to the full AbS technique.

The paper is organized as follows. In Section II, the error propagation problem in video transmission is analyzed and the exact calculation of the overall distortion is described. In Section III, a statistical model for a low complexity estimation of error propagation is derived from experimental data. Simulation results in the context of video transmission over DiffServ networks are reported in Section IV. Finally, conclusions are drawn in Section V.

II. Perceptual importance of video packets

In a video bitstream not all bits are perceptually equally important. The distortion can be partially reduced by *error concealment* which exploits the residual correlation present in the compressed stream to reliably estimate the lost pixels. Errors introduced in the current frame may propagate to other frames because of temporal prediction [12]; error propagation is delimited by the presence of I–frames (provided that they are not in turn corrupted) and, therefore, we assume it affects only the current group of pictures (GOP). A well–known method to evaluate the perceptual importance of a video packet is to consider the overall distortion introduced in the decoded video in the case of its loss with respect to the uncorrupted version of the stream.

At the encoder side the set of lost macroblocks is not known; therefore, the impact on the overall distortion caused by the loss of a given macroblock should be computed as the weighted sum of the distortion values corresponding to the all loss patterns containing that macroblock. The distortion introduced by each loss pattern can be exactly computed during compression by simulating the concealment process and decoding the whole GOP. Let N be the number of macroblocks per GOP. Assuming that each macroblock may be lost independently of the others, 2^N different loss patterns should be considered. The exact calculation of the expected value of distortion for each macroblock is therefore a prohibitive task even for a modern CPU.

The Analysis–by–Synthesis (AbS) method [2] approximates the exhaustive approach by considering single (isolated) packet losses. If a macroblock is lost, it will introduce an overall distortion d_t that depends on the replacement data generated by the error concealment technique plus, in case of inter–frame prediction, the distortion due to error propagation to future frames. Such distortion is the sum of the *current–frame distortion*, d_c, introduced in the current frame, and the *future–frame distortion*, d_f, introduced in the frames that use the current one for prediction. Current–frame distortion only depends on how well the lost MB is concealed, while future–frame distortion depends on the error introduced in the macroblocks in other frames that reference the current one. The value of d_c is computed by applying the concealment algorithm which usually exploits information available in the normal encoding process. The exact value of d_t, instead, can be computed only by decoding all the frames that reference the lost and concealed block. Since this process must be repeated for all packets, the AbS distortion computation method still requires a considerable amount of CPU time and memory.

TABLE I

Average and standard deviation of $r = d_c/d_t$ compared with the model–estimated value as a function of frame type and GOP position (k) for all tested sequences.

Sequence	Frame type	k	r μ	r σ	model: $1/(k+1)$
akiyo	I	11	0.085	0.032	0.083
	P1	8	0.112	0.045	0.111
	P2	5	0.172	0.077	0.166
	P3	2	0.380	0.167	0.333
coastguard	I	11	0.109	0.071	0.083
	P1	8	0.133	0.071	0.111
	P2	5	0.180	0.071	0.166
	P3	2	0.349	0.130	0.333
foreman	I	11	0.140	0.100	0.083
	P1	8	0.162	0.084	0.111
	P2	5	0.219	0.084	0.166
	P3	2	0.406	0.134	0.333
mobile	I	11	0.084	0.032	0.083
	P1	8	0.118	0.045	0.111
	P2	5	0.164	0.032	0.166
	P3	2	0.333	0.118	0.333
mother daughter	I	11	0.085	0.045	0.083
	P1	8	0.110	0.055	0.111
	P2	5	0.169	0.071	0.166
	P3	2	0.360	0.148	0.333
silent	I	11	0.098	0.071	0.083
	P1	8	0.121	0.071	0.111
	P2	5	0.181	0.095	0.166
	P3	2	0.357	0.164	0.333
tempete	I	11	0.091	0.071	0.083
	P1	8	0.117	0.077	0.111
	P2	5	0.176	0.095	0.166
	P3	2	0.357	0.152	0.333
container	I	11	0.081	0.028	0.083
	P1	8	0.107	0.045	0.111
	P2	5	0.162	0.075	0.166
	P3	2	0.332	0.164	0.333

III. A statistical model of future distortion

We computed d_t and d_c for about 350,000 macroblocks belonging to several H.264 sequences using the Analysis–

by–Synthesis approach. Video material was encoded with 12 frames per GOP and two B–pictures between I– or P–pictures. The quantization stepsize was kept constant at 28 for I- and P-frames, and at 30 for B-frames; therefore the tested sequences have bitrates ranging from 350 kb/s up to 1.4 Mb/s.

Error concealment was implemented by replacing a missing MB with the MB in the same position, in the closest frame available at the decoder (i.e., the previous I– or P–frame if the loss is in I– or P-frames, and the previous frame in display order if the loss is in a B–frame). The distortion values were computed as the mean square error on luminance samples between the corrupted video and the correctly decoded one.

We studied the statistical distribution of the ratio $r = d_c/d_t$ for each position in the GOP (i.e., I–frame, first P–frame, and so on). The value of r is always non–negative and it is approximately zero when the distortion introduced in the current frame is negligible with respect to future distortions; the maximum value of r is one and is reached when no other macroblock references the current MB (e.g., when it belongs to a B–picture).

Figure 1 shows the distribution of r for the I–frames of the sequence *Coastguard*. Similar distributions have been obtained for all I–frames of the other sequences in the database.

Table I reports mean and standard deviation of the distribution of r for all tested sequences as a function of the frame to which the MB belongs. The mean value of r strongly depends on the frame position in the GOP, while it is almost independent of the sequence. The small values of standard deviation suggest that the distribution of r in each frame is quite compact and, therefore, the mean value of r can be used to estimate d_t from d_c. Moreover, it can be observed that the mean value of r is consistently quite close to $1/(k+1)$, where $k+1$ is the number of frames potentially affected if the MB gets lost, that is, the current frame and the k frames that use the current one for prediction (with 12 frames per GOP k=11, 8, 5, and 2 for I, P1, P2, and P3, respectively). We, therefore, propose to approximate the mean value of r with $1/(k+1)$. This value can be used to estimate the overall distortion d_t from the current distortion d_c thus avoiding the complete decoding of the GOP for each loss event being considered. The proposed approximation can be straightforwardly extended to any GOP structure delimited by I–frames.

The complexity of the proposed technique is reduced to the computation of d_c, which consists of the simulation of the error concealment process and the MSE computation (one multiplication per pixel). In our test implementation, error-concealed pixel values are a by–product of the decoder embedded in the encoder.

IV. Packet classification for DiffServ

The performance of the proposed method for distortion estimation has been tested in the context of packet classification for DiffServ networks [1]. The two-class network architecture is one of the simplest DiffServ scenarios. It defines

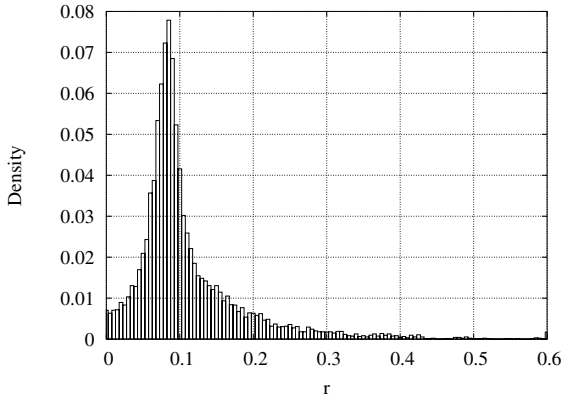

Fig. 1. Distribution of $r = d_c/d_t$ for all I–pictures of sequence *Coastguard*.

two classes of QoS, i.e., a high–cost *premium* service with nearly no losses and low and constant delay, and a regular *best–effort* service that provides the behavior of the current Internet. One of the simplest approaches to assign multimedia packets to the various DiffServ classes is to send the entire flow as high–priority traffic. High–QoS bandwidth, however, is a *limited* as well as an *expensive* resource that needs to be employed efficiently. A more advanced assignment approach could be, in the case of motion–compensated video, to classify packets depending on the frame type (I, P or B) they belong to. If data are coded using a scalable encoder, the resulting layers could be mapped to the DiffServ classes according to their importance. Given the highly *non–uniform perceptual importance* of speech, audio, image and video data, however, a more efficient approach is possible, one that takes into account the perceptual importance of each single packet to be transmitted.

In our simulations video packets have approximately constant size (550 byte) and contain a variable number of macroblocks. For each packet, the sum of the macroblock–level values of d_t is computed. For each GOP, the 20% of packets with the highest values of d_t are assigned to the premium class; the rest is sent as best–effort traffic.

Comparisons are made with respect to the full Analysis–by–Synthesis method. Table II reports the percentage of misclassified packets, i.e., the packets that, using the proposed distortion-estimation method, are assigned differently with respect to the AbS method. Less than 10% of the packets are misclassified for almost all the sequences in our database. Moreover, the distortions associated to the misclassified packets are quite close to the distortions estimated by the AbS method, suggesting that the perceptual impact of the misclassifications should be limited.

Transmission over a DiffServ network was then simulated. Packets belonging to the premium class were subject to 1% random uniformly distributed packet losses, while packets belonging to the best-effort class were subject to 10% packet

TABLE II

PERCENTAGE OF MISCLASSIFIED PACKETS AND PSNR VALUES FOR MODEL–BASED AND FULL ABS DISTORTION ESTIMATION (20% PREMIUM SHARE; 10% PLR IN THE BEST–EFFORT CLASS, AND 1% PLR IN THE PREMIUM CLASS, BERNOULLI PROCESS).

Sequence	misclassified packets (%)	PSNR (dB) model–based	AbS
tempete	4.8	28.4	28.4
foreman	6.2	29.5	29.4
mobile	3.4	24.5	24.5
news	6.7	34.5	34.4
akiyo	9.8	38.4	38.3
salesman	10.2	33.4	33.3
sean	7.7	35.1	35.2
paris	5.0	32.2	32.1

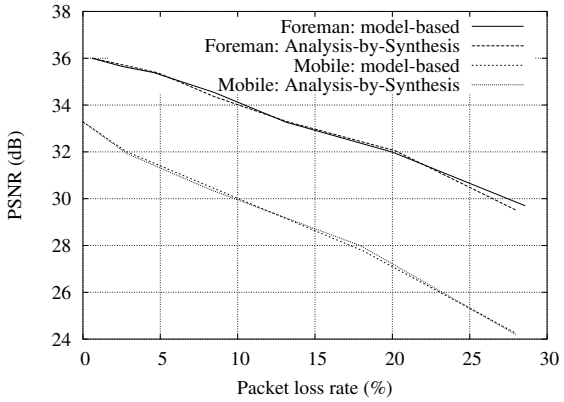

Fig. 2. PSNR as a function of packet loss rate, sequences *Foreman* and *Mobile*.

loss rate. The PSNR values of the decoded video sequence whose packets were classified using the proposed distortion-estimation technique are consistently within 0.1 dB compared to the AbS technique.

Network simulations were also conducted to further study the performance of video transmission over DiffServ based on the proposed distortion estimation technique. A two–class DiffServ network was implemented using a discrete–event simulator; the network has the classical bottleneck topology with two video sources at the same bitrate contending for the link capacity; the bottleneck router has a shared–buffer architecture with per–class FIFO policy and, in case of congestion, it discards packets starting from the lowest–priority class. Figure 2 compares the performance of the model–based approach with respect to the full Analysis–by–Synthesis technique as a function of packet loss rate. Results were obtained by simulating the transmission of the *Foreman* and *Mobile* sequences. The performance of the proposed model–based estimation technique is consistently very close to the performance of the Analysis–by–Synthesis technique with a fraction of its complexity. The same behavior was observed for all other sequences as well.

V. CONCLUSIONS

A method to determine the perceptual importance of video packets has been presented. The distortion introduced in the decoded video by the loss of each macroblock is the sum of two contributions which affect the current frame and future frames, respectively. Distortion results obtained on eight, widely different, H.264 sequences have shown that distortion affecting future frames can be reliably estimated from the current frame distortion and the position of the frame within the GOP. The method, applied to packet classification in a two–class DiffServ network, delivers PSNR results which are consistently within 0.1 dB compared to the Analysis–by–Synthesis technique. Simulations showed that the proposed method can successfully replace the Analysis–by–Synthesis technique for the assessment of the perceptual importance of video packets thus extending the usability of perceptual packet classification even to low complexity scenarios.

REFERENCES

[1] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, "An architecture for differentiated services," *RFC 2475*, December 1998.

[2] E. Masala and J. C. De Martin, "Analysis-by-Synthesis distortion computation for rate-distortion optimized multimedia streaming," in *Proc. IEEE Int. Conf. on Multimedia & Expo*, vol. 3, Baltimore, MD, July 2003, pp. 345–348.

[3] J. C. De Martin, "Source-driven packet marking for speech transmission over Differentiated-Services networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, Salt Lake City, Utah, May 2001, pp. 753–756.

[4] J. C. De Martin and D. Quaglia, "Distortion-based packet marking for MPEG video transmission over DiffServ networks," in *Proc. IEEE Int. Conf. on Multimedia & Expo*, Tokyo, Japan, August 2001, pp. 521–524.

[5] R. Zhang, S. L. Regunathan, and K. Rose, "Video coding with optimal inter/intra-mode switching for packet loss resilience," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 6, pp. 966–976, June 2000.

[6] ——, "Prescient mode selection for robust video coding," in *Proc. IEEE Int. Conf. on Image Processing*, vol. 1, 2001, pp. 974–977.

[7] ——, "Optimized video streaming over lossy networks with real-time estimation of end-to-end distortion," in *Proc. IEEE Int. Conf. on Multimedia & Expo*, vol. 1, Lausanne, Switzerland, August 2002, pp. 861–864.

[8] J. Shin, J. Kim, and C.-C. J. Kuo, "Quality-of-service mapping mechanism for packet video in Differentiated-Services network," *IEEE Transactions on Multimedia*, vol. 3, no. 2, pp. 219–231, June 2001.

[9] J.-G. Kim, J. Kim, J. Shin, and C.-C. J. Kuo, "Coordinated packet level protection employing corruption model for robust video transmission," in *Proc. SPIE Visual Communications and Image Processing*, San Jose, CA, Jan. 2001, pp. 410–421.

[10] I.-M. Kim and H.-M. Kim, "A new resource allocation scheme based on a PSNR criterion for wireless video transmission to stationary receivers over gaussian channels," *IEEE Transactions on Wireless Communications*, vol. 1, no. 3, pp. 393–401, July 2002.

[11] F. Zhang, M. R. Pickering, M. R. Frater, and J. F. Arnold, "Optimal QoS mapping for streaming video over differentiated services networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 5, 2003, pp. 744–747.

[12] B. Girod, N. Färber, "Feedback-based error control for mobile video transmission," *Proceedings of the IEEE*, vol. 87, no. 10, pp. 1707–1723, October 1999.