FEW DECODERS IN THE ENCODER: A LOW COMPLEXITY ENCODING STRATEGY FOR H.26L

Gabriella Olmo, Cristiano Cucco, Marco Grangetto, Enrico Magli

CERCOM - Center for Multimedia Radio Communications
Dipartimento di Elettronica - Politecnico di Torino,
Corso Duca degli Abruzzi 24 - 10129 Torino - Italy
Ph.: +39-011-5644094 - Fax: +39-011-5644149
olmo,(c.cucco,grangetto,magli)@polito.it - www.cercom.polito.it

ABSTRACT

In this paper we propose a reduced complexity technique for the rate-distortion optimization in JVT/H.26L in the presence of packet erasures. It is named *few decoders in the encoder*, and is based on the idea of generating a selected number of error patterns in the encoder, so that a limited number of co-decoding processes can be implemented to estimate the transmission distortion term. The correlation amongst packet erasures is taken into account employing a binary Gilbert model. The proposed algorithm exhibits competitive performance in terms of average PSNR and probability of decoding failure, with very affordable complexity and memory requirements.

1. INTRODUCTION

In this latter years, there has been an increasing interest in multimedia applications over networks subject to bit errors or packet erasures. In order to provide acceptable video quality in such situations, a compromise must be identified between Quality of Service (QoS), bandwidth or transmission rate, and delay. This can be achieved implementing error resilient techniques at the encoder, as well as forward error correction at the network adaptation layer, and concealment of the residual errors at the decoder side. Resilience tools, such as reversible variable length coding (RVLC), insertion of intra-macroblocks, data packetization, are included in all recent video standards such as H.263+ and MPEG-4. However, it is well recognized that most of these methods imply a reduced compression efficiency. A better trade off between bandwidth and resilience can be identified taking the distortion due to transmission errors into account in the rate-distortion (R-D) optimization. In this way, the amount of redundancy necessary to cope with the actual network status is introduced. In [1], intra/inter macroblock (MB) code selection and packetization are jointly optimized, taking into account the distortion due to error propagation. In [2], an optimization of the forced intra MB coding and the placement of synchronization markers is pursued, considering the impact of the data partitioning on error concealment. Also the emerging JVT/H.26L video standard [3] aims at achieving error resilience levels suitable for conversational services, without significant increasing the output bit rate. It addresses an optimum R-D optimization for the MB mode selection, where the distortion term takes into account not only the source coding, but also the network behavior and the error concealment strategies implemented at the decoder. In other words, the distortion *at the receiver* is estimated and employed for the optimal coding mode selection on a macroblock basis. As it is not realistic to assume that feedback information on the network status is available in real time, the effect of packet erasures is estimated at the encoder generating a given number K of possible error patterns, based on the network loss probability distribution. The sample distortion related to each error pattern is then evaluated simulating a co-decoding process, and the expected distortion is estimated by averaging over the K available samples. This approach, although being simple and effective, suffers from a relevant computational complexity and huge memory requiements. In fact, in order not to exhibit appreciable suboptimality, quite large K values must be addressed.

In this paper, we start from the described optimization strategy, named *many decoders in the encoder* (MDE), and address the issue of drastically reducing the computational complexity and memory requirements, at the expenses of slight or nearly appreciable performance degradation; this approach is named *few decoders in the encoder*. Moreover, having recognized the need to properly assess the network loss distribution in order to actually achieve optimal performance, we refine the model employed in the encoder for the distortion estimation, assuming that losses are distributed according to a Gilbert model.

2. RATIONALE

Many recent papers address Lagrangian MB mode decision for the selection of the best coding mode in a R-D optimization sense. Formally, the optimization can be expressed by finding the best code option o^* such as

$$o^* = argmin_{o \in \mathcal{D}}(D(o) + \lambda R(o)) \quad (1)$$

where $\mathcal{D}$ detotes the set of all possible coding options o for the current MB, and $D(o)$, $R(o)$ are respectively the overall distortion and the resulting rate when coding mode o is selected. In practice, whereas the evaluation of $R(o)$ is trivial, being simply the number of bits assigned to the current MB when it is encoded with coding option o, the evaluation of the distortion term $D(o)$ is far more complicated. In fact, an overall optimization has to be performed, taking into account also the transmission errors and the concealment operated at the receiver; in other words, the distorsion at the receiver must be evaluated. Whereas the concealment strategy can be assumed known at the receiver side, this is not true as for the network behaviour. Even though feedback mechanisms have been

considered to make the receiver aware of the network conditions, this task cannot be performed in real time; therefore, it cannot be assumed that the receiver has knowledge of the network behavior, as for the current MB being coded. Following the same notation of [4, 5], let us define the network behavior when the n-th frame is transmitted as a binary sequence $c_{\pi(n)}$, where $\pi(n)$ is the number of packets necessary for transmitting the n-th frame; a value "0" in the channel sequence means that the packet has been correctly received, whereas a value "1" denotes packet erasure. As the network behavior is not known at the receiver, it is modelled as a random variable (RV) $C_{\pi(n)}$.

In [4, 5] the channel distortion is estimated at the receiver, on a pixel-by-pixel basis. Let $s_{n,m,i}$ and $\hat{s}_{n,m,i}(o, C_{\pi(n)})$ be respectively the true and the reconstructed values of the i-th pixel in the m-th MB in frame n, with $i = 1, \cdots, I$, $m = 1, \cdots, M$, $n = 1, \cdots, N$. The reconstructed value is obviously dependent on the selected coding option o and on $C_{\pi(n)}$. Then, the distortion contribution yielded by this pixel can be written as

$$d_{n,m,i}(o) = |s_{n,m,i} - \hat{s}_{n,m,i}\left(o, C_{\pi(n)}\right)|^2$$

and the overall distortion of the m-th MB in frame n is

$$D_{n,m}\left(o, C_{\pi(n)}\right)) = \frac{1}{I}\sum_{i=1}^{I} |s_{n,m,i} - \hat{s}_{n,m,i}\left(o, C_{\pi(n)}\right)|^2$$

Clearly, the overall distortion is a RV depending on the network behavior; the problem is then to estimate its expected value. In the MDE technique, it is assumed that K realizations $c_{\pi(n)}(k)$, $k = 1, \cdots, K$ of the RV $C_{\pi(n)}$ are available. If K is large enough, the expected value of the distortion is estimated by averaging the sample distortion terms, evaluated for each sample sequence $c_{\pi(n)}(k)$:

$$E_{C_{\pi(n)}}\left\{D_{n,m}\left(o, C_{\pi(n)}\right)\right\} \approx \frac{1}{K}\sum_{k=1}^{K} D_{n,m}\left(o, c_{\pi(n)}(k)\right) \quad (2)$$

This approximation is accurate in the limit $K \to \infty$, and in practice it holds for K values "large enough". This implies huge computational burden and memory requirements at the encoder side. In fact, K co-decoding processes must be implemented for each MB being encoded. Whether such complexity can be afforded in practice or not is a rather questionable point.

3. FEW DECODERS IN THE ENCODER

Assuming the MDE method as a starting point, we want to devise an alternative technique, with the goal of drastically reducing the computational complexity and the memory requirements, even at the expenses of a slight performance degradation.

The MDE is almost optimal, as it takes into account potentially all the possible network behaviors to estimate the expected distortion value. The key idea leading to the FDE algorithm is to consider only a properly selected subset of random error patterns, making some assumptions in order to limit the number of co-decoding simulations for each encoded MB.

Assumption 1: a window of N frames preceeding the current one is considered, and no error propagation occurs from data belonging to frames not included in such window. In general, this assumtion is not exactly satisfied; for example, in the common situation when only the first frame in the stream is intra coded, there is potential error propagation from any frame in the stream. However, the assumption can be made adequately precise by selecting a proper value of N.

Assumption 2: the motion vectors of m-th MB in the current frame n are null. This implies that the possible errors impacting on the distortion of this MB are only those localized in the spatially correspondent MB in frames l, $n-N \leq l \leq n-1$. It is worth noticing that this assumption, although being rather abrupt, is coherent with the principle of selecting a very efficient encoding method.

Under these assumptions, all the possible 2^N error patterns *at the macroblock level* (or *MB error patterns*) can affect the distortion of the m-th MB. In fact, only the spatially correspondent MB belonging to the N frames in the window are involved in the prediction of the MB at hand. Therefore, in the FDE algorithm, the 2^N MB error patterns are generated in the form of binary sequences of length n, where a "1" denotes that the MB has been lost, whereas a "0" means that the MB has been correctly received; these sequences are used to estimate the expected distortion. This approach is different from that addressed in [5], where the network behavior is simulated at the packet level. Moreover, the packet loss patterns are considered equiprobable, as can be inferred from Eq. 2, where the expected distortion is evaluated as a not weighted arithmetic mean value. Here, we associate to each MB error pattern a probability of occurrence $P(k)$, $k = 0, \ldots, 2^N - 1$, and the distortion is evaluated as:

$$E_{C_{\pi(n)}}\left\{D_{n,m}\left(o, C_{\pi(n)}\right)\right\} \approx \sum_{k=0}^{2^N-1} P(k) D_{n,m}\left(o, c(k)\right)$$

where $D_{n,m}\left(o, c(k)\right)$ is the MB distortion given the k-th error pattern at the MB level, and $P(0)$ is the probability that the MB is correctly received. The probability of occurrence of the MB error patterns is evaluated invoking the network characteristics at packet level, and inferring the statistical properties at MB level exploiting the rule employed to map frames into packets. More details on this point are given in Sect. 4, where the underlying network model is described.

In practical terms, the main difference bewteen FDE and MDE is that the former requires 2^N instead of K simulated co-decoding processes at the encoder. Since the impairment due to transmission errors somehow decays with time [5], a limited window duration of 2-3 frames is expected to yield satisfactory performance. This is confirmed by the simulation results presented in Sect. 5. Therefore, FDE is expected to exhibit a drastically reduced computational complexity with respect to MDE.

The selection of the Lagrange multiplier λ in Eq. 1 devises some attention. It is well known [4] that the following expression holds:

$$\lambda = (1-p)p_c\lambda_0$$

where λ_0 is the Lagrange multiplier for error-free transmission, p is the MB loss probability, and p_c is the probability that the referenced image part is correct. It is recognized that the estimation of this latter parameter is not straightforward. In [4], it is stated that this probability decreases with increasing distance to an intra refresh of a certain region, and also depends on p. In this paper, we estimate p_c under the assumptions that the motion vectors of all inter coded MBs are null. In this case, p_c is the probability that the m-th MB in all frames between l and $n-1$ are correctly received, l being the nearest frame whose m- th MB is Intra coded. The

direct evaluation of this probability is performed using the packet loss model addressed in Sect. 4.

4. PACKET LOSS MODEL

Another aspect to be controlled in order to assure a good approximation of the overall R-D optimization is the model of packet losses which describes the underlying network behavior. A common assumption is that losses are statistical independent of each other, and identically distributed (IID); in this case, the only relevant parameter is the probability of packet loss, which can be easily estimated. In spite of its simplicity, such a model does not adequately represent those contexts where losses occur in bursts; this behavior is typical of the Internet, where congestion and excessive delay in packet delivery are responsible of the packet losses, as well as of transmission over wireless channels affected by fading and shadowing. In this paper, we take into account the correlation among subsequent packet losses and employ the simple yet effective Gilbert model for evaluating the probability of each loss pattern realization. The Gilbert model [6] is a 2-state Markov chain; in the G (good) state, packets are lost with probability p_g, whereas in the B (bad) state they are lost with probability p_b. In its simplest binary version, also adopted in this paper, $p_g = 0$ and $p_b = 1$. The evolution of the Markov chain is dictated by the probabilities α and β of sojourn in either state. If W_B is a random variables denoting the time of sojourn in the B state, it can be shown that its expected values is $L_B = E\{W_B\} = (1-\alpha)^{-1}$ This parameter represents the expected length of error bursts, and it one of the two parameters employed to describe the Gilbert model. The other significant parameter is the expected loss probability, which, in the binary case, can be written as

$$p = \frac{1-\beta}{2-\alpha-\beta}$$

In this case, it also coincides with the probability P_B that the system is in state B.

This model has been employed to evaluate the probability $P(k)$ of the k-th error pattern at the MB level addressed in the FDE algorithm. Clearly, this probability depends on the packet formation strategy, on the actual position of the MB, within the packet stream, and on the network behavior during the transmission of the preceeding packets. Therefore, the network statistics at the packet level must be translated into statistics at the MB level. Taken into account that the motion vectors of m-th MB are considered null, $P(k)$ can be formalized as a chain of states assumed by the network during the transmission of a number of packets necessary to embed the N frames included in the window. More details on the probability evaluation can be found in [7].

5. SIMULATION RESULTS

The proposed algorithm has been tested using "Foreman" (QCIF, 7.5 fps) and "Mother and Daughter (M-D)" (QCIF, 10 fps) encoded with the H.26L test model JM1.4 The frames are divided into slices, and each slice is mapped into a single network adaptation layer packet (NALP). A trade off has been identified between global overhead and packet length, considering that the probability that a packet is affected by at least one error increases with the packet length. These considerations led to the selection of 512 bytes for Foreman and 400 bytes for Mother and Daughter as for the average slice dimension. The NAL overhead, as well as the RTP/UDP/IP compressed header (3 bytes), the PPP/PDCP header (2 bytes) and the RLC (4 bytes) are taken into account in the bit rate constraints. Entropy coding based on UVLC is addressed, and the simple "previous frame concealment" method is applied. As for the packet loss statistics, both the independent loss model and the binary Gilbert model are considered.

The algorithm has been tested using the common test conditions for packet switched conversational or streaming services over 3G mobile networks. In particular, we have employed Patterns 2 and 3, the former being representative of severe channel conditions, whereas the latter represents milder conditions, suitable for conversational services with retransmissions not allowed. The characteristics of the patterns employed are summarized in Table 1. The parameters P_B and L_B of the Gilbert model employed at the encoder have been evaluated based on those error patterns, leading to:

Pattern 2: Foreman: $P_B = 0.095$, $L_B = 1.9$; M-D: $P_B = 0.075$, $L_B = 1.9$.

Pattern 3: Foreman: $P_B = 0.025$, $L_B = 1.5$; M-D: $P_B = 0.023$, $L_B = 1.6$.

Pattern	Bit rate	duration	BER	Speed
2	64 kbps	60 s	9.3e-3	3 km/h
3	64 kbps	180 s	5.1e-4	3 km/h

Table 1. Packet loss patterns (after [8]).

The effect of employing a Gilbert model in the encoder has been evaluated. To this end, the MDE algorithm with several values of K has been compared with a modified version including a Gilbert model in the encoder. The performance have been compared in terms of the average PSNR and the probability of failure P_f, defined as the probability that the average PSNR drops below a predetermined threshold, so that the quality is considered unacceptable. The situations of failure have not been included in the PSNR computation. The threshold has been set at 24 dB for Foreman and 30 dB for M-D, this latter being characerized by a higher PSNR than the former encoded at the same rate, in the absence of transmission errors. The rate is controlled via the quantization parameter QP, which has been set to 20 for Foreman, and 13 for Mother and Daughter. The results obtained employing the error pattern 3 are summarized in Table 2.

It can be noticed that, as expected, the performance of MDE increases with K, although a saturation effect can be appreciated when K exceeds 40. This is also confirmed from simulations performed employing the error pattern 2, not reported for brevity. The Gilbert model turns out to be advantageous, showing a performance improvement in the range 0.1 - 0.6 dB with respect to the standard MDE algorithm. This quality improvement is obtained at the expenses of a slightly larger rate; this means that the use of the Gilbert model in the encoder leads to the selection of a different optimization point of the R-D curve.

The FDE algorithm is then validated, comparing its performance with MDE and different number of simulated co-decoding processes, K. The results are summarized in Table 3. For both pattern 2 and 3, the two sequences Foreman and M-D have been encoded using

- FDE with 4 and 8 simulated decoders (corresponding to $N = 2$ and $N = 3$ respectively)

Loss model	K	Rate (kbps)	PSNR (dB)	P_f (%)
Foreman				
MDE(BSC)	20	54.44	31.09	7.00
MDE(Gilbert)	20	57.79	31.48	5.17
MDE(BSC)	40	57.57	31.47	5.30
MDE(Gilbert)	40	58.92	31.57	6.03
MDE(BSC)	60	58.77	31.38	5.93
MDE(Gilbert)	60	60.99	31.65	4.41
MDE(BSC)	80	58.45	31.54	4.07
MDE(Gilbert)	80	63.52	32.11	1.17
M & D				
MDE(BSC)	20	50.33	37.04	1.75
MDE(Gilbert)	20	52.81	37.41	0.50
MDE(BSC)	40	50.58	37.42	1.56
MDE(Gilbert)	40	52.56	37.52	2.17
MDE(BSC)	60	54.10	35.53	1.63
MDE(Gilbert)	60	53.72	37.58	1.94
MDE(BSC)	80	50.69	37.27	1.69
MDE(Gilbert)	80	52.79	37.62	2.14

Table 2. Performance of standard and modified MDE, this latter employing the Gilbert model; error pattern 3.

- MDE with $K = 8$, and $K = 80$; the former has been selected as it allows for direct comparison with FDE and $N = 3$, the latter being representative of asymptotic performance.

The comparisons are made in terms of achieved rate, average PSNR, and P_f as previously defined. The quantization parameter QP is also reported.

¿From the presented results, it can be observed that a window duration $N = 3$ is adequate. If the FDE algorithm with $N = 3$ is compared with MDE and $K = 8$, it yields superior performance in terms of both PSNR (with a gain between 0.42 and 1.45 dB) and P_f, which is always considerably lower. It must be recalled that the two algorithms are directly comparable in terms of computational complexity, this latter being linearly dependent on K. On the other hand, the performance of MDE with $K = 80$ is superior with respect to FDE with $N = 3$. This is not surprising, as MDE with an adequate number of simulated co-decoding processes approaches the optimum R-D performance. In detail, FDE exhibits a loss ranging between 0.27 and 0.77 dB in the considered situations. However, this performance gap is limited, considering that it is achieved at the expenses of a computational complexity which is approximatively ten times larger for MDE with respect to FDE.

6. CONCLUSIONS

In this paper we propose a reduced complexity method, named *few decoders in the encoder*, for the rate-distortion optimization in JVT/H.26L in the presence of packet erasures. The algorithm exhibits competitive performance in terms of average PSNR and probability of decoding failure, with very affordable complexity and memory requirements. Future research can be in the direction of including more sophisticated concealment strategies in the optimization task, and the combination of rate control schemes.

Method	K	Rate (kbps)	PSNR (dB)	P_f (%)
PATTERN 2				
Foreman ($QP = 23$)				
FDE	4	56.45	28.97	11.48
FDE	8	60.15	29.67	7.39
MDE	8	53.32	28.91	16.13
MDE	80	63.83	30.01	8.25
M & D ($QP = 14$)				
FDE	4	55.81	35.11	5.70
FDE	8	61.30	36.59	1.63
MDE	8	48.61	35.14	8.57
MDE	80	55.94	36.86	5.14
PATTERN 3				
Foreman ($QP = 20$)				
FDE	4	57.51	30.51	9.10
FDE	8	60.93	31.34	4.50
MDE	8	55.44	30.92	9.77
MDE	80	63.52	32.11	1.17
M & D ($QP = 13$)				
FDE	4	52.20	36.61	3.38
FDE	8	54.07	37.18	1.63
MDE	8	50.59	36.76	2.94
MDE	80	52.79	37.62	2.14

Table 3. Comparison between MDE, FDE. For both algorithms, K is the number of co-decoding processes simulated at the encoder.

7. REFERENCES

[1] G. Cote, S. Shirani, and F. Kossentini, "Optimal mode selection and synchronization for robust video communications over error-prone networks, " *IEEE Journal on Sel. Areas in Comm.*, vol. 18, no. 6, pp. 952-965, June 2000.

[2] K. H. Yang, D. W. Kang, and A. F. Faryar, "Efficient intra refreshment and synchronization algorithms for robust transmission of video over wireless networks," in *Proc. 2001 Int. Conf. on Image Proc. (ICIP)*, vol. 1, pag. 938 -941, 2001.

[3] T. Wiegand, *JVT working draft 2*, JVT-B118r7, Apr. 2002.

[4] T. Stockhammer, D. Kontopodis, and T. Wiegand, "Rate-Distortion optimization for JVT/H.36L video coding in packet loss environment," in *Proc. 12th Int. Packet Video Workshop*, Pittsburg, PY, May 2002.

[5] T. Stockhammer, M. M. Hannuksela, and T. Wiegand, "Error Protection and Feedback Strategies for Wireless Progressive Video Transmission," *IEEE Trans. on Circuits and Systems for Video Tech.*, Sept. 2002 (to appear).

[6] E. N. Gilbert, "Capacity of a burst noise channel," *Bell Systems Tech. Journal*, vol. 3, pp. 1253-1265, Sept. 1960.

[7] C. Cucco and G. Olmo, "Codifica robusta per transmissione video su reti a pacchetto: soluzioni per lo standard H.26L,", Internal Report no. 2002/DELEN/TLC/IPL/57-A, May 2002 (in Italian; available on demand).

[8] ITU-T VCEG-N37 "Common test conditions for RTP/IP over 2GPP/2GPP2 -Software and amendments," VCEG(SG16,Q6), XIV Meeting, S. Barbara, Sept. 2001.