Improving bit-rate and quality control for MPEG-2 video sources

Giancarlo Cicalini*, Lorenzo Favalli*, Alessandro Mecocci**

*Università di Pavia,- Dipartimento di Elettronica via Ferrata, 1, I-27100 Pavia (PV) Italy;
Tel: +39-382-505923; fax: +39-382-422583; e-mail: lorenzo@comel1.unipv.it

**Università di Siena,- Facoltà di Ingegneria; via Roma, 77, I-53100 Siena (SI), Italy
tel: +39-577-2636041 fax: +39-577-263602; e-mail: mecocci@comel1.unipv.it

ABSTRACT.

In video compression techniques, it is very important to implement the most efficient bit allocation strategy in order to achieve the best quality with the minimum number of bits. This paper presents a new feedback/feedforward controller, for MPEG-2 coding, that dynamically tunes the quantization parameters by analysing the image sequence from a psycovisual point of view. The analysis is carried out on an 8x8 pixels block basis to determine the visual characteristic of each macroblock. This pre-analysis classifies macroblocks and assigns quantization parameters to them according to a proposed scale measuring their visual relevance. A post-analysis procedure provides the final tuning. The system generates images with higher quality with respect to the standard Test Model 5.

1 INTRODUCTION

The need for a digital format for images has been growing in the recent years and has produced a series of standards for specialised applications such as video telephony [1], and still pictures [2]. Moving from these standards, new efforts were started in the early 1990s to define a new standard which could cover a wide variety of applications. The work was co-ordinated among ITU-T and ISO groups Since the very beginning of the standardisation process, it was the aim of the working committees to achieve an agreement on a video coding standard capable to cope with the emerging multimedia applications as well as with the traditional video activities: storage and broadcasting. The push to the standardisation work (usually a vary slow process) came from the rushing increase of the demand for video services that was leading to a set of (incompatible) proprietary structures.

If network transmission is a target for digital video, the standardisation process needed to take into account the emerging network standard, ATM [3] with all the constraints on timing and actual bandwidth control schemes imposed by the transmission of high bit-rate variable load video services [4] on a packet network.

The result of these efforts has been MPEG-2 [5], which is a video compression and transmission standard for digital image sequences with bit-rates in the range 2-30Mbit/sec (broadcast to high definition image quality). The main characteristics of the MPG-2 coder are outlined in section II of the paper: a more comprehensive description can be found in [6], [7].

The MPEG-2 standard specifies the techniques to be used and general format of the bit stream, but allows a form of freedom in the choice of some key parameters, such as quantization strategies. Together with the standard, a series of Test Models (TM) [8] have been provided to compare the performances of the different proposed implementations. In this paper, an MPEG-2 coder structure is proposed that is fully compatible with the standard requirements, but allows a better bit allocation strategy and results in better image quality and an almost exact matching with the objective bit rate that should be enforced for network transmission. The structure of the proposed coder is described in section III and the performances are reported in section IV for the colour sequences *calendar and mobile, flowers* and *table tennis* at PAL resolution using TM5 as a benchmark. We regret we are not able to include "visual" results in the paper, but the printing process would occlude a clear evaluation of the comparison.

2 THE MPEG-2 STRUCTURE

MPEG-2 is a coding standard with a layered structure that divides a sequence into *Groups Of Pictures* (GOPs). Each picture in a GOP is in turn divided into 16x16 pixels *Macro Blocks* (MBs), that are further decomposed into 8x8 pixel *blocks*. *Slices* of an arbitrary number of adjacent macroblocks can be introduced for synchronisation purposes. Motion estimation and Discrete Cosine Transform (DCT) are used to reduce respectively the temporal and spatial correlation. In each GOP at least one frame is a "reference frame" coded without motion estimation: this frame is said to be an *Intra-coded* frame (*I*-frame). For all other frames, forward/backward interpolation/prediction is used, whichever appropriate, to achieve the highest compression. *Predicted* frames (*P*-frames) are derived directly from the last available I-frame. P-frames don't follow immediately the I-frames, they are also used to derive the intermediate frames which can be coded using both I- and P-frames. Since these frames can be obtained from a past image (the I-frame) and from a subsequent frame (the P-frame) they are named *bi-directional* (*B*-) frames. Since the decoder will need both the I- and the P-frame to decode the B-frame, the transmission order is altered from the original one and the P-frame is transmitted before the B-frames so that the twelve frames GOP IBBPBBPBBPBBI is transmitted as IPBBPBBPBBPIBB.

The coefficients resulting from the DCT transform of the 8x8 pixel blocks are quantized according to a *quantization* *matrix* that is scaled by a *quantization factor* (q_n), to achieve the target average-bit-rate.

An output controller then monitors the amount of bits produced and dynamically changes the quantization parameters. In fact, it often happens, due to estimation errors, that too many bits are allocated to a certain image or to part of it. In this case the coder must reduce the number of bits used for the remaining frames in the GOP. Consequently, the quality of the reconstructed sequence is not uniform and visually disturbing effects are induced.

3 THE PSYCOVISUAL CODER

The bit allocation and rate control strategy defined by TM5 is based on the assumption that there exists an inverse proportionality relationship between the bit number and the quantization parameter. The resulting parameter is called *complexity* and is supposed to be constant. When the total number of bits per GOP is known, a first level strategy for allocating the bits to the different classes of frames (i.e. I-frames, P-frames, B-frames) is implemented by considering the frame-class complexity given by the tree equations

$$X_I = C_I \cdot Q_I; \quad X_P = C_P \cdot Q_P; \quad X_B = C_B \cdot Q_B;$$

where X_x represents the complexity of the corresponding type of frame, C_x represents the number of bits required to code that frame, and Q_x represents the average value of the quantization parameters used to quantize the MBs. It can be shown that the number of bits to be allocated to the different types of frames are given by

$$T_I = \frac{C_I}{\left(C_I + N_P \cdot C_P + N_B \cdot C_B\right)};$$

$$T_B = \frac{R \cdot C_B}{\left(N_P \cdot C_P + N_B \cdot C_B\right)};$$

$$T_P = \frac{R \cdot C_P}{\left(N_P \cdot C_P + N_B \cdot C_B\right)};$$

where N_x represents the number of I, P, and B frames in the current GOP, while R is the number of remaining bits for the GOP.

This hypothesis is somewhat simplified: a more detailed evaluation shows that complexity cannot be considered a general characteristic but exhibits a slightly different behaviour for different sequences. The inverse proportionality relationship can be applied at a MB level and can only be considered constant for small variations of the n-th MB quantization parameter q_n. Unfortunately, since MPEG-2 uses a DCT, the actual number of bits depends on the image energy distribution, therefore it is not possible to define a priori the *quantization parameter* q_n that grants the required bit-rate.

The complexity X_n of the n-th MB can be obtained by pre-coding it (performing the DCT) using a *profile* p_n representing a good initial estimate of the quantization parameter. At this point the number s_n of bits actually needed for the n-th MB is known: if T is the total number of bits assigned to the frame, we propose to allocate

$$t_n = s_n \cdot \frac{T}{\sum_{i=1}^{MB\#} s_i}$$

bits to the n-th MB. Being the complexity almost invariant, it is possible to obtain the quantization parameter that gives the desired number of bits, from the following formula (since s_n, p_n, and t_n are known).

$$X_n = s_n \cdot p_n = t_n \cdot q_n$$

In order to maximise the image quality for a given amount of allocated bits, it is necessary to keep in mind that the number of bits required to achieve a predefined level, is strictly related to the image type. By iterating this concept, we can state that the number of bits can be tied to the MB type. It is consequently of paramount importance to define a methodology capable to determine a sufficiently fine classification of the MBs based on the visual impact of distortions introduced by the coding process. By doing this we introduce a *psycovisual* description of the image which is carried out by a *pre-analysis* step. The pre-analysis section classifies each luminance block (8x8 pixels) using statistical and perceptual parameters instead of classical energy-based measures. The classification is performed by comparing the luminance gradient with a threshold derived from the Weber-Fechner law and from the Visual Masking Factor [9]. A fast tree-based classifier is used for classification. To account for the different sensibility of the human eyes to different spatial patterns, other statistical properties of the block are computed by evaluating the number of horizontal and vertical variations of the luminance gradient. The information of the four blocks that belong to, say, the n-th MB are then combined according to the following formula to give a single final estimate *($Dist_n$)* of the visual importance of that MB

$$Dist_n = \alpha \sum_{i=0}^{3} (log(1 + loc\_act_i/112) \cdot pv\_fact_i$$

where α is a scaling factor (experimentally set to 5 in our tests). pv_fact_i is a parameter that accounts for the eye sensitivity to the i-th block class (0.7 if it is an edge block, 1.0 if it is a uniform block, and 1.2 if it is a textured block). Loc_act_i is the i-th block activity. The result is to identify the MB based on the properties of the blocks inside it as either *edge, uniform,* or *textured.* :

- if at least three blocks in the MB are uniform the MB is classified as uniform;
- if three or more blocks are edge blocks, then the MB is considered an edge MB;
- if there are at least three texture blocks, the MB is said to be textured;
- if neither of the above conditions is verified, the MB is mixed

The MBs with a low value of $Dist_n$ need more bits, i.e. higher quantization resolution.

To define the image complexity during the pre-coding step, it is necessary to know the quantization profile p_n for each MB. In the proposed system, the profile for the current frame is obtained by taking the average value of the quantization profile in the previous frame of the same type (I-P-B-frame) and then by scaling it according to the formula

$$Pv\_mod_n = W_{\mathbf{min}} + \frac{\left(W_{\mathbf{max}} - W_{\mathbf{min}}\right)}{(30 - \alpha)^2} \cdot Dist_n^2$$

where W_{max} and W_{min} define the scaling range and have been set to 2.2 and 0.35 respectively, during the simulations. The quantization profile for the n-th MB is given by $p_n = Q \cdot Pv\_mod_n$. This approach does not change the MPEG-2 compression scheme and modifies only the output bit-rate control routines that are left free in the standard. As a final remark, the proposed strategy keeps memory of previous frames to estimate the quantization profiles. This rises some problems in the case of scene changes. To grant a fast convergence to the correct values after a scene change, the quantization parameter is computed according to the slightly different formula

$$q_n = p_n \cdot \sqrt{(s_n / t_n)}\,.$$

The experiments show that, by using this formula, only one frame delay is needed to converge to the new true parameters after a scene change (see Figure 3 frame #11, for example).

The final tuning of the quantization parameter is carried out by means of a *post-analysis* block constituted by a feedback controller of the integral type. The controller tunes the q_n for each MB. A simple linear relationship is considered between q_n and the number of bits produced by using this parameter. The non linearities are modelled as an *additive noise*. By linearizing the functions for the q_n around their mean value q_{med} it is possible to obtain the following formula

$$G = [ds / dq]_{q_{med}} = -X_{med} / q_{med}^2$$

and consequently $\Delta s = G \Delta q$ that leads to a simple integral controller.

Finally, a *smoothing buffer* has been added to obtain a nearly constant output bit-rate for applications such as the transmission of MPEG-2 video over ATM networks where rate control algorithms are implemented [10]. The buffer control strategy is fully compliant with the MPEG-2 standard coding scheme, that limits to three frames the maximum allowed delay.

The proposed coder structure is summarised in Figure 1. It is important to note that the standard MPEG-2 bit-stream is left unchanged. In fact, the additional blocks of the proposed scheme run in parallel to the usual scheme and only affect (dynamically) the parameters of the standard coder.

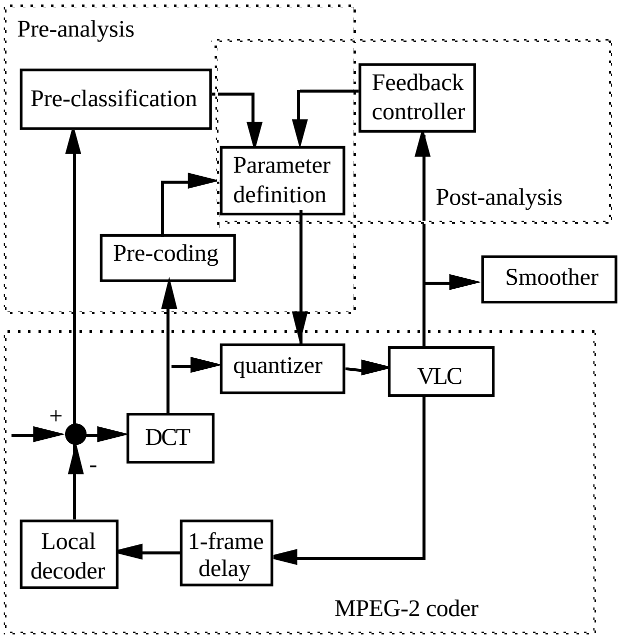

Figure 1. The proposed Coding Scheme

IV. EXPERIMENTAL RESULTS

Experiments have been carried out by using standard video sequences widely adopted for testing, namely: *calendar*, *flowers*, and *tennis table* at PAL resolution. The results of the proposed coding scheme have been compared to those obtained with MPEG-2 Test Model 5 at various target bit-rates. The simulations show that the proposed scheme performs much better then MPEG-2/TM5 and gives a constant error level at all bit-rates. This is different from what happens by using the TM5 that gives very poor results when the allowed bit-rate is low (see Table 1 and Fig. 2 and 3). The fidelity of the coding/decoding process has been evaluated in terms of Peak-SNR (see Figure 4): The PSNR has been defined as

$$PSNR = 10Log_{10} \frac{(255)^2}{\sum_{i=0}^{nx} \sum_{j=0}^{ny} \left(255 - rec[i][j]\right)}$$

where *nx* and *ny* are the number of columns and rows of the image and *rec[i][j]* is the received pixel. This numerical evaluation has been supported by subjective tests, performed by means of a control group: these tests have shown that for the same PSNR the perceived quality of the images is much better than that of the TM5.

Finally, as can be noted in Fig. 5, the output bit-rate is almost perfectly constant also in presence of scene changes (frame #11); this fact proves the effectiveness of the overall control strategy. Moreover, the proposed system is also suitable for renegotiation strategies, where the ATM could dynamically change the allocated bandwidth decreasing or increasing it depending on the current network load.

GOP	Prevision error (bits): new scheme	Proposed scheme %	Prevision error (bits): TM5	TM5 %
Calendar 3Mbit/s				
1° GOP	-503	-0.044%	-155,237	-12.39%
2° GOP	200	0.013%	5,615	3.58%
Calendar 5Mbit/s				
1° GOP	-712	0.029%	-20,639	-1.03%
2° GOP	-238	0.001%	6,762	0.28%
Flowers 3Mbit/s				
1° GOP	1.049	0.087%	-199,733	-16.44%
2° GOP	314	0.021%	5,2198	3.62%
Flowers 12Mbit/s				
1° GOP	553	0.011%	-1,257	-0.026%
2° GOP	-184	-0.0038%	-6,024	-0.1%

Table 1 Comparative performance between TM5 and the proposed scheme

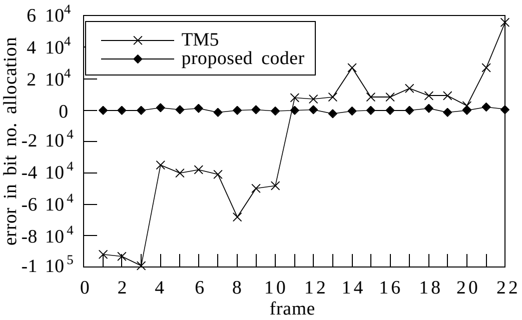

Figure 2. Error in the allocation of the number of bits: sequence calendar at 3 Mbit/s.

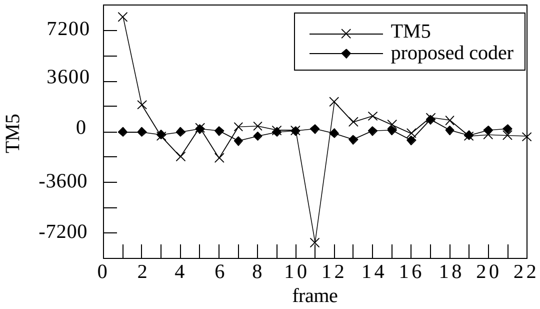

Figure 3. Error in the allocation of the number of bits: sequence flowers at 9 Mbit/s.

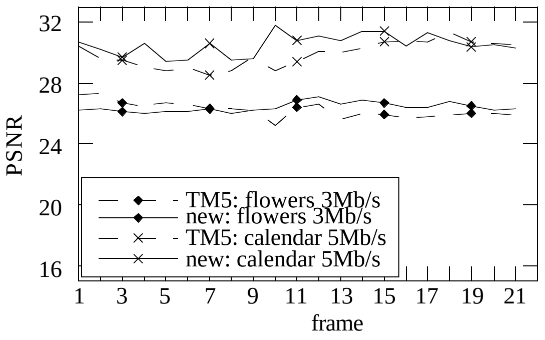

Figure 4. PSNR at 5Mbit/sec and at 3Mbit/sec

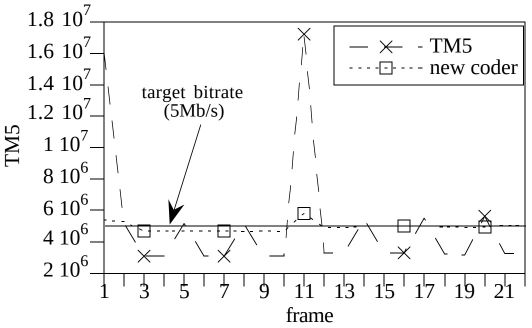

Figure 6. Output bit-rate at 5Mbit/sec (change of scene at frame#11)

CONCLUSIONS

The proposed system solves the problem of bit allocation in MPEG-2 by means of a dynamic allocation strategy that assigns more bits to those MBs to which the human eye is more reactive. Conversely, the number of bits is reduced in those MBs that are less important (from a visual-distortion point of view). In this way, the available bits are used more efficiently. Moreover, a feedback, depending on the actual output-rate, is used to redistribute the bits throughout each GOP; this fact leads to a constant visual quality.

REFERENCES

[1] ITU-T Recommendation H.261, "Video codec for audiovisual services at $p\times 64$kbit/s," December 1990–March 1993 (revised).

[2] G.K. Wallace, "The JPEG still picture compression standard," Communications of the ACM, vol. 34, No. 4, April 1991.

[3] M. DePrycker, “Asynchronous Transfer Mode,” Ellis Horwood Ltd. Ed., London, 1993

[4] N. Ohta, “Packet video,” Artech House, Inc., Norwood, MA, 1994

[5] ITU-T Recommendation H.262, also ISO/IEC 13818-2, Information Technology - generic coding of moving pictures and associated information, 1994.

[6] D. Le Gall: “MPEG: Video Compression Standard for Multimedia Application”; Communications of ACM, 34(4): 305-313, April 1991.

[7] S. Okubo, K. McCann, A. Lippmann, "MPEG-2 requirements, profiles and performance verification - Framework for developing a generic video coding standard," Signal Processing: Image Communications, 7 (1995), pp 201-209.

[8] ISO/IEC JTC1/SC29/WG11/93-400 Test Model Editing Committee, MPEG-2 Test Model 5, April 1993

[9] A. N. Netravali, B.G. Haskell, “Digital Picture Representation and Compression”, Plenum Press, New York, 1988

[10] J.J. Bae, T. Suda, “Survey of Traffic Control Schemes and Protocols in ATM Networks”, Proc. of IEEE, vol. 79, pp. 170-189, Feb. 1991