Proceedings, 2nd IEEE International Conference on Image Processing (ICIP-95), Arlington, Virginia, October 1995.

LOW BIT RATE MODEL-ASSISTED H.261-COMPATIBLE CODING OF VIDEO

Alexandros Eleftheriadis[1]

Department of Electrical Engineering
Columbia University
New York, NY 10027, USA
eleft@ctr.columbia.edu

Arnaud Jacquin

AT&T Bell Laboratories
600 Mountain Avenue, P.O. Box 636
Murray Hill, NJ 07974, USA
arnaud@research.att.com

ABSTRACT

We describe a method of object-selective quantizer control in a standard coding system based on MC DCT—CCITT Recommendation H.261. The approach is based on two novel algorithms, namely *buffer rate modulation* and *buffer size modulation*. By forcing the rate control algorithm to transfer a relatively small fraction—about 10–15% on average—of the total available bit rate from the coding of the non-facial to that of the facial area in head-and-shoulders videoteleconferencing sequences, images with better-rendered facial features are obtained; i.e. blocky artifacts in the facial area are less pronounced and eye contact and lip-sync are preserved. The improvement was found to be perceptually significant on video sequences coded at the rate of 64 kbps, with 48 kbps allocated for the input (color) video signal in QCIF format.

1. INTRODUCTION

CCITT Recommendation H.261 [1] describes an algorithm for video coding at the rates of $p \times 64$ kbps, where $p = 1, 2, \ldots, 30$. The algorithm is a hybrid of Discrete Cosine Transform (DCT) and DPCM schemes, with block-based motion estimation (ME) and compensation (MC). Although it lacks many of the coding features of algorithms designed by later standardization efforts (MPEG-1 and -2), it is widely used in ISDN-based video conferencing systems. The normative part of the standard includes the specification of the decoder only; the encoder design is not specified, and hence significant flexibility is provided to its designer.

The performance of H.261 at its lowest rate of 64 kbps suffers from a significant amount of coding artifacts, as the complexity of the task exceeds the capabilities of the algorithms used. In order to mitigate this effect, most current implementations aim at keeping a fairly "constant" coded picture quality at the expense of temporal resolution. When the sequence motion is moderate to high, temporal subsampling down to a frame rate as low as 2 fps is usually required. This in turn results in synchronization problems between audio and video (as perceived by users), and in particular between lip movement and speech (lip-sync).

[1] Work performed while the author was a UR Intern in the Signal Processing Research Department, AT&T Bell Laboratories, Murray Hill, June–August 1994.

Model-assisted coding [2] is based on feature location detection and area-selective bit allocation, and aims at enhancing the quality of perceptually important image regions (e.g. face, mouth, eyes). In this paper we present a novel model-assisted rate control framework, which maintains full H.261 decoder compatibility. A description of the feature detection algorithm used can be found in [3]. The design of H.261 for low bit rates, and the capability of its syntax to allow the specification of different quantization levels at a sufficiently fine scale (macroblock), were important factors in our selection of H.261 as a base for the implementation of a model-assisted coding system.

The organization of the paper is as follows. In Section 2 we briefly describe the structure of the RM8 H.261 encoder [4], which is used as a reference design for comparison purposes. In Section 3 we describe in detail the proposed model-assisted rate control framework, consisting of the concepts of *buffer rate* and *buffer size modulation*. Finally, in Section 4 we present coding results based on plain and model-assisted RM8, which verify that the proposed algorithms are very effective in significantly increasing the overall perceived image quality. Although results are presented here in the context of H.261, the algorithms are applicable to any coding scheme that employs the classical buffer occupancy feedback rate control architecture.

2. REFERENCE MODEL 8

H.261 is a block-based, motion-compensated transform coding (DCT) design. It provides support for intra (I) and predicted (P) pictures, but not for bidirectionally interpolated (B) pictures. It prescribes the quantization of DCT coefficients using identical uniform quantizers with dead zones for all AC coefficients, and 8–bit uniform quantization for the DC coefficients (with a step size of 8). The AC coefficient quantizer step size is determined as twice the value of a parameter Q_p (or MQUANT, as it is referred to in the standard), which can be indicated at up to the macroblock (MB) level. A rectangular array of 11×3 MBs defines a group of blocks (GOB). The source material used in our experiments had a resolution of 180×120, resulting in a total of $2\frac{1}{3}$ GOBs per picture (frame).

Reference Model 8 [4] is a "reference implementation" of an H.261 encoder that has been used not only for the development of the H.261 standard, but has also been extensively utilized in comparative experiments in the literature. It features "variable thresholding" of DCT coefficients, MC/no-MC, intra/non-intra MB decision procedures, as well as skipping of all-zero MBs with zero motion vectors.

The rate control strategy in RM8 is as follows. The very first picture, which is an I-picture, is coded with a constant Q_p of 16. The output buffer is then set at 50% occupancy. For the remaining pictures Q_p is adapted at the start of each line of MBs within a GOB (i.e. Q_p will be adapted three times within each GOB). The buffer occupancy is examined after the transmission of each MB and, if overflow occurs, the next MB is skipped. The update of Q_p is "linear" with the buffer occupancy according to the relation:

$$Q_{p_i} = \min\left\{31, \quad \left\lfloor \frac{B_i}{B_{\max}/32} \right\rfloor + 1\right\} \tag{1}$$

where Q_{p_i} is the value of Q_p selected for MB i, B_i is the output buffer occupancy just prior to coding MB i, and $B_{\max}$ is the output buffer size. A buffer size of $6400 \times q$ bits is used, given a bit rate of $q \times 64$ kbps for the video signal only. Hence in our experiments a buffer size of 6400 bits was employed.

This formula performs reasonably well for moderate to low complexity frames, but can result in serious artifacts when highly detailed features and/or moderately high motion activity are present. In particular, forced skipped macroblocks that may be caused due to buffer overflow may result in "shearing"—an object appears split (at a macroblock boundary) and its pieces shifted, even under moderate motion. The facial area is particularly susceptible to this kind of artifacts, since its complexity can quickly drive the output buffer to overflow.

3. MODEL–ASSISTED RATE CONTROL

The basic premise of the concept of model-assisted coding [2] is to assign different "quality levels" to identifiable portions of an image that bear different perceptual significance to a viewer. Although a theoretically optimal (under specific assumptions) approach similar to [5] could be followed (with appropriate modifications to allow for a spatially weighted distortion measure), it would have the significant drawback of high complexity and high delay.

We developed an approach that maintains the sequential processing structure of RM8, i.e. macroblocks are coded in their regular left to right, top to bottom order within each GOB, and quantizer selection is based on the current buffer occupancy level. The precise spatial location of a MB, however, now plays an important role in the process.

3.1. Buffer Rate Modulation

In order to be able to spend more bits in regions of interest while staying within a prescribed bit budget (or avoiding buffer overflow), it is necessary to spend less bits on the remaining image areas. We assume that there are M regions of interest $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_M$ in an image, with corresponding areas $A_1, A_2, \cdots, A_M$. We require that the regions are non-overlapping, i.e.

$$\mathcal{A}_i \cap \mathcal{A}_j = \emptyset, \text{ when } i \neq j. \tag{2}$$

The rectangular region encompassing the whole image is denoted by $\mathcal{A}$, and its area by A. We also assume that the coding of each macroblock uses β bits on the average, when the target budget rate is R and the buffer size is $B_{\max}$. Let $\beta_1, \beta_2, \ldots, \beta_M$ denote the target average number of bits per macroblock for the coding of each of the regions of interest. Note that in general we would require $\beta_i > \beta$, i.e. an improved quality within the regions of interest. Let also $\mathcal{A}_0$ denote the portion of the image that belongs to none of the regions of interest, with a corresponding area A_0 and average number of bits per macroblock β_0. In order to satisfy the given average bit budget, the following relation must hold:

$$\sum_{i=0}^{M} \beta_i A_i = \beta A \tag{3}$$

If the parameters $\beta_1, \beta_2, \ldots, \beta_M$ are given, it follows from Eq. (3) that:

$$\beta_0 = \frac{\beta A - \sum_{i=1}^{M} \beta_i A_i}{A - \sum_{i=1}^{M} A_i} \tag{4}$$

This formula defines the "equivalent average quality" for the image region that is exterior to all objects (henceforth called exterior region for brevity), and is uniquely specified by the desired average coding quality of the coded regions and their sizes. In most cases, it is more convenient to express (4) in terms of the relative average qualities $\gamma_i \equiv \beta_i/\beta$, $i = 0, \ldots, M$:

$$\gamma_0 = \frac{A - \sum_{i=1}^{M} \gamma_i A_i}{A - \sum_{i=1}^{M} A_i} \tag{5}$$

Note that if $\gamma_i > 1$ for all $i > 0$, then $\gamma_0 < 1$, as should be expected.

Let us assume that the rate control operation of the generic (not model-assisted) encoder is governed by the function:

$$Q_{p_i} = f(B_i), \tag{6}$$

of which a particular example is Eq. (1). The function $f(\cdot)$ can be quite general, and may also depend on the input signal. It is generally assumed that it is at least an increasing function of B_i. The output buffer evolution can be described by:

$$B_i = B_{i-1} + c(i-1) - r, \tag{7}$$

where B_i is the buffer occupancy prior to coding MB i, r is the average target rate (in bits per MB), and $c(i)$ is the number of bits spent to code the i-th MB and all its immediately preceding overhead information (headers etc.). The function $c(i)$ depends on the input signal, as well as the current value of Q_p; the latter depends on the selection of the function $f(\cdot)$.

In order to convert Eqs. (6) and (7) to provide location-dependent, model-assisted operation, we introduce the concept of *buffer rate modulation*. The idea is to modulate the target rate in Eq. (7) so that more bits are spent for MBs that are inside regions of interest, and less for MBs that are not. The rate r in Eq. (7) now becomes location-dependent, given by:

$$r_i = \gamma_{\zeta(i)} r, \tag{8}$$

where the region index function $\zeta(i)$ associates the position of MB i with the region in which it belongs[1]. The buffer evolution can now be described by:

$$B_i = B_{i-1} + c_{\zeta(i-1)}(i-1) - \gamma_{\zeta(i)} r, \tag{9}$$

where the number of bits spent $c_{\zeta(i)}(i)$ is now region-dependent. Assuming stationarity for $c_k(i)$ for fixed k, and taking expectations on both sides of Eq. (9), we obtain the average rate for region k as:

$$\bar{c}_k = \gamma_k r \tag{10}$$

If the values of γ_i satisfy the budget constraint given by Eq. (3), then the total average rate per MB will be exactly r.

Obviously, an actual system will operate with a regular, un-modulated output buffer which will be emptied at the constant rate r. Consequently, both Eqs. (7) and (9) have to be tracked in order to avoid overflow or underflow (the latter is of importance only if alternate synchronization mechanisms are not available). The modulated, "virtual" buffer is used to drive the evolution of Q_p via the function $f(\cdot)$ of Eq. (6), while the actual buffer is monitored to force MB skipping in cases of overflow. When the virtual buffer overflows, no particular action is taken and Q_p is typically assigned its maximum value (depending on $f(\cdot)$).

Note that in general the γ_i do not have to be all positive. In fact, best results were consistently obtained when the γ_i of the exterior region was negative. Although a negative modulated buffer rate is somewhat surprising and counter-intuitive (as it means that the buffer is actually receiving bits from the channel instead of always transmitting them), it helps to severely constrain bit allocation in the exterior, particularly in pictures which are difficult to code.

We should note that Eq. (10) is valid for stationary processes (or, more generally, for those processes possessing the ergodic property) and in steady-state (after a large number of steps). In practice, stationarity can only be approximately assumed. In addition, since the image regions $\mathcal{A}_i$ have in general rather small dimensions, one can not expect that Eq. (9) will converge while operating inside each of the regions. Furthermore, taking into account the particular pattern with which macroblocks are scanned within each image (due to GOB-based partitioning), it is possible that only a very small number of macroblocks of each region is continuously processed at a time.

For example, consider the case when while scanning the macroblocks we move from region k to l with $\gamma_k \ll \gamma_l$. If we reached steady-state conditions in region k, then the (virtual) buffer occupancy should be high, making it difficult to quickly reach a lower value while in region l. If the portion of l that is being processed is small (1-2 macroblocks), the effect of rate modulation will not be significant.

Consequently, we see that there is a tradeoff between long-term convergence and convergence speed. The former is desirable, in order to satisfy the a priori rate allocation given by Eq. (3), while the latter is necessary in order to accomodate region segments of small size.

[1] A macroblock is considered to belong to a region if at least one of its pixels is inside that particular region.

3.2. Buffer Size Modulation

To mitigate this problem, the concept of *buffer size modulation* was introduced. It is similar to the rate modulation approach discussed above, with the difference that we are now modulating the buffer occupancy (or size[2]) instead of its rate. This is achieved by modifying Eq. (6) as follows:

$$Q_{p_i} = g(B_i, i) \equiv f\left(\frac{B_i}{\mu_{\zeta(i)}}\right) \tag{11}$$

where the μ_i's are the modulation factors for each region. In effect, this function operates in regions of low interest ($\gamma_i < 1$, $\mu_i < 1$) as if the buffer occupancy was higher that it actually is, while in regions of high interest ($\gamma_i > 1$, $\mu_i > 1$) it operates as if the buffer occupancy was lower. The result is that Q_{p_i} is "pushed" to a higher or lower value, depending on the position of the MB within the image. Since here we are interested in boosting the transient behavior, the μ_i values for each region are time-dependent, converging to the value of 1 within each step. This "fading" behavior ensures that the steady state behavior of Eq. (9) is not affected. The precise values of μ_i are empirically obtained, but we have found that they are insensitive to the sequence content and target coding rate.

Of particular importance are the implications of the above scheme regarding potential actual buffer overflows. Although there is obviously no problem in regions with $\mu \leq 1$, since the buffer occupancy is overestimated, buffer overflow in regions with high μ_i can potentially occur rather quickly. We should note, however, that when entering an object region from a lower coding quality region, the buffer occupancy is low (e.g. less than B_{max}/γ_0 on the average for the exterior). This leaves ample buffer space to absorb the rapid increase in the number of bits generated while coding blocks inside an object region.

Applying Eq. (11) to (1), we obtain:

$$Q_{p_i} = \min\left\{31, \quad \left\lfloor \frac{B_i/\mu_{\zeta(i)}}{B_{\max}/32} \right\rfloor + 1\right\} \tag{12}$$

In addition to RM8's updates of Q_p at the start of each line of MBs at each GOB, in this case Q_p is also updated for each macroblock that is inside a region with $\gamma_i > 1$.

In summary, buffer rate modulation allows one to force the rate control algorithm to spend a specified number of bits more in regions of interest, while buffer size modulation ensures that these bits are evenly distributed in the macroblocks of each region. Both techniques can be applied in general to any rate control scheme, including ones that take into account activity indicators etc. Furthermore, complete decoder compatibility is maintained as the region location information does not have to be transmitted to the decoder, and is only used in the encoder.

[2] As described here, the buffer occupancy B_i is modulated; since, however, the function $f(\cdot)$ typically involves B_i as part of the fraction $B_i/B_{\max}$, one can equivalently consider that the buffer size $B_{\max}$ is modulated.

4. CODING RESULTS

Representative frames from the sequence "roberto," coded at 48 kbps and constant frame rate of 5 fps, are shown in Figure 1. The left-hand side images show the results of RM8, while the right-hand side depict the results with model-assisted coding. Noticeable improvement is evident, particularly in the eyes and mouth areas. For example, the contour and interior of the eyes is clearly visible in the model-assisted case. Similarly, the mouth area is more clearly delineated. Significant improvement is also visible in the moving sequence. In constrast with the RM8 results, maintaining eye contact with the subjects is not hindered by severe coding artifacts.

It should be noted that in the moving sequence, other artifacts (color bleeding, mosquito effects, etc.) present in the sequences are particularly annoying, and help to (unfavorably) mask the effects of model-assisted bit allocation. To mitigate this problem, an edge-preserving smoothing post-filter [6] was used.

A perceptual quality evaluation experiment [7] involving twenty subjects, and including the sequence "roberto," resulted in a non-negligible increment of Mean Opinion Score (MOS) of half a point on a five point scale, elevating subjective audiovisual quality from less-than-fair to fair-to-good.

5. REFERENCES

[1] Draft revision of recommendation H.261: video codec for audiovisual services at $p \times 64$ kbit/s. *Signal Processing: Image Communication*, 2(2):221–239, 1990.

[2] A. Eleftheriadis, A. Jacquin, Model–Assisted Coding of Video Teleconferencing Sequences at Low Bit Rates. *Proc. ISCAS '94*, 3:177–180, May–June 1994.

[3] A. Jacquin, A. Eleftheriadis, Automatic Location Tracking of Faces and Facial Features in Video Sequences. *Proc. IWAFGR95*, June 1995.

[4] Description of Reference Model 8 (RM8). CCITT SGXV WG4, Specialists Group on Coding for Visual Telephony, Doc. 525, June, 1989.

[5] A. Ortega, K. Ramchandran, and M. Vetterli, Optimal Trellis-Based Buffered Compression and Fast Approximations. *Trans. on Image Processing*, 3(1):26–40, January 1994.

[6] J. Apostolopoulos, N.S. Jayant, Pre and Postprocessing for Very Low Bit Rate Video Compression. Private Communication.

[7] A. Jacquin, Perceptual Quality Evaluation of Low Bit Rate Model–Assisted Video. *Proc. International Symp. on Multemedia Communications and Video Coding*, October 1995.

Figure 4: Stills from sequence "roberto" coded at 48 kbps, 5 fps, with RM8 (left), and Model-Assisted RM8 (right).