EFFICIENT MODE SELECTION FOR BLOCK-BASED MOTION COMPENSATED VIDEO CODING

Thomas Wiegand, Michael Lightstone, T. George Campbell and Sanjit K. Mitra

Center for Information Processing Research
Department of Electrical and Computer Engineering
University of California, Santa Barbara 93106

ABSTRACT

A method for efficiently selecting the operating modes within a block-based multi-mode video compression system is described. For a given image region, the optimum combination of modes is selected so as to minimize the overall distortion for a given bit-rate budget. Necessary conditions for optimizing the encoder operation are derived within a rate-constrained product code framework [1]. When rate and distortion dependencies exist between adjacent blocks, the ensuing encoder complexity is surmounted using a dynamic programming strategy based on the Viterbi algorithm so as to achieve the optimum selection of macroblock modes. Results are provided for the emerging H.263 video coding standard.

1. INTRODUCTION

A key problem in high compression video coding is the operational control of the encoder. Whereas most video standards uniquely stipulate the bit-stream syntax and, in effect, the decoder operation, the exact nature of the encoder is generally left open to user specification. Ideally, the encoder should balance the quality of the decoded images with channel capacity. This problem is compounded by the fact that the most effective existing video coders utilize several modes of operation which are selected on a block-by-block basis. Specifically, in most standards the current frame is subdivided into unit regions called macroblocks that may contain, for example, a single 16×16 luminance block and two 8×8 chrominance components. As such, a given macroblock can be intra-frame coded, inter-frame coded using motion prediction, or simply replicated from the previously decoded frame. As a further complication, the resulting rate and distortion for a given macroblock are often dependent on the mode selection in adjacent macroblocks. For instance, a rate-coupling may result if the motion vectors, rather than being coded independently, are coded jointly using prediction. Likewise, overlapped block motion compensation leads to a distortion dependency between neighboring macroblocks.

In this paper, we use a rate-constrained product code framework [1] to formalize the problem of optimizing the encoder operation for a given region in the current frame of a video sequence. An associated Lagrangian formulation leads to an unconstrained cost function and, in the special case of mode selection, a non-diverging trellis whose associated paths correspond to all possible operational rate-distortion points for the specified image region. The optimal path in the trellis can be efficiently located using a dynamic programming solution based on the Viterbi algorithm [2].

For application of the mode selection strategy, we consider the emerging H.263 video coding standard [3], the scope of which is the coding of digital video at rates suitable for transmission over public switched telephone network (PSTN) lines.

2. MODE SELECTION

Currently, many block-based video compression strategies employ a multi-mode methodology to obtain more efficient coding results. For example, block-based motion compensation followed by quantization of the prediction error (inter-frame coding) is generally regarded as an efficient means for coding image sequences. On the other hand, coding a particular macroblock directly (intra-frame coding) may be more productive in situations when the block-based translational motion model breaks down. For relatively dormant regions of the video, simply copying a portion of the previously decoded frame into the current frame may be preferred. Intuitively, by allowing multiple modes of operation, we expect improved rate-distortion performance if the modes are allowed to cater to different types of scene statistics, and especially if the modes can be applied judiciously to different spatial and temporal regions of an image sequence. Consequently, in the context of multi-mode video coders two key issues need to be addressed: 1) the design of efficient modes, and 2) the means for selecting the proper mode for different portions of the video. While in this paper we directly address the latter issue, its solution provides an avenue for evaluating the usefulness of future video coding modes.

2.1. Efficient Mode Switching

Consider an image region partitioned into a group of blocks (GOB) given by $\mathcal{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_N)$. For a multi-mode video coder, each macroblock in $\mathcal{X}$ can be coded using only one of K possible modes given by the set $\mathcal{I} = \{I_1, \ldots, I_K\}$. Let $M_i \in \mathcal{I}$ be the mode selected to code block X_i. Then

This work was supported in part by the Ditze Foundation, a National Science Foundation Graduate Fellowship and in part by a University of California MICRO grant with matching supports from Hughes Aircraft, Signal Technology Inc., and Xerox Corporation.

In the Proceedings of ICIP '95, Oct. 23-26, 1995, Washington, D.C., U.S.A.

for a given GOB, the modes assigned to the elements in $\mathcal{X}$ are given by the N-tuple, $\mathcal{M} = (M_1, \ldots, M_N) \in \mathcal{I}^N$. The problem of finding the combination of modes that minimizes the distortion for a given GOB and a given rate constraint R_c can be formulated as

$$\begin{array}{c} \min\limits_{\mathcal{M}} D(\mathcal{X}, \mathcal{M}) \\ \text{subject to } \; R(\mathcal{X}, \mathcal{M}) \leq R_c. \end{array} \tag{1}$$

Here, $D(\mathcal{X}, \mathcal{M})$ and $R(\mathcal{X}, \mathcal{M})$ represent the total distortion and rate, respectively, resulting from the quantization of the GOB $\mathcal{X}$ with a particular mode combination $\mathcal{M}$. To simplify this constrained optimization problem, we can employ a rate-constrained product code framework [1]. Assuming an additive distortion measure, the cost function and rate constraint can be simultaneously decomposed into a sum of terms over the elements in $\mathcal{X}$ and rewritten using an unconstrained Lagrangian formulation so that the objective function becomes

$$\min_{\mathcal{M}} \sum_{i=1}^{N} J(\mathbf{X}_i, \mathcal{M}), \tag{2}$$

where $J(\mathbf{X}_i, \mathcal{M})$ is the Lagrangian cost function for block $\mathbf{X}_i$ and is given by

$$J(\mathbf{X}_i, \mathcal{M}) = D(\mathbf{X}_i, \mathcal{M}) + \lambda \cdot R(\mathbf{X}_i, \mathcal{M}). \tag{3}$$

It is not difficult to show that each solution to Eq. (2) for a given value of the Lagrange multiplier λ corresponds to an optimal solution to Eq. (1) for a particular value of R_c [4, 5]. Unfortunately, even with the simplified Lagrangian formulation, the solution to Eq. (2) remains rather unwieldy due to the rate and distortion dependencies manifested in the $D(\mathbf{X}_i, \mathcal{M})$ and $R(\mathbf{X}_i, \mathcal{M})$ terms. Without further assumptions, the resulting distortion and rate associated with a particular block in the GOB is inextricably coupled to the chosen modes for every other block in $\mathcal{X}$. On the other hand, for many video coding systems, constraints are often imposed that can further simplify the optimization problem.

For example, in the simplest case we can restrict the codec so that both the rate and distortion for a given image block are impacted only by the content of the current block and its respective operational mode. As a result, the rate and distortion associated with each block can be computed without consideration for the operational modes of the other macroblocks, resulting in a simplified Lagrangian given by $J(\mathbf{X}_i, \mathcal{M}) = J(\mathbf{X}_i, M_i)$. In this case, the optimization problem of Eq. (2) reduces to

$$\min_{\mathcal{M}} \sum_{i=1}^{N} J(\mathbf{X}_i, M_i) \; = \; \sum_{i=1}^{N} \min_{M_i} J(\mathbf{X}_i, M_i), \tag{4}$$

and, as a result, can be easily minimized by independently selecting the best mode for each macroblock in the GOB. The drawback is that this structural constraint is rather restrictive and leads to relatively poor rate-distortion performance. As might be expected, most video coding standards such as MPEG, MPEG-2, and H.263 are not so prohibitive in their bit-stream syntax. Typically, a block-to-block dependency is permitted such that the rate term for a given macroblock is dependent not only on the current mode but on the modes of adjacent blocks. For overlapped motion compensation (as found in H.263), the dependency manifests itself in the distortion terms as well.

For instance, consider the situation when the total influence on rate and distortion for any particular macroblock is limited to that from the immediately preceding macroblock. In other words, the rate and distortion for block $\mathbf{X}_i$ is dependent on the mode selected for both blocks $\mathbf{X}_i$ and $\mathbf{X}_{i-1}$, in which case each Lagrangian term can be written as

$$J(\mathbf{X}_i, \mathcal{M}) = J(\mathbf{X}_i, M_{i-1}, M_i). \tag{5}$$

Under this assumption, we can obtain the solution to Eq. (2) by viewing the search for the best combination of N modes in the GOB as an equivalent search for the best path in a trellis of length N. In this case, the nodes in the trellis for $i = 1, \ldots, N$, are given by the elements in $\mathcal{I}$, and the transitional costs from node M_{i-1} to node M_i are given by Lagrangian cost terms specified in Eq. (5). This trellis can be efficiently searched using the Viterbi algorithm to obtain the optimal solution to Eq. (2).

The Viterbi algorithm can also be implemented to obtain an optimal path through the trellis when the rate and distortion terms are dependent not only on the mode selected for the immediately preceding macroblock, but on the immediately ensuing macroblock as well. Assuming that the influence of the previous block can be separated from the influence of the subsequent block (which is often the case), we have

$$\begin{aligned} J(\mathbf{X}_i, \mathcal{M}) &= J(\mathbf{X}_i, M_{i-1}, M_i, M_{i+1}) \\ &= J'(\mathbf{X}_i, M_{i-1}, M_i) + J''(\mathbf{X}_i, M_i, M_{i+1}). \end{aligned} \tag{6}$$

As a consequence, the transitional cost from node M_{i-1} to node M_i is given by the sum of two terms, $J'(\mathbf{X}_i, M_{i-1}, M_i)$ and $J''(\mathbf{X}_{i-1}, M_{i-1}, M_i)$, and just as before, the optimal path can be efficiently determined using dynamic programming. Note that in our analysis, we have excluded the case of non-successive mode dependencies in order to keep the problem tractable.

A final consideration with regards to mode selection is the determination of the Lagrange multiplier λ. Recall that while the solution to the unconstrained Lagrangian cost function for any value of λ results in minimum distortion for some rate, the final rate cannot be specified a priori. Often it is desirable to find a particular value for λ so that upon optimization of Eq. (2), the resulting rate closely matches a given rate constraint R_c. Because of the monotonic relationship between λ and rate, a possible solution is the bisection search algorithm [6]. However, the computation associated with the re-optimization of Eq. (2) for numerous values of λ may preclude such a search in a practical encoder. As an alternative, we have considered a variety of successful heuristics including a frame-to-frame update of λ using least-mean-squares (LMS) adaptation.

2.2. Parameter Optimization

A problem intrinsically related to that of mode switching is the parametric optimization of the modes, themselves. Whereas in Section 2.1 we outlined an efficient procedure for determining the best macroblock modes for a given GOB, the optimization inherently assumed fixed rate-distortion behavior for each possible mode. However, for many multi-mode video coders the rate-distortion characteristics of certain modes are permitted to vary as a function of a finite set of defining parameters. In addition, the parameters, themselves, are usually restricted to a finite set of values. For example, in H.263 the quality of the intra-frame and inter-frame modes is dependent on the parameter QUANT which specifies the quantization step size for the AC transform coefficients. Specifically, this value must lie in the set $\{1, 2, \ldots, 31\}$ (corresponding to step sizes between 2 and 62), and once selected applies to all macroblocks in the current GOB[1]. As stated, the best choice for QUANT requires a full search over all allowable values because no monotonic relationship exists between the parameter and the Lagrangian cost function.

More precisely, consider a set of parameters given by $\{P_i;\ i = 1, \ldots, L\}$ which impact the rate and distortion for certain modes in $\mathcal{I}$. Furthermore, let each P_i take on values from the set $Q_i = \{1, \ldots, N_i\}$ with the restriction that each parameter must remain fixed for all macroblocks in a given GOB. Define a particular collection of these parameters by $\mathcal{P} = (P_1, \ldots, P_L)$. As such, we can modify the unconstrained Lagrangian minimization problem described by Eq. (2) to include the optimization of the parameters $\{P_i\}$ as well, resulting in

$$\min_{\mathcal{P}} \left[\min_{\mathcal{M}} \sum_{i=1}^{N} J(X_i, \mathcal{M}, \mathcal{P}) \right]. \tag{7}$$

Note that the minimization of this cost function requires an exhaustive search over all $\mathcal{P} \in Q_1 \times \cdots \times Q_L$. As an alternative, we can employ a reduced complexity multigrid descent strategy described in [1] that guarantees a locally optimal solution to Eq. (7) for a finite number of iterations. The basic idea of this approach is to hold $L-1$ of the parameters fixed and minimize the total cost function over the remaining free parameter. Once optimized, the current parameter is frozen and the process is repeated. Experimental results have shown that this strategy typically converges in just a few iterations.

3. APPLICATION TO H.263

We now consider the application of the rate-constrained mode switching algorithm described in Section 2.1 to the H.263 video coding standard [3]. The H.263 video coding standard is a descendant of the motion-compensated DCT methodology prevalent in several existing standards such as H.261, MPEG-1, and MPEG-2. As is the case with the other standards, in H.263 each frame of the image sequence is first subdivided into unit regions called macroblocks. A macroblock relates to 16 pixels by 16 lines of the luminance component and the spatially corresponding 8 pixels by 8 lines of both chrominance components. Each macroblock can be coded using any one of a number of possible modes.

The recommendation for the standard contains two picture coding types, INTRA and INTER which specify the possible macroblock modes for an entire frame. The INTRA picture type is more limiting in that it only allows intra coding for macroblocks. In this paper, we concern ourselves with the INTER picture type because within this picture type, individual macroblocks can be coded using a large variety of macroblocks modes, including intra and inter. Specific to H.263 is an additional capability called Advanced Prediction which enforces overlapped motion compensation and permits the use of four motion vectors per macroblock. For our simulations we include the following standard and optional macroblocks modes: intra (I-mode), inter with one motion vector (P-mode), inter with four motion vectors ($P4$-mode), and uncoded (U-mode) which we now briefly describe.

In the I-mode, the luminance and chrominance components are quantized using a "JPEG-like" coding scheme. In contrast, for the P-mode the current macroblock is first predicted using a single, half-pixel accurate motion vector. The resulting motion-compensated prediction error is transformed and quantized in a similar manner to the I-mode. If Advanced Prediction is set, as is the case for our simulations, overlapped motion compensation is employed. The $P4$-mode is different in that it specifies four motion vectors per macroblock. The U-mode (which is indicated by just a single bit) specifies that the current macroblock is to be represented by simply duplicating the contents of the corresponding macroblock in the previous frame.

According to the standard, "the criteria for choice of mode and transmitting a block are not subject to recommendation and may be varied dynamically as part of the coding control strategy." In what follows, we consider the application of the mode selection strategy described in Section 2.1 as an encoder control solution for the H.263 standard. Our goal is to determine the optimum mode selection for a given GOB. For all simulations, the GOB is defined as a single, horizontal macroblock stripe across a given frame. For example, a 176 × 144 QCIF-image consists of 9 macroblock stripes, each containing 11 macroblocks. We restrict ourselves to this scenario so that dependencies only arise between successive macroblocks for the purpose of employing the Viterbi algorithm.

We note that whereas, in general, the coding of a given macroblock in H.263 is influenced by the selected mode of neighboring blocks, there are two notable exceptions for this type of dependency: the I-mode and the U-mode in which the mode selection can be carried out independently of the surrounding macroblocks. For the P-mode, the rate term is dependent on three neighboring macroblocks due to the differential encoding of the motion vectors. By restricting the GOB to a horizontal macroblock stripe, we can eliminate the impact on the trellis from above and need only consider those dependencies resulting from the immediately preceding macroblock. Consequently, we can assign a transitional cost from the previous node to the current node using Eq.(5). In the case of Advanced Prediction, for both the P and $P4$-mode, rate and distortion are dependent on the previous choice for the macroblock mode, while the distortion is dependent on the succeeding macroblock mode as well. In this case, we assign the transitional costs using

[1] As an aside, we note that in some standards the bit-stream syntax does permit certain parameters to vary on a macroblock-by-macroblock basis. However, we neglect this special case because of the associated complexity required for its optimization.

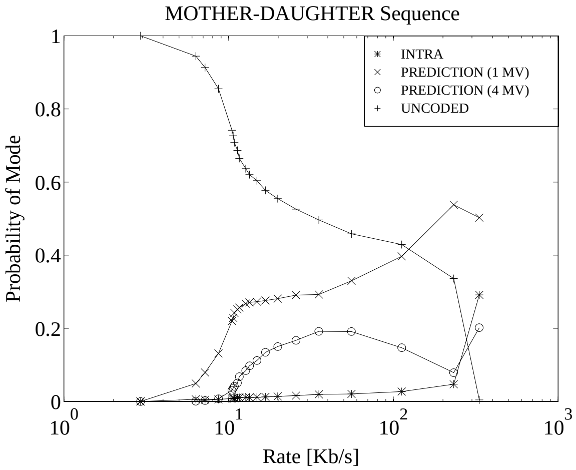

Figure 1: Probability of mode versus rate for the "Mother-Daughter" sequence.

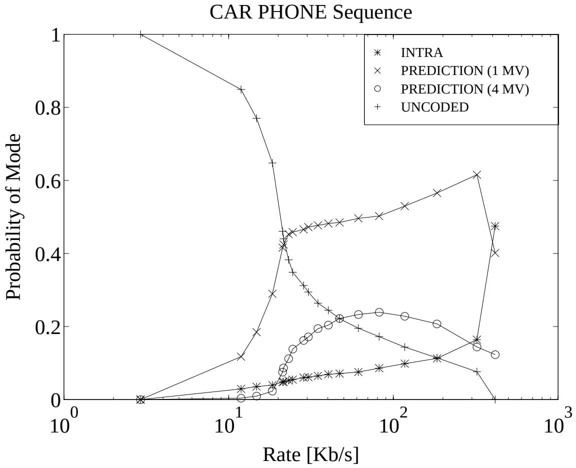

Figure 2: Probability of mode versus rate for the "Car Phone" sequence.

Eq. (6). Finally, the parameter QUANT for each GOB is optimized using the strategy outlined in Section 2.2.

4. RESULTS

Simulation results for the proposed mode switching strategy as applied to H.263 are provided in Figures 1–4. In the first experiments, the frame rate is held constant at 8.33 frames per second and λ is varied to generate coded sequences with an overall average rate from 2.9 Kbits per second (Kb/s) to 400 Kb/s. Figures 1 and 2 demonstrate the probability of selecting the I, P, $P4$, and U modes after running the algorithm on the well-known "Mother-Daughter" and "Car Phone" sequences, respectively. In the next simulations, the frame-skip is adaptive, and the LMS algorithm is used to update λ on a frame-by-frame basis in order to generate a more constant rate. Figures 3 and 4 contain sample still images from the coded sequences. For further experimental results and a more complete delineation of the algorithm as it is applied to H.263, please see [7].

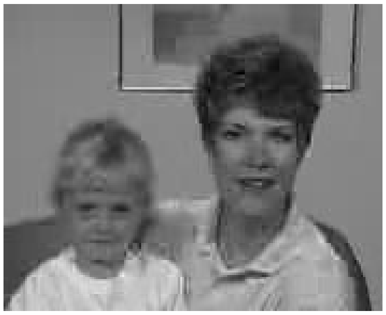

Figure 3: Frame number 188 of the "Mother-Daughter" sequence. The frame is the closest to the mean distortion of the overall coded sequence of 400 frames (0...399). The maximum rate is 12 Kb/s and the average rate is 11.0 Kb/s.

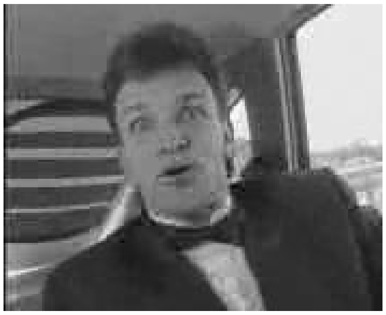

Figure 4: Frame number 66 of the "Car Phone" sequence. The frame is the closest to the mean distortion in the overall coded sequence of 350 frames (0...349). The maximum rate is 20 Kb/s and the average rate is 18.3 Kb/s.

5. REFERENCES

[1] M. Lightstone, D. Miller, and S.K. Mitra, "Entropy-constrained product code vector quantization with application to image coding", in *Proceedings of the First IEEE International Conference on Image Processing*, Austin, Texas, Nov. 1994, vol. I, pp. 623–627.

[2] G. D. Forney, "The Viterbi algorithm", *Proceedings of the IEEE*, vol. 61, pp. 268–278, Mar. 1973.

[3] ITU-T Recommendation H.263, "Video coding for narrow telecommunication channels at less than 64 kbit/s", (Draft), April 1995.

[4] H. Everett III, "Generalized lagrange multiplier method for solving problems of optimum allocation of resources", *Operations Research*, vol. 11, pp. 399–417, 1963.

[5] Y. Shoham and A. Gersho, "Efficient bit allocation for an arbitrary set of quantizers", *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 36, pp. 1445–1453, Sept. 1988.

[6] K. Ramchandran and M. Vetterli, "Best wavelet packet bases in a rate-distortion sense", *IEEE Trans. on Image Processing*, vol. 2, no. 2, pp. 160–175, Apr. 1993.

[7] T. Wiegand, M. Lightstone, T.G. Campbell, and S.K. Mitra, "A rate-constrained encoding strategy for H.263 video compression", in *Proceedings of the IEEE Symposium on Multimedia Communications and Video Coding*, New York, NY, Oct. 1995, To be published.