Rate-Distortion Optimized Mode Selection for Very Low Bit Rate Video Coding and the Emerging H.263 Standard

Thomas Wiegand, Michael Lightstone, *Member, IEEE,* Debargha Mukherjee, T. George Campbell, and Sanjit K. Mitra, *Fellow, IEEE*

Abstract— **This paper addresses the problem of encoder optimization in a macroblock-based multimode video compression system. An efficient solution is proposed in which, for a given image region, the optimum combination of macroblock modes and the associated mode parameters are jointly selected so as to minimize the overall distortion for a given bit-rate budget. Conditions for optimizing the encoder operation are derived within a rate-constrained product code framework using a Lagrangian formulation. The instantaneous rate of the encoder is controlled by a single Lagrange multiplier that makes the method amenable to mobile wireless networks with time-varying capacity. When rate and distortion dependencies are introduced between adjacent blocks (as is the case when the motion vectors are differentially encoded and/or overlapped block motion compensation is employed), the ensuing encoder complexity is surmounted using dynamic programming. Due to the generic nature of the algorithm, it can be successfully applied to the problem of encoder control in numerous video coding standards, including H.261, MPEG-1, and MPEG-2. Moreover, the strategy is especially relevant for very low bit rate coding over wireless communication channels where the low dimensionality of the images associated with these bit rates makes real-time implementation very feasible. Accordingly, in this paper, the method is successfully applied to the emerging H.263 video coding standard with excellent results at rates as low as 8.0 Kb per second. Direct comparisons with the H.263 test model, TMN5, demonstrate that gains in peak signal-to-noise ratios (PSNR) are achievable over a wide range of rates.**

I. 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 typical video sequences contain widely varying content and motion that can be more effectively quantized if different strategies are permitted to code different regions. Currently, the most effective video coders address this problem by utilizing several modes of operation which are selected on a block-by-block basis. The advantage of the multimode approach is that its inherent adaptability lays the foundation for better coding results.

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 intraframe coded, interframe coded using motion compensated 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.

Past papers on video coding have applied rate-distortion theory to improve the performance of an MPEG encoder by optimizing the frame type and/or the quantizer selection [1], [2]. One potential drawback with these approaches is that the problem of selecting the best encoding strategy for a frame is not considered at the macroblock level. Rather, the optimization is accomplished by assuming a fixed number of quantization choices for each frame. For a given number of frames, a diverging trellis is generated whose paths correspond to all possible combinations of quantization choices. The diverging trellis results because interframe dependencies over the entire group of frames are taken into account, and as such, decisions for the current frame impact all future decisions. Thus, the job of the encoder is to determine which set of quantization decisions or, equivalently, which path in the tree, has the lowest total cost in the rate-distortion sense. Unfortunately, due to the interframe dependencies, the size of the tree grows exponentially with the tree depth, and only if the number of quantization choices is relatively small can the optimal solution be feasibly found. For systems like H.263 [3], and even MPEG [4], [5], this scenario constrains the inherent multimode flexibility of the standards so as to significantly

Manuscript received July 24, 1995; revised December 11, 1995. This paper was recommended by Associate Editor H. Gharavi. 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.

T. Wiegand is with the Telecommunications Institute, University of Erlangen-Nuremberg, Cauerstr. 7/NT, 91058 Erlangen, Germany.

M. Lightstone is with Chromatic Research, Inc., Mountain View, CA 94043-4030 USA.

D. Mukherjee and S. K. Mitra are with the Center for Information Processing Research, Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106 USA.

T. G. Campbell is with Compression Labs, Inc., San Jose, CA 95134-1900 USA.

Publisher Item Identifier S 1051-8215(96)03017-0.

lessen the number of possible quantization choices for each frame.

In this paper, we employ a rate-constrained product code framework [6] to formalize the problem of optimizing the encoder operation on a macroblock-by-macroblock basis within each frame of a video sequence. An associated Lagrangian formulation leads to an unconstrained cost function and, in the special case of mode selection, a nondiverging 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 [7]. It is important to note that our objective is simply to make the best possible coding choices for the current frame given that the coding decisions for the previous frame have already been made. As a consequence, it is not possible for the method to fully exploit all interframe dependencies since the impact of current decisions on future frames is not explicitly weighed. The benefit is that this approach facilitates an intrinsically more tractable solution while simultaneously reducing the overall frame delay and permitting a larger number of parameters to vary as part of the optimization. Earlier work of ours on this subject has appeared in [8] and [9].

For application of the mode selection strategy, we consider the emerging H.263 video coding standard [3], the original scope of which has been the coding of digital video at rates suitable for transmission over public switched telephone network (PSTN) lines. Fast modems suited for this application typically run at 28.8 Kb/s within which video, audio, data, and overhead must be transmitted. This places a demanding rate constraint on the video coder which in most cases must operate at less than 20 Kb/s. In terms of wireless mobile networks whose capacities are often less than 19.2 Kb/s [10], this range of operation is also very conducive. Not surprisingly then, in addition to traditional telephony, there has been a significant and growing interest in the extension of the H.263 standard to mobile and wireless applications [11]–[13]. This circumstance further motivates the mode-selection strategy of this paper, which offers benefits in addition to excellent rate-distortion behavior, such as the ability to adjust rapidly to channels with time-varying capacity.

This paper is organized as follows. In Section II-A, we first formulate the mode selection problem as it pertains to a general block-based multimode video coding system, and then derive a solution for obtaining the best achievable performance in the rate-distortion sense. The problem of jointly optimizing the mode selection with the available mode parameters is addressed in Section II-C. Next, in Sections III and IV, a brief overview of the available modes within H.263 is provided, and results of the mode selection strategy as applied to this standard are analyzed and compared with TMN5, the current H.263 test model.

II. Mode Selection

Currently, many block-based video compression strategies employ a multimode methodology to obtain more efficient coding results. For example, block-based motion compensation followed by quantization of the prediction error (interframe coding) is generally regarded as an efficient means for coding image sequences. On the other hand, coding a particular macroblock directly (intraframe 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 multimode 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 modes proposed for future video coding systems.

A. Rate-Distortion Optimization

Consider an image region which is partitioned into a group of macroblocks (GOB) given by $\mathcal{X} = (\mathbf{X}_1, \cdots, \mathbf{X}_N)$. For a multimode video coder, each macroblock in $\mathcal{X}$ can be coded using only one of K possible modes given by the set $\mathcal{I} = \{I_1, \cdots, I_K\}$. Let $M_i \in \mathcal{I}$ be the mode selected to code macroblock $\mathbf{X}_i$. Then for a given GOB, the modes assigned to the elements in $\mathcal{X}$ are given by the N-tuple, $\mathcal{M} = (M_1, \cdots, M_N) \in \mathcal{I}^N$. 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

$$\min_{\mathcal{M}} D(\mathcal{X}, \mathcal{M}) \quad \text{subject to } R(\mathcal{X}, \mathcal{M}) \leq R_c. \tag{1}$$

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 [6]. 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 macroblock $\mathbf{X}_i$ 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 (2) for a given value of the Lagrange multiplier λ corresponds to an optimal solution to (1) for a particular value of R_c [14], [15]. Unfortunately, even with the simplified Lagrangian formulation, the solution to (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 macroblock in the GOB is inextricably coupled to the chosen modes for every other macroblock in $\mathcal{X}$. On the other hand, for many video coding systems, the bit-stream syntax imposes additional constraints 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 macroblock are impacted only by the content of the current macroblock and its respective operational mode. As a result, the rate and distortion associated with each macroblock can be computed without consideration for the operational modes of the other macroblocks, resulting in a simplified Lagrangian given by

$$J(\mathbf{X}_i, \mathcal{M}) = J(\mathbf{X}_i, M_i). \tag{4}$$

In this case, the optimization problem of (2) reduces to

$$\min_{\mathcal{M}} \sum_{i=1}^{N} J(\mathbf{X}_i, M_i) = \sum_{i=1}^{N} \min_{M_i} J(\mathbf{X}_i, M_i) \tag{5}$$

and, as a result, can be easily minimized by independently selecting the best mode for each macroblock in the GOB. For this particular scenario, the problem formulation is equivalent to the bit allocation problem for an arbitrary set of quantizers proposed earlier by Shoham and Gersho in [15], and specifically for video coding by Wu and Gersho in [16]. The drawback is that this structural constraint is rather restrictive and does not correspond to the way macroblocks are coded in most video coding standards such as H.261, MPEG-1, MPEG-2, and especially H.263. 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 macroblocks. For overlapped block 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 macroblock $\mathbf{X}_i$ is dependent on the mode selected for both macroblocks $\mathbf{X}_i$ and $\mathbf{X}_{i-1}$, in which case each Lagrangian term can be written as

$$J(\mathbf{X}_i, \mathcal{M}) = J(\mathbf{X}_i, M_{i-1}, M_i). \tag{6}$$

Under this assumption, we can obtain the solution to (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 $i = 1, \cdots, N$ 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 (6). This trellis, shown in Fig. 1(a) for $K = 4$, can be efficiently searched using the Viterbi algorithm to obtain the optimal solution to (2). We note that a similar dynamic programming solution has also been independently studied by Ortega and Ramchandran for the related problem of quantization parameter assignment in an MPEG environment [17].

Finally, 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 macroblock can be separated from the influence of the subsequent macroblock (which is often the case), we have

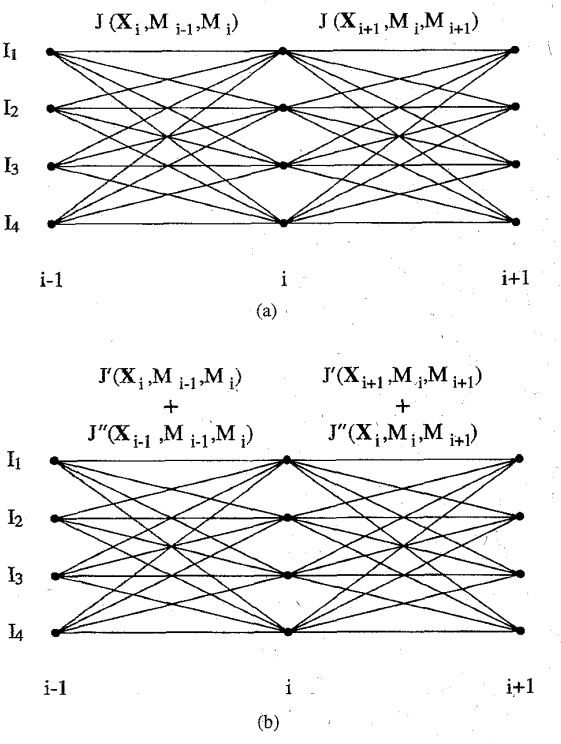


Fig. 1. Resulting multimode trellis for the cases when the rate and distortion dependencies are (a) on past macroblocks and (b) on past and future macroblocks.

$$\begin{aligned} J(\mathbf{X}_i, \mathcal{M}) &= J(\mathbf{X}_i, M_{i-1}, M_i, M_{i+1}) \\ &= J'(\mathbf{X}_i, M_{i-1}, M_i) + J''(\mathbf{X}_i, M_i, M_{i+1}). \end{aligned} \tag{7}$$

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 $J''(\mathbf{X}_{i-1}, M_{i-1}, M_i)$, which constitute the contribution from the preceding and ensuing macroblock, respectively. The corresponding trellis is described in Fig. 1(b), 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 nonsuccessive mode dependencies in order to keep the problem tractable.

B. Lagrange Multiplier Determination

A final critical consideration with regard 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 (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 described in [18] and [19]. However, this approach typically requires a variable number of iterations which introduces additional delay to the encoded bitstream if a single target rate is desired for the entire frame. This may be viewed as a disadvantage in certain environments, such as wireless networks. For example, in many implementations of H.263 for mobile applications [12], [13], feedback channels from the decoder to encoder are employed to better react to changing channel conditions, in which case the round trip delay from encoder to decoder becomes an important issue. As an alternative, we have considered a variety of potential approaches including a frame-to-frame update of λ using least-mean-squares (LMS) adaptation [20]. In some of our previous experiments (found in [8] and [9]), we have incorporated a method for determining the LMS step size dynamically for each frame or GOB (indexed by k) of the video sequence [21]. The strategy effectively reduces the bursty behavior of adaptation and results in an update procedure for the Lagrange multiplier given by

$$\begin{aligned}\lambda_{k+1} &= \lambda_k + \frac{1}{R_k^2}(R_c - R_k)R_k \\ &= \lambda_k + \left(\frac{R_c}{R_k} - 1\right).\end{aligned} \tag{8}$$

Another alternative procedure for setting λ can be found in [22] where the authors design a feedback mechanism so that λ becomes a function of the current output buffer state.

In summary, it is important to note that no matter which algorithm is utilized for selecting the Lagrange multiplier, the fine-tuning of rate is accomplished via a single parameter, λ, with the desirable outcome that—no matter what bit rate results—the distortion of the GOB will be minimum for that rate. This is in striking contrast to other encoder strategies that typically scale a single parameter such as the quantizer step size to control the instantaneous rate, but cannot guarantee any type of optimal rate-distortion performance.

C. Parameter Optimization

A problem intrinsically related to that of mode switching is the parametric optimization of the modes themselves. Whereas in Section II-A 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 multimode 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 intraframe and interframe 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 $\{1, 2, \cdots, 31\}$ (corresponding to step sizes between two 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 $\{P_i;\ i = 1, \cdots, L\}$ which impact the rate and distortion for certain modes in $\mathcal{I}$. Furthermore, let each P_i take on values from the set $Q_i = \{1, \cdots, N_i\}$ with the restriction that each parameter must remain fixed for all macroblocks in a given GOB. Define a particular collection of these parameters by $\mathcal{P} = (P_1, \cdots, P_L)$. As such, we can modify the unconstrained Lagrangian minimization problem described by (2) to include the optimization of the parameters $\{P_i\}$ as well, resulting in

$$\min_{\mathcal{P}} \left[\min_{\mathcal{M}} \sum_{i=1}^{N} J(\mathbf{X}_i, \mathcal{M}, \mathcal{P}) \right]. \tag{9}$$

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 [6] that guarantees a locally optimal solution to (9) 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.

III. Application to H.263

We now consider the application of the rate-constrained mode switching algorithm described in Section II-A to H.263, the International Telecommunication Union's (ITU) draft recommendation for video coding over narrow telecommunications channels [3].

A. The Modes of H.263

The H.263 video coding standard is a descendant of the motion-compensated discrete cosine transform (DCT) methodology prevalent in several existing standards such as H.261 [23], MPEG-1 [4], and MPEG-2 [5]. Together, their primary applications span the gamut from low bit-rate video telephony to high quality high definition television (HDTV) with H.263 focusing (at least initially) on the low bit-rate end. 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. As shown in Fig. 2, a macroblock relates to 16 pixels by 16 lines of the luminance component (Y) and the spatially corresponding 8 pixels by 8 lines of both chrominance components (C_B and C_R). As part of H.263, each macroblock can also be coded using any one of several possible modes, the allowable set of which is determined by the picture coding type.

The recommendation for the standard contains two picture coding types, INTRA and INTER, which specify the possible macroblock modes that may be used for the current frame. The INTRA picture type is more limiting in that it only allows intra coding for macroblocks. It is typically used only for special purposes, e.g., coding the first frame of a video sequence. In

[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.

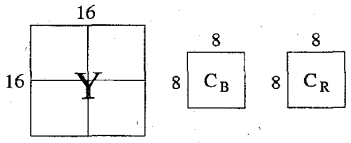


Fig. 2. Macroblock separation.

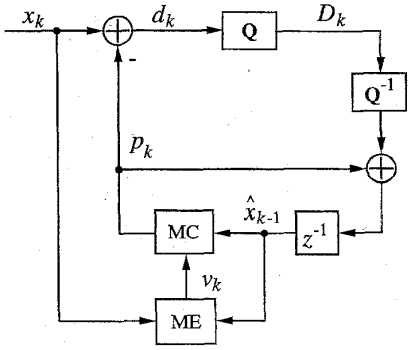


Fig. 3. Prediction loop. The motion vectors v_k which are estimated (ME) using the current original frame x_k and the previous decoded frame $\hat{x}_{k-1}$ are variable length coded using a fixed coding table. Then, v_k and $\hat{x}_{k-1}$ are used to predict the frame p_k by the motion compensation algorithm (MC) and subtracted from the current original frame. The difference image d_k is DCT transformed and quantized (Q) into D_k which is variable length encoded.

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 macroblock modes, including intra and inter. Specific to H.263 is an additional capability called advanced prediction which enforces overlapped block motion compensation and permits the use of four motion vectors per macroblock. This function can be set by a single bit and impacts the macroblock modes for an entire frame. 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. The components are initially segmented into 8 × 8 blocks which are subsequently transformed by the DCT. All ac transform coefficients are then identically scalar quantized with an even step size value ranging from two to 62. Next, the coefficients are "zig-zag" scanned and losslessly encoded using a lookup table that exploits long runs of zeros. Special attention is paid to the quantization of the dc transform coefficient as it is uniformly scalar quantized using an eight bit codeword. Typically, the quantizer step size is fixed for all macroblocks in a GOB. However, as part of the H.263 standard, the encoder can set a two-bit option in the macroblock header which permits a change in the quantizer step-size of ±1 or ±2 for all succeeding macroblocks. As we already mentioned in Section II-C, this type of macroblock-by-macroblock parameter adjustment is not considered for now due to the associated complexity required for its optimization, though in principle, it is not a fundamental obstacle.

In the P-mode, the current macroblock is first predicted using a single, half-pixel accurate motion vector. Each motion

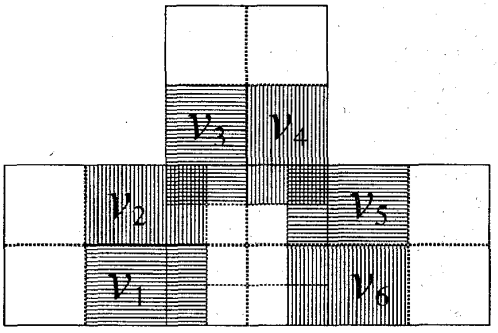
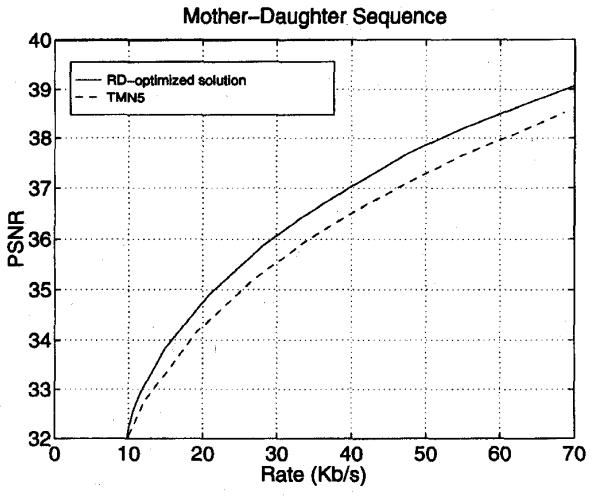


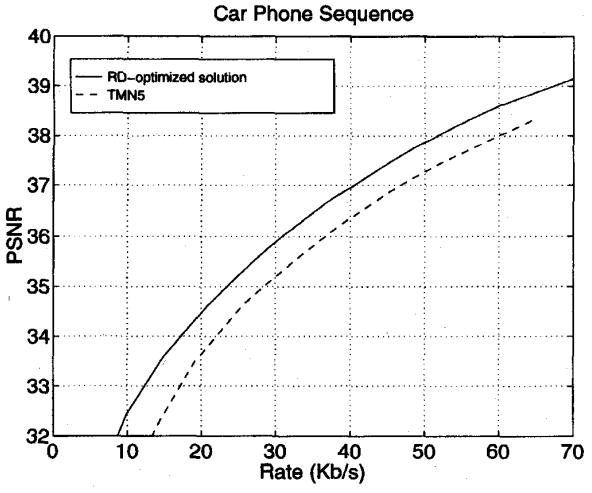
Fig. 4. Overlapped motion compensation. The lapping into the current macroblock is performed by a weighted superposition of the predicted current macroblock and the surrounding macroblocks with a depth of four pixels. Only the luminance component is affected by the overlapped block motion compensation.

vector points to a 16 × 16 luminance region and two 8 × 8 chrominance regions in the previously decoded frame within a horizontal and vertical range of −16 to +15.5 pixels. Once determined, the motion vectors are differentially encoded after each vector is first predicted using the median of three candidate vectors. The candidate vectors correspond to the three surrounding motion vectors located directly above, above and to the right, and directly left of the current motion vector, respectively. Each motion-error term is encoded without loss using a single variable-length codeword from a fixed look-up table. Next, the resulting motion-compensated prediction error is transformed and quantized in the same manner as the I-mode, with the exception that the dc coefficient is not treated separately. The incremental modification of the quantizer step size for individual macroblocks, while allowed by the H.263 standard, is not considered in this paper. A block diagram summarizing the basic operation the P-mode is provided in Fig. 3.

When advanced prediction is turned off, both the I- and P-modes act very similarly as in past standards such as H.261 and MPEG. In contrast, when the advanced prediction bit is set, the P-mode is modified to include overlapped block motion compensation [24], [25]. Moreover, by flipping this bit, an additional macroblock mode can be utilized that not only includes overlapped motion compensation, but also specifies four motion vectors per macroblock. In this mode, which we refer to as the $P4$-mode, the macroblock is segmented into four smaller 8 × 8 blocks, each compensated by one of the four specified motion vectors in the same manner that the larger 16 × 16 blocks are compensated in the P-mode. An important point is that the $P4$-mode must be used in conjunction with another special functionality of H.263, called the unrestricted motion vector mode, in order to allow the lapping of pixels located outside the frame boundaries. This function is similarly set by a single bit for an entire frame and is defined such that the pixels from the border of the picture are copied to the regions outside. The lapping from the outside into the current macroblock is depicted in Fig. 4. The vectors $\{\mathbf{v}_1 \cdots \mathbf{v}_6\}$ are the motion vectors from the neighboring macroblocks, and the lapping is performed using fixed weighting windows. Within a macroblock, each of the four smaller luminance blocks is

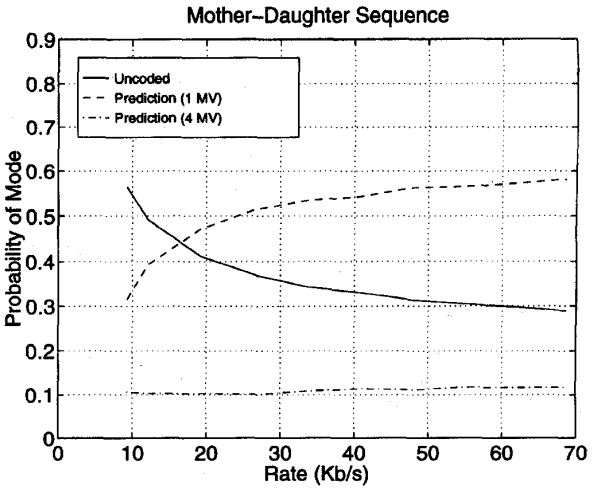


(a)

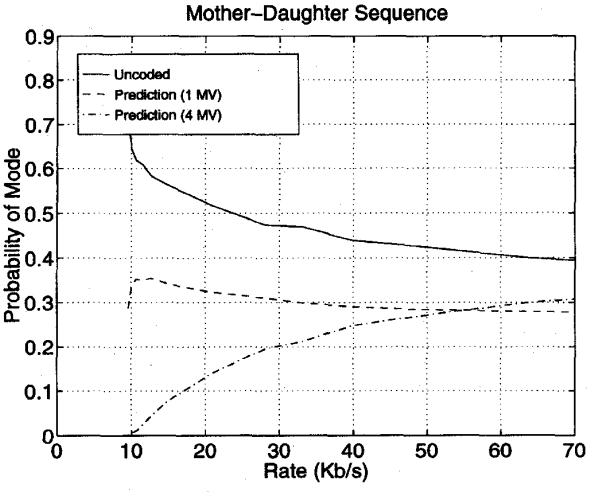


(b)

Fig. 5. Comparison in coding performance between TMN5 and the proposed encoding strategy using rate-distortion optimization. Plots compare average rate versus average PSNR for the first 150 frames of the (a) "Mother-Daughter" and (b) "Car Phone" video sequences. Note: the frame skip is held constant at two for a frame rate of 8.33 frames per second.



(a)



(b)

Fig. 6. Probability of mode versus average rate for the first 150 frames of the "Mother-Daughter" sequence. Results shown are for (a) TMN5 and (b) the proposed encoding strategy using rate-distortion optimization.

similarly predicted by internally applying overlapped block motion compensation between the blocks. The exact procedure for differentially encoding the four motion vectors is detailed in the recommendation. Otherwise, the same prediction loop as depicted in Fig. 3 is applied, and the quantization is performed as explained for the P-mode.

The uncoded mode (U-mode) (which is indicated by just a single bit for a given macroblock) specifies that the current macroblock is to be represented by simply duplicating the contents of the corresponding macroblock in the previous frame.

B. Mode Switching in H.263

According to the standard [3], "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 II-A 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 quarter-common-intermediate format- (QCIF)-image consists of nine 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. This approach also lends itself to wireless scenarios in that the generation of GOB's on a regular interval facilitates the recovery from bit errors which are more likely in the wireless environment.

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. Because there is no transitional cost between modes, the costs for these nodes can be assigned using (4). 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 (6).

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. Using (7), we can compute the cost for the incoming and outgoing transitions of the current node assigned for the P and $P4$-modes as follows. As described in Fig. 4, the distortion of the left half of the macroblock is only influenced by the motion vectors of the macroblock to the left and from the above. The macroblocks' modes from above are fixed because they are determined in the previous GOB, and thus, we need only consider the influence from the left when computing the distortion component of $J'(\cdot)$ in (7). Analogously, all distortion influences except those from the right can be eliminated when computing the distortion component in $J''(\cdot)$. Likewise, the distortion for both chrominance components is equally distributed to the in and outgoing transitions. In terms of rate, the cost assignment to the trellis branches is slightly more complicated because the motion vectors on the right half of the $P4$-mode are predicted from the motion vectors to the left. Consequently, a dynamic update for $J'(\cdot)$ and $J''(\cdot)$ based on the decisions for the incoming transitions is required. Finally, the quantizer step size parameter, QUANT, is optimized using the strategy outlined in Section II-C for each GOB.

IV. Coding Results

Simulation results for the proposed mode switching strategy are provided in Figs. 5–7 for the H.263 video coding standard. For these experiments, the frame rate is held constant at 8.33 frames per second and the Lagrange multiplier λ is varied to generate coded sequences with an overall average rate from roughly 8 Kb/s to 64 Kb/s. As part of the encoding process, both the mode and the quantizer step-size are selected using the procedures outlined in Section II so as to optimize (9) for each macroblock slice. For a frame of reference, these coding results have been compared with coded sequences generated by TMN5, the video codec test model for the H.263 standard. For fairness, both video encoders employ the same negotiable options, namely the unrestricted motion vector mode and advanced prediction. In addition, both methods are constrained to encode the same frames from each video sequence.

Empirically, we have found the proposed mode selection strategy using rate-distortion optimization to outperform

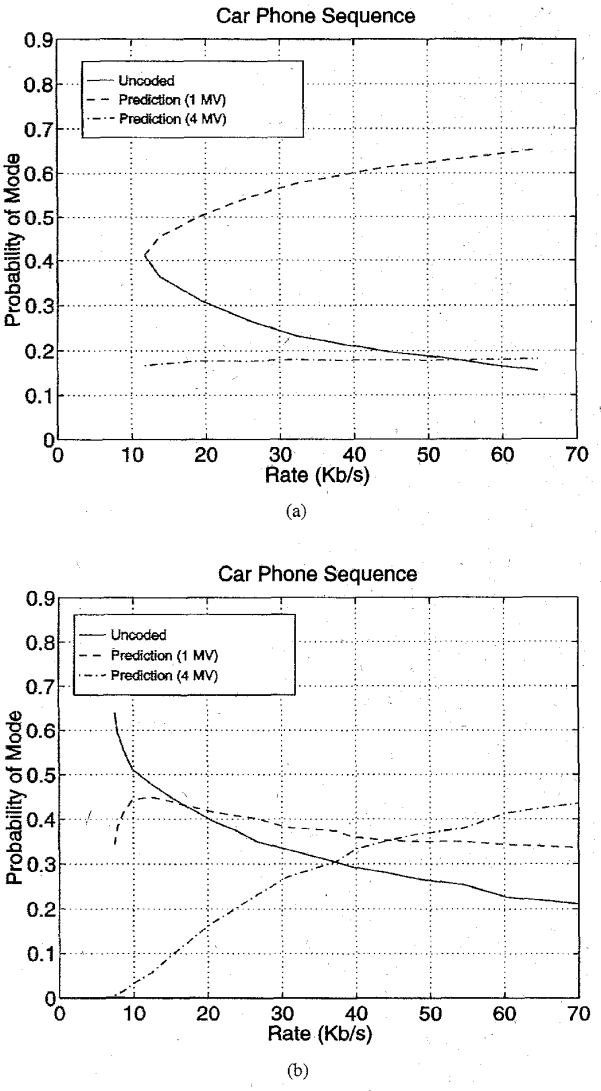


Fig. 7. Probability of mode versus average rate for the first 150 frames of the "Car Phone" sequence. Results shown are for (a) TMN5 and (b) the proposed encoding strategy using rate-distortion optimization.

TMN5 for all test sequences and all rates considered. In some cases, the gains are reasonably significant as compared to TMN5 with improvement up to 1.2 dB in peak signal-noise-ratio (PSNR) observed for a given bit rate. Fig. 5(a) and (b) summarize this performance for the well-known "Carphone" and "Mother-Daughter" sequences, respectively. In these plots, the PSNR is computed from the average distortion contribution for all six 8×8 DCT blocks in each macroblock of the video sequence. Since four of the six blocks in a macroblock correspond to the luminance component and two of the six correspond to the chrominance components, this strategy, in effect, weighs the luminance component by two thirds and the each chrominance component by a sixth. It is entirely possible (though it is not examined here) that other scalar weights may lead to a more perceptually valid distortion measure. In any case, the gains in PSNR confirm what was claimed earlier,

i.e., that a single parameter λ can simultaneously control the instantaneous bit rate and generate excellent performance over a wide range of average rates. Unlike other potential rate-controlling parameters such as the quantizer step size, the method guarantees that no matter what value of λ is selected, the distortion of each GOB is minimum for the resulting rate. Further experimental evaluations regarding the proposed mode switching strategy can be found in [8] and [9].

Though fixing λ for the video sequence does not represent an entirely practical implementation since the maximum instantaneous rate is not constrained, it does provide a means for assessing the relative importance of each mode at different bit rates. For example, Figs. 6 and 7 demonstrate the probability of selecting the P, $P4$, and U modes after encoding the "Mother-Daughter" and "Car Phone" sequences using both TMN5 and the proposed mode selection strategy.[2] Upon close examination, several intuitively appealing aspects of the proposed encoder are confirmed by the plots. For instance, the probability of the U-mode, as expected, tends toward zero at high rates for both sequences. Though not shown here, for $\lambda = 0$, the probability, in fact, becomes exactly zero. In contrast, the more accurate, but also more expensive (in terms of rate) $P4$ mode is chosen with increasing frequency as the rate increases. In between the two extremes is the P mode, which is initially selected more often as rate increases, but begins to taper off after 13 Kb/s as the $P4$ mode begins to pick up momentum.

Finally, it is interesting to note that the relationship between λ and distortion (for rates above 10 Kb/s) is rather consistent between the sequences that we have encoded, i.e., the same value of λ corresponds roughly to the same value of PSNR in all cases. If the primary objective is a constant-distortion coder, then this is good news, implying that the Lagrange multiplier need not be substantially modified from one frame to the next. Unfortunately, the same desirable relationship does not manifest itself for rate and λ. In fact, depending on the sequence, the same value of λ may correspond to widely varying bit rates. Thus, if the goal is coding for a specified rate, which is more often the case (especially in wireless scenarios), a method for controlling the Lagrange multiplier is required.

V. Conclusion

In this paper we have presented a new method for selecting the operating modes of a block-based video coding system that optimizes (for a given GOB) overall performance in the rate-distortion sense. The strategy has been successfully implemented for the H.263 video coding standard with excellent results in terms of the fidelity of the decoded video at bit rates as low as 8 Kb/s. While the algorithm requires some additional complexity over past ad-hoc approaches (due primarily to dynamic programming) that may preclude its usefulness in certain applications, there are many scenarios where the added complexity may not be an issue. For example, in very low bit rate video coding applications (<24 Kb/s), the dimensionality of the image frames is often substantially less than other applications (176 × 144 for a QCIF image in H.263), and as a result, the additional memory and complexity of dynamic programming are much less of an issue. Another potential area conducive to mode-switching is the storage of video onto CD-ROM, in which case the encoding process is performed only once, and off-line. For these cases, the additional encoding complexity is generally not a factor as long as the quality of the decoded images can be improved.

In general, our method provides a means for upper-bounding the achievable performance of various video standards such as H.261 and MPEG, and consequently, can be used to measure the capabilities of existing, heuristically-designed approaches. For instance, it may be useful to know that a particular method is already operating "close enough" to the best possible performance so that no modifications are necessary. Furthermore, using the rate-distortion optimized multimode encoding strategy, it is possible to measure the utility of proposed or optional operating modes, such as advanced prediction in H.263. By implementing the algorithm both with and without a given mode, it becomes very straightforward to assess its relative value. In this sense, existing standards can be streamlined by eliminating modes of operation that are shown to be superfluous. This type of analysis may be beneficial, in general, or for particular classes of image sequences.

[2]The probability of selecting the I mode is not shown since in both sequences it is chosen less than 2% of the time at rates below 100 Kb/s.

Acknowledgment

The authors are grateful for the helpful comments of J. D. Kim. They would also like to thank D. Miller, E. Steinbach, and B. Girod for useful discussions.

References

[1] K. Ramchandran, A. Ortega, and M. Vetterli, "Bit allocation for dependent quantization with applications to multiresolution and MPEG video coders," *IEEE Trans. Image Processing*, vol. 3, no. 5, pp. 533–545, Sept. 1994.

[2] J. Lee and B. W. Dickinson, "Joint optimization of frame type selection and bit allocation for MPEG video coders," in *Proc. ICIP*, 1994, vol. 2, pp. 962–966.

[3] ITU-T Recommendation H.263, "Video coding for low bitrate communication," Dec. 1995.

[4] ISO/IEC 11172-2, "Information technology-coding of moving picture and associated audio for digital storage media at up to about 1.5 mbit/s: Part 2 Video," Aug. 1993.

[5] ITU-T Recommendation H.262—SO/IEC 13818-2, "Information technology-generic coding of moving picture and associated audio for digital storage media at up to about 1.5 mbit/s: Video," Draft, Mar. 1994.

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

[7] G. D. Forney, "The Viterbi algorithm," *Proc. IEEE*, Mar. 1973, vol. 61, pp. 268–278.

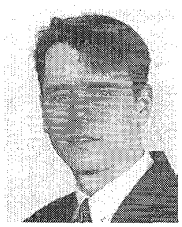
[8] T. Wiegand, M. Lightstone, T. G. Campbell, and S. K. Mitra, "A rate-constrained encoding strategy for H.263 video compression," in *Proc. of the Symp. on Multimedia Communications and Video Coding*, Polytechnic University, Brooklyn, NY, Oct. 1995, to be published.

[9] ——, "Efficient mode selection for block-based motion compensated video coding," in *Proc. of the 1995 IEEE Int. Conf. on Image Processing (ICIP '95)*, Washington, DC, vol. II, Oct. 1995, pp. 559–562.

[10] ITU-T, SG15, WP15/1, Expert's group on Very Low Bitrate Video Telephony, LBC-95-193, Delta Information Systems, "Description of mobile networks," June 1995.

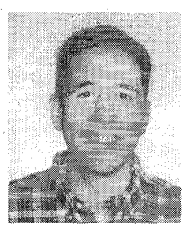
[11] ITU-T, SG15 WP15/1, LBC-95-194, R. Bosch GmbH, "Suggestions for extension of recommendation H.263 toward mobile applications," June 1995.

[12] ITU-T, SG15 WP15/1, LBC-95-267, "Robust H.263 compatible video transmission for mobile applications," University of Erlangen-Nuremberg, Oct. 1995.
[13] ITU-T, SG15 WP15/1, LBC-95-309, "Sub-videos with retransmission and intra-refreshing in mobile/wireless environments," National Semiconductors Corporation, Oct. 1995.
[14] H. Everett III, "Generalized Lagrange multiplier method for solving problems of optimum allocation of resources," *Operations Res.*, vol. 11, pp. 399–417, 1963.
[15] Y. Shoham and A. Gersho, "Efficient bit allocation for an arbitrary set of quantizers," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 36, pp. 1445–1453, Sept. 1988.
[16] S. W. Wu and A. Gersho, "Rate-constrained optimal block-adaptive coding for digital tape recording of HDTV," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 1, no. 1, pp. 100–112, Mar. 1991.
[17] A. Ortega and K. Ramchandran, "Forward-adaptive quantization with optimal overhead cost for image and video coding with applications to MPEG video coders," in *Proc. of IS&T/SPIE, Digital Video Compression: Algorithms and Technologies*, vol. 2419, San Jose, CA, Feb. 1995, pp. 129–138.
[18] J. E. Dennis and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Englewood Cliffs, NJ: Prentice-Hall, 1983.
[19] K. Ramchandran and M. Vetterli, "Best wavelet packet bases in a rate-distortion sense," *IEEE Trans. Image Processing*, vol. 2, no. 2, pp. 160–175, Apr. 1993.
[20] S. Haykin, *Adaptive Filter Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1991.
[21] M. Rupp, "Bursting in the LMS algorithm," 1995, submitted for publication.
[22] J. Choi and D. Park, "A stable feedback control of the buffer state using the controlled Lagrange multiplier method," *IEEE Trans. Image Processing*, vol. 3, no. 5, pp. 546–558, Sept. 1994.
[23] ITU-T Recommendation H.261, "Video codec for audiovisual services at $p \times 64$ kbit/s," Dec. 1990, Mar. 1993 (revised).
[24] H. Watanabe and S. Singhal, "Windowed motion compensation," in *Proc. of the SPIE Conf. on Visual Communication and Image Processing*, 1991, vol. 1605, pp. 582–589.
[25] M. T. Orchard and G. J. Sullivan, "Overlapped block motion compensation: An estimation-theoretic approach," *IEEE Trans. Image Processing*, vol. 3, no. 5, pp. 693–699, Sept. 1994.



Thomas Wiegand received the Dipl.-Ing. in electrical engineering from the Technical University of Hamburg-Harburg, in 1995. He is currently working towards the Dr.-Ing. at the University of Erlangen-Nurmeberg.

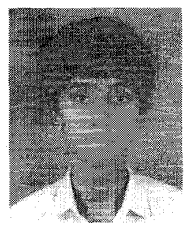
From 1993 to 1994 he was a Visiting Researcher with Kobe University, Japan, In 1995, he was a Visiting Scholar at the University of California at Santa Barbara, USA. His main research interests include information and estimation theory, digital video compression and coding, telecommunications, and multidimensional and multirate signal processing.



Michael Lightstone (S'90–M'94) received the B.S. degree in electrical engineering from the University of Illinois at Urbana-Champaign in 1990 and the M.S. and Ph.D. degrees in electrical engineering from the University of California at Santa Barbara in 1992 and 1995, respectively.

During his undergraduate studies, he interned at a number of companies including McDonnell Douglas Aircraft Company in St. Louis, MO; AT&T Bell Laboratories in Naperville, IL; and Andersen Consulting in Chicago, IL. From 1991 to 1993 he was a visiting researcher at the Tampere University of Technology in Tampere, Finland; the Advanced Video Technology Department at AT&T Bell Laboratories in Murray Hill, NJ; and the Image Processing and Analysis Group at the Jet Propulsion Laboratory in Pasadena, CA. He is currently with Chromatic Research, Inc., in Mountain View, CA. His research interests are in the areas of image and video compression and processing.

Dr. Lightstone is a member of Eta Kappa Nu and Tau Beta Pi.



Debargha Mukherjee was born in Calcutta, India, in 1970. He received the B.Tech. degree from the Indian Institute of Technology, Kaharagpur, India, in 1993. He is currently working toward the M.S. and Ph.D. degree in electrical and computer engineering from the University of California, Santa Barbara.

He was employed as a software engineer in Tata Information Systems Ltd. (an IBM and Tata company in India) between July 1993 and December 1994. He is currently a Research Assistant at the Image Processing Laboratory of University of California, Santa Barbara. His current research interests include signal compression, video compression, and pattern recognition.

T. George Campbell holds the degree of Doctor of Technology from the Tampere University of Technology, in Finland.

He has been in the field of digital video for a number of years and holds several patents in this area. He has conducted research in this area for the University of California and the Swiss Federal Institute of Technology. He has represented Finland, Switzerland, and Japan in Video Standards bodies. In Japan, he worked at the NEC, Central Research Labs. He is currently a Staff Scientist at CLI in San Jose, CA.

Sanjit K. Mitra (SM'69–F'74) received the B.Sc. (Hons.) degree in physics in 1953 from Utkal University, Cuttack, India, the M.Sc. (Tech.) degree in radio physics and electronics in 1956 from Calcutta University, the M.S. and Ph.D. degrees in electrical engineering from the University of California, Berkeley, in 1960 and 1962, respectively.

From June 1962 to June 1965, he was at Cornell University, Ithaca, NY, as an Assistant Professor of Electrical Engineering. He was with AT&T Bell Laboratories, Holmdel, NJ, from June 1965 to January 1967. He has been on the faculty at the University of California since then, first at the Davis campus and more recently at the Santa Barbara campus as a Professor of electrical and computer engineering, where he served as Chairman of the Department from July 1979 to June 1982.

Dr. Mitra served as the President of the IEEE Circuits and Systems Society in 1986. He is currently a member of the editorial boards of the *International Journal on Circuits, Systems and Signal Processing, International Journal on Multidimensional Systems and Signal Processing, Signal Processing,* and the *Journal of the Franklin Institute*. He is the recipient of the 1973 F. E. Terman Award and the 1985 AT&T Foundation Award of the American Society of Engineering Education, the Education Award of the IEEE Circuits and Systems Society in 1989 and the Distinguished Senior U.S. Scientist Award from the Alexander von Humboldt Foundation of West Germany in 1989. In May 1987, he was awarded an Honorary Doctorate of Technology degree from the Tampere University of Technology, Tampere, Finland. He is a Fellow of the AAAS and SPIE and a member of EURASIP and ASEE.