MINMAX Optimal Video Summarization

Zhu Li, *Member, IEEE*, Guido M. Schuster, *Member, IEEE*, and Aggelos K. Katsaggelos, *Fellow, IEEE*

***Abstract*—The need for video summarization originates primarily from a viewing time constraint. A shorter version of the original video sequence is desirable in a number of applications. Clearly, a shorter version is also necessary in applications where storage, communication bandwidth and/or power are limited. In this paper, our work is based on a MINMAX optimization formulation with viewing time, frame skip and bit rate constraints. New metrics for missing frame and video summary distortions are introduced. Optimal algorithm based on dynamic programming is presented along with experimental results.**

***Index Terms*—Dynamic programming, rate-distortion optimization, video analysis, video summarization.**

I. Introduction

THE DEMAND for video summarization originates from viewing time constraints, as well as communication and storage limitations, in security, military, and entertainment applications. For example, in an entertainment application, a parent may want to stream a video summary of the child's soccer game to the other part at voice data rate over the existing 2G/2.5G wireless network. In a security application, a supervisor might want to see a 2-min summary of what happened at airport gate B20, in the last 10 min. In a military situation a soldier may need to communicate tactical information utilizing video over a bandwidth-limited wireless channel, with a battery energy limited transmitter. Instead of sending all frames with severe frame signal-to-noise ratio (SNR) distortion, a better option is to transmit a subset of the frames with higher SNR quality. A video summary generator that can select frames based on an optimality criterion is essential for these applications.

There are also video skim and video retrieval results that incorporate object and semantic level information/preference in generating video skims/summaries. An example is soccer video summarization in [4] and video skim work in [29]. In this paper, we do not cover this type of problems and the formulations do not depend on extracting object level and semantic information.

The solutions to the summarization problem are typically based on a two step approach: first identifying video shots from the video sequence, and then selecting "key frames" according to some criterion from each video shot, [9], [15], [20]. A comprehensive review of past video summarization results can be found in the introduction sections of [8] and [31], and specific examples can be found in [2], [3], [6], [8], [24], [28], and [32]. Some of the main ideas and results among the previously published results are briefly discussed next.

Zhuang *et al.* [32] proposed an unsupervised clustering method. A video sequence is segmented into video shots by clustering based on color histogram features in the HSV color space. For each video shot, the frame closest to the cluster centroid is chosen as the key frame for the video shot. Notice that only one frame per shot is selected into the video summary, regardless of the duration or activity of the video shot.

Hanjalic *et al.* [8] developed a similar approach by dividing the sequence into a number of clusters, and finding the optimal clustering by cluster-validity analysis. Each cluster is then represented in the video summary by a key frame. The main idea in this paper is to remove the visual redundancy among frames.

DeMenthon *et al.* [2] proposed an interesting alternative based on curve simplification. A video sequence is viewed as a curve in a high dimensional space, and a video summary is represented by the set of control points on that curve that meets certain constraints and best represent the curve.

Doulamis *et al.* [3] also developed a two step approach according to which the sequence is first segmented into shots, or scenes, and within each shot, frames are selected to minimize the cross correlation among frames' features.

Sundaram and Chang [28] use Kolmogorov complexity as a measure of video shot complexity, and compute the video summary according to both video shot complexity and additional semantic information under a constrained optimization formulation.

For the approaches mentioned above, various visual features and their statistics have to be computed to identify video shot boundaries and determine key frames by thresholding and clustering. In general such techniques require two passes and are rather computationally involved. They do not have smooth distortion degradation within a video shot and are heuristic in nature.

Since a video summary inevitably introduces distortions at the play back time and the amount of summarization distortion is related to the "conciseness" of the summary, we formulate this problem as a rate-distortion optimization problem. Rate here can be either the temporal rate, which is the ratio of the number of frames selected in the video summary versus that in the original sequence or the actual bit rate. We assume that the summarization distortion is introduced by the missing frames. We introduce a new frame distortion metric based on principal component analysis (PCA). The sequence temporal

Manuscript received May 30, 2004; revised December 22, 2004. An initial version of this work was presented at the International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), Lisboa, Portugal, April, 2004. This paper was recommended by Associate Editor P. van Beek.

Z. Li is with the Multimedia Research Lab (MRL), Motorola Laboratories, Schaumburg, IL 60196 USA, and also with the Image and Video Processing Laboratory (IVPL), Department of Electrical and Computer Engineering, Northwestern University, Evanston, IL 60209 USA (e-mail: zhu.li@motorola.com; zli@ece.northwestern.edu)

G. M. Schuster is with the Hochschule für Technik Rapperswil (HSR), CH-8640 Rapperswil, Switzerland (e-mail: guido.schuster@hsr.ch).

A. K. Katsaggelos is with the Department of Electrical and Computer Engineering, Northwestern University, Evanston, IL 60208-3118 USA (e-mail: aggk@ece.northwestern.edu).

Digital Object Identifier 10.1109/TCSVT.2005.854230

distortion is then modeled as the maximum frame distortion introduced by the summarization, hence the name MINMAX optimal summarization.

For a given temporal rate constraint, we formulate the optimal video summary problem as finding a pre-determined number of frames that minimize the temporal distortion. On the other hand, for a given temporal distortion constraint, we formulate the problem as finding the smallest number of frames that satisfy the distortion constraint. The formulation is also extended to include bit rate constraint as well.

The paper is organized as follows. In Section II, we present the formal definitions and the rate-distortion optimization formulations of the optimal video summary generation problem. In Section III, we present our optimal video summary solution to the temporal distortion minimization formulation. In Section IV, we discuss the optimal video summary solution for the rate minimization formulation. In Section V, we present and discuss some of our experimental results for various algorithms. In Section VI, we draw conclusions and discuss future research directions.

II. MINMAX Rate-Distortion Optimization: Definitions and Formulations

A video summary is a shorter version of the original video sequence. Video summary frames are selected from the original video sequence and form a subset of it. The reconstructed video sequence is generated from the video summary by substituting the missing frames by the previous frames in the summary (zeroth-order hold). Clearly, if we can afford more frames in the video summary, or more bits to encode the summary, the distortion introduced by the missing frames will be less severe. On the other hand, more frames in the summary take longer time to view, require more bandwidth to communicate and more memory to store them. To express this tradeoff between the quality of the reconstructed sequences and the number of frames in the summary, we introduce first certain definitions and assumptions and then proceed with the problem formulations.

A. Summarization Rate and Distortion Definitions

Let a video sequence of n frames be denoted by $V = \{f_0, f_1, \ldots f_{n-1}\}$. The video sequence is either obtained in uncompressed format directly from video camera, or in the form of completely decoded sequence from compressed bit streams. Let the sequence V's video summary of m frames be $S = \{f_{l_0}, f_{l_1}, \ldots f_{l_{m-1}}\}$, in which l_k denotes the kth summary frame. The summary S is completely determined by the frame selection process $L^m = \{l_0, l_1, \ldots, l_{m-1}\}$, which has an implicit constraint that $l_0 < l_1 < \cdots < l_{m-1}$.

The reconstructed sequence $V'_S = \{f'_0, f'_1, \ldots f'_{n-1}\}$ is obtained from the summary S by substituting missing frames with the most recent frame that belongs to the summary S, that is

$$f'_k = f_{i=\max(l):\text{s.t. } l\in\{l_0,l_1,\ldots,l_{m-1}\},\quad i\le k}. \tag{1}$$

Let the distortion between two frames j and k be denoted by $d(f_j, f_k)$. Clearly, there are various ways to define the frame distortion metric $d(f_j, f_k)$ (an example will be presented in Section V). However, the optimal solutions developed in this paper are independent of the definition of this frame metric. To characterize the sequence level summarization distortion, we use the maximum frame distortion between the original sequence and its reconstruction, given by the *summarization distortion* as

$$D(S) = \max_{k\in[0,n-1]} d(f_k, f'_k). \tag{2}$$

The maximum distortion criterion is chosen instead of the average distortion criterion in this case, because it results in more uniformly distributed frame distortion, as is discussed in [27]. The maximum distortion criterion is also found to be a good metric that matches the subjective perception of the distortion. The *summarization temporal rate* of the summarization process is defined as the ratio of the number of frames selected into the video summary m, over the total number of frames, in the original sequence n that is

$$R(S) = R_T(S) = \frac{m}{n}. \tag{3}$$

Notice that the temporal rate $R_T(S)$ is in the range $(0, 1]$. In our formulation we also assume that the first frame of the sequence is always selected into the summary, *i.e.*, $l_0 = 0$. Thus, the rate $R_T(S)$ can only take values from the discrete set $\{(1/n), (2/n), \ldots, (n-1/n), 1\}$. This initial condition is necessary, because in computing the summarization distortion, we need to compute the frame distortion between f_k and the reconstructed frame f'_k, as in (1). If the first frame of the summary is some f_k, with $k > 0$, then the frame distortion $d(f_t, f'_t)$ at $t = 0, 1, \ldots, k-1$ is not defined.

For example, for the video sequence $V = \{f_0, f_1, f_2, f_3, f_4\}$ and its video summary $S = \{f_0, f_2\}$, the reconstructed sequence is given by $V'_S = \{f_0, f_0, f_2, f_2, f_2\}$, the temporal rate is equal to $R(S) = 2/5 = 0.4$, and the temporal distortion computed from (2) is equal to $D(S) = \max\{d(f_0, f_1), d(f_2, f_3), d(f_2, f_4)\}$.

Alternatively, we can use the actual number of bits to express the rate of the summary, that is

$$R(S) = R_B(S) = \sum_{t=0}^{m-1} b(f_{l_t}) \tag{4}$$

where $b(f_{l_t})$ is the number of bits required to encode the summary frame f_{l_t}. Notice that the actual number of bits needed in encoding a summary frame depends on the coding scheme and desired distortion level for the frame. We assume a bit number is assigned to each frame according to some rate profiler [10] to achieve desired, constant, or near constant peak SNR (PSNR) distortion level among summary frames. Studies show that large fluctuations in PSNR distortion among video frames are very annoying to viewers. Furthermore, for a given bit budget per frame, and a given encoding scheme (for example, motion JPEG, or H.263), the allocation of bits to achieve the desired distortion level can be performed in a rate-distortion optimal manner as described in [23] and [26].

B. Rate-Distortion Optimization Formulations

Video summarization can be viewed as a lossy compression process and a rate-distortion framework is well suited for solving this problem. Classical rate-distortion theory [1] characterizes the relation between bit rate and reconstruction distortion for information sources with known distribution

and quantization-coding scheme. However, for video sources, where the precise distribution of the source is not known and there is no closed form expression of relation between rate and distortion, operational rate-distortion (ORD) theory and schemes are used to achieve good rate-distortion performance. Examples can be found in [25]–[27].

In [17] and [18], we formulate and solve the summarization problem as an operational rate-distortion problem using the average frame distortion as the summarization distortion. However, using the maximum frame distortion as the summary distortion is more appropriate in cases where near constant perceived distortion is desired. Using the definitions introduced in the previous section, we now formulate the video summarization problem as a rate-distortion optimization problem. For a given constraint on the maximum summarization distortion $D_{\max}$, we try to minimize the summarization rate, that is,

Formulation I: Minimum rate optimal summarization (MROS)

$$S^* = \arg\min_S R(S), \text{ s.t. } D(S) \le D_{\max} \tag{5}$$

where $D(S)$ and $R(S)$ are defined in (2), and (3) or (4) respectively. The optimization is over all possible video summary frame selections $\{l_0, l_1, \ldots, l_{m-1}\}$, that contain no more than $m = nR_{\max}$ frames for the temporal rate constrained case. For the bit rate constrained case, the optimization is over both the total number of summary frames, m, and the summary frame selection, $\{l_0, l_1, \ldots, l_{m-1}\}$, such that the total bits needed to encode S does not exceed $R_{\max}$.

In addition to the rate constraint, we may also impose a constraint on the maximum number of frames that can be skipped between successive frames in the summary S. Such a constraint imposes a form of temporal smoothness and can be a useful feature in various applications, such as surveillance. Its MROS formulation can be written as

$$S^* = \arg\min_S R(S), \text{ s.t. } D(S) \le D_{\max}$$
$$l_k - l_{k-1} \le K_{\max} + 1, \ \forall\, k. \tag{6}$$

Alternatively, we can formulate the optimal summarization problem as a summarization distortion minimization problem. For a given summarization rate constraint $R_{\max}$ in either temporal rate or bit rate form, the optimal video summary is the one that MINimizes the MAXimum frame distortion (MINMAX), that is:

Formulation II: Minimum distortion optimal summarization (MDOS)

$$S^* = \arg\min_S D(S), \text{ s.t. } R(S) \le R_{\max}. \tag{7}$$

The optimization is over the summary length m, and all possible summary frame selections $\{l_0, l_1, \ldots, l_{m-1}\}$. We may also impose a skip constraint $K_{\max}$ on the MDOS formulation, as given by

$$S^* = \arg\min_S D(S), \text{ s.t. } R(S)$$
$$\le R_{\max} l_k - l_{k-1} \le K_{\max} + 1, \ \forall\, k. \tag{8}$$

The solutions to the MROS and MDOS formulations are given in Sections III and IV, respectively.

III. Solution to the MROS Formulation

For the MROS formulation in (5), if there are n frames in the original sequence, and can only have m frames in the video summary, there are $\binom{n-1}{m-1} = ((n-1)!/((m-1)!(n-m)!))$ feasible solutions, assuming the first frame is always in the summary. When n and m are large the computational cost in exhaustively evaluating all these solutions becomes prohibitive. Clearly, a more efficient solution is needed. We observe that the MROS problem has a certain built-in structure that can be exploited to find such an efficient solution.

A. Rate Minimization Recursion

The MROS problem can be solved in stages. For a given current state of the problem, future solutions are independent from past solutions. Exploiting this structure, a dynamic programming (DP) solution based on [30] and [27] is developed.

Let the summarization distortion state for the video sequence segment starting with the summary frame selection l_t and ending with the next summary frame l_{t+1} be

$$D_{l_t}^{l_{t+1}} = \left\{ \begin{array}{ll} \max\limits_{j \in [l_t, l_{t+1}]} d(f_{l_t}, f_j), & \text{if } l_{t+1} - l_t \le K_{\max} + 1 \\ \infty, & \text{else} \end{array} \right\} \tag{9}$$

which is the distortion introduced by dropped frames following frame l_t up to the next summary frame selection l_{t+1}. Notice that in (9) when the maximum frame skip constraint $K_{\max}$ is present, the distortion state of the segment is set to infinity if the skip constraint is violated. Let the rate of this sequence segment be

$$R_{l_t}^{l_{t+1}} = \left\{ \begin{array}{ll} r(f_{l_t}), & \text{if } D_{l_t}^{l_{t+1}} \le D_{\max} \\ \infty, & \text{otherwise} \end{array} \right\} \tag{10}$$

which means that if the sequence segment distortion is larger than the maximum allowable distortion, there is no feasible rate solution for the segment. If the sequence segment has an admissible distortion state, i.e., $D_{l_t}^{l_{t+1}} \le D_{\max}$, the rate of the segment is represented by the cost of including frame f_{l_t} into the summary. If the rate constraint is in the form of the temporal rate, $R_T(S)$, the cost is represented by the number of frames. If the rate constraint is in the form of the bit rate $R_B(S)$, the cost is represented as the number of bits needed to encode frames. Therefore, we have (11) shown at the bottom of the page.

$$r(f_{l_t}) = \left\{ \begin{array}{ll} 1, & \text{if counting frames} \\ b_{l_t}, & \text{if counting bits, intracoding } f_{l_t} \\ b_{l_t, l_{t-1}}, & \text{if counting bits, intercoding } f_{l_t}, \text{ with MC from } f_{l_{t-1}} \end{array} \right\} \tag{11}$$

Notice that in (11) we assume that some rate profiler (for example, [10]) can provide the bits estimate needed to encode frames to achieve certain constant PSNR quality. With this rate definition for the segment, the MROS problem in (5) is therefore equivalent to the unconstrained minimization problem of

$$\min_{l_1,l_2,\ldots l_{m-1}} \left\{ R_0^{l_1} + R_{l_1}^{l_2} + \cdots + R_{l_{m-1}}^{n} \right\}. \tag{12}$$

In (12), f_n is a virtual final frame, which does not incur any rate cost in computation, that is, $r(f_n) = 0$. The minimization is over both m and the frame selection $\{l_1, l_2, \ldots, l_{m-1}\}$. Problem (12) can be solved recursively. Let the minimum rate for the video summary segment starting with frame f_0 and ending with the summary frame choice $l_t = k$ be

$$J_{l_t=k} = \min_{l_1,l_2,\ldots l_{t-1}} \left\{ R_0^{l_1} + R_{l_1}^{l_2} + \cdots + R_{l_{t-1}}^{k} + r(f_k) \right\}. \tag{13}$$

Notice that the first and the last frame choices $l_0 = 0$ and $l_t = k$ are given in (13), therefore the minimization is over $\{l_1, l_2, \ldots, l_{t-1}\}$. Then for the video segment ending with the summary frame choice l_{t+1}, the minimum rate is given by

$$\begin{aligned} J_{l_{t+1}} &= \min_{l_1,l_2,\ldots l_t} \left\{ R_0^{l_1} + R_{l_1}^{l_2} + \cdots + R_{l_{t-1}}^{l_t} + R_{l_t}^{l_{t+1}} + r(f_{l_{t+1}}) \right\} \\ &= \min_{l_t} \left\{ R_0^{l_1} + R_{l_1}^{l_2} + \cdots + R_{l_{t-1}}^{l_t} + r(f_{l_t}) + r(f_{l_{t+1}}) \right\} \\ &= \min_{l_t} \left\{ J_{l_t} + r(f_{l_{t+1}}) \right\} \\ &= \begin{cases} \min_{l_t}\{J_{l_t} + 1\}, & \text{if counting frames} \\ \min_{l_t}\{J_{l_t} + b_{l_{t+1}}\}, & \text{if counting bits, intracoding} \\ \min_{l_t}\{J_{l_t} + b_{l_{t+1},l_t}\}, & \text{if counting bits, intercoding.} \end{cases} \end{aligned} \tag{14}$$

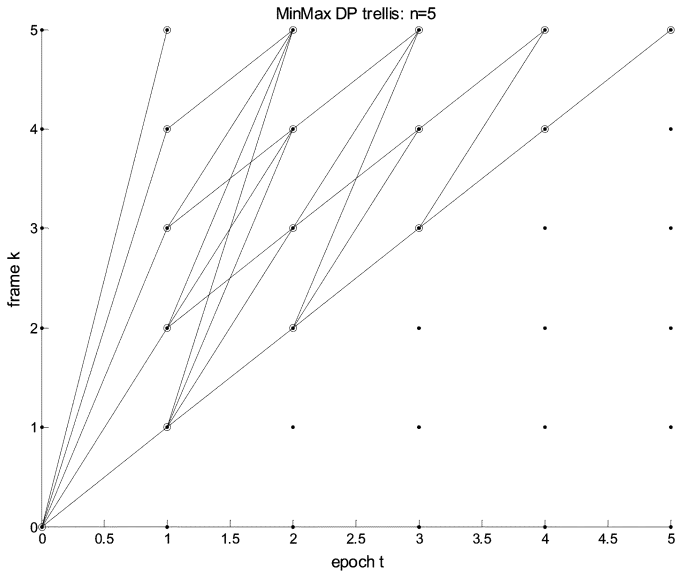
The minimization is over all feasible frame choice l_t. The initial condition is given by

$$J_{l_1} = \begin{cases} 1+1, & \text{if } D_0^{l_1} \le D_{\max}, \text{and counting frames} \\ b_0 + b_{l_1}, & \text{if } D_0^{l_1} \le D_{\max}, \text{and counting intracoding bits} \\ b_0 + b_{l_1,0}, & \text{if } D_0^{l_1} \le D_{\max}, \text{and counting intercoding bits} \\ \infty, & \text{otherwise.} \end{cases} \tag{15}$$

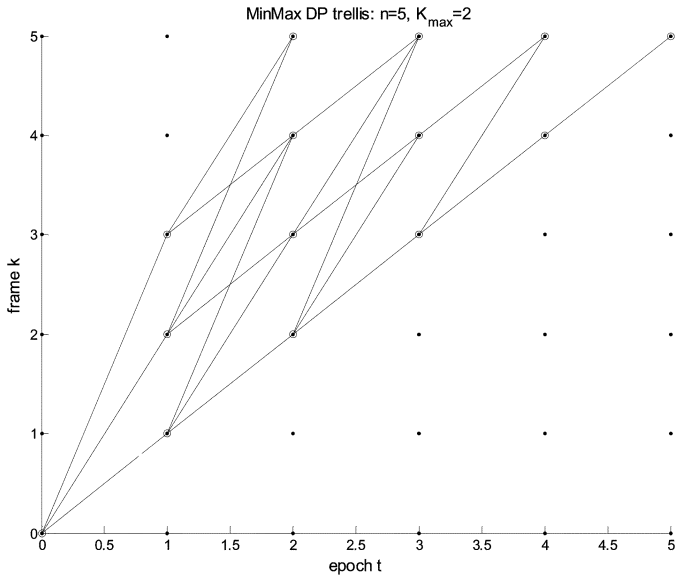
In (15) we assume that the first frame f_0 is always selected for the summary and is intracoded. Equations (14) and (15) give us the recursion we need to compute the solution trellis for a Viterbi algorithm [30] like optimal solution, which is discussed in Sections III-B.

B. DP Solution

With the minimum rate recursion developed in (14) and (15), we propose a DP solution to the MROS problem. The optimal rate recursion starts with the frame node f_0 and expands over all frames that introduce admissible segment distortions and meet the frame skip constraint. A full trellis without distortion and frame skip constraints for $n = 5$ with all possible frame transition arcs is shown in Fig. 1(a). Each arc $A(j,k)$, from frame f_j to f_k represents the rate cost of adding frame f_k to the summary ending with frame f_j in



(a). n=5, and no frame skip constraint



(b). n=5, with frame skip constraint K_{max}=2

Fig. 1. MINMAX MROS DP trellis examples. (a) $n = 5$, and no frame skip constraint. (b) $n = 5$, with frame skip constraint $K_{\max} = 2$.

$$A(j,k) = \begin{cases} 1, & \text{if } D_j^k \le D_{\max}, \text{ and counting frames} \\ b_k, & \text{if } D_j^k \le D_{\max}, \text{ and counting intracoding bits} \\ b_{k,j}, & \text{if } D_j^k \le D_{\max}, \text{ and counting intercoding bits} \\ \infty, & \text{otherwise.} \end{cases} \tag{16}$$

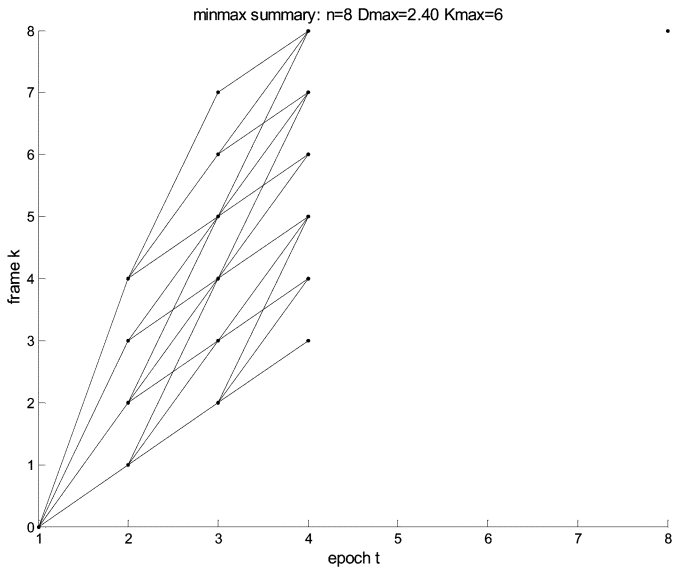


Fig. 2. MINMAX MROS solution trellis example.

It is assumed that no arc can have infinite value in the trellis. A similar trellis is shown in Fig. 1(b), in which a skip constraint with $K_{\max} = 2$ is imposed. A node $N(t,k)$ at epoch t and frame k represents the minimum rate $J_{l_t=k}$. It is computed by

$$N(t,k) = \min_{j \in l_{t-1}} \{N(t-1,j) + A(j,k)\}, \quad \forall\, A(j,k) < \infty. \tag{17}$$

The optimal incoming arc to node $N(t,k)$ in (17) is stored for future use. There is a virtual final frame f_n at each epoch (in Fig. 1, it is f_5.). The trellis expansion stops at the virtual final frame, and the arcs with transition into the virtual final frame is computed as

$$A(j,n) = \left\{ \begin{matrix} 0, & \text{if } D_j^n \le D_{\max} \\ \infty, & \text{else} \end{matrix} \right\}. \tag{18}$$

The optimal solution to the MROS problem is therefore found by selecting the virtual final frame nodes $\{N(t,n)\}$ for $t = 1, 2, \ldots, n-1$, with the minimum rate, and backtracking with the stored optimal incoming arcs for the optimal summary frame selection.

For the temporal rate based formulation, the minimum rate virtual final frame node is the one with smallest epoch t, as indicated in Fig. 2 from an example MROS summary generation for the "foreman" sequence, frames 150–157 ($n = 8$), with maximum skip constraint $K_{\max} = 6$, and maximum summary distortion constraint $D_{\max} = 2.4$. The minimum temporal rate achieved in this case is $R(S) = 3/8$.

From Fig. 2 it is clear that the optimal MROS solution is not unique. Multiple solutions like $\{f_0, f_4, f_7\}, \{f_0, f_4, f_6\}, \ldots, \{f_0, f_2, f_5\}$ are all optimal solutions to the MROS formulation. Additional constraints like the minimum coding cost in bits, and/or the minimum average frame distortion can be applied to determine the unique solution, if necessary.

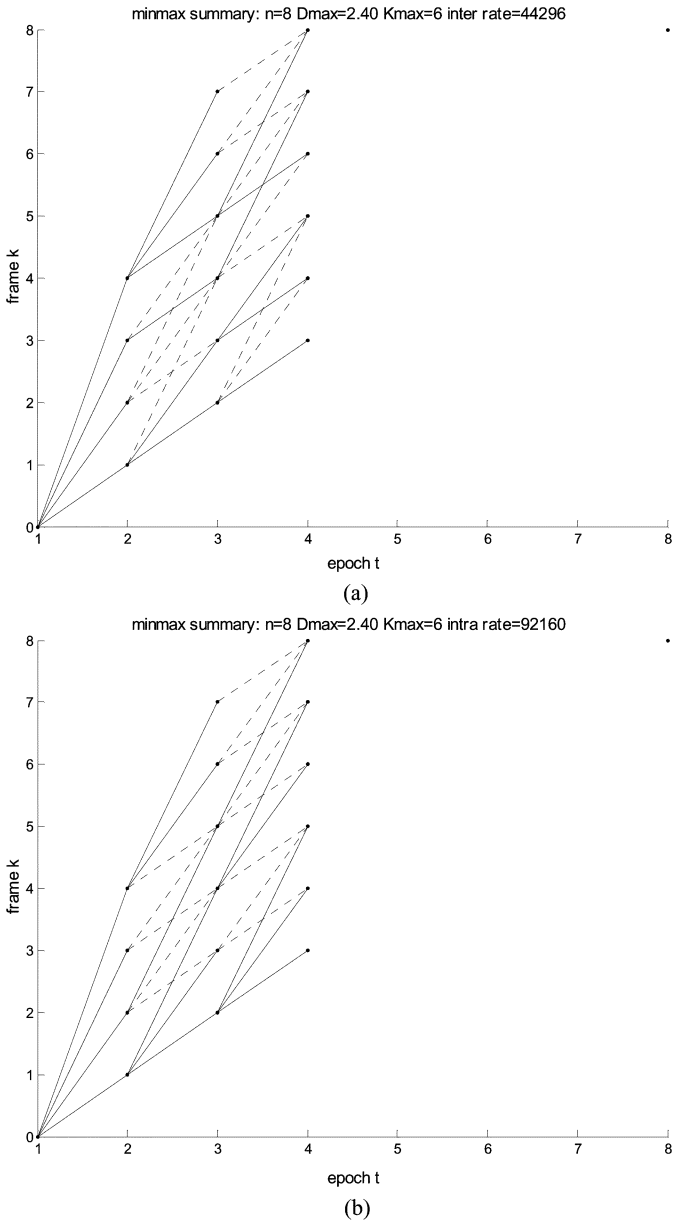


Fig. 3. MINMAX MROS solution trellis example. (a) Optimal path for the intercoded summary. (b) Optimal path for the intracoded summary.

For the bit rate based formulation, we assume that either intracoding or intercoding with an $IPPP\ldots P$ pattern has been implemented. A rate profiler finds the appropriate bit allocation b_k for intracoding frame f_k, and $b_{k,j}$ for feasible intercoding frame f_k based on prediction on frame f_j. Then the DP trellis is built recursively with (16) and (17). The optimal solution is found by identifying the virtual final frame nodes with the minimum $N(t,n)$, and backtracking by utilizing the stored arcs for the optimal MROS summary frame selection.

The example for the "foreman" sequence in Fig. 2 is shown in Fig. 3(a) for the intercoded case, and Fig. 3(b) for the intracoded case, respectively. The dotted lines represent arcs which were active for the temporal rate case but are removed for the bit rate case. Unique solutions are obtained in both cases. The solution $\{f_0, f_4, f_5\}$ in Fig. 3(a) is for the intercoded case, and the solution $\{f_0, f_2, f_5\}$ is for the intracoded case in Fig. 3(b).

C. Computational Complexity

The computational complexity in terms of the number of arc evaluations in the DP solution, for an n-frame MROS problem, with frame skip constraint $K_{\max}$, can be upper-bounded by

$$C(n, K_{\max}) \leq n + K_{\max} \sum_{k=1}^{n-1} k = \frac{(n-1)(n-2)K_{\max}}{2} + n$$

since there are total $n-k$ nodes at each epoch k, and for each node there can be at most $K_{\max}$ incoming arcs. It is clear that the DP solution has a polynomial complexity $O(n^2)$. Obviously, if we have a more stringent constraint on the frame skip (smaller $K_{\max}$), the number of arcs in the DP trellis can be reduced, as shown in Fig. 1. But this may make some lower temporal rate summarization not feasible because some potentially useful arcs are removed. If we have a larger $K_{\max}$, more arcs need to be evaluated in the DP trellis, which could potentially give a lower rate solution.

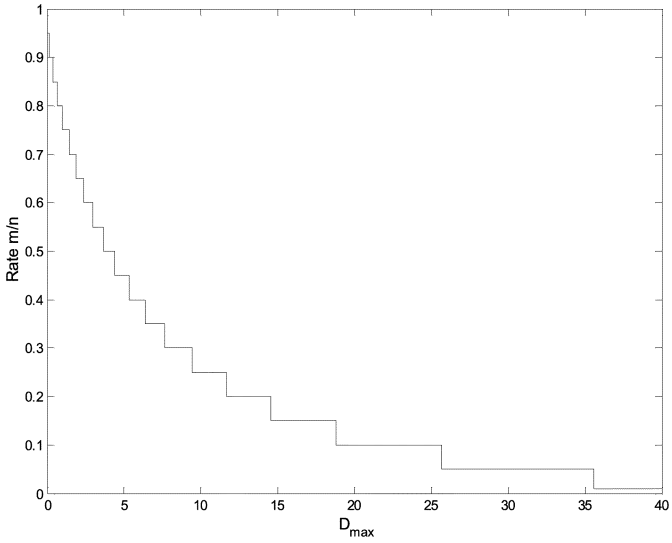


Fig. 4. Example of the operational temporal rate-distortion function.

IV. Solution of the MDOS Formulation

For the MDOS formulation, we minimize the maximum distortion of the video summary for a given rate constraint in either number of frames in the summary or bits available for summary encoding. Considering the temporal rate constrained case, for a given temporal rate constraint of $R_{\max} = m/n$, one solution is to search all $\binom{n-1}{m-1}$ possible frame selection combinations to find the one with minimum maximum distortion. Clearly, this is not practical due to the exponential increase in computational complexity with the number of frames for the problem. This approach is not practical for the bit rate constrained case either.

A solution is to consider the MROS problem and solve the MDOS problem by searching through the hull of the ORD function. The ORD function provides the achievable rate-distortion performance of a specified coding scheme. In the context of MINMAX video summarization, the ORD function is defined as

$$R^*(D_{\max}) = R(S^*), \text{ s.t. } S^* = \arg\min_S D(S) \leq D_{\max} \quad (19)$$

which is the minimum rate achievable for a given maximum distortion constraint, $D_{\max}$. The rate can be the bit rate or the temporal rate. An example of the ORD function using the temporal rate for the "foreman" sequence is shown in Fig. 4.

The ORD function is not continuous, as a range of $D_{\max}$ values can result in the same optimal rate. One important property of the ORD function is that it is nonincreasing with $D_{\max}$.

Lemma 1: $R^*(D_{\max})$ is a nonincreasing function.

Proof: For the MROS problem with distortion constraint D_1, let the optimal summary be S_1^* with frame selection $L_1^* = \{0, l_1, l_2, \ldots, l_{m-1}\}$, for some $1 < m < n$, and the resulting minimum rate be $R_1 = R(S_1^*)$. Then for a new summarization distortion constraint $D_2 > D_1$, the new optimal summary S_2^* can not have $R(S_2^*) > R(S_1^*)$. This can be proved by contradiction. Because arcs $A(l_{t-1}, l_t)$ in the solution path L_1^* are all satisfying the constraint $A(l_{t-1}, l_t) \leq D_1$, and we have $D_1 < D_2$, then we have $A(l_{t-1}, l_t)D_2$. Therefore, L_1^* is embedded in the DP trellis of the MROS problem with new distortion constraint D_2, and since $R(S_2^*)$ is the minimum rate by definition, $R(S_1^*)$ cannot be smaller than $R(S_2^*)$. This contradicts the assumption of $R(S_2^*) > R(S_1^*)$.

Lemma 1 is quite intuitive, since relaxing the distortion constraint always opens up more feasible frame transition arcs in the DP trellis, thus makeing a solution path with smaller rate possible. Since the ORD function is a nonincreasing function of the distortion threshold $D_{\max}$, the MDOS problem can be solved efficiently by a bisection search [5] on the ORD function.

For a given rate constraint of R_0, the algorithm starts with an initial maximum frame distortion bracket $[D^{\text{lo}}, D^{\text{hi}}]$ and initial rate bracket $[R^{\text{lo}}, R^{\text{hi}}]$, such that R_0 is in the initial rate bracket. Then a new distortion middle point is computed as $D_{\text{new}} = (D^{\text{lo}} + D^{\text{hi}})/2$. Solve for its optimal rate $R_{\text{new}} = R^*(D_{\text{new}})$, with the MROS DP algorithm, and find the new rate bracket by replacing either R^{lo} or R^{hi} with R_{new}, such that the rate constraint R_0 is within the new rate bracket. The distortion bracket is then replaced with the corresponding distortion pair $[D^{\text{hi}}, D^{\text{lo}}]$. The process will continue until the rate bracket boundaries converge to R_0. At this point, the search may stop, and the final MROS solution is chosen as the solution to the MDOS problem.

Since the ORD function is a piecewise constant function, we would like to avoid $D_{\max}$ values that do not result in rate change in the MROS problem, especially when the bisection search is close to the solution point. This is achieved by creating a sorted discrete set of frame distortion values given by $DV = \{d_{j,k} = d(f_j, f_k) | j < k, j = 0, 1, 2, \ldots, n-2, k = 1, 2, \ldots, n-1\}$. The bisection search is therefore performed on DV with a maximum size of $n*(n-1)$ for the nonframe skip constrained case, and $n*(K_{\max}+1)$ for the frame skip constrained case. The resulting MDOS problem solution is quite efficient with a computational complexity of $O(\log(n))$ of that of the MROS problem.

V. Experimental Results

A. Frame Distortion Metric

A perceptually meaningful frame distortion metric $d(f_j, f_k)$ with reasonable computational cost is important in our summarization effort. There are a number of ways to compute the frame distortion. Although the proposed DP solution does not depend on any specific distortion metric, we discuss in this section various choices and the one we developed and adopted in our summarization experiments.

Mean squared error (MSE) has been widely used in image processing. However, as is well known, it does not represent well the visual quality of the results. For example, a simple one-pixel translation of a frame with complex texture will result in a large MSE in the original frame size, although the perceptual distortion is negligible. There is work in the literature addressing perceptual quality issues, but such algorithms primarily address the distortion between an image and its quantized versions, not the distortion among different frames.

The color histogram-based distance is also a popular choice [31], but it may not perform well either, since it does not reflect changes in the layout and orientation of images. For example, if a large red ball is moving in a green background, even though there are a lot of "changes," the color histogram will stay relatively constant. The computational cost of generating color histogram is also high for larger frame sizes. In [16], we also experimented with a frame distortion metric that utilizes both color distance and motion activity.

For a frame distortion metric that is effective in reflecting the subjective perception of the distortion among different summary frames, we use the weighted Euclidean distance in the PC space of the scaled video frames. The video frames are first scaled into smaller sizes of 8×6, 11×9, or 16×12 pixel frames. The benefit of the scaling is to reduce noise and local variance such that the frame distortion is evaluated at a proper resolution. It also benefits the subsequent PCA by reducing the dimensionality of the data. The number of sample frames available for PCA is always limited and the reduced dimensionality makes the covariance matrix estimation from the limited data more accurate.

The PCA transform T is found by diagonalizing the covariance matrix of the frames [14], [21], and selecting the desired number of dimensions corresponding to the largest eigenvalues. Therefore, the frame distortion metric is given by

$$d(f_j, f_k) = ||T(D(f_j)) - T(D(f_k))|| \tag{20}$$

where D denotes the scaling process, and T is the truncated PCA transform. In our experiment we randomly selected 3200 frames from various video clips and scaled the frames to 8×6 pixels before performing PCA. The resulting 48 eigenvalues are plotted in Fig. 5. Notice that most of the energy is captured by the bases corresponding to the first 6 (83% energy) to 12 (92.3% energy) largest eigenvalues. Therefore, our adopted PCA transform matrix T has dimension 6×48.

Experimental results with this frame distortion metric are shown as frame-by-frame distance plots $d(f_k, f_{k-1})$ in Fig. 6 for the "foreman" sequence in the upper plot and the "mother–daughter" sequence in the lower plot. The curves seem to reflect well the perceptual changes in the sequences.

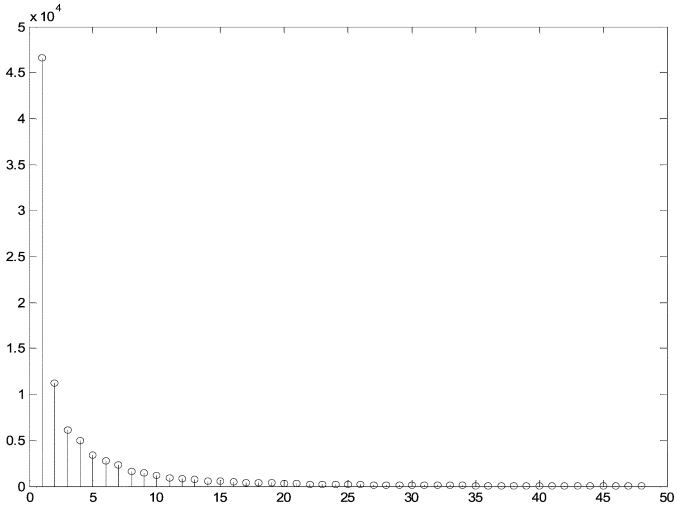


Fig. 5. Eigenvalues resulting from scaling and PCA.

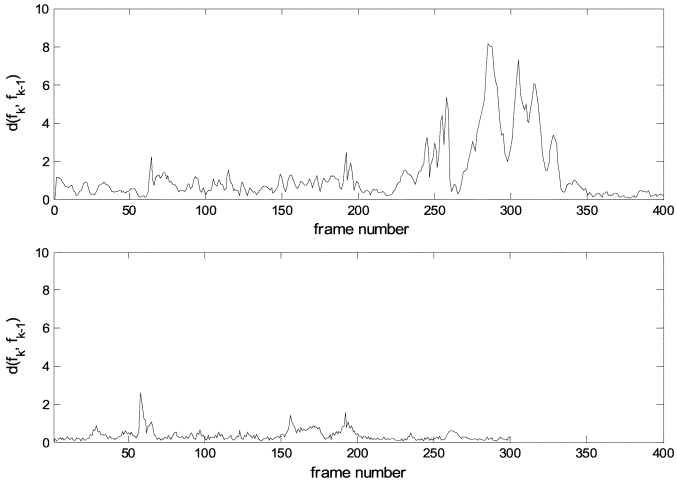


Fig. 6. Frame-by-frame distortion $d(f_k, f_{k-1})$ plot for the "foreman" and "mother–daughter" sequences.

For the "foreman" sequence, frames 1–200 contain a talking head with few visual changes, therefore the frame-by-frame distortion remains low for this period. There is a hand waving occluding the face around frames 253–259, thus we have spikes corresponding to these frames. There is the camera panning motion around frames 274–320, thus we have high values in $d(f_k, f_{k-1})$ for this time period as well. Similar observations for the frame-by-frame distortion curve can also be computed from the "mother–daughter" sequence. Notice that the "foreman" sequence is more "eventful" than the "mother–daughter" sequence and this overall activity level is also captured by the curves.

The PCA feature based frame distortion metric can be made flexible to different applications and user preference. For example, the eigen values learnt from PCA process can be applied to normalize the metric. Also some other local weighting schemes can be trained from user studies to better reflect the human perception.

From the experiment shown in Fig. 6 and other experiments with a variety of video sequences, and it seems that the metric in

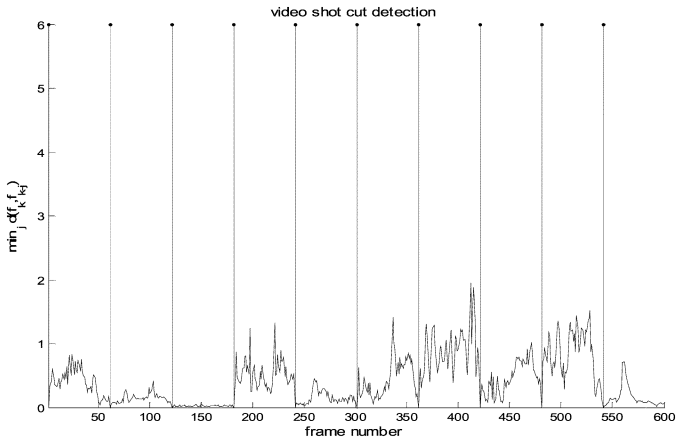


Fig. 7. Video shot cut detection by thresholding on the frame-by-frame distortion curve.

(20) is fairly accurate in depicting the distortion or the dissimilarity between frames, while at the same time keeping the computation at a moderate level for the summarization algorithm.

B. Video Shot Segmentation

Video shot segmentation [9] is a challenging problem. For the purpose of the video summarization, if video shot cut boundaries are detected, then we can break an n-frame video sequence into p video shots of n_j, (for $j = 1, 2, \ldots, p$) frames each. Then the n-frame summarization problem is transformed into multiple summarization problems of smaller sizes n_j. Since the frame distortion across the shot boundary is much higher than the MROS distortion constraint $D_{\max}$, this eliminates the frame transition arc evaluation across the shot boundaries and greatly improves the efficiency of the DP algorithm.

The frame distortion metric developed in the previous section is well suited for the shot cut detection task. For the MROS formulation, we assume that a frame f_k is at a video cut boundary if

$$d_k^{\text{cut}} = \min_{j \in [1, K_{\max}]} \{d(f_k, f_{k-j})\} > D_{\max}^{\text{cut}} \tag{21}$$

where $K_{\max}$ is the frame skip constraint, and $D_{\max}^{\text{cut}} > D_{\max}$ is the cut detection threshold. Using (21) to segment the sequence, it is guaranteed that the optimal summarization solutions obtained for each shot result in an optimal solution for the whole sequence. Clearly, the computational cost has been considerably reduced by this segmentation process, since the complexity is polynomial with respect to the video sequence length n.

A shot cut detection example is shown in Fig. 7. The cut distortion d_k^{cut} and the resulting cut detection is plotted for a mixed sequence of 600 frames with 60-frame segments from ten sequences: "fish," "coast guard," "container," "fun fair," "cubicle," "mother–daughter," "foreman," "table tennis," "toy train," and "weather forecast." The threshold is set at $D_{\max}^{\text{cut}} = 6.0$, and the skip constraint is $K_{\max} = 10$. The actual value of d_k^{cut} at the shot boundaries is truncated to $D_{\max}^{\text{cut}}$. We always assume that the first frame f_0 is at a shot cut. Clearly, the cut detection is accurate in this case.

Notice that the purpose of cut detection here is to help reduce the computational complexity by breaking a large size problem into multiple smaller size ones, while preserving the optimality of the solution. Therefore, even if cuts are missed or misclassified, it will not affect the optimality of the solution.

C. Summarization Simulation Results

We tested the proposed algorithms with a number of sequences. Some of the results are encoded into H.263 streams for subjective evaluation. In the following we report and discuss the results with a 150-frame segment of the "foreman" sequence.

For the MROS formulation with temporal rate, the video summary frame selections and resulting sequence distortions are plotted in Fig. 8. In Fig. 8(a), the results for the "foreman" sequence frames 150–299, with maximum distortion $D_{\max} = 4.0$, and no frame skip constraint are shown. The upper plot is the summary frame selection plotted as vertical lines against the dotted curve of the frame-by-frame distortion $d(f_k, f_{k-1})$, which gives an indication of the activity within the sequence.

It is clear from the plot that more frames are selected into the summary at high activity regions as expected. The bottom plot shows the summarization distortion at each frame, $d(f_k, f_k')$, between the original sequence and the reconstructed sequence from the video summary. The resulting distortions are all below the constraint $D_{\max}$, as indicated in the plots.

The results for the same sequence segment with frame skip constraint $K_{\max} = 10$ are shown in Fig. 8(b). Notice that with the skip constraint more frames ($m = 45$ versus $m = 43$) are needed to achieve the same maximum frame distortion. When the distortion threshold $D_{\max}$ and skip constraint $K_{\max}$ are relaxed, the resulting summary rate goes up, as shown in Fig. 8(c). The pattern of summary frame selection still shows concentration in high activity regions, as expected.

For the MROS problem with bit rate, the results for the same sequence are plotted in Fig. 9(a) for the intercoded case and Fig. 9(b) for the intracoded case. Comparing the results of Fig. 8(b) with that of Fig. 9(a) and (b), we notice that for the same distortion and frame skip constraints, the solution summaries to the MROS problem are different.

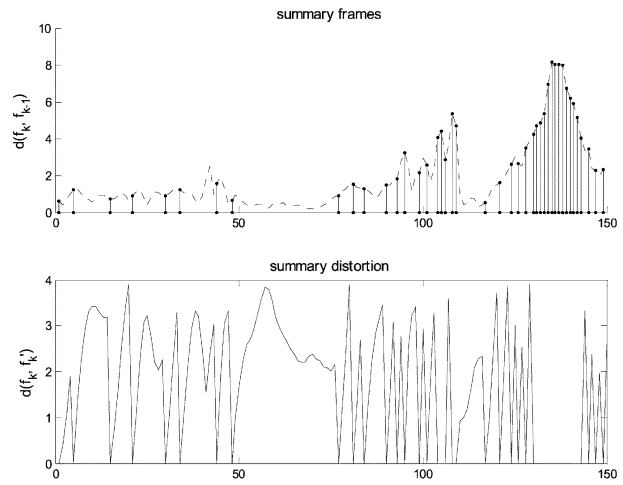
The DP algorithm achieves graceful degradation of the MROS summaries, as the distortion threshold $D_{\max}$ is relaxed. This is shown in Fig. 10 as the operational temporal rate-distortion function curves for the "foreman" sequence segment ($n = 150$, $K_{\max} = 10$) and the "mother–daughter" segment ($n = 150$, $K_{\max} = 10$). Notice that the ORD function curve is content dependent. For the same distortion the higher activity sequence ("foreman") requires larger temporal rate as expected.

For subjective evaluation of the summarization results, we generated video summaries at different distortion levels for various sequences. Their summarization distortion level $D(S)$, spatial distortion as luminance field PSNR, the number of frames m, temporal rate $R_T(S)$, and resulting bit rates $R_B(S)$ are summarized in Table I below.

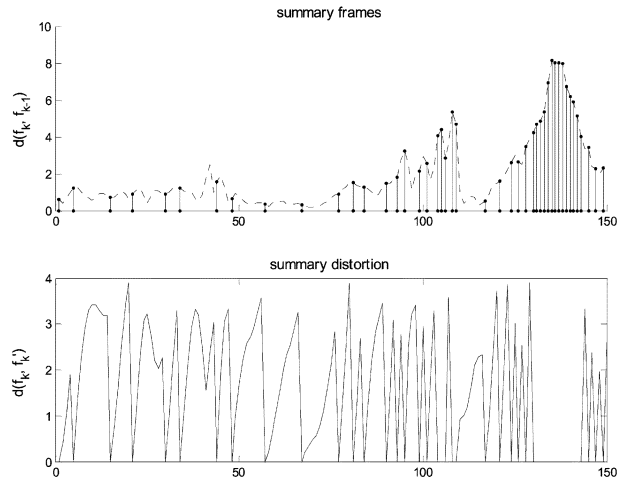
The summaries in Table I are encoded using the TMN implementation of H.263.[1]

Overall, the summaries exhibit a graceful degradation of their visual quality when $D_{\max}$ is relaxed, as shown in Table I for the "foreman," "Stefan," and "mother–daughter" sequences.

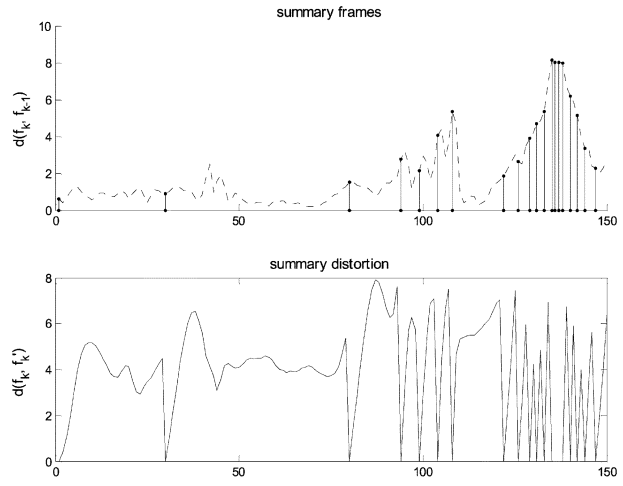
[1]The summary bit streams are available for subjective evaluation at: http://ivpl.ece.northwestern.edu/~zli/new_home/demo/minmax/minmax.html.



(a). "foreman" sequence frames 150-299, n=150, D_{max}=4, *no skip constraint*. Results: m=43, distortion: max=3.91, mean=1.77



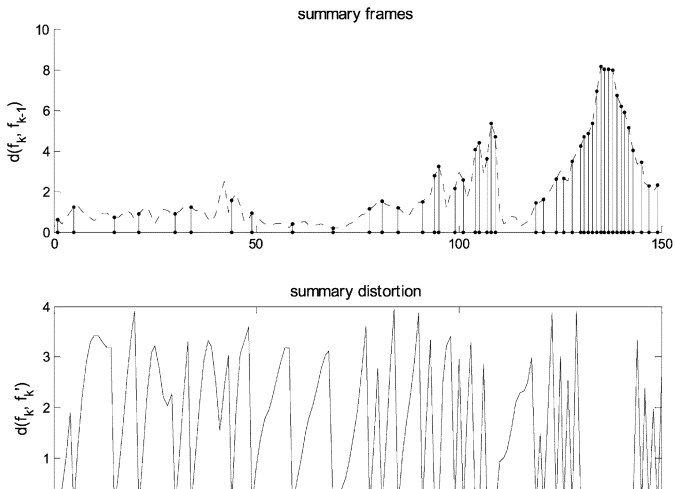
(b). "foreman" sequence frames 150-299, n=150, D_{max}=4, K_{max}=*10*. Results: m=45, distortion: max=3.91, mean=1.60



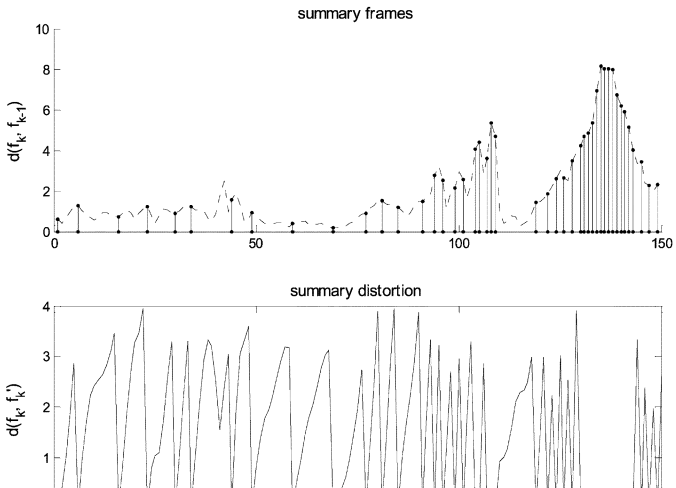
(c). "foreman" sequence frames 150-299, n=150, D_{max}=8, K_{max}=50,. Results: m=20, distortion: max=7.92, mean=4.04

Fig. 8. MROS summarization results with temporal rate. (a) "foreman" sequence frames 150–299, $n = 150$, $D_{\max} = 4$, *no skip constraint*. Results: $m = 43$, distortion: $\max = 3.91$, mean $= 1.77$. (b) "foreman" sequence frames 150–299, $n = 150$, $D_{\max} = 4$, $K_{\max} = 10$. Results: $m = 45$, distortion: $\max = 3.91$, mean $= 1.60$. (c) "foreman" sequence frames 150–299, $n = 150$, $D_{\max} = 8$, $K_{\max} = 50$,. Results: $m = 20$, distortion: $\max = 7.92$, mean $= 4.04$.

The summaries can be encoded at low bit rates for video deployment on 2G/2.5G networks.



(a). "foreman" sequence frames 150-299, n=150, D_{max}=4, K_{max}=*10*. Results: m=45, inter coding bits=467064, distortion: max=3.94, mean=1.63



(b). "foreman" sequence frames 150-299, n=150, D_{max}=4, K_{max}=*10*. Results: m=45, intra coding bits=1235136, distortion: max=3.94, mean=1.57

Fig. 9. Summarization experimental results. (a) "foreman" sequence frames 150–299, $n = 150$, $D_{\max} = 4$, $K_{\max} = 10$. Results: $m = 45$, intercoding bits $= 467\,064$, distortion: $\max = 3.94$, mean $= 1.63$. (b) "foreman" sequence frames 150–299, $n = 150$, $D_{\max} = 4$, $K_{\max} = 10$. Results: $m = 45$, intracoding bits $= 1\,235\,136$, distortion: $\max = 3.94$, mean $= 1.57$.

D. Summarization Computational Cost

In our simulation, we use uncompressed, QCIF sized YUV sequence as input. The algorithms are implemented in Matlab and running on a 2.4-GHz Pentium PC. The code is not optimized for speed. For the "foreman" sequence, starting at frame 150, we generated 4 MINMAX summaries with various numbers of frames. The associated computational cost is summarized in Table II below.

The MROS parameters are given as n, $K_{\max}$ and $D_{\max}$, while the number of frames in the resulting summaries is m. T_{dp} denotes the summarization algorithm execution time, and T_{coding} is the time spent on encoding the summaries. It takes about 0.5, 3, and 23 s to summarize 50, 100, and 200-frame sequences, respectively. Obviously, the performance can be further improved by a more efficient implementation. Notice that as the size n of the sequence doubles, the computational cost increases seven

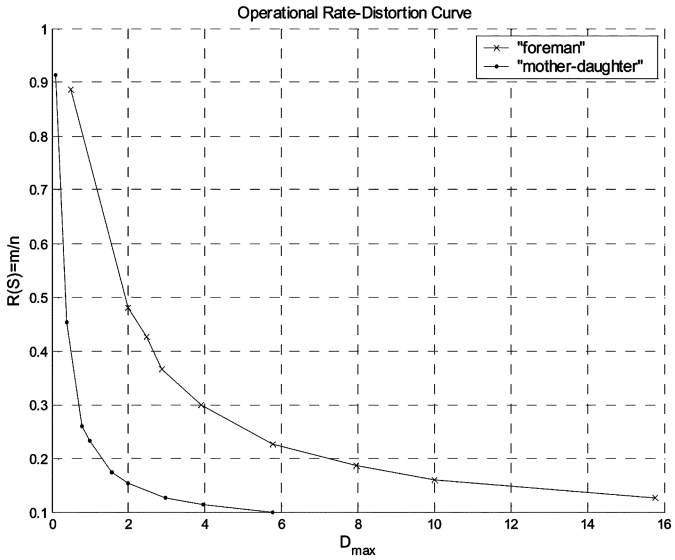


Fig. 10. Operational temporal rate-distortion curves.

TABLE I
RATES AND DISTORTION LEVELS OF MROS SUMMARIES

Sequence: [offs, n]	D(S)	PSNR_Y(luma)	m	$R_T(S)$	$R_B(S)$
"foreman": [150, 240]	4.0	27.9 dB	71	0.29	37.5 kpbs
"foreman": [150, 240]	6.0	27.8 dB	45	0.19	27.6 kpbs
"foreman": [150, 240]	8.0	27.9 dB	35	0.15	24.1 kpbs
"foreman": [150, 240]	10.0	27.7 dB	28	0.12	22.1 kpbs
"mom-daughter": [20, 240]	1.0	31.0 dB	50	0.21	7.04 kpbs
"mom-daughter": [20, 240]	2.0	31.0 dB	23	0.09	4.88 kpbs
"mom-daughter": [20, 240]	4.0	31.1 dB	14	0.05	4.03 kpbs
"mom-daughter": [20, 240]	6.0	31.1 dB	9	0.04	3.29 kpbs
"Stefan":[20, 240]	2.4	25.4 dB	33	0.14	54.0 kbps
"Stefan":[20, 240]	3.2	25.4 dB	22	0.09	39.1 kbps
"Stefan":[20, 240]	4.0	25.5 dB	15	0.06	28.8 kbps
"Stefan":[20, 240]	6.0	25.6 dB	11	0.04	21.5 kbps

TABLE II
SUMMARIZATION COMPUTATIONAL COST FOR THE "FOREMAN" SEQUENCE

(n, K_{max})	D_{max}	m	T_{dp} (sec)	T_{coding} (sec)
(50, 30)	2.0	17	0.5	0.70
(50, 30)	4.0	7	0.5	0.31
(100, 30)	2.0	28	3.2	1.16
(100,30)	4.0	15	3.3	0.61
(200,30)	2.0	83	22.7	4.62
(200,30)	4.0	47	22.8	1.97

times. This is the motivation for performing shot segmentation so that multiple smaller size problems are solved instead of solving the original large size problem in one shot.

When the computational power and/or buffer size are very limited, a greedy constrained skip algorithm similar to [16] can be applied to achieve near optimal results. The algorithm operates as follows: it starts by selecting frame f_0 into the summary, and setting the last summary frame indicator as $L = 0$. Then frames are skipped until frame f_k satisfies, $d(f_L, f_k) > D_{\max}$,

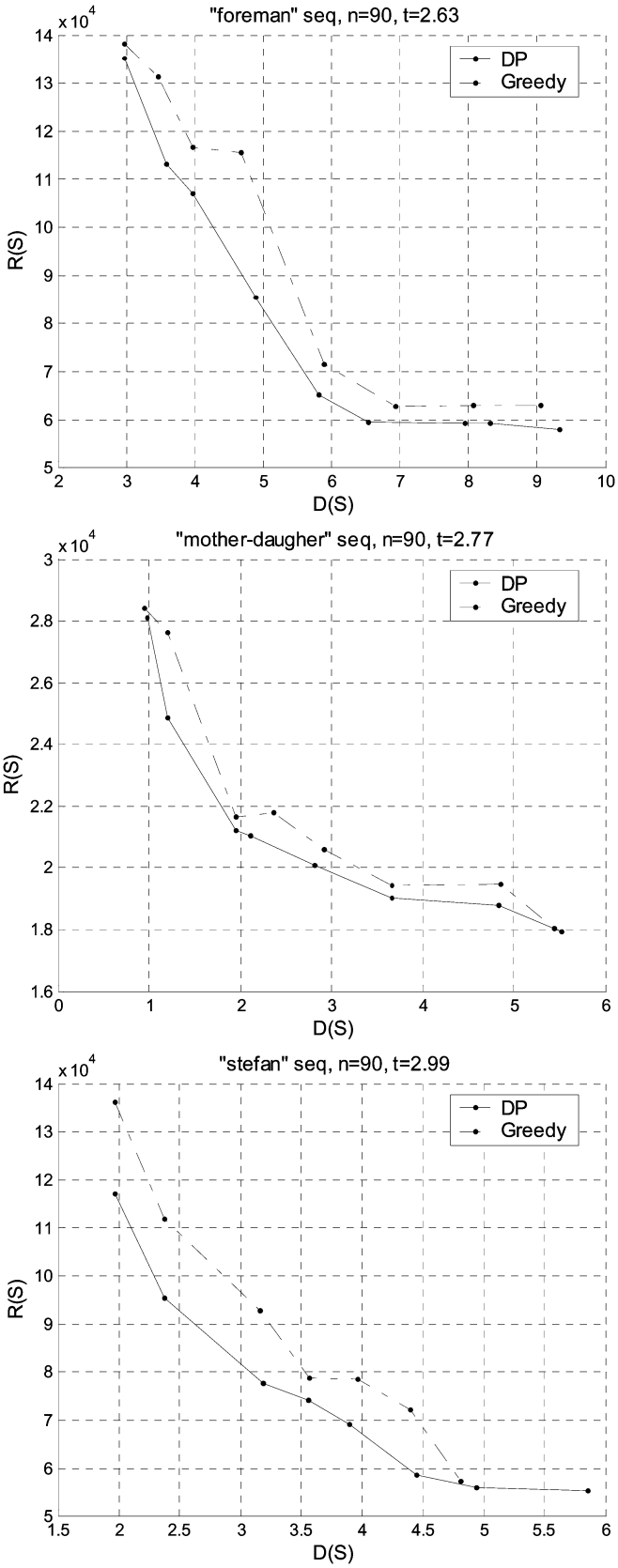


Fig. 11. Performance of the greedy algorithm.

or $k - L > K_{\max}$; then we update $L = k$, and repeat the process until the last frame is reached. The greedy algorithm performance is summarized in Fig. 11 below, in comparison with the

optimal DP-based solution, for segments from the "foreman," "mother–daughter," and "Stefan" sequences.

The greedy algorithm offers relatively good performance, especially when the sequence is not very "active" like, for example, the "mother–daughter" sequence. The computational cost of the greedy algorithm is extremely low, and therefore it can be a valuable alternative to the optimal solution in applications where computational power and buffer size are extremely limited, e.g., a video mobile phone.

VI. Conclusion and Future Works

In this paper we proposed a MINMAX rate distortion optimization framework for the optimal video summary generation problem. We introduced a new frame distortion metric that is well suited for video summarization and video shot cut detection. We developed the optimal algorithms to solve both the rate minimization (MROS) and the distortion minimization (MDOS) formulations with different summarization rates. The resulting summaries are optimal in the MINMAX sense, and the subjective evaluation of the summaries showed that the algorithm can operate at very low bit rate yet offer reasonably good visual quality.

For the case when the computational and buffering capabilities are extremely limited, a heuristic constrained skip algorithm is also developed, which provides a valuable alternative solution, especially for low activity sequences.

We are currently investigating the optimal coding problem in conjunction with the optimal summarization problem. A strategy is being developed for the optimal coding of a video sequence to minimizing the temporal-spatial distortion based on both MSE and summarization distortions.

Acknowledgment

The authors would like to thank the reviewers for their comments and suggestions that helped them to improve the quality of the paper, especially the discussion on the greedy algorithm. They also thank Mr. K. J. O'Connell, manager of the Motorola Multimedia Research Lab (MRL), for his encouragement and support of this work.

References

[1] T. Cover, *Elements of Information Theory*. New York: Wiley, 1991.

[2] D. DeMenthon, V. Kobla, and D. Doermann, "Video summarization by curve simplification," in *Proc. 6th Int. ACM Multimedia Conf.*, Bristol, U.K., 1998, pp. 211–218.

[3] N. Doulamis, A. Doulamis, Y. Avrithis, and S. Kollias, "Video content representation using optimal extraction of frames and scenes," in *Proc. IEEE Int. Conf. Image Processing*, Chicago, IL, 1998, pp. 875–878.

[4] A. Ekin, A. M. Tekalp, and R. Mehrotra, "Automatic soccer video analysis and summarization," *IEEE Trans. Image Processing*, vol. 12, no. 7, pp. 796–807, Jul. 2003.

[5] C. F. Gerald and P. O. Wheatley, *Applied Numerical Analysis*, 4th ed. Reading, MA: Addision Wesley, 1990.

[6] A. Girgenshohn and J. Boreczky, "Time-Constrained key frame selection technique," *Proc. IEEE Multimedia Computing and Systems (ICMCS)*, pp. 756–761, 1999.

[7] Y. Gong and X. Liu, "Summarizing video by minimizing visual content redundancies," in *Proc. Int. Conf. Multimedia and Expo*, 2001, pp. 607–610.

[8] A. Hanjalic and H. Zhang, "An integrated scheme for automated video abstraction based on unsupervised cluster-validity analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, pp. 1280–1289, Dec. 1999.

[9] A. Hanjalic, "Shot-boundary detection: Unraveled and resolved?," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, no. 2, pp. 90–105, Feb. 2002.

[10] Z. He and S. K. Mitra, "A unified rate-distortion analysis framework for transform coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, pp. 1221–1236, Dec. 2001.

[11] A. K. Jain, *Fundamentals of Digital Image Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1989, pp. 11–20.

[12] N. Jayant, J. Johnston, and R. Safranek, "Signal compression based on models of human perception," *Proc. IEEE*, vol. 81, no. 10, pp. 1385–1422, Oct. 1993.

[13] S. Jeannin and A. Divakaran, "MPEG-7 visual motion descriptors," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 6, pp. 720–724, Jun. 2001.

[14] H. Karhunen, *On Linear Methods in Probability Theory*. Santa Monica, CA: Rand, 1960.

[15] I. Koprinska and S. Carrato, "Temporal video segmentation: A survey," *Signal Process. Image Commun.*, vol. 16, pp. 477–500, 2001.

[16] Z. Li, A. K. Katsaggelos, and B. Gandhi, "Temporal rate-distortion based optimal video summary generation," in *Proc. Int. Conf. Multimedia and Expo*, Baltimore, MD, 2003, pp. 693–696.

[17] Z. Li, G. Schuster, A. K. Katsaggelos, and B. Gandhi, "Rate-Distortion optimal video summarization: A dynamic programming solution," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Montreal, QC, Canada, 2004, pp. 457–460.

[18] Z. Li, G. Schuster, A. K. Katsaggelos, and B. Gandhi, "Rate-distortion optimal video summarization with a bit rate constraint," presented at the Proc. IEEE Int. Conf. Image Processing (ICIP), Singapore, 2004.

[19] Z. Li, G. Schuster, A. K. Katsaggelos, and B. Gandhi, "MINMAX optimal video summarization," in *Proc. Int. Workshop Image Analysis for Multimedia Interactive Services (WIAMIS)*, Lisboa, Portugal, 2004, p. 1.

[20] R. Lienhart, "Reliable transition detection in videos: A survey and practioner's guide," *Int. J. Image Graphics*, vol. 1, no. 3, pp. 469–486, 2001.

[21] M. Loeve, *Fonctions Aldatories de Seconde Ordre*. Paris, France: Hermann, 1948.

[22] D. G. Luenberger, *Optimization by Vector Space Methods*. New York: Wiley, 1969.

[23] A. Ortega and K. Ramchandran, "Rate-distortion methods for image and video compression," *IEEE Signal Process. Mag.*, vol. 15, no. 6, pp. 23–50, Nov. 1998.

[24] Y. Qi, A. Hauptmann, and T. Liu, "Supervised classification for video shot," in *Proc. Int. Conf. Multimedia and Expo*, Baltimore, MD, 2003.

[25] K. Ramchandran, A. Oretega, and M. Vetterili, "Bit allocation for dependent quantization with applications to multi-resolution and MPEG video coders," *IEEE Trans. Image Process.*, vol. 3, no. 9, pp. 533–545, Sep. 1994.

[26] G. M. Schuster and A. K. Katsaggelos, *Rate-Distortion Based Video Compression, Optimal Video Frame Compression and Object Boundary Encoding*. Norwell, MA: Kluwer, 1997.

[27] G. M. Schuster, G. Melnikov, and A. K. Katsaggelos, "A review of the minimum maximum criterion for optimal bit allocation among dependent quantizers," *IEEE Trans. Multimedia*, vol. 1, no. 1, pp. 3–17, Mar. 1999.

[28] H. Sundaram and S.-F. Chang, "Constrained utility maximization for generating visual skims," in *Proc. IEEE Workshop Content-Based Access of Image and Video Library*, 2001, pp. 124–131.

[29] C. Taskiran, J.-Y. Chen, A. Albiol, L. Torres, C. A. Bouman, and E. J. Delp, "ViBE: A compressed video database structured for active browsing and search," *IEEE Trans. Multimedia*, vol. 6, no. 1, pp. 103–118, Feb. 2004.

[30] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Trans. Inf. Theory*, vol. IT-13, no. 2, pp. 260–269, Apr. 1967.

[31] Y. Wang, Z. Liu, and J.-C. Huang, "Multimedia content analysis using both audio and visual clues," *IEEE Signal Process. Mag.*, vol. 17, no. 11, pp. 12–36, Nov. 2000.

[32] Y. Zhuang, Y. Rui, T. S. Huan, and S. Mehrotra, "Adaptive key frame extracting using unsupervised clustering," in *Proc. Int. Conf. Image Processing*, Chicago, IL, 1998, pp. 866–870.

Zhu Li (M'01) received the B.S. and M.S. degrees in computer science from Sichuan University, Chengdu, China, and the University of Louisiana at Lafayette in 1992 and 1997, respectively, and the Ph.D. degree in electrical and computer engineering from Northwestern University, Evanston, IL, in 2004.

He has been with the Multimedia Research Lab (MRL), Motorola Laboratories, Schaumburg, IL, since January 2000, and is now a Senior Staff Research Engineer. He is also with the Image and Video Processing Laboratory (IVPL), Department of Electrical and Computer Engineering, Northwestern University, Evanston, IL. His research interests include image/video analysis and coding, optimization, and machine learning.

Dr. Li received a scholarship for graduate study from the Hong Kong University of Science and Technology (HKUST) in 1995.

Guido M. Schuster (M'96) received the Ing. HTL degree in elektronik, mess- und regeltechnik in 1990 from the Neu Technikum Buchs (NTB), Buchs, St. Gallen, Switzerland. He then received the M.S. and Ph.D. degrees, both from the Department of Electrical and Computer Engineering, Northwestern University, Evanston, IL, in 1992 and 1996, respectively.

In 1996, he joined the Network Systems Division, U.S. Robotics, Mount Prospect, IL (later purchased by 3Com). He co-founded the 3Com Advanced Technologies Research Center and served as its Associate Director. He also co-founded the 3Com Internet Communications Business Unit and developed the first commercially available SIP IP Telephony system. He was promoted to the Chief Technology Officer and Senior Director of this Business Unit. During this time, he also served as an Adjunct Professor with the Electrical and Computer Engineering Department, Northwestern University. He is currently a Professor of Electrical and Computer Engineering at the Hochschule für Technik Rapperswil (HSR), Rapperswil, St. Gallen, Switzerland, were he focuses on digital signal processing and Internet multimedia communications. He holds 51 U.S. patents in fields ranging from adaptive control over video compression to Internet telephony. He is the co-author of the book *Rate-Distortion Based Video Compression* (Boston, MA: Kluwer) and has published over 55 peer-reviewed journal and proceedings articles. His current research interests are operational rate-distortion theory and networked multimedia.

Dr. Schuster is the recipient of the gold medal for academic excellence at the NTB, the winner of the first Landis and Gyr fellowship competition, the recipient of the 1999 3Com inventor of the year award, and the recipient of the IEEE Signal Processing Society Best Paper Award 2001 in the multimedia signal processing area.

Aggelos K. Katsaggelos (S'80–M'85–SM'92–F'98) received the Diploma degree in electrical and mechanical engineering from the Aristotelian University of Thessaloniki, Thessaloniki, Greece, in 1979 and the M.S. and Ph.D. degrees, both in electrical engineering, from the Georgia Institute of Technology, Atlanta, in 1981 and 1985, respectively.

In 1985, he joined the Department of Electrical and Computer Engineering, Northwestern University, Evanston, IL, where he is currently Professor, holding the Ameritech Chair of Information Technology. He is also the Director of the Motorola Center for Communications. From 1986 to 1987, he was an assistant professor with the Department of Electrical Engineering and Computer Science, Polytechnic University, Brooklyn, NY. He is an Editorial Board Member of Academic Press, the Marcel Dekker Signal Processing Series, *Applied Signal Processing*, and *Computer Journal* and was an area editor for the journal *Graphical Models and Image Processing* from 1992 to 1995. He is the editor of *Digital Image Restoration* (Springer-Verlag, Heidelberg, Germany, 1991), co-author of *Rate-Distortion Based Video Compression* (Boston, MA: Kluwer, 1997), and co-editor of *Recovery Techniques for Image and Video Compression and Transmission* (Boston, MA: Kluwer, 1998) and is the co-inventor of eight international patents.

Dr. Katsaggelos is an Ameritech Fellow, a member of the Associate Staff, Department of Medicine, at Evanston Hospital, and a member of SPIE. He is a member of the Publication Board of the PROCEEDINGS OF THE IEEE, the IEEE Technical Committees on Visual Signal Processing and Communications, and Multimedia Signal Processing. He served as editor-in-chief of the *IEEE Signal Processing Magazine* from 1997 to 2002, a member of the Publication Boards of the IEEE Signal Processing Society and the IEEE TAB Magazine Committee, an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING from 1990 to 1992, a member of the Steering Committees of the IEEE TRANSACTIONS ON IMAGE PROCESSING (from 1992 to 1997) and the IEEE TRANSACTIONS ON MEDICAL IMAGING (from 1990 to 1999), a member of the IEEE Technical Committee on Image and Multidimensional Signal Processing from 1992 to 1998, and a member of the Board of Governors of the IEEE Signal Processing Society from 1999 to 2001. He received the IEEE Third Millennium Medal in 2000 and the IEEE Signal Processing Society Meritorious Service Award and an IEEE Signal Processing Society Best Paper Award, both in 2001.