Rate-Distortion Optimal Video Summary Generation

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

***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. The summarization process inevitably introduces distortion. The amount of summarization distortion is related to its "conciseness", or the number of frames available in the summary. If there are m frames in the original sequence and n frames in the summary, we define the summarization rate as m/n, to characterize this "conciseness". We also develop a new summarization distortion metric and formulate the summarization problem as a rate-distortion optimization problem. Optimal algorithms based on dynamic programming are presented and compared experimentally with heuristic algorithms. Practical constraints, like the maximum number of frames that can be skipped, are also considered in the formulation and solution of the problem.**

***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 user may want to browse summaries of his/her personal video taken during several trips. In a security application, a supervisor might want to see a 2 minutes summary of what happened at airport gate B20, in the last 10 minutes. 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 SNR distortion, a better option is to transmit a subset of the frames with higher SNR quality. A video summary generator that can "optimally" select frames based on an optimality criterion is essential for these applications.

The solution to this problem is typically based on a two step approach: first identifying video shots from the video sequence [7], [12], [15], [17], and then selecting "key frames" according to some criterion from each video shot. A comprehensive review of past video summarization results can be found in the introduction sections of [6] and [24], and specific examples can be found in [1]-[5], [7], [21] and [25]. Some of the main ideas and results among the previously published results are briefly discussed next.

Zhuang *et al*. [25] proposed an un-supervised 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*. [6] 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 key idea in this paper is to remove the visual redundancy among frames.

DeMenthon *et al*. [1] 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*. [2] 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 [21] 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

Manuscript received February 13th, 2004, revised Septmember 8th, 2004.

Zhu Li is with the Multimedia Research Lab (MRL), Motorola Labs, Schaumburg, IL 60196, USA (phone: 847 576 6942, e-mail: zhu.li@motorola.com).

Guido M. Schuster is now with the Department of Physics, Colorado State University, Fort Collins, CO 80523 USA (e-mail: guido.schuster@hsr.ch).

Aggelos K. Katsaggelos is with the Dept of Electrical & Computer Engineering, Northwestern University, Evanston, IL 60260, USA (e-mail: aggk@ece.northwestern.edu).

Bhavan Gandhi is with the Multimedia Research Lab (MRL), Motorola Labs, Schaumburg, IL 60196, USA (e-mail: abg035@motorola.com).

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 distortion is related to the "conciseness" of the summary, we formulate the summarization problem as a temporal rate-distortion optimization problem. The temporal rate is the ratio of the number of frames selected in the video summary versus that in the original sequence. It characterizes the "conciseness" of the video summary. The summarization distortion is introduced by missing frames. Clearly if all frames are included into the summary, there will be no summarization distortion, and the amount of summarization distortion is determined by the number of missing frames and their locations in the original sequence. We introduce a new frame distortion metric between different frames, and the summarization temporal distortion is then modeled as the average, or equivalently the total frame distortion between the original and reconstructed sequences. Clearly, if we can afford more frames in the summary, the summarization temporal rate will be higher and the summarization distortion will be lower.

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 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 temporal 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. 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 (zero-order hold). Clearly if we can afford more frames in the video 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 trade off between the quality of the reconstructed sequences and the number of frames in the summary, we introduce certain definitions and assumptions for our formulations.

A. Temporal Rate and Distortion

Let a video sequence of n frames be denoted by $V = \{f_0, f_1, \ldots, f_{n-1}\}$. Let its video summary of m frames be $S = \{f_{l_0}, f_{l_1}, \cdots f_{l_{m-1}}\}$, in which l_k denotes the k-th frame selected into the summary S. The summary S is completely determined by the frame selection process $\{l_0, l_1, \ldots, l_{m-1}\}$, which has an implicit constraint that $l_0 < l_1 < \ldots < l_{m-1}$.

The reconstructed sequence $V_S' = \{f_0', f_1', \cdots f_{n-1}'\}$ from the summary S is obtained by substituting missing frames with the most recent frame that belongs to the summary S, that is,

$$f_j' = f_{i=\max(l):\, s.t.\, l\in\{l_0,l_1,\cdots,l_{m-1}\},\, i\le j}. \quad (1)$$

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). The optimal solutions developed in this paper are independent from the definition of this frame metric. To characterize the sequence level summarization distortion, we use the average frame distortion between the original sequence and the reconstruction, given by the *temporal distortion* as,

$$D(S) = \frac{1}{n}\sum_{j=0}^{n-1} d(f_j, f_j'). \quad (2)$$

The *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

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

Notice that the temporal rate $R(S)$ is in 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=1. Thus the rate $R(S)$ can only take values from the discrete set $\{\frac{1}{n}, \frac{2}{n}, \cdots, \frac{n-1}{n}, 1\}$.

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 $D(S) = (1/5)[d(f_0,f_1) + d(f_2,f_3) + d(f_2,f_4)]$.

.

B. Rate-Distortion Optimization Formulations

Video summarization can be viewed as a lossy temporal compression process and a rate-distortion framework [18], [19], [20] is well suited for solving this problem. Using the definitions introduced in the previous section, we now formulate the video summarization problem as a temporal rate-distortion optimization problem. If a temporal rate constraint R_{max} is given, resulting from viewing time, or bandwidth and storage considerations, the optimal video summary is the one that minimizes the sequence temporal distortion. Thus we have:

Formulation I: Minimum Distortion Optimal

Summarization (MDOS):

$$S^* = \arg\min_S D(S),\ s.t.\ R(S) \le R_{\max}, \quad (4)$$

where *D(S)* and *R(S)* are defined by (2) and (3) 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. We call this an (*n*-*m*) summarization problem.

In addition to the rate constraint, we may also impose a constraint on the maximum number of frames, K_{max}, 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. We call this the (*n*-*m*-K_{max}) summarization problem, and its MDOS formulation can be written as,

$$S^* = \arg\min_S D(S),\quad s.t.\ R(S) \le R_{\max}, \text{ and } l_k - l_{k-1} \le K_{\max} + 1,\ \forall k \quad (5)$$

Alternatively we can formulate the optimal summarization problem as a rate-minimization problem. For a given constraint on the maximum distortion D_{max}, the optimal summary is the one that satisfies this distortion constraint and contains the minimum number of frames. Thus we have:

Formulation II: Minimum Rate Optimal Summarization (MROS):

$$S^* = \arg\min_S R(S),\ s.t.\ D(S) \le D_{\max} \quad (6)$$

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

$$S^* = \arg\min_S R(S),\quad s.t.\ D(S) \le D_{\max}, \text{ and } l_k - l_{k-1} \le K_{\max} + 1,\ \forall k \quad (7)$$

The solutions to the MDOS and MROS formulations will be given in Sections III and IV respectively.

III. RATE-DISTORTION OPTIMIZATION: DEFINITIONS AND FORMULATIONS

For the MDOS formulation in (4), if there are *n* frames in the original sequence, and can only have *m* frames in the summary, there are $\binom{n-1}{m-1} = \frac{(n-1)!}{(m-1)!(n-m)!}$ 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. To have an intuitive understanding of the problem, we discuss a heuristic greedy algorithm first before presenting the optimal solution.

A. Greedy Algorithm

Let us first consider a rather intuitive greedy algorithm. For the given rate constraint of allowable frames *m*, the algorithm selects the first frame into the summary and computes the frame distortions. It then identifies the current maximum frame distortion index as $k^* = \max_k \{d(f_k, f_k')\}$ and selects frame f_{k^*} into the summary. The process is repeated until the number of frames in the summary reaches *m*. The resulting solution is sub-optimal. The frames selected into the summary tend to cluster around the high activity regions where the frame-by-frame distortion $d(f_k, f_{k-1})$ is high. The video summary generated is "choppy" when viewed. Clearly we need to better understand the structure of the problem and search for an optimal solution.

B. Distortion State Definition and Recursion

We observe that the MDOS problem has a certain built-in structure and can be solved in stages. For a given current state of the problem, future solutions are independent from past solution. Exploiting this structure, a Dynamic Programming (DP) solution [19], [20] is developed next. An initial version is reported in [14].

Let the distortion state D_t^k be the minimum total distortion incurred by a summary that has *t* frames and ended with frame f_k ($l_{t-1}=k$), that is,

$$D_t^k = \min_{l_1, l_2, \cdots, l_{t-2}} \sum_{j=0}^{n-1} d(f_j, f_{i=\max(l):\, s.t.\ l \in \{0, l_1, l_2, \cdots, l_{t-2}, k\},\, i \le j}) \quad (8)$$

Notice that $l_0=0$ and $l_{t-1}=k$, and they are therefore removed from the optimization. Since $0<l_1<\ldots<l_{t-2}<k$, and $i \le j$, (8) can be re-written as,

$$D_t^k = \min_{l_1, l_2, \cdots, l_{t-2}} \{\sum_{j=0}^{k-1} d(f_j, f_{i=\max(l):\, s.t.\ l \in \{0, l_1, l_2, \cdots, l_{t-2}\},\, i \le j})\} + \sum_{j=k}^{n-1} d(f_j, f_k), \quad (9)$$

in which the second part of the distortion depends on the last summary frame f_k only, and it is removed from the minimization operation. By adding and subtracting the same term in (9) we have,

$$\begin{aligned} D_t^k = \min_{l_1, l_2, \cdots, l_{t-2}} \{&\sum_{j=0}^{k-1} d(f_j, f_{i=\max(l):\, s.t.\ l \in \{0, l_1, l_2, \cdots, l_{t-2}\},\, i \le j}) \\ &+ \sum_{j=k}^{n-1} d(f_j, f_{i=\max(l):\, s.t.\ l \in \{0, l_1, l_2, \cdots, l_{t-2}\},\, i \le j}) \\ &- \sum_{j=k}^{n-1} d(f_j, f_{i=\max(l):\, s.t.\ l \in \{0, l_1, l_2, \cdots, l_{t-2}\},\, i \le j})\} \\ &+ \sum_{j=k}^{n-1} d(f_j, f_k) \end{aligned} \quad (10)$$

We now observe that since $l_{t-2} < k$, we have that

$$\sum_{j=k}^{n-1} d(f_j, f_{i=\max(l):\, s.t.\ l \in \{0, l_1, l_2, \cdots, l_{t-2}\},\, i \le j}) = \sum_{j=k}^{n-1} d(f_j, f_{l_{t-2}}) \quad (11)$$

Therefore the distortion state can be broken into two parts as,

$$D_t^k = \min_{l_1,l_2,\cdots,l_{t-2}} \{\sum_{j=0}^{k-1} d(f_j, f_{i=\max(l):\text{s.t. } l\in\{0,l_1,l_2,\cdots,l_{t-2}\},i\le j}) + \sum_{j=k}^{n-1} d(f_j, f_{i=\max(l):\text{s.t. } l\in\{0,l_1,l_2,\cdots,l_{t-2}\},i\le j}) - \sum_{j=k}^{n-1} d(f_j, f_{l_{t-2}})\} + \sum_{j=k}^{n-1} d(f_j, f_k) \quad , \quad (12)$$

$$= \min_{l_1,l_2,\cdots,l_{t-2}} \{\sum_{j=0}^{n-1} d(f_j, f_{i=\max(l):\text{s.t. } l\in\{0,l_1,l_2,\cdots,l_{t-2}\},i\le j}) - \underbrace{\sum_{j=k}^{n-1} [d(f_j, f_{l_{t-2}}) - d(f_j, f_k)]}_{e^{l_{t-2},k}}\}$$

where the first part represents the problem of minimizing the distortion for the summaries with *t-1* frames and ending with frame l_{t-2}, and the second part represents the "edge cost" of the distortion reduction, if frame *k* is selected into the summary of *t-1* frames ending with frame l_{t-2}. Therefore we have,

$$D_t^{l_{t-1}=k} = \min_{l_{t-2}} \{ \min_{l_1,l_2,\cdots,l_{t-3}} \{\sum_{j=0}^{n-1} d(f_j, f_{i=\max(l):\text{s.t. } l\in\{0,l_1,l_2,\cdots,l_{t-2}\},i\le j})\} - e^{l_{t-2},k}\} = \min_{l_{t-2}} \{D_{t-1}^{l_{t-2}} - e^{l_{t-2},k}\} \quad (13)$$

The relation in (13) establishes the distortion state recursion we need for a DP solution. The back pointer saves the optimal incoming node information from the previous stage. For state D_t^k, it is saved as,

$$P_t^k = \begin{cases} 0, & t = 2 \\ \arg\min_{l_{t-2}} \{D_{t-1}^{l_{t-2}} - e^{l_{t-2},k}\}, & t > 2 \end{cases} \quad (14)$$

Since we assume that the first (0-th) frame is always selected into the summary, P_2^k is set to 0, and the initial state D_1^0 is given as,

$$D_1^0 = \frac{1}{n}\sum_{j=1}^{n-1} d(f_0, f_j) \quad (15)$$

Now we can compute the minimum distortion D_t^k for any video summary of *t* frames and ending with frame *k* by the recursion in (13) with the initial state given by (15). This leads to the optimal DP solution of the MDOS problem.

C. Dynamic Programming Solution for the n-m Summarization Problem

Considering the *n-m* summarization problem case where the rate constraint is given as exactly *m* frames allowed for the summary out of *n* frames in the original sequence, the optimal solution has the minimum distortion of

$$D^* = \min_k \{D_m^k\}, \quad (16)$$

where *k* is chosen from all feasible frames for the *m*-th summary frame. The optimal summary frame selection $\{l_0,l_1,\ldots,l_{m-1}\}$ is therefore found by backtracking via the back pointers $\{P_t^k\}$, similar to the Viterbi algorithm [23],

$$\begin{cases} l_{m-1} = \arg\min_k \{D_m^k\} \\ l_t = P_{t+2}^{l_{t+1}}, \text{ for } t \in [1, m-2] \\ l_0 = 1. \end{cases} \quad (17)$$

As an illustrative example, the distortion state trellis for *n=5* and *m=3* is shown in Fig. 1. Each node represents a distortion state D_t^k, and each edge $e^{j,k}$ represents the distortion reduction if frame f_k is selected into the summary which ended with frame f_j. Note that the trellis topology is completely determined by *n* and *m*. According to Fig. 1, node D_2^4 is not included, since *m*=3 therefore f_4 (the last frame in the sequence) cannot be the second frame in the summary.

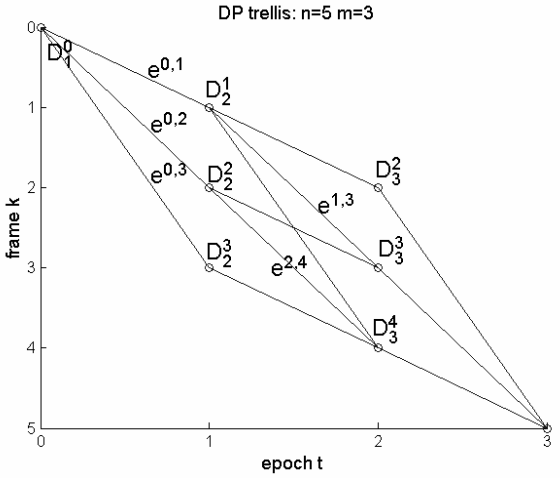


Fig. 1. DP trellis example for n=5, m=3.

Once the distortion state trellis and back pointers are computed recursively according to Eqs. (13) –(15), the optimal frame selection can be found by Eqs. (16) and (17). The number of nodes at every epoch *t>0*, or the depth of the trellis, is *n-m+1*, and we therefore have a total of *1+(m-1)(n-m+1)* nodes in the *n-m* trellis that need to be evaluated.

D. Skip Constraint

The frame skip constraint in summarization is a desirable feature. It limits the maximum number of frames that can be skipped between any two summary frames and can be used to ensure certain degree of temporal smoothness in the video summary playback. When the maximum frame skip constraint K_{max} is imposed in the summary as in (5), the DP trellis topology is affected. As K_{max} is becoming smaller, the number of nodes and edges is also decreasing, which results in lower computational complexity. The resulting solution is optimal subject to the skip constraint, but clearly the resulting distortion is larger (at best equal) to the distortion resulting from the MDOS formulation without the skip constraint in Eq. (4).

The new DP trellis with the skip constraint is denoted as an (n-m-K_{max}) trellis and is completely determined by these three parameters. From each node, the feasible out-going edges are limited by K_{max} in addition to the *n-m* trellis constraint. The

values of K_{max} are in the range [*1, n-m+1*]. Example trellises for *n=9, m=3* are shown in Fig. 2.

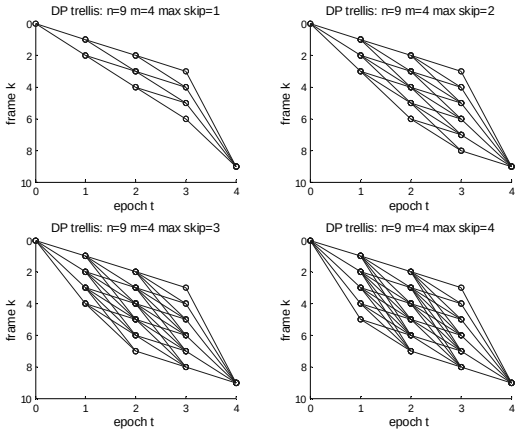


Figure 2. Example DP trellises with various skip constraints

When the maximum skip constraint is not active, or equivalently when $K_{max} > n\text{-}m\text{+}1$, there are *n-m+1* edges from stage 0 to stage 1, and *1+2+…+ (n-m+1)* edges for the remaining stages. The resulting total number of edges is therefore given by,

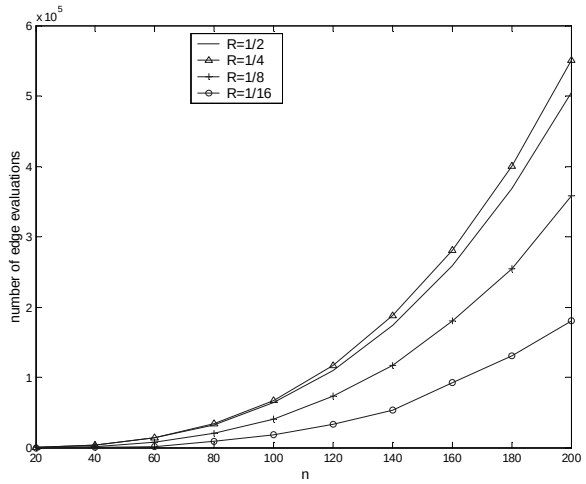
$$E(n,m) = 2(n-m+1) + \frac{(m-2)(n-m+1)(n-m+2)}{2}, \quad (18)$$

or equivalently with rate $R=m/n$, we have,

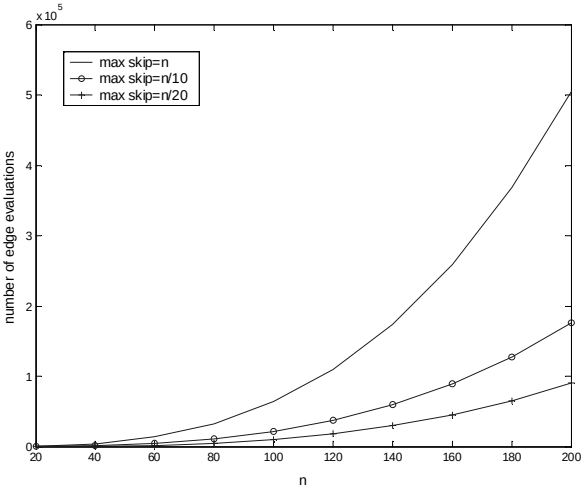
$$\begin{aligned} E(n,R) &= 2(n-Rn+1) \\ &\quad + \frac{(Rn-2)(n-Rn+1)(n-Rn+2)}{2} \\ &= \frac{1}{2}[R(1-R)^2 n^3 + (7R-7R^2-2)n^2 \\ &\quad + (9R-6)n + 6] \end{aligned} \quad (19)$$

For a given *R*, the computational complexity for the evaluation of edges grows with the problem size *n* as $O(n^3)$. *E(n,R)* is shown in Fig. 3a for various values of *R* and *n.* The effect of the skip constraint on the number of edge evaluation for the MDOS problems for a fixed rate *R=0.5* and variable size *n* is shown in Fig. 3b. Notice that for large *n* and small values of the maximum skip constraint, the reduction in the computational complexity becomes significant.

For a given rate *m*, the implicit maximum skip constraint is $K_{max}=n\text{-}m$, which is imposed by the topology of a full *n-m* DP trellis. On the other hand, if K_{max} is smaller than *n/m*, the DP trellis will not be able to consider all *n* frames into the optimization. Therefore for K_{max} to be meaningful, it should belong to the range [*n/m, n-m+1*]. This can be a rather wide range depending on the values of *n* and *m*. Although no specific guidelines are provided for the choice of K_{max}, its value in general should be closer to *n/m* rather than *n-m+1*, in order to address both benefits of reduced computational load and smoothness in the resulting summary.



(a)



(b)

Fig. 3. Computation complexity of the DP solution as a function of the number of frames *n* and (a) rate *R;* (b) the maximum skip constraint

IV. SOLUTION OF THE MROS FORMULATION

For the MROS formulation (6), we minimize the temporal rate of the video summary, or select the smallest number of frames possible that satisfy the distortion constraint. There are two approaches to obtain the optimal solution. According to the first one, the optimal solution results from the modification of the DP algorithm for the MDOS problem. The DP "trellis" is not bounded by the *m*, (length or number of epochs), and its depth equal to *(n-m+1)*, anymore; it is actually a tree with root at D_1^0 and expanding in the *n x n* grid. The only constraints for the frame selection process are the "no look back" and "no repeat" constraints. The algorithm performs a Breadth First Search (BFS) on this tree and stops at the first node that satisfies the distortion constraint, which therefore has the minimum depth, or the minimum temporal rate. The computational complexity of this algorithm grows exponentially and it is not practical for large size problems.

To address the computational complexity issue of the first

algorithm, we propose a second algorithm that is based on the DP algorithm for the solution of the MDOS formulation. Since we have the optimal solution to the MDOS problem, and we observe that feasible rates {*1/n, 2/n, … n/n*} are discrete and finite, we can solve the MROS problem by searching through all feasible rates, and for each feasible rate *R=m/n*, solve the MDOS problem to obtain the minimum distortion $D^*(R)$. The operational *rate-distortion* function $D^*(R)$ resulting from the MDOS optimization is given by,

$$D^*(R) = D^*(m/n) = \min_{l_1, l_2, \cdots, l_{m-1}} (1/n) \sum_{j=0}^{n-1} d(f_j, f_j'), \quad (20)$$

that is, it represents the minimum distortion corresponding to the rate *m/n*. An example of this function is shown in Fig. 4.

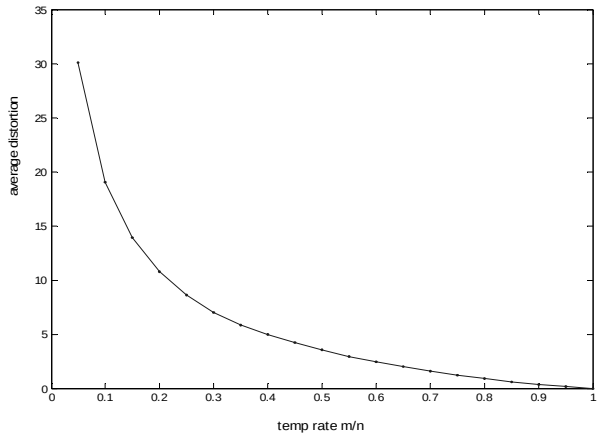


Fig. 4. An example of the operational rate-distortion function

If the resulting distortion $D^*(R)$ satisfies the MROS distortion constraint, the rate *R* is labeled as "*admissible*". The optimal solution to the MROS problem is therefore the minimum rate among all admissible rates. Therefore, the MROS problem with distortion constraint D_{max} is solved by,

$$\min_{R \in \{\frac{1}{n}, \frac{2}{n}, \cdots, \frac{n}{n}\}} R, \; s.t. D^*(R) \le D_{\max} \quad (21)$$

The minimization process is over all feasible rates. The solution to (21) can be found in a more efficient way, since that the rate-distortion function is a non-increasing function of *m*, that is,

Lemma1:

$D^*(m_1/n) \le D^*(m_2/n)$, *if* $m_1 > m_2$, *for* $m_1, m_2 \in [1, n]$

Proof: If we prove that $D^*(m+1/n) \le D^*(m/n)$, then since we have that $D^*(m/n) \le D^*(m-1/n) \cdots \le D^*(1/n)$, Lemma 1 is true. Let $D^*(m/n)$ be the minimum distortion introduced by the optimal *m*-frame summary solution $L^* = \{0, l_1, l_2, \ldots, l_{m-1}\}$, for some $1<m<n$. Since $m<n$, there exists an l_t such that the previous frame to frame f_{l_t}, i.e., f_{l_t-1} (clearly between frames $f_{l_{t-1}}$ and f_{l_t}) does not belong to the summary solution L^*. If frame f_{l_t-1} were to be included in the summary, a new summary with frame selection $L_N = L^* \cup \{f_{l_t-1}\}$ would be generated with resulting distortion $D^{L_N} = D^*(m/n) - d(f_{l_{t-1}}, f_{l_t-1})$. Since $d(f_{l_{t-1}}, f_{l_t-1}) \ge 0$, we have $D^{L_N} \le D^*(m/n)$. Since the resulting *m+1* frame summary (with the inclusion of frame f_{l_t-1}) is not necessarily optimal, we have that $D^*(m+1/n) \le D^{L_N} \le D^*(m/n)$.

Lemma 1 is quite intuitive, since adding a frame to the summary always reduces, or at least keeps the resulting distortion the same. Also because the operational distortion-rate function $D^*(m/n)$ is a discrete and non-increasing function as established in Lemma 1, the MROS problem in (21) can be solved efficiently by a bi-section search [3] on the $D^*(m/n)$.

The algorithm starts with an initial rate bracket of $R_{lo}=1/n$ and $R_{hi}=n/n$, and computes its associated initial distortion bracket of $D_{lo}=D^*(R_{lo})$, and $D_{hi}=D^*(R_{hi})$. If the MROS distortion constraint $D_{max} > D_{lo}$, then the optimal rate is *1/n*. Otherwise we select a middle rate point $R_{new} = \lfloor (R_{hi} + R_{lo})/2 \rfloor$, compute its associated distortion $D^*(R_{new})$, and find the new rate and distortion bracket by replacing either the R_{lo} or the R_{hi} point with R_{new}, such that the distortion constraint D_{max} is within the new distortion bracket. The process is repeated until the rate bracket converges, i.e, $R_{hi}=m^*/n$, $R_{lo}=(m^*-1)/n$, for some m^*. At this point the optimal rate is found as $R^*=m^*/n$, and the optimal solution to the MROS problem is the solution of the n-m^* summarization problem as discussed in Section IIIC. The computational complexity of the bi-section search algorithm is *O(log(n))* times the complexity of the DP *n-m* summarization algorithm.

V. Experimental Results

A. Frame Distortion Metric

The rate-distortion optimal summarization formulation we developed does not depend on a specific frame distortion metric. This offers additional flexibility in summarization solutions. However, an effective and computationally efficient frame distortion metric is also essential to the success of the summarization algorithm.

There are a number of ways to compute the frame distortion $d(f_j, f_k)$. The Mean Squared Error (MSE) has been widely used in image processing. However, it is well-known that the MSE type metric 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, although the perceptual distortion is negligible. There is work in the literature addressing the perceptual quality issues, (for example, [9] and others), however such works are addressing primarily the distortion between an image and its quantized versions.

The color histogram-based distance is also a popular choice [25], 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.

In our previous work on heuristic summarization [13], we adopted a frame distortion metric that is based on the weighted sum of color change and motion activity. The color change is computed from the MPEG-7 color layout feature [26], which not only account for color distribution in YCbCr color space, but also the layout, or the spatial distribution of color. This addresses a problem of histogram-based color features. The motion activity [10] is computed from the variance of the magnitude of the motion vectors between frames. The results are satisfactory in general, but the computation of motion activity is quite expensive.

For a summarization frame distortion metric that reflects the human perception well while can be computed efficiently, we developed a metric that is based on the scale and user preference. The scale *W*x*H* is the spatial resolution at which we want to differentiate frames. Video frames are first scaled into smaller sizes of *W*x*H*, for e.g., 8x6, 12x9 or 16x12. The benefit of this scaling process is to reduce noises and local variances. In our simulation we use both the 8x6 and 11x9 scales. The latter can be obtained in a very fast way from the compressed QCIF size sequences by extracting the DC components.

After scaling, video frames are projected through Principal Component Analysis (PCA) to a linear subspace that preserves most information while further reduces the feature dimension for easier manipulation. The PCA transform *T* is found by diagonalizing the covariance matrix of the frames [11][16], and selecting the desired number of dimensions with the largest eigen-values. The frame distortion is therefore computed as the weighted Euclidean distance between two frames in the PCA subspace, where weight reflects user preference, and is given by,

$$d(f_j, f_k) = \| T(S_{WxH}(f_j)) - T(S_{WxH}(f_j)) \|_V \qquad (22)$$
$$V = A^T A$$

In Eq. (22), S_{WxH} denotes the scaling operation, *T* is the PCA transform. The weight *V* is obtained from user preference, which can also expressed as another linear projection *A*. When large amount of labeled video frames are available, a supervised learning process like Fisher Discriminant Analysis (FDA) can be applied to learn the projection *A*.

In our experiment we collected 3200 frames from various video clips for the PCA. The resulting eigen-values of the PCA for the 8x6 scale are plotted in Fig. 5. Notice that most of the energy is captured by the bases corresponding to the 8 largest eigen-values. Therefore our adopted PCA transform matrix *T* has dimension 8 by 48. Since we don't have any preference at this time, so the weights are uniform.

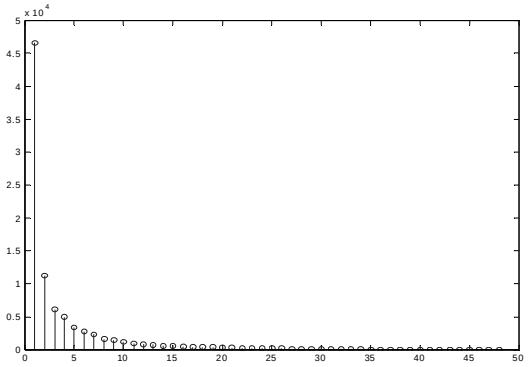


Figure 5. Eigen values resulting from scaling and PCA

Experimental results with this frame distortion metric are shown as a frame-by-frame distance plot $d(f_k, f_{k-1})$ in the upper plot in Fig. 6 for the "foreman" sequence. It seems to reflect well the perceptual change of the sequence, since for the "foreman" sequence, frames 1-200 contain a talking head with little 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. In the lower plot of Fig. 6, the frame-by-frame distortion is plotted for the "mother-daughter" sequence, which is a lower-activity sequence compared with the "foreman" sequence. This is well reflected by the overall lower values in the frame-by-frame plot. Similar interpretation of events can also be found in this example; for example, there is a spike around frames 58-72, which corresponds to the mother touching her daughter's hair. The PCA space Euclidean distance metric has performed better than any other metric we tried and similarly to the color change and motion activity based metric developed in [13], at a lower computational cost.

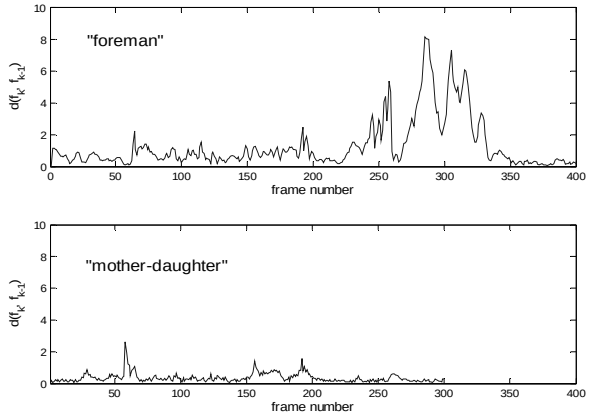


Figure 6. Frame-by-frame distortion $d(f_k, f_{k-1})$ plot for the sample sequences.

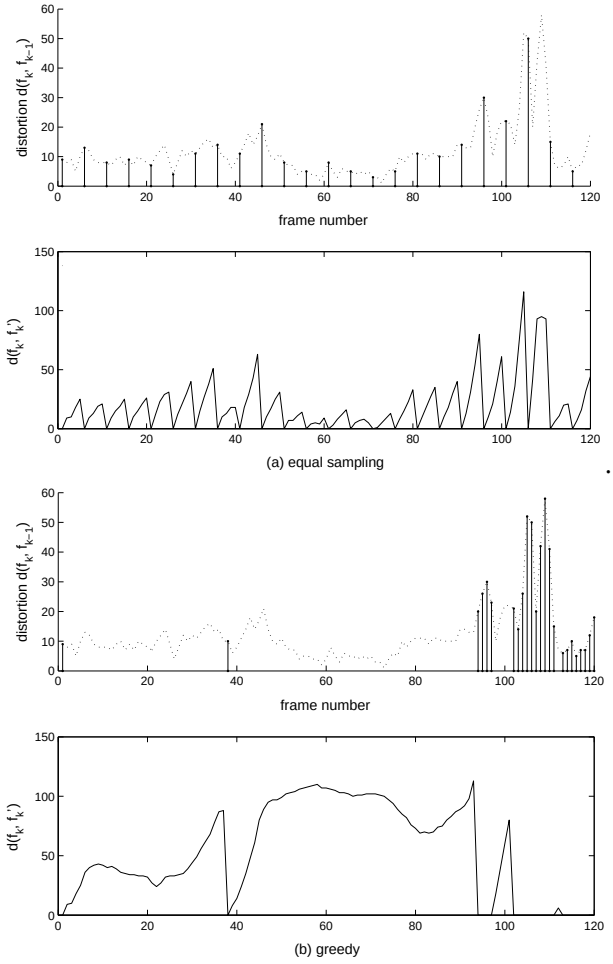
From this experiment it is clear that the metric function in (22) is fairly accurate in depicting the distortion or the dissimilarity of different frames. The computation of this metric does not involve motion estimation. For compressed

sequences, the scaling can be down efficiently from extracting DC values. Overall the computation is moderate.

B. Simulation Results

We tested the proposed DP algorithm with and without skip constraints, as well as the Greedy algorithm described in Section IIIA, and the content blind equal sampling solution on the "foreman" sequence. For the segment with n=120 frames (frames 150-269), the MDOS optimal video summary frame selections and resulting sequence distortions are plotted in Fig. 7. The rate constraint is $R(S)=0.2$, with the number of summary frames given as m=24.

In Fig. 7a, the results from the equal sampling solution 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. Notice that the distortion is high in the high activity region around frame number 100, since the selection is content-blind. The bottom plot shows the per frame distortion, $d(f_k, f_k')$, between the original sequence and the reconstructed sequence from the video summary. The corresponding two plots obtained by the application of the Greedy algorithm are plotted in Fig. 7b. The summary frames are concentrated around the high activity region in this case. This solution does adapt to the content, but it is obviously sub-optimal.



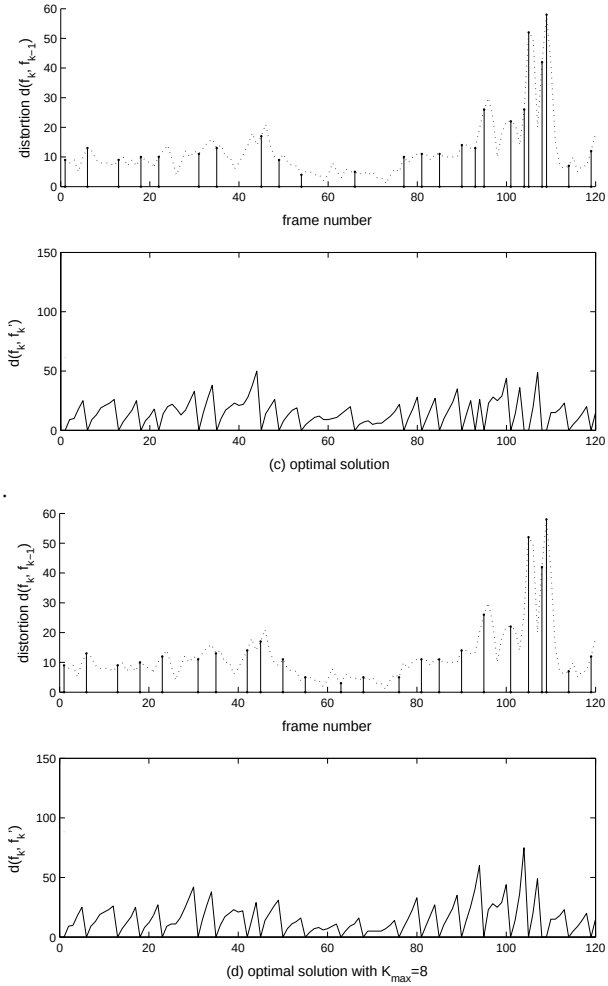


Figure 7. Summaries generated for the "foreman" sequence segment (frames 150-269)

The results from the application of the optimal DP algorithm without the skip constraint are plotted in Fig. 7c. For the given frame budget of $m=24$, this solution offers the minimum distortion. Notice that the summary frames are rather evenly distributed but more frames are selected from the high activity region. Fig. 7d shows the optimal solution with a maximum frame skip constraint equal to 8. The solution is very similar to the solution in Fig 7c, but the distortion incurred is slightly larger than that of the optimal solution without the skip constraint.

It is clear that the optimal algorithm performs better than the heuristic solutions for the same temporal rate constraint. We tested the proposed algorithms with a number of sequences. The distortion performances for the "foreman" sequence, frames 150-269, and the "flower" sequence frames 20-139, both for a matched rate of R=0.2, are summarized in Tables 1 and 2 respectively. In addition to the average distortion the maximum frame distortion is shown, as well as the standard deviation of the frame distortions. Besides minimizing the average distortion, the DP based solutions also result in smaller maximum distortion and standard deviation of the distortion.

TABLE I
DISTORTION PERFORMANCE FOR THE "FOREMAN" SEQUENCE: N=120, M=24, AND MAX SKIP=8.

Algorithm	Distortion (Avg)	Distortion (Max)	Distortion (Std)
Greedy	53.70	113.00	39.38
Equal	19.48	116.00	22.01
DP	14.53	50.00	11.16
DP+Skip	14.98	75.00	13.21

.

TABLE II
DISTORTION PERFORMANCE FOR THE "FLOWER" SEQUENCE: N=120, M=24, AND MAX SKIP=8.

Algorithm	Distortion (Avg)	Distortion (Max)	Distortion (Std)
Greedy	86.62	217.00	70.51
Equal	16.18	94.00	18.24
DP	12.96	40.00	10.56
DP+Skip	12.98	50.00	11.28

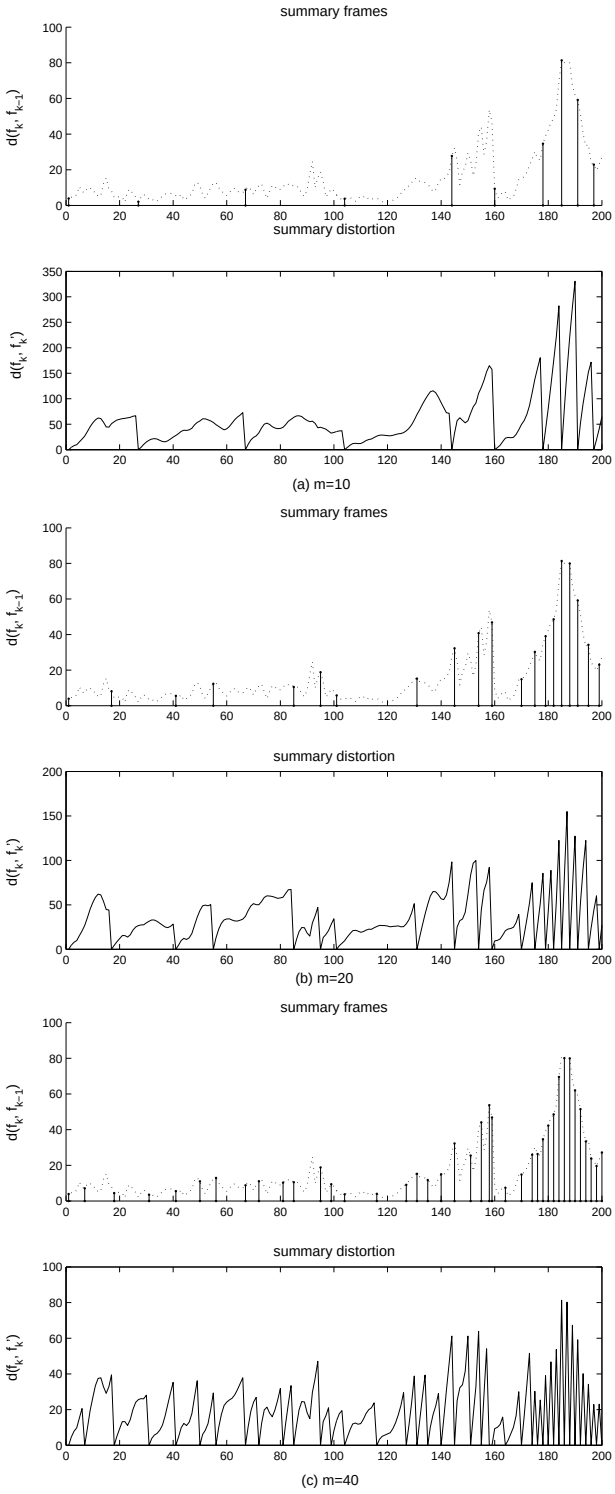
.

We also obtained summarization results for the same sequence at different rates. The summarization results for the 200-frame "foreman" sequence segment between frames 100-299 at summarization rates equal to 0.05, 0.1, 0.2 and 0.4 are plotted in Figs. 8a, 8b, 8c, and 8d, respectively. As expected, as the summarization rate goes up, the summarization distortion goes down.

TABLE III
DISTORTION PERFORMANCE FOR THE "FOREMAN" SEQUENCE: N=200, AT DIFFERENT RATES.

$\boldsymbol{R}(\boldsymbol{S})$	Distortion (Avg)	Distortion (Max)	Distortion (Std)
0.05	56.46	330.60	50.93
0.1	34.95	155.02	26.80
0.2	18.50	81.35	16.17
0.4	7.40	29.53	7.80

The resulting distortion statistics are shown in Table 3. The above summaries clips compressed with a H.263 [22] video coder are also available upon request from interested readers.



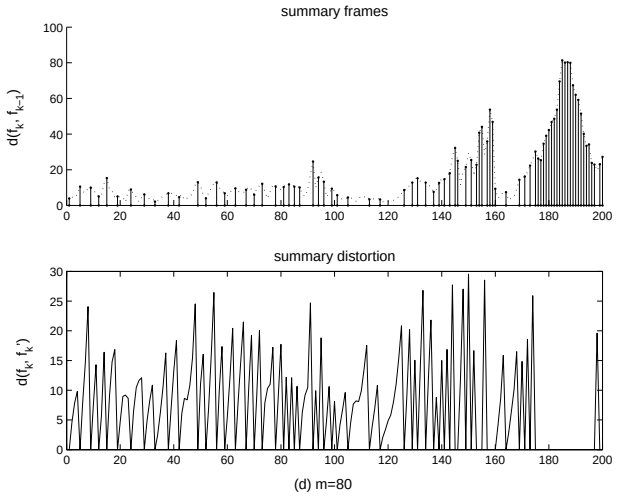


Figure 8. Summarization results at different rate levels for a segment of the "foreman" sequence.

Overall the DP based algorithms produce reconstructed sequences that degrade gracefully as the temporal rate decreases. Imposing a skip constraint makes the video summary smoother and at the same time reduces the computational complexity. The computational complexity involved is moderate and can be optimized for both off-line summarization and on-line video trans-coding into summaries.

VI. Conclusion and Future Works

In this paper we proposed a 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. A recursive distortion state transition is found based on which a dynamic programming solution is developed to solve the minimum distortion optimal summarization (MDOS) formulation. A bi-section search solution is developed to solve the minimum rate optimal summarization (MROS) formulation. The effectiveness of the developed framework is demonstrated via simulations.

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 with control of both the temporal and spatial (PSNR) distortion trade-off.

Acknowledgement

We would like to thank Mr. Kevin J. O'Connell, manager of the Motorola MRL lab for his encouragement and support of this work.

References

[1] D. DeMenthon, V. Kobla and D. Doermann, "Video Summarization by Curve Simplification", *Proceedings of ACM Multimedia Conference* 1998, Bristol, U.K.

[2] N. Doulamis, A. Doulamis, Y. Avrithis and S. Kollias, "Video Content Representation Using Optimal Extraction of Frames and Scenes", *Proc. of Int'l Conference on Image Processing*, Chicago, Illinois, 1998.

[3] C. F. Gerald and P. O. Wheatley, *Applied Numerical Analysis*, 4th edition, Reading MA, Addisson, 1990.

[4] A. Girgenshohn and J. Boreczky, "Time-Constrained Key frame Selection Technique", *Proc. of IEEE Multimedia Computing and Systems (ICMCS)*, 1999.

[5] Y. Gong and X. Liu, "Video Summarization with Minimal Visual Content Redundancies", *Proc. of Int'l Conference on Image Processing*, 2001.

[6] A. Hanjalic and H. Zhang, "An Integrated Scheme for Automated Video Abstraction Based on Unsupervised Cluster-Validity Analysis", *IEEE Trans. on Circuits and Systems for Video Technology*, vol.9, December 1999.

[7] A. Hanjalic, "Shot-Boundary Detection: Unraveled and Resolved?", *IEEE Trans. on Circuits and Systems for Video Technology*, vol.12, No. 2, Februrary 2002.

[8] A. K. Jain, *Fundamentals of Digital Image Processing*, Prentice Hall, Englewood Cliffs, N.J., 1989, pp 11~20.

[9] N. Jayant, J. Johnston, and R. Safranek, "Signal Compression Based on Models of Human Perception", *Proceedings of IEEE*, vol. 81, pp. 1385-1422, October, 1993.

[10] S. Jeannin and A. Divakaran, "MPEG-7 Visual Motion Descriptors", *IEEE Trans. on Circuits and Systems for Video Technology*, vol.11, June 2001.

[11] H. Karhunen, "On Linear Methods in Probability Theory", English translation, Doc. T-131, Rand Corp, Santa Monica, CA, 1960.

[12] I. Koprinska, S. Carrato, "Temporal Video Segmentation: a survey", *Signal Processing: Image Communication*, vol.16, pp. 477-500, 2001.

[13] Z. Li, A. K. Katsaggelos and B. Gandhi, "Temporal Rate-Distortion Optimal Video Summary Generation", *Proceedings of Int'l Conference on Multimedia and Expo*, 2003, Baltimore, USA.

[14] Z. Li, G. Schuster, A. K. Katsaggelos and B. Gandhi, "Rate-Distortion Optimal Video Summarization: A Dynamic Programming Solution", *Proceedings of Int'l Conference on Acoustics, Speech, and Signal Processing* (ICASSP), Montreal, Canada, 2004.

[15] R. Lienhart, "Reliable Transition Detection in Videos: A Survey and Practioner's Guide", *International Journal of Image and Graphics*, Vol.1, No.3, pp. 469-486, 2001.

[16] M. Loeve, *Fonctions aldatories de seconde ordre*, Hermann, Paris, 1948.

[17] Yanjun Qi, Alexander Hauptmann and Ting Liu, "Supervised Classification for Video Shot", *Proceedings of Int'l Conference on Multimedia and Expo*, 2003, Baltimore, USA

[18] K. Ramchandran, A. Oretega, and M. Vetterili, "Bit Allocation for dependent quantization with applications to multi-resolution and MPEG video coders", *IEEE Trans. Image Processing*, vol. 3, September, 1994.

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

[20] 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. on Multimedia*, vol. 1, No. 1, March 1999.

[21] H. Sundaram and S-F. Chang, "Constrained Utility Maximization for Generating Visual Skims", *IEEE Workshop on Content-Based Access of Image & Video Library*, 2001.

[22] University of British Columbia, H.263 Reference Software Model: TMN8.

[23] A. J. Viterbi, "Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm", *IEEE Trans. on Information Theory*, April 1967, vol. IT-13, pp. 260-269.

[24] Y. Wang, Z. Liu and J-C. Huang, "Multimedia Content Analysis", *IEEE Signal Processing Magazine*, vol. 17, November 2000.

[25] Y. Zhuang, Y. Rui, T. S. Huan, and S. Mehrotra, "Adaptive Key Frame Extracting Using Unsupervised Clustering", *Proc. of Int'l Conference on Image Processing*, Chicago, Illinois, 1998

[26] B. S. Manjunath, J-R. Ohm, V. V. Vasudevan and A. Yamada, "Color and Texture Descriptors", *IEEE Trans. on Circuits and Systems for Video Technology*, vol.11, June 2001.

Zhu Li (M'01) received the B.S. and M.S. degrees in Computer Science from Sichuan University, Chengdu, China, in 1992, and the University of Louisiana at Lafayette in 1997 respectively. He received the PhD degree in Electrical & Computer Engineering from Northwestern University, Evanston,

in 2004. He has been with Multimedia Research Lab (MRL), Motorola Labs since 2000 and is now a senior staff research engineer. His research interests include video analysis and machine learning, video coding and communications.
Dr. Li received a graduate scholarship from the Hong Kong University of Science & Technology in 1995.

Guido M. Schuster (S'94–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, Illinois, in 1992 and 1996, respectively. In 1996 he joined the Network Systems Division of U.S. Robotics in Mount Prospect, Illinois (later purchased by 3Com). He co-founded the 3Com Advanced Technologies Research Center and served as the Associate Director of the Center. 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 in the Electrical and Computer Engineering Department at Northwestern University. He is currently a Professor of Electrical and Computer Engineering at the Hochschule fur Technik Rapperswil (HSR), Rapperswil, St.Gallen, Switzerland, where he focuses on Digital Signal Processing and Internet Multimedia Communications.
Guido M. Schuster holds 44 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", published by Kluwer Academic Publishers and has published 53 peer reviewed journal and proceedings articles. Furthermore he is the recipient of the gold medal for academic excellence at the NTB, the winner of the first Landis & Gyr fellowship competition, the recipient of the 3Com inventor of the year 1999 award and the recipient of the IEEE Signal Processing Society Best Paper Award 2001 in the multimedia signal processing area. His current research interests are operational rate-distortion theory and networked multimedia.

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 Engineering and Computer Science at 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. During the 1986–1987 academic year, he was an Assistant Professor with the Department of Electrical Engineering and Computer Science, Polytechnic University, Brooklyn, NY. His current research interests include image and video recovery, video compression, motion estimation, boundary encoding, computational vision, and multimedia signal processing and communications.
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 editorial boards of PROCEEDINGS OF THE IEEE, the Marcel Dekker, Signal Processing Series, Applied Signal Processing, the Computer Journal, and a member of the IEEE Technical Committees on Visual Signal Processing and Communications, and Multimedia Signal Processing. He has served as an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING (1990–1992), an area editor for the journal Graphical Models and Image Processing (1992–1995), a member of the Steering Committees of the IEEE Transactions on Image Processing (1992–1997) and the IEEE TRANSACTIONS ON MEDICAL IMAGING (1990–1999), a member of the IEEE Technical Committee on Image and Multidimensional Signal Processing (1992–1998), the Board of Governors of the Signal Processing Society (1999–2001), the Publication Board of the IEEE Signal Processing Society (1997–2002), the IEEE TAB Magazine Committee (1997–2002), and editor-in-chief of the IEEE Signal Processing Magazine (1997–2002). He is the editor of Digital Image Restoration (Berlin, Germany, Springer-Verlag, 1991), co-author of Rate-Distortion Based Video Compression (Kluwer Academic Publishers, 1997), and co-editor of Recovery Techniques for Image and Video Compression and Transmission, (Kluwer Academic Publishers, 1998). He has served as the General Chairman of the 1994 Visual Communications and Image Processing Conference (Chicago, IL), and as technical program co-chair of the 1998 IEEE International Conference on Image Processing (Chicago, IL). He is the co-inventor of nine international patents, the recipient of the IEEE Third Millennium Medal (2000), the IEEE Signal Processing Society Meritorious Service Award (2001), and an IEEE Signal Processing Society Best Paper Award (2001).

Bhavan Gandhi (M'01) received the B.S (with Honors) and M.S. degrees in Electrical Engineering from University of Illinois at Urbana-Champaign in 1986 and 1988 respectively. He was a Research Scientist at Eastman Kodak Company until 1998, where he was primarily involved in developing image compression technology.

He is currently a Distinguished Member of Technical staff in the Center for Applications, Content, and Services Research within Motorola Laboratories in Schaumburg, IL. His research interests are in the areas of image/video compression, multimedia analysis, and content-based retrieval systems. He has seven issued patents and is a co-author of eleven publications in these areas. He is currently a Member of Tech Staff at Motorola Labs.