OPTIMAL VIDEO SUMMARIZATION WITH A BIT BUDGET CONSTRAINT

[1,3]Zhu Li, [2]Guido M. Schuster, [1]Aggelos K. Katsaggelos, and [3]Bhavan Gandhi

[1]Department of Electrical & Computer Engineering, Northwestern University, Evanston
[2]Hochschule für Technik Rapperswil, Switzerland
[3]Multimedia Communication Research Lab (MCRL), Motorola Labs, Schaumburg

ABSTRACT

The need for video summarization originates primarily from a viewing time or a bit budget 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, which translates into a bit budget constraint. Our work is based on a bit rate-summary distortion optimization formulation. New metrics for video summary distortion are introduced. An optimal algorithm based on Lagrangian relaxation and dynamic programming is presented.

1. INTRODUCTION

The demand for video summary work originates from a viewing time constraint as well as bit budget constraint for communication and storage limitations in security, military and entertainment applications. For example, 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, and then selecting "key frames" according to some criterion from each video shot to generate a video summary for the sequence. Examples of past work are listed in [1]-[9], [13],[16]. With these approaches, various visual features and their statistics are computed to identify video shot boundaries and determine key frames by thresholding and clustering. In general most such techniques require two passes, are rather computationally involved, do not have uniform temporal resolution within a video shot, and they are heuristic in nature.

Since a video summary inevitably introduces distortions at the play back stage and the amount of distortion is related to the "conciseness" of the summary, we formulate this problem as a rate-distortion optimization problem. The rate here is represented by the bits spent to code the video summary with a given coding strategy. The temporal distortion is introduced by the missing frames from the summary. We introduce a temporal frame distortion metric and the temporal distortion is then modeled as the frame distortion between the original and the reconstructed sequences. The optimal solution is obtained using Lagrangian relaxation and dynamic programming.

The paper is organized into the following sections. In section 2 we present the formal definitions and the rate-distortion optimization formulations of the optimal video summary generation problem. In section 3 we discuss our optimal video summary solution to the temporal distortion minimization formulation. In section 4 we discuss the optimal video summary solution for the temporal rate minimization formulation. In section 5 we present and discuss some of our experimental results. In section 6 we draw conclusions and outline our future work.

2. DEFINITIONS AND FORMULATIONS

A video summary is a shorter version of the original video sequence. Video summary frames form a subset of the frames selected from the original video sequence. The reconstructed video sequence is generated from the video summary by substituting the missing frames with the previous frames in the summary (zero-order hold). To state the trade off between the quality of the reconstructed sequences and the number of frames in the summary, we have the following definitions.

Let a video sequence of n frames be denoted by $V= \{f_0, f_1,\ldots,f_{n-1}\}$, and its video summary of m frames by $S= \{f_{l_0}, f_{l_1}, \cdots f_{l_{m-1}}\}$, in which l_k denotes the k-th summary frame's location in the original sequence V. 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, $f_0'=f_0$, and

0-7803-8554-3/04/$20.00 ©2004 IEEE.

$$f_j' = f_{i=\max(l):\, s.t.\, l\in\{l_0,l_1,\cdots l_{m-1}\},\, i\le j}, \qquad j=1,2,\cdots n-1 \tag{1}$$

Let the distortion between two frames j and k be denoted by $d(f_j, f_k)$; then the sequence distortion introduced by the summarization is given by the total (or average) distortion,

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

The summary bit rate is defined as the total number of bits used to code the summary frames in S,

$$R(S) = \sum_{t=0}^{m-1} b(f_{l_t}) = \begin{cases} \sum_{t=0}^{m-1} r^{l_t}, & \text{intra - coding} \\ r^{l_0} + \sum_{t=1}^{m-1} r_{l_{t-1}}^{l_t}, & \text{inter - coding} \end{cases} \tag{3}$$

where r^{l_t} represents the number of bits spent to intra-code frame l_t, and $r_{l_{t-1}}^{l_t}$ the bits to inter-code frame l_t based on motion prediction from the frame l_{t-1}.

With these definitions we formulate the rate-distortion optimal video summarization problem as a constrained optimization problem of minimizing the summary distortion $D(S)$ subject to the bit rate constraint, that is, the MDOS (Minimum Distortion Optimal Summarization) formulation,

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

The minimization is over the number of frames m, and all possible summary frame locations $\{l_0, l_1, \ldots, l_{m-1}\}$.

We also consider the dual problem of minimizing the video summary bit rate $R(S)$ subject to the summary distortion constraint, or the MROS (Minimum Rate Optimal Summarization) formulation,

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

Notice that we have the implicit constraint that the frame selection for the summary is sequential in time, that is, $l_0<l_1<\ldots l_{m-1}$. We also assume that the first frame of the sequence is always selected, i.e, $l_0=0$.

3. SOLUTION TO THE MDOS PROBLEM

It is not feasible to solve the MDOS formulation (4) directly by exhaustive search, since the total number of possible summary choices is $\sum_{m=1}^{n}\binom{n-1}{m-1}$, which grows exponentially with the problem size. Instead, we first relax the constrained MDOS minimization problem with a Lagrangian multiplier [12], that is,

$$S_\lambda^* = \arg\min_S \{D(S) + \lambda R(S)\} \tag{6}$$

If there exists a λ^* such that $R(S_{\lambda^*}^*) = R_{\max}$, then the solution $S_{\lambda^*}^*$ is also the optimal solution to the original MDOS formulation (4), as shown in [12].

We further observe that the relaxed MDOS problem (6) has a certain built-in structure and can be solved in stages. For a given current state, the future solution is independent from the past solution. This structure will give us an efficient Dynamic Programming (DP) solution following [10][11].

Let the distortion state for the sequence segment starting with frame selection l_t and ending with the frame $l_{t+1}-1$ be,

$$G_{l_t}^{l_{t+1}} = \sum_{j=l_t}^{l_{t+1}-1} d(f_{l_t}, f_j) \tag{7}$$

The summary then including t frames and ending with the last frame selection $l_{t-1}=k$ has minimum distortion,

$$D_t^k = \min_{l_1,l_2,\cdots l_{t-2}} \{G_0^{l_1} + G_{l_1}^{l_2} + \cdots + G_{l_{t-2}}^{k} + G_k^n\} \tag{8}$$

and associated bit rate,

$$R_t^k = \sum_{j=0}^{t-1} b(f_{l_j}) \tag{9}$$

With Lagrangian relaxation, the objective becomes,

$$J_\lambda^{t,k} = \min_{l_1,l_2,\cdots l_{t-2}} \{D_t^k + \lambda R_t^k\} \tag{10}$$

For the summary with $t+1$ frames and $l_t=k$, we have

$$\begin{aligned} J_\lambda^{t+1,k} &= \min_{l_1,l_2,\cdots l_{t-1}} \{D_{t+1}^k + \lambda R_{t+1}^k\} \\ &= \min_{l_1,l_2,\cdots l_{t-1}} \{G_0^{l_1} + \cdots + G_{l_{t-1}}^k + G_k^n + \lambda[b(f_0) + b(f_1) \\ &\qquad + \cdots b(f_{l_{t-1}}) + b(f_k)]\} \\ &= \min_{l_1,l_2,\cdots l_{t-1}} \{G_0^{l_1} + \cdots + G_{l_{t-2}}^{l_{t-1}} + G_{l_{t-1}}^n - G_{l_{t-1}}^n + G_{l_{t-1}}^k + G_k^n \\ &\qquad + \lambda[b(f_0) + b(f_1) + \cdots b(f_{l_{t-1}})] + \lambda b(f_k)\} \\ &= \min_{l_1,l_2,\cdots l_{t-1}} \{G_0^{l_1} + \cdots + G_{l_{t-1}}^n + \lambda[b(f_0) + \cdots b(f_{l_{t-1}})] \\ &\qquad - [\underbrace{G_{l_{t-1}}^n - (G_{l_{t-1}}^k + G_k^n)}_{e^{l_{t-1},k}}] + \lambda b(f_k)\} \\ &= \begin{cases} \min_{l_{t-1}}\{J_\lambda^{t,l_{t-1}} - e^{l_{t-1},k} + \lambda r^k\}, & \text{if intra coding} \\ \min_{l_{t-1}}\{J_\lambda^{t,l_{t-1}} - e^{l_{t-1},k} + \lambda r_{l_{t-1}}^k\}, & \text{if inter coding} \end{cases} \end{aligned} \tag{11}$$

The "edge cost" $e^{l_{t-1},k}$ is the distortion difference if frame k is selected into the summary ending with frame l_{t-1}, given by,

$$e^{l_{t-1},k} = \sum_{j=k}^{n-1} [d(f_j, f_{l_{t-1}}) - d(f_j, f_k)] \quad (12)$$

We can split the minimization in (11) because the quantities $e^{l_{t-1},k}$, r^k and $r^k_{l_{t-1}}$ do not depend on the previous frame selections L_{t-2}= $\{l_{t-2}, l_{t-3}, \ldots l_0\}$. This is true for $r^k_{l_{t-1}}$ only if frame f_k is predicted from the original frame $f_{l_{t-1}}$. But in video coder implementations like H.263 [14] we use, the prediction is actually from the reconstructed frame $\hat{f}_{l_{t-1}}$, which can have multiple versions depending on L_{t-2}, and this introduces small variations in $r^k_{l_{t-1}}$. We can either force the prediction on $f_{l_{t-1}}$, or more practically, by using a constant PSNR coding strategy, to keep the variation of $\hat{f}_{l_{t-1}}$ low such that the variance in $r^k_{l_{t-1}}$ is negligible.

The initial condition is given as,

$$J^{1,0}_{\lambda} = \{G^n_0 + \lambda r^0\} \quad (13)$$

From (11)-(13) we established the recursion needed for the Dynamic Programming (DP) solution. The algorithm will build the trellis with this recursion starting from frame f_0, it will add more frames to the summary, and stop when the last (virtual) frame f_n is reached. For each epoch t, the final node is computed as,

$$J^{t,n}_{\lambda} = \min_{k \in F_{t-1}} \{J^{t-1,k}_{\lambda}\} \quad (14)$$

where F_{t-1} is the feasible frame set at epoch $t-1$ that can have transition to the last (virtual) frame f_n. The optimal solution for the relaxed MDOS problem (6) with a particular λ is therefore found by selecting the smallest $J^{t,n}_{\lambda}$ and backtracking for the optimal summary [15].

The operational rate-distortion function is non-increasing and actually convex in most cases. It is known that the Lagrangian multiplier λ is the inverse slope of the operational rate-distortion function convex hull. As λ goes from zero to infinity, the solution of the problem in (6) traces out the convex hull of the operational rate-distortion curve. The solution to the original MDOS problem, $S^*_{\lambda^*}$ is therefore found by a bi-section on λ.

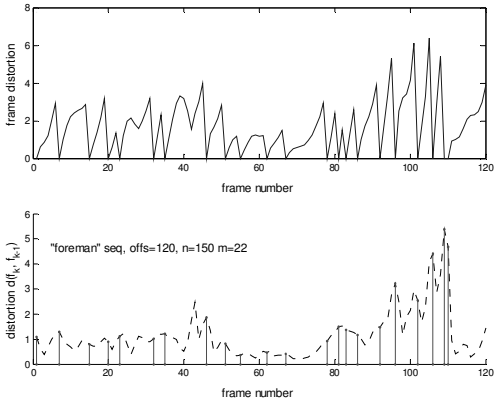
The DP solution to the relaxed problem (6) has complexity of $O(n^2)$. The bi-section search on λ is efficient because the edge and bit costs in the recursion (11) do not change as λ changes, therefore need only be computed once in the bi-section search loop.

4. SOLUTION TO THE MROS PROBLEM

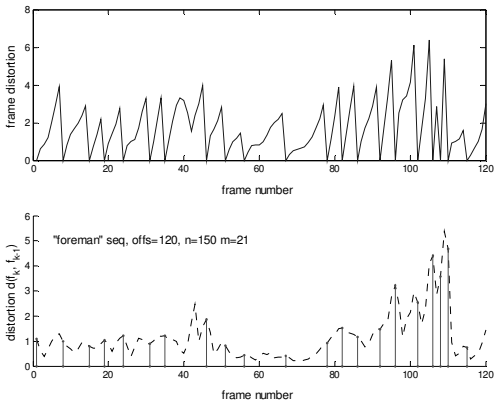
With the Lagrangian relaxation in (6), the MROS problem can be similarly solved by a bi-section search on λ. Let λ^* be the target value such that $D(S^*_{\lambda^*}) = D_{\max}$, then the solution $S^*_{\lambda^*}$ is also the optimal solution to the original MROS formulation (5), as shown in [12].

5. EXPERIMENTAL RESULTS

We implemented the proposed DP algorithm with Lagrangian relaxation for the summarization. For the MROS formulation with a distortion constraint of D_{max}=205, the summarization results for the "foreman" sequence frames 150-270 are shown in Fig 1a for intra coding and Fig 1b for inter coding cases. We use the TMN8 H.263 video coder [14] with fixed QP=10 for the encoding.



(a) inter-coded results



(b) intra-coded results

Figure 1. Summarization Results

In both Fig 1a and Fig 1b, the upper plot is the frame distortion introduced by the summary. Notice that it goes to zero at time location where a summary frame is selected into the summary. The lower plot is the differential frame-by-frame distortion $d(f_k, f_{k-1})$ plot with the summary frame

selections indicated by vertical lines. For the inter coded case in Fig 1a, the rate $R(S)$= 23972 bits, the distortion $D(S)$= 201.8. For the intra coded case in Fig 1b, the rate $R(S)$=61456 bits, the distortion $D(S)$= 202.9. To achieve roughly the same distortion level in the summarization, inter coding uses only 1/3 of bits of that of the intra coding case.

The proposed formulation does not depend on a specific frame distortion metric, however, the success of the summarization clearly depends on the frame distortion metric. One that correlates well with people's perception would be highly desireable. In our implementation, we first reduce the original frame to a desired scale of w x h by repeated low pass filtering and down-sampling process (denoted by D). Then a pre-trained Principal Component Analysis (PCA) transform A is applied to further reduce the w x h frame to a d dimensional vector in the principal component space. The distortion between two frames is therefore computed as the $l2$ distance between two principal component space vectors,

$$d(f_j, f_k) = \| AD_{wxh}(f_j) - AD_{wxh}(f_k) \| \qquad (15)$$

In our implementation we choose 11x9 scale and d=6 dimensions. The effectiveness of this metric is demonstrated by the differential frame-by-frame distortion plot in Fig.1 for the "foreman" sequence, which captures the frame-by-frame changes in that sequence quite well.

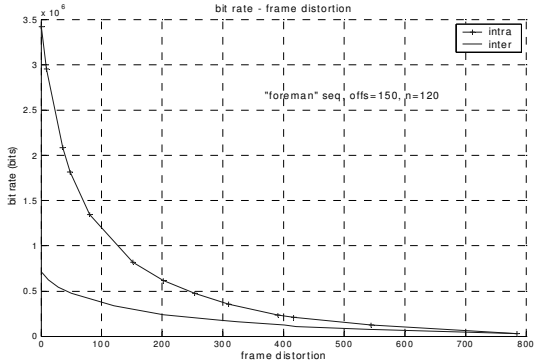


Figure 2. Rate-Distortion Curves

The overall rate-distortion performance for the above mentioned test sequence is plotted in Fig.2 for both the inter and intra coded cases. A subjective evaluation of the results can also be found at:
http://ivpl.ece.northwestern.edu/~zli/new_home/demo/minavg/minavg.html.

6. CONCLUSION AND FUTURE WORK

In this paper we formulated the optimal video summarization problem as a rate-distortion optimization problem and presented the optimal DP solution based on Lagrangian relaxation. The experimental results demonstrated the effectiveness and efficiency of the proposed approach, which can therefore be employed in a variety of real world applications, like SDTV/HDTV program summarization and trans-coding for the mobile users. Work is underway to expand the framework to handle the MINMAX criterion for the summary distortion, and investigate optimal as well as practical solutions to this formulation.

7. REFERENCES

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

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

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

[5] 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.

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

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

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

[9] 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.

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

[11] 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.

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

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

[14] TMN8, the software H.263 video coder, the University of British Columbia implementation.

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

[16] 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.