RATE-DISTORTION OPTIMAL VIDEO SUMMARIZATION: A DYNAMIC PROGRAMMING SOLUTION

*+Zhu Li, #Guido M. Schuster, *Aggelos K. Katsaggelos, and +Bhavan Gandhi

+Multimedia Communication Research Lab (MCRL), Motorola Labs, Schaumburg
*Department of Electrical & Computer Engineering, Northwestern University, Evanston
#Hochschule für Technik Rapperswil, Switzerland

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. Our work is based on a temporal rate-distortion optimization formulation for optimal summary generation. New metrics for video summary distortion are introduced. Optimal algorithms based on dynamic programming are presented along with the results from heuristic algorithms that can produce near optimal results in real time.

1. INTRODUCTION

The demand for video summary work originates from a viewing time constraint 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, and then selecting "key frames" according to some criterion from each video shot to generate video summary for the sequence. Examples of past works are listed in [1]-[7], [14]-[16]. For the approaches mentioned above, various visual features and their statistics have been computed to identify video shot boundaries and determine key frames by thresholding and clustering. In general 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 temporal rate-distortion optimization problem. Temporal rate is the ratio of the number of frames in the video summary versus that of the original sequence. We assume that all the information is presented by the frames included in the summary and the temporal distortion is introduced by the missing frames. We introduce a new frame distortion metric and the temporal distortion is then modeled as the average (or equivalently total) frame distortion between the original and the reconstructed sequences. A dynamic programming solution that find the optimal solution is presented.

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 $S = \{f_{l_0}, f_{l_1}, \cdots f_{l_{m-1}}\}$, 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,

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

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

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

The summary temporal rate is defined as the ratio of the number of frames selected into the video summary versus that of the total frames in the original sequence,

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

Notice that the temporal rate is in the range of (0,1] and can only take values from a discrete set $\{1/n, 2/n, \ldots, 1\}$. For example, for the video sequence $V = \{f_0, f_1, f_2, f_3, f_4\}$, and its video summary $S = \{f_0, f_1\}$, the temporal rate R =0.4. The temporal distortion computed from (2) is $D(S) = \frac{1}{5}[d(f_0, f_1) + d(f_2, f_3) + d(f_2, f_4)]$.

With these definitions we can formulate the temporal rate-distortion optimal video summarization problem as a constrained optimization problem of minimizing the summary distortion $D(S)$ subject to the temporal 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 actually over the number of frames m, and all possible summary frame locations $\{l_0, l_1, \ldots, l_{m-1}\}$.

On the other hand we also consider the dual problem of minimizing the video summary temporal 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

To solve the MDOS formulation (4) directly by exhaustive search will not be feasible. The problem complexity grows exponentially with the sequence size. Instead, we observe that the problem 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 inspired by [12][13].

Let the distortion state D_t^k be the minimum distortion incurred by the summary that has t frames and ended with frame k,

$$D_t^k = \min_{l_1, l_2, \cdots l_{t-2}} \sum_{j=0}^{n-1} d(f_j, f_{i=\max(l):\, s.t.\, l \in \{0, l_1, \cdots l_{t-2}, k\}, i \le j}) \tag{6}$$

Notice that $l_0 = 0$ and $l_{t-1} = k$ and they are removed from the optimization. From (6) we have,

$$\begin{aligned} D_t^k &= \min_{l_1, l_2, \cdots l_{t-2}} \{\sum_{j=0}^{k-1} d(f_j, f_{i=\max(l):\, s.t.\, l \in \{0, l_1, \cdots l_{t-2}\}, i \le j}) \\ &\qquad + \sum_{j=k}^{n-1} d(f_j, f_k)\} \\ &= \min_{l_1, l_2, \cdots l_{t-2}} \{\sum_{j=0}^{k-1} d(f_j, f_{i=\max(l):\, s.t.\, l \in \{0, l_1, \cdots l_{t-2}\}, i \le j}) \\ &\qquad + \sum_{j=k}^{n-1} d(f_j, f_{i=\max(l):\, s.t.\, l \in \{0, l_1, \cdots l_{t-2}\}, i \le j}) \\ &\qquad + \sum_{j=k}^{n-1} d(f_j, f_k) \\ &\qquad - \sum_{j=k}^{n-1} d(f_j, f_{i=\max(l):\, s.t.\, l \in \{0, l_1, \cdots l_{t-2}\}, i \le j})\} \end{aligned} \tag{7}$$

We now observe that,

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

Therefore the distortion state D_t^k in (7) can be broken into two parts,

$$\begin{aligned} D_t^k = \min_{l_1, l_2, \cdots l_{t-2}} \{&\sum_{j=0}^{n-1} d(f_j, f_{i=\max(l):\, s.t.\, l \in \{0, l_1, \cdots l_{t-2}\}, i \le j}) \\ &\underbrace{- \sum_{j=k}^{n-1} [d(f_j, f_{l_{t-2}}) - d(f_j, f_k)]}_{e^{l_{t-2}, k}}\} \end{aligned} \tag{9}$$

where the first part is the problem of minimizing the distortion for the summary with $t-1$ frames ending with frame l_{t-2}, while the second part of the minimization 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}. Now we have,

$$D_t^k = \min_{l_{t-2}}\{ \min_{l_1,l_2,\cdots l_{t-3}} \{\sum_{j=0}^{n-1} d(f_j, f_{i=\max(l):s.t.l\in\{0,l_1,\cdots l_{t-2}\},i\le j})\} - e^{l_{t-2},k}\} \quad (10)$$
$$= \min_{l_{t-2}}\{ D_{t-1}^{l_{t-2}} - e^{l_{t-2},k}\}$$

This last relation established the recursion we need for the DP solution. Since we always select the first frame into the summary, the initial state D_1^0 is given as,

$$D_1^0 = \sum_{k=1}^{n-1} d(f_0, f_k) \quad (11)$$

As an example, the trellis for the *n-m* video summary problem, i.e., generate an *m*-frame summary from an *n*-frame sequence, with *n*=5 and *m*=3 is illustrated in Fig.1,

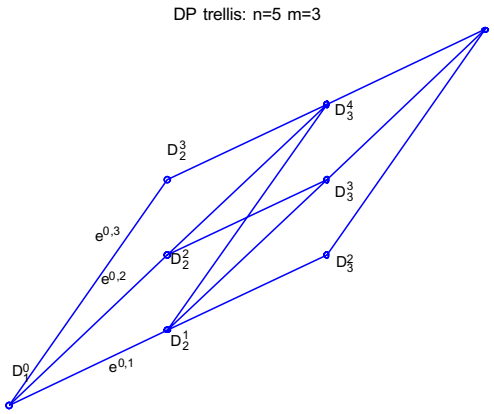

Figure 1. DP trellis for n=5, m=3.

Note that the topology of the trellis is completely determined by the parameters *n* and *m*. As in the Fig 1., node D_2^4 is not included in it since *m*=3 and therefore f_4 (the last frame in the sequence) cannot be the second frame in the summary. In addition a skip constraint can be utilized to constraint the maximum number of frames that can be skipped between any two frames in the summary. Clearly such a constraint further changes the topology of the trellis.

The minimum distortion is min $\{D_m^k\}$ for all feasible *k* at the final stage *m*. (*m*=3 in this example.)The optimal frame selection can be found by backtracking the state trellis similarly to the Viterbi algorithm [15].

The resulting summary has the minimum distortion for the temporal rate of *m/n*. We define the operational distortion-rate function as,

$$D^*(m/n) = \min_{l_1,l_2,\cdots l_{m-1}} (1/n)\sum_{j=0}^{n-1} d(f_j, f_j') \quad (12)$$

Notice that the distortion-rate function is non-increasing with *m*. This is true because adding a frame to the summary problem in (12) always reduces the distortion or at least keeps it the same. The solution to the original MDOS formulation is to find the maximum integer *m* that satisfies the rate constraint R_{max}, and solve the *n-m* summary problem by computing the *n-m* distortion state trellis using the recursion in (10) and backtracking for the optimal frame selection $\{l_0, l_1,\ldots, l_{m-1}\}$.

The computational complexity of the DP solution to the *n-m* summary problem in terms of edge cost evaluation is *(1/2)(m-2)(n-m+1)(n-m+2)+2(n-m+1),* or O(n^2).

4. SOLUTION TO THE MROS PROBLEM

For the RMOS formulation, the optimal solution can be found by a bi-section search on the operational distortion-rate function $D^*(m/n)$.

We start with an initial rate bracket of $R^{lo}=1/n$ and $R^{hi} = n/n$. If the distortion constraint $D_{max} < D^*(R^{hi})$, then there is no feasible solution to the RMOS problem because the distortion constraint is too low. If $D_{max} > D^*(R^{lo})$, then the rate *1/n* is the optimal solution. Otherwise we select a middle point $R_{new} = \left\lfloor \frac{R^{hi}+R^{lo}}{2} \right\rfloor$, compute its associated distortion $D_{new}= D^*(R_{new})$, and find the new rate-bracket by replacing either R^{lo} or R^{hi} with R_{new}, such that the distortion constraint D_{max} is within the new distortion bracket $[D^*(R^{hi}), D^*(R^{lo})]$. The process will continue until the rate bracket boundaries converge. At this point the optimal solution to the MROS problem is found.

Since the feasible rate set is discrete and finite, this algorithm always converges.

5. EXPERIMENTAL RESULTS

The DP algorithm to the *n-m* optimal summary problem was implemented and ran for a number of sequences. For the frame distortion, various distortion metrics can be used for computing $d(f_j, f_k)$. In the reported experiments, we uses the weighted Euclidean distance of the frames in the principle component space similarly to the Color Layout metric [10][11].

As an example, the optimal summary generation for the "foreman" sequence, frames 150-270, with *n*=120 and *m*=24 is shown in Fig. 2. The upper part is the frame distortion introduced by the (*120-24*) optimal summary. Notice that the distortion goes to zero at the frame locations included in the summary. For this case, the average distortion is *14.53*, the max distortion is *50.00*, and the distortion variance is *11.17*. The lower part is the optimal summary frame selection in vertical lines plotted against the frame-by-frame distortion $d(f_k, f_{k-1})$ using the Color Layout metric.

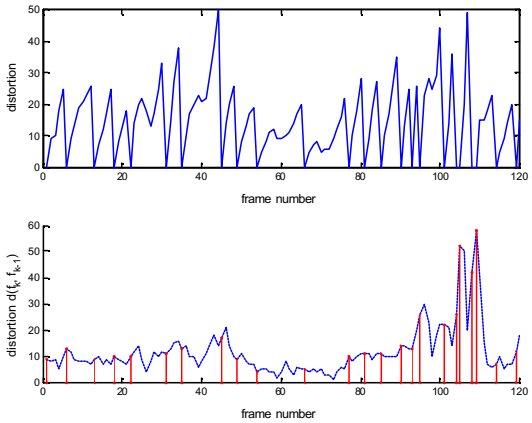

Figure 2. Frame selection and frame distortion

The operational distortion-rate function for the same sequence is plotted in Fig.3. It is convex as expected. The solution from a heuristic Greedy algorithm [8] is also plotted as a comparison. The Greedy algorithm selects frames for the summary iteratively until the frame budget is exhausted. At each iteration step, the frame that introduces the largest distortion is selected into the summary.

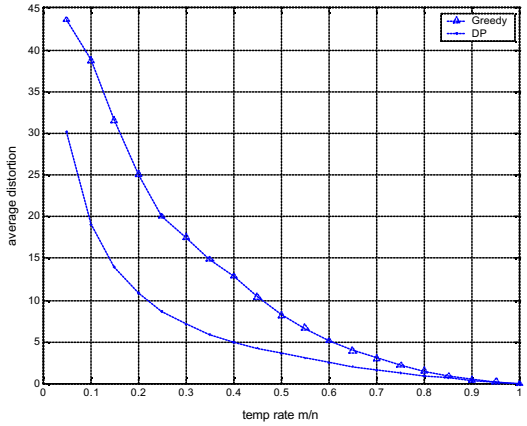

Figure 3. The operational distortion-rate functions

6. CONCLUSION AND FUTURE WORKS

In this paper we formulated the optimal video summarization problem as a rate-distortion optimization problem and presented the optimal solutions to the MDOS and MROS formulations. The experimental results demonstrated the effectiveness and efficiency of the proposed approach, which can therefore be employed in a variety of real world applications.

Work is underway to expand the framework to include the max frame distortion metric and address the issue of optimal coding of the summary.

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] ____, Information Technology – Multimedia Content Description Interface Part 3: Visual, ISO/IEC FCD 15938-3.

[11] 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..

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

[13] 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.

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

[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.