MINMAX OPTIMAL VIDEO SUMMARIZATION

[+*]*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 MINMAX optimization formulation for the optimal summary generation. New metrics for video summary distortion are introduced. Optimal algorithm based on dynamic programming is presented along with the results from a heuristic algorithm 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 frame distortion metric and the temporal distortion is then modeled as the 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 sequence distortion introduced by the summary is given by,

$$D(S) = \max_{j\in[0,n-1]} d(f_j, f_j') \tag{2}$$

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,

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

Notice that the temporal rate is in the range of (0,1] and can only take values from a discrete set $\{1/n, 2/n, \ldots, 1\}$.

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 MROS PROBLEM

To solve the MDOS formulation (4) directly by exhaustive search will not be feasible, since the total number of possible choices is $\binom{n-1}{m-1}$, which grows exponentially with the problem size. Instead, we observe that the MROS problem (5) 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 for the sequence segment started with the frame selection l_t and ended with the frame $l_{t+1} - 1$ be,

$$D_{l_t}^{l_{t+1}} = \max_{j\in[l_t,\, l_{t+1}-1]} d(f_{l_t}, f_j) \tag{6}$$

Let the rate of this sequence segment be,

$$R_{l_t}^{l_{t+1}} = \begin{cases} r(f_{l_t}) = 1, & if\ D_{l_t}^{l_{t+1}} \le D_{\max} \\ \infty, & otherwise \end{cases} \tag{7}$$

Then the MROS problem in (5) is equivalent to the unconstrained problem of,

$$\min_{l_1,l_2,\cdots l_{m-1}} \{\sum_{t=0}^{m-1} R_{l_t}^{l_{t+1}}\} \tag{8}$$

The problem (8) can be computed recursively. Let the minimum rate for the video segment ended with the summary frame choice l_t be,

$$\begin{aligned} J_{l_t} &= \min_{l_1,l_2,\cdots l_{t-1}} \{R_0^{l_1} + R_{l_1}^{l_2} + \cdots + R_{l_{t-1}}^{l_t}\} \\ &= \min_{l_1,l_2,\cdots l_{t-1}} \{\sum_{j=0}^{t-1} R_{l_j}^{l_{j+1}}\} \end{aligned} \tag{9}$$

then for the video segment ended with the summary frame choice l_{t+1}, we have the minimum rate,

$$\begin{aligned} J_{l_{t+1}} &= \min_{l_1,l_2,\cdots l_t} \{R_0^{l_1} + R_{l_1}^{l_2} + \cdots + R_{l_{t-1}}^{l_t} + R_{l_t}^{l_{t+1}}\} \\ &= \min_{l_1,l_2,\cdots l_t} \{\sum_{j=0}^{t-1} R_{l_j}^{l_{j+1}} + R_{l_t}^{l_{t+1}}\} \\ &= \min_{l_t} \{J_{l_t} + R_{l_t}^{l_{t+1}}\} \end{aligned} \tag{10}$$

This gives us the recursion we need to compute the solution trellis for a Viterbi algorithm [15] like optimal solution. The initial condition is given by,

$$J_{l_1} = \begin{cases} 1, & if\ \ D_0^{l_1} \le D_{\max} \\ \infty, & else \end{cases} \tag{11}$$

The recursion started with the frame node f_0 and expand over all frames that introduce admissible segment distortion. A full trellis example for $n=6$ is shown in Fig. 1. Notice that the edges between any frame pair f_j and f_{j+p} is admissible only if

$$\max_{i\in[0,p-1]} \{d(f_j, f_{j+i})\} \le D_{\max} \tag{12}$$

The algorithm will build the trellis with constraint (12) from frame f_0, and stop when the last frame f_n is first reached. The optimal rate is thus the epoch+1 at f_n, and the optimal frame selection is obtained via backtracking. Notice that we have,

$$D_j^k \le D_j^{k+p},\ for\ j \le k \tag{13}$$

which means for all edges originate from node f_j, if some frame f_k is not admissible, then all frames f_{k+p} are not admissible either. This can be used to prune the trellis.

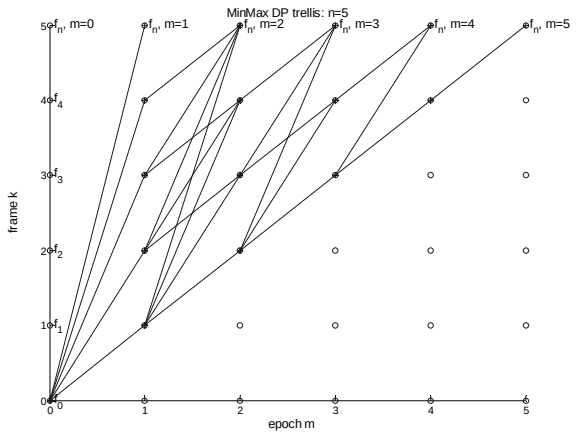

Figure 1. MinMax trellis example

The optimal solution is not unique, as indicated by an example in Fig.2. It is the optimal solution trellis for the "foreman" sequence, frames 10~18, with a max distortion constraint D_{max} = 18. The optimal rate is $m=4$ for this case, yet the optimal frame selections can be $\{f_0, f_3, f_7\}$, $\{f_0, f_2, f_7\}$, $\{f_0, f_2, f_6\}$, …, etc. Additional constraint like the variance of frame skip steps can be imposed to ensure smoother play back of the summary.

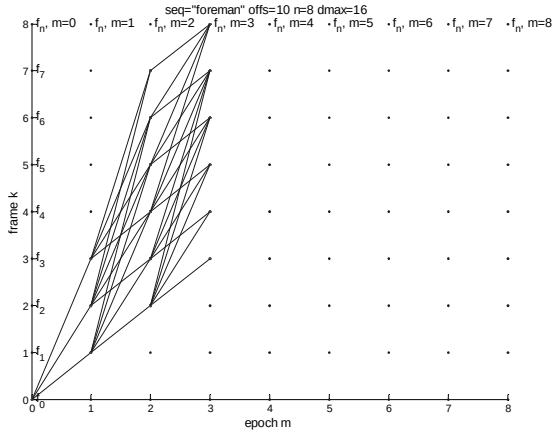

Figure 2. Solution trellis example

Notice that a greedy Distortion Constrained Skip (DCS) solution exists for this particular case in Fig.2, that is the solution $\{f_0, f_3, f_7\}$, which is the inner most path of the trellis. The DCS algorithm operates as follows,

```
L=0, add fL to the summary S
FOR k=1 TO n
  IF d(fL, fk) > Dmax
       L=k, add fL to the summary S
  END
END
```

It skips all frames that introduce acceptable distortion. The DCS algorithm is optimal if the following condition holds,

$$D_{j+p}^{k} \le D_{j}^{k}, \ for\ j+p \le k \tag{14}$$

The condition (14) requires that a shorter sub-segment of the sequences introduce smaller maximum distortion than the longer one with the same last frame. This is true for most natural video sequences. The DCS algorithm is a much faster one-pass solution than the DP algorithm, and will be optimal if (14) holds. Even though (14) may not hold for some sequences, the performance penalty is acceptable. This makes the DCS algorithm an attractive practical alternative for on-line applications like SDTV trans-coding for mobile users.

4. SOLUTION TO THE MDOS PROBLEM

Let the operational max distortion-rate function be,

$$D^*(R) = D^*(m/n) \tag{15}$$

where m^* is the number of frames in the optimal solution of (4). Then we have,

Lemma 1. $D^*(m/n)$ is non-increasing with m.

Proof: Let the frame selections $L=\{0, l_1, l_2,\ldots l_{m-1}\}$ be the optimal solution to (4) with the distortion constraint D, then we can find a frame selection $t^* = \arg\max_{t\in[1,m-1]} \{d(f_{l_{t-1}}, f_{l_t-1})\},\ s.t.\ l_t - 1 \notin L$. Let the new $m+1$ frame selection be $L'=L+\{l_{t^*}-1\}$, then the new distortion $D' < D$, if $d(f_{l_{t^*-1}}, f_{l_{t^*}-1}) < D$ is the $\max_{j\in[0,m-1]} \{d(f_j, f_j')\}$. Otherwise $D'=D$. Let the optimal distortion for $R=m+1/n$ be $D^*(m+1/n)$, then we have,

$$D^*(m+1/n) \le D' \le D^*(m/n) \tag{16}$$

and by the chain rule we have $D^*(m/n)$ is non-increasing with m. With Lemma 1, we also know that the operational distortion-rate function $R^*(D)$ is non-increasing. This gives us a bi-section search solution for the MDOS problem.

For the MDOS formulation, the rate constraint is given as $R_{max}=m_0/n$. We solve the MDOS problem by a bi-section search on feasible distortion thresholds. We start with an initial max distortion bracket of $[D^{lo}=0, D^{hi}]$ and initial rate bracket $[R^{lo}=n/n, R^{hi}]$ such that m_0/n is in the initial rate bracket. Then a new distortion middle point is computed $D_{new}=(D^{lo}+D^{hi})/2$, solve for its optimal rate $R_{new}=m_{new}/n$ with the DP algorithm, and find the new rate bracket by replacing either R^{lo} or R^{hi} with the R_{new}, such that the rate constraint R_{max} is within the new rate bracket. Then replace the distortion bracket with corresponding distortion pair $[D^{hi}, D^{lo}]$. The process will continue until the rate bracket boundaries converge to R_{max}. At this point the optimal solution to the MDOS problem is found.

Since the feasible rate set is discrete and finite, this algorithm always converges. The complexity of the MDOS solution is $O(\log(n))$ times the complexity of MROS problem.

5. EXPERIMENTAL RESULTS

For the frame distortion, various distortion metrics can be used for computing $d(f_j, f_k)$. MSE type metric is an obvious choice, but it does not reflect the subjective perception well. Color and texture features [10][11] are good in image differentiation under certain semantic assumption, i.e, in which feature are we comparing the images, but not very effective as a distortion metric for the summarization problem. In our experiments, we use a metric based on the Euclidean distance of the scaled frames in the principle component space. The benefit of the scaling and PCA processes is to reduce noise and local variance not relevant to human perception.

As an example, the optimal summary generation for the "foreman" sequence, frames 150-270, with n=120 and D_{max}=36 is shown in Fig. 3.

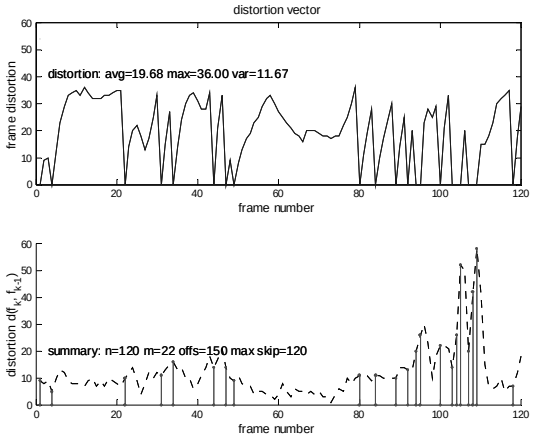

Figure 3. Frame selection and frame distortion

The upper part is the frame distortion $d(f_k, f_k')$ introduced by the MINMAX optimal summary. Notice that the distortion goes to zero at the frame locations included in the summary. The bottom plot is the summary frame selection plotted as vertical lines against the dotted curves of the frame-by-frame distortion of the sequence, which gives an indication of the activity within the sequence. For this case, the resulting max distortion is *36.00*, and average distortion is *19.68*, and the distortion variance is *11.67*.

6. CONCLUSION AND FUTURE WORKS

In this paper we formulated the optimal video summarization problem as a rate-distortion MINMAX optimization problem and presented the optimal DP solution, as well as the near-optimal DCS solution 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 coding cost as an additional constraint 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] ____, 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.