Redundancy removing by adaptive acceleration and event clustering for video summarization

Emilie Dumont and Bernard Mérialdo
Institut Eurécom
2229 route des Crêtes
06904 Sophia Antipolis, FRANCE
{Emilie.Dumont,Bernard.Merialdo}@eurecom.fr

Abstract

In this paper, we propose a novel approach to summarize rushes. Our processing is composed of several steps. First, we remove unusable content and we dynamically accelerate video according to motion activity to maximize the content per time unit. Then, one-second video segments are clustered into similarity clusters. The most important non-redundant pieces of shot are selected such that they maximize the coverage of those similarity clusters. The produced summaries have been evaluated by an automatic method with a strong positive correlation with the TRECVID campaign evaluation.

1. Introduction

With rapid advances in the technology of digital video documents and although powerful technologies now exist to create, play, store and transmit those documents, the analysis of the video content is still an open and active research challenge. In this paper, we focus on video summarization. The automatic creation of video summaries is a powerful tool which allows synthesizing the entire content of a video while preserving the most important or most representative sequences. A video summary will enable the viewer to quickly grab the essence of the document and decide if it is useful for its purpose or not.
Over the last number of years, various ideas and techniques have been proposed towards the effective summarization of video contents. Overviews of these techniques appear in [5], [10], [7]. In [1], [2], [11] redundancy is removed via clustering : a maximum of one shot is retained from a cluster of visually similar shots. Authors in [6], [9] and [3] compute elements such as color contrast, intensity contrast, and orientation contrast to model the human attention level to a particular image. Visual features, in particular color histograms, are often used to measure the similarity between frames or shots, for example authors in [4] remove redundancy by selecting only contiguous frames that maximize the average similarity to a video. In [8], seqnences are selected wihtout redundancy by a model for video content inspired by the TF/IDF model for natural language.
Rushes contain many frames or sequences of frames that are highly repetitive, e.g., many takes of the same scene redone due to errors (e.g. an actor gets his lines wrong, a plane flies over, etc.), long segments in which the camera is fixed on a given scene or barely moving, etc. A significant part of the material might qualify as stock footage - reusable shots of people, objects, events, locations, etc. So, after rushes cleaning by removing junk frames and temporal redundancy, we propose to make a selection of the most relevant segments by maximizing non-redundant information. We begin the selection process by partitioning the video into one-second segments, then we cluster the one-second segments with an agglomerative hierarchical clustering approach. The clustering stops at a threshold which is adapted to the video, based on a measure of quality for the available clusters. Finally, the clusters are used to compute a relevance score for each segment and select a set of relevant segments to be included in the summary. The major steps are illustrated in figure 1.

2 Irrelevant frames removal

Since rushes are raw material used to produce a video, they contain a significant part of uninteresting sequences. For example, rushes contain many frames or sequences of frames that will not be used to produce the final video like test pattern frames, black frames, movie clapper board frames, etc, see figure 2. Those junk sequences of frames are removed in an initial prepossessing step.

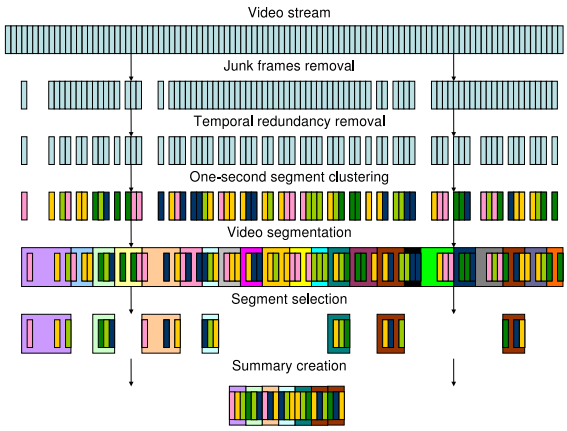

Figure 1. General Scheme

Rushes contain also many frames or sequence of frames that are highly temporal redundant, for example long segments in which the camera is fixed on a given scene or barely moving, etc. So, we perform a dynamically acceleration according to motion activity to remove this temporal redundancy.

Figure 2. Junk frames

2.1 Junk frames

Test pattern frame A test pattern frame is very particular: it is composed of stripes with various colors. Those frames generally have always the same presentation. To detect them, we use a training set of test pattern frames. For each frame in the training set, we extract a hue histogram, and we average histograms of all training frames to build a detector vector T. To remove test pattern frames in a video, we compare frame vectors with the vector T by Euclidean distance and we remove frames with a distance larger than a predefined threshold.

Uniform color frame A second style of junk frame is uniform color frame. To detect them, we compute the entropy of the distribution of color pixels in HSV color space and we remove frames with an Euclidean distance larger than a predefined threshold.

2.2 Temporal redundancy

The gap between rush shot duration and movie shot duration is high : in rush, a landscape shot may last several few minutes, but a fight shot may last just a few seconds. The idea of dynamic acceleration is to show a sequence during a time proportional to its motion activity.
We compute the motion activity $activity(f)$ for each frame f. So we can compute a jump threshold for a video v by

$$jump(v) = \sum_{f \in v} activity(f)/F * acc$$

Where acc is the mean acceleration for entire video, and F the number of frame in v. We keep only one frame every $jump(v)$ activity.

MRS043400	MRS157443	MS212920	MRS035133	MRS150148	MRS157475	MS237650
0.001	0.071	0	0.141	0.049	0.012	0.154
0.013	0.001	0.011	0	0.017	0	0.029
0.312	0.26	0.32	0.19	0.27	0.32	0.15

Table 1. Percentage of irrelevant frames
The first line is the percentage of frames detected like test pattern frame, the second the percentage of frames detected like uniform color frame and the third the percentage of frames selected for next steps.

3 Segment selection

After rushes prepossessing, we perform a selection of the most relevant segments. The idea is to select non-redundant sequences of frames, whose content overlaps as little as possible.We divide the video stream into one-second second segments, and we cluster those segments by agglomerative hierarchical clustering. And finally, we select a set of sequences which covers a maximum of visual interest.

3.1 Hierarchical clustering

In order to detect the visual redundancy of video, we partition video into one-second segment, e.g. into 25 frames. Each one-second segment is represented by a HSV histogram of those frames. The algorithm starts with as many clusters as there are one-second segments, then at each step of the clustering, the number of clusters is reduced by one by merging the closest two clusters, until all segments are finally in the same cluster. The distance between two one-second segments is computed as the Euclidean distance, and the distance between two clusters is the average distance across all possible pairs of segments of each cluster.

3.2 Visual interest

The weight is intuitively related to the importance of the content of a cluster, as :

- the appearance of people. A face detector is used, for each one-second segment s, we extract the normalized average of the face number $face(s)$
- the activity. For each one-second segment s, we compute the normalized average activity $act(s)$
- the color. For each one-second segment s, we compute the normalized entropy of color $ent(s)$

For each visual attribute, we affect a weight represented its importance : Wf, $Wact$ and $Went$ fixed manually. So, we can compute a visual attribute coefficient for each cluster c by :

$$Vatt(c) = \sum_{s \in c} (face(s) * Wf + act(s) * Wact + ent(s) * Went)$$

If we define $0 \leq \alpha \leq 1$ as the weight of the visual interest, we can compute the visual interest of a cluster by :

$$Vint(c) = \alpha * \frac{Vatt(c)}{(Wf + Wact + Went)} + (1 - \alpha)$$

3.3 Segment selection

Each iteration of the algorithm provides a different clustering of one-second segments. The idea is to choose the clustering level which best represents the visual redundancy of the video. We want to choose a level where each cluster contains only similar segments and all similar segments are in the same cluster. A segment is just a predefined number of successive one-second segments.
For each level, we select the most important and non redundant segments for the summary by an iterative algorithm. The weight of a segment is defined as the sum of the weights of the clusters it contains, and have not yet been selected. We iteratively select the most important segment, and mark its clusters as selected. This process is repeated until all clusters have been selected.
And finally, we select the clustering level with the duration the closest to the required percentage.

4 Results and discussion

4.1 Automatic evaluation

In the development of video summarization systems, one of the main problem remains the method of evaluation, every system presents different and not comparable performances. A new evaluation campaign proposes to solve this problem : the TRECVID 2007 BBC rushes summarization evaluation pilot [7]. A human judge was given the summary, and a chronological list of up 12 topics from a ground truth. The assessor viewed the summary and determined which topics are presented, so the percentage of topics found by assessor determine the main measure : IN. This manual evaluation was evaluated in [7], and in conclusion they found that the agreement between assessor looks high by looking pairs of assessors in their binary judgments of individual included topic for a given summary.
This evaluation of summaries is presently manual. Then, this manual method has a number of disadvantages, in particular, the difficulty to reproduce experiments with other data. So in [1], and in [2], authors propose to evaluate IN by an automatic process : if at least 25 frames (one second) of the topic is in the summary, so the system considers the topic found. In the second paper, authors showed a strong correlation between manual evaluation and automatic evaluation. So, to experiment, we choose to use IN like the recall and the precision is computed by the number of topics found in the summary divided by the number of segments selected.

4.2 Experimental results

We experimented our method on 7 videos proposed by BBC Rushes Task 2007 in TRECVID. It consists of unedited video footage, shot mainly for five series of BBC drama programs. One of the main parameter of our process is the number of one-second segments contained in a segment. This length represents, intuitively, the minimal length of a topic. If this value is small, summary will be a sum of very short element and the summary will be unpleasant to show, while if this value is big, summary will be a sum of only one element, so pleasant to show, but with only one topic. Curve 3 shows a segment length equals to 4 seconds is the best trade-off to have a high recall and a high precision in the same time. We study the optimum length for the summary, in other words which is the ideal percentage of initial video to show most topics and least redundancy. The curve 4 shows a length of 0.03 percent of the initial video is optimum.

Now we know optimum size for the summary length and for the ideal number of one-second segments to compose a segment, we compare our summaries with baselines and eurecom summaries proposed during TRECVID 2007 with a same summary length equals to 0.04%, see table 2. We show our method selects less topics than baselines but the summary contents is more relevant and, clearly, this method is better than the one proposed in TRECVID 2007 (eurecom in the table 2) whereas eurecom recall were goods (and precision was not evaluated). In Eurecom system, video is segmented into shots, and the most important and

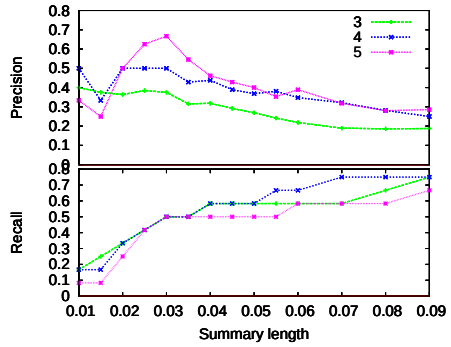

Figure 3. Recall-Precision for MRS150148
For segment length equals to 3, 4 and 5 seconds, we compare the recall/precision values for different summaries. Summary length is the percentage of the summary length in relation to the initial video length.

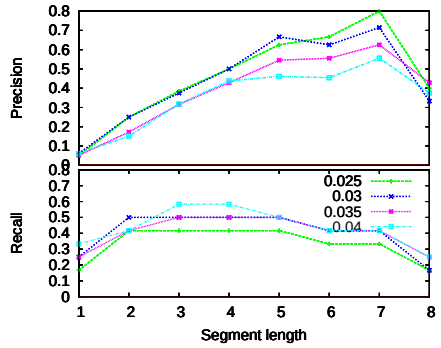

Figure 4. Recall-Precision for MRS150148
For summary length equals to 0.025, 0.03 , 0.035 and 0.04, we compare the recall/precision values for different summaries. Segment length is the number of one-second segments used to compose a segment.

non-redundant shots were selected for inclusion in the summary. During presentation in the summary, shots were dynamically accelerated according to the motion activity, and presented using a split-screen display, through this did not rate highly with assessors because to present 4 shots in the same time, we perform a too fast acceleration.

	baseline1	baseline2	eurecom	our method			
				α=0.5 Wf=0.5 Went=0.5 Wact=0.5	α=0 Wf=1 Went=0 Wact=0	α=0 Wf=0 Went=1 Wact=0	α=0 Wf=0 Went=0 Wact=1
recall	0.63	0.63	0.4	0.37	0.38	0.38	0.4
precision	0.11	0.13	0.06	0.13	0.15	0.13	0.14

Table 2. Comparison of different methods

5 Conclusion

Video summarization is a process of removing redundant contents and selecting video segments to create a condensed video. Our process begins by a removing of junk frames, and temporal redundancy. A one-second segments clustering leads the video segment selection, in order to create a summary with the most relevant and non-redundant segments. Experimentation show the optimum values for parameters, and we show our method selects a good proportion of topics and summary contents is relevant.

6 Acknowledgement

The research leading to this paper was supported by the Institut Eurecom and by the European Commission under contract FP6-027026, Knowledge Space of semantic in-ference for automatic annotation and retrieval of multimedia content - K-Space.

References

[1] E. Dumont and B. Mérialdo. Split-screen dynamically accelerated video summaries. In *MM 2007, 15th international ACM conference on multimedia, September 24-29, 2007, Augsburg, Germany*, Sep 2007.

[2] A. G. Hauptmann, M. G. Christel, W.-H. Lin, B. Maher, J. Yang, R. V. Baron, and G. Xiang. Clever clustering vs. simple speed-up for summarizing rushes. In *TVS '07: Proceedings of the international workshop on TRECVID video summarization*, pages 20–24, New York, NY, USA, 2007. ACM.

[3] S. Lee and M. Haye. An application for interactive video abstraction. In *Proceedings of the ICASSP Conference*, 2004.

[4] B. Mérialdo and B. Huet. *Automatic video summarization*. Chapter in "Interactive Video, Algorithms and Technologies" by Hammoud, Riad (Ed.), 2006, XVI, 250 p, ISBN: 3-540-33214-6.

[5] A. Money and H. Agius. Video summarisation: A conceptual framework and survey of the state of the art. In *Journal of Visual Communication and Image Representation*, 2007.

[6] C.-W. Ngo, Y.-F. Ma, and H.-J. Zhang. Automatic video summarization by graph modeling. In *Proceedings of the Ninth IEEE International Conference on Computer Vision*, 2003.

[7] P. Over, A. F. Smeaton, and P. Kelly. The TRECVID 2007 BBC rushes summarization evaluation pilot. In *Proceedings of the TRECVID Workshop on Video Summarization (TVS'07)*, pages 1–15, New York, NY, September 2007. ACM Press.

[8] J. Saarela and B. Mérialdo. Using content models to build audio-video summaries. In *SPIE'99, Storage and Retrieval for Image and Video Databases, January 26, 1999 - San Jose, USA*.

[9] M. Shi Lu King, I.Lyu. Video summarization by video structure analysis and graph optimization. In *Proceedings of the International Conference on Multimedia and Expo*, 2004.

[10] B. T. Truong and S. Venkatesh. Video abstraction: A systematic review and classification. *ACM Trans. Multimedia Comput. Commun. Appl.*, 2007.

[11] I. Yahiaoui, B. Merialdo, and B. Huet. Generating summaries of multi-episodes video. In *ICME 2001, International Conference on Multimedia & Expo , August 22-25, 2001 Tokyo, Japan.*