A COLLABORATIVE APPROACH TO AUTOMATIC RUSHES VIDEO SUMMARIZATION

Werner Bailer+, Emilie Dumont, Slim Essid† and Bernard Mérialdo**

+JOANNEUM RESEARCH
Graz, AUSTRIA
werner.bailer@joanneum.at

*Institut Eurécom
Sophia-Antipolis, FRANCE
{dumont, merialdo}@eurecom.fr

†TELECOM ParisTech
Paris, FRANCE
slim.essid@telecom-paristech.fr

ABSTRACT

Video summarization is a useful tool which allows a user to grasp rapidly the essence of a video. In the development of this research topic we propose a new method based on different individual content segmentation and selection tools in a collaborative system. The main innovation of this work is to merge results from different approaches, so as to benefit from their respective qualities. Our system is organized in two phases: first segmentation of the video, second identification of relevant and redundant segments. The final list of selected segments is used to concatenate the video segments and build the final summary. In order to assess the effectiveness of this organization, we evaluate our system with a method based on the TRECVID 2007 BBC rushes summarization evaluation pilot and compare our performance with existing systems.

***Index Terms*— Video Summarization, Rushes, TRECVID, evaluation, MPEG-7**

1. INTRODUCTION

A summary is a shortened version of the original document. The main purpose of such a condensation is to highlight the major points from the original (much longer) subject, e.g. a text, a film or an event. The aim is to help the audience get the gist in a short period of time. Automatic video summarization is a challenge since it requires making decisions about the semantic content and importance of each segment in a video. This factor complicates the development of automatic video summarization systems and evaluation methods.

In the TRECVID 2007 BBC rushes summarization evaluation pilot [1], the task is to automatically create an MPEG-1 summary clip no longer than 4% of the full video that shows the main objects and events in the video. To evaluate the generated summaries, a human assessor views the summary using only the Play and Pause controls and compares visible objects/events with a predefined ground truth. Several indicators are collected during this evaluation, in particular the percentage of ground truth topics found is a measure of the information contained in the summary.

2. COLLABORATIVE APPROACH

This work takes place within the K-Space European Network of Excellence. The general objective of the network is to narrow the gap between low-level content descriptors and high-level human interpretations of audiovisual media [8]. Within the scope of this network, we are collaborating to develop an automatic summarization system. The main idea is to merge results from various approaches, so that the final summary can be decided based on a variety of information. We expect that a combined approach would be more accurate and more robust than individual systems. In order to implement this strategy in the framework of the TRECVID rushes summarization task, we have designed a two-phase architecture:

- First, we build a common segmentation of the video. This is achieved by merging segmentations based on different indicators.
- Second, the common segments are evaluated for redundancy and relevance. Each partner contributes by suggesting lists of relevant and redundant segments. Those lists are fused to compose a final ranked list of selected segments.

Figure 1 gives a graphical overview of this organization. The construction of the video summaries is therefore performed through the following steps:

1. Each partner proposes one or more segmentations of the original video, based on various indicators, and including confidence values for each suggested boundary.
2. Those segmentations are fused to produce a common segmentation of the original video.
3. Each partner analyses the common segments to detect redundancies and assess relevance. Two results are produced. First, a list of redundant segments, which shall not be included in the summary because they do not exhibit interesting content, or because they are

similar to other segments. Second, a ranked list of selected segments, which provide an indication of the importance of each common segment with respect to the information contained in the original video.

4. These lists are fused to produce a ranked list of common selected segments. Redundancy and relevance are taken into account to produce this list.
5. Finally, a video summary is constructed by concatenating the video clips of the selected segments.

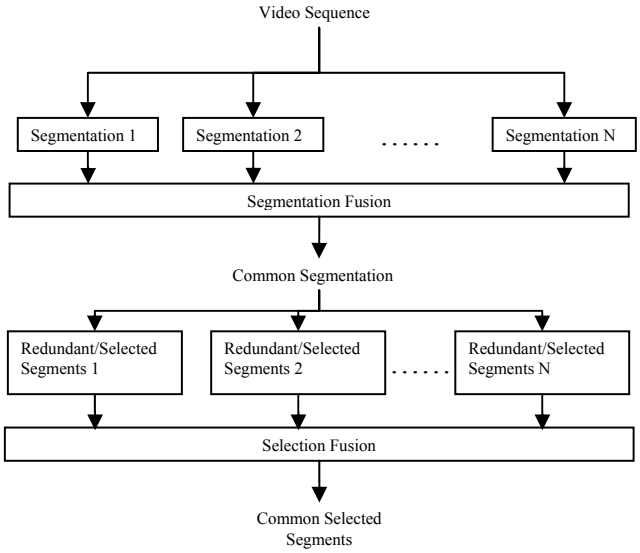

Figure 1: Overview of the collaborative summarization process.

3. COMMON SEGMENTATION

In a first step several segmentations of the original video are produced, based on various indicators and features.

One segmentation is a hard cut detection [2] (as this is the only type of transition appearing in unedited material) based on a SVM on color differences of three subsequent frames. The SVM is implemented through LIBSVM. To train the SVM classifier, ground truth from the TRECVID 2006 shot boundary detection task has been used.

The second segmentation is a shot boundary detection: we consider a sliding window over video frames centered on the current frame. To compute the distance between two frames, we build a 16-region HSV histogram for each frame. For each pre-frame and post-frame, we compute frame similarity between this frame and the central frame. We compare the ranking of pre-frames and post-frames, and we detect a transition when the number of top ranked pre-frames is greater than a predefined threshold.

Then the different segmentations are fused in order to produce a common temporal representation. The task is one of selecting the most relevant segment boundaries among the alternative ones (outputs by the different systems), using the confidence values associated with them.

First, we normalize the confidence values corresponding to each alternative system to make them commensurate. This is done by standardizing the confidence values, i.e. centering them and setting their variance to 1.

We then use an agglomerative clustering algorithm [4][5] to group together the closest boundaries. The algorithm starts with as many clusters as original data objects, where the data objects are the boundary times, measuring the proximities between all pairs of clusters and grouping together the closest pairs into new clusters. The algorithm stops when all the objects lie in a single cluster. The result of this procedure is a graph (called *dendrogram*) which depicts the relations and proximities between the obtained nested clusters. Boundary clusters are then formed by processing the dendrogram in such a way that the distances between all the boundaries grouped together in a cluster remain smaller than 5 seconds.

The last step consists in selecting the best representative of each boundary cluster. This is done as follows:

- for singleton clusters, the candidate boundary is kept only if its (standardized) confidence value is greater than -1;
- for the non-singleton clusters, the boundary exhibiting the highest confidence value is selected.

4. SEGMENT SELECTION APPROACHES

We use two strategies for determining segments to be included into the summary. One is the explicit selection of segments that are found to be relevant. For each of these segments, a relevance value is determined. The other is to determine redundant segments that shall not be included. The redundancy of a segment can be *absolute* (i.e. the content is not needed, e.g. a shot containing a color bar) or *relative* w.r.t. to a set of segments, i.e. these segments contain the same content and only one out of such a set needs to be considered. In the following we describe the approaches we have used to gather lists of selected and redundant segments.

4.1 Relevance Detection Approaches

Approach 1

A list of relevant segments based on visual features is created from visual activity and face detection results. The visual activity is normalized and segments with an activity exceeding 1.5 × standard deviation are used as candidates. The relevance value is reduced if no faces are present.

Temporal filtering is applied to remove outliers and enforce longer segments.

Approach 2

We divide the original video into one second segments, and we cluster these segments by agglomerative hierarchical clustering. We represent the one second segments by a HSV histogram. The distance between two such segments is computed as the Euclidean distance of the histograms, and the distance between two clusters is the average distance across all possible pairs of one second segments of each cluster.

We then iteratively select common segments which cover a maximum of content [3]. The importance of a common segment is defined as the number of clusters it contains.

4.2 Redundancy Detection Approaches

Approach 1

A straightforward approach for determining redundant segments is to identify color bars and monochrome frames. This is done by analyzing the standard deviation of columns of the frames [2]. If it is below a threshold for virtually all frames of a shot, this shot is marked as redundant.

Approach 2

Pattern models are used to detect redundant shots such as bars and monochrome images. Similar segments are detected when they contain the same clusters (as defined in approach 2 for relevance).

Approach 3

The take clustering approach proposed in [6] is used to identify and group several (possibly partial) takes of one scene. The approach is based on matching sequences of visual features using a variant of the Longest Common Subsequence (LCSS) algorithm and applying hierarchical clustering [5] to the resulting distance matrix.

5. FUSION OF SEGMENT SELECTIONS

The fusion step merges the different lists of selected and redundant segments in order to produce an output selection list of segments which shall be included in the final summary. The fusion step (i) transforms relative into absolute redundant segment lists, (ii) combines the relevant and redundant segment lists, (iii) selects an appropriate threshold and adapts the segment selection to the length constraints of the summary.

Relative redundancy information such as that about take clustering cannot be used directly, as not all but only *all but one* segments of a cluster are redundant. The reason for deferring this decision into fusion is that more information is available and other input can be used. We have implemented two approaches: The first option is to use the longest of the alternative takes. This ensures that most of the content of the take is included, even if it is unique to this take (e.g. this could be the only complete take, while the others in the cluster are only partial takes). The disadvantage is that parts of this take may need to be discarded later to fulfill length constraints. The second option is to use a clip that is most representative for the cluster, i.e. is shared by most of the alternative takes and has a high relevance. Applying one of these strategies allows treating the selected take or clip as relevant and the others as redundant.

The lists of relevant and redundant segment are interpreted as relevance functions rel(t) and red(t) over time t. The selection function is then given as

$$\text{select}(t) = \begin{cases} 1, & \text{if } \frac{w_{rel}}{n_{rel}} \sum_{k=1}^{n_{rel}} \text{rel}(t) + \frac{w_{red}}{n_{red}} \sum_{k=1}^{n_{red}} \text{red}(t) \geq \theta \\ 0 & \text{otherwise,} \end{cases}$$

where n_{rel} and n_{red} are the number of input relevant and redundant segment lists, w_{rel} and w_{red} are the relative weights of relevance and redundancy information and θ is a threshold.

The next step is to determine θ, so that $\Sigma_t \text{select}(t)$ is maximum while $\Sigma_t \text{select}(t) \leq T$, with T being the maximum duration of the selected segments (i.e. the maximum length of the summary). The optimization problem is solved using binary search. Very short segments (less than one second)) in the result are discarded, as they are hard to perceive and rather disturbing in the summary.

Especially in cases where few input lists are available, a number of segments with equal total relevance values may exist, so that – depending on the choice of the threshold – all of them are kept or discarded. As a result summaries that are much shorter than the maximum length are created. To avoid this we apply the following strategy, if the length of the summary would be below 90% of the maximum length: θ is set to the relevance of the segments in question, i.e. so that they are included. If several segments from one shot of the common segmentation are selected, only the longest is kept. Of the remaining segments we iteratively select the longest segment and crop beginning and end until the length constraint is matched.

6. EVALUATION

We experiment our method on 7 videos of the TRECVID BBC Rushes Task 2007 test set. It consists of unedited video footage, shot mainly for five series of BBC programs.

In the development of video summarization systems, one of the main problems remains evaluation. We use an automatic evaluation method based on a Bayes Network and proposed in [7]. This method has a strong positive correlation (Pearson's coefficient = 0.77), a moderate agreement (Kappa's coefficient = 0.49) and a weak variability (MSE =

0.043) with the manual evaluation method used for the official TRECVID evaluation [1].

In our method, we automatically evaluate the metric called IN, i.e. the percentage of topics found in the summary. A topic describes a video segment concerning people, things, events, locations, etc. and combinations of the former with a specific camera motion. Table 1 shows IN for different segment selections: Common selection 1 uses the longest of the alternative takes with $w_{rel}=0.7$ and $w_{red}=0.3$, Common selection 2 uses the longest of the alternative takes with $w_{rel}=w_{red}$ and Common selection 3 uses a representative clip of a take cluster with $w_{rel}=w_{red}$.

	MRS157443	MRS035132	MRS150148	MRS157475	MS237650	MRS043400	MS212920	**Mean**
Selection 1	0.25	0.00	0.25	0.33	0.17	0.33	0.00	**0.19**
Selection 2	0.17	0.00	0.25	0.25	0.17	0.11	0.08	**0.15**
Common selection 1	0.33	0.00	0.50	0.33	0.17	0.22	0.33	**0.27**
Common selection 2	0.67	0.00	0.50	0.42	0.25	0.22	0.33	**0.34**
Common selection 3	0.58	0.00	0.50	0.42	0.17	0.33	0.33	**0.33**

Table 1: Collaborative summary evaluation.

Those results show that the fusion of selections generally performs better than each of the selection. The poor results for video MRS035132 are due to a bad estimation of the IN percentage by the Bayes network. A manual evaluation would have given a value of IN equal to 0.33.

In the case of video MS212920, two selections with IN values of 0 and 0.08, are fused into a selection with a IN value of 0.33: the explanation is that the common selection is also considering segments with lower ranks which are not included in each of the selection. When such segments are relevant, they increase the score of the common selection.

The evaluation is limited because of the availability of ground truth data. Despite its limitation, these experiments show a trend for the common selection to improve over the individual ones.

7. CONCLUSION

The current fusion method is a rather straightforward approach. The first results that we have obtained are encouraging, and motivate us to explore such combinations further. We are currently in the process of annotating segment selection ground truth on a larger subset of the TRECVID 2007 data set. Once this is available, segment selection can be treated as a joint labeling and segmentation problem, i.e. producing the output selection segmentation based on the different input sequences. Such a problem could be for example solved using Hidden Markov Models or Conditional Random Fields. With a larger data set for training and evaluation, the fusion mechanisms should be more robust, and lead to more substantial and consistent improvements in the evaluation.

ACKNOWLEDGEMENTS

The research leading to this paper was partially supported by the European Commission under contract FP6-027026, Knowledge Space of semantic inference for automatic annotation and retrieval of multimedia content - K-Space. BBC 2007 Rushes video is copyrighted. The BBC 2007 Rushes video used in this work is provided for research purposes by the BBC through the TREC Information Retrieval Research Collection.

8. REFERENCES

[1] Over, P., Smeaton, A. F., and Kelly, P. 2007. The Trecvid 2007 BBC rushes summarization evaluation pilot. In Proceedings of the international Workshop on TRECVID Video Summarization (Augsburg, Bavaria, Germany, September 28 - 28, 2007). TVS '07. ACM

[2] Bailer, W., Lee, F., and Thallinger, G. 2007. Skimming rushes video using retake detection. In Proceedings of the international Workshop on TRECVID Video Summarization (Augsburg, Bavaria, Germany, September 28 - 28, 2007). TVS '07, ACM

[3] Dumont, E. and Merialdo, B. 2007. Split-screen dynamically accelerated video summaries. In Proceedings of the international Workshop on TRECVID Video Summarization (Augsburg, Bavaria, Germany, September 28 - 28, 2007). TVS '07. ACM

[4] Sergios Theodoridis and Konstantinos Koutroumbas.*Pattern Recognition*. Academic Press, 1998.

[5] Richard Duda and P. E. Hart. *Pattern Classification*. Wiley-Interscience. John Wiley & Sons, 2001.

[6] Werner Bailer, Felix Lee and Georg Thallinger, "Detecting and Clustering Multiple Takes of One Scene," in Proceedings of 14th Multimedia Modeling Conference, Kyoto, JP, Jan. 2008, pp. 80-89.

[7] Emilie Dumont and Bernard Merialdo, "Automatic evaluation method for rushes summarization: experimentation and analysis", CBMI 2008, 6th International Workshop on Content-Based Multimedia Indexing, 18-20 June, London, UK

[8] K-Space Network of Excellence, http://www.k-space.eu/