Rushes Video Summarization Using a Collaborative Approach

Emilie DUMONT[*], Bernard MERIALDO[*], Slim ESSID[†], Werner BAILER[+], Herwig REHATSCHEK[+], Daragh BYRNE[˜], Hervé BREDIN[˜], Noel E. O'CONNOR[˜], Gareth J.F. JONES[˜] Alan F. SMEATON[˜], Martin HALLER[°], Andreas KRUTZ[°], Thomas SIKORA[°], Tomas PIATRIK[‡]

[*]EURECOM
Sophia-Antipolis, FRANCE
{dumont, merialdo}@eurecom.fr

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

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

[˜]Dublin City University
Dublin, IRELAND
smeaton@computing.dcu.ie

[°]Technische Universität Berlin
Berlin, GERMANY
sikora@nue.tu.berlin.de

[‡]Queen Mary University of
London, UK
tomas.piatrik@elec.qmul.ac.uk

ABSTRACT

This paper describes the video summarization system developed by the partners of the K-Space European Network of Excellence for the TRECVID 2008 BBC rushes summarization evaluation. We propose an original method based on individual content segmentation and selection tools in a collaborative system. Our system is organized in several steps. First, we segment the video, secondly we identify relevant and redundant segments, and finally, we select a subset of segments to concatenate and build the final summary with video acceleration incorporated. We analyze the performance of our system through the TRECVID evaluation.

Categories and Subject Descriptors

H.3.1 [Content Analysis and Indexing]: *Abstracting methods*

H.5.1 [Multimedia Information Systems]:*Evaluation/methodology*

General Terms

Algorithms, Measurement, Experimentation

Keywords

video summarization, rushes, TRECVID, evaluation, MPEG-7, fusion.

1. INTRODUCTION

A summary is a shortened version of an 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 this paper, we present a technique for automatic video summarization developed as part of the K-Space Network of Excellence for our participation in the 2008 TRECVID video summarization task. We use power and knowledge of each partner in the idea to merge all individual techniques to acquire a competitive system.

The rest of this paper is organized as follows: Section 2 presents the collaborative approach, Section 3 the video segmentation step, Section 4 the segment selection step and Section 5 the summary presentation step. Finally, we analyze results of the evaluation within the TRECVID BBC task.

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 [1]. Within the scope of this network, we are collaborating to develop an automatic video 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 [2]. In order to implement this strategy, we have designed a three-phase architecture as shown Figure 1:

- First, we build a common segmentation of the video. JRS, Eurecom and TUB proposed one or more segmentations of the original video, based on various indicators, and including confidence values for each suggested boundary. GET's role has been to merge segmentations based on different indicators to produce a common segmentation of the original video.
- Second, the common segments are evaluated for redundancy and relevance. JRS, Eurecom and QMUL analyzed the common segments to detect redundancies and assess

relevance. Two kinds of results have been 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. These lists were fused by JRS to produce a ranked list of common selected segments. Redundancy and relevance are taken into account to produce this list.

- Finally, a video summary is constructed by concatenating the video clips of the selected segments with a video acceleration included and produced by DCU.

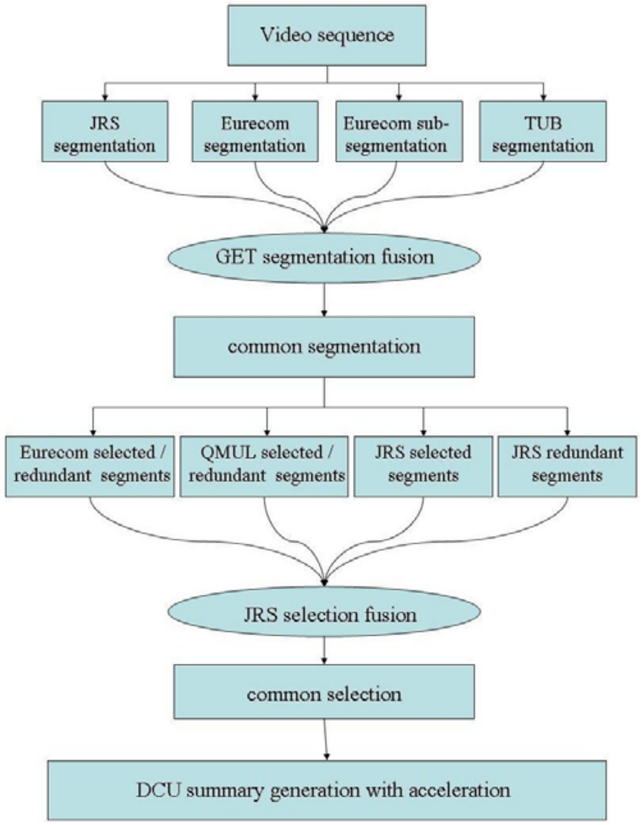

Figure 1: Overview of the collaborative summarization process.

3. VIDEO SEGMENTATION

In a first step, several segmentations of the original video are produced based on various indicators and features. The different segments are fused in order to produce a common temporal segmentation. The task is one of selecting the most relevant segment boundaries among the alternative ones using the confidence values. Table 1 shows the number of sequences and their length in the segmentations proposed by the partners and in the common segmentation.

	Mean number of segments per video	Mean length of segments (in frames)
JRS segmentation	56.125	709.500
Eurecom segmentation	104.075	382.613
Eurecom sub-segmentation	185.725	214.405
TUB segmentation	561.700	70.890
Common segmentation	70.450	565.23

Table 1: Segmentation results.

3.1 Individual segmentation

3.1.1 JRS segmentation

As we are dealing with unedited material we perform a hard cut detection using 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.

3.1.2 Eurecom segmentation

We perform 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. We propose two segmentations: a shot segmentation and a sub-shot segmentation. For the sub-shot segmentation, thresholds are lowered.

3.1.3 TUB segmentation

Camera motion characterization is used to identify segments with consistent camera motion. For this, higher-order global motion estimation parameters are described by appropriate features. These features are then used for multi-class SVM classification of no/left/right pan, no/up/down tilt, and no/in/out zoom [3]. Each identified segment based on these classification results represents a video sequence with the same camera movement. Finally, confidence values required by the fusion stage are determined by using Platt's approach during SVM classification and assigned to the segmentations [4].

3.2 Common segmentation

The common segmentation is obtained in a straightforward unsupervised way by selecting the most relevant segment boundaries among the alternatives output by the individual segmentation blocks. The approach is briefly described in the following.

First, we normalize the confidence values from each individual 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 [6][7] to group together the closest boundaries. Clusters of boundary times are thus formed by ensuring 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 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

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. Table 2 shows results of the segment selection approaches.

	Mean number of selected segments per video	Mean number of redundant segments per video
JRS retake det.	n/a	14.42
JRS face/motion act.	1304.38	n/a
JRS color bars	n/a	49.56
Eurecom selection	3.52	70.45
QMUL selection	5.30	65.50
Common selection1	19.67	n/a
Common selection2	6.68	n/a

Table 2: Selection of segments

4.1 Individual relevance detection

4.1.1 JRS relevance detection

A list of relevant segments based on visual features is created from visual activity and face detection results. The face detector is based on the algorithm developed by Viola and Jones which is implemented in Intel's OpenCV library. The face detection result forms a discrete function. To eliminate short-term false detections (i.e. sudden appearance or disappearance of faces) mathematical morphological closing and opening operators were applied to the sequence of 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.

4.1.2 Eurecom relevance detection

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 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 like in [2]. The importance of a common segment is defined as the number of clusters it contains.

4.1.3 QMUL relevance detection

A fundamental step in our approach is to create the similarity matrix and organize video frames into the tree-structure using ant-tree clustering method. We use a combination of MPEG-7 Color Layout and Edge Histogram descriptors for representing each video frame. On the basis of a root on which the tree is built, frames are gradually fixed to the structure. The movement and fixing of frames in a position depends on the similarity value, temporal information and the local neighborhood of moving frame. In order to find the optimal number of clusters and for increasing the quality of each cluster, the similarity threshold updating procedure is optimized as proposed in [8]. Results of the ant-tree algorithm are clusters which are used in the decision process for classification of relevant/redundant segments. Common video segments which contain representative frames attached to the root of the tree are classified as relevant. The importance of relevant segment is defined by number of frames from redundant segments in the corresponding cluster.

4.2 Individual redundancy detection

4.2.1 JRS redundancy detection

A straightforward approach for determining redundant segments is to identify color bars and monochrome frames. If the standard deviation of columns of a significant number of frames in a shot is below a threshold this shot is marked as redundant.

As a second approach the take clustering algorithm proposed in [5] is used to identify and group several (possibly partial) takes of one scene. The problem of detecting and clustering multiple takes of the same scene shot from the same or similar camera positions is formulated as a problem of matching sequences of visual features (ColorLayout and EdgeHistogram descriptors for every key frame, average visual activity between the key frames) of segments of the input video. The algorithm uses the LCSS (Longest Common Subsequence) model to determine the similarity between two segments. Hierarchical clustering is applied to the resulting distance matrix and yields a set of clusters that correspond to scenes.

4.2.2 Eurecom redundancy detection

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

4.2.3 QMUL redundancy detection

We detect monochrome parts of the video by computing visual deviation of frames within a video segment. Simple threshold comparisons are used to identify static redundant segments. Visually similar scenes or multiple takes of the same scene are grouping together using our aforementioned approach (see section 4.1.3). Basically, all video segments which do not contain the representative frame are classified as redundant.

4.3 Common selection

The fusion step, which is an extended version of that described in [2], 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. 2 shows an overview of the process. Relative redundancy information such as that from 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 (e.g. about junk content). We 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.

A joint relevance timeline is calculated as a weighted sum of the different input relevant/redundant segment lists. The next step is to determine a relevance threshold, so that the duration of segments above the threshold is maximum but below the given length constraint. The optimization problem is solved using binary search. Very short segments (below a user defined threshold) in the result are discarded, as they are hard to perceive and rather disturbing in the summary. Longer remaining segments are split into several shorter ones in order to increase the number of different clips in the selected segments. This produces results with several shorter segments instead of fewer longer ones. select the longest segment and crop it at beginning and end until the length constraint is matched.

The fusion method has the following parameters (the values in parenthesis are those used for K-Space run 1 and 2): the minimum (0.5, 2 secs) and maximum (2, 20 secs) duration of a selected segment, the minimum temporal distance between two selected segments (both 6 secs), the maximum total duration of selected segments (3%, 8% of the input duration) and the weights for combining summed relevances with summed redundancies (both 0.5).

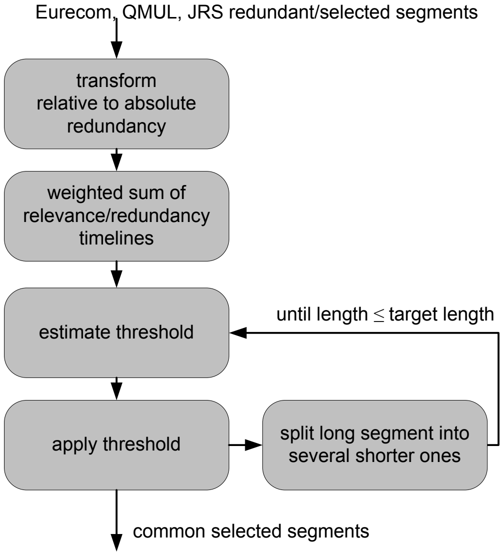
Eurecom, QMUL, JRS redundant/selected segments
transform relative to absolute redundancy
weighted sum of relevance/redundancy timelines
estimate threshold
until length ≤ target length
apply threshold
split long segment into several shorter ones
common selected segments

Figure 2: Fusion of the different lists of selected and redundant segments.

5. SUMMARY PRESENTATION

The previous processing steps provided a set of video selected segments to be included within the final summary. Selected segments for two runs to be submitted for evaluation, were determined. For summaries belonging to the first run, a large number of extremely short shots were selected. Within this run, the duration of these selected segments totaled 3% of the original footage and each segment was no more than 50 frames in length. For the second run, a smaller set of shots were selected with an overall longer duration. The duration of these selected segments totaled 8% of the original footage's length and each segment was at least 250 frames in length, but often significantly more.

As a first stage in the construction of the summary, video and audio for each selected segment was extracted and separated from the original footage. This was achieved using a custom application to transform the output from the previous steps and initiate extraction using FFMPEG. Once the video frames and audio were available they were constructed into a final video summary using Proce55ing [7], an open source programming language specifically designed for electronic arts and visual design. Within the final summary, the selected segments were accelerated uniformly in order to fit within the 2% upper bound for the final submission. For the first run, segments were accelerated 1.5 times their original speed while for Run 2 they were accelerated to 4 times their original speed. Hard cuts were used to join summary segments and as such no transition was employed to mark the transition from one segment to another. Audio was included for each segment in the summary; however, this was not accelerated. Acceleration may have significantly distorted the audio thereby making it difficult to interpret and consequently distracting to the viewer. Instead an audio clip taken from the middle of the selected segment was included and aligned with the video playback. Each summary additionally includes a timeline at the bottom of the screen. The timeline is highly transparent to prevent occlusion of shot based visual information but is sufficiently visible to provide a useful cue to the location of a summary segment within the original footage. The position of the timeline marker updates as the summary moves to a new segment. The presentation of the final summary is illustrated in Figure 3.

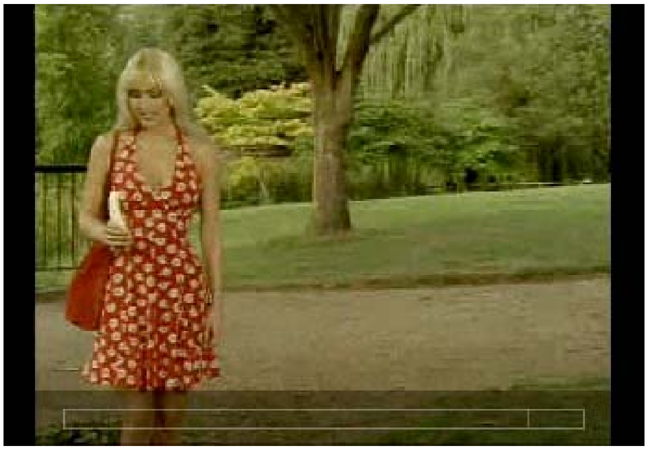

Figure 3: Summary Presentation Format (Example from Summary MRS148090).

6. RESULTS

Summary evaluation proposed by [8] shows results for each of 39 summaries for the two runs. Several criterions are used for summary evaluation including:

- IN - fraction of inclusions found in the summary (0 - 1)
- JU - Summary contained lots of junk: 1 strongly agree - 5 (best) strongly disagree
- RE - Summary contained lots of duplicate video: 1 strongly agree - 5 (best) strongly disagree
- TE - Summary had a pleasant tempo/rhythm: 1 strongly disagree - 5 (best) strongly agree

Figure 4 shows the comparison between the 2 runs of K-Space and the baseline and mean of runs submitted.

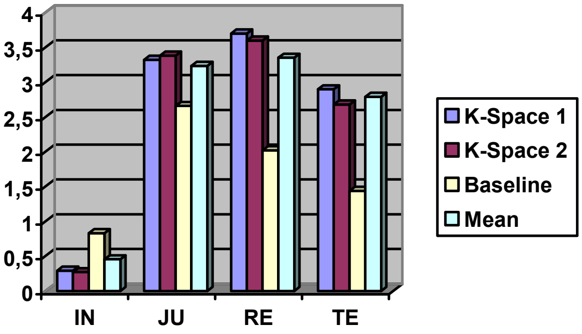

Figure 4: K-Space results.

This suggests that our system is reasonable, especially given the fact that evaluation is not easily reproducible, so it is difficult to have a good training for the different methods, and in particular in our case, because for each step, we use individual method and we perform a fusion, so the training step is very important. Concerning JU and RE, we have good results, this means that individual methods were able to find correctly junk sequences, redundancy and fusion reflect these good detections. The TE is equal to the average of participants. The fraction of inclusions found in our summaries is a little lower than the average value.

The evaluation results shows that we do not show a lot of junk frames and redundancy, so we show a set of interesting sequences with a little acceleration (1.5x) for the first run and a fast acceleration for the second run (4x). And for each segment, we show only 2 seconds. So, to increase IN, we must show more sequences, e.g we should perform a faster acceleration and propose only one second per segment.

7. CONCLUSION

We have presented the K-Space participation in TRECVID 2008 summarization task. This has been our first participation in the TRECVID summarization task as a large group of research teams drawn together in an EU-funded network. We proposed to fuse individual tools in order to profit of all partner knowledge. This first participation was done in a good manner: our results are good and encouraging us to continue our collaboration.

8. ACKNOWLEDGMENTS

The research leading to this paper was supported by the European Commission under contract FP6-027026, Knowledge Space of semantic inference for automatic annotation and retrieval of multimedia content - K-Space. Part of this work was sponsored by BT Group plc. and the Irish Research Council for Science, Engineering and Technology (IRCSET): Funded by the National Development Plan.

BBC 2008 Rushes video is copyrighted. The BBC 2008 Rushes video used in this work is provided for research purposes by the BBC through the TRECVID Collection.

9. REFERENCES

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

[2] Werner Bailer, Emilie Dumont, Slim Essid and Bernard Mérialdo, A collaborative approach to automatic rushes video summarization, ICIP 2008, IEEE International Conference on Image Processing, October 12–15, 2008, San Diego, California, U.S.A.

[3] Martin Haller, Andreas Krutz and Thomas Sikora, "A Generic Approach for Motion-based Video Parsing", in Proceedings of the 15th European Signal Processing Conference (EUSIPCO 2007), Poznań, Poland, Sept. 2007, pp. 713-717

[4] John C. Platt, "Probabilities for SV Machines" in A.J. Smola, P.L. Bartlett, B. Schölkopf, and D. Schuurmans (eds.), *Advances in Large Margin Classifiers*. MIT Press, 2000, pp. 61-74

[5] 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.

[6] E. Dumont, B. Mérialdo, Redundancy removing and event clustering for video summarization, WIAMIS 2008, 9th International Workshop on Image Analysis for Multimedia Interative Services, May 7-9, 2008, Klagenfurt, Austria

[7] C. Reas, B. Fry, and J. Maeda. Processing: A Programming Handbook for Visual Designers and Artists. The MIT Press, 2007.

[8] P.Over, A.F. Smeaton and G.Awad. The TRECVid 2008 BBC rushes summarization evaluation. In TVS'08: Processsings of the International Workshop on TRECVID Video Summarization, 2008.

[9] U. Damnjanovic, T. Piatrik, D. Djordjevic, and E. Izquierdo, "Video Summarisation for Surveillance and News Domain," in Proceedings of 2nd International Conference on Semantics and Digital Media Technologies (SAMT2007)*,* December 5-7, 2007, Genova, Italy