# Split-Screen Dynamically Accelerated Video Summaries

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 describe our approach to the TRECVID 2007 BBC Rushes Summarization task. Our processing is composed of several steps. First the video is segmented into shots. Then, one-second video segments are clustered into similarity classes. The most important non-redundant shots are selected such that they maximize the coverage of those similarity classes. Then shots are dynamically accelerated according to their motion activity to maximize the content per time unit. Finally they are optimally grouped by sets of four to be presented using split-screen display. The summaries produced have been evaluated in the TRECVID campaign. We present a first attempt at automating the evaluation process.

## Categories and Subject Descriptors

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

## General Terms

Algorithms, Performance

## 1. INTRODUCTION

Digital video documents are now widely available. 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 [15], [8]. A key element is the process of redundancy elimination. Visual features, in particular color histograms, are often used to measure the similarity between frames or shots, for example authors in [2] and [7] remove redundancy by selecting only contiguous frames that maximize the average similarity to a video, while authors in [6] propose a set of methods of audio-visual attention model features. Authors in [9], [14] and [5] compute elements such as color contrast, intensity contrast, and orientation contrast to model the human attention level to a particular image. Authors in [12] extract high-contrast scenes to include in movie summary. Redundancy can also be removed via clustering, as in [1], [3] and [4] in which a maximum of one shot is retained from a cluster of visually similar shots.

This paper is organized as follows : the next section explains our motivation and approach. In the following sections, we describe the details of our method. Finally, we will present the evaluation results provided by TRECVID, and propose a first attempt to automate this evaluation.

## 2. GENERAL APPROACH

We now present the major steps of our approach, as illustrated in figure 1. First, since rushes are raw material used to produce a video, they contain a significant part of uninteresting shots. 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. Those uninteresting shots are removed in an initial preprocessing step.

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, we propose to make a selection of the most relevant shots by maximizing non-redundant information. We begin the selection process by partitioning the video into one-second segments, then we cluster the 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 shot and select a set of relevant shots to be included in the


*TVS'07,* September 29, 2007, Augsburg, Bavaria, Germany.

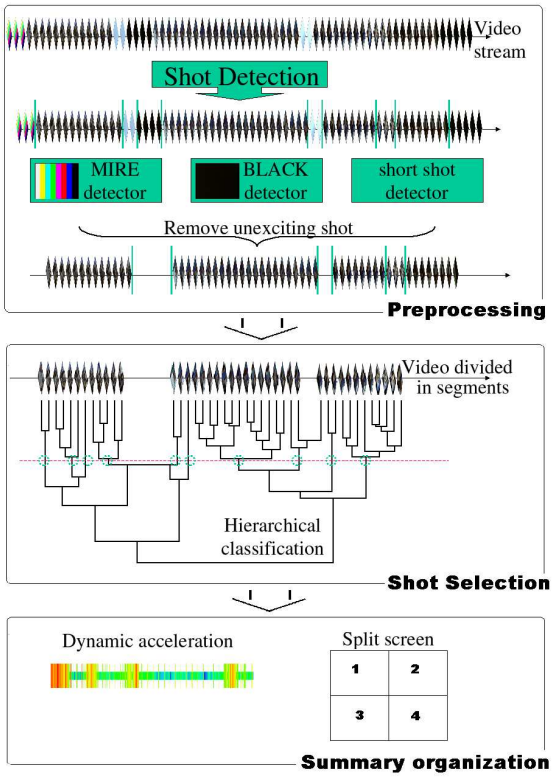


**Figure 1: Scheme of proposed approach**

summary.
To present the selected shots, we propose two original techniques.

- First, shots are dynamically accelerated according to the motion activity, so that a maximum amount of content may be presented in a minimum amount of time.
- Second, we group shots by sets of 4 and display them together using a split screen display. The grouping should follow rules in order to maximize the presentation efficiency.

# 3. PREPROCESSING

As a first step, we remove unexciting shots, such as black shots, test pattern shots, etc... using specific detectors. The table 1 shows the results of the preprocessing.

## 3.1 Shot boundary detection

We perform shot boundary detection using a method similar to the one proposed in [16]. We consider a sliding window over video frames, centered on the current frame. To compute the distance between frames, we build a 16-region HSV histogram for each frame, we remove the 4 central regions, and use the Euclidean distance between those vectors as a measure for frame similarity. For each pre-frame and post-frame, we compute the frame similarity between this frame and the central frame. For hard cuts, we compare the ranking of pre-frames and post-frames, and we detect a cut when the number of top ranked pre-frames is maximum. For gradual transitions, we compute the average similarity of pre-frames and post-frames, and we detect the end of a transition when this ratio is minimal.

| | Videos | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | MRS025913 | MS210470 | MRS158385 | MRS042543 | MRS148800 | MRS157475 | MRS159023 | MRS157478 | MRS043400 | MS237650 |
| Shots | 76 | 56 | 63 | 42 | 47 | 48 | 144 | 89 | 85 | 13 |
| TP | 0 | 1 | 1 | 1 | 2 | 1 | 5 | 1 | 1 | 1 |
| Black | 6 | 7 | 2 | 2 | 2 | 0 | 3 | 2 | 1 | 1 |
| Short | 44 | 31 | 29 | 13 | 20 | 16 | 41 | 52 | 19 | 4 |
| Total | 26 | 17 | 31 | 26 | 23 | 31 | 95 | 34 | 64 | 7 |

**Table 1: Results of preprocessing : this table shows the number of shots, test pattern shots, black shots and short shots detected during the process for 10 videos and the total number of shots selected to continue.**

## 3.2 Removing specific shots

### 3.2.1 *Test pattern shots*

A test pattern shot contains very particular frames, composed of stripes with various colors and greys. 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 HSV histogram, and we average the histograms of all training frames to build a detector vector $T$.
To remove the test pattern shots, we compare all frame vectors of the shot with the detector vector $T$ using the Euclidean distance. If the number of frames similar to $T$ is larger than a predefined threshold, this shot is categorized as a test pattern and removed.

### 3.2.2 *Black shots*

In a similar manner, we compute a characteristic HSV vector for black frames called $BLACK$ and we remove all shots where the number of frames similar to the BLACK vector is larger than a predefined threshold.

### 3.2.3 *Short shots*

Rushes often contain particular events which lead to detect false transitions during shot boundary detection. There are various reasons of this, for example when people pass in front of the camera between two video recordings or when a movie clapper board is used in front of the camera... To cope with such oversegmentation, we remove all shots with less than 25 frames (1 second).

# 4. SHOT SELECTION

After rushes preprocessing, we propose to make a selection of the most relevant shots. The idea is to select non-redundant shots, whose content overlaps as little as possible. This is performed by partitioning the video into

one-second segments, in other words, by partitioning events. Then we cluster segments by agglomerative hierarchical clustering, so a list of clusters represents an event. And finally, we select a set of shots which covers all events. Figure 2 shows some results of this process.

## 4.1 Hierarchical clustering

In order to evaluate the visual redundancy of video, we partition the video sequence into segments of 1 second each (25 frames). We cluster segments by an agglomerative hierarchical clustering algorithm. Each segment is represented by a HSV histogram of the central frame. The distance between two 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. 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.

Each iteration of the algorithm provides a different clustering of the 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. For this purpose, we assign a coefficient of perceptual duration, and we select the level with a perceptual duration equals 16% of video duration.

## 4.2 Cluster weight

The weight is intuitively related to the importance of the content of a cluster. As the appearance of people is generally an important part of the content, a face detector is used [13]. For each cluster, and for each segment, we extract the number of faces, so we can compute face probability $P(face/c)$ of a cluster $c$ by dividing the number of segments containing a face by the number of segments. Also, we extract the normalized average entropy $Ent(c)$. And finally, we compute the weight of a cluster $c$ by :

$$weight(c) = \frac{1 + 0.5 * P(face/c) + 0.5 * Ent(c)}{2}$$

## 4.3 Shot selection

We select the most important and non redundant shots for the summary by an iterative algorithm. The weight of a shot 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 shot, and mark its clusters as selected. This process is repeated until all clusters have been selected.

## 4.4 Perceptual duration

For each level $l$, we evaluated the perceptual duration from the motion activity explained in [10] by :

$$PD(l) = \frac{\sum_{s \in S} activity(s)}{\sum_{s \in V} activity(s)}$$

where $S$ is the set of segments of selected shots and $V$ the set of video segments. The summary should represent a maximum of 4% of the original content, presented in a 4 split screen, so we select the level which leads to a perceptual duration equals to 16% of the video perceptual duration.

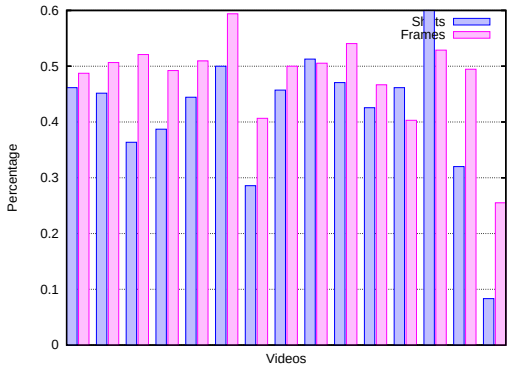


**Figure 2: Results of shot selection : percentage of selected shots and selected frames for each video.**

# 5. SUMMARY PRESENTATION

Once the shots in the summary have been selected, they have to be assembled in a single video, which represents a maximum of 4% of the original content, as stated in the TRECVID BBC Rushes guidelines. We propose two original ideas for this assembly: - shots are dynamically accelerated based on their content, so that we maximize the content displayed by time unit, - shots are grouped 4 by 4 and presented in a split-screen display, so that 4 shots are visible at the same time.

## 5.1 Dynamic acceleration

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 secondes. The idea of acceleration is to show a sequence during a time proportional to its motion activity.

We compute the motion activity $activity(f)$ for each frame. For the whole video, the set of frames is $F$, and the global motion is $Gactivity = \sum_{f \in F} activity(f)$. The maximum number of frames for the summary is $Tframes$, so that we can compute the number of frames for each shot $s$ by :

$$frame(s) = Tframes * \frac{\sum_{f \in s} activity(f)}{Gactivity}$$

To select frames for the summary, we store the $frame(s)$th decreasing motion activity value of shot frames in $threshold(s)$. And we select frames with a larger value than $threshold(s)$, see figure 3 .

## 5.2 Split screen organization

The display is split in 4 sections as shown in figure 4. 4 shots of the summary are presented simultaneously, one in each section. We cluster all the shots of the summary by groups of 4, based on the temporal similarity and visual dissimilarity.
The split screen technique allows to present a lot of

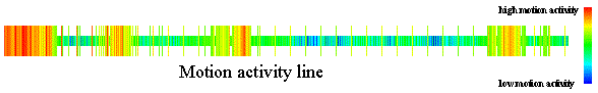


**Figure 3: Example of a dynamic acceleration on video MRS048780 (shot 13). 198 of 555 frames are in the summary : frames with a vertical small line have a motion activity lower than the threshold, and other are selected to create the summary.**

simultaneous information to the viewer. We found that presenting visually similar shots at the same time was sometimes confusing and differences were difficult to detect. This is why we chose to present shots that are as visually dissimilar as possible.

**Figure 4: Example of a split screen: frame 200, extracted MRS157484 summary.**

# 6. RESULTS AND DISCUSSION

## 6.1 Experimental results

The evaluation is based on several measures : DU duration of the summary, XD difference between target and actual summary size, TT total time spent judging the inclusions, VT total video play time (versus pause) judging the inclusions, IN fraction of inclusions found in the summary, EA "Was the summary east to understand", RE "Was there a lot of duplicate video". The complete evaluation for all TRECVID participants was done by [11], the table 2 shows the Eurecom results compared with the average ones.

We focus our discussion on the three following measures: IN, EA and RE, which we feel are specially interesting. Figures 5, 6 and 7 show the results of those measures for the first 20 videos. The inclusion rate IN is generally better than the average, which is expected since the split-screen techniques allows to display more information per time unit. The easyness EA is quite low, very often the worst observed in the experiments. This may be due on one side, to the fact that watching four running videos at the same time is very difficult, and requires an extreme and exhausting

| | DU | XD | TT | VT | IN | EA | RE |
|---|---|---|---|---|---|---|---|
| Eurecom | 42 | 18 | 119 | 43 | 0.53 | 1.97 | 3.02 |
| Average | 50.5 | 9.3 | 93 | 52 | 0.48 | 3.18 | 3.65 |
| Maximal | 64 | 33.8 | 119.3 | 66.6 | 0.68 | 3.6 | 3.98 |
| Minimal | 26 | -4.34 | 61.7 | 28.4 | 0.25 | 1.97 | 3.02 |

**Table 2: Comparison between mean results of Eurecom and the average, minimal and maximal results of participants.**

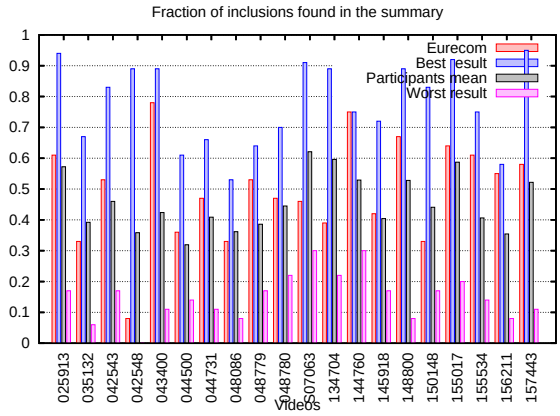


**Figure 5: Comparison between Eurecom results and the average, best, and worst results of participants for the fraction of inclusions found in the summary.**

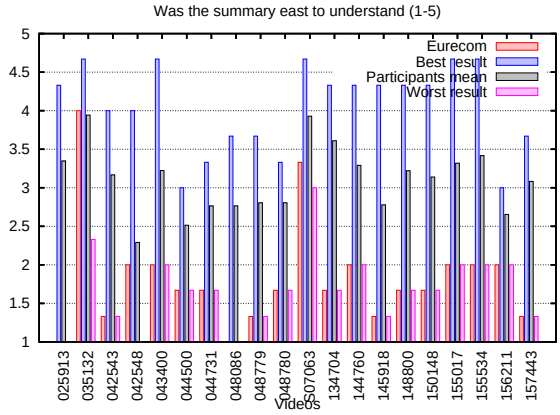


**Figure 6: Comparison between Eurecom results and the average, best, and worst results of participants for EA "Was the summary easy to understand".**

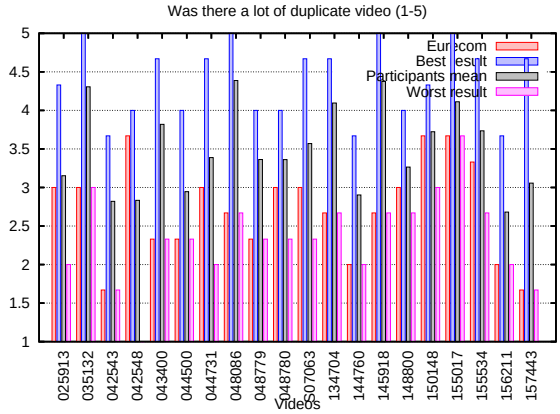


**Figure 7: Comparison between Eurecom results and the average, best, and worst results of participants for RE "Was there a lot of duplicate video".**

attention from the user. This is probably also partly due to

our acceleration algorithm, which is some cases could lead to accelerations that are above an admissible rate. In such situation, a topic might not be detected by the evaluator even if it is effectively present in the video. The redundancy RE is low too, and this is probably due to the use of entire shots as selection units. This prevents redundancy within a shot to be removed.

## 6.2 Automatic evaluation

The TRECVID evaluation of summaries is presently manual. This has a number of disadvantages, in particular, the difficulty to reproduce experiments with other data. In an attempt to automate the evaluation, we manually add to the list of topics for a video the frame number intervals where this topic appears, together with an estimation of the minimal duration required to notice the topic while viewing. This information allows to automatically estimate the IN measure by :

$$IN = \frac{\sum_{t \in T} \frac{\lfloor \min(RF(t), Nf(t)) \rfloor}{RF(t)}}{\sharp T}$$

where $T$ is the set of topics, $RF(t)$ is the minimum number of frames to detect the topic $t$ and $Nf(t)$ is the number of frames of the topic $t$ selected in the summary.

We experimented this method on the video MRS157443. $RF(t)$ is set to the same value for every topic, and we can see how the IN measures changes when this value varies. Results are shown in figure 8. We can see that when $RF(t) \in [2 : 22]$, the IN value has a value similar to the one found in the TRECVID evaluation. This is just a preliminary result, this approach has now to be validated on a larger set of videos.

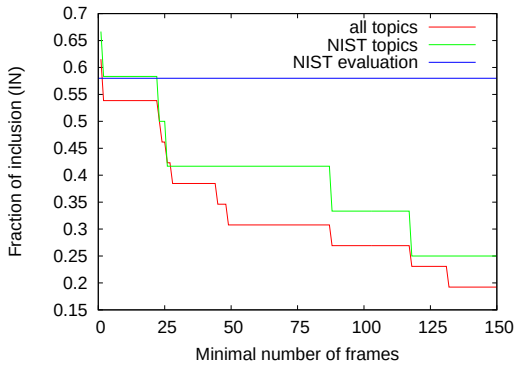


**Figure 8: Automatic evaluation for IN**

# 7. CONCLUSIONS AND FUTURE WORK

A new approach has been proposed for summarizing video rushes. We proposed to select the most interesting shots to create the summary. To present it, we propose two original approachs : to accelerate shots according to motion activity and to split the screen in four sections. Summaries were evaluated by TRECVID and compared with other methods.

This suggests several improvements that we hope to investigate in future work. Summaries are hard to understand, so to improve the visibility of summaries is an interesting investigation. A second improvement would be to take into account a notion of redundancy during the classification. Currently, we are also investigating a method to work with a selection unit shorter than shots.

# 8. 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.
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.

# 9. REFERENCES

[1] R. L. A. Hanjalic and J. Biemond. Automated high-level movie segmentation for advanced video retrieval systems. In *IEEE Trans. Circ. Syst. Video Technolo. 9, 4 June, 580-588*, 1999.
[2] M. Cooper and J. Foote. Summarizing video using non-negative similarity matrix factorization. In *IEEE Workshop on Multimedia Signal Processing*, 2002.
[3] Y. Gong and X. Liu. Summarizing video by minimizing visual content redundancies. In *Proceedings of the International Conference on Multimedia and Expo*, 2001.
[4] Y.-H. Gong. Summarizing audio-visual contents of a video program. In *EURASIP J. Appl. Signal Process.: Special Issue on Unstructured Info. Manage. Multimedia Data Sources, 2 (Feb.)*, 2003.
[5] S. Lee and M. Haye. An application for interactive video abstraction. In *Proceedings of the ICASSP Conference*, 2004.
[6] Y.-F. Ma, L. Lu, H.-J. Zhang, and M. Li. A user attention model for video summarization. In *Proceedings of the tenth ACM international conference on Multimedia*, 2002.
[7] 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.
[8] 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.
[9] 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.
[10] J. Oh and P. Sankuratri. Computation of motion activity descriptors in video sequences. In *Proc. of the 2nd WSEAS International Conference on Multimedia, Internet and Video Technologies*, pages 139–144, Skiathos Island, Greece, 2002.
[11] 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.
[12] S. Pfeiffer, R. Lienhart, S. Fischer, and W. Effelsberg. Abstracting digital movies automatically. In *Journal of Visual Communication and Image Representation*, 1996.
[13] H. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. In *Computer Vision and Pattern Recognition '96*, June 1996.
[14] 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.
[15] B. T. Truong and S. Venkatesh. Video abstraction: A systematic review and classification. *ACM Trans. Multimedia Comput. Commun. Appl.*, 3(1):3, 2007.
[16] T. Volkmer, S. Tahaghoghi, and H. E. Williams. RMIT University at TREC 2004. In E. M. Voorhees and L. P. Buckland, editors, *Proceedings of the TRECVID 2004 Workshop*, 2004.