# MULTI-EPISODES VIDEO SUMMARIES

Benoit Huet, Itheri Yahiaoui and Bernard Merialdo
Institut Eurecom
2229 Route des cretes,
06904 Sophia-Antipolis, France.
{Huet,Yahiaoui,Merialdo}@eurecom.fr

## ABSTRACT

This paper presents a novel concept in the context of automated video summaries creation. The idea is to build summaries of multiple video sequence such as soap opera or TV series where similar scenes are present in several episodes. Here we will aim at exposing within the summaries the scenes that make each episode unique. The videos are first segmented into shots and a representative frame is extracted from each shot. The set of colour histograms of those frames is the basis for our selection method. Indeed, the clustering of the key-frames colour histogram will provide us with information about which shots belong to a single, some, many or all video episodes. We will present our results along with a sensitivity evaluation of the proposed method.
*Keywords: Video Summaries, Content Management*

## 1 INTRODUCTION

Multimedia information and particularly video is becoming more and more common as the power of computing devices and the bandwidth of transmission channels increase along with the ever decreasing cost of storage. However, accessing a particular scene from a video or a particular video from a collection is non trivial due to the temporal structure of the data. It is therefore necessary to develop methodologies for summarising and presenting the video to the user in a more suitable format.

There have been many efforts devoted by the research community to address the shot detection and segmentation problems [2, 1]. A shot is a "continuous" sequence of frames from a video recorded from a single camera operation. It is the basis of much research in video content analysis and the basis of the work presented here.

In this paper we describe a new methodology used to construct video summaries particularly adapted to TV Series. Indeed, in the context of TV series it seems of particular interest to concentrate on scenes that are specific to an episode rather than scenes occurring in many of the episodes under consideration. One could easily foresee that a summary containing the most common scenes will almost certainly contain the opening and ending scene, since they are present in all episodes. Those scenes are, however, not of interest to a viewer attempting to determine whether or not a particular episode has already been screened. Our objective is therefore rather different to the one of "conventional" video summaries research [5, 3, 4]. In the context of multi-episode summaries, summarising videos independently from one another does not seem appropriate as there might be a lot of redundant information in the summaries. In this context information about what is common, what is unique between episodes should be identified and employed in the automatic summarisation process.

This paper is divided in 4 sections. The second section describes the methodology devised to perform automatic episode specific summaries. The performance measure employed to compare the "quality" of the candidate summaries is reported in the third section of this paper. In section 4, the results of the algorithm will be presented. Additionally, a sensitivity evaluation of the clustering technique will be reported. Finally, conclusions about the work achieved are presented along with future improvements.

## 2 THE METHODOLOGY

The approach we have developed in order to perform multi-episode video summarisation is divided in three major steps. The first of those is the segmentation of all the videos under consideration.

Once the videos have been segmented a representative key-frame is selected from each shot. The key-frames represent the salient content of the shot. The use of key-frames reduces the complexity of video content processing. Many techniques have been presented in the literature for key-frame selection [8, 7]. They vary greatly in term of the number of frames representing each shot and their computational complexity. As far as the work presented here is concerned, each

shot will be represented using its median key-frame. If the threshold used to perform the shot segmentation is small enough all the frames within a shot should look very much alike. Taking the median frame appears like a reasonable assumption in this case.

Colour histograms are employed to capture the colour distribution characteristics of each key-frame. The similarity between any pair of shots is computed by comparing their corresponding colour histograms. This is a similar approach to the one of Swain and Ballard for content-based image retrieval [6]. In order to capture some locality information key-frames are divided in nine equal regions from each of which a colour histogram is computed. As a result, characteristic key-frames are represented using a vector based on the concatenation of the nine histograms. As an alternative or in addition to the colour histograms, one could consider using histograms representing the quantity of movement within each shot.

It is important to note at this point that since shots are represented by key-frames (a single image) the duration information associated is missing. The solution adopted here is one where we retain the duration of each shot as a weight parameter together with the vector representation.

Having completed the shot representation step, further processing needs to be done in order to organise all the shots according to their similarity. We achieve this by first assigning key-frames from all the episodes under consideration to the cluster where the distance to the centre key-frame is less than a given threshold and creating a new one if no existing clusters satisfy this condition. In an attempt to refine the "quality" of the cluster a "K-means like" clustering algorithm is iterated until no key-frame change clusters. Key-frames are therefore assigned to clusters where other key-frames have a similarity value based on histogram comparison smaller than a threshold. We choose a small threshold value in order to enforce strong similarity among shots within clusters. The comparison is performed using the standard $L_2$ norm:

$$L_2(H_1, H_2) = \sqrt{\sum_{i=1}^{n} (H_1(i) - H_2(i))^2}$$

where $H_1$ and $H_2$ are two key-frame vectors of size $n$.

When all the shots $s$ have been assigned to a cluster $C(s)$, the selection of the most suitable shots for each summary $r(v)$ can be performed. This is the final step of our algorithm. The task, here, is to locate clusters containing only shots that belong to the episode for which we wish to create a summary. The best cluster $C$ contains shots originating only from the episode $v$ for which the summary $S$ is being created and has the most important duration with respect to the other clusters. The duration of a cluster is computed using the number of key-frames contained and the length of the shots they represent. To be more formal we can write the weight of a class, its specific duration $W(C, v)$ as:

$$W(C, v) = \sum_{s \in v} u(C, v) d(s)$$

where $d(s)$ is the duration of shot $s$ and with

$$u(C, v) = \begin{cases} 1 & \text{if } \forall s \in C \ s \in v \\ 0 & \text{otherwise} \end{cases}$$

The concatenation of representative shots from cluster for which $W(C, v) \neq 0$ constitutes the most specific summary for that video episode. However, it is not practical to present all those shots to the user, so only the best $k$ are selected to constitute the summary $r(v)$. This process is repeated for all the video episodes $v$ under consideration. Finally, the key-frames closest to the centre of the selected clusters are used to constitute the still visual summary as shown in figure 1 and 2.

It is clear that in such a framework the value of the threshold may be of capital importance. For example, a "small" threshold will generate a lot of "small" clusters. As the value of the threshold increases the number of classes (clusters) decreases, it is therefore more difficult to find clusters representing only one video. The extreme case is one where all key-frames belong to the same and unique cluster. We will study the effect the clustering threshold value in the experimental section.

# 3 EVALUATION MEASURE

The evaluation of a summary is a sensitive issue and there remain many uncertainties about the method to employ. Two, rather opposite, approaches have been reported in the literature. The first relies on users evaluating the quality of the resulting summaries. The second is based on the computation of a mathematical criteria. It is clear that the later is far more convenient to use. It is however not always clear how the performance value attributed to a summary can be interpreted.

Here, we make use of the idea of summary coverage to compute the quality of the summary. The coverage of a summary is directly related to the duration of the classes (clusters) it is composed of. In order to easily compare the coverage of two summaries, a normalisation is performed (using the total length of the corresponding video).

# 4 EXPERIMENTS

In this section we present the results obtained on summarising six episodes from the TV serie Friends. The

data was digitally recorded from a television channel in MPEG1 format at a frame rate of 14 frames/sec. We fixed the size of the summaries to six key-frames for convenience reasons (a six by six grid fits perfectly a standard TV/Computer display). The summaries are computed following the steps described in section 2.

Figure 1 and figure 2 show the resulting summaries for a given threshold. Each line corresponds to a video with the left-most frame representing the most pertinent shot (key-frame).

Figure 1: Summary created with a threshold of 6000. Coverage=33% of the original videos.

One can see by comparing the two sets of summaries that as expected the threshold value used during the clustering process has some effect on the selection of the classes used to build the summaries. For example summary frames from the first episode remain practically unchanged whereas for the sixth one, only a single frame is common to both summaries. However the coverage of both summaries isn't dramatically altered by this modification of the clustering threshold value. We shall look in some more details how much effect the choice of the clustering criterion has on the coverage of the summaries with respect to the videos.

It is interesting to visualise the manner in which the episode specific classes are distributed over the videos. Figure 3 shows the position of the shots which belong to classes containing only shots from the same video with the dark zones (red) corresponding to the episode specific shots (in other words, those for which $W(C, v) \neq 0$ is satisfied).

Figure 4 shows the effect of the clustering threshold on the coverage of the summaries (diamond) and the episode specific classes (square) with respect to the six

Figure 2: Summary created with a threshold of 14000. Coverage=34% of the original videos.

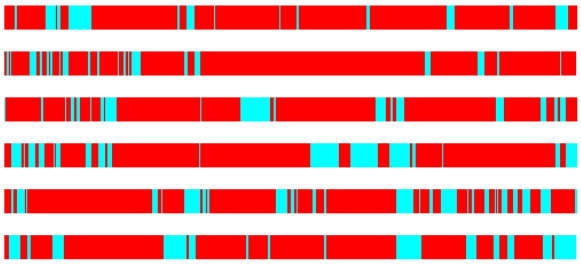

Figure 3: Distribution of the episode specific shots (dark zone or red) on the six video

episode. In other words, for the diamond curve the sum of the duration of the 36 clusters is plotted as a function of the clustering threshold. Similarly, the curve marked with squares indicates the duration of all the clusters containing key-frames from a unique video. It is worth pointing out that the total duration of the six videos of "Friends" used in this experiment is 83150. This plot indicates that there is a range of threshold values for which the coverage of the summaries remains rather stable. Judging from the summaries shown in figure 1 and 2 this does not necessarily implies that the same classes and key-frames will be selected. However, it should be pointed out at this point that the reason for this plateau is directly related to the fact that in the videos considered there are some very long shots. Those long shots are likely to constitute classes with a single key-frame for small threshold. It is therefore clear that even for very small clustering threshold values the coverage will remain rather high. The coverage

of the classes containing only representative from a single video reaches over 99.99% for a clustering threshold of 5000. This indicates that the threshold is so small that only a very small fraction of the shots from a video can still be seen as similar to shots from other episodes.

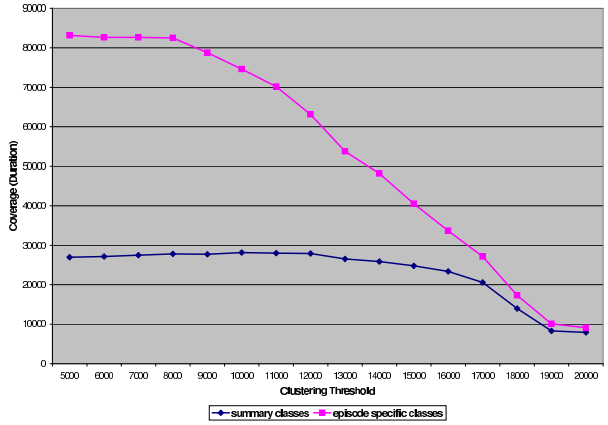


Figure 4: Effect of the clustering threshold value on summaries coverage

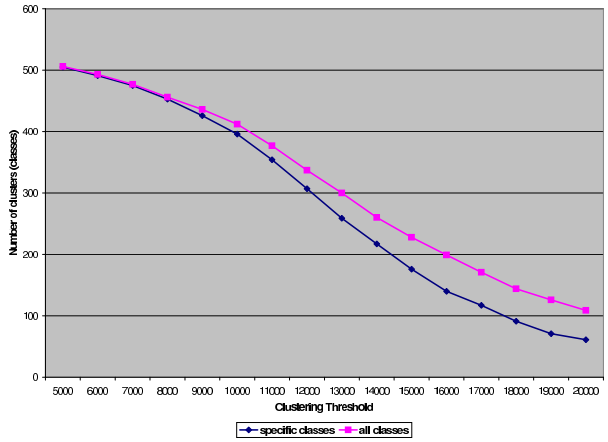


Figure 5: Effect of the clustering threshold value on the clusters

The plot presented in figure 5 depicts the effect of the clustering threshold on the clusters and their content. This plot provides some insight on how many classes are created (square marked curve) and how many are available to build the episode specific summaries. As expected, as the threshold value increases clusters contain less similar key-frames and therefore classes become less specific.

# 5 CONCLUSION

A novel approach to automated video summaries creation has been presented. It is based upon the concept that in the context of multiple TV series it is more interesting to find out about what is specific to a particular episode than what is common to all episodes under consideration. We showed that even for very small summaries (six key-frames per video) a high level of original video coverage is achievable. Additionally, an investigation about the effect of the thresholding criterion indicated that the range of possible value is quite large. Indeed, using threshold value between 5000 and 12000 resulted in a summary coverage of over 33% of the original videos. We also presented that despite a rather constant coverage value, the key-frames contained in the summaries were not necessarily the same for different clustering parameters.

# 6 ACKNOWLEDGEMENTS

This research was supported by Eurecom's industrial members: Ascom, Cegetel, France Telecom, Hitachi, ST Microelectronics, Motorola, Swisscom, Texas Instruments, and Thales.

# References

[1] J.S. Boreczky and L.A. Rowe, "Comparison of video shot boundary detection techniques." In *SPIE Conference on Visual Communication and Image Processing*, 1996.

[2] M.R. Naphade, R. Mehrotra, A.M. Ferman, J. Warnick, T.S. Huang and A.M. Tekalp "A high performance algorithm for shot boundary detection using multiple cues.", In *IEEE ICIP*, vol. 1, pages 884–887, 1998.

[3] N. Vasconcelos and A. Lippman, "Bayesian modeling of video editing and structure: semantic features for video summarization and browsing.", In *IEEE ICIP*, vol. 3, pages 153–157, Oct 1998.

[4] S. Uchihashi and J. Foote, "Summarizing video using a shot importance measure and a frame-packing algorithm.", In *IEEE ICASSP*, vol. 6, pages 3041–3044, 1999.

[5] M.A. Smith and T. Kanade, "Video skimming and characterization through the combination of image and language understanding.", In *IEEE International Workshop on Content-Based Access of Image and Video Database*, pages 61–70, 1998.

[6] M.J. Swain and D.H. Ballard, "Color indexing.", *International Journal of Computer Vision*, 7(1):11–32, 1991.

[7] W. Wolf, "Key frame selection by motion analysis.", In *IEEE ICASSP*, vol. 2, pages 1228–1231, 1996.

[8] Z. Yueting, R. Yong, T.S. Huang and S. Mehrotra, "Adaptive key frame extraction using unsupervised clustering." In *IEEE ICIP*, vol. 1, pages 866–870, 1998.