GENERATING SUMMARIES OF MULTI-EPISODES VIDEO

Itheri Yahiaoui, Bernard Merialdo, Benoit Huet

Institut Eurécom
Département Communications Multimédia
BP 193 – 06904 Sophia-Antipolis- France
{Itheri.Yahiaoui,Bernard.Merialdo,Benoit.Huet}@eurecom.fr

ABSTRACT

The growing availability of multimedia data such as video on personal computers and home equipment creates a strong requirement for efficient tools to manipulate this type of data. Automatic summarization is one of these tools, which automatically creates a short version or subset of key-frames which contains as much information as possible as in the original video. Several approaches have been proposed to define and identify what is the most important content in a video.

In this paper, we propose a new approach for the automatic creation of summaries for multi-episode videos, such as TV series. In this case, it is necessary to identify similarities and differences among videos (what's common, what's unique, how they differ) in order to find which elements best characterize a particular video with respect to the others. We describe our proposed method, provide the results of its application on a sample set of videos, and suggest a new criterion to evaluate the quality of the summaries that have been created.

1. INTRODUCTION AND RELATED WORK

With increased computing power, electronic storage capacity and bandwidth of transmission, multimedia information and particularly digital video is becoming more and more common and very important for education, entertainment and many other applications. This large amount of multimedia data has fueled efforts to provide and develop techniques for efficiently processing and manipulating this type of data. In particular, automatic summarization is a useful tool which allows a user to grasp rapidly the essential content of a video, without the need for watching the entire document.

A number of techniques have been proposed and developed to automatically create video summaries [4][12]. They fell in two categories:

- A mathematically oriented approach uses similarities within the video to compute a relevance value of video segments or frames. Possible criteria for computing this relevance include the duration of segments, the inter-segment similarities, and combination of temporal and positional measures. Examples of this approach are the use of the SVD (Singular Value Decomposition) by Gong and Liu [14], or the shot-importance measure by Uchihashi and Foote [10].
- A rule-based approach combines evidences from several types of processing (audio, video, natural language) to detect certain configuration of events, which are included in the summary. Examples of this approach are the "video skims" of the Informedia Project by Smith and Kanade [7], and the movie trailers of the MoCA project by Lienhart et al [9]. Sometimes multiple characteristics of the video stream are employed simultaneously; the video itself but also the audio signal (speech, music, noise, etc...) and even the textual information contained in closed caption. In such a case, some rules have to be defined to combine the different characteristics in order to identify pertinent segments. The automatic creation of movie trailers is a possible application of such methods.

In both approaches, a first phase is generally introduced to pre-process the video at a low level and identify segments, either through frame clustering and/or cut detection. The result of the summarization process can be a video itself, or a sequence of key-frames from the selected segments.

Our approach follows the mathematical pattern presented above. However, as we are considering the summarization of several videos at once, we introduce new criteria to define the relevance of video segments based on the analysis of all segments from all the videos.

The paper is organized as follows. Our approach for multi-episode video summarization is introduced in section 2 and described in section Then we report some experiments to validate our method, and introduce a new criterion to evaluate the quality of the summaries.

2. MULTI-EPISODE VIDEO SUMMARISATION

Summarizing video content is important to several applications including archiving and providing access to video teleconferences, video mail, video news, etc... Whereas summarization of a single video has received increasing attention [1][4][7][8][9][10], comparatively few investigators have examined the problem of multi-episode video summarization [2][6]. The situation of multi-episode videos occurs for example in soap operas or TV series, where much information (scenes, actors) is present in the same or in similar manner in several videos. In this case, summarizing videos independently one from each other might include redundant information in the summaries. New methods have to be designed to take advantage of these similarities to produce more efficient set of summaries.

Our approach identify similarities and differences among videos (what's common, what's unique, how they differ) by comparing and classifying representations of video content. Because the same information often appears in slightly different forms in multiple segments, we have developed algorithms that can

eliminate redundant information across series to provide a concise summary.

Our approach for multi-episode video summarization is divided in five steps:

The first step is a pre-processing of video streams. Because the opening (jingle) and ending (credits) scenes are present in all episodes, they are important elements of the video set. Those scenes are however not of interest to a viewer attempting to understand the content of a particular episode. Therefore, we eliminate the opening and ending scenes from the video data to be processed.

The next step consists of analyzing the content of the video to create characteristic vectors to represent visual information included in the video frames. We achieve this using color histograms to capture the color distribution of individual frames. We also capture some locality information by dividing each frame into nine equal regions on which the color histograms are computed. These nine histograms are then concatenated to make up the characteristic vector of the corresponding frame. In order to reduce computation and memory cost, we sub-sample the video such that only one frame per second is processed. Additionally, when consecutive frames are very similar, i.e. the difference of their color histograms is below a given threshold, we keep only the first of them, and we preserve the duration information by adding a counter to the histogram.

We now cluster frames (represented by their color histograms) with an initial step where we create a new cluster when the distance of a frame to existing clusters is greater than a threshold, followed by several k-Means type steps to refine the clusters. This clustering operation produces classes of video frames with similar visual content. The frequency of occurrence of frames from each video within classes allows to compute the importance of the various classes.

Then we select for each episode the most pertinent classes. There are two possible methods for the selection of the classes used to construct the summary. More details will be reported in the next sub-section.

Finally, the global summary can be constructed and presented to the user, as an hypermedia document composed of representative images of videos content selected in the previous step or as an audio-video sequence of reduced duration, obtained by concatenating video segments corresponding to the selected frames. In this paper summaries are presented in the form of a table of images (frames extracted from the video) where each row represents a particular episode. The number of rows in the table is the number of different episodes under consideration. The number of images describing each episode (the number of columns in the table) is however entirely user definable.

Video Segment selection

Once video frames have been clustered, the videos might be described as sets of frame classes. We consider that important classes are those which appear most often, as they represent a longer portion of the video. We define the value of the summary as:

$$\sum_{v} \sum_{c \in r(v)} d_v(c)$$

where, v represents the videos, c represents the different classes obtained from clustering, r(v) is the summary of the video v and $d_v(c)$ is the number of frames from video v which belong to class c.

This maximization is subject to certain constraints. First, we would like to let the user define the size of the summary. This is because, in many applications, the size of an interesting summary depends on the time that the user has available for watching. Therefore, we specify that each summary contains k segments, with k provided as an external constraint (given by the user). The problem then becomes:

$$\underset{r}{Max} \sum_{c,v} r(c,v) d_v(c)$$

with

$$r(c,v) = \begin{cases} 1 & if \quad c \in r(v) \\ 0 & if \quad c \notin r(v) \end{cases}$$

and the constraint that

$$\sum_{c} r(c,v) = k \quad \forall v$$

We have defined two variants of the selection process:

- In the **independent selection**, we allow a class to appear in several different summaries. This corresponds to a straight application of the previous formulation, where nothing prevents a class from being selected for several video summaries. Of course, when a class is selected for a specific video v, the video segment which will be included in the summary of v is the segment of v whose frames are closest from the centroid of the class (frames in this class but belonging to other videos, are excluded).
- In the **dependent selection**, we impose that classes should not be used more than once. Since we do not allow for classes to represent more than a single episode, the affectation of classes to videos is realized so that the frequency of occurrence is maximized. This corresponds to the previous formulation with the following additional constraint:

$$\sum_{v} r(c,v) \leq 1 \quad \forall c$$

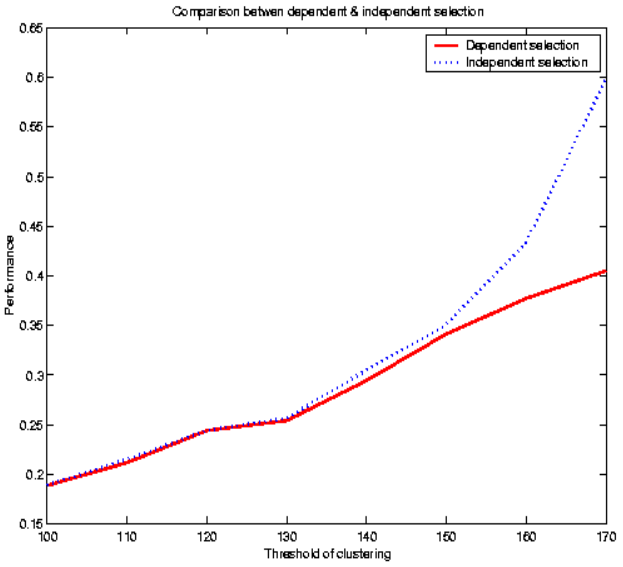
3. EXPERIMENTS

In this section we present our results on multi-episode video summarization. As test data, we recorded six episodes of the TV serie "*Friends*". These recordings were Mpeg1 compressed, with a digitization rate of 14 frames/sec. The distance used for clustering color histograms is the Euclidean distance. We fixed the size of the summaries to six segments, (which provides a convenient display on a screen). The selection of segments for the summaries is performed in the following way:

- In the independent selection case, classes are sorted by decreasing order of frequency in video $d_v(c)$ and the first six are selected for each summary.
- The dependent selection summaries are constructed from the independent selection summaries by looking at ambiguous

classes. When an ambiguous class is found (a class which has been selected for several videos) it is assigned to the summary of the video for which it is most frequent and removed from the other summaries. This process is repeated until ambiguous classes are found.

The following graph shows the coverage of the summaries for the two methods and various values of the clustering threshold.



As an illustration, the following image shows the dependent selection summaries that have been built for a threshold of 140:

Summary created with Dependent selection

This threshold was manually selected by observing pairs of frames with various distances, so as to be as consistent as possible with a human judgement of frame similarity. In this example, 8 classes (out of 36) were removed from the independent selection summaries to build the dependent selection summaries.

4. SIMULATED USER EVALUATION

The problem of evaluating video summaries is very delicate, and serious questions remain concerning the appropriate methods and type of evaluation [5][11]. There are currently two approaches:

- A *user-based* evaluation requires a number of users to judge the summaries, or perform a given task with the knowledge of the summary. This is the definite method, as it can directly measure the effect of the summaries on the task for which they have been designed. However, this method is very difficult to put into practice because of its very high cost. It is therefore only employed in rare cases where users are easily available, or where it is very important to validate the summarization method.
- A *criteria-based* evaluation judges a summary with respect to some mathematical criteria, for example a distortion or frequency measure. Often, this is the same criterion as the one used in the construction of the summary. This method is very simple to implement, because it only requires computer calculations, and it allows to very easily compare the performance of several algorithms. However, it has one serious drawback, namely that the numerical value that represents the performance has often little meaning to the user, so that it is difficult for a user to interpret the quality of a summary from this value.

We propose a novel approach, which we call the Simulated User Evaluation, which is based on experiments which simulate the behavior of users. If we place some assumptions and rules about how a real user could react in front of a specific task, we can evaluate the performance of those virtual users in an automatic way.

The experiment we design is the following:

- A user looks at all the summaries,
- We show the user one image extracted from the videos, and the user has to guess which video it comes from,
- We assume that the user tries first to find identical images (equality threshold) than if none are found, he looks for similar images (similarity threshold).

For this experiment, and with these assumptions, we can automatically compute the performance of such a user on this task. This is achieved using the following algorithm.

1- We try all frames in the video as sample frame, and the (simulated) user tries to guess which video they come from.

2- We compute the distance between this sample frame and all the frames in the summaries,

3- If there exist a summary frame for which the distance is smaller than a given image equality threshold (selected experimentally), the user will indicate the corresponding video as answer,

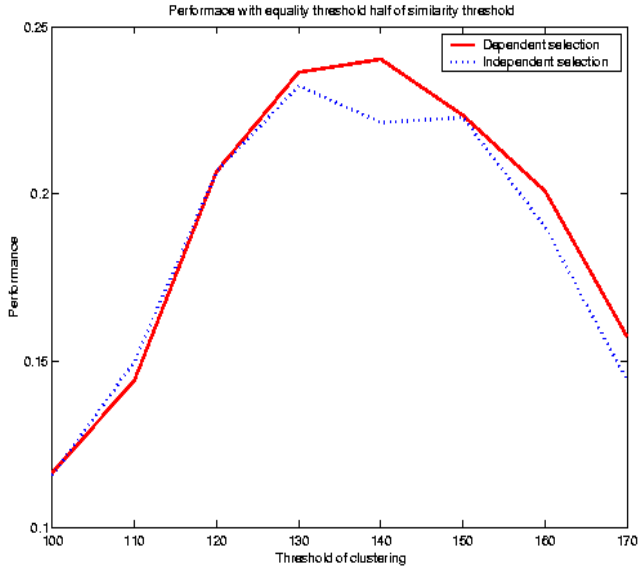
If no summary frame is closer than the identity threshold, but there is one for which the distance is below a similarity threshold, the user will indicate the corresponding video as answer.

In other situations, the user will provide no answer.

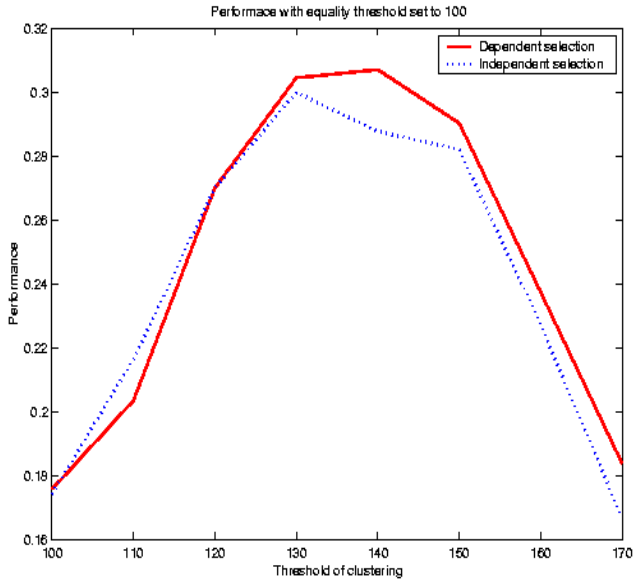
4- We evaluate the performance by counting the number of correct answers. When the answer could be ambiguous, (more than a single frame below a threshold) the performance is

increased by only 1/n, where n is the number of frames causing the ambiguity.

The graphs below show the performance measure with this new method. In the first one, the equality threshold is set to half the similarity threshold.



In the second graph, the equality threshold is set to a constant value which has been experimentally obtained.



These results show that by watching the summaries, a user is able to identify a video from a single frame with a probability of 0.25 - 0.3.

5. CONCLUSION

Automatic video summarization is a very important tool. In this paper, we have proposed a novel technique to automate multi-episode video summary creation. We have presented two distinct selection methods. To evaluate the quality of the summaries, we have devised a new approach which simulates a user's human behavior. This allows an automatic evaluation which leads to performance levels which are hopefully easier to understand than other approaches.

Acknowledgements:

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

6. REFERENCES

[1] Anastasios D. Doulamis, Nikolaos D. Doulamis and Stefanos D. Kollias. Efficient Video Summarization based on a Fuzzy Video Content Representation. IEEE International Symposium on Circuits and Systems, vol. 4, pp. 301-304, 2000.

[2] Bernard Merialdo. Automatic Indexing of Tv News. Workshop on Image Analysis for Multimedia Integrated Services, pp. 99-104, 1997.

[3] Emile Sahouria and Avideh Zakhor. Content Analysis of Video Using Principal Components. IEEE Transactions on Circuits and Systems for Video Technology, vol. 9, no. 8, pp. 1290 -1298, Decembre 1999.

[4] Giridharan Iyengar and Andrew B. Lippman. Videobook: An Experiment in Characterization of Video. IEEE International Conference on Image Processing, vol. 3, pp. 855-858, 1996.

[5] Inderjeet Mani and Mark T. Maybury. Advances in Automatic Text Summarization. The MIT Press, 1999.

[6] Mark T. Maybury and Andrew E. Merlino. Multimedia Summaries of Broadcast News. IEEE Intelligent Information Systems, pp. 442 -449, 1997.

[7] M.A. Smith and T. Kanade. Video Skimming and Characterization through the Combination of Image and Language Understanding. IEEE International Workshop on Content-Based Access of Image and Video Database, pp. 61-70, 1998.

[8] Nuno Vasconcelos and Andrew Lippman. Bayesian Modeling of Video Editing and Structure: Semantic Features for Video Summarisation and Browsing. IEEE International Conference on Image Processing, vol. 3, pp. 153-157, 1998.

[9] Rainer Lienhart, Silvia Pfeiffer and Wolfgang Effelsberg. Video Abstracting. In Communications of ACM, vol. 40, no. 12, pp 54-62, December 1997.

[10] Shingo Uchihashi and Jonathan Foote. Summarizing V Video Using a Shot Importance Measure and a Frame-Packing Algorithm. IEEE International Conference on Acoustic, Speech and Signal Processing, pp. 3041-3044, 1999.

[11] Udo Hahn and Indejeet Mani. The Challenges of Automatic Summarization. IEEE Computer, vol. 33, no. 11, pp: 29-36, November 2000.

[12] V.Di Lecce, G.Dimauro, A. Guerriero, S.Impedovo, G.Pirlo, A.Salzo. Image Basic Features Indexing Techniques for Video Skimming. IEEE International Conference on Image Analysis and Processing, pp. 715-720, 1999.

[13] Yueting Zhuang, Yong Rui, Thomas S. Huang and Sharad Mehrotra. Adaptive Key Frame Extraction Using Unsupervised Clustering. IEEE International conference on Image Processing, vol. 1, pp. 866-870, 1998.

[14] Yihong Gong; Xin Liu. Generating Optimal Video Summaries. IEEE International Conference on Multimedia and Expo, vol. 3, pp. 1559-1562, 2000.