Automatic construction of multi-video summaries

Itheri Yahiaoui , Bernard Merialdo , Benoit Huet

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

Abstract:

In this paper, we present a novel methodology for the automatic construction of summaries for multi-episode videos, such as TV series. This work is based on the Simulated User Principle to evaluate the quality of a video summary in a way, which is automatic, yet related to user's perception. The method is studied for the case of multi-episode video, where we don't only select what is important in a video, but rather what distinguishes this video from the others. Experimental results are presented to support the proposed ideas.

Keywords: Video analysis, Automatic Summarization, Evaluation

Résumé:

Dans cet article, nous présentons une nouvelle approche pour la création automatique de résumés de plusieurs vidéos, comme par exemple des épisodes de séries télévisées. Cette méthodologie est basée sur le principe d'utilisateur simulé afin d'évaluer la qualité du résumé vidéo d'une manière automatique mais inspirée de la perception humaine. Il faut noter que pour les résumés multi-vidéos, il est nécessaire non seulement d'identifier les informations qui sont importantes dans une vidéo, mais aussi celles qui caractérisent cette vidéo par rapport aux autres. Afin de valider les idées proposées, des résultats expérimentaux sont présentés.

Mots-clés: Analyse de vidéo, Résumé automatique, Evaluation

1 Introduction & related work

With increased computing power, electronic storage capacity and communication bandwidth, 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 fuelled efforts to provide and develop techniques for efficiently processing and manipulating this type of data. In particular, automatic summarisation 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.

Automatic summarization is subject to very active research, and several approaches have been proposed to define and identify what is the most important content in a video. Existing approaches to video summarization can be classified in two categories, one where a rule-based approach is employed and the other where mathematically oriented techniques can be found. The rule based approaches combine 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 [Smi,1998], and the movie trailers of the MoCA project by Lienhart et al [Lie,1997]. The automatic creation of movie trailers is a possible application of such methods. The mathematically oriented approaches, on the other hand, use 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 [Gon,2000], or the shot-importance measure by Uchihashi and Foote[Uch,1999]. The methods we propose here fall in the later category.

A key issue in automated summary construction is the evaluation of the quality of the summary with respect to the original data. There is no ideal solution to this problem, so a number of alternative approaches are available. With user based evaluation methods, a group of user is asked to provide an evaluation of the summaries, either directly or by comparison between several summarization methods. In this case, the evaluation is directly computed from the user's responses. Another user based evaluation is to ask a group of users to accomplish certain tasks (i.e. answering questions) with or without the knowledge of the summary, and measure the effect of the summary on their performance. Alternatively, for summaries created using a mathematical criterion, the corresponding value can be used as a measure of quality for the summary. However, all these evaluation techniques present drawbacks; User-based one's are difficult and expensive to set-up and their bias is non trivial to control, whereas mathematically based one's are difficult to interpret and compare to human judgement.

Finally, while summarization of a single video has received increasing attention [Dou,2000] [Iye,19996] [Smi,1998] [Vas,1998] [Lie,1997] [Uch,1999], comparatively little work has been devoted to the problem of multi-episode video summarization [Mer,1997][Gon,2000]. which raises other interesting difficulties.

In this paper, we propose a new approach for the automatic creation and evaluation of summaries based on the Simulated User Principle. This method addresses the problem related to the evaluation of the summary and is applicable to both cases of single video and multi-episode videos.

This paper is organized as follows. Section 2 describes some basics about video summaries and the simulated user principle approach. Then, in section 3, we present

some experimental results. Section 4 studies the robustness of summaries. Conclusions and possible future extensions of the work are then presented in section 5.

2 Simulated User Principle for Multi-Episode Summarization

A video summary should contain the most important information included in the video it originates from. The aim is to enable a person to get an idea of what the original data is about simply by viewing its summarized version. The situation is more complex when we deal with multi-episode videos, such as TV series. Imagine for example that your set-top box has recorded a number of episodes of your favorite series, and you ask for a summary of these recordings. In this case, the summary for each episode should contain elements which best characterize this episode with respect to the others. Summarizing videos independently from one another, like in the single video case, does not seem appropriate, as this is likely to generate summaries with a lot of redundant information. It is therefore necessary to identify what the similarities and differences are among the videos (what's common, what's unique, how they differ) in order to build efficient summaries.

In the introduction we have reported a number of alternatives for creating and evaluating summaries. In this paper, we create and evaluate summaries using the Simulated User Principle based on a mathematical criterion which emulates a real user's video summary evaluation. In the Simulated User Principle, we define a real experimentation, a task that some user has to accomplish, and on which a performance measure can be defined. Then, we use reasonable assumptions to predict what the user behaviour could be on this task. In other words, we use a Simulated User to accomplish this task, of which we can exactly predict how he will behave. This allows us to mathematically define the performance of this Simulated User on the experiment.

As an application of the Simulated User Principle to the problem of multi-episode video summarization, we propose the following scenario for the Simulated User Experiment:

- The user is shown all the summaries,
- He is shown a randomly chosen excerpt of a randomly chosen video,
- He is then asked to guess which video this excerpt was extracted from.

The simulated behavior of the user is the following:

- If the excerpt contains images which are similar to one or several images in a single summary, he will provide the corresponding video as an answer (but it is not certain that this is the correct answer),
- If the excerpt contains images which are similar to images in several summaries, the situation is ambiguous and the user cannot provide a definite answer,
- If the excerpt contains no image which is similar to any image in any summary, the user has no indication and cannot provide a definite answer.

We define the performance of the user in this experiment as the percentage of correct answers that he is able to provide when he is shown all possible excerpts of all

videos with predefined duration. Note that only in the first case described above is the user able to identify a particular video. But this answer might not be necessarily correct, because an image in an excerpt of one video can be similar to an image in the summary of another video.

This Simulated Experiment assumes that the user has a perfect visual memory, so that he can immediately identify similar images when they are shown to him. It also assumes that we can closely approximate the user judgement on similarity. Note that variants of this procedure could be considered; for example, in the ambiguous second case described above, it could be possible to use frequency information alternatives to rank the various summaries and still select a video. We did not explore such issues yet.

2.1 Automatic Summarization

Now that the Simulated User Experiment is defined, we need an algorithm to automatically construct summaries with good (and if possible optimal) performance for this experiment. The multi-episode summary building process proposed is divided into five phases: video streams pre-processing, feature vectors construction, classification, selection and summary presentation. The fourth phase that performs the selection of the elements to include in the summaries is specific to our evaluation criterion.

Video Streams Pre-processing:

The first phase 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 manually eliminate the opening and ending scenes from the video data to be processed.

Feature Vectors Construction:

The next phase consists of analysing the content of the video to create characteristic vectors to represent visual information included in the video frames. We achieve this using colour histograms to capture the colour distribution of individual frames. We also capture some locality information by dividing each frame into nine equal regions on which the colour 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 colour 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.

Classification:

We now cluster frames (represented by their colour 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 like” steps to refine the clusters. This clustering operation produces classes of video frames with similar visual content.

Video Segment Selection:

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

Summary Presentation:

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 classes. 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.

2.2 Selection Method:

Once video frames have been clustered, the videos might be described as sets of frame classes. The most pertinent classes will be kept. We shall now see the methodologies devised to compute this pertinence value.
Assume that the excerpts that we consider have duration d. If the video contains N frames, there are N-d+1 different excerpts:

- E_1 contains frames f_1, f_2, … f_d,
- E_2 contains frames f_2, f_3, … f_{d+1},
- And so on, up to E_{N-d+1} which contains frames f_{N-d+1}, f_{N-d+2}, … f_N.

Two frames are considered to be similar if and only if they belong to the same class:

$$f_i \text{ and } f_j \text{ similar} \Leftrightarrow C(f_i) = C(f_j)$$

Figure 1 illustrates the relations between excerpts, frames and classes.

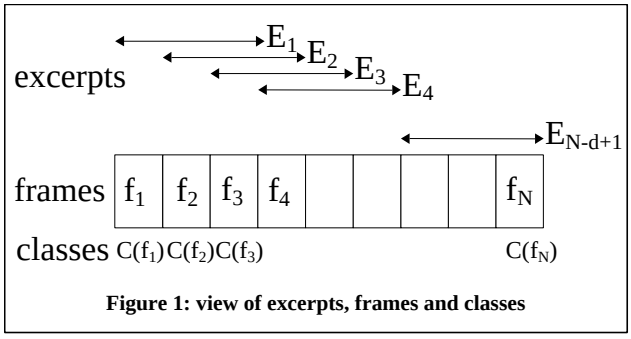

Figure 1: view of excerpts, frames and classes

If we denote by E_i^v an excerpt of a video v, and S_v a summary for video v, the cases that have been described above can be formally characterized by the properties:

- Unambiguous case:

$$\exists v'\ \exists j \quad f_j \in E_i^v \text{ and } C(f_j) \in S_{v'}, \forall v'' \neq v' \forall f_j \in E_i^v \quad C(f_j) \notin S_{v''}$$

- Ambiguous case:

$$\exists v' \exists v'' \neq v'\ \exists j \quad f_j \in E_i^v \text{ and } C(f_j) \in S_{v'} \text{ and } C(f_j) \in S_{v''}$$

- Unknown case:

$$\forall v' \quad \forall f_j \in E_i^v \quad C(f_j) \notin S_{v'}$$

The performance of the user is the number of correct answers, so it is the number of unambiguous cases for which v'=v:

$$\text{Card}\left\{ \begin{array}{l} (i,v) : \exists j \quad f_j \in E_i^v \text{ and } C(f_j) \in S_v, \\ \qquad \forall v' \neq v \forall f_j \in E_i^v \quad C(f_j) \notin S_{v'} \end{array} \right\}$$

The construction of the summaries is delicate, because when we choose to add a class in a summary, we have to consider not only the coverage of this class on this video, which should be high, but also take into account the coverage of this class on the other videos, which should be low to minimize erroneous choices. The coverage of a class on a video v is defined as:

$$\text{Cov}_v(C) = \text{Card}\left\{ i : \ \exists j \quad f_j \in E_i^v \quad \text{and} \quad C(f_j) = C \right\}$$

An exhaustive enumeration of all possible sets of summaries is computationally untractable, so we use a sub-optimal algorithm to build a good set of summaries. Our algorithm proceeds as follows:

- Each summary is initially empty,
- We select each video v in turn, and add to its current summary S_v the one class C with maximal value:

$$\text{value}_v(C | \{ S_v \}) = \text{Cov}_v(C|S) - a \sum_{v' \neq v} \text{Cov}_{v'}(C|S)$$

  where S is the set of all classes already included in any of the summary:

$$S = \bigcup_v S_v$$

- When all summaries have the desired size, we iteratively replace any chosen class if we can find another class with better value.

The coefficient α is used to impose a penalty to classes whose coverage on other videos is large, because they are likely to generate ambiguous or erroneous cases in the simulated experiment.

Now, that our proposed algorithm is described and in order to compare it with other algorithms, we propose two variants based on a similar idea and two others based on a different approach. This four algorithms are presented just below:

- To compare dependant and independent selection and as a baseline experiment to validate the importance and specificity of multi-episode video summaries, we construct single-video summaries of each video (using global similarity classes). In such approach, when we select classes to be included in the summary for a video, we

ignore which classes are also present in the other summaries. Therefore, a class can be present twice or more in the complete set of summaries.

- In order to eliminate all ambiguous cases in the simulated experiment, we develop an algorithm based on the computation of coverage, similarly to the previous ones, but which is more sensitive to ambiguous cases. During the selection phase, candidate classes should not be present in other summaries and should not be present in excerpts containing previously selected classes of other videos.

- Based on the work of Uchihashi and Foote [Uch,1999], who defined a measure to compute the importance of shots, we adapted our multi-episode summarisation method. Here, shots are constructed based on our classification by concatenation of successive frames belonging to the same class. The shot importance measure is slightly modified from the original work such that the weight Wi of a class, which is the proportion of shots from the whole videos that are in cluster i, is computed as $W_i = S_i \Big/ \sum_{j=1}^{C} S_j$ where C is the number of classes based on all frames from all video episodes under consideration and S_i is the total length of all shots in cluster i, found by summing the length of all shots in the cluster. Thus the importance I_j of shot j (from cluster k) is $I_j = L_j \log 1/W_k$ where L_j is the length of the shot j.

A shot is important if it is both long and not similar to most other shots. In our case, in order to represent each video by specific shots and the longest possible, we compute the importance shot factor for all possible shots, and then we select the most important shots from each video to be included to the corresponding summary.

- The major idea of this method is to do a parallel with text summarization methodologies [Man,1999], where the TF_IDF formula has proven to be very effective. For text summarization this approach is based on terms which represent the items, whereas for multi-video summaries items are classes. In the video case, the importance I of class c is computed as $I_c = L_c \log n/nc$ where L_c is the length (total duration) of the class c, n the number of videos and nc the number of videos containing at least one frame from the class c.

Having computed the importance of each class, we select the most important ones to be included in the global summary. In the case where the class is present in more than one video, we have to determine to which summary it should be affected. We do this by computing for each video the proportion of frames belonging to this class that are present in this video, and we take the most probable one.

3 Coverage experiments

In this section we present the evaluation results using the Simulated User Principle on multi-episode video summaries created with different algorithms. As test data, we recorded six episodes of the TV serie “Friends”. These recording were Mpeg1 compressed, with a digitization rate of 14 frames/sec. We fixed the size of the summaries to six segments (which provides a convenient display on a screen).

The following graph (figure 1) shows the respective performance of these methods when the duration of the excerpt used for evaluation varies.

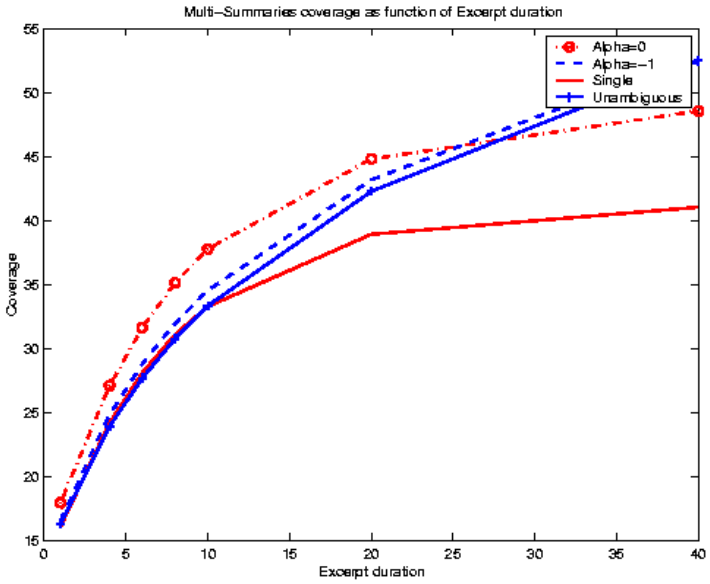

Figure 1

We note that the first two methods that build summaries based on a mathematical criterion inspired for the evaluation criterion itself with alpha equal respectively to 0 and –1 give the best performance. We also note that the multi-episode summaries (methods 1 and 2) are more efficient than the single video summaries (method 3). Method 5 based on shot importance performs very poorly: this is due to the fact that shots are selected on their length and low number of occurrence, and that rare shots are likely to have little coverage over a video. Method 6, inspired from TF-IDF provides rather reasonable results when compared with others. It should also be noted that results obtained with method 4 (unambiguous) can be compared to those of method 2, and that both give the best coverage for large excerpts duration.

4 Robustness of summaries

Having constructed multi-episode video summaries using a number of methods it is of interest to evaluate the performance of the summaries for unrestricted excerpt duration. To study the robustness of these methods, summaries were built using given excerpts duration and then evaluated using other excerpts duration. Figure 3 presents the results of this experiment for summaries based on method 1. Note that the construction method itself suggests that the coverage should be at maximum when identical excerpt duration is employed for both construction and evaluation. Except in the case of summaries created with excerpt duration of 1 second, all remaining methods provide rather similar and good performance.

This figure provides evidence that the choice of the excerpt duration is not a crucial factor for the performance of the summaries. Summaries optimized for a given duration will still provide a reasonable performance for others duration.

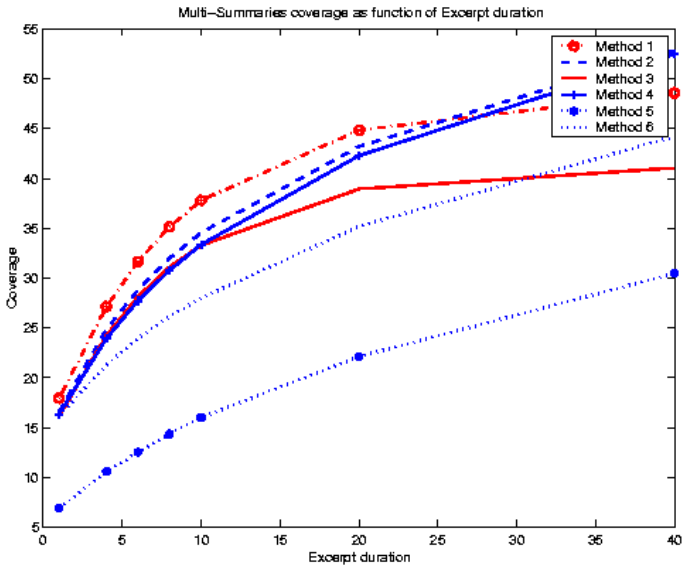

Figure 2

As an illustration, figure 2, show the six summaries constructed using the first method. Each row of the figure corresponds to a video episode and frames are displayed in temporal order

Figure 3

5 Conclusion

A comparison of some approaches to construct automatically multi-video summaries has been presented. Based on the Simulated User Principle we evaluate the results obtained with six alternative methodologies. Our experiments demonstrate that when both construction and evaluation are performed with the same principle the best results are achieved. Our proposed method clearly outperforms both the method of Uchihashi and Foote [Uch,1999] and a method inspired from the TD-IDF formula. Our evaluation of the robustness of the summaries shows that it is possible to obtain reasonable results with summaries created for specific excerpt duration. We envisage the creation of optimal summaries independently of the excerpt duration in order to achieve high coverage performance for any selected excerpt.

6 References

[Dou 2000] Anastasios D. Doulamis, Nikolaos D. Doulamis and Stefanos D. Kollias. Efficient video summarization based on a fuzzy video content representation. ISCAS 2000 – IEEE *International Symposium on Circuits and Systems*, May 28-31, 2000, Geneva, Switzerland.

[Iye 1996] Giridharan Iyengar and Andrew B. Lippman. Videobook: An experiment in characterization of video, IEEE *Intl. Conf. on Image Processing*, , September 1996.

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

[Vas 1998] Nuno Vasconcelos and Andrew Lippman. Bayesian modeling of video editing and structure: Semantic features for video summarisation and browsing. IEEE *Intl. Conf. on Image Processing*, 1998.

[Lie 1997] Rainer Lienhart, Silvia Pfeiffer and Wolfgang Effelsberg. Video abstracting. *Communications of* ACM, Dec 1997.

[Uch 1999] Shingo Uchihashi and Jonathan Foote. Summarizing video using a shot importance measure and a frame-packing algorithm. IEEE ICASSP 1999.

[Mer 1997] Bernard Merialdo. Automatic indexing of TV news. *Workshop on Image Analysis for Multimedia Integrated Services*, June 1997.

[May 1997] Mark T. Maybury and Andrew E. Merlino. Multimedia Summaries of Broadcast News. IEEE *Intelligent Information Systems*, 1997.

[Gon 2000] Yihong Gong; Xin Liu. Generating optimal video summaries. ICME 2000.

[Man 1999]I Mani and M. T. Maybury. *Advances in Automatic Text Summarization.* MIT Press, 1999.