COMPARISON OF MULTI_EPISODE VIDEO SUMMARISATION ALGORITHMS

Itheri Yahiaoui – Bernard Merialdo – Benoit Huet
Institut Eurecom
Departement Communication Multimedia
BP 193 – 06904 Sophia – Antipolis- France
{Itheri.Yahiaoui,Bernard.Merialdo,Benoit.Huet}@eurecom.fr

Abstract

This paper presents a comparison of some methodologies for the automatic construction of video summaries. The 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 describe what is important in a video, but rather what distinguishes this video from the others. Experimental results are presented to support the proposed ideas.

1 INTRODUCTION & RELATED WORK

The ever-growing availability of multimedia data, creates a strong requirement for efficient tools to manipulate and present data in an effective manner. Automatic video summarization tools aim at creating with little or no human interaction short versions which contains the salient information of original video. The key issue here is to select what should be kept in the summary and how can this relevant information be automatically extracted. To perform this task we consider several algorithms and compare their performance to define the most appropriate one for our application.

A number of approaches have been proposed to define and identify what is the most important content in a video. However, most have two major limitations. First, evaluation is difficult, in the sense that it is hard to judge the quality of a summary, or, when a performance measure is available, it is hard to understand what is its interpretation. Secondly, while summarization of a single video has received increasing attention [1,2,3,4,5,6], little work has been devoted to the problem of multi-episode video summarization [7,8] which raises other interesting difficulties.

Existing video summarization approaches can be classified in two categories. The rule based approaches combine evidences from several types of processing (audio, video, text) to detect certain configuration of events to include in the summary. Examples of this approach are the "video skims" of the Informedia Project [3], and the movie trailers of the MoCA project [5]. The mathematically oriented approaches, on the other hand, use similarities within the video to compute a relevance value of video segments or frames. Possible relevance criteria include segments duration, inter-segment similarities, and combination of temporal and positional measures. Examples

of this approach include the use of Singular Value Decomposition [9], and shot-importance measure [6]. 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. Since there is no ideal solution 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. Another method 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 directly as a measure of quality. 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.

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 the simulated user principle approach. In section 3, we describe the different algorithms used to construct multi-episode summaries. Experimental results and a study of summary robustness are presented in section 4 and 5. Conclusions and future extensions of the work are presented in section 6.

2 SIMULATED USER PRINCIPLE

In the Simulated User Principle, we define a real experimentation, a task that some user has to accomplish, and on which a performance measure is defined. Then, we use reasonable assumptions to predict the Simulated User behavior on this task. The performance of the Simulated User on the experiment is defined mathematically.

Applying the Simulated User Principle to the problem of multi-episode video summarization leads to the following scenario for the Simulated User Experiment:

- Show all the summaries to the user,
- Show a randomly chosen excerpt of a randomly chosen video,
- Ask the user 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,
- 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 similar image to any image in any summary, the user has no indication and cannot provide a definite answer.

The performance of the user in this experiment is the percentage of correct answers that he is able to provide when he is shown all possible excerpts of all videos. 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.

3 ALGORITHMS COMPARISON

In this paper, we present several algorithms that we use to automatically construct multi-episode video summaries. The simulated user principle is then used to evaluate the "quality" of the summaries. Finally we compare and discuss evaluation results to define the most appropriate algorithm for this application.

First we describe briefly the principle of three algorithms which are based on some ideas experimented before (more details are available in [10]), secondly we explicit the remaining algorithms that are original to the presented work.

Each multi-episode summary building process is divided into five phases: video streams pre-processing, feature vectors construction, classification, selection and summary presentation. The first three and the last one are the same for the six algorithms, nevertheless the fourth phase which performs the selection of the elements to include in the summaries is specific to each method.

Video Streams Pre-processing: Opening and ending scenes, common to all episodes are removed from further processing since of not interest to a viewer attempting to understand the content of a particular episode.

Feature Vectors Construction: The next phase consists of analyzing the content of the video to create characteristic vectors to represent visual information included in the video frames. Frames are divided into nine equal regions on which the color histograms are computed to capture both locality information and color distribution. The 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.

Classification: Frames are clustered 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.

Video Segment Selection: For each episode we select the most pertinent classes based on six alternative methods. More details are 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 or as an audio-video sequence of reduced duration. 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 images describing each episode (columns) is however entirely user definable.

3.1 Several Methods for selection:

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 a number of methodologies devised to compute this pertinence value.

Method 1: Based on the evaluation criterion we use a measure of coverage, so we attribute a coverage value to each class, this coverage value represent the number of excerpts of predefined length that contain this class. In this method the coverage is computed by using only the current video for which we select a class to add. A class

must be selected only once so it cannot represent two videos in the same global summary. To respect this constraint we use a conditional coverage. All excerpts containing classes that have already been selected will be neglected.

Method 2: This method is almost identical to the first one. The only difference is that coverage of candidate classes on other videos is taken into account during selection. To restrict ambiguous or erroneous cases, we use a negative coefficient to impose some penalty on classes with a large coverage on other videos.

Method 3: 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). When we select classes to be included to summary for a video we ignore classes present in the other summaries. Therefore, a class can be present twice or more in the global summary constituted of the concatenation of the different single summaries.

Method 4: 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.

Method 5: Based on the work of Uchihashi and Foote [6], who defined a measure to compute the importance of shots, we adapted our multi-episode summarization 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 of a class Wi , 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 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.

Method 6: The major idea of this method is to do a parallel with text summarization methodologies [11], 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. Therefore 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.

4 COVERAGE EXPERIMENTS

In this section we present the evaluation results using the simulated user principal on multi-episodes video summaries created with six 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).

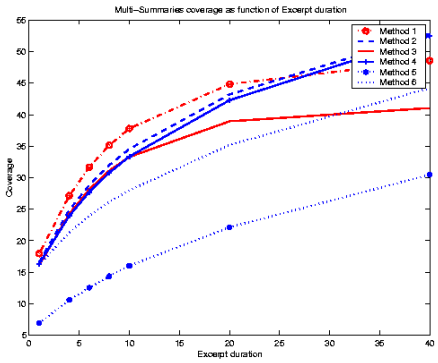

Figure 1

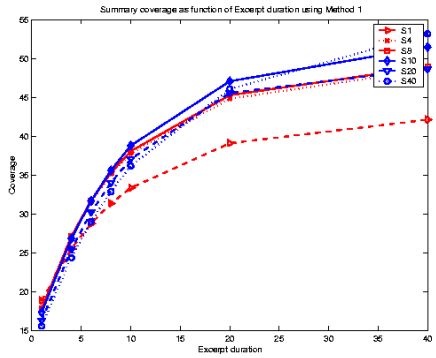

Figure 2

The graph in figure 1 shows the respective performance of these six methods when the duration of the excerpt used for evaluation varies. We note that the two first methods that build summaries based on a mathematical criterion inspired for the evaluation criterion itself give the best performance. We note also that the multi-episode summaries (methods 1 and 2) are more efficient than the single video summaries (method 3). As expected the 5th method performs very poorly. This is due to the fact that shots are selected on their length and low number of occurrence. Obviously, rare shot are likely to have little coverage over a video. Method 6, inspired from TF-IDF provides rather average results when compared with others. It should also be noted that results obtained with method 4 can be compared to those of method 2, and that both give the best coverage for large excerpts duration.

5 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. The first four methods are dependent on this excerpt duration whereas the last two are not. To study robustness, summaries were built for various excerpts duration and then evaluated using various excerpts duration. Figure 2 presents the results of this experiment for summaries based on method 1. Note that the construction method itself suggests that the coverage should be 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.

6 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 [6] 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.

7 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] 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.

[3] 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.

[4] 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.

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

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

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

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

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

[10] I Yahiaoui, B. Merialdo and B. Huet. Generating Summaries of Multi-episodes Video. To appear in Proc of IEEE ICME 2001.

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