AUTOMATIC VIDEO SUMMARIZATION

Itheri Yahiaoui, Bernard Merialdo, Benoit Huet
Multimedia Communications Department, Institut EURECOM
BP 193, 06904 Sophia-Antipolis, FRANCE
{Itheri.Yahiaoui,Bernard.Merialdo,Benoit.Huet}@eurecom.fr

ABSTRACT

In this paper, we present a new approach for the automatic construction of video summaries. We introduce the Simulated User Principle to evaluate the quality of a video summary in a way which is automatic, yet related to user perception. We present experimental results to support our ideas.

1. INTRODUCTION

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 such tools, which automatically creates a short version or subset of keyframes which contains as much information as possible as the original video. Summaries are important because they can provide rapidly users with some information about the content of a large video or set of videos. From a summary, the user should be able to evaluate if a video is interesting or not, for example if a documentary contains a certain topic, or a film takes partly place in certain location. 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. However, most approaches currently have the limitation that evaluation is difficult, so that it is hard to judge the quality of a summary, or, when a performance measure is available, it is hard to understand what the interpretation of this measure is.

In this paper, we propose a new approach for the automatic creation of summaries based on the Simulated User Principle, to address this problem, and we show some experiments that demonstrate how it can be used for video summarization.

1.1 Simulated User Principle

Existing approaches to video summarization can be classified in two categories:

- Some methods use pattern recognition algorithms to identify elements in the video, and rules to qualitatively select important elements to be included in the summary.
- Other methods use mathematical criteria, such as frequency of occurrence, to quantitatively evaluate the importance of video segments.

When one wants to evaluate the quality of a summary, several approaches are again available:

- One can ask a group of users to provide an evaluation of the summaries, either directly, or by comparison between several summarization methods.
- A quite realistic evaluation is to have a group of users to accomplish certain tasks (for example answering questions) with or without the knowledge of the summary, and measure the effect of the summary on their performance.

In the case of a mathematical criterion, the corresponding value can be used as a measure of quality for the summary.

Approaches which involve real users are the ones which provide the most realistic results, because one can have a clear view of the reasons for the quality of the summary. Unfortunately, these approaches are also very difficult and expensive to set up. To obtain results which have some statistical significance, many users are needed, and, when different summaries have to be compared, a user who has seen summary A can no longer be used to evaluate summary B because he cannot forget his previous knowledge. This greatly limit the number of experimentations which can be done with a given user group, so that it is difficult to compare many variations of the same method.

Therefore, a mathematical criterion is often used to evaluate the quality of the summary. In this case, evaluation is cheap, unambiguous, with no or little statistical variation, and can be used as often as needed. The difficulty is that it is not always easy to understand the meaning of this performance measure, because it is not related to any user activity. Therefore the real importance of an increase in the performance measure remains mysterious.

Our proposal is to use the Simulated User Principle to avoid the difficulties just mentioned. 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 behavior could be on this task. In other words, we use a Simulated User to accomplish this task, of which we can exactly know how he behaves. This allows us to mathematically define the performance of this Simulated User on the experiment. This leads to a mathematical criterion for which we try to build the best summary.

With our approach, we try to combine the bests of the approaches already mentioned:

- The use of a mathematical criterion allows easy rigorous multiple evaluations,
- This criterion is the performance of a user on a certain experiment, so that its interpretation is clear.

Of course, not all types of experiments are suitable for the application of this principle. In this paper, we provide one example that we feel is realistic enough to demonstrate the validity of this principle.

1.2 Automatic Video Summaries

Two types of video summaries are generally considered:

- "Video skims" are video sequences composed with portions of the original video, to form a shorter version,
- "keyframes sets" are selected images from the video, which can be displayed, for example, in a hypermedia document to access internal parts of the video.

Video sequences are often decomposed into consecutive shots, and shots are often represented by one or several keyframes, so that the difference between these two types of summaries is not very important, at least when we consider automatic procedures to select the best keyframes or shots.

Our approach is based on similarities of images in the videos, so that the same procedure can be used to select keyframes or video segments. Our algorithms are therefore able to construct summaries under the form of either a set of keyframes, or video skims. For the illustration of our methods in this paper, we will use the keyframe form of the summaries.

We also assume that the expected duration of the summary is a parameter which is given to the summarization process. This corresponds to the practical case where a user wants to spend so much time (say 20 seconds) watching the summary, but not more, or when space to display keyframes is limited. In our experiments, we use a summary length of six keyframes, because we found that this was a good compromise for a display on a computer or TV screen (six images of reasonable size can be displayed on a screen width). If a video skim version should be built, we would consider video segments of five seconds each (short video segments are hard to watch and remember, long video segments with uniform content are a waste of time), so that our choice of six keyframes roughly corresponds to thirty second summaries.

2. RELATED WORK

Summarizing video content is important to several applications including archiving and providing access to video teleconferences, video mail, video news, etc... [1] [5] [8] [9] [10] [11], [2] [7].

As mentioned earlier, approaches fall in two broad categories:

- Rule-based approaches. 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 [8], and the movie trailers of the MoCA project by Lienhart et al [10]. 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.
- Mathematically oriented. Such approaches 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 [15], or the shot-importance measure by Uchihashi and Foote [11].

3. VIDEO SUMMARIZATION

3.1 Simulated User Experiment

To apply the Simulated User Principle to the problem of the summarization of a single video, we propose the following simulated experiment:

- The user is shown the summary of a video,
- He is shown a randomly chosen excerpt of this video,
- He is then asked whether this excerpt originates from the same video as the summary or not.

The simulated behavior of the user is the following:

- If at least one image in the excerpt is "similar" to an image in the summary, then he can answer positively (and this answer is correct),
- If this is not the case, he is in doubt and cannot provide any answer.

The probability of a correct answer is the expected performance of a user on this experiment. It is based on the assumption that the user has a perfect visual memory of images in the summary, and that he does not know in advance if the excerpt comes from the same video or not (although he is actually only shown excerpts from the same video).

To compute the expected performance in an automatic way, the only difficult point is to have a definition of image similarity which is consistent with the user's definition. We will describe our approach for this definition in the experiment section.

Note that one might think that a complete experimentation should display to the user excerpts coming from the current video as well as excerpts from other videos. However, this is difficult to implement in practice, because it would make the experiment dependent of the choice of these other videos. If we consider image similarity with a rather strict interpretation, then it is very unprobable that an image from a video will be similar to an image from another video (except maybe if the same clip is used in several videos), so that adding extra videos would not change the number of positive answers in the simulated experiment.

We consider that the excerpt is chosen randomly in the video with a uniform probability distribution. This gives the same importance to the various parts of the video. If one feels that this should not be the case, for example that the sequences in the beginning of the video are more important, or that the end of the video should not be disclosed, it is possible to use a non-uniform probability distribution to achieve these requirements.

We also consider that all excerpts have the same duration (which is a parameter is the summarization process). It is not necessarily the case, and we could design an experiment where both the location and the duration of the excerpt are chosen at random. However, we have no reasonable interpretation for a variation in

duration, neither a reasonable duration probability distribution to suggest.

3.2 Automatic Summarization

Now that the Simulated User Experiment is defined, we need a process to automatically construct a summary with good (and if possible optimal) performance for this experiment. In the case of single video summarization, this turns out to be relatively simple.

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

We assume that frames have been clustered into "similarity classes", so that two frames are considered to be similar if and only if they belong to the same class:

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

This is a very strong assumption, and we will explain in the next section how these similarity classes are built. Figure 1 illustrates the relations between excerpts, frames and classes.

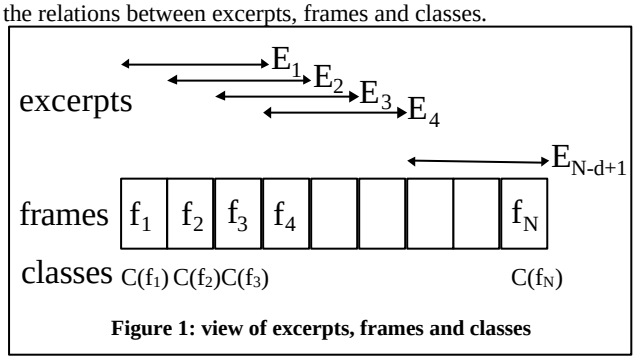


Figure 1: view of excerpts, frames and classes

Let us define the coverage Cov(C) of a class C as the number of excerpts which contain at least one frame from class C:

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

The coverage of a set of classes $C_1, C_2, \dots C_k$ is the number of excerpts which contain at least one frame from one of the classes:

$$\text{Cov}(C_1, C_2, \dots C_k) = \text{Card}\{i : \exists j, l \quad f_j \in E_i \quad \text{and} \quad C(f_j) = C_l\}$$

If a video summary is composed of frames $f_1, f_2, \dots f_k$, it induces a performance in the simulated user experiment which equals:

$$\text{Cov}(C(f_1), C(f_2) \dots C(f_k))/(N-d+1)$$

Therefore, the optimal summary is simply one which maximizes:

$$S = \arg\max_{f_1, f_2, \dots f_k} \text{Cov}(C(f_1), C(f_2) \dots C(f_k))/(N-d+1)$$

This can be achieved in two steps:

- First find a set of classes with maximal coverage,
- Second select a representative frame in each class.

3.3 Summary Construction

The optimal summary can be found by enumerating all the sets of k classes $\{C_1, C_2, \dots C_k\}$ and keeping the best one. Because the enumeration can be computer intensive, it is profitable to select carefully the order in which classes are selected, so that the best solutions are found early.

If a class C_m is added to an existing set $\{C_1, C_2, \dots C_{m-1}\}$, we can define the "conditional coverage" as its contribution to the coverage of the final set:

$$\text{Cov}(C_m \mid C_1 C_2 \dots C_{m-1}) = \text{Cov}(C_1 C_2 \dots C_m) - \text{Cov}(C_1 C_2 \dots C_{m-1})$$
$$= \text{Card}\left\{ i : \begin{array}{l} \exists j \quad f_j \in E_i \text{ and } C(f_j) = C_m \\ \text{and} \quad \forall f \in E_i \quad \forall j = 1,2,\dots m-1 \quad C(f) \neq C_j \end{array} \right\}$$

Then, the coverage of a set of classes $\{C_1, C_2, \dots C_k\}$ can be computed as:

$$\text{Cov}(C_1 \dots C_k) = \text{Cov}(C_1) + \text{Cov}(C_2 \mid C_1) + \dots + \text{Cov}(C_k \mid C_1 \dots C_{k-1})$$

The algorithm to construct the optimal summary proceeds as follows:

- Step 1: start with an empty set of classes,
- Step 2: order the classes that have not yet been selected by decreasing conditional coverage with respect to the current set,
- Step 3: try to add each class in turn to the current set. If the desired size for the summary is reached, replace the current solution by the current set if its coverage is larger. Otherwise, recursively go to step 2 to continue enumeration.
- Step 4: when all classes have been tried, backtrack to the previous level to continue enumeration.

During this backtracking procedure, it is possible to avoid some enumeration by noting that the following relation always hold if m<k:

$$\text{Cov}(C \mid C_1 C_2 \dots C_{m-1} \, C_m \dots C_k) \leq \text{Cov}(C \mid C_1 C_2 \dots C_k)$$

so that:

$$\begin{aligned} &\text{Cov}(C_1 C_2 \dots C_{m-1} \, C_m \dots C_k) = \\ &\quad \text{Cov}(C_k \mid C_1 C_2 \dots C_{k-1}) + \text{Cov}(C_{k-1} \mid C_1 C_2 \dots C_{k-2}) + \dots \\ &\qquad + \text{Cov}(C_m \mid C_1 C_2 \dots C_{m-1}) + \text{Cov}(C_1 C_2 \dots C_{m-1}) \\ &\leq \text{Cov}(C_k \mid C_1 C_2 \dots C_{m-1}) + \text{Cov}(C_{k-1} \mid C_1 C_2 \dots C_{m-1}) + \dots \\ &\qquad + \text{Cov}(C_m \mid C_1 C_2 \dots C_{m-1}) + \text{Cov}(C_1 C_2 \dots C_{m-1}) \end{aligned}$$

This inequality provides an upper bound for the best solution that can be constructed by extending the set $\{C_1, C_2, \dots C_{m-1}\}$. If this upper bound is less that the coverage of the current best solution, then enumeration can be cut off at this level. This reduces the amount of computation required, while preserving the optimality of the algorithm.

Note that the algorithm starts by selecting the class C_1 with maximal coverage, then C_2 with maximal conditional coverage with respect to C_1, and so on until C_k. The first complete solution found is then the result of a series of greedy choices, and it appears that, experimentally, it is often the optimal choice over all possible combinations.

Once that the best set of classes has been found, it only remains to select a representative frame for each class. Since all frames in a class are supposedly similar, this choice should have no influence on the quality of the summary. The representative frame is selected as the one whose feature vector is the closest to the centroid of the class.

4. EXPERIMENTS

4.1 Video Processing

In our experiments, we are using Mpeg1 encoded videos. Because we are only interested in a global analysis of the videos and not in short duration details, videos are sub-sampled, so that one frame per second is retained. For each frame, a feature vector is built as the color histogram for nine rectangular regions of equal size in the frame. To further simplify processing, consecutive frames whose feature vectors are very close (with a strict threshold) are confused and only the first one is kept, together with a duration information to preserve temporal information.

4.2 Frame similarity

Frame similarity is based on the distance between feature vectors. Frames are considered similar if the distance between their features vectors is less than a threshold. We use the L_1 distance with a threshold which provides the best agreement with a manual judgement of similarity.

4.3 Similarity classes

Now that the distance and threshold have been defined, frames whose distance is less than the threshold are considered as being similar. We further construct similarity classes by clustering features vectors using a two step procedure:

- First each vector is either added to an existing class if its distance to the centroid is less than the similarity threshold, or used to create a new class if this is not the case,
- Then these classes are iteratively refined using a few iterations of a k-means like reassignment step.

4.4 Summary performance

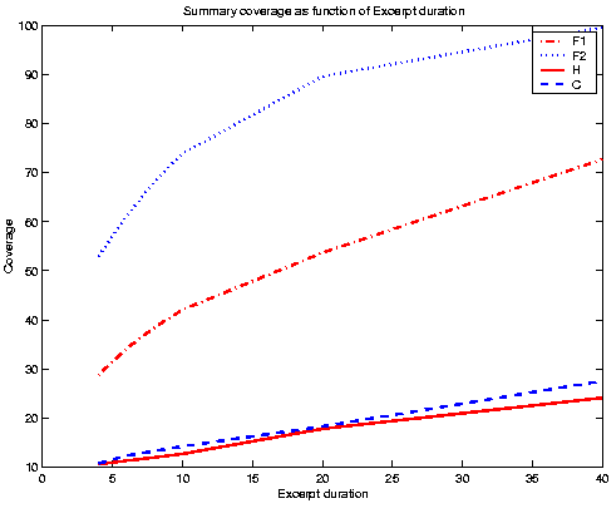
We applied our summarization algorithm to several videos:

- two episodes from the TV serie "*Friends*" (F1 and F2),
- a documentary "Histoire d'eau" (video H),
- a fiction "Chapeau melon et bottes de cuir" (video C).

The episodes of "Friends" were recorded in our laboratory from a regular TV channel. The other two videos are part of a video corpus distributed by the INA (French National Institute for Audio-Visual). The following table presents the duration (in seconds) of the different videos and the number of similarity classes.

	F1	F2	H	C
Duration	1310sec	1298sec	2118sec	3035sec
Classes	248	165	933	1065

Each video is summarized independently of the others. The following graph shows the coverage of the summaries (expressed in percentage) for various durations of the excerpts. All summaries are constructed with a given size of six frames.



As expected, the coverage of summaries increases as the duration of excerpts increases. We observe that different videos have different behavior. Summaries of episodes of the TV series appear to provide high coverage, starting at about 30%, while documentary and fiction provide a coverage starting at around 10% only. This is largely due to the fact that the documentary and fiction exhibit a greater diversity (similarity classes) than the TV series episodes. Of course, this is also partly due to the fact that the episodes are much shorter than the other two videos.

We ran the summarization algorithm with various duration for the excerpts. In all of our experiments but one, the first solution found by the greedy construction without backtracking was also the optimal one. Therefore, it appears that in practice, an enumeration of all solutions is not necessary to provide quasi-optimal summaries.

5. CONCLUSION

Automatic video summarization is a very important tool. In this paper, we have proposed a novel approach to automate the creation of video summaries. The Simulated User Principle aims at providing a practical way to create and evaluate video summaries, yet it uses a performance measure which is based on a simulated experiment whose results are easily interpretable. We have shown how this principle can be applied to the creation of single video summaries, but the same idea can be extended to the construction of multi-episode video summaries. Experiments demonstrate the feasibility of our approach. The algorithms proposed in this paper can be used to automatically generate summaries for video recordings, as the ones that can be available in set-top boxes, or in multimedia databases.

6. REFERENCES

[1] 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.

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

[3] Bernard Merialdo, Kyung Tak Lee, Dario Luparello, and Jeremie Roudaire. Automatic construction of personalized TV news programs. In *ACM Multimedia conference*, November 1999.

[4] 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, December 1999.

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

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

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

[8] 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, pages 61-70, 1998.

[9] 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.

[10] Rainer Lienhart, Silvia Pfeiffer and Wolfgang Effelsberg. Video abstracting. In Communications of ACM, December 1997.

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

[12] Udo Hahn and Indejeet Mani. The challenges of automatic Summarization. IEEE Computer, November 2000.

[13] V.Di Lecce, G.Dimauro, A. Guerriero, S.Impedovo, G.Pirlo, A.Salzo. Image basic features indexing techniques for video skimming. IEEE Intl. Conf. on Image Processing, 1999.

[14] Yueting Zhuang, Yong Rui, Thomas S. Huang and Sharad Mehrotra. Adaptive key frame extraction using unsupervised clustering. IEEE Intl. Conf. on Image Processing, 1998

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