Automatic Video Summarization

Benoit Huet and Bernard Merialdo

Institut EURECOM, 2229 route des crêtes, 06904 Sophia Antipolis, France
benoit.huet@eurecom.fr and bernard.merialdo@eurecom.fr

1 Introduction

Due to the ever increasing number of multimedia documents one is potentially confronted with everyday, tools are eagerly awaited to ease the navigation through massive quantities of digital media files. Summaries provide an interesting solution to this problem. Indeed, by looking at a summarized version of a document one is able to quickly identify interesting or relevant documents.

In this chapter, we present a brief review of recent approaches in video summarization, and then we propose our approach based on the Maximum Recollection Principle. We show that this approach is supported by reasonable assumptions, and that this principle can be applied in diverse situations. In particular, we describe how it can be applied to the summarization of a single video sequence, a set of video sequences, and a combined audio-video sequence. For all these cases, we present some experimentation and discuss implementation issues for the corresponding algorithms.

2 State of the art in video summarization

The automatic creation of multimedia summaries is a rather powerful tool which allows to synthesize the entire content of a document while preserving the most important or most representative parts. Here, we concentrate on video summarization. In this respect, the creation of a video summary will result in a new document which may consist on an arrangement of video sequences or an arrangement of images. In other word, a video summary may take the form of a dynamic or a static document. The original document represented in such an abstract manner may find various perusals. It may help, for example, to get a quick feel about the content of a document or even about the general content of an entire database of multimedia documents. Another example of possible usage, particularly well suited for multi-episode TV series, is the ability to identify documents which have already been watched. Along

the same line a video summary should enable the viewer to decide whether the content of the original is relevant or not. This leads us to the obvious fact that a document can be summarized in a number different ways. Each of these individual summaries may have equal quality with respect to their intentional usage despite being different. This clearly exposes the difficulties associated with the task of automatic video summarization. In the context of text summarization, Mani and Maybury [1] have identified three important factors for summary creation and evalutation; Conciseness, Context and Coverage. He et al. [2] address the same issue for video summarization and identify a fourth factor; Coherence. Video summarization has started to receive interest from the research community in the mid nineties [3][4][5][6][7]. Since then, the topic has received an ever increasing attention. The approaches found in the literature are extremely varied, and can be organized along a number of potential axes, such as the modalities employed to create the summary, the type of summary created (static vs dynamic), the method used for the creation (the selection process), whether the method offers generic properties or has suitable for a specific type of video. Here, we will divide the literature into 2 main categories according to the type of summary created by the method. This choice is motivated from the fact that some application may fit more closely one type of summary than the other. Having said that, it is possible to transpose a dynamic summary into a static one by performing key-frame selection. The opposite, converting a static summary into dynamic one, is also achievable by recovering shots from which key-frames were selected in order to create the video skim. We shall now report some of the ap-proaches from the literature for dynamic and static summary creation.

2.1 Dynamic Summaries

Dynamic summaries are often referred to as video-skims. Video-skims may be seen video preview where shots or scenes which have been classified as less important or less relevant are skipped. This type of summaries has the advantage over their static counterpart to combine images (video) and audio. This allows the summary to convey more information about the original content of the multimedia document. In its most simplistic form a video-skim is created by extracting pieces of video of fixed duration at intervals uniformly distributed over the video [9]. Nam and Tawfik [8] have proposed an approach which extends the basic scheme by sub-sampling the video non-linearly. The rate depends directly on the amount of visual activity measured within the shots. Others including [10] and [11], have proposed to basically fast-forward through the video in a uniform or adaptive manner. The major drawback of such approaches is the distortion caused to the original material. Overall, the common task of more advanced algorithms is the selection of the excerpts to retain for the summary. Obviously, this will essentially depend on the objective and the application domain of the summary. Some view the summarization process as one where the objective is to remove redundant scenes or shots

from the original document. In [12], a self similarity matrix of video features is employed to select and adjust video excerpts length. This selection process may also be addressed like a clustering problem. In [12], visual features grouped according to their similarity and excerpts which lie closest to cluster centers are used to construct the summary. Similar approaches [13][14][15] have extended this idea with the use of additional modalities such as audio, text, motion, etc In the case of domain specific methods, it might be possible to detect particular events, such as goals in a soccer game or action scenes in a movie. In [16], Lienhart et al. studied the properties generally found in movie trailer, which resulted in a number of event detectors based on video, audio and text features. The location of events detected in the movie indicates which shots/scenes should be present in the trailer. Another event based technique was proposed by Chang et al. [17]. In this work, baseball game highlights are detected using HMM models trained on 7 different game actions. Another class of methods achieves video summarization by looking at the evolution through time of a single or a set of features. In effect, a score (feature value) is computed and associated with each temporal video element; A shot, a frame, a caption word, a sentence, etc depending on the modality and the method. The selection scheme for video-skims candidates based on temporal element feature value may be threshold based, maxima based [18] or obtained in a greedy manner [19][20]. In such approaches, the challenge is to identify the right set of features.

2.2 Static Summaries

As opposed to dynamic summaries, it is possible to present static summaries differently. The static summary may be viewed like a story-board (or a film strip), a mosaic of key-frames, a slideshow or a flowchart. Its major advantage over video-skims is the possibility to present the content with an emphasis on its importance or relevance rather then in a sequential manner. The most basic way to create a static summary is to perform some sub-sampling on the video, at a rate based upon the number of key-frame desired [4][5]. The major drawback of summarization through direct sub sampling is that there are little guarantees that the selected key-frames have some sort of relevance. A step toward improving the selection process is to detect content change in the video and retain key-frames from the segmented shots [3][21][22]. The difference between such approaches resides in the method employed to select the representative frame or frames for each shot. This may be realized in a systematic manner (i.e. by taking the first frame of each shot [3], or competitive process over shot frames [23]. An obvious way to select candidate frames to summarize a video is to cluster features extracted from the video and identify representative key-frames from each cluster. Other approaches, view the video as a curve in a multi-dimensional space of feature where each video frame is represented as a point. In this framework proposed by DeMenthon et al. [24] the process of summarization corresponds to the selection a

set of points (frames) on the curve for which retain as much as possible the general shape of the curve. In effect, polygonal approximation techniques can be employed to provide solutions. In [24], a recursive binary curve splitting algorithm is used to this end but alternative algorithms such as discrete contour evolution [25] may also be employed. Event detection is yet another way to capture and identify important video frames. The most common attribute employed for event detection is motion. In [26], Lui et al. extract representative frames based on the analysis of the motion patterns within shots. Others [27][28], base the selection process on characteristic motions of extracted region from frames. Content characteristic may also be of importance in order to determine the importance of a shot. For example, knowing that a frame contains people [38] or specific objects with a given behaviour [29] can take a part in the summary creation process. Approaches relying on event detection are generally too specific to deal with arbitrary videos and are therefore application domain limited. In an attempt to obtain summaries with as much fidelity as possible to the original multimedia document, Chang et al. [30] introduced the idea of using frame with maximum frame coverage as summary representative. This idea has then been extended by Yahiaoui et al. [31] to the selection of the set of frames which are the most frequently found (or sufficiently similar to at least one frame) in excerpts of a given duration. A variant of this approach [32] has been developed for multi-episode video summarization in an attempt to exhibit the major differences between episodes of TV series. This approach insures that the resulting summaries will have little redundancy while covering as many different aspects of the video as the number of key-frames selected.

2.3 Summary Evaluation

Evaluation of video summaries is an issue often overlooked by researcher. This is probably due to the fact that is there is no standard measure to access the quality of a summary. Moreover, the quality of a summary depends greatly on it intended purpose as well as the application domain, thus it is not possible to define a general performance measure. Furthermore, the process of summary evaluation is a highly subjective one. The most common evaluation found in the literature [33][34] consists in presenting results of the approach for a number of multimedia documents and providing some motivation for the selected sequences or key-frames. Some researchers go through the time consuming process of involving users in the evaluation process. This more realistic evaluation procedure may be performed in three different manners. In the first scenario [26][38], users or experts are asked to summarize some document in order to obtain a ground truth which can then by compared with the automatically created one. In the second scenario [35], they are asked to judge or access of the quality of computer generated summaries with respect to the original videos. In the last scenario [36][37], the summaries are presented to the evaluation users along with a set of tasks or questions. The quality of the

answers is then analyzed to grade the summary and therefore the underlying construction methodology. It is nonetheless possible to define a metric in order to access the quality of the summaries. This metric is in most case directly derived from the fidelity factor used to perform the selection process and is therefore often biases toward the newly proposed approach [30][33][2]. In an effort to provide common metric for video summarization algorithms, DeMenthon et al. [39] have proposed an automatic performance evaluation based on performance evaluation metrics used in the field database retrieval. The review presented as introduction to this chapter about video summarization is by no mean exhaustive. For a more comprehensive review of the field we invite the interested readers to have a look at the following papers [40][41][42].

3 Maximum Recollection Principle

3.1 Definition

The idea for the Maximum Recollection Principle (MRP) was suggested by the situation where some people randomly zap to a TV channel, watch a few seconds and are able to recognize a movie that they have already seen. The formalization of this idea leads to the following statement:

> The summary of a document should contain such information to maximize the probability that a user would recognize the document when exposed to an extract of the document.

This statement provides the basis for a sound framework to define optimal summaries, while leaving much flexibility in the application to various types of documents. Several arguments support the use of the MRP:

- first, it is a reasonable objective for a summary, as the "zapping example" is a very common situation,
- second, it provides a measurable criterion (probability of recognition), so that an optimal summary can be defined,
- third, it leaves open the precise definition of an extract, which information from the document is being displayed, (we have found that a random choice of the extract is a good start, but the duration of the extract is still a parameter of the summarization),
- finally, the concept of "recognition" can be implemented in a number of different ways, leading to variations which can be adapted to numerous situations (for example, different similarity measures or different document types).

The application of the Maximum Recollection Principle to video sequences gives a simple illustration of the principle. A video sequence is a sequence of images $V = I_1, I_2, \ldots I_T$. A summary is a selection of key-frames: $S = I_{s(1)}, I_{s(2)}, I_{s(k)}$ (we assume that the size k of the summary is fixed, either

by the system or by the user). We suppose that a virtual user has seen the summary and is presented a random excerpt $E(r,d) = I_r, I_{r+1}, I_{r+d}$, of the video. The user will recognize the video V if at least one of the images of the excerpt $E(r,d)$ is similar to an image in the summary. Images I and I' are similar is the value of a similarity function $sim(I,I')$ is less than a predefined threshold θ. The quality of the summary can then be measured as the percentage of excerpts for which the recognition occurs. Formally,

$$\mathtt{perf(S)} = \#\{\mathtt{E(r,d)} : \exists \mathtt{i,j}\ \mathtt{sim}(\mathtt{I_{s(i)}}, \mathtt{I_{r+j}}) < \theta\} \tag{1}$$

(this number can be also normalized by the total number of excerpts).

3.2 Illustration

As an example, suppose that images have been clustered into similarity classes, so that two images are considered similar if and only if they belong to the same class. The summary is composed of a number of similarity classes (there is no need for two images of the same class in the summary). If we consider excerpts of length one, then the probability that the excerpt will be similar to an image of the summary is simply the sum of the frequencies of the class in the summary. Therefore, the optimal summary (in this simplistic case) is composed of the most frequent similarity classes. If we consider excerpts with length greater than two, then the situation is more complex, as two frequent similarity classes may often correspond to the same excerpts, and therefore be redundant in the summary. For example, assumes that the similarity classes are named A, B, C, ..., and the video is composed of images forming the sequence:

A B C . A B C . A B . A .

(where the period indicates another similarity class than A, B, C). With excerpts of length 1, we can draw the following performance table for summaries $\mathtt{S}_1 = \{\mathtt{A},\mathtt{B}\}$ and $\mathtt{S}_2 = \{\mathtt{A},\mathtt{C}\}$.

Video	A	B	C	.	A	B	C	.	A	B	.	A	.	Perf
$\mathtt{S}_1 = \{\mathtt{A},\mathtt{B}\}$	+	+	-	-	+	+	-	-	+	+	-	+	-	7
$\mathtt{S}_2 = \{\mathtt{A},\mathtt{C}\}$	+	-	+	-	+	-	+	-	+	-	-	+	-	6

The best summary is evidently the one composed with the most frequent similarity classes, in this example, A and B.

If excerpts have a length of two, then excerpts overlap, so that the performance table becomes: (where a + indicates a match with the excerpt starting at this position).

In this case, the best summary is $\mathtt{S}_2$, despite the fact that class C is less frequent than class B. The reason is that the performance criterion will only

Video	A	B	C	.	A	B	C	.	A	B	.	A	.	Perf
$S_1 = \{A, B\}$	+	+	-	+	+	+	-	+	+	+	+	+	-	10
$S_2 = \{A, C\}$	+	+	+	+	+	+	+	+	+	-	+	+	-	11

count one when several classes contribute to the same excerpt. Therefore, the optimal summary will be based on a selection of classes which exhibit an optimal mix of high frequency but also high spread throughout the video.

3.3 Experiments

In order to validate our approach, we have performed a number of experiments. Several issues are considered:

- experiment with reasonable similarity measures,
- design efficient summary construction algorithms,
- evaluate the performance of the summaries with respect to the excerpt length.

Visual similarity is a very difficult and complex topic. In order to define a simple and reasonable similarity measure, we have manually labeled pairs of images as visually similar or not, then we have compared the re-sults of several simple similarity measures. We have found that the measure which provides the most coherent results with our manual labeling was based on blob histograms. We have used this measure throughout our experiments. We have focused our work on the efficient construction of summaries. The performance criterion that we have defined requires a combinatorial enumeration to select the optimal summary. The reason is that, if we try to build the optimal summary by successively adding keyframes, the selection of a keyframe may completely rearrange the importance of the remaining keyframes. To tackle this problem, we have explored suboptimal ways of gradually constructing the summary. With our approach we build the summary in a greedy fashion by selecting the best keyframe at each step, then, when the desired size is reached, we try to replace every frame with a better frame if it exists. We found that this process was near optimal from a performance perspective, while being much faster than complete enumeration from a computational perspective.

We have also experimented with the importance of the excerpt duration on the quality of summaries. In our experiments, we have used different videos: F1 and F2 are 16 minutes episodes of a TV series, H is a 50 min documentary, and C is a 45 minutes fiction. We also used the first 16 minutes H1 and C1 of H and C. We have computed the performance of summaries composed of 6 keyframes, for excerpt durations varying from 4 to 40 seconds. The results are indicated in figure N. They show that the performance increase with excerpt duration, which is obvious as there are more chances to recognize a similarity with a summary keyframe when the excerpt is longer. TV series provide a better performance, which can probably explained by the fact that the setting of the

scene is generally fixed or limited, which increases the chances of similarity. Of course, the duration of the video is also of importance, and the shorter version H1 and C1 have better performance than H and C, although still lower than the performance for TV series of the same duration.

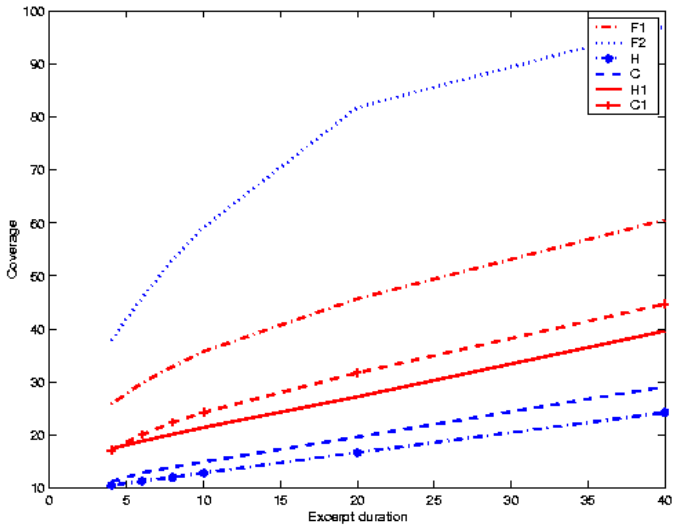

Fig. 1. Summary coverage with respect to excerpt duration

4 Multi-video summarization

4.1 Definition

In this section, we describe the application of the MRP to the case of the simultaneous summarization of a set of videos. Of course, an easy solution is to construct independent summaries for each video. But in the case where the videos are related, for example different episodes of the same TV series, this could lead to redundant information included in several summaries. In the case of related videos, we would like the summaries to contain only the information which is particular to each video, and not to contain information which is common in several or all episodes.

Assume that the user has seen the summaries of all the videos and is presented with an excerpt of an unknown video. Three cases may happen:

- if the excerpt has no similarity with any of the summaries, then the unknown video remains unknown,
- if the excerpt has some similarities with one of the summaries S, then the video is identified as the corresponding video,

- if the excerpt has some similarities with several of the summaries, then the case is ambiguous and the video cannot be uniquely identified.

According to the MRP, we should try to build a set of summaries which maximize the probability of correct answer occurring in case two, and minimize the other cases. The optimal construction of summaries becomes much more difficult in the case of multi-videos. If we try to build the summaries progressively, by adding keyframes one by one, it may happen that a keyframe which would be very good for the performance of a summary would also create many ambiguities with other previously selected keyframes. We have experimented with several strategies to select heuristically interesting keyframes to be added, the idea being to jointly maximize the number of similarities that can be found in the current video and minimize the number of confusions that it may create in other videos.

4.2 Experiments

We have evaluated our algorithms on a set of 6 episodes of a TV series. The construction of the summaries is iterative, one keyframe being added to each summary in turn. The following figure shows the number of correct identifications, incorrect identifications and ambiguities in the summaries built for various excerpt durations.

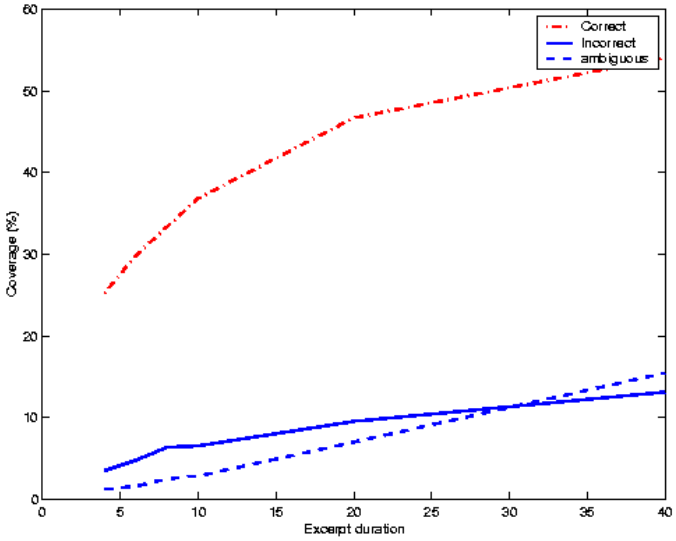

Fig. 2. Multi-Summaries evaluation with respect to excerpt duration

We also evaluated the robustness of the summaries that are built with respect to the excerpt duration. For this, we consider the summaries built by the

best method for a given excerpt duration, and we evaluate their performance for a different excerpt duration. The results are given in the figure below:

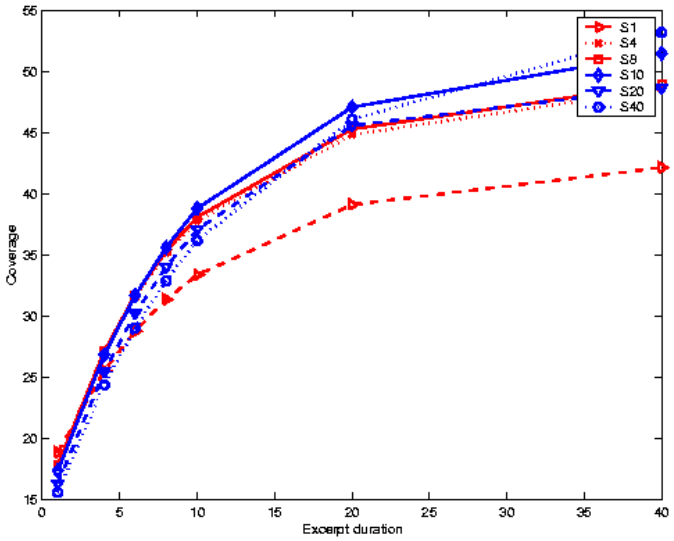

Fig. 3. Summary coverage with respect to excerpt duration

These results show that, except for summaries built for a duration of 1 second, the performance remains quite stable. For example, a summary built for a duration of 4 seconds has almost the same performance on a duration of 20 seconds as the optimal summary.

5 Joint Video and Text summarization

In this section, we assume that the documents are composed of both video and text, the text being the synchronized transcription of the audio channel. An excerpt of the document will contain both the video and the corresponding audio. A summary will be composed of a set of keyframes and keywords. The Maximum Recollection Principle can be applied easily again in this case. An excerpt of the document will be recognized if an image of the video is similar to a keyframe of the summary, or a word in the audio is similar to a keyword of the summary. The performance of the summary is the percentage of excerpts that are recognized. One may argue about the relevance of the similarity criterion for key-words. While for images, similarity is generally a convincing argument for the similarity of the videos, this is not really true for keywords, except maybe for very rare words. In fact, keywords can be more considered as hints for similarity than complete evidence. We have explored

this issue in other work (not presented here, because it deals with keywords only), for example by considering the number of documents on the Internet that are retrieved from a single keyword as an indicator of the pertinence of the keyword. Though, for joint video and text summarization, we have found that the usage of keyword similarity provides a sensible way of selecting keywords from the document. Although this issue deserves deeper investigations, we consider our current approach as a useful step in this direction. The text from the audio track is filtered (common words are removed, as in information retrieval systems, and words are stemmed). The similarity between keywords is simply the identity of the stems. The construction process of the summaries remains similar to the case of video only: the process starts with an empty summary and tries to add elements one by one. An element can be either a keyframe or a keyword. At each step, the element which best improves the performance of the summary is selected.

5.1 Experiments

We have tested our approach on several videos. We have obtained the transcription of the audio track through the caption channel. This provides us with keywords and timecode information that relates them to the corresponding keyframes. As an example, the following figure shows a summary with 10 elements for a documentary about the Amazonian forest.

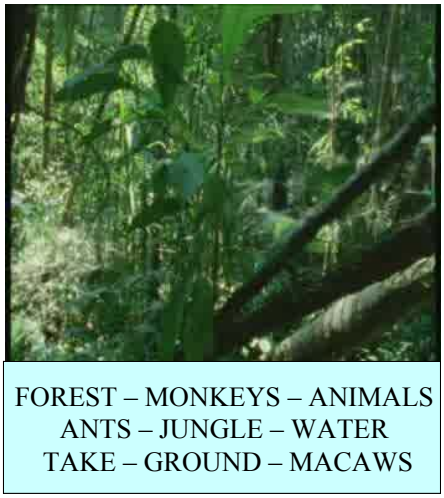

Fig. 4. An exemple of multi-modal summary

Here, the construction algorithm has chosen to include one image and 9 keywords in the summary. This selection is based on the efficiency of the element in terms of similarity with the possible excerpts. We can also impose to

use a predefined number of images and key-words in the summary. This may happen for example, when the summary should be displayed inside a predefined template. The following figure shows the performance of the summaries on several documents, as a function of the number of images and an excerpt duration of 20 seconds.

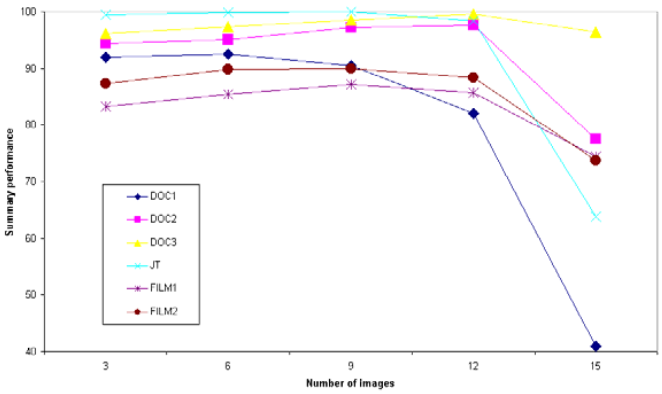

Fig. 5. Summary performance vs number of images

6 Constrained display summaries

During the course of a project, we got the request to produce summaries which could fit on a PDA display. Given some space limitation, we wanted to produce the best summary, in this case, this means displaying as much information as possible within this space. Therefore, we modified our construction algorithm by assigning an occupancy factor to images and key-words, for example images take 5 units of space, while keywords use only 1 unit. The algorithm is then able to find the best summary which respects the global occupancy constraint. At each step, the element which gets the best performance/occupancy ratio is selected. The process stops when there is no more available space to fill. In this case, the number of elements is not fixed, but chosen by the algorithm.

7 Home video network interaction

In the previous sections, a generic approach to multimedia documents summarisation has been proposed. We will now describe a system; developed within the European Project SPATION [44] (Services Platforms and Applications for Transparent Information management in an in-hOme Network) where

video summaries are used effectively. With the ever increasing storage, networkability, and processing power of consumer electronic devices (CE-devices), developing a common platform would result in a home network with tremendous possibilities. The challenging task of storing and retrieving information in the home network was the focus of the project. Today's CE-devices are able to store hundreds of hours of video, thousands of songs, and ten thousands of photographs, etc. Solutions are required to organize and retrieve the various documents (mp3, mpeg, jpeg, etc) in a user friendly manner. To address the issue of navigation and content selection, multimedia summaries are introduced to provide an effective solution. In the SPATION project two different types of summaries are automatically computed on the devices; a video trailers and visual overviews consisting of representative frames [43]. Those different types of generated summaries can be used in future CE-devices in a number of ways. The automatically generated summaries can be of great assistance to users whether they wish to locate a particular program or a scene within a program or they are trying to remember if they have already watched a program. It may also help a user to decide whether a programme should be deleted, archived on DVD or simple kept on the local hard drive without having to watch at it entirely. Summaries can also be extremely useful in the context of video on demand in order to decide if the entire movie should be downloaded, or simple to ease the process of choosing what to watch next out of the hundreds of hours of recorded material. In the case where summaries can be computed on the fly, when one turns a channel on and the broadcast of the program has already started if would be possible to display a preview of what has happened so far (using picture in picture view for example). Among the important functionalities related to the viewing of trailer like summaries, the ability of pause, stop, fast-forward or rewind and skip to the next one are very desirable. Indeed this would allow "zapping" through summaries as easily as zapping through broadcasts or DVD scenes. In the case of mosaic like (still picture) summaries, it is possible to refine the level of detail of summaries by increasing the number of representative images, or by selecting one of the summary thumbnails in order to access directly the corresponding scene in the original document. Another extremely useful feature relies on the fact that is it possible to download, via Bluetooth or WiFi, summaries on a mobile handheld device. The summaries may then be displayed on the PDA or Philip's iPronto while the user is on the move. For the SPATION demonstrator, which is depicted in the figure below, a Philips iPronto is used. This device also operates as an advanced remote control, providing access to the home CE-devices and their multimedia documents. Having the summaries available on the mobile device allows viewing shorter versions of the programs while the user if away from the home. Indeed, this is also possible in the case where the program (photo, song or even video) does not fit entirely on the device permanent storage (hard drive or flash memory based storage) to replace it with a summarised version adapted to the amount of storage available.

Fig. 6. The SPATION demonstrator, presenting the user with a static summary of a broadcasted program

The SPATION demonstrator shows a realistic application for video summaries in an interactive environment. It also, proves the feasibility of such a system and clearly gives some insight in some of the functionalities of future home CE-devices.

8 Conclusion

In this chapter, we have presented our approach to video summarization, based on the Maximal Recollection Principle. Our proposal provides a criterion for the automatic evaluation of summaries, and thus allows to define and construct optimal multimedia summaries. We have shown that it can be applied in a variety of situations, in particular it can be used to generate optimal summaries containing both video and text. Moreover, the demonstrator developed during the SPATION project exposes some of the many potential usages of multimedia summaries. Indeed, summaries will un-doubtedly play an important role within future interactive multimedia devices.

Acknowledgements

The work presented in this book chapter has been funded by the European Commission under the SPATION FP5 project. Additionally, the authors wish to thank Itheri Yahiaoui whose PhD work greatly contributed to this book chapter.

References

1. Inderjeet Mani and Mark T. Maybury. Advances in Automatic Text Summarization. The MIT Press, 1999.
2. He, L., Sanocki, E., Gupta, A., and Grudin, J. Auto-summarization of audio-video presentations. In ACM Multimedia (ACMMM'99). Orlando, Florida, 489-498. 1999.
3. Tonomura, Y., A. Akutsu, K. Otsuji, and T. Sadakate. Videomap and videospaceicon: Tools for anatomizing video content. ACM INTERCHI'93, pages 131-141, 1993
4. Arman, F., Depommier, R., Hsu, A., and Chiu, M. Y. Content-based video indexing and retrieval. In ACM Multimedia (ACMMM'94). San Francisco, California, 97-103. 1994.
5. Smoliar, S. and Zhang, H. Content-based video indexing and retrieval. IEEE Multimedia Magazine 1, 2, 62-72. 1994.
6. Pfeiffer, S., Lienhart, R., Fischer, S., and Effelsberg, W. Abstracting digital movies automatically. Journal of Visual Communication and Image Representation 7, 4 (Dec.), 345-353, 1996.
7. Smith, Michael A. and Takeo Kanade. Video skimming and characteriza-tion through the combination of image and language understanding. IEEE International Conference on Computer Vision and Pattern Recognition, pages 775-781, 17-19 June 1997.
8. Nam J. and Ahmed H. Tewfik. Video abstract of video. IEEE 3rd Workshop on Multimedia Signal Processing, pages 117-122, 13-15 September 1999.
9. Di Lecce, V., G. Dimauro, A.Guerriero, S.Impedovo, G.Pirlo, and A.Salzo. Image basic features indexing techniques for video skimming. IEEE Inter-national Conference on Image Analysis and Processing, pages 715-720, 27-29 September 1999
10. Omoigui, N., L. He, A. Gupta, J. Grudin and E. Sanocki, "Time-compression: System concerns, usage, and benefits", Proc. of ACM Confer-ence on Computer Human Interaction, 1999.
11. Peker, K. A. and Divakaran, A. Adaptive fast playback-based video skim-ming using a compressed-domain visual complexity measure. In ICME'04. Taipei, Taiwan. 2004.
12. Cooper, M. and Foote, J. Summarizing video using non-negative similarity matrix factorization. In IEEE Workshop on Multimedia Signal Processing (MMSP'02). St. Thomas, US Virgin Islands, 25-28, 2002.
13. Gong, Y. and Liu, X. Video summarization and retrieval using singular value decomposition. ACM Multimedia Systems Journal 9, 157-168, 2003.
14. Gong, Y.-H. Summarizing audio-visual contents of a video program. EURASIP J. on Applied Signal Processing: Special Issue on Unstructured Information Management from Multimedia Data Sources 2003, 2.
15. Ngo, C.-W., Ma, Y.-F., and Zhang, H.-J., Automatic video summarization by graph modeling. In International Conference on Computer Vision (ICCV'03). Vol. 1. Nice, France, 2003.
16. Sundaram, H. and Chang, S.-F., Video skims: Taxonomies and an optimal generation framework. In IEEE International Conference on Image Process-ing (ICIP'02). Rochester, NY, 2002.
17. Lienhart Rainer, Silvia Pfeiffer, and Wolfgang Effelsberg. Video abstract-ing. Communications of ACM, 40:55-62, December 1997.

18. Chang, P., Han, M., and Gong, Y. Extract highlights from baseball game video with hidden markov models. In IEEE International Conference on Image Processing (ICIP'02). Rochester, NY. 2002.
19. Xiong, Z., Radhakrishnan, R., and Divakaran, A. Generation of sports highlights using motion activity in combination with a common audio feature extraction framework. In ICIP'03. Vol. 1. Barcelona, Spain, 5-8. 2003.
20. Taskiran, C. M., Amir, A., Ponceleon, D. B., and Delp, E. J. Automated video summarization using speech transcripts. In Storage and Retrieval for Media Databases. Proc. SPIE, vol. 4676. 371-382. 2001.
21. Li, Y., Narayanan, S., and Kuo, C.-C. J. Movie content analysis, indexing and skimming via multi-modal information. In Video Mining, A. Rosenfeld, D. Doermann, and D. DeMenthon, Eds. Kluwer Academic Publishers, Chapter 5. 2003.
22. Lienhart, R., Silvia Pfeiffer, and Wolfgang Effelsberg. Scene determination based on video and audio features. IEEE Conference on Multimedia Computing and Systems, pages 07-11, 1999.
23. Ueda, H., T. Miyatake, and S. Yoshizawa. An interactive natural motion-picture dedicated multimedia authoring system. ACM SIGCHI 91, pages 343-350, 1991.
24. Ferman, A.M. and A. Murat Tekalp. Multiscale content extraction and representation for video indexing. SPIE on Multimedia Storage and Archiving Systems II, 3229:23-31, 1997.
25. DeMenthon, D., V. Kobla, and D. Doermann. Video summarization by curve simplification. ACM International Conference on Multimedia, pages 211-218, August 1998.
26. Calic, J. and Izquierdo, E. Efficient key-frame extraction and video analy-sis. In ITCC'02. 28-33, 2002.
27. Liu, T., Zhang, H.-J., and Qi, F. A novel video key-frame extraction algo-rithm based on perceived motion energy model. IEEE Transaction on Cir-cuits and Systems for Video Technology 13, 10 (Oct.), 1006-1013, 2003.
28. Pope, A., R. Kumar, H. Sawhney, and C. Wan. Video abstraction: summarizing video content for retrieval and visualization. Conference Record of the Thirty-Second Asilomar Conference, I:915-919, 1998.
29. Calic, J. and Thomas, B. T. Spatial analysis in key-frame extraction using video segmentation. In Workshop on Image Analysis for Multimedia Inter-active Services (WIAMIS'2004). Lisboa, Portugal, 2004.
30. Stefanidis, A., A. Partsinevelos, and A. Doucette. Summarizing video data-sets in the spatiotemporal domain. 11th International Workshop on Data-base and Expert Systems Applications, pages 906-912, 4-8 September 2000.
31. Chang, H. S., Sull, S., and Lee, S. U. Efficient video indexing scheme for content-based retrieval. IEEE Transactions on Circuits and Systems for Video Technology 9, 8 (Dec.), 1269-1279. 1999.
32. Yahiaoui, I., Mrialdo, B. and Huet, B. Optimal video summaries for simu-lated evaluation. CBMI 2001 - European Workshop on Content-Based Mul-timedia Indexing, September 19-21, 2001 Brescia, Italy.
33. Yahiaoui, Itheri, Mrialdo, B. and Huet, B. Comparison of multi-episode video summarization algorithms. EURASIP Journal on applied signal proc-essing Special issue on multimedia signal processing - Volume 2003 N1, January 2003 , pp 48-55.

34. Zhuang, Y., Rui, Y., Huang, T., and Mehrotra, S. Adaptive key frame extraction using unsupervised clustering. In International Conference on Im-age Processing (ICIP'98). Chicago, Illiois, 866-870. 1998.
35. Yu, X.-D., Wang, L., Tian, Q., and Xue, P. Multi-level video representation with application to keyframe extraction. In International Conference on Multimedia Modeling (MMM'04). Brisbane, Australia, 117-121. 2004.
36. Diklic, D., D. Petkovic, and R. Danielson. Automatic extraction of representative keyframes based on scene content. Conference Record of the Thirty-Second Asilomar Conference on Signals, Systems Computers, I:877 -881, 1998.
37. Ding, W., G. Marchionini, and T. Tse. Previewing video data: browsing key frames at high rates using a video slide show interface. International Symposium on Research, Development and Practice in Digital Librairies (ISDL'97), pages 425-426, 1997.
38. Yahiaoui, Itheri, Mrialdo, B. and Huet, B. Generating summaries of multi-episodes video. ICME 2001, International Conference on Multimedia and Expo , August 22-25, 2001 Tokyo, Japan.
39. Frederic Dufaux. Key frame selection to represent a video. International onference on Image Processing, II:275-278, 2000.
40. Huang, M., Ayesh Mahajan, and Daniel DeMenthon, Automatic Perform-ance Evaluation for Video Summarization, University of Maryland Tech-nical Report LAMP-TR-114, CAR-TR-998, CS-TR-4605, UMIACS-TR-2004-47, June 2004.
41. Yahiaoui, I., Construction automatique de rsums videos. PhD Thesis, In-stitut Eurecom, 2003.
42. Ying Li, Tong Zhang, Daniel Tretter. An Overview of Video Abstraction Techniques. Imaging Systems Laboratory HP Laboratories Palo Alto, Tech-nical Report: HPL-2001-191, July 31st, 2001.
43. Truong B.T. and Venkatesh S. Video Abstraction: A Systematic Review and Classification. The ACM Transactions on Multimedia Computing, Communications, and Applications, Vol. V, No. N, May 2005.
44. Mekenkamp, G., Barbieri, M., Huet, B. and Yahiaoui, I. and Merialdo, B. and R. Leonardi, Generating TV Summaries for CE-devices, MM'02, 10th International ACM Conference on Multimedia, Juan-les-Pins, France , pp 83-84, December 1-6, 2002.
45. Homepage of the SPATION project, http://www.extra.research.philips.com/euprojects/spation