EURASIP Journal on Applied Signal Processing 2003:1, 48–55

Comparison of Multiepisode Video Summarization Algorithms

Itheri Yahiaoui
Department of Multimedia Communications, Institut Eurécom, BP 193, 06904 Sophia Antipolis, France
Email: itheri.yahiaoui@eurecom.fr

Bernard Merialdo
Department of Multimedia Communications, Institut Eurécom, BP 193, 06904 Sophia Antipolis, France
Email: bernard.merialdo@eurecom.fr

Benoit Huet
Department of Multimedia Communications, Institut Eurécom, BP 193, 06904 Sophia Antipolis, France
Email: benoit.huet@eurecom.fr

Received 26 April 2002 and in revised form 30 September 2002

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 that is automatic, yet related to the user's perception. The method is studied for the case of multiepisode video, where we do not describe only 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 and phrases: multimedia content analysis, video summaries, image similarity, automated evaluation.

1. INTRODUCTION

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 identify what should be kept in the summary and how relevant information can be automatically extracted. To perform this task, we consider several algorithms and compare their performance to define the most appropriate one for our application.

2. RELATED WORK

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 its interpretation. Secondly, while the summarization of a single video has received increasing attention [1, 2, 3, 4, 5, 6], little work has been devoted to the problem of multiepisode 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]. Among the most recent and noticeable work, the work of the Sundaram and Chang [10, 11] based on both audio and video content, developed a measure of complexity and comprehension time of the shot based on user evaluation. The summary presented as a skim of the original is constructed, thanks to a function which describes the utility of each shot. The methods we propose in this paper fall in the same category. However, our approach does not involve users but, instead, is fully automated.

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 users 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 to 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 ones are difficult and expensive to set up and their bias is nontrivial to control, whereas mathematically based ones are difficult to interpret and compare to human judgment.

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 multiepisode 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 multiepisode summaries. Experimental results and a study of summary robustness are presented in Sections 4 and 5. Conclusions and future extensions of the work are presented in Section 6.

3. SIMULATED USER PRINCIPLE

In the simulated user principle, we define a real experimentation, a task that some user has to accomplish, 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 multiepisode video summarization leads to the following scenario for the simulated user experiment:

(i) show all the summaries to the user,
(ii) show a randomly chosen excerpt of a randomly chosen video,
(iii) ask the user to guess which video this excerpt was extracted from.

The simulated behavior of the user is the following:

(i) 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,
(ii) 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,
(iii) if the excerpt contains no image similar 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.

4. COMPARISON OF ALGORITHMS

In this paper, we present several algorithms employed to automatically construct multiepisode video summaries. We intend to compare the quality of the summaries created with our new method (and its possible variations) and other well-known techniques for video or text summarization. The simulated user principle is then used to evaluate the "quality" of the different summaries. Finally, we compare and discuss evaluation results to define the most appropriate algorithm for the task in hand.

Each multiepisode summary building process is divided into five phases: video streams preprocessing, 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 preprocessing

The opening and ending scenes, common to all episodes, are removed from further processing since they are not of 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 subsample 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 Section 4.1.

Summary presentation

Finally, the global summary can be constructed and presented to the user as a 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 (column) is, however, entirely user definable.

4.1. Alternative selection methods

Once video frames have been clustered, the videos might be described as sets of frame classes. The most pertinent classes will be kept. We now see a number of methodologies devised to compute this pertinence value.

First, we describe the principle of the first four methods which are based on our newly proposed ideas. Secondly, the fifth algorithm based on closely related published work is exposed. Finally, we describe method six inspired from the TF-IDF which is commonly employed for the construction of text summaries.

The basic method

Having described the principle of the simulated user, we may now formally describe the summary creation methodology. 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 we consider have duration d. If the video contains N frames, there are the following $N-d+1$ different excerpts:

(i) E_1 contains frames $f_1, f_2, \ldots, f_d$,
(ii) E_2 contains frames $f_2, f_3, \ldots, f_{d+1}$,
(iii) and so on up to E_{N-d+1}, which contains frames f_{N-d+1}, $f_{N-d+2}, \ldots, f_N$.

We assume that the 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

$$f_i \text{ and } f_j \text{ similar} \iff C(f_i) = C(f_j). \tag{1}$$

This is a very strong assumption and the similarity classes are built as described in Section 4.

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

We define the coverage $\mathrm{Cov}(C)$ of a class C as the number of excerpts which contain at least one frame from class C,

$$\mathrm{Cov}(C) = \mathrm{Card}\,\{i : \exists j,\ f_j \in E_i \text{ and } C(f_j) = C\}. \tag{2}$$

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

$$\begin{aligned}\mathrm{Cov}\,(C_1, C_2, \ldots, C_k)&\\ = \mathrm{Card}\,\{i : \exists j, r\ f_j \in E_i \text{ and } C(f_j) = C_1\}&\\ \text{and } 1 \le r \le k.&\end{aligned} \tag{3}$$

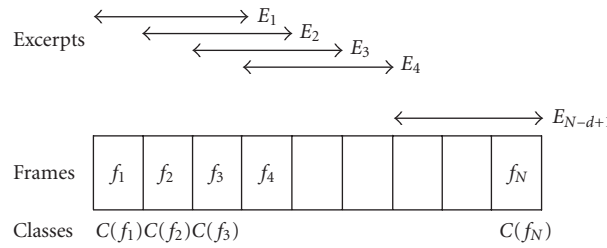

Figure 1: View of excerpts, frames, and classes.

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

$$\mathrm{Cov}\,(C(f_1), C(f_2), \ldots, C(f_k))/(N-d+1). \tag{4}$$

Therefore, the optimal summary is simply one which maximizes

$$\begin{aligned}S = \arg\max_{f_1, f_2, \ldots, f_k} &\mathrm{Cov}\,(C(f_1), C(f_2),\\ &\ldots, C(f_k))/(N-d+1).\end{aligned} \tag{5}$$

This can be achieved in two steps as follows:

(i) first, find a set of classes with maximal coverage,
(ii) second, select a representative frame in each class.

Summary construction

The optimal summary can be found by enumerating all the sets of k classes $\{C_1, C_2, \ldots, C_k\}$ and keeping the best one. Because the enumeration can be computer intensive, it is profitable to carefully select 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 $\{C_1, C_2, \ldots, C_{m-1}\}$, we can define the "conditional coverage" as its contribution to the coverage of the final set

$$\begin{aligned}&\mathrm{Cov}\,(C_m \,|\, C_1 C_2 \cdots C_{m-1})\\ &\quad = \mathrm{Cov}\,(C_1 C_2 \cdots C_m) - \mathrm{Cov}\,(C_1 C_2 \cdots C_{m-1})\\ &\quad = \mathrm{Card}\,\{i : \exists j,\ f_j \in E_i \text{ and } C(f_j) = C_m \text{ and } \forall f \in E_i,\\ &\qquad\qquad \forall r = 1, 2, \ldots, m-1,\ C(f) \neq C_r\}.\end{aligned} \tag{6}$$

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

$$\begin{aligned}\mathrm{Cov}\,(C_1 \cdots C_k) = \mathrm{Cov}\,(C_1) + \mathrm{Cov}\,(C_2 \,|\, C_1)&\\ + \cdots + \mathrm{Cov}\,(C_k \,|\, C_1 \cdots C_{k-1}).&\end{aligned} \tag{7}$$

In order to enable multiple video to be considered at once for summarization, we denote by E_i^v an excerpt of a video v, and S_v a summary for video v; the various cases that have been described in the simulated user behavior can be formally characterized by the following properties:

(i) unambiguous case:

$$\begin{aligned} &\exists v', j, \quad f_j \in E_i^v, \quad C(f_j) \in S_{v'}, \\ &\forall v'' \neq v', \ \forall f_j \in E_i^v, \quad C(f_j) \notin S_{v''}; \end{aligned} \tag{8}$$

(ii) ambiguous case:

$$\exists v' \ \exists v'' \neq v' \ \exists j, \quad f_j \in E_i^v, \quad C(f_j) \in S_{v'}, \quad C(f_j) \in S_{v''}; \tag{9}$$

(iii) unknown case:

$$\forall v', \ \forall f_j \in E_i^v, \quad C(f_j) \notin S_{v'}. \tag{10}$$

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

$$\begin{aligned} \text{Card}\,\{(i,v) : \exists j, \ f_j \in E_i^v \text{ and } C(f_j) \in S_v, \\ \forall v' \neq v, \ \forall f_j \in E_i^v, \ C(f_j) \notin S_{v'}\}. \end{aligned} \tag{11}$$

In the multivideo case, similarity classes are globally defined for all the videos at once. The construction of the summaries becomes more difficult 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 or ambiguous choices. The coverage of a class on a video v is defined as

$$\text{Cov}_v(C) = \text{Card}\,\{i : \exists j, \ f_j \in E_i^v \text{ and } C(f_j) = C\}. \tag{12}$$

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

(i) each summary is initially empty,

(ii) we select each video v in turn and add to its current summary S_v the one class C with maximal value

$$\text{value}_v(C \mid \{S_v\}) = \text{Cov}_v(C|S) - \alpha \sum_{v' \neq v} \text{Cov}_{v'}(C|S), \tag{13}$$

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

$$S = \bigcup_v S_v, \tag{14}$$

(iii) 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 on classes whose coverage on other videos is large, because they are likely to generate ambiguous or erroneous cases in the simulated experiment.

Method 1. This method is based on the coverage value to each class as described previously. In this method, the coverage is computed considering only the excerpts from the video we intend to create a summary for. In other words, for this method, the coefficient α in (13) is set to 0. However, a class may only be selected once, so it cannot represent two videos in the same global summary. In order 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 $\alpha = 1$ in (13) 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 multiepisode 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 multiepisode 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 W_i, which is the proportion of shots from the whole videos that are in cluster i, is computed as $W_i = S_i / \sum_{j=1}^{C} S_j$, 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 [12], where the TF-IDF formula has proven to be very effective. For text

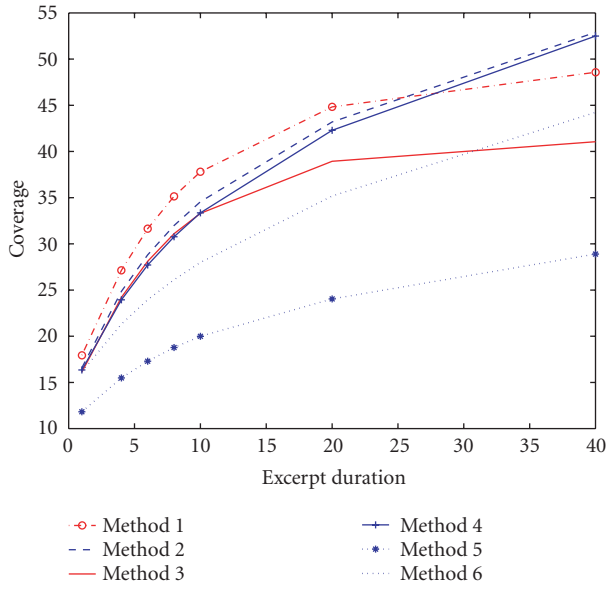

FIGURE 2: Multisummaries coverage as function of excerpt duration.

summarization, this approach is based on terms which represent the items, whereas for multivideo 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.

5. COVERAGE EXPERIMENTS

In this section, we present the evaluation results using the simulated user principal on multiepisodes video summaries created with six different algorithms. As test data, we recorded six episodes of the TV series "Friends." These recordings were compressed Mpeg1 with a digitization rate of 14 frames per second. We fixed the size of the summaries to six segments (which provides a convenient display on a conventional TV 4/3 or 16/9 screen).

The graph in Figure 2 shows the respective performance of the six methods presented in Section 4 when the duration of the excerpt used for both construction and evaluation varies. For additional information the exact data values obtained are presented in Table 1. We note that the first two methods that build summaries, based on a mathematical criterion inspired for the evaluation criterion itself, give the best performance. We note also that the multiepisode summaries (Methods 1 and 2) are more efficient than the single video summaries (Method 3).

Method 5 consists in selecting the rarest and longest shots. Despite the fact that our performance criterion is based on the rate of correct simulated user responses (excerpts for which the simulated user has correctly guessed the corresponding video) and the fact that the rarest shots are by definition highly specific of the video they originate from, the results based on this method are quite poor. The reason is that the rarest shots are likely to be nonredundant (they occur only once) within the video and therefore unlikely to be widely spread across the video. This greatly reduces the number of possible excerpts covering the selected shots and, consequently, low performance results, based on coverage information, are obtained.

Method 6, inspired from TF-IDF, provides rather average results when compared with others. It should also be noted that the results obtained using Method 4 can be compared to those of Method 2 and that both give the best coverage for large excerpts duration.

As a general comment, we can observe on Figure 2 that, for excerpt duration of approximately over 27 seconds, the results of Method 2 (with $\alpha = 1$) and Method 4 (with no ambiguity requirement) perform better than the basic method (Method 1), whereas it was the opposite for shorter excerpts. A possible explanation consist in considering that, for the longest excerpt, the probability of finding similar images in other videos increases. In this approach, the simulated user is more likely to make an incorrect guess, leading to a performance reduction which does not occur with the other two methods.

Similarly, the third method offers reasonable performance results for short excerpt. As the duration of the excerpt used for construction increases, the number of ambiguous cases increases. This is due to the fact that key frames from the same similarity class may be employed in many summaries.

6. ROBUSTNESS OF THE SUMMARIES

Having constructed multiepisode video summaries using six alternative methods, it is of interest to evaluate the performance of the summaries for unrestricted excerpt duration. The first four methods are dependent on the 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 3 presents the results of this experiment for summaries based on Method 1. The choice was made to perform this study using the first method as it is the one which provided the best overall results in the performance experiments.

Note that the construction method itself suggests that the coverage of the summary over the video should be maximum when the same excerpt duration is employed for both construction and evaluation. However, the results shown in Table 2 do not follow the same trend. Indeed, for each of the summaries, the best coverage results are obtained for large evaluation excerpt size. Except in the case of summaries created with excerpt duration of 1 second, all the remaining

Table 1: Summary coverage results for the six methods under consideration.

		Method 1	Method 2	Method 3	Method 4	Method 5	Method 6
Excerpt duration	1	18.0%	16.6%	16.1%	16.4%	11.8%	16.2%
	4	27.1%	24.9%	24.2%	23.9%	15.5%	21.3%
	6	31.6%	28.8%	28.1%	27.7%	17.3%	24.0%
	8	35.2%	32.0%	31.1%	30.8%	18.8%	26.1%
	10	37.8%	34.6%	33.3%	33.4%	20.0%	28.0%
	20	44.8%	43.2%	38.9%	42.3%	24.0%	35.2%
	40	48.6%	52.9%	41.1%	52.5%	28.9%	44.2%

Table 2: Summary coverage results based on Method 1 with different construction and evaluation excerpt size.

		Construction excerpt duration						
		1	4	6	8	10	20	40
Evaluation excerpt duration	1	18.9%	18.0%	17.8%	17.8%	17.4%	16.3%	15.6%
	4	25.6%	27.1%	27.1%	27.1%	26.8%	25.5%	24.4%
	6	28.8%	31.6%	31.7%	31.7%	31.7%	30.2%	29.0%
	8	31.4%	35.2%	35.3%	35.3%	35.6%	34.0%	32.9%
	10	33.4%	37.8%	38.1%	38.1%	38.8%	37.1%	36.2%
	20	39.1%	44.8%	45.3%	45.3%	47.1%	45.6%	46.1%
	40	42.1%	48.6%	49.0%	49.0%	51.5%	48.7%	53.2%

Table 3: Computation time for the selection process.

	Construction excerpt duration						
	1	4	6	8	10	20	40
Speed (second)	47 (second)	50 (second)	53 (second)	56 (second)	58 (second)	71 (second)	97 (second)

methods provide rather similar performance. The coverage increases in a similar manner for all summaries, which indicates that, for reasonable well-chosen creation excerpt duration, the methods provides robust summaries with respect to evaluation conditions.

7. RUNNING TIMES

Although we have not speed-optimized our code, we provide some informative elements of the complexity of our algorithms. The execution times given in this section have been obtained while running the complete process of automatic summarization (as described in Section 3 on a SUN Ultra10 workstation). It took 5 minutes to compute and store the 2705 region histograms representing all 99 minutes worth of video material (the six Friend episodes). Classification of the representative key frame into 1285 classes based on histogram similarity took approximately 3.5 hours.

The selection method used here is the basic one (Method 1), and the computation time required for various excerpt duration is shown in Table 3. It is interesting to note that, thanks to the implementation of a greedy selection mechanism, computation time increase sublinearly. Finally, evaluation is extremely quick as it takes only about a second to compute the performance of a summary.

Despite the fact that we do not perform automatic video summarization in real time, we believe that, with the appropriate optimization and the use of hardware tools, the construction of multivideo summaries is achievable on demand.

8. CONCLUSION

A comparison of some approaches to automatically construct multivideo summaries has been presented. Based on the newly defined 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.

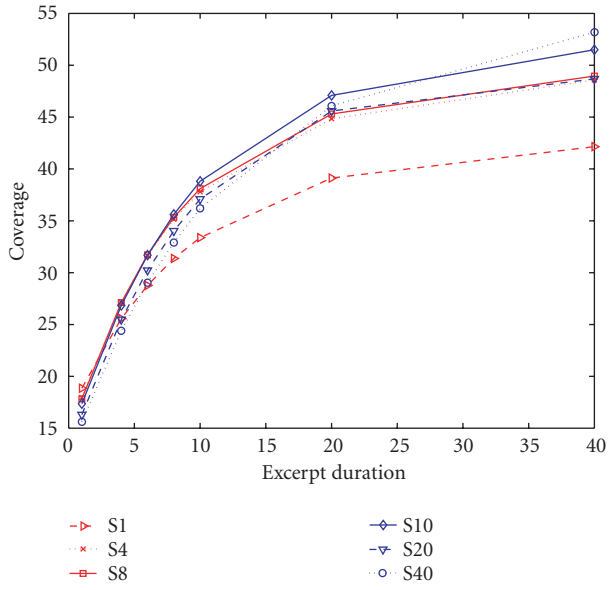

FIGURE 3: Summary coverage as function of excerpt duration using Method 1.

An 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.

ACKNOWLEDGMENT

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

REFERENCES

[1] A. D. Doulamis, N. D. Doulamis, and S. D. Kollias, "Efficient video summarization based on a fuzzy video content representation," in *Proc. IEEE Int. Symp. Circuits and Systems*, vol. 4, pp. 301–304, Geneva, Switzerland, May 2000.

[2] G. Iyengar and A. B. Lippman, "Videobook: An experiment in characterization of video," in *Proc. IEEE International Conference on Image Processing*, vol. 3, pp. 855–858, Lausanne, Switzerland, September 1996.

[3] M. A. Smith and T. Kanade, "Video skimming and characterization through the combination of image and language understanding," in *Proc. IEEE International Workshop on Content-Based Access of Image and Video Databases*, pp. 61–70, Bombay, India, January 1998.

[4] N. Vasconcelos and A. Lippman, "Bayesian modeling of video editing and structure: Semantic features for video summarization and browsing," in *Proc. IEEE International Conference on Image Processing*, vol. 3, pp. 153–157, Chicago, Ill, USA, 1998.

[5] R. Lienhart, S. Pfeiffer, and W. Effelsberg, "Video abstracting," *Communications of the ACM*, vol. 40, no. 12, pp. 54–62, 1997.

[6] S. Uchihashi and J. Foote, "Summarizing video using a shot importance measure and a frame-packing algorithm," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 6, pp. 3041–3044, Phoenix, Ariz, USA, 1999.

[7] B. Merialdo, "Automatic indexing of tv news," in *Workshop on Image Analysis for Multimedia Integrated Services*, pp. 99–104, Louvain-la-neuve, Belgium, June 1997.

[8] M. T. Maybury and A. E. Merlino, "Multimedia summaries of broadcast news," in *Proc. IEEE International Conference on Intelligent Information Systems*, pp. 442–449, Grand Bahama Island, Bahamas, December 1997.

[9] Y. Gong and X. Liu, "Generating optimal video summaries," in *Proc. IEEE International Conference on Multimedia and Expo*, vol. 3, pp. 1559–1562, New York, NY, USA, 2000.

[10] H. Sundaram and S.-F. Chang, "Constrained utility maximization for generating visual skims," in *Proc. IEEE International Conference on Multimedia and Expo*, Tokyo, Japan, August 2001.

[11] S.-F. Chang, "Optimal video adaptation and skimming using a utility-based framework," in *Tyrrhenian International Workshop on Digital Communications*, Capri Island, Italy, September 2002.

[12] I. Mani and M. T. Maybury, Eds., *Advances in Automatic Text Summarization*, M. I. T. Press, Cambridge, Mass, USA, 1999.

Itheri Yahiaoui graduated in the Department of Computer Science and Engineering in 1997 from the University of Batna, Algeria. She received the Diplome d'Etudes Approfondies in Robotics in June 1999 from the University of Pierre et Marie Curie, Paris VI, France. In September 1999, she joined the Multimedia Communications Department at Institut Eurécom to study toward the Ph.D. degree under the supervision of Professor Bernard Merialdo. Her work focuses on the automatic construction of video summaries. Her research interests include multimedia indexing (video analysis, segmentation and classification), content-based retrieval, and image processing.

Bernard Merialdo graduated from the École Normale Supérieure (Mathématiques) in 1975. He received the Ph.D. degree in computer science from Paris 6 University in 1979 and an "Habilitation à Diriger des Recherches" from Paris 7 University in 1992. He first taught at the Faculty of Sciences in Rabat, Morocco. In 1981, he joined the IBM France Scientific Center in Paris, where he led several research projects on natural language processing and speech recognition using probabilistic models. From 1988 to 1990, he was a Visiting Scientist in the IBM TJ Watson Research Center in Yorktown Heights, NY, USA. In 1992, he joined the Multimedia Communications Department of the Institut Eurécom. He is now a full Professor and head of this department. His current research topics are multimedia indexing (video segmentation, analysis and classification, information filtering) and multimedia collaborative applications. His group is involved in several research projects in partnership with other universities and industrial companies. He is an Associate Editor of the IEEE Transactions on Multimedia. He participates in numerous scientific organizations, program committees, expert boards; for example, he is the General Chairman of the ACM Multimedia 2002 conference.

Benoit Huet received his B.S. degree in computer science and engineering from the École Supérieure de Technologie Electrique (Groupe ESIEE, France) in 1992. In 1993, he was awarded the M.S. degree in artificial intelligence from the University of Westminster, UK, with distinction, where he then spent two years working as a Research and Teaching Assistant. He received his Ph.D. degree in computer science from the University of York, UK, in 1999 for his research on the topic of object recognition from large databases. He is currently an Assistant Professor in the Multimedia Information Processing Group of the Institut Eurécom, France. He has published some 35 papers in journals, edited books, and refereed conferences. His research interests include computer vision, content-based retrieval, multimedia data indexing (still and/or moving images), and pattern recognition. He is involved in major scientific organizations, reviewing panels, and program committees.