# Optimal Video Summaries for Simulated Evaluation

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

**Abstract:** The ever-growing availability of multimedia data, such as video, creates a strong requirement for efficient tools to manipulate and present this type of data in an effective manner. Automatic summarization is one of those tools. The idea is to automatically and with little or no human interaction creates a short version or subset of key-frames which contains as much information as possible as in the original video. The key issue here is what should be kept in the summary and how can this relevant information be automatically extracted.

This paper presents a novel methodology for the automatic construction of video summaries. The work is based on the Simulated User Principle to create and evaluate the quality of a video summary in a way, which is automatic, yet related to user's perception. The method is studied for both the case of single video summaries and the more complex case of summarization involving multi-episode videos. The goal, then becomes not only to describe what is important in a video, but what distinguishes this video from the others. Experimental results are presented to support the proposed ideas.

## 1. Introduction

With the increasing computing power, electronic storage capacity and bandwidth of transmission, multimedia information and more particularly digital video is becoming more and more common and very important for education, entertainment and many other applications. The vast quantities of multimedia data have triggered the need and the efforts to provide methodologies and tools for dealing with this type and amount of data. In particular the creation of summarized version of digital videos is a useful tool for allowing a viewer to rapidly grasp the essential content of a video or set of videos. 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 two major limitations. The first of these is the difficult evaluation, in the sense that it is hard to judge the quality of a summary, or, when a performance measure is available, it is hard to understand what the interpretation of this measure is. Secondly, summarization of a single video has received increasing attention [1] [5] [8] [9] [10] [11], comparatively little work has been devoted to the problem of multi-episode video summarization [1] [7] which give rise to interesting issues.

Existing approaches to video summarization can be classified in two categories. A category where rule-based approach are employed and the other where mathematically oriented techniques can be found. The rule based approaches combine evidences from several types of processing (audio, video, natural language) to detect certain configuration of events, which are included in the summary. Examples of this approach are the "video skims" of the Informedia Project by Smith and Kanade [8], and the movie trailers of the MoCA project by Lienhart et al [10]. The automatic creation of movie trailers is a possible application of such methods. The mathematically oriented approaches, on the other hand, use similarities within the video to compute a relevance value of video segments or frames. Possible criteria for computing this relevance include the duration of segments, the inter-segment similarities, and combination of temporal and positional measures. Examples of this approach are the use of the SVD (Singular Value Decomposition) by Gong and Liu [15], or the shot-importance measure by Uchihashi and Foote [11]. The method we propose here falls in the later category.

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

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

This paper is organized as follows. In section 2, describes the basics about video summaries and the simulated user principle approach. In section 3, we describe the application of the principle for single video summarization and present some experimental results. Then, in section 4, we propose the adapted method for the summarization of multi-episode videos along with the corresponding experiments. Conclusions and possible future extensions of the work are then presented in section 5.

## 2. Video Summaries and The Simulated User Principle

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

In general, two types of video summaries are considered:

- “Video skims” are video sequences composed with portions of the original video, to form a shorter version,
- “Key-frames 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 represented by one or several key-frames, 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 key-frames or shots.

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

We also assume that the expected duration of the summary is a summarization process parameter. This corresponds to the practical case where a user wants to spend a specific amount of time (i.e. 20 seconds) watching the summary, or when space to display key-frames is limited. In our experiments, we use a summary length of six key-frames, because we found that this was a good compromise for a display on a computer or TV screen (six images of reasonable resolution 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 and long video segments with uniform content are not appropriate in a summary), so that our choice of six key-frames roughly corresponds to thirty second summaries.

In the introduction we have reported a number of algorithm alternative for creating and evaluating summaries. Our approach, based on the Simulated User Principle is a mathematical criterion which emulates a real user's video summary evaluation. With this approach, we attempt to combine the best of both the creation and the evaluation process. Indeed the use of a mathematical criterion which represents a user-based evaluation, simplifies the "quality" estimation process of the summary. Additionally, our simulated user criterion can be used as a constraint in an optimization process in order to determine the best key-frames for the summary.

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. Of course, not all types of experiments are suitable for the application of this principle. In this paper, we will provide examples of Simulated User Principle for both the cases of single video and multi-episode videos.

## 3. Simulated User Principle for Single Video Summarization

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 have, after experimenting with the most common histogram distance measure ($L_1$ and $L_2$) opted for the $L_2$ norm as described in the experiment section.

We consider that all excerpts have the same duration (which is a parameter is the summarization process) and that they all have the same probability of being chosen. 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 within some probability distribution. However, we have no reasonable interpretation for a variation in duration, neither a reasonable duration probability distribution to suggest.

## 3.1 Automated Optimal Summarization

Having defined Simulated User Principle for single video summaries, a process to automatically construct a summary with optimal (or near optimal) performance for this experiment can be devised.. Assume that the excerpts that we consider have duration d. If the video contains N frames, there are N-d+1 different excerpts:

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

Figure 1 illustrates the relations between excerpts and frames.

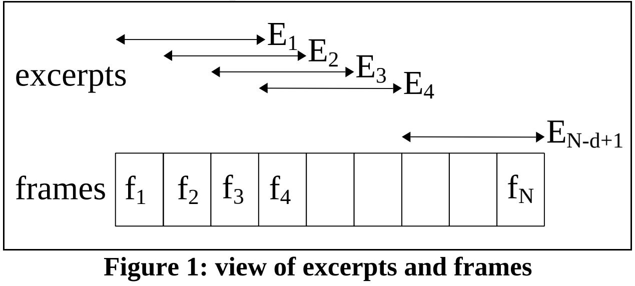

**Figure 1: view of excerpts and frames**

Let us define the coverage Cov(*f*) of a frame f as the number of excerpts which contain at least one frame similar to the frame *f*:

$$\mathrm{Cov}(f) = \mathrm{Card}\{\mathrm{i} : \ \exists \mathrm{j} \ \ \mathrm{f}_{\mathrm{j}} \in \mathrm{E}_{\mathrm{i}} \quad \text{and} \ \ \mathrm{f}_{\mathrm{j}} \ \textit{similar to } f\}$$

The coverage of a set of frames $f_1$, $f_2$, … $f_k$ is the number of excerpts which contain at least one frame similar to one of this frames:

$$\mathrm{Cov}(f_1,\ f_2, \dots\ f_{\mathrm{k}}) = \mathrm{Card}\{\mathrm{i} : \exists \mathrm{j}, \mathrm{l} \ \ \mathrm{f}_{\mathrm{j}} \in \mathrm{E}_{\mathrm{i}} \quad \text{and} \quad \mathrm{f}_{\mathrm{j}} \ \textit{similar to } f_{\mathrm{l}}\}$$

If a video summary is composed of frames *f₁*, *f₂*, … *fₖ*, it induces a performance in the simulated user experiment which equals:

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

Therefore, the optimal summary is simply one which maximizes:

$$S = \underset{f_1,\, f_2,\dots\, f_k}{\arg\max}\ \mathrm{Cov}(f_1,\, f_2 \dots f_\mathrm{k})/(\mathrm{N - d + 1})$$

## 3.2 Summary Construction

The optimal summary can be found by enumerating all the sets of k frames $\{f_1, f_2, \dots f_\mathrm{k}\}$ and keeping the best one. Since the enumeration of all possible combinations is computationally demanding, we use a greedy criterion to restrict selection of the sub-optimal k-best frames. If a frame $f_\mathrm{m}$ is added to an existing set $\{f_1, f_2, \dots f_\mathrm{m-1}\}$, we can define the "conditional coverage" as its contribution to the coverage of the final set:

$$\begin{aligned}\mathrm{Cov}\big(f_\mathrm{m} \mid f_1 f_2 \dots f_\mathrm{m\text{-}1}\big) &= \mathrm{Cov}(\ f_1 f_2 \dots f_\mathrm{m}) - \mathrm{Cov}(\ f_1 f_2 \dots f_\mathrm{m\text{-}1}) \\ &= \mathrm{Card}\left\{\mathrm{i}: \begin{array}{l} \exists \mathrm{j}\ \ \mathrm{f_j} \in \mathrm{E_i}\ \text{and}\ \mathrm{f_j}\ \textit{similar to}\ f_\mathrm{m} \\ \text{and}\ \ \forall \mathrm{f} \in \mathrm{E_i}\ \ \ \forall \mathrm{j} = 1,2,\dots \mathrm{m}-1\ \ \ \mathrm{f}\ \ \textit{not similar to}\ f_\mathrm{j} \end{array}\right\}\end{aligned}$$

Then, the coverage of a set of frames $\{f_1, f_2, \dots f_\mathrm{k}\}$ can be computed as:

$$\mathrm{Cov}(\ f_1 \dots f_\mathrm{k}) = \mathrm{Cov}\big(f_1\big) + \mathrm{Cov}\big(f_2 \big| f_1\big) + \dots + \mathrm{Cov}\big(f_\mathrm{k} \big| f_1 \dots f_\mathrm{k-1}\big)$$

The algorithm to construct the optimal summary proceeds as follows:

- Step 1: start with an empty set of frames,
- Step 2: order the frames that are not similar to those already selected by decreasing conditional coverage with respect to the current set,
- Step 3: add the best frame (according to step 2),
- Step 4: repeat steps 2 and 3 until the summary is complete.

Note that the algorithm starts by selecting the frame $f_1$ with maximal coverage, then $f_2$ with maximal conditional coverage with respect to $f_1$, and so on until $f_\mathrm{k}$. The complete solution found is then the result of a series of greedy choices.

## 3.3 Experimental Results

In our experiments, we are using Mpeg1 encoded videos and since we are only interested in a global analysis of the videos and not in short duration details, videos are sub-sampled to one frame per second. For each frame, a feature vector is built as a 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 its duration to preserve temporal information.

### *3.3.1 Frame similarity*

Color histograms are employed to capture the color distribution characteristics of each key-frame. The similarity between any pair of key-frames is computed by comparing their corresponding color histograms. This is a similar approach to the one of Swain and Ballard for content-based image retrieval [16]. In order to capture some locality information key-frames are divided in nine equal regions from each of which a color histogram is computed. As a result, characteristic key-frames are represented using a vector based on the concatenation of the nine histograms. Key-frames are said to be similar if their histogram difference is smaller than a threshold. We choose a small threshold value in order to enforce strong similarity. The comparison is performed using the standard $L_2$ norm.

### *3.3.2 Summary performance*

We applied our summarization algorithm to the following videos:

- two episodes from the TV serie "*Friends*" (videos F1 and F2),
- a documentary "Histoire d'eau" (video H),
- a fiction "The Avengers" (video C),
- video H1 and C1 are respectively subsequence of video H and C whose duration has been limited to a comparable length than F1 and F2 (˜ 1300sec).

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). Table 1 presents the duration (in seconds) of the different videos.

Each video is summarized independently of the others using the algorithm described in section 3.2. Figure 2 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 the behavior of the summary varies depending on the video. Summaries of episodes from TV series (F1 and F2) 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 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. This statement is re-enforced by the fact that the summaries constructed from subsequence of both H and C (respectively H1 and C1) obtain higher coverage performance than the original videos.

| Video | Duration (in seconds) |
|---|---|
| F1 | 1310 |
| F2 | 1298 |
| H | 2118 |
| C | 3035 |
| H1 | 1284 |
| C1 | 1287 |

**Table 1: video duration**

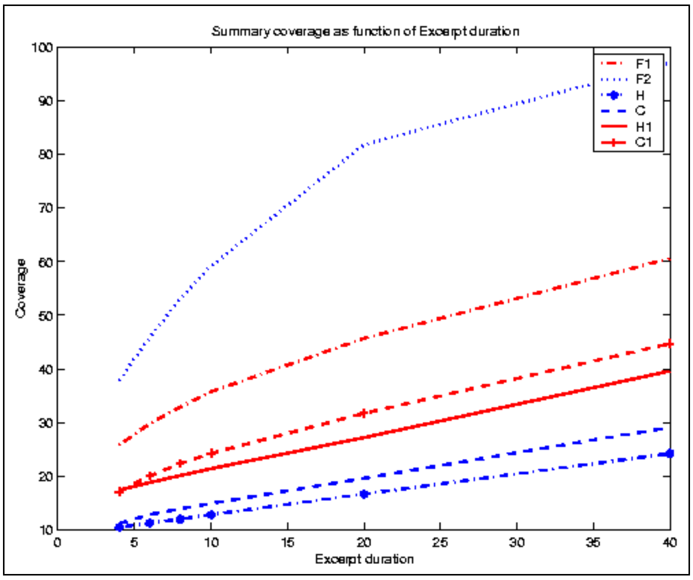

**Figure 2: single video summary coverage**

We ran the summarization algorithm with many duration for the excerpts. 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 near-optimal summaries.

## 4. Simulated User Principle for Multi-Episode Summarization

We now focus on the problem of multi-episode video summarization. As we stated in section 2 the aim here is rather different than in the single video case. In this case, much information (scenes, actors) is present in the same or in similar manner in several videos, so that summarizing videos independently one from each other could include redundant information in the summaries. Specific methods have to be designed to take advantage of those similarities and produce more efficient set of summaries.

Now we adapt the Simulated User Principle that was proposed for single video summarization to take into account the particular case of multi-episode videos. We propose the following scenario for the new Simulated User Experiment:

- All the summaries are shown to the user,
- He is then shown a randomly chosen excerpt of a randomly chosen video,
- He is then asked to guess which video this excerpt was extracted from.
- The simulated behavior of the user is the following:
- If the excerpts contains images which are similar to one or several images in a single summary, he will provide the corresponding video as an answer (but it is not certain that this is the correct answer),
- If the excerpt contains images which are similar to images in several summaries, the situation is ambiguous and the user cannot provide a definite answer,
- If the excerpt contains no image which is similar to any image in any summary, the user has no indication and cannot provide a definite answer.

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.

Once again, this assumes that the user has a perfect visual memory, so that he can immediately identify similar images when they are shown to him. It also assumes that we can closely approximate the user judgement on similarity.

## 4.1 Automated Optimal Summarization

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

- Unambiguous case:

$$\exists v' \ \exists j \quad f_j \in E_i^v \ \exists f_m \text{similar to } f_j \text{ and } f_m \in S_{v'}, \forall v'' \neq v' \forall f_j \in E_i^v \ \forall f_m \text{similar to } f_j \quad f_m \notin S_{v''}$$

- Ambiguous case:

$$\exists v' \exists v'' \neq v' \ \exists j \quad f_j \in E_i^v \quad \exists f_m, f_n \text{ similar to } f_j \text{ and } f_m \in S_{v'} \text{ and } f_n \in S_{v''}$$

- Unknown case:

$$\forall v' \quad \forall f_j \in E_i^v \ \forall f_m \text{ similar to } f_j \quad f_m \notin S_{v'}$$

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

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

We rely on the same image similarity definition as in the single video case. The construction of the summaries becomes more difficult, because when we choose to add a frame in a summary, we have to consider not only the coverage of this frame on this video, which should be high, but also take into account the coverage of this frame on the other videos, which should be low to minimize erroneous choices. The coverage of a frame on a video v is defined as:

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

An exhaustive enumeration of all possible sets of summaries is not computationally tangible, so we use a sub-optimal algorithm to build a good set of summaries. Our algorithm proceeds as follows:

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

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

where S is the set of all frames already included in any of the summary: $S = \bigcup_v S_v$

- When all summaries have the desired size, we iteratively replace any chosen frame if we can find another frame which provides a better value.

## 4.2 Experimental Results

We now consider as test data recordings six episodes from the TV series “Friends”. The videos are processed in a similar manner to the one described for the single video experiments:

- Videos are sub-sampled at one frame per second,
- A feature vector is built for each frame as the color histogram on nine regions of the image,
- Consecutive frames are collapsed into one when their feature vectors are very low.

### *4.2.1 Summary performance*

We ran our summarization procedure on the six “Friend” videos for excerpt duration varying from 4 seconds to 40 seconds. The following table and its respective graphs show the performance of our method.

| Excerpt duration | % correct | % ambiguous | % incorrect |
|---|---|---|---|
| 4 sec | 25.25 | 1.27 | 3.53 |
| 6 sec | 29.87 | 1.61 | 4.79 |
| 8 sec | 33.36 | 2.51 | 6.38 |
| 10 sec | 36.82 | 2.86 | 6.54 |
| 20 sec | 46.70 | 7.02 | 9.54 |
| 40 sec | 54.06 | 15.47 | 13.14 |

**Table 2: coverage over the original videos**

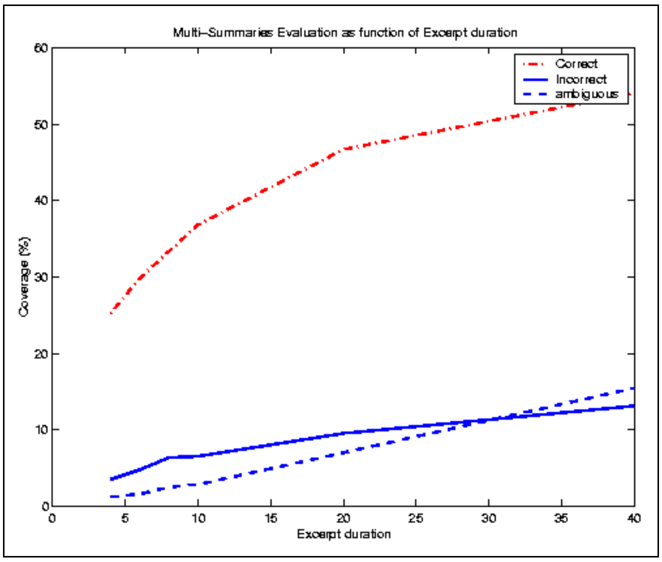

**Figure 3: Evaluation of the summaries**

It can be noticed that as in the single video case, as the duration of the excerpts increases the percentage of correct video identification by the simulated user also increases. The percentage of correct answers is not the only possible measure of evaluation. We can also look at the percentage of incorrect answers and ambiguous cases. It appears that except in the case of large excerpt summaries are more likely to cause incorrect identification than ambiguity.

As an illustration of our summarization process, figure 4 shows the summaries of the six videos for different excerpt duration. Each line contains the six key-frames for one episode (key-frames have been slightly distorted to fit inside the column). Although some differences can be seen between both sets of summaries the majority of the frame remain unchanged for small variation of the excerpt duration.

## 5. CONCLUSION

In this paper, we have proposed a novel approach to automate the creation of video summaries based on the Simulated User Principle. It was devised as a practical way to create and evaluate video summaries, using a performance measure which is based on a simulated user experiment whose results are easily interpretable. We have shown how this principle can be applied to the creation of single video summaries, as well as multi-episode video summaries. Experiments demonstrate the approach is feasible, and that despite their very small size (6 frames) summaries are able to cover at least 30% of the original videos. The algorithms proposed in this paper can be used to automatically generate summaries for video recordings, as the ones that will be available in set-top boxes, or in multimedia databases.

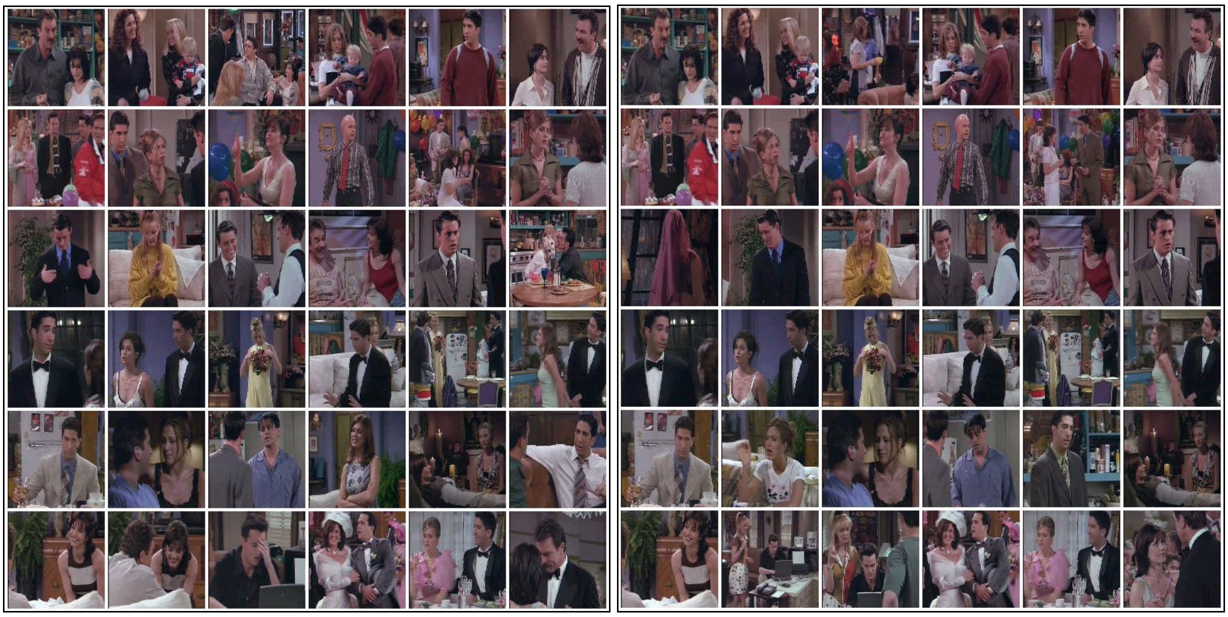

**Figure 4: near-obtimal summaries for excerpt of 6 seconds and 10 seconds respectivelly**

## 6. REFERENCES

[1] Anastasios D. Doulamis et al.. Efficient Video Summarization based on a Fuzzy Video Content Representation. IEEE Intl. Symposium on Circuits and Systems, vol. 4, pp. 301-304, May 28-31, 2000.

[2] Bernard Merialdo. Automatic indexing of TV news. Workshop on Image Analysis for Multimedia Integrated Services, pp. 99-104, 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, pp. 1290 -1298, December 1999.

[5] Giridharan Iyengar and Andrew B. Lippman. Videobook: An experiment in characterization of video, IEEE Intl. Conf. on Image Processing, vol. 3, pp. 855-858, 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, pp. 442 -449, 1997.

[8] M.A. Smith and T. Kanade. Video skimming and characterization through the combination of image and language understanding. IEEE Intl. Workshop on CB Access of Image and Video DB, pp 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 Img. Proc., vol. 3, pp. 153-157, 1998.

[10] Rainer Lienhart, Silvia Pfeiffer and Wolfgang Effelsberg. Video abstracting. Communications of ACM, vol. 40, no. 12, pp 54-62, Dec 1997.

[11] Shingo Uchihashi and Jonathan Foote. Summarizing video using a shot importance measure and a frame-packing algorithm. IEEE International Conference on Acoustic, Speech and Signal Processing, pp. 3041-3044, 1999.

[12] Udo Hahn and Indejeet Mani. The challenges of automatic Summarization. IEEE Computer, vol. 33, no. 11, pp: 29-36, 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 International Conference on Image Analysis and Processing, pp. 715-720, 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, vol. 1, pp. 866-870, 1998

[15] Yihong Gong; Xin Liu. Generating optimal video summaries. IEEE International Conference on Multimedia and Expo, vol. 3, pp. 1559-1562, 2000.

[16] M.~J. Swain and D.~H. Ballard. Color indexing. International Journal of Computer Vision, 7(1):11-32, 1991.