

Evaluation of Automatic Video Summarization Systems

Cuneyt M. Taskiran

Motorola Labs, Multimedia Research Lab
Schaumburg, IL 60196

ABSTRACT

Compact representations of video, or video summaries, data greatly enhances efficient video browsing. However, rigorous evaluation of video summaries generated by automatic summarization systems is a complicated process. In this paper we examine the summary evaluation problem. Text summarization is the oldest and most successful summarization domain. We show some parallels between these two domains and introduce methods and terminology. Finally, we present results for a comprehensive evaluation summary that we have performed.

Keywords: video summarization, video skimming, summary evaluation methods

1. INTRODUCTION

Deriving compact representations of video sequences that are intuitive for users and let them easily and quickly browse large collections of video data is fast becoming one of the most important topics in content-based video processing. In this paper we collectively refer to such representations as *video summaries*. With the proliferation of personal video recorder devices such as hand-held cameras and video cell phones users can easily generate many times more video footage than they can digest. On the other hand, programs such as sports and news must be processed quickly for production or their value quickly diminishes. Thus, there is a growing need for automatic methods to generate video summaries from both user and production viewpoints.

Automatic summarization, be it for video, audio, or a text document, is a challenging and ill-defined task since it requires the summarization system to make decisions about the semantic content and relative importance of parts of the document with respect to each other. A factor that complicates development of automatic video summarization algorithms is that the evaluation of the resulting summaries is problematic, since it is hard to derive quantitative measures of summary quality. Much of the complexity of summary evaluation arises from the fact that it is difficult to specify what one really needs to measure, and why, without a clear formulation of what precisely the summary is aimed to capture.

In this paper we examine the problem of evaluation of automatically generated video summaries. Since automatic video summarization is still an emerging field, serious questions remain concerning the appropriate methodology in evaluating the quality of the generated summaries. Although there are many summarization algorithms proposed in the literature, works that include a systematic evaluation of the proposed summaries are relatively rare. This makes it difficult, if not impossible, to compare different algorithms and assess their relative merits. Automated text summarization dates back at least to Luhn's work at IBM in the 1950s,¹ which makes it the most mature area of media summarization. We believe that the evaluation approaches and methodology developed in the text summarization domain can be useful in developing evaluation schemes for video summarization.

2. CHARACTERIZATION OF VIDEO SUMMARIES

2.1. Parallels with Text Summarization Systems

Researchers working in the automated text summarization field have identified three distinct stages in the summarization process²: topic identification, interpretation, and summary generation. In the topic identification stage the system identifies the most important units of the text, e.g., words, sentences, paragraphs. During the interpretation stage the identified topics are fused and expressed using concepts and words not found in the original text. Finally, in the summary generation stage the coherence of the text is improved by handling dangling references and repeated material. Almost all text summarization systems today only contain the first stage.³ Therefore they are *text extraction* systems as opposed to *text abstraction* systems that contain the other

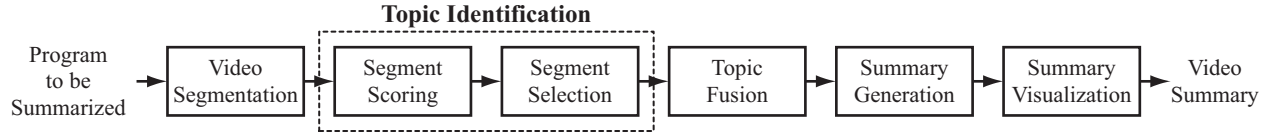


Figure 1. Block diagram of a general video summarization system.

two stages and that can paraphrase the original text to create an abstract. This limitation is due to the fact that the interpretation stage necessary to perform text abstraction requires domain knowledge which is hard to build.

The above concepts can also be applied to video summarization. However, there seems to be no clear parallel in the video summarization domain to text abstraction, since “paraphrasing” a video segment is not as well-defined as text. Hence, practically all video summarization systems are extraction systems. Motivated by the above ideas from text summarization, a block diagram of a video summarization system is given in Figure 1, In such a system, the first processing block is the temporal segmentation of video. The size of the video segments to be used is an open issue. Some systems use shot boundaries while others divide the video on speech pause boundaries.⁴ Still other systems forego the temporal segmentation step and either use individual frames as the processing unit^{5,6} or sample the video at fixed instances⁷ followed by clustering to obtain segments.

The next processing step is to calculate several importance scores for segments by video, audio, and text analysis modules working in parallel. One segment scoring method that is widely used in text summarization is to use positional criteria, that use text structure to determine sentence importance, e.g., choosing the first sentence of each paragraph. A similar approach may be applied for video summarization, based on a particular domain. For example, only anchor shots may be selected for the summary of a news program. Another approach is to cue words to detect important parts of a document. This method was applied to video summarization by a number of researchers using the closed-caption program text and cue phrases such as “please welcome”⁸ as well as named entity recognition.^{9,10} Segment scores may also be using data from other modalities. For videos with event-based content (see Section 2.3) audio and visual events in the segments can be detected and used to derive scores. The works cited in^{11,12} are examples of this approach to soccer video summarization.

After a set of scores are calculated for segments the segment selection module combines the scores for each segment into a single segment quality score and returns a selection of the highest scoring segments, according to the summary length requested by the user. The selected segments may then be processed further by the topic fusion module that employs domain knowledge if it is present. For example, using soccer domain knowledge, events detected in soccer videos may be collected together and summarized further. As pointed out in the discussion of text summarization systems, this step is currently skipped in most systems, since it requires complex domain knowledge. In the summary generation module the viewability of the summary is improved by removing repetitive video segments and other discrepancies, i.e., references to segments of video that were not included in the summary. Because of its complexity, this processing step is also skipped in most video summarization systems.

Finally, the generated summary has to be displayed to the user in an intuitive and compact manner, which is done by the summary visualization module. Video visualization schemes proposed in the literature mainly fall into two categories: Storyboards^{5,6,13} based on keyframes extracted from video and video skims^{4,14-16} where portions of the source video are concatenated to form a much shorter video clip. Storyboards are more compact and present “at a glance” information about video content; however, they also some present fundamental drawbacks, since they do not preserve the time-evolving nature of video programs. They are somewhat unnatural and hard to grasp for non-experts, especially if the video is complex. Most techniques just present keyframes to the user without any additional metadata, like keywords, which can make the meaning of keyframes ambiguous. Storyboards also make all shots appear equally important to the user and the representation becomes unpractically large for long videos. This last problem may be alleviated by sizing the keyframes according to their score. This approach have been used in¹⁷ and¹³ where keyframes from segments are arranged in a “video poster” using a frame packing algorithm.

2.2. Summary Application Domains

The properties of a video summary depends on the application domain, the characteristics of the sequences to be summarized, and the purpose of the summary. Some of the purposes that a video summary might serve are listed below.

- Intrigue the viewer to watch the whole video. Movie trailers, prepared by highly skilled editors and with high budgets, are the best examples of summaries of this type.
- Let the viewer decide if the complete program is worth watching. Summaries of video programs that may be used in personal video recorders are examples of this category of summaries.
- Help the viewer locate specific segments of interest. For example, in distance learning, students can skip parts of a video lecture that they are familiar with and, instead, concentrate on new material.
- Let users judge if a video clip returned by a video database system in response to their query is relevant. In content-based image database applications the results of a user query are shown as thumbnail images, which can be judged at a glance by the user for relevance to the query. Judging the relevance of query results is time-consuming for video sequences, since search results may contain long sequences containing hundreds of shots.
- Enable users of pervasive devices, such as personal digital assistants, palm computers, or cellular phones, to view video sequences, which these devices otherwise would not be able to handle due to their low processing power. Using summaries also result in significant downloading cost savings for such devices.
- Give the viewer all the important information contained in the video. These summaries are intended to replace watching the whole video. Executive summaries of long presentations or videoconferences would be an example of this type of summary.

The above list, while not exhaustive, illustrates the wide range of different types of video summaries one would like to generate. For most applications video summaries mainly serve two functions: the *indicative function*, where the summary is used to indicate what the original program is about while providing minimum content; and the *informative function*, where the summaries are used to cover the information in the source program as much as possible, subject to the summary length. Clearly these two summary functionalities are not independent. Video summarization applications often will be designed to achieve a mixture of the two functionalities. The viewer's available time and the environment where the summary will typically be consumed, as well as the display characteristics of the device used, are also important factors in determining the type of summary to generate. For example, a summary of a lecture must cover the main results and conclusions, while a movie trailer must not reveal the punch line. Naturally, there is no single approach, either manual or automatic, that will apply to the generation of all types of video summaries.

2.3. Types of Program Content

An important distinction we make when considering video summarization algorithms is the type of the program content that will be summarized. For the purposes of video summarization we categorize video program content into two broad classes:

- *Event-based content.* Video programs of this type contain easily identifiable story units that form either a sequence of different events or a sequence of events and non-events. Examples of the first kind of programs are talk shows and news programs where one guest or news story follows another and their boundaries are well-defined. The best example of programs where sequence of events and non-events occur are sports programs. Here, the events may correspond to important instances in games, such as touchdowns, home runs, or goals.

- *Uniformly informative content.* These are programs which cannot easily be broken down to a series of events as event-based content. For this type of content, many parts of the program may be equally important for the user. Examples of this type of content are sitcoms, presentation videos, documentaries, soap operas, and home movies.

Note that the distinction introduced above is not clear cut. For example, for sitcoms one can define events according to audience laughter in the soundtrack. Movies are another example: most action movies have a clear sequence of action and non-action segments.

For event-based content, since the types of events of interest are well-defined, one can use knowledge-based video event detection techniques. In this case, the processing is generally domain-specific and a new set of events and event detection rules must be derived for each application domain, which is a disadvantage. However, the summaries produced will be more reliable than those generated using general-purpose summarization algorithms.

3. EVALUATION OF VIDEO SUMMARIES

3.1. General Summary Quality Criteria

Evaluation of the quality of automatically generated video summaries is a complicated task since it is difficult to derive objective quantitative measures for summary quality. In order to be able to measure the effectiveness of a video summarization algorithm one first needs to define features that characterize a good summary, given the specific application domain being studied. In general, a good video summary must have the following characteristics

- it must be considerably shorter than the original video sequence;
- it must contain the important information of the original video, where the importance measure depends on the particular application; and
- it must be easy for users to grasp and follow.

Note that these requirements are by necessity loosely defined, since, as described in Section 2, summaries serve many different application domains, and each domain may have different quality criteria. Furthermore, it may be impossible to maximize all three of the above characteristics. The second requirement is related to the *coverage* of the summary while the third one is related to its *detail*. Ideally one would like the generated summaries to be detailed and to have high coverage, that is, cover most of the important topics of the full program while providing most of the details for each individual topic. Clearly, the creation of a short summary imposes a trade-off between coverage and detail.

Given a full program F and a summary based on it, S , we can define two measures that reflect the first two constraints given above as³

$$\text{Compression Ratio (CR)} = \frac{\text{length of } S}{\text{length of } F} \quad (1)$$

$$\text{Retention Ratio (RR)} = \frac{\text{information in } S}{\text{information in } F} \quad (2)$$

A good summary may then be defined to be one where CR is low and RR is high. Summarization systems may be characterized by plotting these two ratios, a number of examples are shown in Figure 2. The most desirable curve is the one in Figure 2(c) where the system has incorporated most of the important information using a short summary.

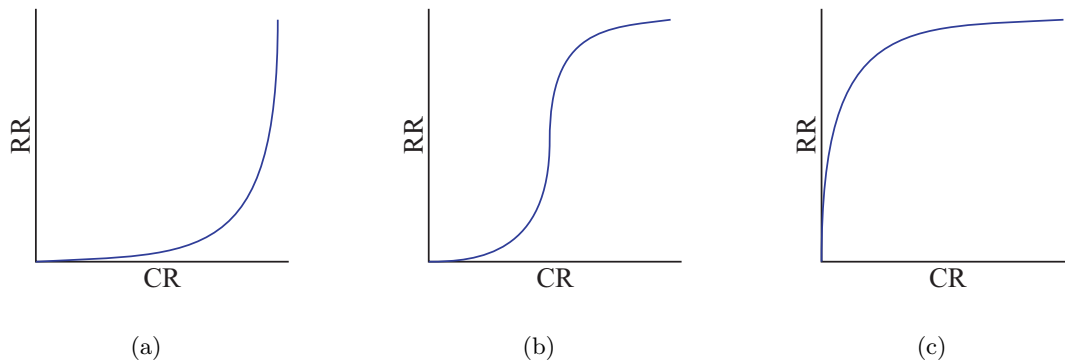


Figure 2. Compression Ratio (CR) versus Retention Ratio (RR) curves for different types of summarization systems.

3.2. Intrinsic versus Extrinsic Evaluation Methods

In text summarization, researchers distinguish between two different approaches to evaluate systems: *intrinsic* and *extrinsic* evaluation methods.^{18,19} In intrinsic evaluation methods, the quality of the generated summaries is judged directly based on the analysis of summary. The criteria used may be user judgment of fluency of the summary, coverage of key ideas of the source material, or similarity to an “ideal” summary prepared by humans. On the other hand, in extrinsic methods the summary is evaluated with respect to its impact on the performance for a specific information retrieval task. Both approaches have their shortcomings. In intrinsic evaluation it is hard to derive ideal summaries, since even professionals may not agree on which parts of the document to include in the summary, except for very short summaries. Also, summaries receive different scores with different measures, or when compared to different (but presumably equivalent) ideal summaries created by humans.²⁰ The major problem in extrinsic evaluation is to ensure that the metric applied correlates well with task performance efficiency, otherwise summaries will score lower on harder tasks regardless of summary quality.³

The main difficulty in summary evaluation is to measure the Retention Ratio (RR) of a summary, since this value depends on the interesting information content of the summary. There are several ways to estimate RR . For event-based content, e.g., a sports program, where interesting events are unambiguous, summaries might be judged on their coverage of these events in the source video. Two such evaluations are given in²¹ and²² Ekin and Tekalp²¹ give precision and recall values for goal, referee, and penalty box detection, which are important events in soccer games. In Ferman and Tekalp’s study,²² the video summary was examined to determine, for each shot, the number of redundant or missing keyframes in the summary. For example, if the observer thought that an important object in a shot was important but no keyframe contained that object, this resulted in a missed keyframe.

For uniformly informative content, where events may be harder to identify, different evaluation techniques have been proposed. One common approach is the quiz method, where RR is estimated by determining how well S allows subjects to answer questions derived from F . In their study He *et al.*¹⁶ gave users a multiple choice quiz derived from the original videos of presentations before and after watching a video skim extracted from them. The quizzes consist of questions prepared by the presentation speakers and are assumed to reflect the key ideas of the presentation. The quality of the video skims were judged by the increase in quiz scores. A similar technique was used by Taskiran *et al.*⁴ in evaluating video skims extracted from documentaries. The quiz method has some drawbacks: First, it was found that this approach may have difficulty differentiating between different summarization algorithms depending on program content.^{4,16} Second, it is not clear how quiz questions can be prepared in an objective manner, except, perhaps, by authors of presentations who are usually not available. Finally, the concept of a “key idea” in a video program is ambiguous and may depend on the particular viewer watching the skim. Another extrinsic evaluation method was used by Christel *et al.*²³ In this study video skims extracted from documentaries were judged based on the performance of users on two tasks: fact-finding, where users used the video skims to locate video segments that answered specific questions; and gisting, where users

matched video skims with representative text phrases and frames extracted from source video.

In intrinsic evaluation of text summaries generally summaries created by experts are used.¹⁸ Using a similar approach would be much more costly and time consuming for video sequences. Another scheme for evaluation may be to present the segments from the original program to a large number of viewers and let them select the segments they think should be included in the summary, thereby generating a ground truth. This seems to be a promising approach although agreement among human subjects becomes an issue for this scheme. A commonly used intrinsic evaluation method is to have users rate the skims based on subjective questions, e.g., “Was the summary useful?” and “Was the summary coherent?” Such surveys were used in.^{15, 16, 23}

4. EXPERIMENTS AND RESULTS

We have developed a video summarization algorithm where segments are determined by pause boundaries in speech. Segment scores are calculated using term frequencies obtained using ASR and a dispersion measure that maximizes summary coverage. The algorithm is described in detail in.²⁴ In order to test the quality of the summaries produced we conducted experiments using both intrinsic and extrinsic evaluation methods. The number of questions derived from the full program that the summary contains was deemed to be one of the intrinsic evaluation measures. We have also asked subjects to rate the summaries. As the extrinsic evaluation method we used the quiz method described in Section 3.2. Below we summarize some of our results, full results are given in.²⁴

4.1. Sequences and Algorithms Used

We chose three documentary programs to use in our study. The original lengths of these programs were 60 minutes, 210 minutes, and 93 minutes for the *Seahorse* (SH), *MarkTwain* (MT), and *WhyDogsSmile* (WDS) documentaries, respectively. In order to have summaries of manageable length and to have all full programs have approximately the same length, we selected a 20 minute portion of each program near the beginning and used these as our full-length programs. In broadcast documentaries a summary of the whole documentary is sometimes presented at the beginning of programs. Since this would have interfered with our evaluation, we have taken care not to include such documentaries in the experiment. We selected program content to minimize the chance of subjects’ having prior knowledge about the programs.

Three algorithms, abbreviated as **FREQ**, **RAND**, and **DEFT**, were used on the full programs to generate three different skims for each full program. We denote the summary of the *Seahorse* program using the **RAND** algorithm by **SH+RAND**, and similarly for the other summaries. The algorithm **FREQ** our summarization algorithm. The **RAND** algorithm detects the segments in the program, using the same pause detection algorithm as was used for **FREQ**, but randomly selects the segments to be included in the summary. The **DEFT**, or default, algorithm is the simplest video summarization algorithm possible, consisting of subsampling the video program temporally at fixed intervals. This algorithm divides the full program into $\lfloor T_{\text{summary}}/T_s \rfloor$ temporal intervals, and selects the first T_s seconds of each interval for the summary. Our preliminary tests suggest a value of $T_s = 5$ seconds to be the minimum value that viewers feel comfortable with. The main difference between **RAND** and **DEFT** is the segmentation of video on audio pauses. In²³ pilot testing of the video summaries was used to derive the conclusion that users preferred summary segments based on audio pauses. However, this factor was not tested in isolation to determine its effect on an extrinsic task. We included the **RAND** algorithm to study this factor.

A factor that can influence evaluation results is the value of the summarization ratio used to obtain the video skims. The evaluation results of the same summarization system can be significantly different when it is used to generate summaries of different lengths.¹⁸ In our experiments we have kept the summarization ratio constant at $f = 0.1$.

4.2. Experimental Design

The 48 subjects participating in the experiment were randomly divided into three groups of 16. Each subject watched three video skims. In order to minimize learning and other unforeseen cross-over effects, both the order of the content and the algorithms were different for the three groups, as shown in Table 1. Each subject received \$10 for participation in the experiment and participated in the experiment individually. After watching each

Table 1. Experimental summary viewing order.

Group	first	second	third
Group 1	MT+FREQ	SH+RAND	WDS+DEFT
Group 2	WDS+RAND	MT+DEFT	SH+FREQ
Group 3	SH+DEFT	WDS+FREQ	MT+RAND

Table 2. Quiz performance for 10 questions.

	FREQ	RAND	DEFT
est. mean, $\hat{\mu}$	6.542	5.313	4.604
est. std. deviation, $\hat{\sigma}$	1.650	1.652	1.865
est. std. error = $\hat{\sigma}/\sqrt{n}$	0.238	0.238	0.269

video skim, the subjects first answered three questions about the quality of the summary they just watched. Then, they answered ten multiple choice questions derived from the full program and containing important facts about the content presented in the skim. While watching the summaries the users were not able to pause the video or jump to a specific point in it.

The questions that were presented to the subjects were determined by two judges. One judge was the first author and the other judge was naive about the algorithms used. Each judge independently marked the parts of the closed-caption transcripts of the programs that they deemed were the important points of the programs without watching any of the summaries. The intersection of these two lists of marked locations were used to generate the questions.

4.3. Results for Extrinsic Evaluation with the Quiz Method

The number of correct answers out of 10 questions that the subjects scored after watching the summaries generated by the algorithms, summed over the three experimental groups are listed in Table 2.

From these results we can see that the FREQ algorithm performed better than RAND and DEFT. In order to assess the statistical significance of the results in Table 2, we list pairwise comparison p -values in Table 3 obtained using a two-sided t -test. The p -value refers to the probability of incorrectly rejecting the hypothesis that two populations have identical mean values. The true p -values are larger, however. The reason is that Table 2 shows the results of testing three hypotheses, not one. Using a cut-off value equal to $p = 0.05/3 = 0.0167$, the differences in scores between FREQ and RAND, and FREQ and DEFT are statistically significant, while the difference between RAND and DEFT is not statistically significant.

In order to investigate the effect of program content on the performance of the algorithms, the mean correct scores that the subjects obtained from the multiple choice questions for the three documentary programs are compared in Figure 3. Assuming the value $p = 0.05$ as the cut-off value for deciding statistical significance, and correcting for the fact that we are performing nine tests, statistically significant differences were found between FREQ–RAND and FREQ–DEFT for *MarkTwain*, and among all three algorithms for *WhyDogsSmile*. This implies that the FREQ algorithm produced more informative summaries.

The theoretical maximum scores, c_{\max} , that the subjects can obtain for the three programs are shown as horizontal lines on the bars in Figure 3. These values were calculated using the relation

$$c_{\max} = c_{\text{inc}} + 0.25(10 - c_{\text{inc}}),$$

where c_{inc} is the number of answers determined by the judges included in the summary of each program, which are listed in Table 4, and 0.25 is the probability of getting a correct answer by guessing. This simple model assumes that the subjects have perfect memory and full attention, therefore always correctly answering the questions that were covered by the summary. As seen in Figure 3, in all cases but one the mean score of the subjects is higher than the theoretical maximum. These differences are mainly due to subjects’ previous knowledge on the topics, which is impossible to completely eliminate and “guesstimation” of the correct answer using visual cues and inference.

Table 3. Pairwise p -values using a two-sided t -test.

	FREQ	RAND
RAND	4.348×10^{-4}	
DEFT	5.189×10^{-7}	5.183×10^{-2}

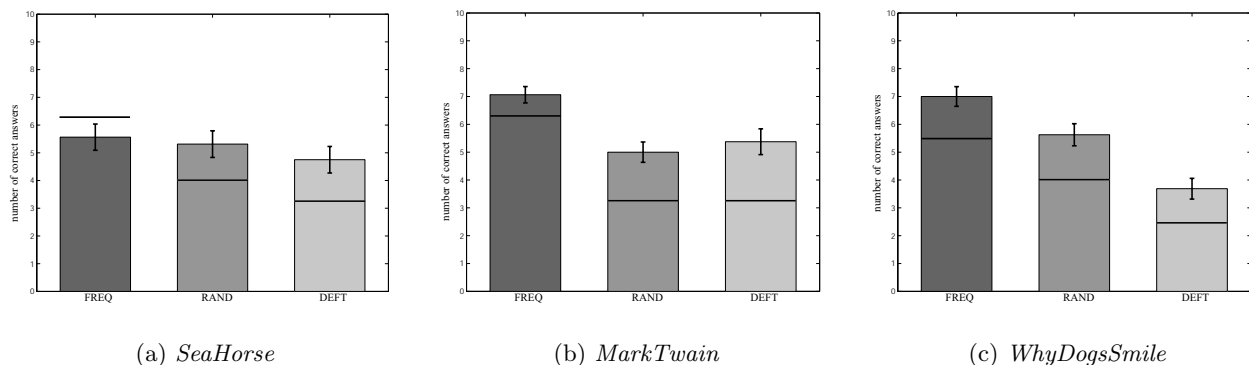


Figure 3. Effect of program content on scores for the three test programs.

4.4. Results for Intrinsic Evaluation

We examined each of the nine summaries for each documentary–algorithm pair and determined how many answers each summary contains. These numbers are tabulated in Table 4. As can be seen from this table, the information covered by summaries generated using FREQ contain significantly more information about the full programs compared to the other two algorithms. Since the summarization ratio was $f = 0.1$ we would expect the random summaries to contain one answer on the average. The results for RAND and FREQ in Table 4 agree well with this prediction.

As the second intrinsic evaluation scheme, after watching each summary we asked the subjects to answer two assessment questions. The questions were “I found the summary to be clear and easy to understand” and “I feel that I can skip watching the whole program because I watched this summary.” Subjects rated both of these statements on a scale of 1 to 5. The medians of the subjects ratings for the three algorithms, are shown in Table 5. Although the interpretation of such subjective quality assessments may be difficult and error-prone, from this data we can make the observation that the correlation between the performance in extrinsic task we used (question answering) and the subjective assessment of subjects is not very strong.

5. CONCLUSIONS

In this paper examined the problem of evaluation of summaries generated by automatic video summarization systems. We showed some parallels from the text summarization domain and introduced some methods and terminology that are used in the field. Finally, we presented results for a comprehensive evaluation summary that we have performed.

REFERENCES

1. P. H. Luhn, “Automatic creation of literature abstracts,” *IBM Journal*, vol. 2, no. 2, pp. 159–165, 1958.
2. K. Sparck-Jones, “Automatic summarizing: Factors and directions,” in *Advances in Automatic Text Summarization* (I. Mani and M. Maybury, eds.), pp. 1–13, Cambridge, MA: MIT Press, 1999.
3. E. Hovy, “Text summarization,” in *The Oxford Handbook of Computational Linguistics* (R. Mitkov, ed.), pp. 583–598, New York, NY: Oxford University Press, 2003.
4. C. M. Taskiran, A. Amir, D. Poncelon, and E. J. Delp, “Automated video summarization using speech transcripts,” *Proceedings of SPIE Conference on Storage and Retrieval for Media Databases 2002*, vol. 4676, 20–25 January 2002, San Jose, CA, pp. 371–382.

Table 4. Answer coverage of algorithms.

	FREQ	RAND	DEFT
Seahorse	5	2	1
MarkTwain	5	1	1
DogSmile	4	2	0

Table 5. Subjective summary assessment results.

	FREQ	RAND	DEFT
Question I median	3	3	2
Question II median	2	2	1

5. D. DeMenthon, V. Kobla, and D. Doerman, "Video summarization by curve simplification," *Proceedings of the ACM Multimedia Conference*, September 12-16 1998, Bristol, England, pp. 211–218.
6. A. Stefanidis, P. Partsinevelos, P. Agouris, and P. Doucette, "Summarizing video datasets in the spatiotemporal domain," *Proceedings of the International Workshop on Advanced Spatial Data Management (ASDM'2000)*, September 6-7 2000, Greenwich, England.
7. M. Cooper and J. Foote, "Summarizing video using non-negative similarity matrix factorization," *International Workshop on Multimedia Signal Processing*, December 9-11 2002, St. Thomas, US Virgin Islands.
8. L. Agnihotri, K. V. Devera, T. McGee, and N. Dimitrova, "Summarization of video programs based on closed captions," *Proceedings of SPIE Conference on Storage and Retrieval for Media Databases 2001*, vol. 4315, January 2001, San Jose, CA, pp. 599–607.
9. M. J. Pickering, L. Wong, and S. M. Rueger, "ANSES: Summarization of news video," *Proceedings of the International Conference on Image and Video Retrieval*, vol. LNCS 2728, July 24-25 2003, Urbana, IL, pp. 425–434.
10. H. D. Wactlar, "Informedia - search and summarization in the video medium," *Proceedings of the IMAGINA 2000 Conference*, January 31 - February 2 2000, Monaco.
11. B. Li, H. Pan, and I. Sezan, "A general framework for sports video summarization with its application to soccer," *Proceedings of IEEE International Conference on Acoustic, Speech and Signal Processing*, April 6-10 2003, Hong Kong, pp. 169–172.
12. R. Cabasson and A. Divakaran, "Automatic extraction of soccer video highlights using a combination of motion and audio features," *Proceedings of SPIE Conference on Storage and Retrieval for Media Databases 2003*, vol. 5021, 20-24 January 2003, Santa Clara, CA, pp. 272–276.
13. S. Uchihashi, J. Foote, A. Girgenson, and J. Boreczky, "Video Manga: Generating semantically meaningful video summaries," *Proceedings of ACM Multimedia '99*, October 30 - November 5 1999, Orlando, FL, pp. 383–392.
14. J. Oh and K. A. Hua, "An efficient technique for summarizing videos using visual contents," *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME'2000)*, July 30-August 2 2000, New York, NY.
15. R. Lienhart, "Dynamic video summarization of home video," *Proceedings of SPIE Conference on Storage and Retrieval for Media Databases 2000*, vol. 3972, January 2000, San Jose, CA, pp. 378–389.
16. L. He, E. Sanocki, A. Gupta, and J. Grudin, "Auto-summarization of audio-video presentations," *Proceedings of the 7th ACM International Multimedia Conference*, 30 October - 5 November 1999, Orlando, FL, pp. 489–498.
17. M. M. Yeung and B.-L. Yeo, "Video visualization for compact presentation and fast browsing of pictorial content," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 7, no. 5, pp. 771–785, October 1997.
18. H. Jing, R. Barzilay, K. McKeown, and M. Elhadad, "Summarization evaluation methods: Experiments and analysis," *Proceedings of the AAAI Symposium on Intelligent Summarization*, March 23 - 25 1998, Palo Alto, CA.

19. I. Mani, D. House, G. Klein, L. Hirschman, L. Obrst, T. Firmin, M. Chrzanowski, and B. Sundheim, "The TIPSTER SUMMAC text summarization evaluation," tech. rep., National Institute of Standards and Technology, October 1998.
20. R. L. Donaway and L. M. K.W. Drummey, "A comparison of rankings produced by summarization evaluation measures," *NAACL Workshop on Text Summarization*, 2000, Seattle, WA, pp. 69–78.
21. A. Ekin and A. M. Tekalp, "Automatic soccer video analysis and summarization," *Proceedings of SPIE Conference on Storage and Retrieval for Media Databases 2003*, vol. 5021, 20-24 January 2003 2003, Santa Clara, CA, pp. 339–350.
22. A. M. Ferman and A. M. Tekalp, "Two-stage hierarchical video summary extraction to match low-level user browsing preferences," *IEEE Transactions on Multimedia*, vol. 5, no. 2, pp. 244–256, 2003.
23. M. G. Christel, M. A. Smith, R. Taylor, and D. B. Winker, "Evolving video skims into useful multimedia abstractions," *Proceedings of the ACM Computer-Human Interface Conference (CHI'98)*, April 18-23 1998, Los Angeles, CA, pp. 171–178.
24. C. M. Taskiran, Z. Pizlo, A. Amir, D. Ponceleon, and E. J. Delp, "Automated video program summarization using speech transcripts," *IEEE Transactions on Multimedia*. to appear in 2006.