Automatic and User-Centric Approaches to Video Summary Evaluation

Cuneyt M. Taskiran and Frank Bentley

Motorola Labs, Applications Research Center
1295 Algonquin Road, Schaumburg, IL 60196
{cuneyt.taskiran, f.bentley}@motorola.com

ABSTRACT

Automatic video summarization has become an active research topic in content-based video processing. However, not much emphasis has been placed on developing rigorous summary evaluation methods and developing summarization systems based on a clear understanding of user needs, obtained through user centered design. In this paper we address these two topics and propose an automatic video summary evaluation algorithm adapted from teh text summarization domain.

Keywords: automatic video summary evaluation, user centered design

1. INTRODUCTION

Thanks to the proliferation of personal video recording devices, such as hand-held cameras, webcams, and video enabled cell phones, users today are able to easily generate video content. This has led to the creation of large Internet video repositories that are growing at a rapid rate, e.g., 65,000 video clips were being added daily to the popular Internet video site YouTube in August 2006. On the other hand, developments in the delivery of television programming, such as digital video recorders, bundled video and data services, and the advent of Internet Protocol television (IPTV), have started to radically alter traditional television viewing patterns. Deriving compact representations of video sequences that are intuitive for average users and let them efficiently browse large collections of video data has become more important than ever.

Due to the factors outlined above, automatic video summarization has become an active research topic in content-based video processing and various summarization algorithms and summary visualization schemes have been proposed. However, the majority of work done on video summarization has two important issues: first, there is no commonly used video data collection to train and test proposed summarization algorithms; second, generally little attention is given to developing rigorous summary evaluation methodologies. Without these two components, not only it is hard to compare the performance of different algorithms, but it also becomes difficult to make statements about the quality of the summaries produced with respect to human judgment. Recently more researchers have started to address these problems in video summary evaluation.

We believe that are two important issues that need to be addressed in the evaluation of automatically generated summaries:

- *Automatic summary evaluation.* Without well-defined methods to automatically evaluate summary quality, improvement of summarization algorithms becomes a cumbersome and bias-prone process, since each time the output needs to be evaluated by human judges.
- *User-centric design.* The generated video summaries are meant to be consumed by users. Therefore, summarization systems designed without a clear understanding of what users need and how they interact with summaries will have a low probability of adoption in practical applications.

These issues are concerned with approaches that attack the problem of summary evaluation from opposite ends of the spectrum. The development of practical video summarization systems strongly depends both on the rapid development and comparison of algorithms based on automatic evaluation methods as well as a clear understanding of what users want from such a system.

Much work has been done on developing methods for evaluation of text summaries and evaluating the effectiveness of these methods. Although some of the results obtained are unique to the text domain, many methods and findings are directly applicable to video summarization; however, researchers in video summarization field may not be familiar with this work. Therefore, we discuss some results from text summary evaluation work that are applicable to the evaluation of video summaries.

The paper is organized as follows: In Section 2 we provide a brief overview of the main approaches used in previous work on video summary evaluation. We provide motivations for development of automatic summary evaluation methods and briefly discuss parallel ideas from text summarization in Section 3 Then, in Section 4, we present an algorithm adapted from the text summarization domain to automatically calculate a goodness score for a summary, based on a set of reference summaries created by human experts. We address the issue of user-centric design in Section 5. Finally, a framework for future work based on the ideas presented is given in Section 6.

2. PREVIOUS APPROACHES TO VIDEO SUMMARY EVALUATION

It is beyond the scope of this paper to provide a comprehensive review of all previous work in summary evaluation, our aim is rather to provide an outline of key approaches that were used. More complete surveys are available elsewhere.[1, 2]

We borrow terminology from the text summarization domain[3] and classify previous video summary evaluation methods into two categories: *intrinsic* and *extrinsic*. In intrinsic evaluation methods the quality of the generated summaries is judged directly based on the analysis of summary, based on criteria such as fluency and coverage of key ideas, while in extrinsic methods the summary is evaluated with respect to its impact on the performance for a specific information retrieval task.

2.1. Intrinsic Evaluation Methods

Summaries created by human experts are often used in intrinsic evaluation of text summaries. A similar approach is rarely used in evaluating video summaries, since creating video summaries is more ambiguous and costly. For video content where events that are interesting to users are unambiguous, e.g., goals, touchdowns, summaries may be judged on their coverage of these events.[4]

Another commonly used intrinsic evaluation method is to use Likert scale questionnaires where users rate their level of agreement with statements about summaries such as "I found the summary to be clear and easy to understand" or "I feel that I can skip watching the whole program because I watched this summary" .[1, 5, 6]

2.2. Extrinsic Evaluation Methods

Various extrinsic evaluation methods have been proposed to evaluate summary quality and summary visualization schemes. A common approach is the quiz method, where a summary's coverage of key ideas is tested using a multiple choice quiz derived from the full-length video.[1, 5]

Extensive extrinsic evaluation studies were performed over the years within the scope of the Informedia Project[2] including fact-finding, where users utilized video summaries to locate video segments that answered specific questions; and gisting, where users matched video summaries with representative text phrases and frames extracted from source video.

2.3. Problems with Current Video Summary Evaluation Approaches

There are some problems associated with most approaches used in the evaluation of video summaries. The precision and recall values used as an intrinsic evaluation method provide a quantitative summary goodness measure; however, it is not clear if these values correlate strongly with summary quality judgment of users. Although simple and widely used, the questionnaire method has several drawbacks. Studies comparing the informativeness of these methods with other extrinsic evaluation methods have found that subjective assessments do not correlate strongly with users' performance on information retrieval tasks.[1, 7] Furthermore, as will be discussed in depth in Section 5, rating the "usefulness" of a summary is an ill-defined task which makes the results hard to interpret.

Extrinsic methods have their own problems. The level of the information retrieval task to be used is difficult to choose: it is easier to attribute the differences of user performance to summary quality if a low-level task is used but these bear little resemblance to real-world tasks. On the other hand, if complex tasks mimicking real-world situations are used, it becomes harder to isolate the effect of summary quality from other factors.[2] For the commonly used quiz method it is not clear how to derive the set of questions used since the list of important points in a video program will depend on the person preparing the list.

3. MOTIVATIONS FOR AUTOMATIC SUMMARY EVALUATION

Automated text summarization dates back at least to Luhn's work at IBM in the 1950s,[8] which makes it the most mature area of media summarization. A number of government sponsored text summarization and evaluation efforts, such as the Text Summarization Evaluation (SUMMAC)[3] and Document Understanding Conferences (DUC),[9] has enabled researchers in this field to conduct large-scale experiments and compare results on the same data collection. We believe that most results obtained during these extensive testing and evaluation efforts in the text summarization domain can be applied to the problem of video summary evaluation. Since researchers in this field may not be familiar with the work on text summarization, in this section we present a brief overview of the main evaluation approaches and key findings.

Automatically generated text summaries are usually evaluated through comparison with summaries produced by human experts. Traditionally this evaluation was performed manually through human judgment to compare an automatically produced summary with a single reference summary that was generated by a human expert (e.g., all summary evaluation for the Document Understanding Conferences in 2001–2003 was done using this approach[9]). In the manual method of summary evaluation a human expert performs a binary comparison between an automatically generated summary and a reference summary created by an human expert. There are many problems associated with this approach, some of which are listed below.

- *Using a single reference summary is not adequate.* Reference summaries produced by equally skilled experts have some overlap in content but it is highly likely that they will also have content that is unique to each summary. In one study it was found that agreement among summaries created by experts was low (40%) when summarizing a single document and was even lower (29%) when summarizing multiple documents on the same topic.[10] Other studies have confirmed that only a few of the key ideas from the full text is included in each reference summary.[11, 12]

- *Human judges have difficulty assessing overlap between reference summaries.* If one tries to remedy the above problem by having judges consider multiple reference summaries another problem emerges, the instability of human judgments of "information overlap".[10] Overlap between different text summaries is not a binary relation and is hard to judge. Simple precision–recall type measures are not adequate to reflect the overlap between a given summary and a set of reference summaries.

- *Manual evaluation is costly and time-intensive.* Obtaining evaluations from human experts is expensive and can take weeks to complete. This is an important problem since in the development of new algorithms there is a need to monitor the results produced in a quick and inexpensive way, so that bad ideas can be eliminated and algorithm parameters can be fine tuned.

Due to the shortcomings of the manual evaluation approach listed above, development of automatic methods for evaluating the quality of automatically generated text summaries has become an area of active research.[13] The strong need for developing automatic evaluation methods is not limited to the text summarization area, it is an important problem in all areas of language technology where text is produced automatically. It is a common observation that, when a language technology develops a standard text collection on which to develop its techniques and an automated methods of evaluating results to derive development, progress tends to speed up dramatically.[14] For example, the recent development of the BLEU score,[15] which correlates well with human rankings, has been important in the machine translation area.

Methods that have been proposed for automatic summary evaluation can be divided into two categories based on the granularity considered:

- *Word based methods.* These methods are based on the observation that if two summaries are semantically close they are likely to share many words and phrases. The notion of sharing is quantified by comparing n-grams of a candidate summary with the n-grams of the reference summaries and counting the number of matches. The ROUGE measure[16] for automatic text summary evaluation uses this approach as well as the BLEU measure[15] for evaluation of machine translation systems.

- *Text segment based methods.* In reference summaries alternate phrasings for the same semantic content is commonly encountered, which may reduce the usefulness of word based similarity measures. Methods comparing the similarity of sub-sentence text units of varying lengths has been produced to alleviate this problem. These include the Pyramid method[12] and comparison using Basic Elements.[14]

Many of the above ideas are also applicable to video summary evaluation. One important difference between text and video is that language has a discrete and syntactic nature that makes it different from video, although there is the film grammar approach in film theory that tries to find parallels between the two. Due to this difference in nature, evaluation methods based on text n-grams have no clear counterpart in video summary evaluation. However, we believe that evaluation methods based on text segments can be readily adapted to the video summary domain.

4. THE PYRAMID ALGORITHM FOR VIDEO SUMMARY EVALUATION

In this section we describe an algorithm to calculate evaluative scores for automatically generated video summaries based on a set of reference summaries created by human experts. The algorithm described here is an adaptation of a method proposed to evaluate text summaries.[12, 13]

In automatic video summarization systems the first step is generally the segmentation of the video sequence to be summarized into temporal components. However, the granularity of the components to be used for summaries varies between different systems that have been proposed. In our algorithm we use video segments consisting of contiguous video frames as summary components, rather than video shots. Using segments rather than shots as summary components provides several advantages:

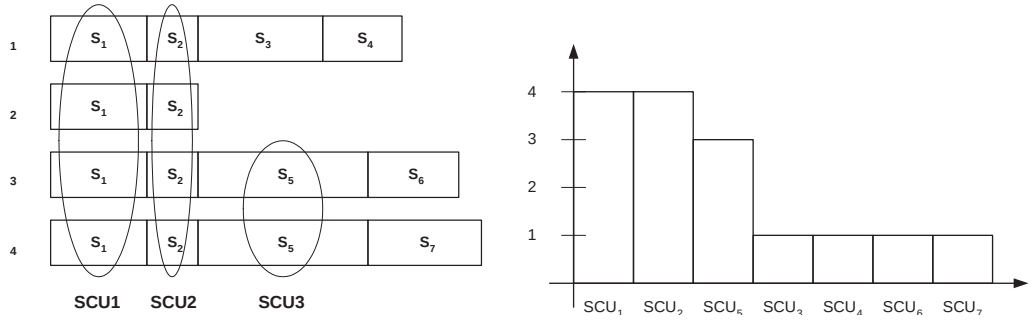
- Shot boundaries in video may not be aligned with audio boundaries, that is, a shot boundary may occur in the middle of a sentence in the audio. In previous video summary evaluation user studies it was found that users find this to be annoying and distracting.[5]

- Some video summarization algorithms[17, 18] do not perform shot boundary detection as a preliminary step, but rather use a bottom-up approach where video frames are clustered. Using video segments rather than shots makes our evaluation algorithm also applicable to these methods.

- Certain video content, such as presentations and meeting videos, contains long shots, making it hard to use shots as summary components fro such content. Long video shots also occur as establishing shots in programs as diverse as documentaries and talk shows.

In our approach a video summary, Σ, is represented as a time ordered sequence of n video segments, that is, $\Sigma = \{S_1, \ldots, S_n\}$. We assume that a *reference set* of L video summaries created by human experts are available, $\mathcal{R} = \{\Sigma^1_{\text{ref}}, \ldots, \Sigma^L_{\text{ref}}\}$.

The automatic summary evaluation problem may then be stated as follows: Given a video summary Σ, derived from the same content as the reference set $\mathcal{R}$, calculate a score for Σ that reflects its quality as compared to the summaries in $\mathcal{R}$. Note that this goodness score gives little information when considered in isolation; it is better utilized as a tool to rank different summaries. The goal is to produce a goodness score that ranks summaries close to the ranking produced by human judges.



(a) SCU clusters in the reference set

(b) SCU frequencies

Figure 1. An example illustrating how SCU frequencies are calculated for a set of four reference summaries.

4.1. The Pyramid Video Summary Evaluation Algorithm

The proposed algorithm consists of the following steps which are described in detail below.

Preliminary Identify *summary content units* (SCUs, defined below) in $\mathcal{R}$. Then, assign a weight to each unique SCU found based on its frequency in $\mathcal{R}$. This step needs to performed only once for a given $\mathcal{R}$.

Match to Reference For each possible video segment in Σ, find the most similar SCU in $\mathcal{R}$.

Select Covering From the set of candidate video segments found in the previous step, derive a covering set of segments for Σ that has maximum overall similarity with $\mathcal{R}$.

Calculate Score Based on the covering, calculate a goodness score for Σ using the weights of chosen SCUs.

4.1.1. Calculating Reference Summary Content Unit Frequencies

In this step the reference summaries in $\mathcal{R}$ are analyzed to identify clusters of video segments that correspond to approximately the same content, which are referred to as summary content units (SCUs).[13]

Figure 1 provides a simple example where there are four reference summaries. After the segments that cover similar content are identified, it is found that segments $\{S_{1,1}, S_{2,1}, S_{3,1}, S_{4,1}\}$ in these summaries approximately correspond to the same content, so this cluster of segments is labeled as SCU_1 (the match is approximate since the segments will not span identical durations of the full-length video). Similarly, segments $\{S_{1,2}, S_{2,2}, S_{3,2}, S_{4,2}\}$ are clustered to form SCU_2 and segments $\{S_{3,3}, S_{4,3}\}$ are clustered to form SCU_3. The remaining segments are not shared across reference summaries, corresponding to singleton SCUs. Based on this analysis, SCUs are assigned weights which are proportional to their frequency counts in $\mathcal{R}$, as shown in the histogram in Figure 1. The distribution of the SCU frequencies generally will follow a power law, since there will be a few SCUs with high frequencies (i.e. shared across many reference summaries) and a large number of SCUs with a frequency of one or two (i.e., unique to one or two reference summaries).

Since the rest of the evaluation process depends on the results of this step, the identification of the SCUs have to be done with high accuracy, which requires some manual intervention in analysis of the reference summaries. However, this analysis step needs to be performed only once for a data set.

4.1.2. Matching Candidate Segments to Reference SCUs

The evaluation process for a summary Σ starts by matching each video segment S in Σ with a SCU in $\mathcal{R}$ with which it shares the most content. Since each SCU represents a cluster of video segments, this step is equivalent to merging S to the closest cluster. Any one of the many methods proposed to compare the similarity of video sequences to can be used in calculating the similarity of S with the segments in the SCU being considered.

4.1.3. Deriving a Covering for the Candidate Summary

After all the video segments in the summary, Σ, to be scored have been assigned to an SCU, the similarity scores are used to derive a set of disjoint segments that covers Σ and maximizes the overall similarity score. If the summary contains n segments there are $O(2^n)$possible such sets, so dynamic programming is used to find the set that optimizes the similarity.

4.1.4. Calculating Score for the Candidate Summary

Finally, each selected score is assigned the frequency score of the corresponding SCU and these are summed to calculate the total score for Σ. This raw score is normalized by the maximum score possible for an "ideal" summary, which contains as many high-weight SCUs as possible in a summary of the same size as Σ.

4.2. Comparison with Previous Automatic Video Summary Evaluation Algorithms

The authors are aware of only two other methods that were proposed to automatically evaluate video summaries. The method proposed by Yahiaoui et al.[19] for multiepisode video summary evaluation is based on the following simulated user experiment: after watching all summaries, the user is shown a randomly chosen excerpt from one of the full-length videos and is asked to guess which video the excerpt was extracted from. They then develop a model of summary coverage to quantify summary quality. One drawback of this approach is that the summarization algorithm is tied to the evaluation method, reducing the applicability of the evaluation method to diverse summarization methods. Also, as pointed out before, it is unclear how strongly the coverage score correlates with human judgment of summary quality.

The SUPERSIEV automatic video evaluation system proposed by Huang et al.[20] is similar to our approach in that they also adapt ideas from text summarization domain and evaluate summaries based on a mainstream summary synthesized from a set of references summaries, similar to the idea of SCUs that we described. After this mainstream summary is constructed, generated summaries are evaluated by single frame matching and sequence alignment. The synthesis of the mainstream summary is performed automatically through k-means clustering of frames. Compared with the manual creation of the SCUs in our method, automatically creating the mainstream summary seems more desirable; however, the summary so produced will be much less reliable (note that some SCUs may span multiple shots, these will be broken into multiple clusters when clustering is used). The quality of the mainstream summary is crucial for the accuracy of summary evaluation and such effects may decrease the accuracy.

5. APPLYING USER-CENTERED METHODS TO VIDEO SUMMARIZATION EVALUATION

In this section we change the focus from automatic summary evaluation to evaluation strategies trying to determine what user information needs video summaries fulfill. While previous approaches to evaluating video summarizations have helped researchers improve their algorithms and optimize for given scenarios, we are unaware of a study employing the methods of user-centered design (UCD)[21] to understand the larger questions of why people want video summarized and their needs for summarizations in different contexts. These methods have been demonstrated to be instrumental for learning how to design for actual use as well as to evaluate systems across domains as diverse as industrial design,[22] user interface design,[23] and healthcare.[24] The typical UCD lifecycle starts with observing users, creating requirements for design, designing and building a system, and then evaluating it by observing users interacting with the system in real-world tasks (as seen in Figure 2). This is truly a cycle, and iterative design and evaluation has been shown to lead to applications that better meet user needs.

We believe that the field should take a step back from algorithmic evaluation or "usefulness" questionnaires and take a look at the tasks in people's everyday lives that cause frustration or can be helped by summarization techniques. By building systems to help with these specific tasks, the usefulness of a given summary can be analyzed with respect to the actual situation in which it would be used, with the people who would actually be using it. This rich analysis can shed insights onto parts of summarization algorithms that could be improved as well as how algorithms work in different domains with users who have different purposes for viewing the summary. By making the evaluations as realistic as possible, researchers can gain better understandings of how summaries would be used in real life situations.

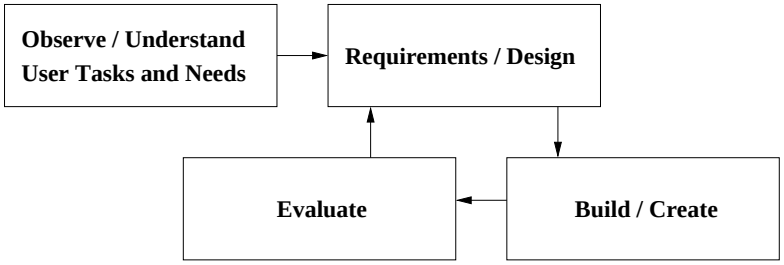


Figure 2. A typical example of a full user-centered design process. User observation and task analysis lead into a cycle of iterative design, development, and evaluation.

5.1. Understanding User Needs

In following the UCD process, it is important to first understand the needs of users in the real world and the tasks that they carry out. This is often obtained from following methods such as Contextual Inquiry[25] or through rapid ethnographic-style inquiry.[26] These methods allow researchers to understand user needs and breakdowns in current systems as well as the tasks and interactions currently required to achieve a certain goal. By performing studies in several domains, researchers can develop requirements for systems in different environments and develop an understanding of potential users and their needs. Perhaps a person on a train who is interested in politics has different needs for summarizations than a person sitting in their living room deciding if it is worthwhile to purchase an on-demand movie. Perhaps there are similarities in both tasks that could be discovered to build better all-purpose summarizers.

These needs can be understood by direct observation of uses in their environments or through cultural probe analysis.[27] Cultural probes allow researchers to see the environments in which people interact and get closer to a person's life than most researchers could in person. Photo diaries, post cards, scrapbooking, and other physical items could be used to see the situations in which people want to see video clips and time diaries can help document the amounts of time that people have available for watching content.

5.2. Conducting Evaluation

Once a set of requirements are created, researchers can choose from the many video summarization methods available (or create a new one to meet unforeseen needs) and use any of a variety of methods to evaluate the performance of the chosen summarization algorithm(s). More than simply asking for a Likert-scale opinion of things like "usefulness," these methods allow researchers to see how summarization technologies fit into everyday lives and meet the real needs of users.

For each particular task identified through earlier parts of the research (e.g. "getting the news before my train gets home" or "finding out if I'd be interested in watching this given documentary") it is important to find users who would actually perform these tasks in real life and to choose content that is interesting and relevant to them. Participants not interested in the material or who would never perform a given task cannot provide accurate evaluation of the usefulness or desirability of the summary for that scenario. If possible, these tasks should be performed in the actual environment that the participant would be using the final application so that any environmental effects could be taken into consideration.

After viewing the summaries, researchers should ask participants about how well the summary helped them with the task that was trying to be accomplished. Additional questions should help researchers answer their specific research questions and develop a qualitative understanding of what worked and what did not in the summary. Perhaps the participant would want more talking scenes or perhaps the few plays leading up to a goal in a soccer match are as important as the goal itself. Simple measures such as "useful" cannot capture these rich attributes that can come out in an interview after direct use of the system.

This process of analyzing summarization techniques based on expected use and real-world tasks will help ground analysis in the reality of everyday users and situations. Concepts such as precision and recall might not be important to a user whose ultimate goal is to watch the entire video. Likewise if a participant knows that

there will be a quiz at the end of the video, they will likely watch it differently from the way in which they would in a real situation. It is important to match the questions and method for evaluation to the task being accomplished.

Following these principles it is possible to gain a deeper understanding into why videos should be summarized and how they will be used in the real world. While space does not allow us to go into the detail of any of these methods, we hope that readers will attend to the references for full explanations of these methods and examples on how they have been successfully applied. Likely, this will lead to insights that can help develop new approaches for summarization and will better meet user needs.

6. FRAMEWORK FOR FUTURE RESEARCH IN VIDEO SUMMARY EVALUATION

Although there has been an increasing focus in video summarization work on evaluating the quality of generated summaries, the area lacks large-scale experiments, data sets, and objective methods of evaluation for summarization algorithm comparison that are available in the text summarization domain. Below, we outline a framework that we believe can guide future research on video summary evaluation.

- *Generation of a common summary evaluation data set.* The TRECVID evaluation has a large collection of video content from diverse sources. It has already been suggested that this collection can be used for summarization work.[2] Especially the so-called BBC rushes, which contain unedited video footage, can serve as an ideal evaluation test data. However, reference summaries need to be created for the data collection, which can be done collaboratively within the TRECVID community.

- *Development of automatic summary evaluation measures.* As explained before, fast, metrics-based comparison between algorithms can significantly increase rate of innovation in the field, such increases have been observed in the areas of text summarization and machine translation when automatic evaluation methods that correlate well with human rankings were introduced.

- *Having large-scale evaluation workshops.* Specialized evaluation workshops similar to DUC or TRECVID energize the community and provide a platform where ideas may be compared and shared quickly. A summarization focused track can be added to the TRECVID evaluation or it can held as an independent workshop.

- *Increased emphasis on user-centric methods.* More so than text summaries, video summaries are meant to be consumed by the average user on various consumer electronics devices, e.g., TVs, PCs, cell phones. For the proposed summarization systems to be adopted by users they should be based on a clear understanding of user needs and interaction patterns.

7. CONCLUSIONS

In this paper we have addressed two important issues related to the evaluation of video summaries, namely deriving automatic methods fro summary evaluation and developing summarization systems based on user centered design.

REFERENCES

1. C. M. Taskiran, Z. Pizlo, A. Amir, D. Ponceleon, and E. J. Delp, "Automated video program summarization using speech transcripts," *IEEE Transactions on Multimedia*, vol. 8, no. 4, pp. 775–791, August 2006.
2. M. G. Christel, "Evaluation and user studies with respect to video summarization and browsing," *Proceedings of the IS&T/SPIE Conference on Multimedia Content Analysis, Management, and Retrieval*, 17–19 January 2006, San Jose, CA.
3. 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.

4. 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.
5. 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.
6. H. Sundaram, L. Xie, and S.-F. Chang, "A utility framework for the automatic generation of audio-visual skims," *Proceedings of the ACM Multimedia Conference*, 1–6 December 2002, Juan Les Pins, France, pp. 189–198.
7. 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.
8. P. H. Luhn, "Automatic creation of literature abstracts," *IBM Journal*, vol. 2, no. 2, pp. 159–165, 1958.
9. "http://duc.nist.gov/."
10. C.-Y. Lin and E. Hovy, "Manual and automatic evaluation of summaries," *Proceedings of the Workshop on Automatic Summarization, ACL 2002*, 11–12 July 2002, Philadelphia, PA.
11. H. Halteren and S. Teufel, "Examining the consensus between human summaries: Initial experiments with the factoid method," *Proceedings of the HLT-NAACL Text Summarization Workshop*, June 2003, Edmonton, Canada.
12. A. Nenkova and R. Passonneau, "Evaluating content selection in summarization: The Pyramid Method," *Proceedings of the Human Language Technology conference (HLT/NAACL 2004)*, 2–7 May 2004, Boston, MA.
13. A. Harnly, A. Nenkova, R. Passonneau, and O. Rambow, "Automation of summary evaluation by the pyramid method," *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-2005)*, 21–23 September 2005, Borovets, Bulgaria.
14. E. Hovy, C.-Y. Lin, and L. Zhou, "Evaluating DUC 2005 using basic elements," *Proceedings of the Document Understanding Conference (DUC 2005)*, 9–10 October 2005, Vancouver, Canada.
15. K. Papieni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, July 2002, Philadelphia, PA, pp. 311–318.
16. C.-Y. Lin and E. Hovy, "Automatic evaluation of summaries using N-gram co-occurrence statistics," *Proceedings of the HLT-NAACL Text Summarization Workshop*, June 2003, Edmonton, Canada.
17. 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.
18. Y. Gong and X. Liu, "Video summarization using singular value decomposition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 13-15 June 2000, pp. 174–180.
19. I. Yahiaoui, B. Merialdo, and B. Huet, "Comparison of multi-episode video summarization algorithms," *EURASIP Journal on Applied Signal Processing*, vol. 3, no. 1, pp. 48–55, 2003.
20. M. Huang, A. B. Mahajan, and D. F. DeMenthon, "Automatic performance evaluation for video summarization," Tech. Rep. LAMP-TR-114,CAR-TR-998,CS-TR-4605,UMIACS-TR-2004-47, University of Maryland, College Park, June 2004.
21. D. A. Norman and S. W. Draper, *User Centered System Design; New Perspectives on Human-Computer Interaction*. Mahwah, NJ, USA: Lawrence Erlbaum Associates, Inc., 1986.
22. T. P. Tom Kelley, Jonathan Littman, *The Art of Innovation: Lessons in Creativity from IDEO*. Doubleday, 2001.
23. A. W. Ellen Isaacs, *Designing from Both Sides of the Screen: How Designers and Engineers Can Collaborate to Build Cooperative Technology*. Sams, 2001.
24. C. Burns and F. Dust, *Extra Spatial*. San Francisco, CA, USA: Chronicle Books, 2003.
25. H. Beyer and K. Holtzblatt, "Contextual design," *interactions*, vol. 6, no. 1, pp. 32–42, 1999.
26. J. Beebe, *Rapid Assessment Process: An Introduction*. Altamira Press, 2001.
27. B. Gaver, T. Dunne, and E. Pacenti, "Design: Cultural probes," *interactions*, vol. 6, no. 1, pp. 21–29, 1999.