MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Subjective Assessment of Consumer Video Summarization

Clifton Forlines, Kadir A. Peker, Ajay Diivakaran

TR2006-002 January 2006

Abstract

The immediate availability of a vast amount of multimedia content has created a growing need for improvements in the field of content analysis and summarization. While researchers have been rapidly making contributions and improvements to the field, we must never forget that content analysis and summarization themselves are not the user's goals. Users' primary interests fall into one of two categories; they normally either want to be entertained or want to be informed (or both). Summarization is therefore just another tool for improving the entertainment value or the information gathering value of the video watching experience. In this paper, we first explore the relationship between the viewer, the interface, and the summarization alborithms. Through an understanding of the user's goals and concerns, we present means for measuring the success summarization tools. Guidelines for the successful use of summarization in consumer video devices are also discussed.

SPIE Conference Multimedia Content Analysis, management and Retrieval

Subjective Assessment of Consumer Video Summarization

Clifton Forlines, Kadir A. Peker, Ajay Divakaran,
Mitsubishi Electric Research Laboratories, Cambridge, MA 02139

ABSTRACT

The immediate availability of a vast amount of multimedia content has created a growing need for improvements in the field of content analysis and summarization. While researchers have been rapidly making contributions and improvements to the field, we must never forget that content analysis and summarization themselves are not the user's goals. Users' primary interests fall into one of two categories; they normally either want to be entertained or want to be informed (or both). Summarization is therefore just another tool for improving the entertainment value or the information gathering value of the video watching experience. In this paper, we first explore the relationship between the viewer, the interface, and the summarization algorithms. Through an understanding of the user's goals and concerns, we present means for measuring the success summarization tools. Guidelines for the successful use of summarization in consumer video devices are also discussed.

Keywords: Video Summarization, Subjective Assessment of Summarization

1. INTRODUCTION

Prior work on video summarization has focused more on the generation of summaries and less on the assessment of user satisfaction with summaries. Many of these video summarization tools require large CPU resources for analysis, high-resolution screens for displaying the results, and rich input devices for manipulating and exploring the summaries. Our focus is on consumer video devices such as the Personal Video Recorder. The end-user or consumer has a variety of goals while watching content, which are enabled by the combination of the summary generating tool(s) and the user-interface. We see video summarization as an unobtrusive aid to content navigation/traversal. We therefore frame the problem of identifying and applying success criteria to video summarization as identification of criteria for user goal satisfaction. Thus we focus on video summarization as providing a segmentation of the content that enables the consumer to satisfy goals such as getting a digest of the content, skipping over certain segments, content selection, quick review of what has been watched so far etc.

The rest of this paper is organized as follows. In section 2, we briefly describe some related work in the field of summarization. In section 3 we define the problem that we are attempting to address, noting that content analysis and summarization themselves are *not* the user's goals but rather tools that can enable the user to complete their goals. In section 4 we address the user's concerns that have arisen with recent changes to consumer video availability, and describe why the successful use of summarization is badly needed and how consumers are likely to reward the designers of systems that address these concerns. In section 5 we address user goals and sub-goals and enumerate many of the jobs that current interfaces are not fully getting done. In section 6 we address the role of the presentation/interface, and in section 7 we list some of the specific considerations that need to be taken into account when working with summarization on consumer devices. In section 8 we cover some guidelines for the use of summarization in an interface. In section 9 we describe an example system that uses summarization to address some goals of the user, and in section 10 we describe how we plan to evaluate this and other systems. Finally, we conclude with a look toward the future in section 11.

2. RELATED WORK

The goals of multimedia content summarization are two-fold. One is to capture the essence of the content in a succinct manner and the other is to provide top-down access into the content for browsing. Towards achieving these goals, signal processing & statistical learning tools are used to generate a suitable representation for the content using which summaries can be created. For content that is carefully produced & edited (scripted content) such as news, movie,

drama, etc., a structure is imposed during the production of the content in terms of semantic units such as news stories, scenes etc. Therefore, a representation that captures the sequence of semantic units that constitute the content would be useful. The user can browse the content using abstractions of each of the semantic units such as keyframes, skims etc. Hence, past work on summarization of scripted content has mainly focused on coming up with a Table of Contents (ToC) representation. In unscripted content such as sports & surveillance, interesting events happen sparsely in a background of usual events. Therefore, if the analysis is focused on detecting specific events of interest, a summary can be generated using a combination of what is typical in the content and what are "interesting" events in the content. For examples of approaches to video summarization see [].

3. DEFINING THE PROBLEM – WHAT TO MEASURE?

While researchers have been rapidly making contributions and improvements to the field, we must never forget that content analysis and summarization themselves are *not* the user's goals. Users' primary interests fall into one of two categories; they normally either want to be entertained or want to be informed (or both). Summarization is therefore just another tool for improving the entertainment value or the information gathering value of the video watching experience.

Figure 1 illustrates the relationship between the summarization algorithm, the user interface, and the user (goals and abilities). Much of the previous work in the field involves looking at the summarization algorithm portion of the diagram, and previous studies often compare different summarization methods outside of the context in which they are used. Evaluation of consumer devices must be driven from the user side of the diagram and take the specifics of the interface into account. The interface, which mediates the relation between the user and the summarization algorithms, is capable of turning am error ridden summarization into a useful tool for the viewer. Similarly, the wrong interface can take even the most sophisticated analysis and produce an unpleasant experience for the user. Only through an understanding of the goals of the user, will the designers of this type of system be able to determine the measures by which to compare different interface / algorithm combinations.

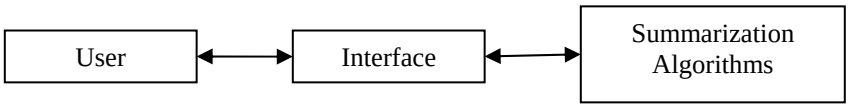

Figure 1: The relationship among the user, the interface, and the algorithms. The user never interacts with the algorithms themselves, so it is the role of the interface to mediate use of and correct for mistakes in the summarization. Evaluation of a summarization algorithm itself does little to predict its usefulness to the user.

4. USER CONCERNS – THE GRASS IS ALWAYS GREENER

Anxiety is not a word that one would normally associate with an entertainment experience; however, recent changes in television watching have put a lot of stress and pressure on the user. These changes stem from a common root – an increase in the programming choices available to consumers. While choice is normally a good thing, once one's choices become unmanageable, one often becomes paralyzed by choice. Three such changes are:

Growth in the number of available channels.

When viewers had three live network broadcasts to choose between, they could reasonably be expected to browse the current offerings and make an informed decision about what to watch. With hundreds of channels available, how does one choose (or even brose through) the offerings? Flipping through 500 channels and spending 3-4 seconds on each channel takes about 30 minutes, which means the programs one is browsing through have finished before one can make a choice about what to watch.

Time is no longer a constraint.

The effect of the growth in the number of available channels is amplified by the addition of time-shifting provided by personal video recorders. With a PVR, one not only has access to hundreds of live channels, but also has access (give a

little foresight) to anything previously broadcasted on these channels. If one lacks this foresight, many PVRs are helpful enough to automatically record those programs that one would have recorded had they only had more time.

Competing media.

We must not forget to look at consumer video devices in the context of the variety of activities and media that are competing for our user's time and attention. Video rental services, the Internet, satellite radio, books, newspapers and periodicals, podcasts, video games, etc. are all competing for consumer's leisure time. These media are experiencing similar growth patterns to television broadcasts.

People are rarely satisfied with their own situation and often err in thinking that others have it better than they do. This "the grass is always greener on the other side of the fence" mentality extends to entertainment experiences too. When given three network broadcasts to choose between, a consumer can be relatively confident that they chose "the best" program to watch, having browsed all the options and made an informed decision. Given hundreds and hundreds of available programs, how confident are people that they made the right choice in deciding what to watch? Do they feel that they're missing out on something? What if other people are having a better time than they are?

Unmanageable choice has the potential to lead to an anxiety ridden experience, a quality that conflicts with the primary goal of the television watching activity – to be entertained. The designers of interfaces and summarization methods that help relieve the anxiety and pressure brought on by choice will certainly be rewarded by their customers.

5. GOALS OF THE USER – WHAT JOBS NEED TO BE DONE?

In the introduction, we stated that the high-level goals of the user are to be entertained or to be informed, or both. In this section, we will discuss some of the sub-goals of these high-level goals and discuss how summarization could be used to address these goals. Recall that we stated at the outset that while video summarization is formulated as digest creation in the summarization generation literature, as per our discussion in section 3, the user goals are not limited to exactly that application. The metadata generation tools at our disposal are Summarization and Indexing, and both can be applied to satisfy a variety of consumer goals. These sub-goals generally fall into three categories, goals that arise during the *consumption of content*, goals that arise during the *selection of content*, and goals that arise during *content management.* In this section, we present the goals in terms of statements that the user might say to themselves.

It is important to note that the user's goals are often a mixture of the goals below, that these goals often overlap, ad that a user's goals may change over the course of a video watching session. Furthermore, there are often more than one user sharing a consumer video device, and they may have different goals while interacting with the device at the same time.

5.1 Content Management

This section, we enumerate some of the goals that arise during the management of recorded content. With a finite amount of space available in which to store recorded content, consumers need to manage this space and prioritize the recorded material in it.

"Did the system record what I wanted it to record?"

Knowing that the system is behaving in the way that one expects it to behave is an important step in trusting a device. While virtually all PVRs list the title and date of the recorded programs, this information may not be sufficient in answering the above question. Could a brief summary or overview of the program better address this question?

"Is it safe to delete this?"

Again, the title and date may not be enough to make a decision when it comes to program triage. An overview or summarization could help prioritize one program over another when deciding what to delete to free up space on the device.

5.2 Selection of Content

In this section, we enumerate some of the goals that arise during the selection of recorded content to consume.

"What programs are available?"

When beginning a television watching session, a user may have a specific program in mind, or may wish to browse through the available programs in order to pick something to watch. When the list of available programs becomes large, this goal becomes daunting. Summarization could aid in giving the user an overview of what programs are available to them, and greatly aid in answering the above question.

"Do I want to watch this more than I want to watch anything else that is available?"

When browsing through recorded content on a PVR, most interfaces present the title of the program, a brief text description, and some details about the network and recording date and time. This is the information that the viewer has available when they try to answer the question of what to watch. This complex task, which involves balancing the changing moods and time constraints of one or more people may benefit from summaries that will give a better understanding of the fit between the viewers' mood and the content. The goal should be to help the user feel they have found "the best" program available, not just on that is "good enough".

"Have I seen this before?"

Most PVR interfaces indicate whether or not a program has been previously viewed in the listing of recorded shows; however, this indication only tells if *this particular recording* of a program has been previously viewed. The system has no knowledge of viewing that may have taken place on another device, live during the broadcast, or even, in the case of reruns and revivals, months or years previously. The viewer will not always be able to answer the question "Have I seen this before?" from the information that is typically given. A summary that includes enough detail should help the user remember if they have viewed this content before.

"Do I want to watch all of this?"

Many programs naturally split into well defined sections, like the stories in a news cast, the acts on a variety show, and the guests on a talk-show. When selecting content, it would be very valuable to help the user decide if they want to watch the entire program, or if they plan to view only parts of it. Given an answer to the question above, the user could begin playback by selecting from among several different summarized playbacks given their level of interest.

5.3 Consumption of Content

In this section, we enumerate the goals that arise during the consumption of recorded content. While many decisions about the use of summarization may occur before playback begins, situational and mood changes may prompt the user's goals to shift during playback of recorded content.

"I want to move quickly through this part, but I'm still a little interested"

Fast-forwarding is a familiar summarization technique with which a user can gain an overview of a section of video without watching the entire portion in real-time. Much understanding can be gained through watching accelerated playback that would be lost were the user to instantly skip to the next section. With linear playback on VCRs, fast-forward was used for both skipping and accelerated playback. With random access DVRs, skipping by chapter or by commercial length increments leaves fast-forward for accelerated playback.

"I want to move directly to the next part"

Many programs are naturally thought of in segments, and the understanding of many of these segments do not rely upon having viewed earlier segments. For example, in a game show, each question asked of the participant is its own miniature show and can be watched and enjoyed without knowledge of the questions asked earlier. A viewer who is uninterested in a particular segment of such a show may want to skip to the next question, mini-game, etc. Other examples include browsing through the stages in a how-to program ("I know how to do this part, let's move to the part I don't know how to do."), browsing through acts in a variety show, or songs in a concert, browsing through the stories in a news program, browsing through the guests in a morning show, etc.

"I don't have time to watch all of this, just give me the good parts."

People often skim the stories in a newspaper, the articles in a magazine, and the text on web pages – why not recorded programs on a PVR? Skimming is not only used as a method for content selection, but also as a method of consumption when time and interest are in short supply. This skimming is different from accelerated playback in that some sections are skipped over in favor of others rather than the entire program being presented at an accelerated rate. By skipping over more or less sections of a video, a user can view the length of portions that correspond to their interest.

"Please fill this gap in my schedule"

Time-shifting (recording a live program at one time and then viewing it at another time) is popular in part because television broadcasts are changed to fit into the user's schedule, not the other way around. It's common to hear the purchaser of a PVR state that, "they'll never go back to watching TV the old way." When looked at this way, time-shifting allows television watching to fit into the gaps in people's lives [3], and time-shifting meets this goal, but only part way. Unfortunately, the gaps in one's life often do not divide neatly into 30 minute segments, which implies that summarization can help. Consumers will likely reward the designers of systems that can fill an 18 minute gap in one's day with the best 18 minutes of an hour-long program. It is this example that we explore in detail in a later section.

6. THE ROLE OF INTERFACE / PRESENTATION

In our view, the primary role of the user interface to summarization is to provide smooth recovery from mistakes made by the automatic summarization without overly burdening the viewer. This brings up the issue of whether a superb presentation coupled with a middling summarization algorithm will make the user happier than would a highly accurate summarization coupled with a middling interface. The answer likely lies with what job the user was trying to get done when they system designers thought a summarization would be helpful. We hope to resolve such questions with our proposed user studies.

Each screen in the interface needs to be designed with the goals of the user in mind. By knowing what questions the user might ask himself within each portion of the interface, designers can best use summarization to aid the fulfillment of these goals. A 60-second summary of a sporting event, which shows the most exciting moments, the game-winning play, and the final score, is an excellent solution to the goal in which the user says to herself, "I don't have time to watch all of this, just give me the good parts." This same summary is a terrible choice in the situation in which the user is asking, "Do I want to watch this?" In one case, the interface is using summarization to fulfill the goal of the user, which is to gain as much of the entertainment value from the game as they can in a short time span. In the other case, the interface has failed in that it has ruined the entertainment value of the program by giving away the ending.

7. SPECIAL CONSIDERATIONS FOR CONSUMER DEVICES

Integrating video summarization and browsing technologies into consumer electronic devices is an added challenge on top of the technical challenges of content analysis itself. A TV environment and the user's interaction with it are very different than a PC and the user's interaction with it. There is very limited interaction, it is mostly expected to be a "sit back and enjoy" mode of use. The remote control is a very limiting user input device compared to the mouse and keyboard. The menu system, the screen resolution, etc. are all limiting in the same way. The summarization and browsing technology has to be presented in a very simple, intuitive way. It shouldn't break the normal modes of operation the user is accustomed to. The target population is very varied and the common denominator leaves us with very little base assumptions. Furthermore, the TV application itself is expected to be simple and straightforward.

Additionally, consumer electronics platforms have lower computational power than the PCs. Many of the components are custom hardware, especially those for decoding the video etc. This makes adding new algorithms difficult. Cost constraints are very tight, any extra hardware is strongly avoided, especially for new and unproven technologies like video summarization. (For these reasons, although there has been a lot of research on summarizing video using visual features, only audio based solutions have made to the CE device. The visual pipeline is implemented in hardware and is a closed box. Any additional hardware component is extra cost that is avoided. However the audio is implemented on programmable DSP chips, where the software can be altered at no additional cost, to provide summarization functionality.) Although there are more and more PC-based appliances such as media centers, they are not quite mainstream yet. They, too, have computational constraints as there are many other demanding processes (anything relating to video) running on the system.

Many computational models that are taken granted on the PC platform (and during the algorithm development, since the researcher is very much accustomed to that model) do not hold true for consumer devices. For instance, the hard disk access is very different for AV (1394) discs compared to EIDE hard disks. Memory management and access, buffers, busses and associated constraints can be very different. Hence, there usually is a quite long way to the actual product even if the algorithm is mature on a different (e.g. PC) platform.

Operational issues also need to be carefully considered. At what stage and in what logic unit the summarization takes place depends on the whole operation of the system. Storage of and access to the extracted summaries, how much to preprocess and how much to dynamically compute is also a design problem that depends on the system and the usage scenarios.

8. USER INTERFACE DESIGN GUIDELINES

In this section, we describe some interface design guidelines that drive our development of the use of summarization in the user interface.

Guideline 1: Default to what the viewer is familiar with.
While the use has many goals, they rarely have the goal, “I want to spend time learning X.” Designers should be very careful not to break the television watching experience through the addition of features. If in doubt, default to what the viewer is experienced with using. Allow the user to gradually gain and understanding of the extra features through a controlled exposure to the power of the system.

Guideline 2: Allow the viewer to be as passive or as active as they want to be.
Television watching in most often meant to be a leisure-time activity. If the designers have excessive expectations for the role of the viewer in controlling the television watching experience, our design will almost certainly have a negative impact on their enjoyment. Television critics are quick to point out the remedial value of watching television; however, television watching is popular, in part, because of the extremely low level of effort that it requires. In general:

Payoff = Intrinsic Value of the Activity / Level of Effort

For every increase in the level of effort that our interface demands of the viewer, we must have an equal increase in the value of the activity.

People’s moods also change over time. The same viewer who wants to participate actively in the control of the summarization of a program on Saturday afternoon may want only to lay back and just watch that Saturday night. Designers should aim to support an active engagement in the interface, as well as a passive use of the new features, as well as the levels in-between.

Guideline 3: Design the interface to be forgiving of mistakes made by the summarization algorithm.
While many are impressive, few summarization algorithms are perfect. Designers should aim for interfaces that reduce the impact of errors made by the algorithms. A viewer needs to “trust” the recommendations made by the summarization UI, and this trust will never develop if errors have a highly negative impact on the viewing experience.

9. EXAMPLE SYSTEM – MELCO DVD RECORDER

Mitsubishi Electric's latest DVD Recorder (Raku-reko) provides a "highlight playback" capability. The highlights are extracted during the recording by analyzing the audio channel and looking for a characteristic mixture of cheering and the commentator's excited speech. Each second of the program is assigned an importance level based on the percentage of the "highlight audio" detected in a ten second window centered at the point of interest. The user interface consists of a plot of the importance level of each second of the program, as shown in figure 2. The interface enables the user to set an importance threshold so only the portions that exceed the threshold are played. The length of the summary corresponding to the choice of threshold is displayed as shown, so the user can choose a desired summary length by moving the threshold up or down as needed. In this case, the interface allows the user to easily recover from errors in the summarization because he can easily lengthen the summary if he misses an event. Furthermore, the user can toggle

between highlight and normal playback which enables seamless incorporation of highlight playback into existing video playback.

Figure 2: A screenshot from the Mitsubishi Electric "Raku-reko" DVD Recorder. This interface uses summarization during playback of a recorded sports program, allowing the user to view only the exciting portions of the program.

This interface addresses several user goals that arise during the consumption of content, as described in section 5. The viewer can fill any length gap in their schedule. By adjusting the threshold level (shown as the horizontal line in figure 2), the length of the summary can be set to fill the allotted time period. The upper right corner of the screen shows both the original length of the video and the length of the currently selected summarization. If the user feels that "I don't have time to watch all of this, just give me the good parts", then they can select an excitement threshold that matches the level of excitement they are interested in, ignoring the resulting length of the summary.

10. PROPOSED USER STUDIES

Recall our view that a good interface can compensate for a middling summarization. Past evaluations measure differences between systems as a whole. More controlled studies would look at variations in one component (such as different summarization algorithms, or different interfaces) at a time to control for confounding effects. Additionally, the means to compare variations in these components are difficult without a choice of what goal summarization is aiding in a specific portion of the UI. Recall the 60-second summary, complete with game winning play and final score, that makes for an excellent brief presentation of a program for a viewer who is on-the-run, but a horrible spoiling preview of an event for a viewer who is looking through their recorded content with the aim of picking something out to watch.

Our plan is to further investigate the goals that users have with the aim of identifying the jobs that current DVR devices are not getting done. These goals will be added to those presented in section 5. Preferably, this investigation will take place in the context of the users' television watching environment. With this list in hand, we will be in a position to investigate specific changes to the summarization algorithms and interface that support a specific goal. Because entertainment is a subjective activity, our measurements will be largely qualitative as well and will likely require repeated visits with individual users to build up a rhetoric and understanding of their impressions with the tools that we design.

11. DISCUSSION AND CONCLUSION

Looking forward, we have the following avenue for further research – first, look into using summarization in many parts of the interface to address a variety of user goals, second, systematically evaluate changes to the algorithms and interface in support of these goals. While prior work has focused mainly on the creation of summaries and

segmentations of video, and comparing these segmentations and summaries against one another as well as "ideal" segmentations and human generated summaries, we want to focus on the measurements to use when comparing summaries in terms of user satisfaction. The consumer has a complex and changing mix of goals when interacting with the device, and these goals must help us decide upon the measurements to use when comparing summarization techniques. Then, only through looking at the combination of the interface and the summarization can we make intelligent decisions that allow our customers to complete the jobs that they wish to get done. Consumers will likely reward those designers who can combine a summary with an interface that allows them to control that summary for the purpose of fulfilling a goal that is either difficult or impossible using current devices.

REFERENCES

1. Divakaran, K. A. Peker, R. Radhakrishnan, Z. Xiong and R. Cabasson, "Video Summarization using MPEG-7 Motion Activity and Audio Descriptors," Video Mining, eds. A. Rosenfeld, D. Doermann and D. DeMenthon, Kluwer Academic Publishers, 2003.

2. Lu, X., Ma, Y.-F., Zhang, H.-J. and Wu, L., An Integrated Correlation Measure for Semantic Video Segmentation. IEEE International Conference on Multimedia and Expo, ICME, (Lausanne, Switzerland, 2002).

3. Mark Stringer and Jessica Ward (2001). Who Wants to Play? Beginning Research into Mobile Gaming. Designing Ubiquitous Computing Games Workshop. http://www.viktoria.se/play/workshops/ubigame.ubicomp/papers/stringer.pdf (15 May. 2002).

4. Z. Xiong, R. Radhakrishnan, A. Divakaran, "Generation of Sports Highlights Using Motion Activity in Combination with a Common Audio Feature Classification Framework" to appear in International Conference on Image Processing, Barcelona, Spain 2003.

5. Yeung, M.M. and Yeo, B.L. Time-Constrained Clustering for Segmentation of Video into Story Units *Proceedings of the International Conference on Pattern Recognition* (ICPR '96) Volume III-Volume 7276 - Volume 7276, IEEE Computer Society, 1996.