Exploratory Video Search by Augmented Playback

Ullas Gargi, Somnath Banerjee
HP Laboratories Palo Alto
HPL-2006-155
October 31, 2006*

video browsing, video search

Exploratory search or browsing interfaces to digital video have traditionally followed the approach of navigating a pre-computed index based on content analysis, annotational metadata, or both. Examples include video storyboards, video skims, etc. We propose a new paradigm for exploratory video search based on augmenting video playback with metadata channels and flexible playback modes. Our method uses content analysis as well as optional user input to determine higher semantic importance segments of video and varies the playback rate of video accordingly. A past and future keyframe buffer as well as a fluid user control interface allow the user to keep track of the video while still moving through it at high speed. Preliminary results in this ongoing project are presented.

* Internal Accession Date Only
Approved for External Publication

Exploratory Video Search by Augmented Playback

Ullas Gargi
Hewlett-Packard Labs
1501 Page Mill Road
Palo Alto, California, USA
ullas.gargi@hp.com

Somnath Banerjee
Hewlett-Packard Labs
24 Salarpuria Arena, Hosur Main Road
Bangalore, India
somnath.banerjee@hp.com

ABSTRACT

Exploratory search or browsing interfaces to digital video have traditionally followed the approach of navigating a pre-computed index based on content analysis, annotational metadata, or both. Examples include video storyboards, video skims, etc. We propose a new paradigm for exploratory video search based on augmenting video playback with metadata channels and flexible playback modes. Our method uses content analysis as well as optional user input to determine higher semantic importance segments of video and varies the playback rate of video accordingly. A past and future keyframe buffer as well as a fluid user control interface allow the user to keep track of the video while still moving through it at high speed. Preliminary results in this ongoing project are presented.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Algorithms, Human Factors

Keywords

Video indexing, video browsing

1. INTRODUCTION

Efficient access to large digital video collections is a known and increasingly important information management problem. The media information retrieval community has traditionally solved this by video indexing—the analysis of video into scenes and shots, the determination of one or more representative keyframes for each shot and the creation of a storyboard, or visual index, of keyframes. Video browsing then involves the user navigating this storyboard to access the segment(s) of interest. Another approach has been the creation of video skims [2] which use multiple channels of information extracted from a video, such as keyframes, audio signal, closed–caption transcript, etc.

While these approaches have their merits, they have one large downside: they convert the experience of viewing video into a static one of navigating an index. The storyboard approach in particular is inflexible in that it assumes a perfect index is created. If the scene or shot segmentation is incorrect or the keyframe selection imperfect, there is no way for a user to be aware of it and recover from the system's error.

We propose an exploratory search mechanism for digital video that involves augmented fast playback. In Section 2 we describe the idea and its motivation. Section 3 covers the use of automatically–generated and optional user–supplied metadata channels. The control interface is described in Section 4. Our implementation and preliminary results of using our method on a small data set are presented in Section 5. Relevant prior work is cited in Section 6. A discussion and plans for future work are contained in the conclusion.

2. MOTIVATION AND OVERVIEW

Large digital video collections are increasingly being made accessible to people due to increased device storage, network transmission bandwidth capabilities and playback device processing capabilities. Some of this video is heavily annotated by professionals, for example video created in the entertainment and broadcast industries. An increasingly larger portion of video however is being created by consumers to preserve their personal memories, to share with others online or as a means of self-expression. These video files have less metadata but share the same problem of being a linear medium.

Standardization efforts such as MPEG-7 have created means to express video metadata, both content-based and otherwise, in a very general and flexible manner. Automatic methods to create the metadat are still lagging which is why the standards are not much used. However, our thesis is that even in the presence of rich metadata, such as either from professional annotation, content analysis or by collaborative tagging by a human community, the current state of the art in video search only results in the retrieval of one or more video segments for playback in response to a query. These video segments are still browsed, experienced or played back in the traditional way. There has been less work on how to use whatever metadata is available to improve the exploratory search or browsing experience with digital video.

This paper deals with how to use metadata for digital video to improve the exploratory search, or browsing, experience. Our intuition is that video playback can be augmented to make it akin to a visualization experience by using *metadata channels* that are created, both offline and online, in combination with realtime user control. We adopt a sequence of principles in designing this system:

- Fast browsing can beat search for medium or small-sized collections: our approach is not intended to scale to querying extremely large video stores. In those cases a pure query system makes sense. It is instead targeted

toward a person or family's personal video collection which is likely to be of the order of at most a few hundred hours.

- Instead of creating a detailed browsing index, rely on the human visual system's capacity to grasp the semantic content of images quickly and provide the user with fast tools for browsing based on extracted metadata.
- Any automatic method of extracting metadata from video is going to fail sometimes with both false positive and false negative errors.
- Give the user control and fallback modes so that algorithmic errors can be compensated for by the human if need be.

We assume that some measure of the *interestingness* of the video is available. This may be obtained by some combination of traditional video content analysis, including face detection, temporal segmentation, image frame quality measures and the like; user supplied metadata such as favorite points obtained either explicitly or transparently by user tracking; and collaborative filtering or social tagging where the explicit tagging or implicit playback behavior of a community enriches the video with another metadata channel.

Given such a measure, the playback rate is dynamically altered from normal to extremely high–speed adaptive to the current interestingness as well as user input. Ideally, boring, unimportant portions of the video will be played back at high speed, essentially skipping most of the frames, while important segments are played at more normal speeds. If the experience can be made seamless and fluid, the user may be able to glean almost all the value from the video while spending only a fraction of time viewing it. Methods to determine important segments of the video are presented in the next section.

We propose to use extremely high–speed playback (of the order of 100x) to solve the information overload problem inherent in many hours of linear video. This could raise issues of its own related to human short term visual memory and cognitive load associated with trying to keep up with rapid changes. Therefore we also use a context display to tell the user, effectively, what just happened and what is about to happen in the video. This is covered in Section 4.

3. METADATA CHANNELS

As mentioned above, the interestingness value for the current playback point in the video is used to determine the instantaneous playback rate. This interestingness value is derived from the three different sources of metadata mentioned above: automatic analysis, user annotation, and social. We refer to sources of annotation for a video as a metadata channel to emphasize the similarity to audio, graphical effect, or other parallel tracks used in digital video editing.

In our current implementation we use an automated video content analysis module. Given an input video file, this performs shot change detection and face detection and generates a metadata channel with these events tied to frame number. These video events are used to select the keyframes that will be displayed in the keyframe context buffer. The keyframe information can also be used to determine the importance level of a segment.

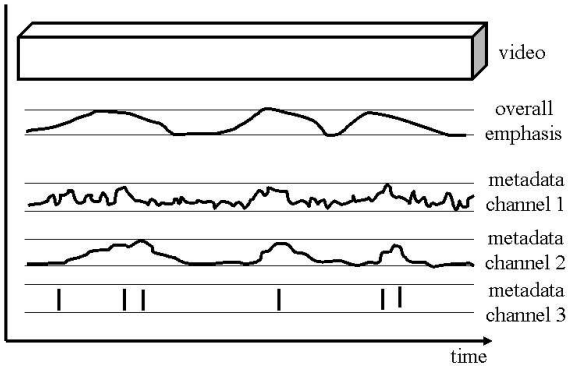

Figure 1: Illustrating metadata channels combined to compute emphasis

In addition to automatically generated metadata channel we allow the overlay of a user annotation metadata channel. Our current implementation uses the open source video player Mplayer which as a facility for the user to mark points in video. We adapt this feature to allow the user to mark segments as favorites or as being of high importance or alternatively as unimportant. These two metadata channels are combined with any social usage metadata (other people's annotations) to generate the per-frame importance level.

4. THE CONTROL INTERFACE

In any system that attempts to compact or select information for faster browsing by a user there is an increased emphasis on certain items or portions of data (those that are selected for representation) and a decreased emphasis on the elided portions. The problem with any automatic scheme to determine which portions are which is that the algorithms may make errors. This could be because of classification errors, such as marking a face where there is none, or vice versa. But it could also be because the assumption embedded in the algorithm, that faces are important, is untrue for a particular user. Therefore we need to provide a mechanism for the user to override and possibly invert the systems judgement. The system should be able to fallback to a mode where it does no compaction whatsoever. And ideally the user can move fluidly between a range of modes.

Given a set of metadata channels such as described above, we have a sparse list of events in the video. The default baseline behavior of the system is to play video marked unimportant at high speed, by seeking ahead a certain number of frames and to play important segments at normal speed. However, the user may disagree with the systems notion of importance. In addition the user may wish to increase or decrease the overall speedup. To allow this we have designed a control interface with two axes shown in Figure 2. This depicts a two-dimensional input control space that the user can specify. The co-ordinate on the horizontal axis controls the negative emphasis placed on unimportant segments, i.e. it represents the amount of speedup of the video for unimportant segments while the vertical co-ordinate determines the relative emphasis placed on important segments. Note that increasing the positive emphasis while keeping the negative emphasis constant results in a lower interestingness threshold being required for normal speed (emphasized) playback. These axes are mapped to the position of the mouse within the software player video window. The user can thus control both the overall and relative speedups. Positioning the

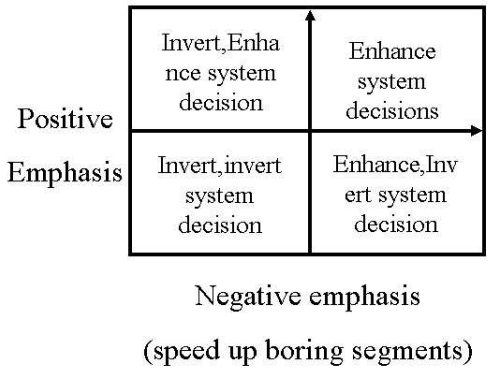

Figure 2: Design of the user control space

Figure 3: The augmented player interface

mouse at the center of the control UI nullifies all system playback rate adaptation and reduces the display to normal video playback. Moving horizontally increases the speedup of uninteresting portions of the video. In the bottom left quadrant, the system's ideas of importance and unimportance are reversed.

A buffer of keyframes is displayed along the vertical margins of the player window. The keyframes are those from the past (i.e. recently played) and the future (yet to be played). Buffering the keyframes of the past allows the user to review content that has been played at high speed and if desired, to go back to view that content. This is useful both to give the user context and for her to recover if the system has incorrectly labeled a semantically important video segment as unimportant.

5. PRELIMINARY RESULTS

Our current implementation uses the open source video player Mplayer's video filter architecture. A photograph of the augmented video player display is shown in Figure 3 playing back a roughly one hour home video file.

The user experience based on our own playing with the prototype on a small set of home video files is satisfactory with scope for planned improvements. The method does allow one to skip through the boring portions of a video while still watching the interesting portions at normal speed. If keyframes are detected for short unimportant segments that are skipped over then the viewer gets a sense of them as well. Currently we do not impose a constraint that skipped segments must contain a minimum positive definite number of keyframes but this may be a useful feature.

Direct access to past or future keyframe segments by clicking on the keyframe proves to be very useful. After clicking on the keyframe, the system can optionally boost the emphais on the segment containing the keyframe. This becomes an online implicit annotation by the user on the selected segment. The user can mark points as being interesting via a keyboard interface. These points are saved to a metadata file and are re-used the next time that video file is played.

One assumption of this work is that the video is stored locally and thus high-speed playback is not subject to the latencies that would be apparent in network playback, at least over networks and media applications that are not aware of each other. Even with local video, the need to read ahead in order to fetch future keyframes creates issues with decoding.

There are issues with breaking the traditional video playback model. Video playback relies on psychovisual image retention in order to maintain the feeling of continuity. By skipping intermediate frames we are increasing the disparity between consecutive images. If one focuses on the video this can be very slightly jarring, although not terribly so. The user always has the option to revert to watching the keyframe buffer as a surrogate instead.

A bigger problem is audio. Rapid skipping when playing back at high speed results in chopped audio segments being played back. While this gave the viewer a better sense of the video, it did create a somewhat jarring effect. One solution would be to mute the audio while playing back at high speed. However this violates our design principle of giving the user as much contextual information as possible. Another solution being considered is to allow the audio track to fall out of synchronization briefly while skipping ahead with the video. Playing back a certain minimum number of consecutive audio data packets segments before allowing a resynchronization should result in more complete audio fragments.

6. RELEVANT PRIOR WORK

Indexing digital video for browsing has generally been achieved in the past by the creation of storyboards. The video is analyzed to detect shot changes, scene changes, and representative keyframes [9]. The keyframes may be clustered or otherwise arranged hierarchically [1, 9] or laid out according to importance [6]. This static browsing index is presented to the user and clicking on a keyframe results in a jump to the specified segment of video.

While a static index has many advantages, not least that it can be printed and used as an index for physical media, it does drop a lot of information in order to be compact and manageable. There is also the experiential downside of converting video to images. Studies comparing static to dynamic visual indexes have found that on measures other than object recognition, static presentation was no better than dynamic presentation [4].

Other representations of video indices include the Video Skim [2]. A video skim is a compacted version of a video formed by retaining segments with high importance, typically identified on the basis of a closed caption transcript or speech recognition as well as face detectors and motion

analysis and removing other segments. In principle this is similar to our use of metadata to highlight interesting segments. However, the video skim is still played back essentially as a series of short clips with no context of past or present. Since less important segments are completely removed, there is no way for the user to override the system's opinion of importance.

Rapid Serial Visual Presentation (RSVP) methods have been proposed for navigating multiple kinds of information [7], especially on constrained screen devices [3]. In particular they have been used for browsing large image collections as well as video browsing [5]. In the latter case, keyframes were statically selected and then displayed in rapid succession. Experiments were performed to test gist determination, object recognition and action recognition. Our work differs from this approach in two ways: we use actual video frames, not keyframes, to display directly. Keyframes are used only to provide context in the past and future. Additionally, the user has complete control over the current playback rate. If the user reduces the overall speedup to 1, then every frame becomes a keyframe and the method defaults to normal video playback. The user may also choose to reverse playback. These two refinements, the keyframe buffer and user control, alleviate one of the issues with RSVP and related techniques generally—that they impose a high cognitive load on the user.

A dynamic (varying spatial layout) RSVP technique for consumer video was presented by Wittenburg et al [8]. A variety of different 3-D effect trails were used to present the user with context in both past and future with a maximum speedup of 11 times realtime. This is similar to the use of the keyframe buffers in our system. It is however an image, i.e. keyframe, based RSVP method which does not revert to normal video playback under user control.

7. CONCLUSIONS

We have presented a new technique for scalable exploratory search of digital video via metadata–augmented video playback. It uses content analysis and user annotation to determine the importance of portions of a video file and varies playback rate adaptive to the importance. It display past and future keyframe buffers to provide context to the user in the form of keyframes, which is especially useful in the case of high–speed playback. It differs from prior work in the use of a novel control interface to allow the user to enhance, override or invert the system's notions of importance and to revert to normal playback. Preliminary results on a small dataset are presented.

Our ongoing work on this project involves making the user interface smoother, handling audio in a better way and performing user studies to refine the prototype. Future enhancements may include the use of metadata channels from other users in a collaborative filtering approach.

8. ACKNOWLEDGMENTS

We would like to thank the authors and maintainers of the open source Mplayer software and the many libraries that it depends upon for building a platform that we could alter to our needs.

9. REFERENCES

[1] J.-Y. Chen, C. Taskiran, E. Delp, and C. Bouman. ViBE: A New Paradigm for Video Database Browsing and Search. In *Proc. IEEE Workshop on Content-Based Access of Image and Video Libraries*, pages 96–100, 1998.

[2] M. Christel, M. Smith, C. Taylor, and D. Winkler. Evolving video skims into useful multimedia abstractions. In *ACM Conference on Computer Human Interaction (CHI)*, pages 171–178, April 1998.

[3] O. de Bruijn and R. Spence. Rapid serial visual presentation: a space-time trade-off in information presentation. In *Working conference on advanced visual interfaces AVI'98*, 1998.

[4] A. Komlodi and G. Marchionini. Key Frame Preview Techniques for Video Browsing. In *ACM International Conference on Digital Libraries*, pages 118–125, 1998.

[5] T. Tse, G. Marchionini, W. Ding, L. Slaughter, and A. Komlodi. dynamic keyframe presentation techniques for augmenting video browsing. In *Working conference on advanced visual interfaces AVI'98*, 1998.

[6] S. Uchihashi, J. Foote, A. Girgensohn, and J. Boreczky. Video manga: generating semantically meaningful video summaries. In *7th ACM international conference on Multimedia*, pages 383–392, 1999.

[7] K. Wittenburg, C. Chiyoda, M. Heinrichs, and T. Lanning. Browsing through rapid-fire imaging: Requirements and industry initiatives, 2000.

[8] K. Wittenburg, C. Forlines, T. Lanning, A. Esenther, S. Harada, and T. Miyachi. Rapid serial visual presentation techniques for consumer digital video devices, 2003.

[9] H. Zhang, C. Low, and S. Smoliar. Video Parsing and Browsing using Compressed Data. *Multimedia Tools and Applications*, 1(1):91–113, 1995.