VIDEOABSTRACT: A HYBRID APPROACH TO GENERATE SEMANTICALLY MEANINGFUL VIDEO SUMMARIES

*Candemir Toklu, Shih-Ping Liou and Madirakshi Das**

Siemens Corporate Research
Multimedia & Video Technology Department
755 College Road East, Princeton, NJ 08536, USA

* Department of Computer Science
University of Massachusetts
Amherst, MA 01003, USA

ABSTRACT

Video summarization is a key component in providing Internet users a way to quickly browse a video clip in different levels of detail, without the need to view the entire video clip. We present a hybrid approach to video summary generation, which automatically process the video, creating a multimedia video summary, while providing easy-to-use interfaces for verification, correction, and augmentation of the automatically generated story segments and extracted multimedia content. Algorithms are developed to solve the sub-problems of story segmentation, story boundary refinement, and video summary generation. The use of automatic processing in conjunction with input from the user allows a user to produce meaningful video summaries efficiently.

1. Introduction

Most video data (even in the compressed form) is too big to transfer over general-purpose networks such as the Internet. It is therefore extremely important for such a system to provide users a way to quickly browse a video clip in different levels of detail, without the need to view the entire video clip. This requires the development of a *video abstraction* or *summarization* system. The objective of video summarization is to capture the structure present in the video, and select the *best* images, sounds, and text chunks to represent each story in a compact form.

Obtaining a video summary to a reasonable degree of perfection automatically, is a complex process. First, video frames must be segmented into a set of units called *shots*. A shot in a video refers to contiguous frames depicting continuous action in time and space. Consecutive shots are then grouped to form a story unit, where various tracks (audio, visual, and text) are aligned. Finally, representative content from the original video must be extracted to create the multimedia summary. Since the generation of video summary comprises a sequence of dependent steps, even one mistake can create useless results.

The research on video summarization has been focused on developing integrated algorithms to segment a video into semantic units, extract key frames and key phrases, and compose a visual or multimedia summary by concatenating all extracted content together. This automatic approach (which requires automatic speech recognition, and natural language and image understanding) is computationally expensive, difficult, and tends to be very domain specific. It is nearly impossible, in practice, to obtain useful video summary based solely on automatic processing.

On the other hand, such information can be manually indexed into the system, creating high quality video structure to produce multimedia summaries. Since this is a manual process, it is be very time-consuming.

What we need is a hybrid approach, i.e., automatically reconstructing a semantic structure for video and creating the video summary in a preprocessing step, while providing an easy-to-use interface for verification, correction and augmentation of the automatically generated video summary. This paper describes such a hybrid approach that includes several automatic algorithms closely integrated with three graphic user interface objects. The automatic algorithms include cut detection, story segmentation based on the closed-captioned text (in the absence of subject change markers), shot grouping, audio segmentation, story boundary refinement, key-frame selection, and video summary generation, whereas the graphic user interface objects consist of the video player, browser, and table of contents viewer.

This paper is organized as follows. Section 2 reviews the past literature and provides the motivation for this work. A brief summary of the novel aspects of the proposed method is given in section 3. Section 4 described the algorithms used in automatic generation of video summaries from the raw video. Section 5 describes the interactive aspects of video summary generation, including the user interface. Finally, it concludes with comments on future directions in section 6.

2. Literature Review

Video summarization methods can be classified into two major approaches: (i) visual, and (ii) multimedia summary generation.

Visual summary generation algorithms focus on finding a smaller set of images, *key-frames*, to represent the visual content in the video, and presenting these key-frames to the user. Key-frames can be chosen at uniform time intervals as in the MiniVideo [1] and Video Magnifier [2], or selected based on the content as in [3]-[8].

To our knowledge Informedia system [8] and the auto-summarization system for audio-video presentations [9] are the only content-based audio/video browsing and retrieval system that addresses multimedia summary generation. Informedia system [8] uses intensive analysis techniques from computer vision, speech analysis, and natural language processing to construct a *video skim*, which represents a short synopsis of the original video. The audio component of the video skim is obtained by adding ranked segments of the audio track. Audio segments are simply words or phrases, and ranking is based on the frequency of the word or phrase in the video. The audio skims generated are fragmented and usually difficult for users to comprehend. The auto-summarization system for audio-video presentations [9] exploits the information in the audio signal (e.g. pitch and pause information), knowledge of slide transition points in the presentation, and information about access patterns of previous users. Therefore, this system is heavily domain dependent.

Yet another study, the automatic system for authoring of hypermedia documents of videos [10], addresses multimedia representation of videos that are accompanied by closed-captioned text. However, this study does not address story segmentation and summarization of closed-captioned content for more compact representation of video.

3. Our Approach

Our approach to video summarization is illustrated in

Figure 1. This hybrid approach consists of a set of automatic algorithms supported by content processing tools and a collection of interactive interfaces. Our approach differs from existing approaches in the following aspects.

First, none of the existing video summarization algorithms provide manual feedback mechanisms during the automatic creation of semantic video structure and generation of video summaries. Ueda *et.al.* [11] discusses a system for video authoring that allows manual feedback, however they do not extend it to video summarization. Errors produced by any one of the automatic algorithms will propagate, resulting in wrong story boundaries. Therefore, it is important to provide interfaces so that a human operator can verify and correct the results produced automatically at every step of the summarization process.

Second, we segment video into stories based on the speaker change markers and keywords in the closed-captioned text. Natural language

0-7803-6536-4/00/$10.00 (c) 2000 IEEE

keywords have much more descriptive power and are much easier to use than abstract image features.

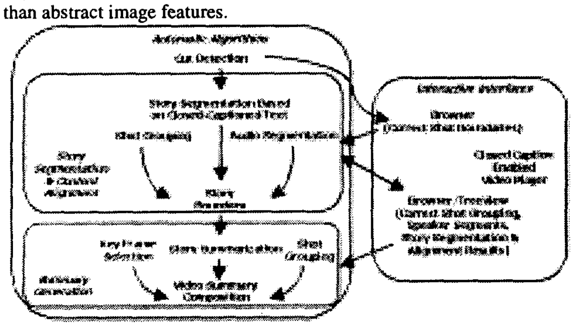
Automatic Algorithms
Cut Detection
Story Segmentation Based on Closed-Captioned Text
Shot Grouping
Audio Segmentation
Story Segmentation & Content Alignment
Story Boundaries
Key Frame Selection
Story Summarization
Shot Grouping
Summary Generation
Video Summary Composition
Interactive Interfaces
Browser (Correct shot boundaries)
Closed Caption Enabled Video Player
Browser /Treeview (Correct Shot Grouping, Speaker Segments, Story Segmentation & Alignment Results)

Figure 1: Illustration of the proposed video summarization method.

Third, since closed-captioned text often lags behind its corresponding audio and visual tracks, we suggest procedures for content alignment. These procedures rely on independent audio and video content analysis, providing needed robustness to the video summarization process.

Fourth, we do not use pre-defined video structures and domain knowledge, such as the broadcasting program and its story narration models, to facilitate story segmentation. Hence, the story segmentation approach is not domain or program specific, and it combines natural language, video and audio processing.

Finally, the content produced from the video and the system allows scalable summary generation. Thus, the video summary can be tailored according to the viewer's profile and needs. In addition, video summarization system also eliminates the redundant information such as the repeating shots in the final multimedia presentation of video.

4. Automatic Video Summary Generation

We have developed a set of automatic algorithms that can produce semantic structures from raw video, and generate multimedia summaries of such structures. The first task in automatic summary generation is to recover the shots present in the video. We incorporate a scene change detection method [12]. Each shot is completely defined by a start and an end frame and a list of all shots in a video is stored in a *shotlist*. A single frame, the *representative frame*, represents each shot.

The more complex task is to organize the shots into a story structure reflecting the semantics of the video as well as align all tracks (text, audio, visual) of the video content. We have automated the process of inferring the semantics of the video using information stored both in the closed-captioned texts and image sequence track of the video. To align different tracks of the video content, we have to first analyze information that exists in all tracks and then fuse the independently processed results.

Once the story structure is obtained and all tracks are properly aligned, video summaries can be generated by extracting *representative* content from the original video. This involves obtaining text summaries at different level of details, selecting key-frames that best represent the visual track, and using shot grouping results to eliminate all key-frames belonging to an anchorperson or a reporter shot.

In the following sub-sections, we will describe the story segmentation, content alignment and summary generation algorithms respectively.

4.1 Story Segmentation and Content Alignment

To automatically group shots into story units requires the understanding of the semantic structure present in the video. Many researchers [5],[13] use strong domain knowledge to constrain this problem. Others [6] identify repeating shots to detect parallel events in the video and construct a scene transition graph that serves as a story summary for video browsing applications. Recently, researchers started integrating information from different media to automatically reconstruct semantic video structures [14],[15]. However, the method of [14] is heavily domain dependent, since they suggest using keywords specific to news broadcast for finding the story boundaries.

Today, many videos are made available with closed-captioned text (or transcripts in Europe), even in the case of live broadcast. In addition to the textual representation of the spoken words from the audio track of the video, closed-captioned text also contains additional markers for change of speakers. *Change of speaker* markers may take different forms depending on the source from which the video was acquired e.g. name of the speaker followed by a colon, or the ">>" symbol. In addition to the *change of speaker* markers, special markers for indicating *a change of subject* may be included in the closed-captioned text. However, a change of subject is less commonly marked than the change of speakers since it is hard to do so without knowing the context of the broadcast at the time of captioning. When they are marked, a ">>>" symbol is often used at the point the subject is being changed. These markers can be used to directly segment the video into stories.

Although closed-captioned text is very useful, it often lags behind its corresponding audio and visual tracks. In addition, audio, due to editing effects, may not align with its associated video frames. Content alignment becomes very important if we would like to compose multimedia summaries at different levels of detail.

In the following sections, we will first describe how to use closed-captioned text to group shots into story units in the absence of change of subject markers. Finally, we will explain how audio can be segmented into different semantic units and the procedure for aligning the closed-captioned text to its corresponding audio and visual tracks.

4.1.1 Story Segmentation

If the special markers for indicating *a change of subject* are included in the closed-captioned text, then we sue them for segmenting the video into stories. If these markers are not present, then we apply the following procedure for finding the story boundaries. The first task is to recover the *speaker change markers* in the closed-captioned text. From the location of the speaker change markers we derive *speaker segments*. Each speaker segment is completely defined by a start and end (frame number of the current and next speaker change marker, respectively), and the list of all speaker segments in the video is stored in a *speakerlist*.

Sentences are considered to be the basic blocks in the closed-captioned text. Hence, we re-produce the sentences for each speaker segment. These sentences are processed to extract *noun phrases* and *proper nouns*, which are useful components for providing clues about the semantics of these sentences. When extracting noun phrases, the longest valid noun phrase is found. Proper nouns found include names of people, organizations, places and times. These data are stored in a data file indexed by the segment and the sentence number within the segment.

After extracting all speaker segments, the system then groups them using proper nouns to obtain story boundaries. This is because the same proper nouns often appear repeatedly across the body of a story. We match these elements to group speaker segments into stories. Starting with the last segment, we try to match proper nouns (if present) with preceding segments going back a maximum of about 2.5 minutes of actual time (which corresponds to about 4500 frames. We localize the search to a limited time band because we do not want widely separated segments to be merged together when there are no common elements found in the segments between them. Segments far from each other can still be in the same story if there are matching proper nouns found in the intermediate segments. This is assured because we search incrementally, going from each segment to its previous one and extending the search to a fixed number of frames in the past.

Proper nouns are matched as sub-strings so that a person's full name matches his/her surname or first name as well. Similarly, a full place name (including the state) matches the name of the place and vice versa. When a match is found, all segments between the current and the matched segment are grouped under one story. This is very useful since all segments that do not have proper nouns to match, but embedded between two matching segments still become a part of the same story.

4.1.2 Shot Grouping

To align story boundaries obtained from the closed-captioned texts with the accompanied visual information, we fuse the story segmentation results described in previous section with the results obtained from an alternating story segmentation approach that is only based on the visual information. In an earlier study [16], we have observed that repeating shots which are similar in some way, alternating or interleaving with other shots, are often used to convey parallel events in a scene or to signal the beginning of a semantically meaningful unit. Hence, we have proposed a shot grouping method and alternative story segmentation based only on visual information in that study. Since we do not use pre-defined models for similar shots, this observation can be used for a wide variety of videos for which organization is particularly relevant e.g. news, sports events, interviews, documentaries etc.

4.1.3 Audio Segmentation

As we have discussed earlier, closed-captioned text often lags behind the corresponding spoken words. Without proper alignment, it is not possible to use original audio track in the constructed multimedia summary.

The first step in aligning audio with closed-captioned text is to segment audio into semantically meaningful units, as we segment the video into shots. In our approach, we divide audio into *speech, non-speech sound*, and *silence units* (*events*) [17].

We have observed that the silence events usually correspond to sentence, speaker, and story boundaries (with increasing length respectively). Thus, the audio track corresponding to each speaker segment or story can be extracted by detecting the silence regions in the neighborhood of the start and the end frames of a speaker or a story segment. This operation also facilitates the audio and closed-captioned text alignment.

4.1.4 Story Boundary Refinement

At this stage we have grouped shots based on their visual appearance and segmented audio into speech, non-speech, and silence units. In addition, the two sets of story boundaries are determined based on the closed-captioned text and the visual content in the video. The boundaries, *visual-story* boundaries, obtained by looking to group of images are aligned to shot boundaries. On the other hand, the stories obtained from the closed-captioned text lags behind the corresponding visual content. We have observed that a speaker or story change marker is usually accompanied by a shot change, e.g., a new story always accompanies a new shot. Based on this observation, we have developed the following procedure to align the story boundaries to the visual-story boundaries. We start with the first story, and find the shot that contains the start frame of this story. After finding the visual-story that contains this shot, we set the start frame of the current story to the start frame of this visual-story. If two or more stories matches to the same visual-story, then their start frames are set to the start frames of the shots that contain their start, except the first one in time that matches to this visual-story. This procedure is repeated for the subsequent stories in the video. Finally, we update the end frames of each story. Since the start frame of the next story in the list determines the end frame of the current story, end frames of the stories are set to the new start frames of the subsequent stories.

4.2 Summary Generation

We generate video summaries as slide-shows with additional audio segments as narration. We could also generate video summaries in the form of composite images comprising many representative images with or without voice comments.

In order to compose a video summary, our system first segments the video into stories as described above, and then extracts the best representative content for each story from the original video. This includes text summaries, representative image frames, and corresponding audio data for each story. Since it is more difficult to semantically summarize visual than textual content, our system creates summaries at different granularities only based on the textual information. Once text summaries are obtained, the system simply uses either clips from the audio track of the original video or text-to-speech synthesis as narration.

4.2.1 Story Summarization

Earlier, we have suggested representing each shot with a representative frame, the image corresponding to the first frame in a shot. Generally speaking, in a shot, where visual effects and object and/or camera motions are prominent, one representative image is not sufficient to represent the entire shot. We need to select additional frames, *key-frames*, from the shot. We employ the content-based key-frame selection method described in detail in [18] for selecting a set of representative frames for each shot.

Given the closed-captioned text, summary sentences are generated using an off-the-shelf text summary generation tool such as Inxight Summarizer™[19]. The number of sentences in the story summary can be specified, depending on the level of detail required in the final summary.

4.2.2 Story Summary Composition

Our system first creates an initial list of images by including all representative frames and key-frames falling within the boundary of the story. This list could include many images representative of the story, but it could also include images of the anchorperson or the reporter, not useful in providing glimpses of the story.

Such anchorperson or reporter images are often the most common images in a news or documentary video and should be removed from the initial list. This process is analogous to the process used in indexing text databases, where the most frequently occurring words are eliminated because they convey the least amount of information.

The remaining list of images is sampled to produce a set of images for the summary composition. The simplest way is to sample uniformly with the sampling interval being determined by the number of images required for the given length of the summary. However, we can also use image frames that correspond to the extracted proper nouns. This is important if a story is about a person, a places, etc. present in the video.

5. Interactive Interfaces

We now describe the tools provided to the user to modify the results of automatic story segmentation at both the shot and story levels. The steps in the interactive generation of the final organized video can be summarized as follows:

1. Automatic construction of *shotlist* from raw video.
2. * Viewing and editing the shot change structure to add new shots or merge shots.
3. Automatic clustering of shots into visually similar groups.
4. * Viewing and editing the shot grouping results generated automatically to create new groups and modify existing groups.
5. Automatic extraction of speaker segments and clustering of speaker segments into stories based on proper nouns.
6. Automatic refinement of the story boundaries based on shots, shot groupings and audio segmentation.
7. Viewing and modifying the story boundaries to reorganize the video.
8. Automatic generation of multimedia video summary

The steps marked with an asterix in the above list need user interaction and therefore, an interface needs to be provided with the required functionality. Since these interfaces work with higher level representations of the video, a separate component is also provided to view the raw video accompanied by its audio. The interactive video organization environment provides three main interfaces to the user, which communicate with each other so that changes made using one component produce the appropriate updates in the other interfaces.

Browser: This interface is used to view and modify the audio classes, shotlist and speakerlist. The video stream is represented as a composite image, *cross-section image*, making it easy to detect shot boundaries visually. The results of automatic cut detection are displayed alongside using colored bars, and can be altered easily. Similar to cross-section images, an *audio spectrogram,* and a curve, *pitch curve,* is used to represent the audio information space. We consider the audio spectrogram as a raw representation of the audio track of the video. Audio spectrograms have been used for visualizing the sounds for quite long. Audio spectrogram also contains information about different events in audio for an experienced user, i.e. spectrum of music and speech has different patterns, useful for correcting the automatic segmentation errors. Segmentation of the audio into events (silence, speech and non-speech segments) further enhances the audio information space. Similarly, the audio events are represented with a colored bar alongside the spectrogram to facilitate audio event correction.

TreeView: This interface serves a dual purpose. Before the story boundaries are computed, it is used to view and alter the automatic grouping of shots based on visual similarity. Once the similarity-based grouping results have been finalized, it is used as an interface to correct story boundaries and align spoken sentences to the sentences obtained from the closed-captioned text. The user is allowed full freedom in changing the story boundaries, since semantic information can often be missed or misinterpreted by automatic processing. The following operations are provided to facilitate changes in the structure of the tree.

- *Move nodes*: Nodes can be moved either one or more at a time by selecting the node(s) to be moved and the destination node. The moved node(s) are added as siblings of the destination node selected.
- *Add nodes*: New leaf nodes can only be added through changes in the *shotlist* using *Browser.* However, all types of non-leaf nodes can be added as parents of existing nodes.
- *Update*: This operation is a utility to make it easier to move large number of nodes. It repeats the previous operation on all siblings of the node that was selected for the previous operation.
- *Interactions*: When changes are made in the order of the shots, audio events, or speaker changes using *TreeView,* and the user wants to see these changes reflected in the *Browser*, (s)he can opt to send a signal to the *Browser* to reload the rearranged shotlist, audio event list and speakerlist after saving it from *TreeView*. The *VideoPlayer* can be called from this interface exactly as in the *Browser* interface.

No explicit delete function is provided in this interface since leaf nodes can only be deleted through changes in the shotlist using the *Browser.* All other (non-leaf) nodes are deleted automatically when they have no children.

The user can invoke these operations to regroup the stories into more meaningful stories and sub-stories. The order of stories can also be changed from their usual temporal order to a more logical sequence. When used along with the *Browser*, all possible changes to the content and organization of the stories are supported.

Though the primary function of the *TreeView* interface is to interact with the high-level structure, the same interface can also be used to view and modify the groups generated by the automatic grouping process. In this case, there are only two types of nodes attached to the root node. If the group contains a single member, the member shot is attached as a leaf node to the root. For groups containing more than one member, an intermediate *Group* node is attached that contains the member shots as its children.

VideoPlayer: This interface is used for playing the video, with audio, from any point in the video. It has the functionality of a VCR including fast forward, rewind, pause and step. It can also display the closed-captioned text.

6. Conclusion

We have proposed a hybrid video summarization system that combines the benefits of automatic processing with human input. Algorithms for producing a multimedia video summary automatically from a raw video have been developed by solving the sub-problems of cut detection, story segmentation, shot grouping, audio segmentation, story boundary refinement, key-frame selection, and video summary generation. Automatic processing reduces the user's work from going through the whole process manually to just provide error checking. We have constructed an environment that provides effective tools to the user to guide the video summary generation process. The system has been tested on news stories and documentary videos and produce good summaries requiring only small alternations from the user.

An important direction of future work is to enhance the interactive interface to allow better feedback from the user to improve the alignment of closed-captioned text with its corresponding audio track. It is also useful to have the system utilize the user feedback in a more intelligent way. For example, corrections made on story boundaries can be used by the system to fine-tune the shot-grouping algorithm. Correction made on speaker boundaries can be used to fine-tune the audio segmentation algorithm. Using the cues from the user will further reduce the work that is needed to modify the story structure automatically.

7. References

[1] Y. Taniguchi, *et. al.*, "An intuitive and efficient access interface to real-time incoming video based on automatic indexing," *Proc. of ACM Multimedia*, pp. 25-33, San Francisco (CA), Nov. 5-9, 1995.

[2] M. Mills, *et. al.*, "A magnifier tool for video data," *Proc. of ACM Human Computer Interface (CHI)*, pp. 93-8, May 1992.

[3] B. C. O'Connor, "Selecting key frames of moving image documents: a digital environment for analysis and navigation," *Microcomputers for Info. Management* 8(2) pp. 119-133, 1991.

[4] M. M. Yeung and B. Liu, "Efficient matching and clustering of video shots," *Proc. IEEE ICIP*, Vol. 1, pp. 338-41, 1995.

[5] H. J. Zhang, *et. al.*, "Video parsing and browsing using compressed data," *Multimedia Tools and Applications,* 1(1), pp. 89-111, 1995.

[6] M. M. Yeung and B.L. Yeo, "Video visualization for compact representation and fast browsing of pictorial content," *IEEE Trans. on Circuits and Systems for Video Tech.*, vol. 7, no. 5, 1997.

[7] S. Uchihashi, *et. al.,* "Video Manga: generating semantically meaningful video summaries," in *Proc. of ACM Multimedia,* 1999.

[8] A. G. Hauptmann and M. A. Smith, "Text, speech, and vision for video segmentation: The Informedia Project," in *Proc. of the AAAI Fall Symp. on Comp. Models for Integrating Lang. and Vision*, 95.

[9] L. He, *et. al.,* "Auto-summarization of audio-video presentations," in *ACM Multimedia Proceedings*, pp. 489-98, 1999.

[10] B. Shahraray and D. C. Gibbon, "Automated authoring of hypermedia documents of video programs," in *Proc. of ACM Multimedia*, 1995.

[11] H. Ueda and T. Miyakate, "Automatic scene separation and tree structure GUI for video editing," in *Proc. of ACM Multimedia*, pp. 405-6, 1996.

[12] S. P. Liou, *et. al.* "Efficient and reliable digital media archive for content-based retrieval," Multimedia Systems, 7(4), pp. 256-68, ACM Press, July, 1999.

[13] D. Swanberg, *et. al*, "Knowledge guided parsing in video databases," *Storage and Retrieval for Image and Video Databases*, SPIE: 1908, 13-25, 1993.

[14] A. Merlino, *et. al.*, "Broadcast news navigation using story segmentation," in *Proc. of ACM Multimedia*, pp. 381-91, 1997.

[15] Q. Huang, *et. al.*, "Automated semantic structure reconstruction and representation generation for broadcast news," in *SPIE Proceedings on SRIVD VII*, vol. 3656, pp. 50-62, Jan. 1999.

[16] M. Das and S. P. Liou, "A new hybrid approach to video organization for content-based indexing," *Proc. IEEE International Conference on Multimedia Systems and Computing,* 1998.

[17] C. Toklu and S.-H. Liou, "Image and audio sequence visualization and interaction mechanisms for structured video browsing and logging" *to appear in IEEE 2000 International Conference on Image Processing.*

[18] C. Toklu and S.-H. Liou, "Automatic key-frame selection for content-based video indexing and access ", *Proceedings of the SPIE for Storage and Retrieval for Media Databases 2000.*

[19] http://www.inxight.com.