INTEGRATING VISUAL, AUDIO AND TEXT ANALYSIS FOR NEWS VIDEO

Wei Qi, Lie Gu, Hao Jiang, Xiang-Rong Chen and Hong-Jiang Zhang

Microsoft Research, China
5F, Beijing Sigma Center
Beijing 100080,P.R.China
E-mail: hjzhang@microsoft.com

ABSTRACT

In this paper, we present a system developed for content-based broadcasted news video browsing for home users. There are three main factors that distinguish our work from other similar ones. First, we have integrated the image and audio analysis results in identifying news segments. Second, we use the video OCR technology to detect text from frames, which provides a good source of textual information for story classification when transcripts and close captions are not available. Finally, natural language processing (NLP) technologies are used to perform automated categorization of news stories based on the texts obtained from close caption or video OCR process. Based on these video structure and content analysis technologies, we have developed two advanced video browsers for home users: intelligent highlight player and HTML-based video browser.

1. INTRODUCTION

TV is one of the most important sources of information at home. However, today's TV technology forces viewers to be passive and follow TV schedules, especially in news watching. Such a situation is being changed with the rapid development of new generation TV technologies, such as WebTV and video over IP and it becomes desirable to provide intelligent recording and interactive browsing functions. One of the application scenarios is a personalized digital news recorder that is capable of recording news stories of interest to a particular user and filtering out others. To achieve this objective, automatic news video structure parsing and content analysis tools are essential.

Due to its importance and usefulness, there have been many research efforts in news video structure and content analysis. The earlier work by Zhang, *et al,* focused on news video story parsing based on well-defined temporal structures in news video [1]. Repetitive patterns of anchor appearance in news video was detected using simple motion analysis based on predefined anchor shot templates and was used as indication of news story boundaries. However, only image data were used in this proposed scheme, and only minimum content-based browsing can be done with such a scheme. The work reported by Shahraray and D. Gibbon [2] used key-frames and text information to provide pictorial transcript of news video, with almost no automatic structural and content analysis. A very impressive work was reported by the *Informedia,* in which speech and image analysis were combined to extract content information and to build indexes of news video [3]. Lately, more research efforts adopted the idea of information fusion such that image, audio and speech analysis are integrated in video content analysis [e.g. 4, 5].

In this paper we present a system developed for content-based news video recording and browsing, mainly focused for home users. In this system, visual, audio and text information are integrated in video structure and content analysis. As shown in Figure 1, we consider that news video has a hierarchical structure of three layers. Based on this structure, we developed a structure parsing system that integrates visual, audio and text analyses to extract structural elements at each layer. At the bottom level, image analysis is performed to segment a video sequence into shots, which forms the basic units of the video. Above 'shot' level is the 'scene' or "group of shots" which are a sequence of shots that share the similar visual or audio contents. The groups of shots are extracted by grouping process using both visual and audio information. The composition unit at the top level of the structure is 'story', which is a complete news item separated lead by an anchorperson shot. Also, we developed a clustering based algorithm that combines the visual and audio method to detect anchorperson in the program as the indicator of story boundaries. Furthermore, by using NLP techniques, stories are classified into pre-defined categories based on text information associated with each news item, and associate stories of each news item are searched and linked with the news story. In case that transcripts or close captions are not available with news video, we use a video OCR technique to extract text from image sequence.

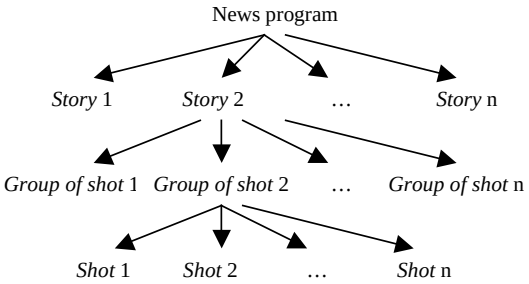

Figure 1 Hierarchical structure for news video.

The rest of the paper is organized as following. Section 2 first presents in detail our approach to news video structure parsing by integrating image and audio analysis. This is followed by a detailed presentation of algorithms for video OCR and text-based news categorization. In Section 3, we present two advanced news video browsers that supporting content-based news browsing. Section 4 concludes the paper.

2. NEWS STRUCTURE PARSING AND CONTENT ANALYSIS ALGORITHMS

The core of our news parsing system is the algorithm for news video structure parsing that integrates audio aided video scene analysis scheme. The main contribution for our method lies in two aspects: more robust and accurate shot grouping and anchorperson detection in news video. This is the major difference of our system from other work on integrating visual and audio information in video structure and content analysis [2, 3, 4, 5]. Also, our system has the capability to automatically categorize news stories to pre-defined news categories.

2.1 News structure parsing algorithms

Figure 2 shows our audio aided video scene analysis system. First, the input news video stream is split into audio stream and video stream. Then, audio stream is classified into four classes: speech, music, environment sound and silence. For speech data, it is further segmented into different elements according to different speaker. At the same time, shot detection and key frame extraction is performed on video stream. Then, color correlation analysis between shots is performed and a so-called expanding window grouping algorithm is applied such that shots whose objects or background are closely correlated, for instance shots occurring in the same environment, are grouped.

At the next step, audio and visual analysis results are fused to refine shot grouping. For this, the result of speaker change detection are combined with that of the color-based shot grouping results to find the 'group of shots' defined in Figure 1. The fusion rule is that the shots within a speaker segment and correlated according color correlation analysis are grouped together and marked as related. In other words, a sequence of shots will be grouped into a scene only when both visual content correlation analysis and audio segmentation detect a common scene boundary.

To extract news story boundaries, robust anchorperson detection is needed. On the audio side, segments marked by speaker change points are further clustered; while on the video side, shots are grouped based on key-frame clustering. Then, we use the following heuristics to detect anchorperson candidates on both audio and video data:

(1) The proportion of the anchorperson speech/image in the news broadcast is usually higher than the other speech/image.
(2) The distribution of the anchorperson speech/image is more disperse in time than other speech/image. That is, it will appear from the beginning to the end of a news broadcast.

In this way, audio and visual analysis results are combined again, resulting in accurate detection of the anchorperson shot. The fusion rule used here is that: do AND operation between anchorperson shots and anchorperson speech segments. The test on hours of CNN news video shows that accuracy of the clustering based news segmentation approach as described above achieve close to 98%[6].

The result of this structure parsing process is a set of meta data about a given video, structured with three layers as illustrated in Figure 1. Such meta data forms the basis for further content analysis of news video and structure-based parsing.

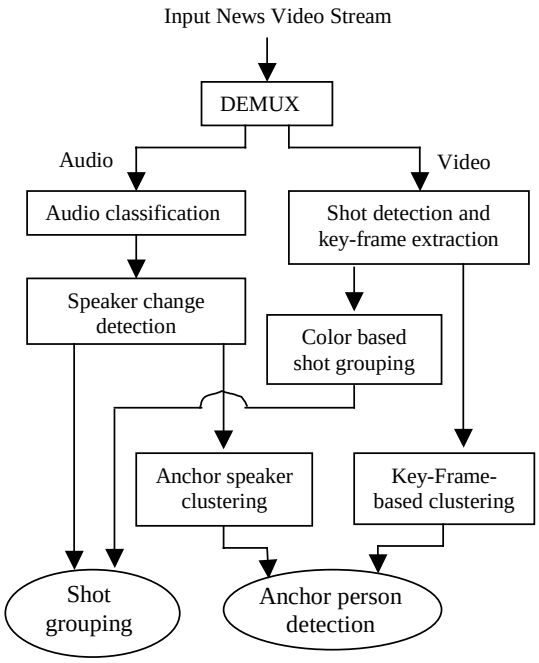

Figure 2: Audio aided news video parsing process.

2.2 Detecting text from video frames

News parsing process as presented in Section 2.1 extract structure information of a news video program. However, to extract semantic content information of news for content-based news indexing and search is a much more challenging task. One of effective way to achieve this solve this problem is to integrate textual information embedded in news video.

Text is an important information source often embedded in news video and it is a useful data for news video content, especially for high-level semantic content analysis, such as news categorization and associate story searching. For broadcast news video, text information may come in the format of caption text strings in video frames, as close caption, or transcripts. While signal noise level is low, text information may also be obtained from speech analysis. However, in many cases, transcripts and close caption of news program may not be available, and speech recognition may not product a text transcript with high accuracy. In such cases, extracting text information directly from image sequence plays a key role in news video content analysis, which often referred as *video OCR*. For this purpose, another important component of our news parsing and browsing system is automated video OCR [3, 7].

There are mainly two steps in video OCR: text extraction and text recognition. We developed a fast and robust video OCR algorithm [7]. As shown in Figure 3, we first apply a horizontal and vertical Sobel differential calculator, followed by an edge thinning process on the original image, to obtain a vertical edge map, and a horizontal edge map of the frame. From the vertical edge map, we obtain candidate text regions. Then, we perform horizontal edge alignment to eliminate false candidate areas. Finally, we use a shape suppression technique based on Bayesian decision theory to avoid false candidates resulting from non-text texture areas. Experiments have shown the proposed approach is very efficient and accurate in text area detection (see Figure 4).

For detected text areas in a video frame, we perform image enhancement using multiple video frames. This will smooth the background of text areas, and enhance the text strings. After this processing, we can then apply conventional OCR engine to the detected text area and extract text information.

We have tested our method on 2 hours CNN TV programs and a 21 seconds segment of MPEG-7 test data. The data contains many different sources, including business news, sport news, commercials, movies, weather reports, etc. With the proposed approach, over 95% of text regions have been detected correctly with a false detection rate lower than 5% in most cases except TV commercials. In TV commercials, characters in a same text region often vary to a large extent in both font and size, thus cannot pass the horizontal edge alignment confirmation.

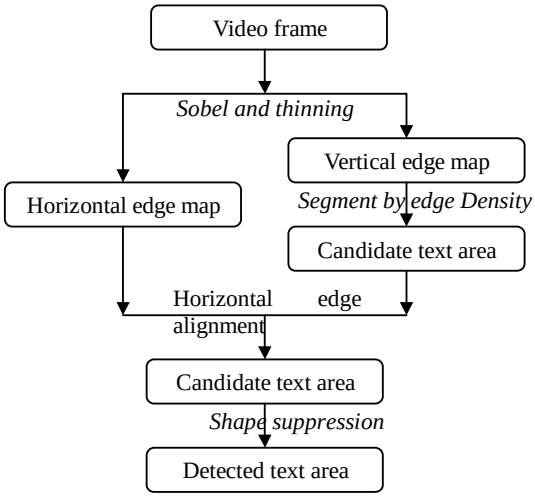

Figure 3: Diagram of the video text extraction algorithm

(a) Business news (b) Movie credit

Figure 4: Some experimental results for text extraction

2.3 News story categorization and associate story search using NLP technology

Text categorization is the process to assign natural language texts to one or more predefined categories based on their content. This is an important component in many information organization and management tasks. For content-based news video browsing and search, it is desirable to categorize each news video stories extracted in the parsing process described in Section 2.1 into pre-defined and common used news categories. This will provide users a table of content similar to that of printed books to facilitate fast navigation and search. For this, a text-based news categorization algorithm using Support Vector Machines [8] has been developed and integrated into our system.

The key idea behind the SVM based text categorization algorithm is to use a training set of labeled instances to learn the classification function automatically. SVM classifiers resemble a vector of learned feature weights. The resulting classifiers have many advantages: they are easy to construct and update, they depend only on information that is easy for people to provide, and they allow users to smoothly tradeoff precision and recall depending on their task.

In news story categorization process, each news story is represented as a vector of words extracted from the close

caption or transcript of the news, as typically done in information retrieval. To reduce the number of features, we first remove features based on overall frequency counts, and then select a small number of features based on their fit to categories. We used simple linear SVMs because they provide good generalization accuracy and because they are fast to learn. The accuracy of such SVM based categorization is as high as averaging 91.3% for the 10 most frequent categories and 85.5% over all 118 categories [8] when tested using the Reuters collection. Our test using CNN news transcription also achieves over 80 % in accuracy. The categorization errors mainly result from very short transcripts. In our system, the category definition follows the news categories defined on CNN News Website.

Same feature vectors of news stories are used as a query to search in a given news database or from other news web site to extract associated news to a given news stories, using traditional textual information retrieval techniques.

3. ADVANCED VIDEO BROWSERS FOR HOME USERS

Based on meta data of video structure and content extracted using the video parsing and content analysis technologies presented in the last section, we have developed two advanced video browsers for home users.

3.1 Intelligent TV browser

The first one is named *Intelligent TV Browser*, which has VCR-like browsing functions, but support shot/scene based fast, non-linear forward /backward browsing. By using this browser, users can choose to skip commercial shots and locate the shot or story that they are interested in. It also provides the capability to play video highlight based on extracted key-frames.

3.2 Html-based video browser

The other one is a HTML-based video browser that allows video news being viewed simultaneously via web browsing devices. The HTML content of the video news is automatically generated with the structure parsing and text processing processes presented in the last section.

As shown in Figure 5, a user is able to start news video browsing from the news categories under which each news storied is automatically assigned by the text categorization process. Clicking a news item under a category, the summary for this news item will appear in a web page on the left, including the key-frame of the news, together with transcripts of the whole story. A user can play the story by clicking on the key-frame window. Also, the associate stories for this news item are listed on the right, which are extracted by searching from news databases or other web-sites.

Figure 5: Html-based video browser interface

4. CONCLUSION

In this paper, we present a system developed for content-based broadcasted news video browsing for home users. This system integrates the audio-visual as well as text detection and NLP technique analysis to extract structure and content information of news video and to organize and categorize news stories. Based on the component technologies, two advanced browsers, intelligent TV browser and html-based browser are provided for more convenient and intelligent browsing. Future work will be focused on extending the system to performing intelligent filtering and recording using machine learning-based user profiling techniques.

5. REFERENCES

[1] H.-J. Zhang, Y.-H. Gong, S.W. Smoliar and S. Y. Tan. *Automatic parsing of news video*. Proc. of the IEEE International Conference on Multimedia Computing and Systems, 1994. pp. 45-54.

[2] B. Shahraray and D. Gibbon, "Automatic authoring of hypermedia documents of video programs," Proc. of ACM Multimedia'95, San Francisco, November 1995, pp.401-409.

[3] A. G. Hauptmann and M. Smith, "Text, Speech and Vision for Video Segmentation: The Informedia Project", *Working Notes of IJCAI Workshop on Intelligent Multimedia Information Retrieval*, Montreal, August 1995, pp.17-22.

[4] J.S.Boreczky and L.D. Wilcox. *A Hidden Markov Model Frame Work for Video Segmentation Using Audio and Image Features*. Proceedings of ICASSP'98, pp.3741-3744, Seattle, May 1998.

[5] T. Zhang and C.-C. J. Kuo. *Video Content Parsing Based on Combined Audio and Visual Information.* SPIE 1999, Vol.IV, pp. 78-89.

[6] H. Jiang, H.-J. Zhang, Audio content analysis in video structure analysis, Technical Report, Microsoft Research, China.

[7] X.-R. Chen and H.-J. Zhang, *Text Area Detection From Video Frames*, Technical Report, Microsoft Research, China.

[8] S. T. Dumais, J. Platt, D. Heckerman and M. Sahami *Inductive learning algorithms and representations for text categorization*. Proc. of ACM-CIKM98.