Middle-East Journal of Scientific Research 23 (Sensing, Signal Processing and Security): 370-376, 2015
ISSN 1990-9233

DOI: 10.5829/idosi.mejsr.2015.23.ssps.206

Analysis on Video Retrieval Using Speech and Text for Content-Based Information

[1]R. Rajarathinam and [2]R. Latha

[1]Research and Development Centre, Bharathiar University, Coimbatore, 641 046, India
[2]Department of Computer Applications, St. Peter's University, Avadi, Chennai, 600 054, India

Abstract: E-Teaching has become more and more popular in the last previous years. The amount of showing video data on the World Wide Web (WWW) is growing rapidly. Therefore, a more useful method for video retrieval in WWW or within large teaching video archives is urgently needed. This paper presents an analysis and approaches followed for automated video indexing and video search in enormous teaching video archives. First of all, we apply automatic video segmentation and key-frame detection to offer a visual guideline for the video content navigation. Parallel, we extract textual metadata by applying video Optical Character Recognition (OCR) technology on key-frames and Automatic Speech Recognition (ASR) on teaching audio tracks. The OCR and ASR transcript as well as detected slide text line types are adopted for keyword extraction, by which both video- and segment-level key words are extracted for content-based video browsing and search. The ideas are captured with behavioural parameters of proposed indexing functionalities are analysed.

Key words: Teaching videos · Automatic video indexing · Content-based video search · Teaching video archives

INTRODUCTION

DIGITAL video has become a popular storage and exchange medium due to the rapid development in recording technology, improved video compression techniques and high-speed networks in the last few years. Therefore audio visual recordings are used more and more frequently in e-teaching systems. A number of universities and research institutions are taking the opportunity to record their teaching materials and publish them online for students to access independent of time and location. As a result, there has been a huge increase in the amount of multimedia data on the Web. Therefore, for a user it is nearly impossible to find desired videos without a search function within a video archive. Even when the user has found related video data, it is still difficult most of the time for him to judge whether a video is useful by only glancing at the title and other global metadata which are often brief and high level. Moreover, the requested information may be covered in only a few minutes, the user might thus want to find the piece of information user requires without viewing the complete video. The problem becomes how to retrieve the appropriate information in a large teaching video archive more efficiently.

Most of the video retrieval and video search systems in social web sites reply based on available textual metadata such as title, genre, person and brief description, etc. Generally, this kind of metadata has to be created by a human to ensure a high quality, but the creation step is rather less time and cost. Moreover, the manually provided metadata is typically brief, high level and Informative. That is, beyond the current approaches, the next generation of video get back systems apply automatically generated metadata by using video analysis technologies. Many content-based metadata could be generated which may lead to two research questions in the e-teaching context:

- Can those metadata assist the learner in searching required teaching content more efficiently?
- If so, how can we extract the important metadata from teaching videos and provide hints to the user?

According to the questions, we proposed the following:

Corresponding Author: Mrs. R. Rajarathinam, Research and Development Centre, Bharathiar University, Coimbatore, 641 046, India.

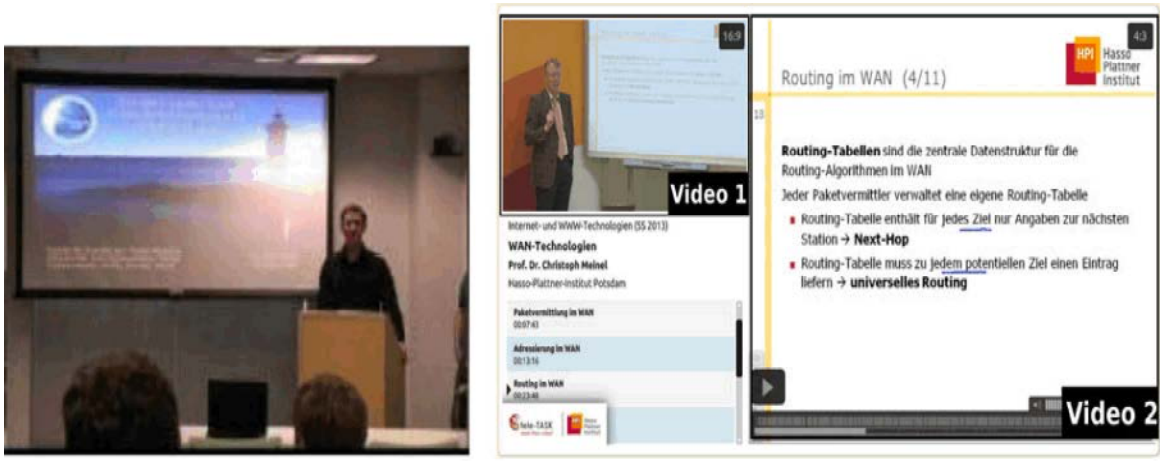

Fig. 1: (a) This shows an example of old type lecture video (b) Exemplary lecture video. Video1 is showing the professor giving his lecture video but presentation is played in Video2.

Propositions

Proposition 1: The relevant metadata can be automatically gathered from teaching videos by using appropriate analysis techniques. They can help a user to find and to understand teaching contents more efficiently and the learning effectiveness canthus be improved. Already used video retrieval based upon visual aid feature extraction cannot be simply applied to teaching recordings because of the homogeneous scene composition of teaching videos.

Fig. 1(a) shows an exemplary teaching video recorded using an outdated format produced by a single video camera. Varying factors may lower the quality of this format.

Example: Motion changes of the camera may affect the size, shape and the brightness of the slide; the slide can be partially obstructed when the speaker moves in front of the slide; any changes of camera focus (switching between the speaker view and the slide view) may also affect the further slide detection process. Nowadays people tend to produce teaching videos by using multi-scenes format (Fig. 1(b)), by which the speaker and his presentation are displayed synchronously.

Proposition 2: This can be achieved either by displaying a single video of the speaker and a synchronized slide file, or by applying a state of the art teaching recording system such as tele-Teaching AnywhereSolution Kit(tele-TASK) [1]. This system suggested by E. Leeuwis, M. Federico and M. Cettolo.

Explanations: Fig. 1(b) illustrates an example of such a system which delivers two main parts of the teaching: the main scene of teaching which is recorded by using a video camera and the second which captures the desktop of the speaker's computer (his presentations) during the teaching through a frame grabber tool. The key benefits of the recent one for a teaching are the flexibility. No added synchronization method between video and slide files is required for indexing and no need of taking care of the slide format.

The main disadvantage is that the video analysis methods may introduce errors. Our research work mainly focuses on those teaching videos produced by using the screen grabbing method. Since two videos are gathered automatically during the recording process. Therefore, the temporal scope of a complete unique slide can be considered as a teaching segment. This way, segmenting two-scenes teaching videos can be achieved by only processing slide video streams, which contain most of the visual text metadata. The extracted slide frames can provide a visual guideline for video content navigation. Text is a high-level semantic feature which has often been used for content-based information retrieval. In teaching videos, texts from teaching slides serve as an outline for the teaching and are very important for understanding. Therefore after segmenting a video file into a list of key frames (all the unique slides with complete contents), the text finding procedure will be executed on each and every key frame and the filtered text objects will be further used in text identifying and slide structure analysis processes.

Importantly, the extracted structural metadata can enable easier video browsing and video search functions. Addressing is one of the most important carriers of information in video teachings. Therefore, it is of distinct advantage that this information can be applied for automatic teaching video indexing. Unfortunately, most of the existing teaching speech recognition systems in the reviewed work cannot achieve a sufficient recognition result, the Word Error Rates having been reported from different analysis are approximately 85%. The poor recognition results not only limit the usability of speech transcript, but also affect the efficiency of the further indexing process.

In our research, we intended to continuously improve the ASR result for German teachings by building new speech training data based on the open-source ASR tool. However, in the open-source context, it lacks method for generating the German phonetic dictionary automatically, which is the one of the most important part of ASR software. Therefore, we developed an automated procedure in order to fill this gap. An enormous amount of textual metadata will be created by using OCR und ASR method, which opens up the content of teaching videos. To enable a reasonable access for the user, the representative keywords are further extracted from the OCR and ASR results. For content-based video search, the search indices are created from different information resources, including manual annotations, OCR and ASR keywords, global metadata, etc.

The rest of the paper is organized as follows: Section 2 reviews related work in teaching video retrieval and content based video search domain. Section 3 describes the automatic video indexing methods. A content-based teaching video search engine using multimodal information resources is introduced in Section 4. Finally, Section 5 concludes the paper with an outlook on future work.

Related Work: The ranked keywords from both segmented video-level can directly be used for video content browsing and video search. Furthermore, the video similarity could be calculated by using the Cosine Similarity Measure based on extracted keywords. Finally, the major contributions of this paper are the following:

- The process of extracting metadata from visual as well as audio resources of teaching videos automatically by applying appropriate analysis techniques. For evaluation purposes we developed several automatic indexing functionalities in a large teaching video portal, which can guide both visually- and text-oriented users to navigate within teaching video. We conducted a user study intended to verify the research hypothesis and to investigate the usability and the effectiveness of proposed video indexing features.
- For visual analysis, we propose a new method for slide video segmentation and apply video OCR to gather text metadata. Furthermore, teaching outline is extracted from OCR transcripts by using stroke width and geometric information. A more flexible search function has been developed based on the structured video text.
- We propose a solution for automatic German phonetic dictionary generation, which fills the gap in open-source ASR domain. The dictionary software and compiled speech corpus are provided for the further research use.
- In order to overcome the solidity and consistency problems of a content-based video search system, we propose a keyword ranking method for multimodal information resources. In order to evaluate the usability, we implemented this approach in a large teaching video portal.
- The developed video analysis methods have been evaluated by using compiled test data sets as well as opened benchmarks. All compiled test sets are publicly available from our website for the further research use.

Teaching Video Retrieval: Information retrieval in the multimedia based learning domain is an active and multidisciplinary research area. Video texts, spoken language, community tagging, manual annotations, video actions, or gestures of speakers can act as the source to open up the content of teachings.

Wang *et al*. proposed an approach for teaching video indexing based on automated video segmentation and OCR analysis. The proposed segmentation algorithm in their work is based on the differential ratio of text and background regions. Using thresholds they attempt to capture the slide transition. The final segmentation results are determined by synchronizing detected slide key-frames and related textbooks, where the text similarity between them was calculated as indicator. Researcher Grcar *et al*. introduced Video Teachings.Net which is a digital archive for multimedia presentations [2]. Similar to the authors also apply a synchronization process between

the recorded teaching video and the slide file, which has to be provided by presenters. Our system contrasts to these two approaches since it directly analyses the video, which is thus independent of any hardware or presentation technology. The constrained slide format and the synchronization with an external document are not required.

Animated Content: Furthermore, since the animated content involvement is often applied in the slide, but has not been considered in analysis of C. Munteanu, G. Penn, R. Baecker and Y. C. Zhang, their system might not work robustly when those effects occur in the teaching video [9]. The final segmentation result is strongly dependent on the quality of the OCR result. It might be less efficient and imply redundancies, when only poor OCR result is obtained. Tuna *et al*. presented their approach for teaching video indexing and search [3]. They segment teaching videos into key frames by using global frame differencing metrics. In [12], the final Segmentation result is strongly dependent on the quality of the OCR result. It might be less efficient and imply redundancies, when only poor OCR result is obtained.

Then standard OCR software is applied for gathering textual metadata from slide streams, in which they utilize some image transformation techniques to improve the OCR result. In our analysis many authors developed a new video player, in which the indexing, search and captioning processes are integrated. Similar to [9], the used global differencing metrics cannot give a sufficient segmentation result when animations or content build-ups are used in the slides. In that case, many redundant segments will be created. Moreover, the used image transformations might be still not efficient enough for recognizing frames with complex content and background distributions. Making use of text detection and segmentation procedures could achieve much better results rather than applying image transformations. Jeong *et al*. proposed a teaching video segmentation method using Scale Invariant Feature Transform (SIFT) feature and the adaptive threshold in one our analysis [4]. In their work SIFT feature is applied to measure slides with similar content. An adaptive threshold selection algorithm is used to detect slide transitions. In their evaluation, this approach achieved promising results for processing one-scene teaching videos as illustrated in Fig. 1(a).Recently, collaborative tagging has become a popular functionality in teaching video portals. Sack Moritz *et al*. apply tagging data for teaching video retrieval and video search [13]. Beyond the keyword based tagging, J. Glass, T. J. Hazen, L. Hetherington and C. Wang proposed an approach to annotate teaching video resources by using Linked Data [6]. Their framework enables users to semantically annotate videos using vocabularies defined in the Linked Data cloud. Then those semantically linked educational resources are further adopted in the video browsing and video recommendation procedures. However, the effort and cost needed by the user annotation-based approach cannot satisfy the requirements for processing large amounts of web video data with a rapid increasing speed. Here, the automatic analysis is no doubt much more suitable. Nevertheless, using Linked Data to further automatically annotate the extracted textual metadata opens a future research direction.ASR provides speech-to-text information on spoken languages, which is thus well suited for content-based teaching video retrieval.

The studies described in the analysis are based on out-of-the-box commercial speech recognition software. Concerning such commercial software, to achieve satisfying results for a special working domain an adaption process is often required, but the custom extension is rarely possible. The authors of [1] and [8] focus on English speech recognition for Technology Entertainment and Design (TED) teaching videos and webcasts. In their system, the training dictionary is created manually, which is thus hard to be Extended or optimized periodically.

Glass *et al*. proposed a solution for improving ASR results of English teachings by collecting new speech data from the rough teaching audio data. Inspired by their work, we developed an approach for creating speech data from German teaching videos. Haubold and Kender focus on multi-speaker presentation videos. In their work speaker changes can be detected by applying a speech analysis method [7]. Overall, most of those lecture speech recognition systems have low recognition rate, the WERs of audio lectures are approximately 40-85 percent. The poor recognition results limit the further indexing efficiency. Therefore, how to continuously improve ASR accuracy for lecture videos is still an unsolved problem. The speaker-gestures-based information retrieval for lecture videos has been studied. The author equipped the lecture speaker with special gloves that enable the automatic detection and evaluation of gestures. The experimental results show that 12 percent of the lecture topic boundaries were correctly detected using speaker-gestures. However, those gesture features are highly dependent on the characteristics of speakers and topics. It might have limited use in large lecture video archives with massive amount of speakers.

Content-Based Video Search: Several content-based video search engines have been proposed recently. D. Lee *et al*. proposed a lecture web cast search system, in which they applied a slide frame segmented to extract lecture slide images [5]. The system retrieved more than 25,000 lecture videos from different resources such as social websites, etc. These search indices are created based on the global metadata obtained from the video hosting website and texts extracted from slide videos by using a standard OCR engine. Since they do not apply text detection and text segmentation process, the OCR recognition accuracy of their approach is therefore lower than our systems.

Furthermore, by applying the text detection process we are able to extract the structured text line such as title, subtitle, key-point, etc., that enables a more flexible Search function. In the previous projects, a content based semantic multimedia retrieval system has been developed. After the digitization of media data, several analysis techniques, e.g., OCR, ASR, video segmentation, automated speaker recognition, etc., have been applied for metadata generation. An entity recognition algorithm and an open knowledge base are used to extract entities from the textual metadata. As mentioned before, searching through the recognition results with a degree of confidence, we have to deal with the solidity and the consistency problem. However the reviewed content-based video search systems did not consider this issue. For content-based video search, the search indices are created from different information resources, including manual annotations, OCR and ASR keywords, global metadata, etc.

Here the varying recognition accuracy of different analysis engines might result in solidity and consistency problems, which have not been considered in the most related work. Therefore, they propose a new method for ranking keywords extracted from various information resources by using the extended Term Frequency Inverse Document Frequency (TFIDF) score [10]. The ranked keywords from both segment and video-level can directly be used for video content browsing and video search. Furthermore, the video similarity can be calculated by using the Cosine Similarity Measure [11] Based on extracted keywords.

Automated Lecture Video Indexing: In this chapter we will present four analysis processes for retrieving relevant metadata from the two main parts of lecture video, namely the visual screen and audio tracks. From the visual screen we firstly detect the slide transitions and extract each unique slide frame with its temporal scope considered as the video segment. Then the video OCR analysis is performed for retrieving textual metadata from slide frames. Based on OCR results, we propose a novel solution for lecture outline extraction by using stroke width and geometric information of detected text lines. In speech-to-text analysis we applied the open-source ASR software to build the acoustic and language model, the author collected speech training data from open-source corpora and our lecture videos. As already mentioned, it lacks method in open source context for creating German phonetic dictionary automatically. We thus developed a solution to fill this gap and made it available for the further research use.

Slide Video Segmentation: Video browsing can be achieved by segmenting video into representative key frames. The selected key frames can provide a visual guideline for navigation in the lecture video portal. Moreover, video segmentation and key-frame selection is also often adopted as a pre-processing for other analysis tasks such as video OCR, visual concept detection, etc. Choosing a sufficient segmentation method is based on the definition of "video segment" and usually depends on the genre of the video. In the lecture video domain, the video sequence of an individual lecture topic or subtopic is often considered as a video segment. This can be roughly determined by analysing the temporal scope of lecture Slides.

Many approaches make use of global pixel-level-differencing metrics for capturing slide transitions. A drawback of this kind of approach is that the salt and pepper noise of video signal can affect the segmentation Accuracy. After observing the content of lecture slides, we realize that the major content as, e.g., text lines, figures, tables, etc., can be considered as Connected Components (CCs).

We therefore propose to use CC instead of pixel as the basis element for the differencing analysis. We call it component-level-differencing metric. This way we are able to control the valid size of the CC, so that the salt and pepper noise can be rejected from the differencing process. For creating CCs from binary images algorithm demonstrated an excellent performance advantage. Another benefit of our segmentation method is its robustness to animated content progressive build-ups used within lecture slides.

Only the most complete unique slides are captured as video segments. Those effects affect the most lecture video segmentation methods mentioned in chapter 2.

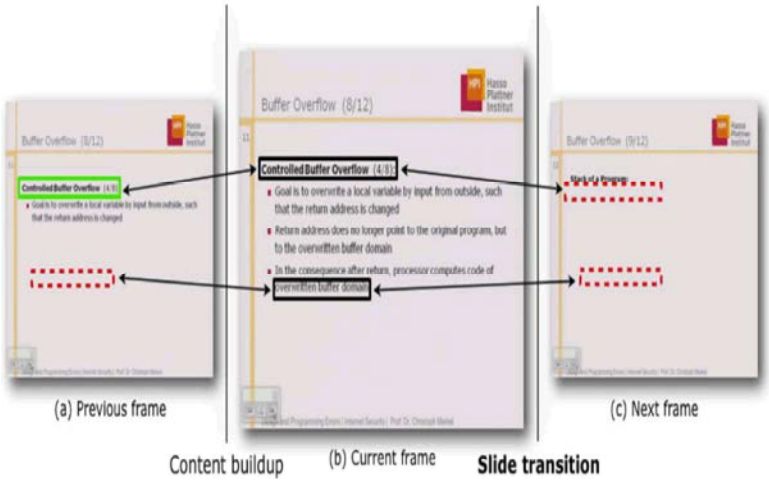

Fig. 2: In frame (a) and (b), a same text line (top) can be found; whereas in frame (b) and (c), the CC-based differences of both text lines exceed the threshold Ts2.A slide transition is thus found between frame (b) and (c).

Fig. 3: (a) Exemplary text detection results (b) Text binarization results

Fig. 4: Visualisation of separated lecture slides for video elements

Video Content Browsing and Video Search: We can do a browsing with lecture key frames and extra lecture outline used for video portal. Figure 4 shows the visualisation of key frames in our lecture video portal. Slide gallery has been provided underneath of the video elements, while the user would like to read the content clearly. Video would be navigated right from the starting of the by clicking on the timeline elements and thumbnails.

CONCLUSION

In this paper we analysed an approach for content based video lecture indexing and retrieval in large lecture video archives. Research hypothesis is verified. Further works are ASR for other language, videos and keyword extraction has to be done as analysis.

REFERENCES

1. Leeuwis, E., M. Federico and M. Cettolo, 2003. "Language modelling and transcription of the ted corpus lectures," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process.
2. Grcar, M., D. Mladenic and P. Kese, "Semi-automatic categorization of videos on videolectures. net," in Proc. Eur. Conf. Mach. Learn. Knowl.
3. Tuna, T., J. Subhlok, L. Barker, V. Varghese, O. Johnson and S. Shah, 2012. "Development and evaluation of indexed captioned searchable videos for stem coursework," in Proc. 43rd ACM Tech. Symp. Comput. Sci. Education.
4. Jeong, H.J., T.E. Kim and M.H. Kim, 2012. "An accurate lecture video segmentation method by using sift and adaptive threshold," in Proc. 10th Int. Conf. Advances Mobile Comput.
5. Lee, D. and G.G. Lee, 2008. "A korean spoken document retrieval system for lecture search," in Proc. ACM Special Interest Group Inf. Retrieval Searching Spontaneous Conversational Speech Workshop.
6. Glass, J., T.J. Hazen, L. Hetherington and C. Wang, 2004. "Analysis and processing of lecture audio data: Preliminary investigations," in Proc. HLT-NAACL Workshop Interdisciplinary Approaches Speech Indexing Retrieval.
7. Haubold, A. and J.R. Kender, "Augmented segmentation and visualization for presentation videos" in Proc. 13th Annu, ACM Int. Conf. Multimedia
8. Curst, W., T. Kreuzer and M. Wiesenh Cutter, "A qualitative study towards using large vocabulary automatic speech recognition to index recorded presentations for search and access over the web," in Proceedings.
9. Munteanu, C., G. Penn, R. Baecker and Y.C. Zhang, 2006. "Automatic speech recognition for webcasts: How good is good enough and what to do when it isn't," in Proc. 8th Int. Conf. Multimodal Interfaces.
10. Salton, G. and C. Buckley, "Term-weighting approaches in automatic text retrieval," Inf. Process.
11. Salton, G., A. Wong and C.S. Yang, (Nov. 1975). A vector space model for automatic indexing, Commun. ACM Journal.
12. Pong, T.C., F. Wang and C.W. Ngo, "Structuring low-quality videotaped lectures for cross-reference browsing by video text analysis," J. Pattern Recog.
13. Sack, H. and J. Waitelonis, 2006. "Integrating social tagging and document annotation for content-based search in multimedia data," in Proc. 1st Semantic Authoring Annotation Workshop.