International Journal for Rapid Research in Engineering Technology and Applied Science
ISSN (Online): 2455-4723
Vol 2 Issue 1 February 2016

Pacify based Video Retrieval System

MrRishikesh S Patil[1], Prof. Chhaya Nayak[2]

1. M.Tech Student, Department of Computer engineering, B.M Technology, Indore-MP

2. Head of Department, Department of Computer engineering, B.M Technology,Indore-MP

R.G.P.V, Madhya Pradesh-India

Abstract

Video is becoming a prevalent medium for e-learning. Lecture videos contain text information in both the visual and aural channels: the presentation slides and lecturer's speech. This paper examines the relative utility of automatically recovered text from these sources for lecture video retrieval. To extract the visual information, we apply video content analysis to detect slides and optical character recognition to obtain their text. We extract textual metadata by applying video Optical Character Recognition (OCR) technology on key-frames and Automatic Speech Recognition (ASR) on lecture 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 keywords are extracted for content-based video browsing and search.

Index Terms—*Lecture videos, automatic video indexing, content-based video search, lecture video archives*

I. Introduction

Digital video has become a popular and Storage medium of exchange because of the rapid development in recording technology, video compression techniques improved and broadband networks in recent years [1].To arrive at identification accuracy that is acceptable for retrieval given the difficult conditions, the different parts of our ASR system must be made as robust as possible so that it is able to cope with those problems that typically emerge when technology is transferred from the lab and applied in a real life context. This is in accordance with our overall research agenda the development of robust. ASR technology that can be ported to different topic domains with a minimum effort [2].

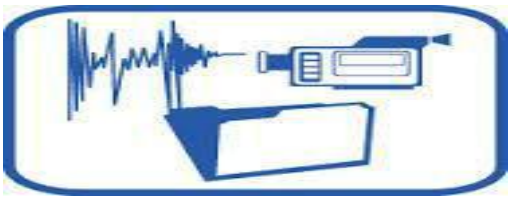


Fig 1.1 Paciy Based Video retrieval

Video is complex data type – audio & video Audio can be handled by query by humming. Voice recognition system using tree like structure to construct all possible substrings of a sentence. Audio is categorized by: speech, music, and sound. Audio retrieval methods: Hidden Markov Model, Boolean Search with multi-query using Fuzzy Logic.

E.g. in our day to day life when we interact with the Google Video Archive selections

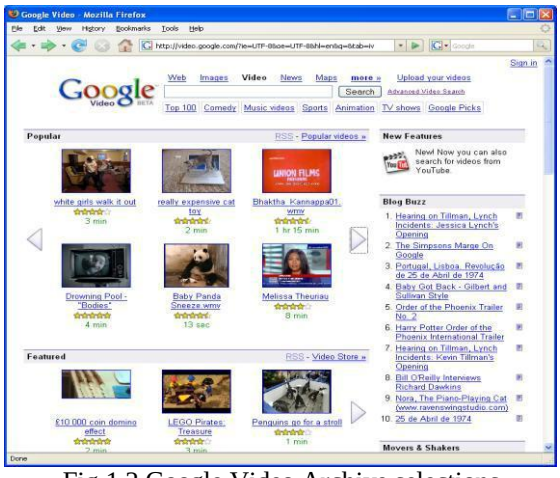


Fig 1.2 Google Video Archive selections

Challenges in this domain are as follows:

- There is an amazing growth in the amount of digital video data in recent years.
- Lack of tools for classify and retrieve video content
- There exists a gap between low-level features and high-level semantic content.
- To let machine understand video is important and challenging.

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 here quires without viewing the complete video. The problem becomes how to retrieve the appropriate information in a large lecture video archive more efficiently.

II. Survey of Literature

Wang et al. proposed an approach for indexing video conferencing based video segmentation and analysis automated OCR, segmentation algorithm proposed in his work is based on the differential relation of the regions of text and background. The use of thresholds in their attempt to capture the slide transition. The final results of segmentation are determined by the timing of slides detected key-frames and related textbooks, when the text similarity between them was calculated as an indicator. Grcar et al. VideoLectures.net introduced in [7], which is a digital file for multimedia presentations. As in [6], the authors also apply a synchronization between video taped lecture and slide file, which must be provided by the presenters. our system In contrast with these two approaches because it directly analyzes the video, which is independent of any hardware or presentation technology. No limited format slides and synchronization with an external document is required. Moreover, since the content animated evolution is often applied on the slide, but has not been considered in [6] and [7], your system may not work firmly when these effects occur in the video conference. In [6], the final Segmentation result is strongly dependent on the quality of OCR results. It might be less efficient and involve layoffs, when only poor OCR result is obtained .Tuna et al. presented its approach to video indexing and searching conference [8]. They conference video segment key frames using global metrics framework differentiation. Then OCR software standard applies to the collection of textual metadata streams slides, in which some image processing techniques are used to improve the recognition result. They developed a new video player, in which the processes of indexing, searching and subtitling are integrated. As in [6], differentiation global metrics used can not give a sufficient result of segmentation when animations or Accumulations content used on slides. In this case, segments are created many redundant. Moreover, the image transformations could be used not yet efficient enough for recognizing frames with complex background content and distributions. Using methods of detecting and segmenting text could achieve much better results in the application of image transformations. Jeong et al. proposed proposed a method of segmentation using video conference Scale invariant functions Transform (SIFT) feature and adaptive threshold [9]. In its role as SIFT work applied to measure slide with similar content. An adaptive selection algorithm is used to detect threshold slide transitions. In its evaluation, this approach has achieved promising results. Sack and Waitelonis and Moritz et al applying the tagging data recovery video conferencing and video search. Beyond the keyword based tagging, Yu et al. proposed an approach to annotate resources to video conference using linked data. Its framework allows users to semantically annotate videos using vocabularies defined in the Linked Data cloud. Then these educational resources related semantically more adopted in video browsing and video recommendation Procedures. However, the effort and cost required for annotation-based approach can not satisfy user requirements for processing large amounts of data from web video with a fast growing speed. Here, the automatic analysis is undoubtedly much more appropriate. However, using Linked Data for more automatically annotate textual metadata extracted opens a line of future research. ASR provides voice information into text into spoken languages, it is therefore very suitable for video retrieval based conference content. In [4]and[12] are Based on recognition software trade-of-the-box voice. As for this type of commercial software, to achieve satisfactory results for a particular domain of work often a process of adaptation is required, but the custom extension is rarely possible. The authors of [2] and [5] focus on speech recognition dictionary for Technology Entertainment and Design (TED) conference videos and webcasts. In his system, the training dictionary is created manually, which is therefore difficult to be periodically extended or optimized. Glass et al. proposed a solution to improve the performance of

ASR conference in English by collecting new data from voice to data from raw audio conference [3]. Inspired by their work, we have developed an approach for creating voice data Germans videos conferences Haubold and Kender videos focus on multi-speaker presentation. In his work changes speakers can be detected by applying a method of analysis of speech. An algorithm for extracting topic sentences ASR highly imperfect results (WER = 75 percent) has been proposed for use during lectures sources such as textbooks and files related slides. In general, most of these systems conference speech recognition have a low rate of recognition, audio conferencing WERS are approximately 40-85 percent. Poor recognition results indexing limit efficiency more.

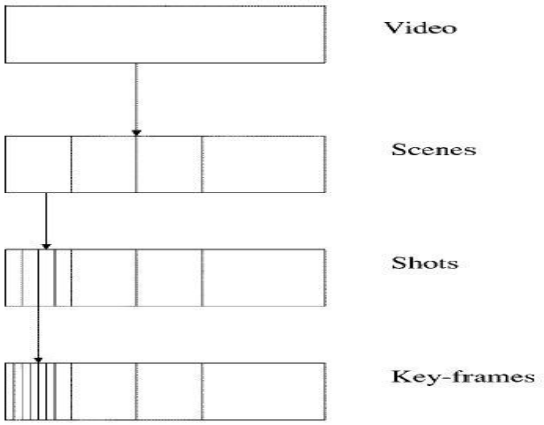
Therefore, how to continuously improve the accuracy ASR lecture videos is still an unsolved problem Information retrieval speaker-gesture-based video conferencing has been studied in [13]. The author tells the speaker conference with special gloves that allow automatic detection and evaluation of gestures. Experimental results show that 12 percent of the limits detected correctly conference theme with speakers gestures. However, these gesture features are highly dependent on the characteristics of the speakers and topics. It may be of limited use in large video files with massive amounts conference speaker.

III. Analysis of our work

A) Techniques in Video Data Management

Video Parsing
Manipulation of whole video for breakdown into key frames.
Video Indexing
Retrieving information about the frame for indexing in a database.
Video Retrieval and Browsing
Users access the db through queries or through interactions.

In video parsing steps are as follows



Text is a high-level semantic feature which has often been used for content-based information retrieval. In lecture videos, texts from lecture slides serve as an outline for the lecture and are very important for understanding [3]. Therefore after segmenting a video file into a set of key frames the text detection procedure will be executed on each key frame, and the extracted text objects will be further used in text recognition and slide structure analysis processes. Especially, the extracted structural metadata can enable more flexible video browsing and video search functions. Speech is one of the most important carriers of information in video lectures. Therefore, it is of distinct advantage that this information can be applied for automatic lecture video indexing. A large amount of textual metadata will be created by using OCR und ASR method, which opens up the content of lecture 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.

B) Video Retrieval Using Speech Recognition

- Speech recognition is the process of converting an acoustic signal, captured by a microphone or a telephone, to a set of words.
- The recognized words can be an end in themselves, as for applications such as commands & control, data entry, and document preparation.
- They can also serve as the input to further linguistic processing in order to achieve speech understanding [4].

Video Retrieval Using Speech Recognition contains following steps

- Signal processing
- Speech recognition
- Semantic interpretation
- Dialog Management
- Response Generation
- Speech synthesis (Text to Speech)

A speech decoder will always try to map a sound segment to a sequence of words, processing non-speech portions of the videos (i) would be a waste of processor time, (ii) introduces noise in the transcripts due to assigning word labels to non-speech fragments, and (iii) reduces speech recognition accuracy in general when the output of the first recognition run is used for acoustic model adaptation purposes.

In our proposed system we will concentrate on the following points:

- A valid text line object must have more than three characters,
- A valid text line object must contain at least one noun
- The textual character count of a valid text line object must be more than 50 percent of the entire string length.

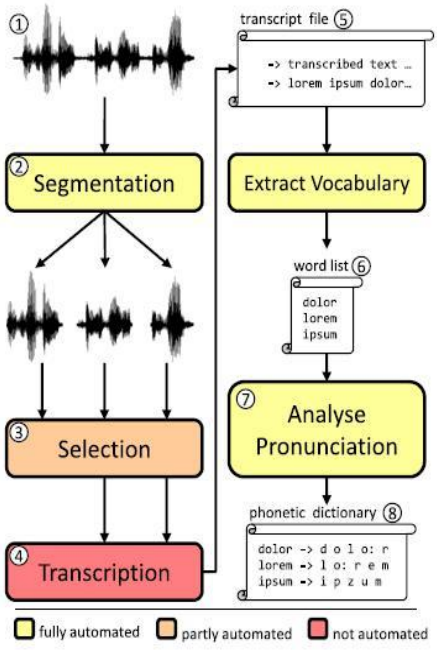


Fig 2.2 Work flow of proposed System &Video Parsing

General proposed architecture for the proposed system will be as follows:

Therefore we need to focus on the speech recognition for retrieving the contents of video which can depict as follows:

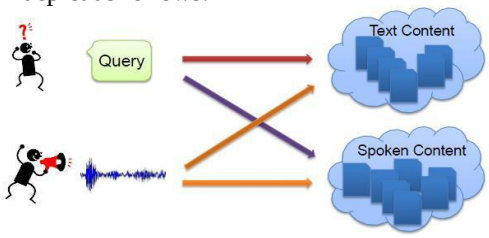


Fig 2.3 Video Content retrieval using Speech Recognition

C) 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 preprocessing 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. In the first step, the entire slide video is analyzed. We try to capture every knowledge change between adjacent frames, for which we established an analysis interval of three seconds by taking both accuracy and efficiency into account[8]. This means that segments with duration smaller than three seconds may be discarded in our

system. Since there are very few topic segments shorter than three seconds, this setting is therefore not critical. Then we create canny edge maps for adjacent frames and build the pixel differential image from the edge maps. The CC analysis is subsequently performed on this differential image and the number of CCs is then used as a threshold for the segmentation.

IV. Conclusion

The first conclusion is that the slide text and spoken text are not the same. Comparison of the ground truth and automatic transcripts reveal substantial differences in the content and volume of slide and spoken text. The overlap is limited even when controlling for recognition errors with manual transcripts. Issuing term queries that are common to the SLIDE and SPOKEN ground truth retrieve different videos among the results using both manual and automatic text search indexes. Secondly, both manually and automatically extracted slide text exhibit greater retrieval precision when compared to manually and automatically transcribe spoken text. We attribute this result to two causes. First, the usage of terms in slides is the product of a deliberate authoring process, while speech is often partially improvised. Less descriptive terms are more common in speech, and in turn more commonly shared with other videos spoken transcripts. This imprecision limits the discriminative power of spoken text for video retrieval. The second factor is the differing recognition error profiles of ASR and OCR. Errors are more frequent in OCR,but occur at the character level producing non-dictionary terms in the transcripts. These errors do not degrade text-based retrieval, since they do not appear as queries. Errors in ASR occur at the word level due to phonetic and out of vocabulary mismatch. The resulting inserted terms tend to be dictionary words that appear in both other video transcripts and search queries. Automated annotation for OCR and ASR results using Linked Open Data resources offers the opportunity to enhance the amount of linked educational resources significantly. Therefore more efficient search and recommendation method could be developed in lecture video archives.

V. References

[1] Haojin Yang and Christoph Meinel, Member, *IEEE* ―Content Based Lecture Video Retrieval Using Speech and Video Text Information‖ *IEEE* TRANSACTIONS ON LEARNING TECHNOLOGIES, VOL. 7, NO. 2, APRIL-JUNE 2014.

[2] E. Leeuwis, M. Federico, and M. Cettolo, ―Language modelingand transcription of the ted corpus lectures,‖ in Proc. *IEEE* Int.Conf. Acoust., Speech Signal Process., 2003, pp. 232–235.

[3] J. Glass, T. J. Hazen, L. Hetherington, and C. Wang, ―Analysis and processing of lecture audio data: Preliminary investigations,‖ in Proc. HLT-NAACL Workshop Interdisciplinary Approaches Speech Indexing Retrieval, 2004, pp. 9–12.

[4] W. H€urst, T. Kreuzer, and M. Wiesenh€utter, ―A qualitative studytowards using large vocabulary automatic speech recognition to index recorded presentations for search and access over the web,‖in Proc. IADIS Int. Conf. WWW/Internet, 2002, pp. 135–143.

[5] C. Munteanu, G. Penn, R. Baecker, and Y. C. Zhang, ―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,2006.

[6] T.-C. Pong, F. Wang, and C.-W. Ngo, ―Structuring low-quality videotaped lectures for cross-reference browsing by video text analysis,‖ J. Pattern Recog., vol. 41, no. 10, pp. 3257–3269, 2008.

[7] M. Grcar, D. Mladenic, and P. Kese, ―Semi-automatic categorization of videos on videolectures.net,‖ in Proc. Eur. Conf. Mach.Learn. Knowl. Discovery Databases, 2009, pp. 730–733.

[8] T. Tuna, 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. Educ., pp. 129–134. [Online].Available:http://doi.acm.org/10.1145/2157136.2157177.

[9]] H. Sack and J. Waitelonis, ―Integrating social tagging and document annotation for content-based search in multimedia data,‖ in Proc. 1st Semantic Authoring Annotation Workshop, 2006.

[10] H. Sack and J. Waitelonis, ―Integrating social tagging and document annotation for content-based search in multimedia data,‖ in Proc. 1st Semantic Authoring Annotation Workshop, 2006.

[11] C. Meinel, F. Moritz, and M. Siebert, ―Community tagging in tele-teaching environments,‖ in Proc. 2nd Int. Conf. e-Educ., e-Bus.,e-Manage. and E-Learn., 2011.

[12] S. Repp, A. Gross, and C. Meinel, ―Browsing within lecture videosbased on the chain index of speech transcription,‖ *IEEE* Trans.Learn. Technol., vol. 1, no. 3, pp. 145–156, Jul. 2008.

[13] J. Eisenstein, R. Barzilay, and R. Davis. (2007). ―Turning lecturesinto comic books using linguistically salient gestures,‖ in Proc.22nd Nat. Conf. Artif. Intell., 1, pp. 877–882. [Online].

Available:http://dl.acm.org/citation.cfm?id=1619645.1619786.

[14] J. Adcock, M. Cooper, L. Denoue, and H. Pirsiavash, ―Talkminer: A lecture webcast search engine,‖ in Proc. ACM Int. Conf. Multimedia,2010, pp. 241–250.

[15] J. Nandzik, B. Litz, N. Flores Herr, A. L€ohden, I. Konya, D.Baum, A. Bergholz, D. Sch€onfuß, C. Fey, J. Osterhoff, J.Waitelonis,H. Sack, R. K€ohler, and P. Ndjiki-Nya. (2012) ― Contentus—technologies for next generation multimedia libraries, MultimediaTools Appl., pp. 1–43, [Online]. Available:http://dx.doi.org/10.1007/s11042-011-0971-2.