ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 9, September 2015

An Efficient Method to Retrieve Content Based Lecture Video Using Speech and Video Text Information

Rupali Khollam[1], Prof. Pratap Singh Sir[2]

M.E Student, Dept. of Computer Engineering, Institute of Knowledge COE, Savitribai Phule Pune University, Pimpale, Jagtap, Shirur, Pune, India

Professor, Dept. of Computer Engineering, Institute of Knowledge COE, Savitribai Phule Pune University, Pimpale, Jagtap, Shirur, Pune, India

ABSTRACT: Recording video lectures and putting them on the Web for access by students is a popular trend at various universities and institutions. Therefore the amount of lecture video data is growing rapidly on the World Wide Web (WWW). The video contains text information in the form of visual (the presentation slides) as well as audio channels (lecturer's speech). So it becomes a need for an efficient method for video retrieval in WWW or within large lecture video archives. To extract the visual information, I apply video content analysis to detect slides and (OCR)Optical Character Recognition to obtain text from them and (ASR) Automatic Speech Recognition is used to extract spoken text from the recorded audio. In this paper I present an approach for automated video indexing and video search in large lecture video archives. Firstly I apply automatic video segmentation and key-frame detection. Then I apply video Optical Character Recognition (OCR) technology on key-frames to extract textual metadata and Automatic Speech Recognition (ASR) on lecture audio tracks.

KEYWORDS: Lecture videos, automatic video indexing, content-based video search, lecture video archives.

I. INTRODUCTION

In the last few years software's for recording technology, improved video compression techniques and high-speed networks are gaining increasing popularity especially at universities. Because of this many universities capture and record live presentations of lectures and put them on net (WWW) for access by the students. It makes presentation video libraries easily accessible and would allow for more efficient retrieval of specific presentation or speaker video clips. E-lecturing is used so that students would be able to quickly access and review their required presentations independent of location and time. It causes a huge increase in the amount of multimedia data on the Web. Due to this it becomes nearly impossible to find desired videos without a search function within a video archive. Also when the user found related video data; it is still difficult for him to decide whether a video is useful for him by only glancing at the title and other global metadata which are brief. Text is a high-level semantic feature used for the content-based information retrieval. Speech is one of the important carriers of information in video lectures. In this paper we have proposed an efficient method to retrieve content based video within a video achieve. To extract text from a video file firstly the algorithm implement segmenting a video file into a set of key frames (all the unique slides with complete contents)then the text detection procedure is executed on each key frame. The extracted text objects are further used in text recognition and slide structure analysis processes. Text are extracted from video file by using OCR (Optical Character Recognition) and from audio tracks by using ASR(Automatic Speech Recognition). For ranking the keywords extracted from various information resources algorithm uses the extended Term Frequency Inverse Document Frequency. The ranked keywords further used for video browsing and video search.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 9, September 2015

II. RELATED WORK

Author [9] proposed an approach for lecture video indexing based on automated video segmentation and OCR analysis. The proposed segmentation algorithm is based on the differential ratio of text and background regions. Final segmentation results are determined by synchronizing detected slide key-frames and related text books. The authors[10]also apply a synchronization process between the recorded lecture video and the slide file, which has to be provided by presenters. The animated content involvement applied in the slide, but has not been considered in [9] and [10], their system might not work robustly when those effects occur in the lecture video. In [9], the final segmentation result is strongly dependent on the quality of the OCR result. It might be less efficient and imply redundancies, when poor OCR result is obtained.[11] presented an approach for lecture video indexing and search. They segment lecture videos into key frames by using global frame differencing metrics and consider some image transformation techniques to improve the OCR result. They developed a new video player, in which the indexing, search and captioning processes are integrated. Though, the used global differencing metrics cannot give a sufficient segmentation result when animations or content build-ups are used in the slides. Many redundant segments will be created. Also the used image transformations were not efficient enough for recognizing frames with complex content and background distributions. Authors [13],[14] apply tagging data for lecture video retrieval and video search. Yu et al. presented an approach to annotate lecture video resources by using Linked Data. These 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. The authors of [1] and [6] focus on English speech recognition for Technology Entertainment and Design (TED) lecture videos and webcasts. In their system, the training dictionary is created manually, which is hard to extend or optimize periodically. Author [3] proposed a solution for improving ASR results of English lectures by collecting new speech data from the rough lecture audio data. The author in[17] applied a slide frame segmented to extract lecture slide images. But they do not apply text detection and text segmentation process; hence the OCR recognition accuracy of their approach is lower. Our proposed system is independent of any hardware or presentation technology. We use keyword ranking method for multimodal information resources and propose a solution to create a German Phonetic Dictionary to improve ASR results. Also by applying text detection process, able to extract structured text lines such as title, subtitle, key point ,etc. which enables more flexible search function. .Overall ,most of the lecture speech recognition system have low recognition rate, word error rate and also result in solidity and consistency problems.

III. PROPOSED SYSTEM

A. Description of the Proposed Algorithm:

In the proposed system we extract metadata from visual as well as audio resources of lecture videos automatically by applying appropriate analysis techniques. For visual analysis, it propose a new method for slide video segmentation and apply video OCR to gather text Metadata. Lecture outline is extracted from OCR transcripts by using stroke width and geometric information. Subsequently, it will 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 and 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. 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.

a. Slide Video Segmentation

Video browsing is achieved by segmenting video into key frames. The selected key frames provide a visual guideline for navigation in the lecture video portal. Also, video segmentation and key-frame selection is often adopted as a preprocessing for other analysis tasks such as video OCR, visual concept detection, etc.

b. Adaptive Binarization Of key frames

Key frame extraction is an important component of content-based video retrieval and directly influences on the efficiency of video retrieval. Nowadays some problems are existed in the algorithms of key-frame extraction, such as

features selected singly, choosing threshold value difficultly and so on. This paper proposes a new key-frame extraction method which based on adaptive threshold detection of multi-features. First, two descriptors of color histogram, edge histogram of the adjacent scale wavelet transform are used to describe visual content, and combine to form a frame difference measure of multi-feature integration. Then key-frames are extracted by the adaptive threshold detection. Finally, experimental results show that the proposed algorithm is efficient.

c. OCR

For retrieving textual data from the video we developed a novel video OCR system for gathering video text. Optical character recognition (OCR) is the mechanical or electronic conversion of images of typewritten or printed text into machine-encoded text. In the detection stage, an edge-based multi-scale text detector is used to quickly localize candidate text regions with a low rejection rate. For the text area verification, an image entropy-based adaptive refinement algorithm used which splits the most text- and non-text-regions into separate blocks. Then we apply Stroke Width Transform (SWT) based verification procedures to remove the non-text blocks. For removing the spelling mistakes resulted by the OCR engine, we perform a dictionary-based filtering process. We use term frequency inverse document frequency for ranking keywords. Ranked keywords used for video content browsing and video search. Video Similarity calculated by using cosine similarity measure based on extracted keywords. We defined a new formula for calculating TFIDF score, as shown by Eq. (1):

$$Tfidf_{seg\text{-}internal}(k\omega)= 1/N(tfidf_{ocr}.1/n_{type}) \ldots\ldots(1)$$

Where kw is the current keyword, $tfidf_{ocr}$ and $tfidf_{asr}$ denote its TFIDF score computed from OCR and ASR resource respectively, w is the weighting factor for various resources, n type denotes the number of various OCR text line types. N is the number of available information resources, in which the current keyword can be found, namely the corresponding TFIDF score does not equal 0.Since OCR text lines are classified into four types in our system we can calculate the corresponding weighting factor for each type and for each information Resource by using their confidence score. Eq. (2) depicts the formula

$$\omega_i=\mu/\sigma_i \ (i=1\ldots n) \ldots\ldots(2)$$

Where the parameter m is set to equal 1 in our system and can be calculated by using the corresponding recognition accuracy of the analysis engine, as shown by Eq. (3)

$$\sigma_i=1\text{-}Accuracy_i \ (i=1\ldots n)\ldots\ldots (3)$$

d. ASR

Texts from the audio data are extracted by using ASR. In computer science and electrical engineering, speech recognition (SR) is the translation of spoken words into text. First we extract the audio data from video files. Then, we transcribed audio recordings using an automatic speech recognition (ASR) engine. First, the recorded audio file is segmented into smaller pieces and improper segments are sorted out. For each remaining segment the spoken text is transcribed manually, and added to the transcript file automatically. As an intermediate step, a list of all used words in the transcript file is created. In order to obtain the phonetic dictionary, the pronunciation of each word has to be represented phonetically.

e. Creation Of German Phonetic Dictionary

The phonetic dictionary is an essential part of every ASR software. For each word that appears in the transcripts, it defines one or more phonetic representations. As our speech corpus is growing continuously, the extension and maintenance of the dictionary becomes a common task. An automatic generator is highly desired. Unfortunately it lacks such a tool in the open-source context. Therefore, we have built a phonetics generator by using a customized phonetic alphabet, which contains 45 phonemes used in German pronunciation. We provide this tool and the compiled speech training data for the further research use.

B. Proposed Algorithm

1. Take input video
2. Extract frames from video
3. For all frames in video do
4. Perform Segmentation on each frame
5. Apply binarization on each frame

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

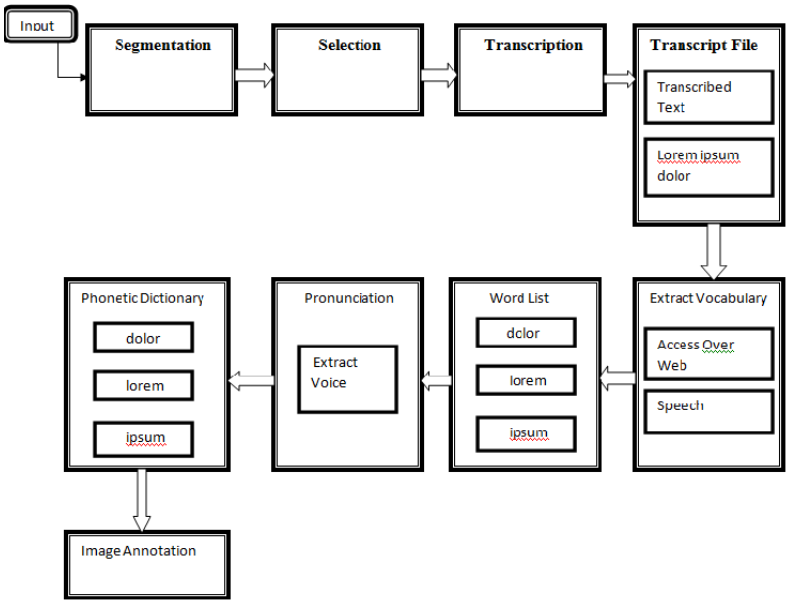
International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 9, September 2015

6. Detecting image features like resolution and inversion.
7. Lines detection and removing.
8. Page layout analysis.
9. Detection of text lines and words.
10. Recognition of characters using OCR.
11. Input Speech.
12. Extract transcript file and compare each audio slide using ASR.
13. Do searching and indexing based on results of OCR and ASR

C. System Architecture



IV. SIMULATION RESULTS

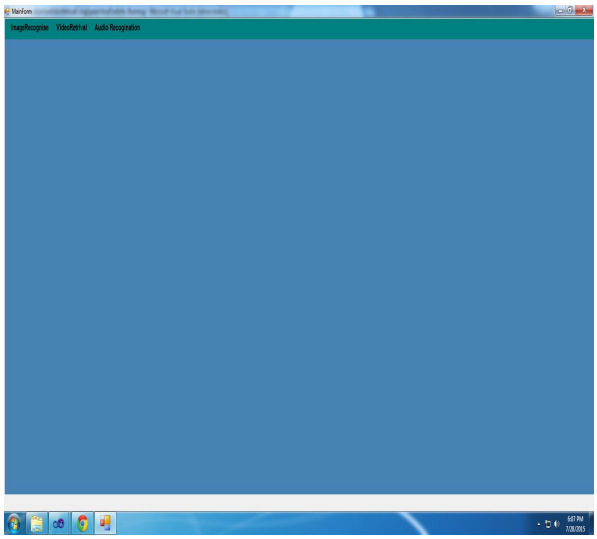


Fig 1.Home Page

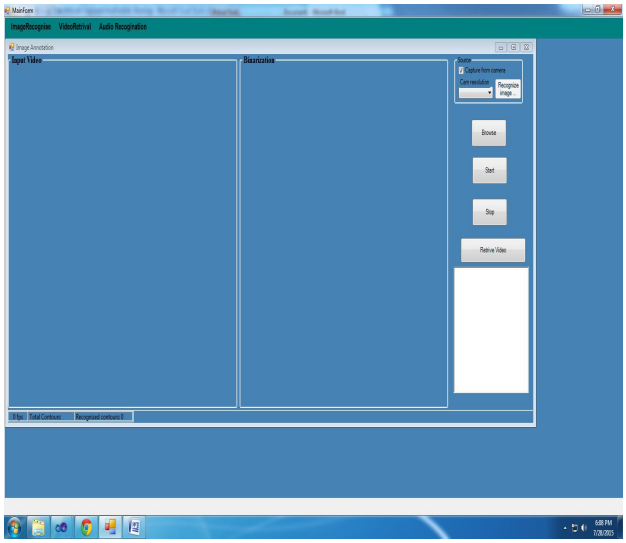


Fig 2.Video Retrieval

Fig.2 gives video retrieval-User browse for a video. It gives binary images of frames from the video.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 9, September 2015

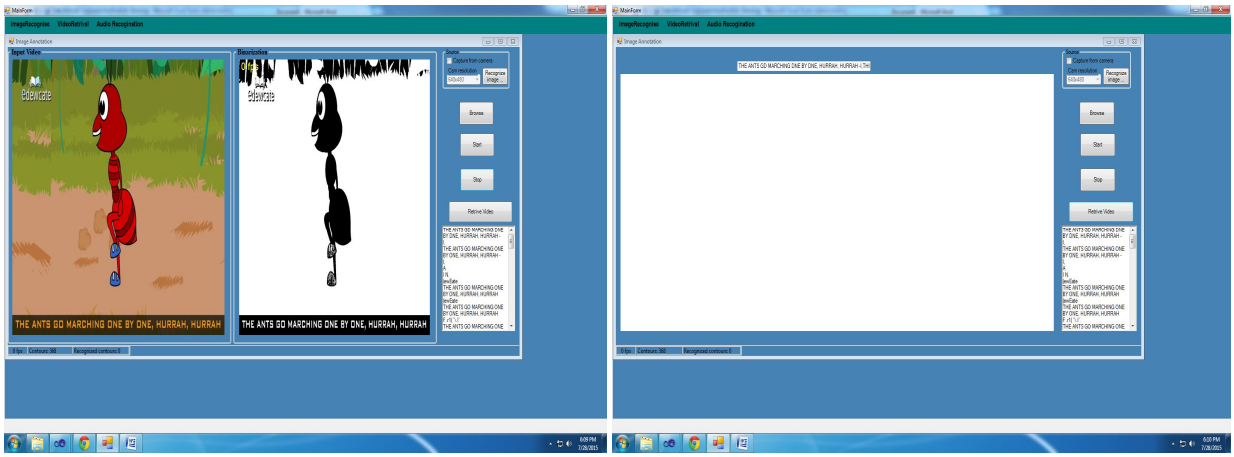


Fig 3.Text Extraction

Fig.3 shows the extraction of text from the video for further retrieval.

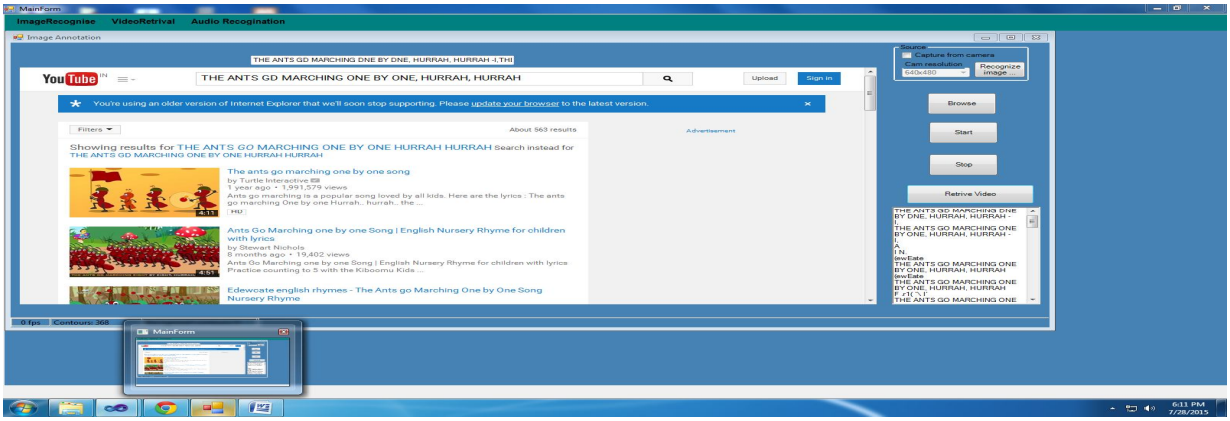


Fig 4.Video Retrieval

Fig.4 Retrieve the video related to the extracted text.

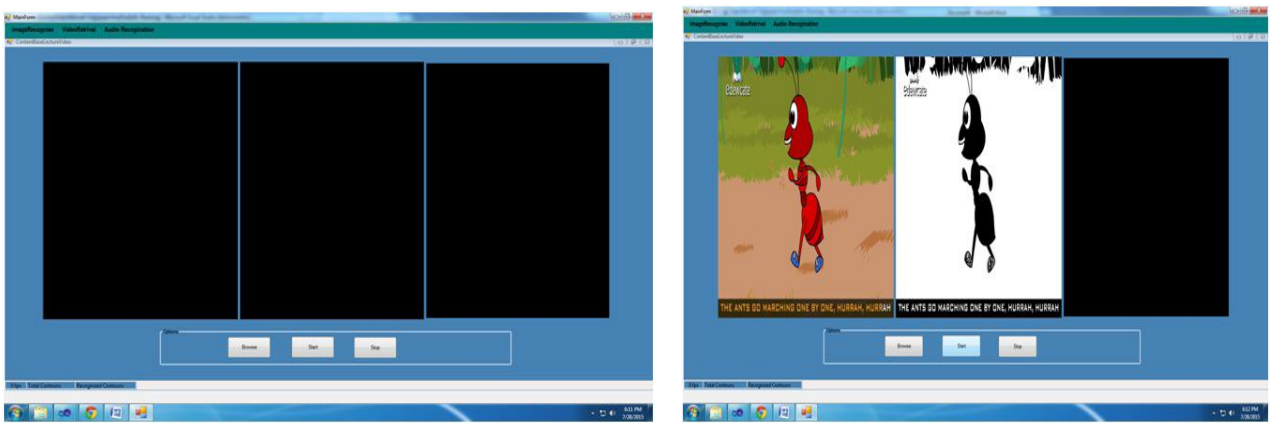


Fig.5 Image Annotation

Fig.5 User can annotate frames from video for efficient retrieval.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 9, September 2015

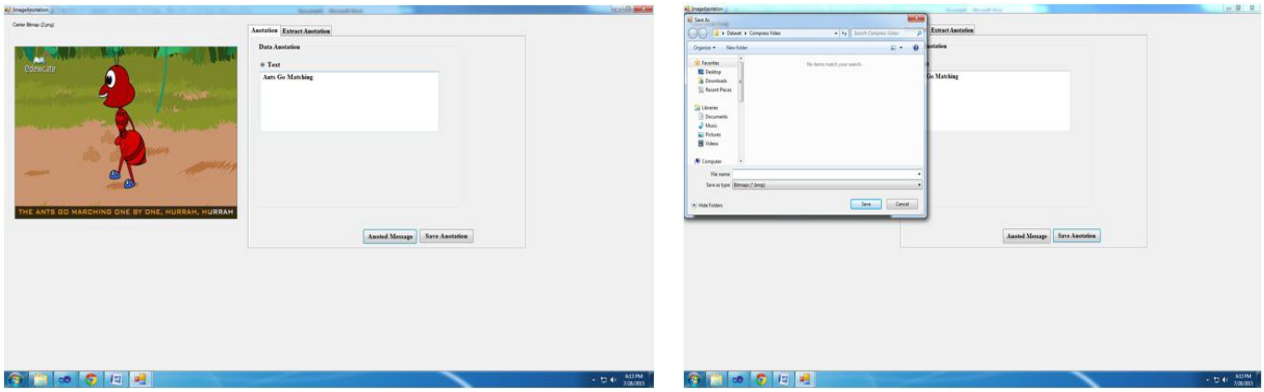


Fig.6 Annotate image and save file

Fig.6 User annotates frame from video and save file for further use.

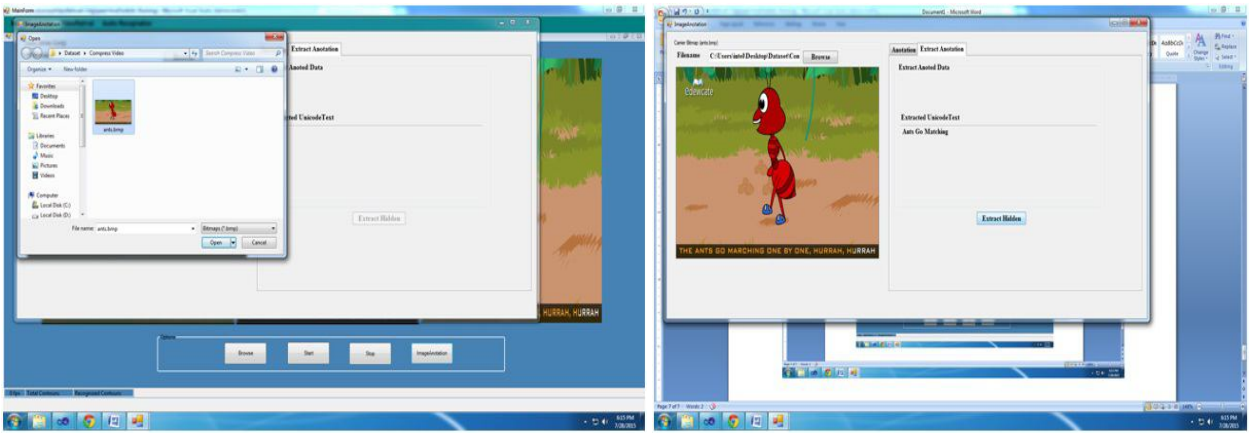
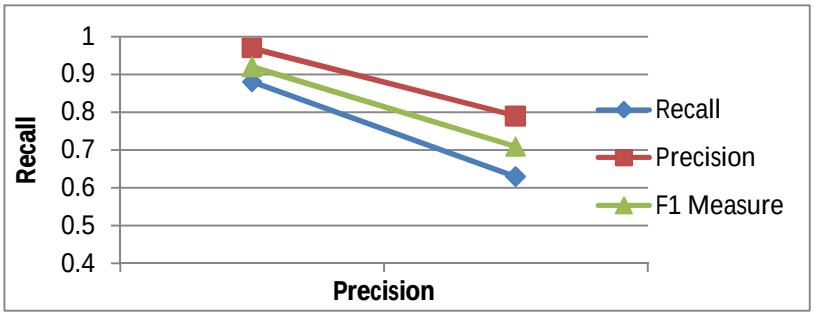


Fig.7 Extract Annotation

Fig. 7 User extract annotation for relevant retrieval of video.

V. CONCLUSION AND FUTURE WORK

This proposed system present an approach for content-based lecture video indexing and retrieval in large lecture video archives. Several novel indexing features have been developed in a large lecture video portal by using those metadata and a user study has been conducted. The relevant metadata can be automatically gathered from lecture videos by using appropriate analysis techniques. It can help a user to find and to understand lecture contents more efficiently, and the learning effectiveness can thus be improved.



ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 9, September 2015

REFERENCES

1. Leeuwis, M. Federico, and M. Cettolo, "Language modeling and transcription of the ted corpus lectures," in Proc. IEEE Int.Conf. Acoust., Speech Signal Process., 2003, pp. 232–235.
2. Lee and G. G. Lee, "A korean spoken document retrieval system for lecture search," in Proc. ACM Special Interest Group Inf. Retrieval
3. Searching Spontaneous Conversational Speech Workshop, 2008. 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. A. Haubold and J. R. Kender, "Augmented segmentation and visualization for presentation videos," in Proc. 13th Annu. ACM Int. Conf. Multimedia, 2005, pp. 51–60.
5. W. H€urst, T. Kreuzer, and M. Wiesenh€utter, "A qualitative study towards 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.
6. A. 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.
7. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," Inf. Process. Manage., vol. 24, no. 5, pp. 513– 523, 1988.
8. Salton, A. Wong, and C. S. Yang. (Nov. 1975). A vector space model for automatic indexing, Commun. ACM, 18(11),pp. 613–620, [Online]. Available: http://doi.acm.org/10.1145/ 361219.361220
9. 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.
10. M. Grcar, D. Mladenic, and P. Kese, "Semi-automatic categorization of videos on ideolectures.net," in Proc. Eur. Conf. Mach. Learn. Knowl. Discovery Databases, 2009, pp. 730–733.
11. 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.
12. J. Jeong, 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., pp. 285– 288. [Online]. Available: http://doi.acm.org/10.1145/2428955.2429011.
13. 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.
14. 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.