International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

A Survey on Content Based Lecture Video Retrieval Using Speech and Video Text information

Rupali Khollam[1], S. Pratap Singh[2]

[1]Savitribai Phule Pune University, Institute of Knowledge COE, Pimpale-Jagtap,Shirur, Tal.-Shirur,Dist.-Pune, India

[2]Professor, Savitribai Phule Pune University, Institute of Knowledge COE, Pimpale-Jagtap,Shirur, Tal.-Shirur,Dist.-Pune,,India

Abstract: ***Recording lectures and putting them on the Web for access by students has become a general trend at various universities. The amount of lecture video data is growing rapidly on the World Wide Web (WWW). Lecture videos contain text information in the visual as well as audio channels: the presentation slides and 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, we apply video content analysis to detect slides and (OCR) Optical Character Recognition to obtain their text and (ASR) Automatic Speech Recognition is used to extract spoken text from the recorded audio. In this paper we present an approach for automated video indexing and video search in large lecture video archives. Firstly we apply automatic video segmentation and key-frame detection. Then; we 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

1. Introduction

Due to the rapid development in recording technology, improved video compression techniques and high-speed networks in the last few years its becoming very popular in universities to capture and record live presentations of lectures. Presentations are delivered with the help of slides that express the author's topical structuring of the content. The system is used regularly for lecture recording by many universities and also by many other institutions. It attracted not only a broad variety of users but also served as a basis for a commercial product. Many research projects have experimented with automatic lecture recording and making the resulting documents available for access by students over the Web. Easily accessible presentation video libraries 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. As a result, there is a huge increase in the amount of multimedia data on the Web. Hence it becomes nearly impossible to find desired videos without a search function within a video data. Also when the user found related video data, it is still difficult 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. Text is a high-level semantic feature used for the content-based information retrieval. In lecture videos, texts from lecture slides serve as an outline for the lecture and are very important for understanding. So after segmenting a video file into a set of key frames (all the unique slides with complete contents), 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. In In the following, we present, a workflow for gathering video textual information, including video segmentation/lecture slide extraction, video OCR, ASR, and keyword extraction from OCR and ASR results. We can detect the unique lecture slides by using a Connected Component (CC)-based segmentation method, The detected slide keyframes are further utilized by a video OCR engine. To develop a content-based video search engine in a lecture video portal, the search indices will be created from different information resources, including manual annotations, OCR and ASR transcripts etc. The varying recognition accuracy of different analysis engines might result in solidity and consistency problems.

2. Automated Lecture Video Indexing

We perform four analysis processes for the retrieval task from visual screen and audio tracks. From the visual screen initially we detect the slide transitions and extract each unique slide frame considered as the video segment. Then video OCR analysis is performed for retrieving textual metadata from slide frames. Based on this OCR results, we propose a novel solution for lecture outline extraction by using stroke width and geometric information of detected text lines.

2.1 Slide Video Segmentation

Firstly, the entire slide video is analyzed and captures every knowledge change between adjacent frames, for this we established an analysis interval of three seconds taking both accuracy and efficiency into account. This means the segments having duration smaller than three seconds may be discarded in our system. Then we create canny edge maps for adjacent frames and build the pixel differential image from the edge maps. The CC analysis (Connected Component) is subsequently performed on this differential image and the number of CCs is then used as a threshold for the segmentation. CC-based method, by which the binary CCs are applied instead of image pixels as the basis element. In this way, high-frequency image noises can be removed in the frame comparison process by adjusting a valid size of CCs.

Licensed Under Creative Commons Attribution CC BY

In second segmentation step the real slide transitions will be captured. First we define the title and content region of a slide frame. Any small changes within the title region may cause a slide transition, e.g. two slides often differ from each other in a single chapter number. If there is no difference found in title region, then we try to detect the first and the last bounding box object in content region vertically and perform the CC-based differencing within the object regions of two adjacent frames In case that the difference value of both object regions between adjacent frames exceed the threshold Ts, a slide transition captured. This method designed for segmenting slide videos, but it is not suitable when the slides include videos with varying genres and are played during the presentation. For solving this problem we extend the original algorithm by using a Support Vector Machine (SVM) classifier and image intensity histogram features. We use the Radial Basis Function (RBF) as kernel

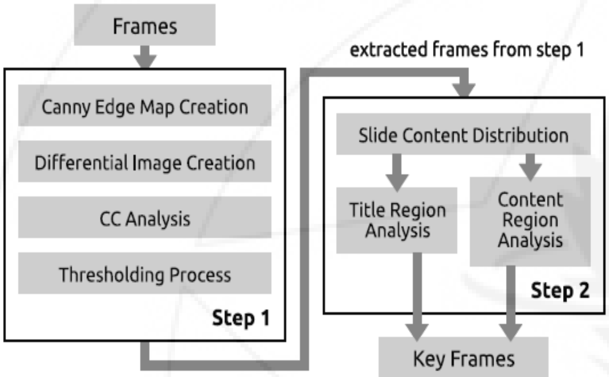

Figure1: Lecture video segmentation workflow.Step1.adjacent frames are compared with each other by applying the CC analysis on their differential edge maps.Step2:slide transitions are captured by performing title and content region analysis.

2.2 Video OCR for Lecture Videos

Texts in lecture slide are closely related to the lecture content and used for there retrieval task. In our approach we developed a novel video OCR system for gathering video 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 subsequent text area verification, an image entropy-based adaptive refinement algorithm not only serves to reject false positives that expose low edge density, but also further 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. But the SWT verifier is not able to correctly identify special non-text patterns such as sphere, window blocks ,garden fence so we adopted an additional SVM classifier to sort out these non-text patterns in order to further improve the detection accuracy. For text segmentation and recognition, we developed a novel Binarization approach, in which we use image skeleton and edge maps to identify text pixels. The proposed method includes three main steps: text gradient direction analysis, seed pixel selection, and seed-region growing. After the seed-region growing process, the video text images are converted into a suitable format for standard OCR engines. For removing the spelling mistakes resulted by the OCR engine, we perform a dictionary-based filtering process. For ranking keywords we use term frequency inverse document frequency. Ranked keywords used for video content browsing and video search. Video Similarity calculated by using cosine similarity measure based on extracted keywords.

2.3 ASR for Lecture Videos

Spoken documents are generated by extracting audio data from lecture 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.

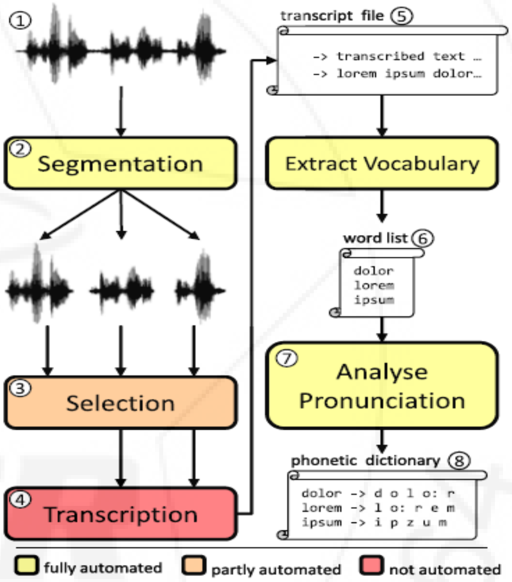

Figure 3: Workflow of our extending speech corpus.

3. Video Content Browsing and Video Search

3.1 Keyword Extraction and Video Search

The lecture content-based metadata is gathered by using OCR and ASR tools. But the recognition results of automatic analysis engines are often error prone and generate a large amount of irrelevant words. Therefore we extract keywords from the raw recognition results. Keywords summarize a document and are widely used for information retrieval. Only nouns and numbers are considered as keyword candidate. The top n words from them are considered as keyword. Segment-level as well as video-level keywords are extracted from different information resources such as OCR and ASR transcripts respectively. For extracting segment level keywords, we consider each individual lecture video as a document corpus and each video segment as a single document, whereas for obtaining video-level keywords, all lecture videos in the database are processed, and each video is considered as a single document. To extract segment-level keywords, we first arrange each ASR and OCR word to an

Licensed Under Creative Commons Attribution CC BY

appropriate video segment according to the time stamp. Then we extract nouns from the transcripts by using the Stan ford part-of-speech tagger and a stemming algorithm is subsequently utilized to capture nouns with variant forms. To remove the spelling mistakes resulted by the OCR engine, we perform a dictionary-based filtering process. We calculate the weighting factor for each remaining keyword by extending the standard TFIDF score. In general, the TFIDF algorithm calculates keywords only according to their statistical frequencies. It cannot represent the location information of keywords that might be important for ranking keywords extracted from web pages or lecture slides. Therefore, we defined a new formula for calculating TFIDF score, as shown by Eq. (1):

$$\mathrm{Tfidf}_{seg}\text{-internal}^{(k\omega)} = 1/N(\mathrm{tfidf}_{ocr}.1/n_{type}\sum_{i=1}^{ntype} \omega i + tfidasr * \omega asr) \quad (1)$$

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

$$\omega_{i=}\mu/\sigma_{i\ (i=1\ldots.n)} \quad (2)$$

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)

$$\sigma_{i}\text{=1-Accuracyi (i=1….n).} \quad (3)$$

References

[1] G. Salton, A. Wong, and C. S. Yang. (Nov. 1975). A vectorspace model for automatic indexing, Commun.ACM, 18(11),pp. 613–620, [Online]. Available: http://doi.acm.org/10.1145/361219.361220
[2] J. Glass, T. J. Hazen, L. Hetherington, and C. Wang, "Analysis and processing of lecture audio data: Preliminary investigations," inProc. HLT-NAACL Workshop Interdisciplinary Approaches Speech Indexing Retrieval, 2004, pp. 9–12.
[3] A. Haubold and J. R. Kender, "Augmented segmentation and visualization for presentation videos," in Proc. 13th Annu. ACMInt. Conf. Multimedia, 2005, pp. 51–60.
[4] 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..

Author Profile

Rupali received the B.E. and pursuing M.E. degrees in Computer Engineering from Institute of Knowledge College of Engineering in 2011 and 2014, respectively. During 2010-2011, she worked on Data Mining project for fulfilment of her Bachelors' Degree.

Licensed Under Creative Commons Attribution CC BY