International Journal for Rapid Research in Engineering Technology & Applied Science
ISSN (Online): 2455-4723
Vol 2 Issue 6 July 2016

Survey Paper on Content Based Video Retrieval by OCR Technique

Paridhi Soni[1], Prof. Abhishek Tiwari[2]
[*1]Research Scholar, Mahakal Institute of Technology, Ujjain, M.P, India.
[2]Associate Professor, Mahakal Institute of Technology, Ujjain, M.P, India.
*Department of Information Technology

Abstract
In the last decade e-lecturing has become more and more popular. The amount of lecture video data on the World Wide Web (WWW) is growing rapidly. Therefore, a more efficient method for video retrieval within large lecture video archives is urgently needed. In this paper the usability and utility study for the video search function in our lecture video portal will be conducted. Automated annotation for OCR 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.

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

1. 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]. Therefore, the e-lecturing system is used frequently for audiovisual (audio & video) recordings. An e-lecture consists of slides with relevant points mentioned by the lecturer. A number of colleges and research institutes are taking a chance to record their lectures and publish them online for students to access free of time and location. As a result, there has been a huge increase in the amount of multimedia data on the Web. The user requested for appropriate information which is covered in only few part of the video, and he wants only that information without viewing the complete video. So the problem is how to retrieve the appropriate information in a large lecture video. There are many video search systems like YouTube, Bing etc. based on available textual metadata such as title, person and brief description etc. 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.
In starting the lecture videos are recorded by a single video camera, which is the cause of lower quality lecture videos. Traditional video retrieval based on feature extraction cannot be efficiently applied to lecture recordings. Lecture recordings are characterized by a homogeneous scene composition. Most of the time, the lecturer is in focus, presenting a topic which is not visible. Thus, image analysis of lecture videos fails even if the producer tries to loosen the scene with creative camera trajectories.

Fig.1. (a) Example of outdated lecture video format

(b) An exemplary lecture video. Video 1 shows the professor giving his lecture, whereas his presentation is played in video 2.
But Nowadays 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 lecture recording system such as tele-Teaching Anywhere Solution Kit (tele-TASK). In it the speaker and his presentation are displayed synchronously.
In this paper, we are presenting Content Based Video Retrieval (CBVR) System it includes various steps: Video Segmentation: it is used for video segmentation, Feature Extraction: Features are extracted for the key frame and stored into feature vector. Histogram Of Gradient is an algorithm which is used for the feature extraction. Video segmentation is first step towards the content based video search aiming to segment moving objects in video

sequences. Segmentation of Video is done with the help of step by step process of video segmentation.

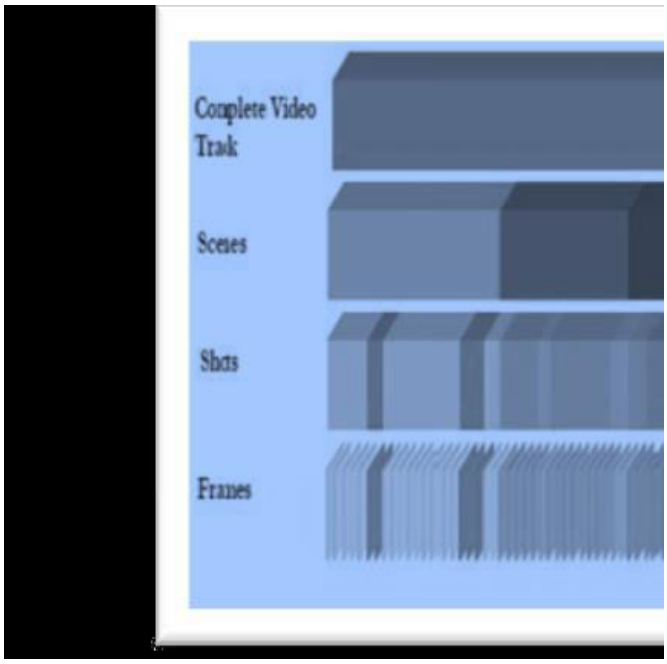

Fig.2. Video Segmentation

The complete video is first converted into scenes, then scenes are converted into shots and finally shots are converted into various frames.

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. We extract textual metadata by applying video Optical Character Recognition (OCR) technology on key-frames and Automatic Speech Recognition (ASR) on lecture audio tracks. OCR is primarily used for converting scanned images of documents into text searchable format. In this situation, this technology is the most useful because it helps do away with manual data capture. The need for OCR technology came about because when dealing with scanned images of paper documents, software just did not have the means to recognize the printed text present on them. It is a translation of scanned image. Automatic speech recognition (ASR) can be defined as the independent, computer-driven transcription of spoken language into readable text in real time. ASR is technology that allows a computer to identify the words that a person speaks into a microphone or telephone and convert it to written text. The ultimate goal of ASR is to allow a computer to recognize in real-time, with 100% accuracy, all words that are intelligibly spoken by any person, independent of vocabulary size, noise, speaker characteristics or accent. Today, if the system is trained to learn an individual speaker's voice, then much larger vocabularies are possible and accuracy can be greater than 90%. A large amount of textual metadata will be created by using OCR und ASR method, which opens up the content of lecture videos. In our research, we intended to continuously improve the ASR result for German lectures by building new speech training data based on the open-source ASR tool. For content-based video search, the search indices are created from different information resources, including manual annotations, OCR and ASR keywords, global metadata, etc. In addition to video OCR, ASR can provide speech-to-text information from lecture videos, which offers the chance to improve the quantity of automatically generated metadata.

2. Related Works

1. In Content based lecture video retrieval using speech and video text information, information retrieval in the multimedia-based learning is an active research area. Video texts, spoken language, video actions, can act as the source to open up the content of lectures. Wang et al. proposed an approach for lecture video indexing based on automated video segmentation. 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. ASR provides speech-to-text information on spoken languages, which is thus well suited for content-based lecture video retrieval. The studies described in [2] are based on out-of-the-box commercial speech recognition software. 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 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. 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. In it 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. Video browsing can be achieved by segmenting video into representative key frames. 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. 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). Another benefit of our segmentation method is its robustness to animated content progressive build-ups used within lecture slides.

2. Text segmentation and recognition, we developed a novel binarization approach, in which we utilize image skeleton and edge maps to identify the text pixels. The proposed method consists of 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. The subsequent spell-checking process will further sort out incorrect words from the recognition results. Generally, in the lecture slide the content of title, subtitle and key point have more significance than the normal slide text, as they summarize each slide. In addition to video OCR, ASR can provide speech-to-text information from lecture videos, which offers the chance to improve the quantity of automatically generated metadata dramatically. A recorded lecture audio stream yields approximately 90 minutes of speech data, which is far too long to be processed by the ASR trainer or the speech decoder at once. Shorter speech segments are thus required.

3. This paper presented an approach for content-based lecture video indexing and retrieval in large lecture video archives. The text is retrieved with the usage of the SVM classification and the HOG feature extraction method. The main process of processing speed of both videos and the feature extraction is analyzed and the effectiveness evaluation is higher compared with the presented one. Histogram of Oriented Gradients (HOG) is feature descriptors which counts occurrences of gradient orientation in localized portions of an image. HOG feature extraction method extracts gradient values of all frames. Finally, Support vector machines (SVM) classifier is used for classification. SVM are supervised learning models with associated learning algorithms. Histogram Of Gradients is an algorithm which is used for the feature extraction. It is extracted based on the histogram of the feature. HOG are feature descriptors used in image processing and computer vision for the purpose of object detection. For evaluation purposes, several automatic indexing functionalities is developed in a large lecture video portal, which can guide both visually and text oriented users to navigate within lecture video. A user study that intended to verify their search hypothesis and to investigate the usability and the effectiveness of proposed video indexing feature. In this paper, we are presenting Content Based Video Retrieval (CBVR) System it includes various steps: Video Segmentation: Adaptive Thresholding algorithm is used for image segmentation, Feature Extraction: Features are extracted for the key frame and stored into feature vector. Histogram Of Gradient is an algorithm which is used for the feature extraction.

4. Automatic speech recognition (ASR) can be defined as the independent, computer-driven transcription of spoken language into readable text in real time. ASR is technology that allows a computer to identify the words that a person speaks into a microphone or telephone and convert it to written text. The ultimate goal of ASR is to allow a computer to recognize in real-time, with 100% accuracy, all words that are intelligibly spoken by any person, independent of vocabulary size, noise, speaker characteristics or accent. Today, if the system is trained to learn an individual speaker's voice, then much larger vocabularies are possible and accuracy can be greater than 90%. The task is to getting a computer to understand spoken language. By “understand” we mean to react appropriately and convert the input speech into another medium e.g. text. Speech recognition is therefore sometimes referred to as speech-to-text (STT). Feature extraction, Acoustic modeling, Pronunciation modeling, Decoder. The process of speech recognition begins with a speaker creating an utterance which consists of the sound waves. These sound waves are then captured by a microphone and converted into electrical signals. These electrical signals are then converted into digital form to make them understandable by the speech-system.

5. OCR is primarily used for converting scanned images of documents into text searchable format. In this situation, this technology is the most useful because it helps do away with manual data capture. The need for OCR technology came about because when dealing with scanned images of paper documents, software just did not have the means to recognize the printed text present on them. It is a translation of scanned image. The advancements in pattern recognition have accelerated recently due to the many emerging applications which are not only challenging, but also computationally more demanding, such evident in Optical Character Recognition (OCR). Optical Character Recognition is classified into two types, Offline recognition and online recognition. In offline recognition the source is either an image or a scanned form of the document whereas in online recognition the successive points are represented as a function of time and the order of strokes are also available [5].

3. Existing System

Existing System presents an approach for automated video indexing and video search in large lecture video archives. This Method applies automatic video segmentation and key-frame detection to offer a visual guideline for the video content navigation. Subsequently, extract textual metadata by applying video Optical Character Recognition (OCR) technology on key-frames. The OCR 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.

4. Proposed System

The usability and utility study for the video search function in our lecture video portal will be conducted. Automated annotation for OCR 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.

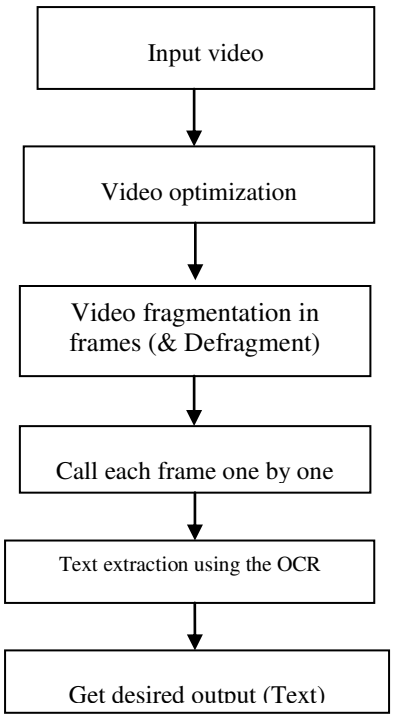

Fig.2. Video Segmentation

5. Conclusion

We can conclude as this research work will fulfill some issues by proposed solution. These are the major issues which are not retrieving the video efficiently from the large amount of video database in existing system:-

1. How to retrieve the appropriate information in a large lecture video archive more efficiently.
2. To let machine understand video is important and challenging.
3. How to continuously improve the accuracy ASR lecture video which stills an unsolved problem.
4. The main problem is that the video analysis methods may introduce errors.

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-JUNE2014.

2] Stephan Repp, Andreas Groß, and Christoph Meinel, Member, *IEEE* ―Browsing within Lecture Videos Based on the Chain Index of Speech Transcription‖ *IEEE* TRANSACTIONS ON LEARNING TECHNOLOGIES, VOL. 1, NO. 3, JULY- SEPTEMBER 2008.

3] Vigneshwari.G, A SURVEY ON CONTENT BASED LECTURING VIDEO RETRIEVAL International Journal of Computer Science and Mobile Computing, Vol.3 Issue.11,November- 2014, pg. 275-282

4] Automatic Speech Recognition http://ijettjournal.org/volume-4/issue-2/IJETT-V4I2P210

5] Optical Character Recognition International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 1, January 2014

6] Bhagwant B Handge et al, Retrieval of Video Using Content (Speech &Text) Information Int. J.Computer Technology & Applications,Vol 5 (6),1939-1944