International Journal of Electrical, Electronics and Computer Engineering **3**(2): 169-173(2014)

ISSN No. (Online): 2277-2626

A Review on: Video Searching Using Speech & Video Text Information

Mohammed Salim A. Khan*, Santosh S. Lomte** and Seema Singh Solanki*

**Research Scholar, Department of Computer, Dr. Seema Quadri Institute of Tech, Aurangabad, Maharashtra, India*

***Professor, Department of Computer, Dr. Seema Quadri Institute of Tech, Aurangabad, India*

****Assistant Professor, Department of Computer, Dr. Seema Quadri Institute of Tech, Aurangabad, India*

(Corresponding author: Mohammed Salim A. Khan)

(Received 05 October, 2014 Accepted 28 December, 2014)

ABSTRACT: Since Last decade e-lecturing has become more and more popular. There is a continuous increment in the number of lecture videos data on the Internet and there is no doubt that in near future it will be booming field of e-lectures. Hence there is a need for highly flexible and best results-yielding algorithms or methods to search appropriate contents. Thus, indexing is an approach which is mainly emphasized in this paper for automated video indexing and video search in large lecture video archives. Primarily, Videos are divided into some segments and then key-frames are identified among these segments for further protocols of video content data extraction and mining. Immediately after that, steps like extraction of textual metadata by applying Optical Character Recognition (OCR) technology on key-frames of videos and Automatic Speech Recognition (ASR) on audio lecture tracks are followed. The OCR and ASR transcripts are the methods from which textual data is extracted and stored into digitized format and can be used as keywords related to videos, matching these keywords with the available database of videos on the Internet and thus will provide relevant Videos and contents. The evaluation of results for performance in index based searching method is remarkable as compared to other searching methods.

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

I. INTRODUCTION

In today's day-to-day environment, Due to the homogenous scene property of the videos formed, appropriate important results based on visual feature extraction cannot be simply applied to lectures recorded in video format. Nowadays people look to create lecture videos by using dual scene format, by which the speaker and his presentations are displayed synchronously. The major advantages of showing presentations for a lecturer will be flexibility and a higher level of understandability. For indexing no extra synchronization between video and slide files are required and we also neglect the existence of slide files. The system lacks its importance just because of the method of video analysis, which introduces errors. Our research work mainly focuses on those lecture videos produced by using screen grabbing method. Since two videos are synchronized automatically during the recording process. Therefore, the temporal scope of complete unique slides can be considered as lecture segment. This way, segmentation of dual scene lecture video can be achieved by processing slide video segmentation.

Extraction of textual metadata is carried out from audio and visual information of the video, by using some of the best applicable algorithms and techniques. A separate large video lecture portal is used on the interpretation of video's automatic indexing features, that will definitely help both visually and text-oriented users to search for the appropriate video lecture. Then, a survey was conducted by us to check the research effectiveness on the change in method of video searching. This data survey helped in analysis of usability of video indexing for searching on Internet. For visual analysis, an approach of OCR is used. OCR is passed with the some number of key frames snapped from videos to gather textual metadata of the current video. The overall key data of the video files is generated using graphical functions. Structured video text provides a more helpful search function. We also had a solution for German dictionary which has automated German phonetic sound to attend, which also assists in ASR domain and auto corrects the general errors. The overall modules of dictionary and compiled speech of software is used for future scope. For maintaining the constraints of solidity and stability problems of a content-based video search system, we propose a different keyword classification method for different models of information resources. In order to measure the usefulness, we tried this method in a large lecture video portal.

II. EXISTING SYSTEM

Prior to this system, it was not an easy task to provide the users with the appropriate results in a definite time.

In general whenever user requests any video, results appearing in the output will have the matching similar names of video for which it was searched. Actually it works by matching the names of files stored in the database with the actual data required and accordingly results are fetched and displayed. Accuracy with this system is adjusted and minimized. In the existing system, results are displayed according to the title of the video so it becomes impossible to view each and every video having similar type of titles. For example, if user wants to study conditional statements in c language then it will search video according to the name of video but it is inconvenient for the user to study that video if it contains only short information about user's query. So, it may give irrelevant and inappropriate results.

In existing system video retrieval based on visual feature extraction cannot be simply applied to lecture recordings because of the homogeneous scene composition of lecture videos. For example, if a camera is in motion then camera may distort the properties of the slide; the slide can be partially blocked if the person is in moving position 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.

III. PROPOSED SYSTEM

We propose the system is to retrieve a video on the basis of its contents rather than retrieving video according to its title and metadata description. So that it will provide an exact result for the user's demand. For this purpose, we have to implement a model which captures the various frames from a video lecture. All the captured frames are then classified according to the duplication property. Key frames segmentation is performed by considering a time interval within two frames. Sometimes it happens that a video lecture contains one slide presentation for a long period of time then maximum time interval is introduced in seconds. Subsequently key frames subdivision occurs in which we fetch all the text from all the frames for further video retrieval system using optical character recognition (OCR) process. Also we make all the voices resulting into text using ASR technique. This all material mined till now (Text and Voice from Video) is used for content based video retrieval system and clustering of video according to their text and voice parameters.

This OCR algorithm is used to extract the characters from the literal information as well as ASR algorithm is used to retrieve the speech information from the video lecture. The OCR and ASR record as well as detect slide text line types are implemented for mining some relevant important words, by which both video and segment-level keywords are extracted for content-relevant video surfing. The presentation and the success measurements of proposed indexing functionalities is proven by evaluation.

IV. FUTURE IMPLEMENTATION

Implementation is the phase of the project when our estimations and works of papers and documentation is converted into a real life work. Thus it can be measured to be the most precarious stage in attaining a successful new system and in assuring the user, with guarantee that the new system will work and be operative. The operation stage contains cautious scheduling, analysis and research of the existing system and it's constraints on implementation, scheming of approaches to complete conversion and evaluation of conversion methods. Two major algorithms which are used for implementation of the proposed system are as follows;

A. Optical Character Recognition (OCR)

The task of recognizing characters in lecture video can be divided into the modules This division is consistent with the standard approach used in OCR-systems First of all the raw signal from the composite video signal is sampled to create a binary image, from which the subtitles can be extracted. When using the signal from the camera it is also required to manage the basic requirements such as alignment of text occurring in frame must be horizontal, which is a precondition for our OCR-algorithm. Next the binary image is profiteered to remove noise and enhance characteristic features. After working on image different statistics like number of lines, words and characters are needed to be stored. All these newly collected statistically analyzed data is matched with all the existing set of databases. After picking the most probable letters in a given word, the word can be related with a dictionary checkup, to verify if the letter combination is likely.

(i) To undergo the study of OCR images are used which processes by sampling the composite signal from the moving pictures like video and applying a threshold. A sample of the subsequent binary picture is shown Fig. 1.

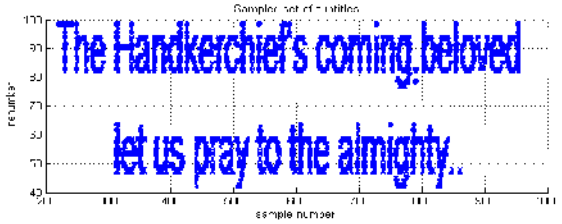

Fig. 1: Sampling and applying a binary threshold the binary picture is created.

(ii) In the case where the images come from the CCD filmed camera, it will usually be essential to bring out some minor typical adaptations, since our later designated feature extraction is not rotational invariant. The subtitles consist of one or two lines of chockfull letters with the similar coordination. This basic information can be used for finding the rotation of the image, which maximizes the horizontal sum in the frame (corresponding to a horizontal orientation of the subtitles).

(ii) The success of the OCR-algorithm depends on the preliminary filtering. The main purpose of this filtering and segmentation is to clearly distinguish between the textual data and making this data as self-describing about the media file. It starts with low pass-filtering the image to eradicate high frequency noise which was not removed in the sample process.

(ii) To compare each region with our database we need to mine relevant features. We use a grouping of simple arithmetical and semantic features, which reduces the extent of calculations. The arithmetical features are relations between area covered in frame, width/height of objects, and background/foreground characteristics.

(ii) The horizontal and vertical projections are compared with the database using cross correlation. When comparing the arithmetical features such as ratio between height/width, foreground/background area during the video and first order moment assumptions are made. For each letter the mean of this circulation is given in the databank and the change is chosen for best detection. An extracted feature such as the pixel area can now be checked towards the distribution of each letter. The results from the projections and the other statistics are finally combined and the letter with similar meaning is chosen. Then the semantic features are used for correction of the most likely misclassifications.

(ii) Even with a good detection rate of each letter can have errors which will occur from time to time. To cater for this mismanagement each word is checked with a vocabulary, containing the most common words. If the term doesn't exist then the most likely matching term is chosen, e.g. by keeping the word with most letters in their required position. It is then assessed which of the 2 words is the most similar in sense, based on the probabilities from the classification.

B. Automatic speech recognition

Automatic speech recognition algorithm extracts speech or voice from lecture video and converts it into textual information for storing it into database. Speech is one of the most significant medium of knowledge sharing in video classes. Thus, it is of unique advantage that this material can be applied for automatic lecture video indexing. Awkwardly, most of the currently available lecture speech recognition systems in the studied work cannot attain a adequate recognition outcome on the Word Error Rates. ASR is meant to empower particular machines to recognize actual voice characters into digitized format without human involvement. These are the basic steps in Automatic Speech Recognition:

(i) User speaks

(ii) System extracts features from the speech.

(iii) Those features statistically match up with phoneme.

(iv) Use the word statistics to go from phoneme ordering towards.

V. MODULES

A. Frame Extraction

This is the first module of our project in which input video is given and that video is segmented into the number of key frames during certain period of time interval in seconds. Sometimes, it happens that a same key frame is displayed for a long period of time then to reduce duplication we will increase the time interval of video segmentation.

B. OCR Module

Optical character recognition module is used to retrieve the text metadata from the extracted key frames of lecture video. When OCR algorithm is applying on the lecture video it has to follow certain steps. At first the raw signal from the composite video signal is sampled to create a binary image, from which the subtitles can be extracted. When using the signal from the camera it is also necessary with some spatial adjustment to ensure that the lines of text are horizontal in the image, which is a precondition for our OCR-algorithm. Next the binary image is pre filtered to remove noise and enhance characteristic features. After the optimal filtering the image can be divided into individual sentences, arguments and characters. Every feature and characteristics whether it is statistical or semantic they are spotted and matched with a previously existing database (words provided as training suite). After choosing the most likely letters in a given word, the word can be compared with a dictionary lookup, to verify if the letter combination is likely.

C. ASR Module

Speech is one of the most important carriers of information in video lectures. Automatic speech recognition algorithm extracts speech or voice from lecture video and converts it into textual information and stores it into database.

D. Retrieval of content based Video

OCR and ASR results of a synchronized timestamp are combined and results are shown according to the contents of video rather than title of video.

VI. SYSTEM ARCHITECTURE

The system is to retrieve a video on the basis of its contents and information it shares. So that it will provide an exact result for the user's request. For this purpose, we have to develop a model which screenshots the in-between frames from a video. All the snapped frames are then arranged according to the replicative characteristics. Then we mine all the text from all the frames for further video retrieval system. Then using ASR the filtering of audio is started to extract textual data from audio and speeches and this is how the video retrieval system will work. This all information extracted from a video media file will be attached as keywords in the video. Thus categorization will be done on the basis of these keywords, so that it will help further to directly search the content on the Internet. We extract metadata from pictorial as well as audial resources of lecture videos automatically by applying suitable analysis and study techniques. For assessment purposes we developed numerous automatic indexing functionalities in a large lecture video portal that can help the users in searching the relevant data. A research survey was held under the proposed system to check the effectiveness and the usability of the video indexing methodology for more accuracy in data mining but only related to videos. For information extraction through videos the OCR will be collecting key data from the key frames snapped in-between by the frame extractor. Furthermore, lecture's postscript is extracted from OCR transcripts by using stroke width and geometric information. Structured video text provides a more helpful search function. We also had a solution for German dictionary which has automated German phonetic sound to attend, which also assists in ASR domain and auto corrects the general errors. The overall modules of dictionary and compiled speech of software is used for future scope In order to overcome the solidity and consistency problems of a content-based video search system, we propose a different keyword ranking method for different models of information resources.

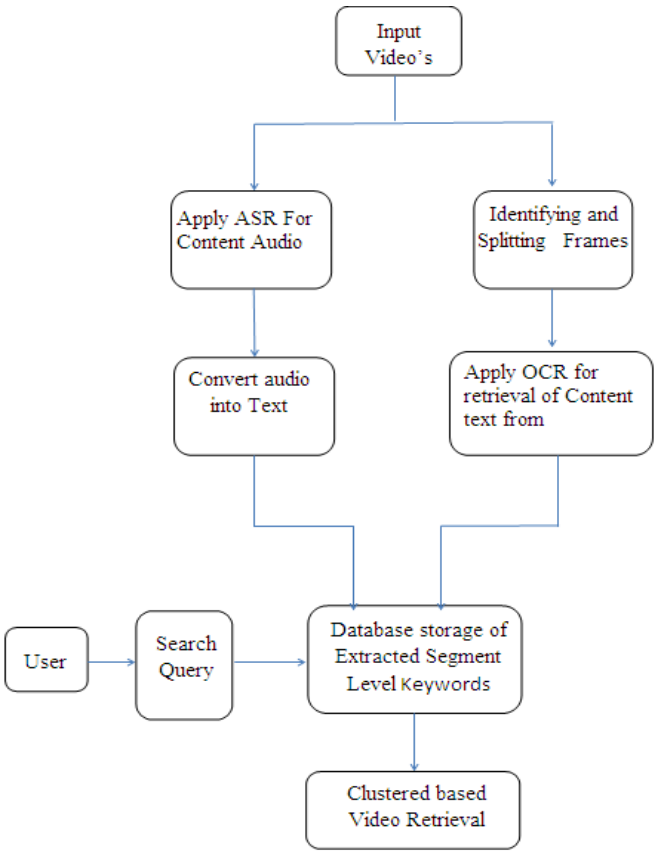

Fig. 2. Process of Video Retrieval.

In order to measure the usefulness and this is verified at the large video portals. 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.

CONCLUSION

We are exposing the methodology for content-based lecture video indexing and retrieval in large lecture video data collections. For performing this research we had applied the methods discussed like using OCR and ASR to sort out the informative key words. The system uses video indexing procedure to sort out the contents of the web which will help in easy and fast retrieval. In this way we have developed a system which retrieves and results a video according to their contents not on the basis of title and metadata description only. The problem of existing system is tried to overcome with this proposed system which reduces the time complexity as user will have to go through a video which are having most applicable stuffs related to user's search query.

REFERENCES

[1]. Haojin Yang and Christoph Meinel "Content Based Lecture Video Retrieval Using Speech and Video Text Information," *IEEE 2014*, Vol. **7**, no.2, pp. 142–154.

[2]. H. Yang, B. Quehl, and H. Sack. (2012), "A framework for Improved video text detection and recognition," *Multimedia Tools Appl.*, pp. 1–29, [Online]. Available: http://dx.doi.org/10.1007/s11042-012-1250-6.

[3]. E. 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.

[4]. D. Lee and G. G. Lee, "A korean spoken document retrieval system for lecture search," in Proc. *ACM Special Interest Group Inf. Retrieval Searching Spontaneous Conversational Speech Workshop, 2008.*

[5]. 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.

[6]. 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.

[7]. 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.

[8]. 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.

[9]. H. 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

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

[11]. J. Eisenstein, R. Barzilay, and R. Davis. (2007). "Turning lectures into 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.