International Journal of Scientific Engineering and Research (IJSER)
www.ijser.in
ISSN (Online): 2347-3878, Impact Factor (2014): 3.05

Clustering and Retrieval of Tele -Lecture Content

Ashwini Y. Kothawade[1], Prof. Dipak R. Patil[2]

[1]Department of Information Technology, K.K. Wagh College of engineering

[2]Department of Information Technology, Amruvahini College of engineering

Abstract: ***Nowadays video lecturing is becoming popular due to its various advantages than classroom learning. Many institutes and organizations are using this method for learning. So there is an enormous amount of data available in video lecturing form. To extract the exact video and exact information through this collection of information from videos is a tedious task. In this paper we introduce the techniques for automatically retrieving the information from video files to collect it as a metadata for those files. For efficient retrieval of text from videos we use the OCR (Optical Character Recognition) tool to extract text from slides and ASR (Automatic Speech Recognition) tool for retrieval of speech information by the speaker. First of all we do segmentation and classification of video files for extracting the key frames. Then the OCR and ASR tool is used for collecting the information and it will be stored as a metadata for the file. At last, we provide the efficient browsing for these videos by using the clustering and ontology concept.***

Keywords: OCR, ASR, Content retrieval, e-learning, tele-lecture, tesseract OCR, indexing to video lectures

1. Introduction

E-learning is popularly used today by people for various purposes like learning for persons with disabilities, for showing demonstrations of practical in medial field, learning advanced techniques for professionals, to provide training for employees located in various branches of the industry etc. So the use of video lecturing is increasing due to its independence of time and location, many institutes and organizations are uploading their video lectures on the internet. Therefore the vast amount of information gets stored into the internet. From this information to collect the desired information becomes difficult without the proper retrieval system.

The existing multimedia video retrieval techniques are not applicable to video lectures because they are based on feature extraction and identifying the similarity between the frames[5], while video lectures are having homologous features between frames with many frames having similar content. So the technique used in multimedia retrieval cannot be used for video lecture retrieval. In the traditional systems, the search for video lectures is provided based on the metadata linked to it which is inserted manually by the creator of the video. There are many disadvantages identified due to this manual insertion technology because limited amount of data can be provided with each video file like its title, creation date, its type, size etc. which is insufficient for large size video with many concepts covered. When the user fires the query for getting the data, the data may not be available in the title but video may contain some kind of information related to it. At this time, the search becomes inefficient due to the limited metadata information. For inserting more information in metadata by this method is time and cost consuming.

So for increasing the efficiency of the search, more advanced technique is needed which collects the data from video files automatically and treat it as a metadata. The research has been done on this which extracts the content by identifying content on the video lecture file and analyzing the words talked by the speaker. These are two major sources of the content, one is from the speaker who speaks and another is from their slideshow content or handwritten text written on the board. The existing technologies are using OCR for slide text retrieval and ASR for retrieving the audio information from the video lectures. It provides more relevant data by assuming that the slides contain the important topics with larger font which needs to be extracted and frequently used words by the speaker are stop words and has to be removed. May technologies and research regarding this are provides the result which may be variable in its accuracy for text extraction.

In this paper, we study the automatic generation of metadata from the video files based on its content information. The slide content from the video files are extracted by optical character recognition tool which we are using Tesseract OCR and audio content are retrieved by the automatic speech recognition method. While extracting the information by using these tools some challenges may have to be faced. Nowadays the video recording is done based on multi-scene format in which the frame may contain multiple scenes at a time, like professor explaining the slide in half part of the frame and slides are shown on next half part as shown in figure 1.

Figure 1: The existing video lecture example

The challenge may occur in OCR system for the variation in creation, entering texts, text in the form of graphs, tables

etc. For the speech analysis, this is very important content of the information to be extracted have also some challenges like the speaker's fluency of the speech, his pronunciation, background noise, handwritten slide which may affect the extraction.

Figure 2: Different types of scenes and different presentation formats in instructional videos.[6]

The extracted texts by these both methods will generate the large amount of data which needs to be reduced. The keywords are extracted further by calculating the term frequency inverse document frequency (TFIDF) score. For providing the linked videos of the given search ranked keywords are clustered for different video files by k-means algorithm with similarity between terms is calculated by Euclidian distance algorithm.

2. Literature Survey

a. Video Content Retrieval

MadhavGitte, Harshal Bawaskar, Sourabh Sethi, Ajinkya Shinde developed a content based video retrieval system which uses multimodal features to extract the videos from multimedia warehouse. The system is not so complex but which extracts the information based on only feature extraction algorithm and then though the clustering method is applied but it is not giving the efficient result in case of video lectures, as content in video lectures may be more and to extract the information by feature extraction may return non text blocks also (like images). [8]

Erwin Leeuwisl, Murcello Federico and Mauro Cettolo [4] focused on English speech recognition in Translanguage English database (TED) corpus. He developed the language model by lecture transcript. The training dataset provided has to be inserted manually so it takes time and also which is hard to be extended or optimized. The word error rate returned by this was nearly 40%.

Wolfgang Hürst, Thorsten Kreuzer, Marc Wiesenhütter [3] proposed a method which extracts audio information from video lectures by developing the different ASR method. Large vocabulary automatic speech recognition (LVASR) is used which handles all kind of audio signal like low quality signal, noisy signal, fast speaker word detection which improves word accuracy than other methods for audio retrieval. But in this system, the ranking to be given to words identified is difficult and the method does not collect any information from the slides in the video which contain most important information and the method also applicable only for the German lecture videos. Xiaoqiang Xiao, JashaDroppo and Alex Acero have suggested the new technique to be use with the ASR which is more efficient than the traditional HMM based ASR system. The author has developed an IR system which identifies the subwords units into the valid word. Initially instead of decoding the speech into word, it is decoded first in subword units [10].

Matthew Cooper has compared the results from both the OCR and ASR techniques. He has shown that there is much difference in texts extracted from ASR and OCR. Also he has proved that the OCR technique is giving more accurate results for text extraction than the ASR technique. Because the ASR has many sources of error like pronunciation of the speaker, phonetic errors etc. and the errors occur in this will directly affect the retrieval performance which not happens in OCR. [9]

Alexander G. Hauptmann, Rong Jin, and Tobun D. Ng shown the different video retrieval technique for the multimedia video data, generate and combine all the metadata. The author had also shown the results for retrieval by using all methods and compared it. The author has proved that the result is more efficient when retrieval is done by using enhanced OCR with preprocessing is done effectively, the probabilistic model is used and if the speech recognition algorithm is also used with it. [2] Tiecheng Liu and John R. Kender had developed a new software tool for the content retrieval of video lectures based on its low level features. The system especially works on real time content retrieval and keyframe extraction from videos which have handwritten slide scene type. The author has used a rule based model for detecting the boundaries between frames. As the handwritten slides are there for extraction so it has a substantial amount of frames to be generated for each event which has more similarity in content. So it becomes difficult to identify the boundaries or shots between frames rather than to find in other scene type like already prepared slides or notes. But the retrieval is only done with the content and no further processing is done for efficiency of the search. [6]

b. Automated Tagging

In [11], the author has extracted information from video lectures by speech recognition and slide transition by the OCR tool which gives the accuracy for audio retrieval as 45%. Also, in [12], the author has used the MPEG-7 as a metadata framework for inserting collaborative information from multimedia video files. Then the annotation and tagging is provided based on this information which is extracted from the vocabulary defined in linked data. The annotation done by this is time consuming and cost inefficient.

In our paper, we are first of all doing the segmentation as a preprocessing task and then applying OCR and ASR methods for creating the metadata. The TFIDF score is calculated and based on the keyword frequency the clusters are formed. While searching for the video file, first of all the clusters are identified then all the video files within the cluster will be returned as a result.

3. Proposed Model

A. System Architectures

The proposed method is used for automating the indexing for video lectures with the information attached to it, created from content. The major sources of information about the video lecture are the lecturers' talk and slides used for explaining the topic.

The objective of the system is to provide the efficient search for the video lectures through the browser which can be done by adding more relevant metadata with the lecture video files. The metadata can be added by capturing the information from audio and video frames. For audio extraction we have used the ASR technique and for content retrieval from the frames we used the OCR technique. Also in this paper we have used clusters for returning the result with linked video files. For this purpose, we have implemented a model which separates frames from a video for keyframe identification. All the captured frames are then classified according to the duplication property. The keyframes are identified and will be used for the information extraction. We fetch all the text from these frames by using OCR strategy. Also we extract audio information by using ASR technique. The collected information (Text and Voice from Video) is used for creating metadata and clustering of terms according to their text and voice parameters.

The architecture of the proposed system can be shown in figure 3

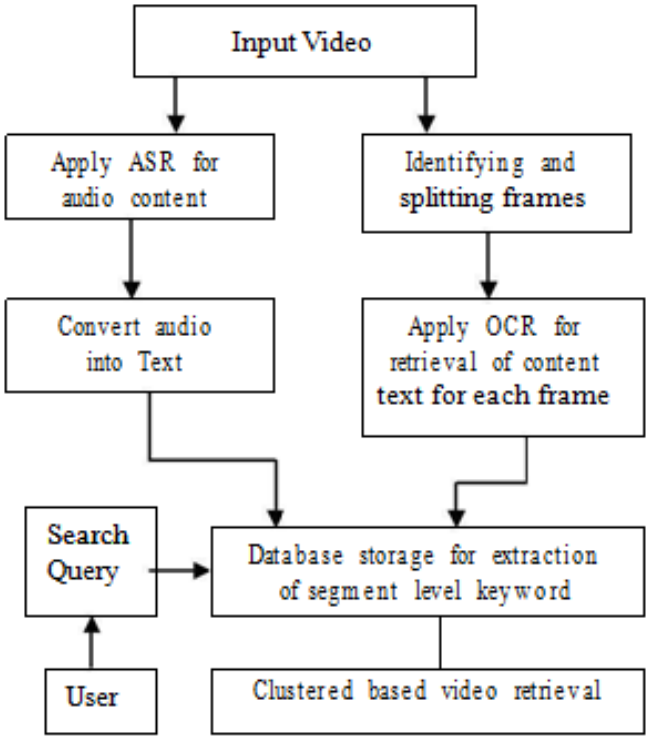

Figure 3: Proposed System Architecture

As shown in the architectural diagram, the metadata extraction is done from visual as well as audio resources of the video lectures by using OCR and ASR technique. The dictionary software is used for identifying valid words collected from both techniques. For the evaluation purpose of the system the existing systems has used the different automatic indexing techniques which are assigning the collected text to that particular file only. In the proposed system, we have used the clustering technique, which automatically forms the clusters depending on the frequency of the extracted words by ignoring stopwords. Based on that, the result will be returned to the user's query which is more appropriate and more related to the user's requirement.

B. Implementation Modules

The modules involved in the proposed system are as described below:

a). Slide Segmentation

When the frames are separated from the lecture video files, many of the frames we get which are similar in its content due to monotonous scene view. This property of the video files is different than other multimedia video files in which the frames are classified based on difference in the scene structure. The segmentation has to be done which will be for the same title or subtitle on the slide. The segmentation method is used for the video lecture files which identifies the difference among frames by connected component analysis. The connected component analysis, similarity is identified based on the group of pixels. The segmentation method consists of following steps:

- The time interval from three seconds to five seconds is provided for analysis of frames. The frames coming after the given time period are considered for further processing and others are discarded, assuming that the frames coming within interval are monotonous. Sometimes, same frame is displayed for a longer period of time then to reduce duplication we will have to increase the time interval of video segmentation.
- Canny edge maps are created, which builds the pixel differential images from edge maps.
- Then, for the similarity detection between the content of the frames is done by connected component analysis. The number of CC will act as a threshold value for segmentation. The further segmentation is done only when the number of CC exceeds this threshold value.
- In the next step of segmentation, the title and content region are first defined. Any small change in the title region may cause slide transition.
- Again the threshold value is calculated and slide transition occurs when the difference among object regions exceeds this threshold value.

b) OCR Technique

Optical character recognition retrieves the text information from images and converts it into editable text. In this paper, we have used the Tesseract OCR which is the open source, freely available and platform independent tool for extracting the data from video slides. For Tesseract the image needs to be converted into binary format. [5] For effective result to be gain by OCR tool, we need to do some pre-processing task to the keyframes. The pre-processing task is done which identifies the keyframes from the video

files.

The steps to be followed by the OCR tool are:

- The first step is Adaptive Thresholding, which converts input image into binary format.
- Next step is connected component analysis, which can be used to detect character outlines.
- Lines and words are analyzed within fixed area or equivalent text size.
- Character outlines are organized into words by two passes. In the first pass, the word is recognized by text and is passed to an adaptive classifier. In the second pass, the adaptive classifier will have training dataset provided which can be used to resolve various issues for text extraction from images.

b) ASR Technique

The automatic speech recognition (ASR) technique extracts speech or voice from multimedia files and converts it into meaningful textual information. 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. Unfortunately, the ASR technique is still under development and is not providing the efficient results expected. The Word Error Rates returned by the existing systems are not as expected. ASR is aimed to enable computers to recognize speaking voice characters without human intervention.

Automatic Speech Recognition model mostly uses the probabilistic approach for identifying original word. When the word or word sequence is pronounced its score is calculated by using acoustic properties of phonemes for matching the word with the speech signal. [13] The ASR model has the following steps:

- Pre-processing
- Feature Extraction
- Decoding
- Post-processing

The preprocessing is done for removing the unnecessary sound in speech like background noise, door closing voice etc. The high pass filtering method can be used for reducing this noise and for identifying the speech and non-speech segmentation. Signal energy based algorithms can be used for identifying the start of the speech segments By this algorithm the speech segment can easily be detected when it crosses the given signal threshold value. As there may be some small energy signals with pauses between the words, the algorithm must be enhanced by time windowing.

For extracting the features, the acoustics observations have to be extracted from a time frame of uniform length 25ms. From this time frame a multidimensional acoustic vector is calculated. Human ear can respond to the non-uniform frequency bands. The band-pass filtering can be used for non-uniform frequency bands by defining the frequencies in Mel scale. By discrete cosine transformations, the spectrums created are correlated. As the first coefficients carry the most significance, they are selected to form feature vector. Resulting features are called as Mel cepstra, for which the further processing is done by cepstral mean subtraction. The vector created will be of high dimensionality. To project it into lower dimension, the algorithms like principle component analysis is used.

Decoding is the process of matching the sequence of words with the acoustic that is represented by feature vector. The prerequisite for the decoding is availability of the dictionary which has words to be spoken with its phoneme sequence. Three information sources must be available for decoding

- An acoustic model with an HMM for each unit
- A dictionary with list of words and phoneme
- A language model with word or word sequences.

For improving the recognition accuracy, rescoring is done by higher order language model. The triagram model based rescoring is done in this system.

d) Database Creation

The databases are created which are collected from OCR, ASR and also from the metadata which is manually created for the video files. This database information will act as a metadata for that video files. The important aspect of our system is we are storing this metadata in memory rather than storing in somewhere else. Due to this, the time for the search gets reduced. When the user will fire the query for accessing the videos, due to in memory storage the search time reduces significantly. The redundancies are checked between the words collected for avoiding the wastage of storage space.

e) Clustering for efficient Search:

The databases collected from the above sources are large in size, in which all characters and words are present including insignificant words like stopwords. The stopword removal technique is used which gathers only the important characters. The search is related to all the video files for the same query for which the dataset collection has to be arranged efficiently. For efficient finding of video files the clustering methodology is used. The clustering is done based on the frequency of the terms. The TF-IDF score is calculated, which gives the value of the word by its term frequency and inverse document frequency. By these calculations we can exactly identify the important words. These important words are again clustered for different video files for returning the result with all its related videos. Here, we are using k-means clustering algorithm which forms the clusters will be formed based on term frequency.

4. Result Analysis

The comparison between the results of existing system and proposed system is as shown in the figure 4 and figure 5 by considering two parameters, accuracy and time for search. The graphic representation in the figure 4 shows the efficiency of the proposed system with the existing system. For the small search query length the accuracy provided by the existing system is nearly 60%. But the proposed system is giving nearly 80%, which is far more efficient than the

existing system. Also as the length of the search query increases, the accuracy results are also get improved proportionally for the proposed system. If time required for search is considered, then too our system providing more efficient result rather than existing systems. This shows the existing system is much efficient for all kind of query length and providing more accurate result than the existing system.

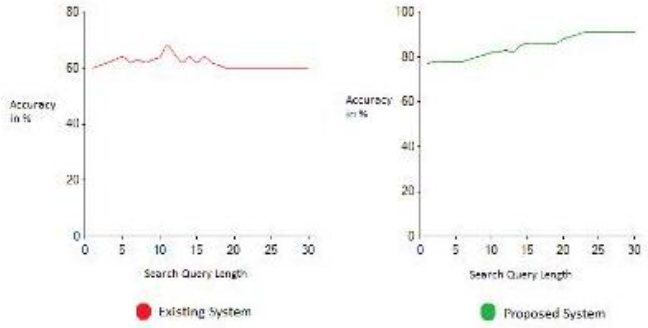

Figure 3: Result comparison between existing and proposed system by considering Accuracy parameter

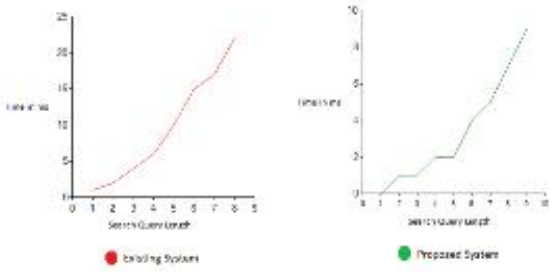

Figure 4: Result comparison between existing and proposed system by considering time for search as a parameter

5. Conclusion

In this paper, we have developed an efficient information retrieval method from video lecture files. The existing systems are doing extraction based on various methods which provides the WER nearly 71%. But as here we have used Tesseract OCR with ASR technique with preprocessing methods, which reduces word error rate much and also the clustering method used after data extraction, which is based on ontology of the terms, provides much related and desired metadata for the file and linked information for the videos. So the video lecture browsing becomes efficient. This metadata created can be treated as indexing to that video file which is automatically created. As the results shows, the proposed method also provides more accuracy with least search time. The future scope of this system will be to retrieve the information for the handwritten text which can be combined with the audio text.

References

[1] Haojin Yang and ChristophMeinel," Content Based Lecture Video Retrieval Using Speech and Video Text Information",
IEEE Transactions On Learning Technologies, Vol. 7, No. 2, April-June 2014.
[2] Alexander G. Hauptmann, Rong Jin, and Tobun D. Ng, "Video Retrieval using Speech and Image Information", Electronic Imaging Conference (EI'03), Storage Retrieval for Multimedia Databases, Santa Clara, CA, January 20-24, 2003
[3] Wolfgang Hürst, Thorsten Kreuzer, Marc Wiesenhütter , "A qualitative study towards using large vocabulary automatic speech recognition to index recorded presentations for search and access over the web"
[4] E. Leeuwis, M. Federico, and M. Cettolo, "Language modelling and transcription of the ted corpus lectures," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., 2003, pp. 232–235.
[5] Madhav Gitte1, Harshal Bawaskar2, Sourabh Sethi3, Ajinkya Shinde4 "Content based video retrieval system", International Journal of Research in Engineering and Technology, Volume: 03 Issue: 06 | Jun-2014
[6] Chirag Patel, Atul Patel, Dharmendra Patel, "Optical Character Recognition by Open Source OCR Tool Tesseract: A Case Study", International Journal of Computer Applications (0975 – 8887) Volume 55– No.10, October 2012
[7] Tiecheng**Liu** and John R. Kender, "Rule-based semantic summarization of instructional videos", IEEE International conference on Image Processing, Vol 1, 2002
[8] Duc Phuong Nguyen, Martin Guggisberg, Helmar Burkhart "Multimedia Information and Mobile-Learning", Eighth IEEE
[9] International Symposium on Multimedia (ISM'06), 2006
[10] Matthew Cooper, "Presentation Video Retrieval using Automatically Recovered Slide and Spoken Text", Multimedia content and mobile devices, SPIE proceedings vol. 8667, 2013
[11] Xiaoqiang Xiao, JashaDroppo and Alex Acero" Information retrieval methods for automatic speech recognition", IEEE international conference on Acoustic Speech and Signal Processing, 2010
[12] Vijaya Kumar Kamabathula, Sridhar Iyer, "Automated Tagging To Enable Fine-Grained Browsing Of Lecture Videos" , IEEE International Conference on Technology for Education (T4E), 2011
[13] Harald Sack, Jörg Waitelonis, "Integrating Social Tagging and Document Annotation for Content-Based Search in Multimedia Data", Proceedings of the 1st Semantic Authoring and Annotation Workshop (SAAW'06), Athens (GA), USA, (November 2006) issn=1613-0073
[14] Dong Yu, Li Deng, "Automatic Speech Recognition", Springer Signal and communication technology, 2012