A Survey on Automatic Video Lecture Indexing

Ashwini Y. Kothawade, Dipak R. Patil

Department of Information Technology, Amruvahini College of engineering, Sangamner, India

---

Abstract - *Popularity of e- learning is increasing day by day due to its various advantages than classroom learning. It gives very efficient and flexible learning because it is independent of location and time. Many organizations are involving in this, by uploading their lectures on internet. The data on the internet is also increasing due to this uploading of lectures by various universities and organizations. Therefore, the accuracy efficient searching becomes less which only searches from metadata attached to the file. The need arises for the effective indexing to each video lecture file. In this paper, the overview of different indexing techniques with its accuracy results has been given studied.*

Key Words: e-lecture, video indexing, OCR, ASR, Content retrieval, e-learning, tele-lecture, video lecture retrieval, video metadata creation.

1. INTRODUCTION

When streaming media is considered, e-learning is popularly use due to easy audiovisual recordings and flexibility given to student. He can learn at anytime, anywhere their interested topics by avoiding the cost of traveling. It can also be made available as an additional material for the course being studied. Near about thousands of videos courses are provided with full content of the course by the different organizations like NPTEL, MOOC, MIT OpenCourseware etc. However, a general video lecture on one hour produce a large video file based on the recording technique used. Therefore, the data is increasing on the net and within this data to extract only the desired information quickly becomes difficult. To avoid this different techniques and approaches are used which can extract the data ontologically from video files and assign indexing to the file based on that extracted contents.

The existing multimedia video retrieval techniques are not applicable to video lectures because the contents extracted are from feature extraction and by 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 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.

Therefore, 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. Many techniques have been developed yet which extracts the content from video lecture file and analyzes the words talked by the speaker. There are two major sources of the information content, one is from the speaker who speaks and another is from their slideshow content or handwritten text written on the board. While retrieving this information many challenges have to be faced by the researcher as described in [8].

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(a) or having discussion after finishing the slide presentation as shown in fig. 1(b). [1][9]

Figure 1(a): A recent video lecture image from MIT

For these kinds of video lectures as in 1(a), MIT has suggested the online version of 2.002, which provides the search by keyword and organized as a set of basic concepts in a tree structure format with subtopics are shown by branches of the tree.[6]

Figure 1(b): The video recording may be of the discussion with audience in the question answer session.

Extensive research has been done for retrieving the content and providing the appropriate indexing. Only to extract low level features or to extract the syntactic features is not efficient for the instructional videos. The sources of information in lecture videos are instead the slides, handwritten text on blackboard and audio, for which the extraction techniques are capturing the keywords from slides, blackboard, whiteboard and audio tracks [8]. Poor handwriting, poor pronunciation, change of presentation format and lack of accuracy in recognizing texts are major challenges in video lecture indexing. Recent video lectures are also demanding for advanced interfaces and browsers for visualizing content and displaying liked information [8].

For retrieving the content the OCR for video frame content and ASR for audio content are used broadly in different ways In this paper, we will discuss about these issues, techniques used for indexing with steps.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, which 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. The scenes may have variability in its composition as shown in figure 2 which makes the retrieval more difficult.

The retrieval of text is especially done by Optical Character Reorganization (OCR) and Automatic Speech Recognition (ASR). The technologies have been developed which are using different tools for efficient retrieval.

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

1.1 General Architecture

Generally, all the systems follow the steps for retrieving the information as shown in figure 3. 1) Video Segmentation which separates the video lecture frames into frames by identifying the keyframes. 2) Text extraction from slides which uses OCR technique for retrieving the text from slide or handwritten text from blackboard. 3) Text extraction by ASR technique which extracts the information from the audio of the video file. 4) Post processing the data which collects the text into efficient way and provides easy retrieval.

The architecture, that is commonly used for indexing is as given in figure 3.

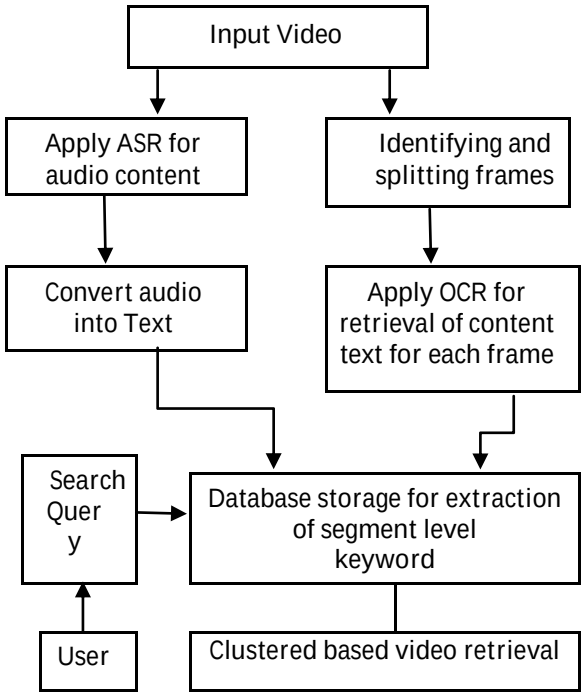

Figure 3: General Retrieval System Architecture

1.2 Overview of results of accuracy

The following table shows the different research works being done with their results. Some systems are retrieving only from either video text or audio while some are retrieving from both, audio and video text but providing less accuracy. In this paper, different approaches are overviewed with their technique.

Table -1: Accuracy results of different lecture video indexing techniques

Reference Number	Year	Accuracy percentage	
		Speech retrieval accuracy	*Video retrieval accuracy*
[19][14]	2003, 2008	About 40%	--
[9]	2005	49%	--
[3]	2002	About 32%	46.77%.
[26]	2014	46.01%	60.99%
[1][2]	2014	About 52%	85%
[28]	2011	--	97.01%
[29]	2011	20-40%	--
[21]	2011	--	74%
[23]	2002	50%	--
[32]	2013	56.3%	36.5%
[33]	2008	35%	70%

2. SEGMENTATION

Segmentation is very important in accessing the content from video. As video lectures are homogeneous in scene type so proper segmentation is needed for accessing keyframes. The segmentation is needed in both audio and video files. For the video segmentation the similarity is calculated based on the content on the slides and for audio files the extracted and converted text are being used.

2.1 Video segmentation

In video segmentation, the frames are separated from the video and keyframes are identified. Keyframes shows the different content than the previous frames. When proper segmentation is done the exact keyframes are selected which doesn't miss any important texts. Generally for video lectures the text contained in frame can be used for keyframe identification. Shot boundary detection is done by identifying the difference among four second or three second windows and which is compared with mean across both windows or the predefined threshold value[1][9][15]. If the difference among slides deviates with this value then the slide transition is made.

The edge detection algorithm is used by compass mass template matching technique with static or adaptive threshold value. The detected edges are localized by determining the location of edge and stored into the edge map. A morphological process dilation and erosion is used for detecting non text region [22].

The connected components are the maximal connected region by the pixels. The edge detection and edge maps are created. And finally the localization scheme will be applied for identifying the texts which separates out non text blocks or background region [3]. Connected component analysis is done for each adjacent frames of the video file. The component-level-differencing metric is used for differentiating the text. The number of connected components is used as a threshold value for segmentation. The new frame will be captured when it exceeds the threshold value. Then slide transition is done which identifies the keyframes by its title and subtitle matching. This matching is done by comparing the height of the text or occupancy of the region of the entire frame.[2]

When the text extraction is related to handwritten text, then the segmentation or preprocessing strategy is different depending on the type of video. The differentiation can be done by separating background region and foreground region. It uses foreground object detection model which detects both, stationary and moving object by its color feature and color co-occurrence feature respectively. The Bayes' theorem will be applied for posterior probability of foreground and background object. [4]

Jacob Eisenstein and Regina Barzilay [13] used a model of conference resolution and gesture salience for selecting keyframes. By predicting the instance where the gesture recognition is necessary author has built a model of conference resolution. The model is then augmented with the hidden variable which decides the efficiency of the gesture feature at each instance and also predicts the importance of gesture at a particular instance.

2.2 Audio segmentation

In audio segmentation, the audio waves are captured, translated into text and then the separation is done to avoid redundancies. Each transition in the speech (e.g. pause) can lead to segment boundary. The wireless microphone is generally used for capturing the audio signals. The audio segmentation mostly done by the speaker, his pauses break during speech etc. The vectors of 13 Mel Frequency Cepstral Coefficients are calculated by sampling the audio signals at regular intervals. Then the Bayesian Information Criterion can be used for detecting speaker changes. If a positive maximum value is

found among these BIC values then the speaker change has been achieved clearly [9][10]. With MFCC, the perceptual features are also used for feature extraction. Lie Lu, Stan Z. Li and Hong-Jiang Zhang[16] divided audio signals into one second audio clip. The support vector machine concept is used which classifies the data into speech and non-speech classes. For each frame, the feature extraction is done by the MFCC and perceptual features like short time energy (STE), zero crossing rates (ZCR), sub-band powers distribution, brightness, and width and the pitched ratio.

The supervised techniques can be used for classifying. The training is provided on the commonly observed characteristics of the lecture speech like variation in delivery style, student interaction etc [17][18].

3. TEXT EXTRACTION

After doing the segmentation, the next important phase for achieving the indexing is the extraction of text from the major sources of video files, that is from the material the speaker is using for teaching reference and the audio speech generated by the speaker.

3.2 Video Text Extraction

Initially, for extracting the content, the teaching focus is used which can be found by gesture of the speaker, teaching time of each slide and speed of speech. This focused analysis provides more accurate results. [27]

The information is collected is either from the slides, notes or handwritten text on the board. The Optical Character Recognition (OCR) technique is used to extract the text content from the video files. After the edge based shot boundary detection, the text extraction is done for the slides which the speaker has more focused. This can be calculated by the time for explaining the slide and the content on which the speaker is emphasizing by his fingers. [15]

After the video segmentation has been performed, the video frame structure is analyzed [20]. The localization verification scheme is used for detecting text regions and characters. Further the stroke width transform (SWT) is used to avoid the non-text region and noise. The SWT algorithm computes the width of the stroke containing the pixel of non-text regions like tree, sphere and window. The output we get from SWT is a feature map where each pixel contains the potential stroke width value of input image pixels [1][20]. The detection accuracy can be further improved by using support vector machine classifier (SVM) with radial basis function as kernel [1].

The text localization can also be done by OCR with weighted DCT (discrete cosine transformation). The DCT text detector does not changes actual font size and style and also gives suitable runtime efficiency. The DCT text detector is computationally efficient compared to classifier text detector which needs a training period [19]. After the text detection by DCT the refinement algorithm is used to extract proper text, which based on vertical and horizontal distribution of image and dynamic thresholding algorithm can be used for deciding the refinement. And the noise can be removed by text box verification procedure [21]. To overcome the poor resolution dynamic contrast and brightness adaption is done enhancing the text quality [2].

The important information can also be hidden video file in the graphical or image form. In [26], author has developed systems for multimodal indexing using bag of subjects' model which improves the accuracy by returning part of interest. Also in cross model the facility is provided for giving input in the graphical form. In [28], the preprocessing is done with image binarization, edge detection, resizing, inversion and then the OCR tool is applied. Tesseract, MODI, GOCR tools are used for extracting texts this preprocessed images which gives more accurate result.

3.2 Audio Text Extraction

The Automatic Speech Extraction (ASR) system can be used for extracting the text from the audio data. There are many challenges in this approach. Julian David Echeverry Correa allows the segmentation by supervised learning and the topic selection is done by TFIDF score with ASR [18]. The Term Frequency and Inverse Document Frequency (TFIDF) calculate the relative frequency of the keywords within a file and it returns the result for the most occurred element excluding stopword. The keywords are identified in the document by the normalization. But, the system is not applicable to singular and plural text similarity. Generally, the ASR system works in two components; the front end and the decoder. For gaining appropriate representation of speech, the front end extracts features from input speech signal which is sampled by time domain waveform. The decoder generates the most probable word sequence from the extracted feature by front end [24].

Most systems are using ASR with IR methods. James Allan [23] claimed that the recognition errors generally occur in audio files though ASR and IR (Information Retrieval) techniques. It can be reduced if small span of text is

available which reduces recognition errors and redundancies. The challenges come in the use of important words in speech and the user interface. With using ASR with IR, the search strategy can be changed by retrieving snippet rather than retrieving through segment [24].

Large vocabulary automatic speech recognition (LVASR) transforms the audio signals into a text with more accuracy as the word length increases. The accuracy can be improved in text recognition by providing 1) the appropriate vocabulary for the words that are often used 2) classification of topic-related words in vocabulary [3].

4 VIDEO INDEXING

After entering the query the search will be done in different levels. Video indexing means providing the access points to facilitate retrieval of information. At first, the relevancy of related slides, then corresponding audio clip and at last the manual transcripts is calculated by TF-IDF (Term frequency-inverse document frequency) [3]. TF-IDF measures the distribution of search terms in the document. Precision and recall values are used for evaluating the performance where precision p and recall r being defined as p = (no. of retrieved relevant documents / total no. of retrieved documents) and r = (no. of retrieved relevant documents / total no. of relevant documents), respectively. For the speech recognition performance word error rate (WER) is used and that depend on the language model used and the training vocabulary provided.

The collected text by OCR and DCT based text detector are further processed by time-based text occurrence information. The analyzed text contents by removing stopwords and the collected keywords are further used for video indexing [2]. The context and dictionary based approach can be used for the accurate detection of text and to identify the required keywords [20]. In [26], results are compared for different indexing techniques like multimodal, cross model, using bag of mixed words and bad of subjects. Within that, the multimodal retrieval with bag of subjects is giving more accurate result.

The clustering can also be used for providing the relevant indexing. The bottom-up approach of collecting the text from audio first and then it is related with the video content. The accuracy in clustering can be increased by additional cues from the video segments like motion activity by the speaker [9]. In speech transcripts, the clustering is used by chaining in which chains are formed by collecting the accumulated keywords that appear in a document. The words are collected by the frequency of occurrence of the word in the whole document [25]. In [29], speech recognition software generates timestamp for each word. The relationship between these timestamp can be used for forming clusters which separates the parts of the video lecture by its topic and then phrasing is done which provides link to specific result and also to the specific part of the video file. The parts of the speech are collected from the clusters.

The IR system used with ASR provides the indexing for efficient retrieval to user's query. After ASR decoding the preprocessing is done which removes stemming, stopwords etc. Part of speech (POS) tagging is applied for identifying the weight of each word in the query text and distribution of word in audio document. This technique is called as spoken document retrieval (SDR). For the Spoken Text Detection (STD), the search separated into two parts; Indexing and Search. In Indexing the lattice of each hypothesis is expanded into a finite state transducer. The start time and end time of the word or word sequence are encoded. At the Search stage, the in-vocabulary queries are compiled into finite state acceptors with zero cost and out of vocabulary queries are compiled in non linear finite state acceptors with different costs [24].

The search can be further improved by providing the community rating to lecture content. The rating is provided in different layers of the video lecture [30]. The vector space model for the vocabulary in which space density computation is performed by computing cosine similarity measures. It is done for assigning automatic indexing to a collection of documents [20].

Another method for assigning the indexing is the use of tagging which is popularly used recently. In [11][12], the tagging is assigned to the lecture videos by capturing the content from speech transcript or slides. And the result will be returned to the users' query in thumbnail form which is focused on the users' interest. For video, the collaborative tagging can be used with MPEG-7 metadata which is assigned for a single scene.

5. CONCLUSIONS

We have presented the overview of different technologies regarding automatic indexing to video lectures. The texts are extracted from the content of video lecture in different ways and again the keywords are collected efficiently and used as a metadata for the file. Instead of extracting the text only from video files if the extraction is done with audio too. It provides more accuracy for the search.

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] Haojin Yang, Maria Siebert, Patrick Luhne, Harald Sack, Christoph Meinel "Lecture Video Indexing and Analysis Using Video OCR Technology4] handwritten text" , 2011 Seventh International Conference on Signal Image Technology & Internet-Based Systems.

[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" , *ICWI, page 135-143. IADIS,* (*2002*).

[4] Ali Shariq Imran, Sukalpa Chanda,Faouzi Alaya Cheikh, Katrin Franke, Umapada Pal "Cursive Handwritten Segmentation and Recognition for Instructional Videos", 2012 Eighth International Conference on Signal Image Technology and Internet Based Systems, 2012.

[5] Madhav Gitte, Harshal Bawaskar, Sourabh Sethi, Ajinkya Shinde "Content based video retrieval system", International Journal of Research in Engineering and Technology, Volume: 03 Issue: 06 | Jun-2014.

[6] Jennifer Chu, April 3, 2013, "A new wrinkle in online education : An experimental online course gives some students scheduling freedom", Retrieved from file:///H:/iPgCon/ME_Project/Ref_Project/Introducti on/ A%20new%20w. rinkle%20in%20online%20education%20_%20MIT% 20News.htm.

[7] Dong Yu, Li Deng, "Automatic Speech Recognition:A Deep Learning Approach", Springer Link, Signals and Communication Technology,2015.

[8] Tiecheng Liu, "Lecture videos for e-learning: current research and challenges", IEEE Sixth International Symposium on Multimedia Software Engineering Proceedings, 2004.

[9] 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

[10] S.S. Chen, P.S. Gopalakrishnan, "Speaker, environment and channel detection and clustering via the Bayesian Information Criterion," *DARPA Broadcast News Transcription and Understanding Workshop*, Landsdowne, VA, pp. 127-132, 1998.

[11] Kamabathula, V.K, "AUTOMATED TAGGING TO ENABLE FINE-GRAINED BROWSING OF LECTURE VIDEOS", IEEE International Conference on Technology for Education (T4E), 2011.

[12] H. 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.

[13] 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.

[14] 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.

[15] Yu-Tzu Lin, "Structuring And Analyzing Low-Quality Lecture Videos", IEEE International Conference on Acoustics, Speech and Signal Processing, 2009. ICASSP 2009.

[16] Lie Lu, Stan Z. Li and Hong-Jiang Zhang ,"Content-Based Audio Segmentation Using Support Vector Machines", IEEE International Conference on Multimedia and Expo, 2001.

[17] Balagopalan, A. ; Balasubramanian, L.L. ; Balasubramanian, V. ; Chandrasekharan, N. ; Damodar, A., "Automatic Keyphrase Extraction and Segmentation", IEEE International Conference on Technology Enhanced Education (ICTEE), 2012.

[18] M.C. Benítez-Ortúzar, J.L. Pérez-Córdoba (eds.). "Applications of Speech Technologies: Talks and Contributions presented at the summer course: Applications of Speech Technologies". J.D. Echeverry as Author of a chapter in book: "Audio-Speech segmentation and Topic Detection for a Speech-based Information Retrieval System". Ed. Universidad de Granada, ISBN 978-84-338-5596-1, 2013, pp 279-291.

[19] 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

[20] G. 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.

[21] Haojin Yang, Maria Siebert, Patrick Luhne, Harald Sack, Christoph Meinel, "Automatic Lecture Video Indexing Using Video OCR Technology", IEEE internation symposium on Multimedia, 2011.

[22] C.P.Sumathi, N.Priya, " A Combined Edge-Based Text Region Extraction from Document Images", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 8, August 2013 ISSN: 2277 128X.

[23] James Allan, "Perspectives on Information Retrieval and Speech", Information Retrieval Techniques for Speech Applications : Lecture Notes in Computer Science Volume 2273, 2002, pp 1-10, 2002.

[24] Justin Chiu ,"Speech Retrieval under Limited Resources and Open Domain Conditions", June 2014.

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

[26] Nhu Van NGUYEN,"Multi-modal and cross-modal for lecture videos retrieval", 22nd International Conference on Pattern Recognition, 2014.

[27] Yu-Tzu Lin, Greg C. Lee, Bai-Jang Yen,Chia-Hu Chang,Huei-Fang Yang, "Indexing and Teaching Focus Mining **of Lecture Videos",** 11th IEEE International Symposium on Multimedia, 2009

[28] Tayfun Tuna, Jaspal Subhlok, Shishir Shah, 'Indexing and Keyword Search to Ease Navigation in Lecture Videos', IEEE Applied Imagery Pattern Recognition Workshop, 2011.

[29] Stephan Repp, Christoph Meinel "Semantic Indexing for Recorded Educational Lecture Videos", IEEE ,Applied Imagery Pattern Recognition Workshop (AIPR), 2011.

[30] Moritz, F. ; Siebert, M. ; Meinel, C., "Improving community rating in the tele-lecturing context ",International Conference for Internet Technology and Secured Transactions (ICITST), 2010.

[31] X. S. Hua, X. R. Chen, L. W. Yin, H. J. Zhang. "Automatic Location of Text in Video Frames," in *Proc. of ACM Multimedia 2001 Workshops: Multimedia Information Retrieval*, 2001,pp. 24–27.

[32] Matthew Cooper, "Presentation Video Retrieval using Automatically Recovered Slide and Spoken Text", SPIE Electronic Imaging 2013.

