Lecture Video Retrieval Using Audio and Text Transcripts

Mrs. S. Vanitha Sivagami (Associate Professor)
Computer Science & Engineering,
Mepco Schlenk Engineering College,
Sivakasi, Tamil Nadu, India
svanitha@mepcoeng.ac.in

Anisha VP (UG Student)
Computer Science & Engineering,
Mepco Schlenk Engineering College,
anishaa.vijayakumar@gmail.com

Jeyavani B (UG Student)
Computer Science & Engineering,
Mepco Schlenk Engineering College,
jaymep.jb14@gmail.com

Abstract- **Nowadays, most of the universities as well as organizations record their lectures and make it available online for learners. Hence, indexing and searching of videos have become a tedious process with increasing amount of video data over the internet. In this lecture video retrieval system, the content inside the video is analyzed and the relevant videos are fetched. First, the videos are split into frames and then subjected to edge detection and connected components analysis to extract keyframes. The text content is extracted from keyframes by using optical character recognition technique. The recognized text is stored as transcripts. Simultaneously, the audio content of the video is also extracted and converted into text using speech to text service. The speech and text transcripts are then subjected to further processing, where keyword extraction is done using part-of-speech tagging. Then, term frequency score is calculated for the extracted keywords. Finally, the training phase ends with storing the keywords along with their frequency score. In the next phase, the user inputs a keyword to the system. The system performs a search of the keywords from the created database, and provides those relevant resultant videos based on the term-frequency score of the keyword in videos.**

Keywords – **video indexing, lecture videos, content based video retrieval, image processing, keyword extraction, transcription.**

I. INTRODUCTION

In most of the video search engines, the searching is based on tags, annotations and the meta data associated with the video. These are usually given by the author who creates it or uploads it to the internet. In some cases, the author might be discussing a topic for only few seconds. This video will also be fetched when a user searches for the particular topic, as it would have been included in the metadata. However, user gets upset after watching an entire video only to find out a meagre content related to the topic specified by him.

In our proposed system, this constraint is eliminated by analyzing the text inside each and every frame of the video, which crisply presents the essence of the subject matter. The audio content is also analyzed so that any keywords in it are not missed.

II. LITERATURE REVIEW

B. V. Patel et. al. (2010) proposed a video retrieval system based on entropy, black and white points on edge, and features of video key frames [1]. By using entropy feature, keyframes from the video are extracted and video feature database is created and also the features are extracted for the video query. Based on the similarity measure the output video is retrieved. But it has very few feature extraction methods which cannot describe all the details in the video.

Another system retrieves similar videos based on local feature detector and descriptor called Speeded-Up Robust Feature (SURF), which is studied by Dipika H Patel (2015) [3]. Here, the SURF feature descriptor is extracted from the identified keyframes. The features are also extracted for input video clip. The retrieved videos are ranked based on its similarity to test clip. However, both detector and descriptor does not use color information which impose a limitation on this method.

Another research by Tuna et. al. (2012) [5] reports on the development and evaluation of ICS (Indexing, Captioning and Search) videos framework and assessment of its value as an academic learning resource. It uses ICS videos, i.e., videos enhanced with Indexing, Captioning, and Search capability that are designed for quick access to video content. The investigation revealed that the preponderance of technical jargon and abbreviations could not provide satisfactory results for lecture videos.

Most of the existing techniques either consider only few features for segmentation to extract text from videos, or consider OCR and ASR separately. Our approach parallelizes the image processing steps, which reduces huge time consumption. Text transcripts are obtained from the key frames of the lecture videos as well as the audio content that is spoken by the lecturer. These transcripts are then used for the process of keyword extraction for retrieving relevant videos. The system proposed provides an efficient content based video retrieval. The next section details the system architecture of the proposed system, and the subsequent section deals with the implemetation results.

III. SYSTEM ARCHITECTURE

This section deals with the architecture of the proposed system, and explains the sequence in which data flows within the system. The system has two major phases as described below.

1) Training Phase

The proposed system has an administrator, who creates a keyword storage from all the training videos to the system during the training phase. This system flow of the training phase is depicted in Fig. 1.

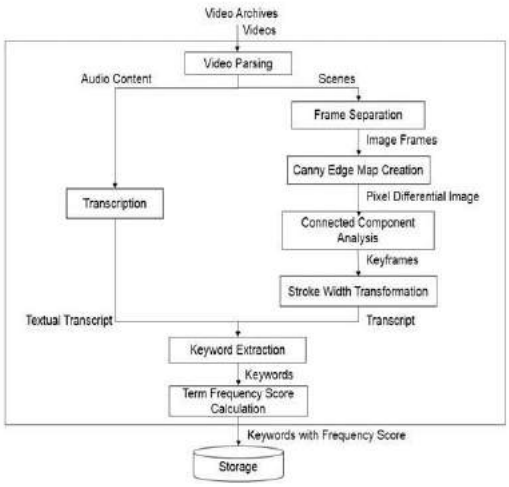

Fig. 1: System flow diagram of the training phase

a) Video Parsing

In this parsing process, first, audio is extracted from each video. Simultaneously, video is also split into a sequence of image frames.The length of videos in the archive differs, hence the number of frames in it. The number of frames in each video is determined by the frame rate of the video. In this system, the jumping interval for the analysis of frame extraction is set as 3 frames, based on the research in [2].

b) Canny Edge Map Creation

The regions with strong intensity contrasts in images are characterized as edges. The process of edge detection filters out useless information and reduces the amount of data significantly. However, the important structural properties in an image are preserved in the process. After edges are created on all the generated image frames, they are fed as input to the connected component analysis module.

c) Connected Components Analysis

Once region boundaries have been detected, connected component analysis is used to extract regions which are not separated by a boundary. The set of connected components partition an image into segments. Image segmentation plays a vital role in the CBVR system in identifying text regions. This segmentation is performed using the contour tracing technique as proposed by Fu Chang et. al. (2004) [4]. Once the components are labelled in each frame, the number of components are compared between consecutive image frames. If this difference is greater than a certain threshold say 25, then the frame is captured as a new segment. This process avoids redundancy among frames, thus easing further processing.

During this process, face recognition is also performed on segments. The area covered by the face is computed and compared with the size of the image. The proportionate size is considered and compared with a threshold value. This segment is then eliminated if it exceeds the threshold, on the assumption that the segment doesn't contain any text.

d) Stroke Width Transformation

Stroke width transformation is a part of Optical Character Recognition (OCR) to extract text from the image. Through OCR the non-text parts are eliminated and the text is retrieved. This CBVR system uses Tesseract OCR Engine for text identification. This engine itself performs the SWT on image frames and retrieves text from the input key frames and stores it as a text document.

e) Transcription

Transcription is the process of converting speech to text document. In this system, speech recognition process is carried out manually using IBM Watson's speech to text service. The audio extracted from the video during video parsing is uploaded to the speech to text service. The service continuously returns and updates the transcription as more speech is heard. The service produces transcribed words as output, which is then stored as a text file. This text file is then subjected to part-of-speech tagging, along with the text file that is obtained from OCR.

f) Keyword Extraction

The speech and textual transcripts are first combined to form a textual corpus. This corpus is processed by a sentence splitter. This splits the sentences by identifying delimiters like period(.) or new-line feeds. Then, the split sentences are sent to a tokenizer, which further splits each sentence to a set of words. These tokens are given as input to a normalizer. It tries to correct any misspelt token, if possible. Finally, all the tokens in the corpus are processed by Stanford POS tagger. The tokens and their corresponding tags are usually obtained as an XML file. For easier processing, this XML format is converted into text using Groovy script. The tokens and tags are stored in alternate lines of a text file. The tags and description resulted from POS tagging are filtered in this process. In this CBVR system, only nouns are considered as keyword candidate. Nouns are denoted using the tags NN, NNS, NNP and NNPS, which represent 'Noun, singular', 'Noun, plural', 'Proper noun, singular', and 'Proper noun, plural' respectively.

First, all the nouns are extracted from the collection of tokens from POS tagging process. The resulting nouns may contain redundancy of tokens within them as the corpus was built by combining the speech and text transcripts. Next, these redundancies are removed and unique keywords are obtained for each video. At the end of this process, a set of unique keywords is generated. For

each of these keywords 't', the term frequency score is computed as follows.

$$TF(t) = \frac{\text{Number of times term t appears in a document}}{\text{Total number of terms in the document}}$$

g) Keyword Storage

In this process, the keywords along with their frequency scores is mapped with the video so that the retrieval process can be effective. This is done by storing the keywords and the video in a matrix format as in Fig.2. Based on the starting letter of the keyword, storage is done in separate structures. Whenever a new keyword is to be added, it is compared with the existing keyword list in the cell array to check if it is already present.

- If it is not present already, then it is added to the cell array as a new column and its term frequency score is entered into the corresponding cell.
- If the keyword already exists then the term frequency score of the keyword will be updated for the specific video.
- If the keyword is not present in a video, then the intersecting cell will contain an empty value. The new video name will be added as a new row in the cell array.

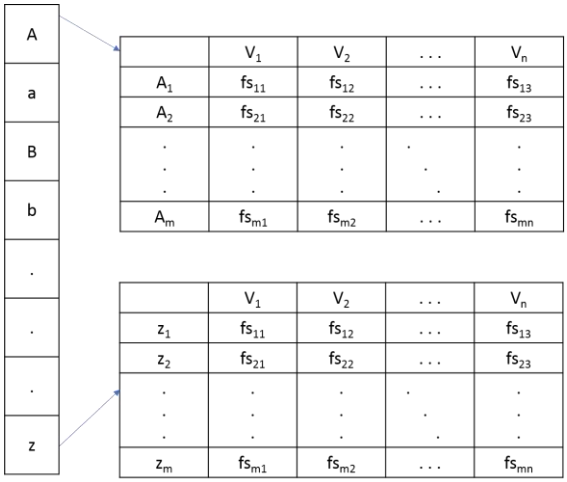

Fig. 2: Keyword – Video matrix for storing term frequency

2) Testing phase

The system flow of the testing phase is depicted in Fig. 3.

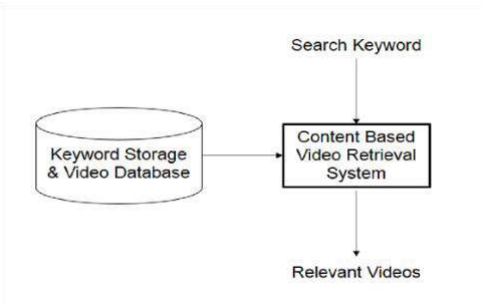

Fig. 3: System flow diagram of the testing phase

The keywords extracted from the training phase is available in the database. Now, the user inputs a query to the system for which the relevant videos are retrieved. The input query fed into the CBVR system is matched with the keywords present in the database. To reduce the retrieval time, search is performed only in the structure that corresponds to the first letter of the inputted keyword. Moreover, the list of keywords is in sorted order. So, binary search on the keyword-video matrix in Fig. 2 provides faster retrieval. Based on the matches, a list of corresponding videos is retrieved. These keywords may be found in one or more videos. The priority of the videos to be displayed first is decided by the term frequency score calculated. The videos that have a high term frequency score are retrieved first.

IV. IMPLEMENTATION RESULTS

Here, the results of the designed CBVR system that operates on speech and text transcripts to establish the system's usefulness. The videos that are trained in the system are collected from NPTEL and Coursera datasets. They belong to 6 different domains like Cryptography, Data Analytics, Principles of Compiler Design, Linux Programming and Scripting, Ground Improvement Techniques (Civil Engineering), and Economics, Management & Entrepreneurship.

The video dataset archived by the administrator is given as input for the process of frame extraction. After the process, image frames are obtained as results as in Fig. 4.

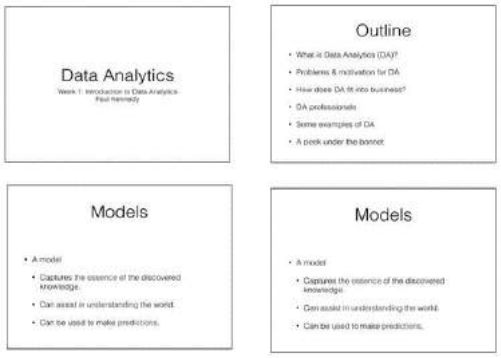

Fig. 4: Frames extracted from a lecture video

The image frames obtained from the frame extraction process are given as input to the Canny edge detection process. The results obtained from Canny process is illustrated in Fig. 5.

Fig. 5: Pixel differentiated images for the corresponding frames from video

The pixel differentiated images are captured as segments based on the differences in adjacent frames. This is performed by the analysis of connected components. Face recognition is also performed during this step to eliminate those slides that predominantly occupy faces of the lecturer. The segmented images after this process is shown in Fig. 6. This phase also reduces redundancy in frames.

Fig. 6: Key frames obtained after performing Connected Components Analysis

Textual content is captured from the key frames obtained. This is stored in a text file as textual transcript. These transcripts for the five key frames shown in Fig. 6 are displayed below from top left to bottom right.

Text from Keyframe 1:

Data Analytics

Week 1: Introduction to Data Analytics

Paul Kennedy

Text from Keyframe 2:

Outline

What is Data Analytics (DA)?

Problems & motivation for DA

How does DA fit into business?

DA professionals

Some examples of DA

A peek under the bonnet

Text from Keyframe 3:

Models

' A model

-Captures the essence of the discovered knowledge.

-Can assist in understanding the world.

-Can be used to make predictions.

The sample of a transcript obtained from ASR for a single audio file is shown below.

"For example going back to the example with the students in the class I might just be trying to understand if this natural. Structure so or substructures of students in the class. For example the students muckraking tip hot time students or full time students. Well and what brightens undergraduate students across graduate students. Well they could be another concept right down such as. Students that have a good mathematical understanding versus you. So. By the time. So, I'm always trying to capture the essence of. The world. Either by making sense or by making. Data mining is applied in lots of different areas. So, it's being applied by business government all of the areas that you can see on the slide. I was collected and the daughters collected up pretty much everything said businesses customers. Prices everything. And what we want to do this startup money well generally water better support manage. So, to make. Evidence based. It's to help them make. We might be trouble so China fund fraudulent behavior. Or we want to try to find out she. Next thing we can look at a self-motivation for doing data analytics."

The tags identified from part-of-speech tagging for a sample sentence "Captures the essence of the discovered knowledge", from the text of 3rd keyframe are listed below.

Captures - VBZ
the - DT
essence - NN
of - IN
the – DT
discovered – VBN
knowledge – NN

In the above list, VBZ, DT, NN, IN and VBN represent the parts-of-speech 'Verb, 3rd singular, present', 'Determiner', 'Noun, singular', 'Preposition', and 'Verb, past participle' respectively.

As only nouns are considered as keywords, the following list provides the keywords for the sentence that is tagged.

• essence
•knowledge

The term frequency for the keywords extracted is calculated for each keyword as the ratio of the number of occurrences of that keyword to the total number of keywords in the specific document.

Performance analysis is accomplished for a set of 50 videos. For this process, a set of 5039 keywords is collected manually from the video dataset, which are treated as expected keywords. The CBVR system has identified 25733 keywords from the same 50 videos. These are considered as predicted keywords. Performance analysis is carried out by comparing the expected keywords with the predicted keywords and vice versa.

The performance of the designed CBVR system is analyzed and the details are provided in Table1.

TABLE I: Contingency Table for Performance Analysis of the CBVR system.

		Predicted Keywords	
		P	***N***
Expected Keywords	***P***	4516 (TP)	523 (FN)
	N	20694 (FP)	0 (TN)

In table 1, TP, FP, FN and TN denote True positive, False positive, False negative and True negative results of the analysis.

The accuracy of the designed CBVR system is computed as **89.62%**.

V. CONCLUSION

In the proposed system, the key frames are extracted from videos using canny edge detection process and connected component analysis. The text from each key frame is obtained by optical character recognition using stroke width transformation process. Simultaneously the audio extracted from the video is converted into text transcripts. POS tagging is performed for keyword selection. The term frequency score for each keyword is calculated and mapped with the video name while storing. When a keyword is inputted to the system, the relevant videos are fetched based on the term frequency score of the keyword, if available.

If the input consists of multiple words, then the resultant videos are fetched not based on the semantic meaning of the keyword. In future, this work can be extended to support multi-word input and semantic analysis of the input keywords while fetching videos. This can also be extended to support video clip as input for the system.

REFERENCES

[1] V. Patel, A.V. Deorankar and B. B. Meshram (2010), 'Content Based Video Retrieval using Entropy, Edge Detection, Black and White Color Features', 2nd International Conference on Computer Engineering and Technology, Chengdu, vol. 6 pp. 272 - 276.

[2] Haojin Yang, and Christoph Meinel (2014), 'Content Based Lecture Video Retrieval Using Speech and Video Text Information', IEEE Transactions on Learning Technologies, vol. 7 Issue 2 pp. 142 - 154.

[3] Dipika H Patel (2015), 'Content based Video Retrieval using Enhance Feature Extraction', International Journal of Computer Applications (0975 – 8887), vol. 119 Issue 19.

[4] Fu Chang, Chun-Jen. Chen, and Chi-Jen Lu (2004), 'A linear-time component-labelling algorithm using contour tracing technique', Computer Vision and Image Understanding, vol. 93 Issue 2 pp. 206 - 220.

[5] T. Tuna, J. Subhlok, L. Barker, V. Varghese, O. Johnson, and S. Shah (2012), 'Development and Evaluation of Indexed Captioned Searchable Videos for STEM Coursework', Proc. of the 43rd ACM technical symposium on Computer Science Education, pp. 129 - 134.