ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 4, Issue 1, January 2016

Content based Video Retrieval and Recommendation Using Speech and Text information

Sarang S. Ajnadkar, Prof. Swati Patil

Research Scholar, GHR, Institute of Engineering and Management, Jalgaon, MS, India

Assistant Professor, Dept. of Computer Engineering, GHR, Institute of Engineering and Management, Jalgaon, MS, India

ABSTRACT: Creating video recordings of events such as lectures or meetings is increasingly less expensive and easy. Thus the Video data is increasing in a great deal on World Wide Web (www) and so thus the need of more efficient and correctly functioning method of video indexing, grouping and video retrieval in WWW or Large video archives is necessary. This paper presents a speech and text based video retrieval and Video search system using Optimal Character Recognition (OCR) and Automated Speech Recognition (ASR). First, we convert the video into key-frames and extract the Audio and Text using OCR and ASR. Following step is to produce a summary presenting key points of the video, by making use of text and audio extracted from the Video. This summary will then be used for grouping and Indexing of videos. This in turn will improve the user's aptitude to quickly review this material. This will make user go through only information that they needed. However, the text in the video may vary in dimension, orientation, style, background, contrast and variations in rhythm, volume of and noise in speech and the differentiating between the key-speeches and unnecessary other sounds used during the recording as well, makes data extraction extremely challenging.

KEYWORDS*:* Video Indexing, OCR, ASR, key-frames, data extraction

I. INTRODUCTION

Digital Video has become a largely used to store and exchange data over the last few years, as recording the events, such as Meetings, Lectures is inexpensive and very easy as well as the rapid development in recording technologies makes it widely available. A number of Universities and organizations are recording their seminars and lectures, and making them available over the World Wide Web (www) for students and researchers to access. This results into a continuously increasing Video data over the www, which in turn generates the large video archives. But when user searches for the videos needed, they need to depend on the information added with the videos like, details, genre, subject etc., by producers. This means, even after finding the related video, the user is unconvinced about the information they will get from that particular video. Or sometimes, the user needs to watch those lengthy and boring seminars and lectures, only to get the information of few seconds or less. For example, most of the video retrieval and video search systems, like Bing, YouTube replies the users with the available textual data, such as title, genre, person, and brief description, etc. Often, this data is added by the user which sometimes can be contemptible. This manually given information, most of the times, is incomplete or irrelevant. Therefore user wants some technique, which will give them fair amount of information without viewing those, lengthy and boring videos by using some automatically generated textual data.

First of all, we apply Video segmentation and automatic key-frame detection, so that we can find out the important frames from the video and avoid repetitiveness. Later, we can separate textual data from frame using Optimal Character Recognition (OCR) technology on each frame. And extract Audio, using Automatic Speech Recognition (ASR) technique [1]. From this extracted data, the keywords are generated for the video, which will give the clear idea about video to the users. Textual data is enormously used nowadays, for content-based information

DOI: 10.15680/IJIRCCE.2016. 0401161

retrieval. Extraction of this information involves detection localization, tracking, extraction, enhancement, and recognition of the text from a given image. However, variations of text due to differences in size, style, orientation, and alignment, as well as low image contrast.

The complex background makes the problem of automatic text extraction extremely challenging [2]. But there is a one aspect which separates text from other elements in the frame is its nearly constant Stroke Width. This can be utilized to get the portion of frame which is likely to contain textual data [3]. Similarly, the variations in speech due to the tempo of the spoke person, clarity in his voice, the noise been added to the video becomes problematic when extracting audio data from the video clip. In the Videos, the texts serve as an outline description for Video and are important for indexing of the videos [1]. There is great number of repetitiveness in frames of one shot. These repetitions are reduced by selecting the best frames from the shot. This selected frames work similar to keywords. They are also important for Indexing of Videos. Once the videos are indexed and grouped properly, the retrieval process is fairly easy. The success of this technique highly depends on the other techniques used for Video Segmentation, which will give the best frames from the video. The Optical Character Recognition algorithm will also play a vital role as it will provide us the textual data from the key frames provided by Video Segmentation techniques, which in turn will used as key-words for the video. Same is the case with Automatic Speech Recognition technique, as it will give resulting key- audio signals for the video. The Video Indexing and Retrieval techniques will also play their role in replying the user with the matches' documents with the user queries

II. RELATED WORK

Literature on Retrieval and Indexing is classified into Video segmentation, Retrieval of textual information, Retrieval of speech, and some methods for Retrieving Videos.

Video Segmentation

There is massive number of repetitions in frames from one shot; therefore some best frames are selected as key-frames [5] to compactly represent the shot. The extracted key-frames should contain as much prominent content of the shot as possible and decrease the repetition.

H. J. Jeong[6] proposed a highly accurate method for video segmentation using SIFT and an Adaptive threshold. Using SIFT, we can easily compare two slides, having similar contents but different backgrounds. And we can calculate frame transition quite accurately by using Adaptive Threshold.

Retrieval of Textual Information

OCR was initially developed for high contrast data images, taken from metal and other surfaces with uneven roughness and reflectivity. The basic technique used for this was, that the impressed characters appeared dark and background light, after reflection of light [7]. A vigorous approach to retrieve text from a color image was given by Y. Zhan [8]. The proposed algorithm uses the multistate Wavelet features and the structural information to locate the text lines. Then a Support Vector Machine (SVM) classifier was used to get the exact text from those previously located text lines.

An efficient and computationally fast algorithm to extracting text from documents was developed by S. Audithan [9]. They used a Haar Discrete Wavelet transformation to detect edges of candidate text regions. Non-text edges were removed using some technique. H. Yang [10] has developed a Skeleton-Based binarization method to separate and extract text from complex backgrounds. These can be processed by standard OCR software. J. Einstein[11] proposed a linguistically-motivated approach to select key-frames from video that contain most important gestures. More specifically, he bootstrap from multiple model reference resolution to identify the key gestures. Then the frames are selected, having these key gestures.

Retrieval of Speech

J. Foote[12] proposed a Large-Vocabulary Recognition System (LVRS), which used a "sub-word" approach, instead of developing an explicit Hidden Markov Models (HMMs)[13] for every one of the one thousand words in the vocabulary, a couple of hundred "sub-word" models are used.

Van Thong [14] has given some experiments showing some high speech recognition and retrieval performance even though the audio signals has different acoustic conditions.

The ASR captures an acoustic signal from video as a representative of speech. By using pattern matching, it will determine the words spoken in the video. Speech recognizers typically have a stored acoustic set and patterns of language models in computer database. These patterns are results of training and stored rules of interpreting a language. These models are checked with the captured singles from video. The elements in the computer databases, some techniques are used to determine the best match from the set of matched contents [15].

W. Hurst [16] identified some basic situations that should be considered when recording a lecture for audio extraction, and audio recognition accuracy is influenced by some easy system modifications. He also showed that, the retrieval performance can significantly increase after considering audio signals rather than textual data from frames.

Methods for Retrieving Videos

Keywords generated from Optimal Character Recognition (OCR) and Automatic Speech Recognition (ASR) summarizes the document or Video. These keywords are used for information retrieval from Video archives [1]. J. Fan [4] has proposed a new Framework, called "Class View" for more advanced content-based video retrieval. The important concept they have proposed is, a hierarchical video classification technique to minimize the difference between low level visual features and high level visual concepts. In conventional retrieval, the Euclidean distance between the database and the query is calculated. Short distance indicates that there are more similarities between query frame and database frame. Using this, it is easier to group and retrieve videos [17].

III. Proposed Method

Project flow is given by figure. 1.

Slide Video Segmentation

Video browsing can be achieved by segmenting video into representative key frames. The selected key frames can provide a visual guideline for navigation in the lecture video portal. Moreover, video segmentation and key-frame selection is also often adopted as a pre-processing for other analysis tasks such as video OCR, visual concept detection, etc. Choosing a sufficient segmentation method is based on the definition of "video segment" and usually depends on the genre of the video. In the lecture video domain, the video sequence of an individual lecture topic or subtopic is often considered as a video segment In the first step, the entire slide video is analyzed. We try to capture every knowledge change between adjacent frames, for which we established an analysis interval of three seconds by taking both accuracy and efficiency into account. In the second segmentation step the real slide transitions will be captured. The title and content region of a slide frame is first defined.

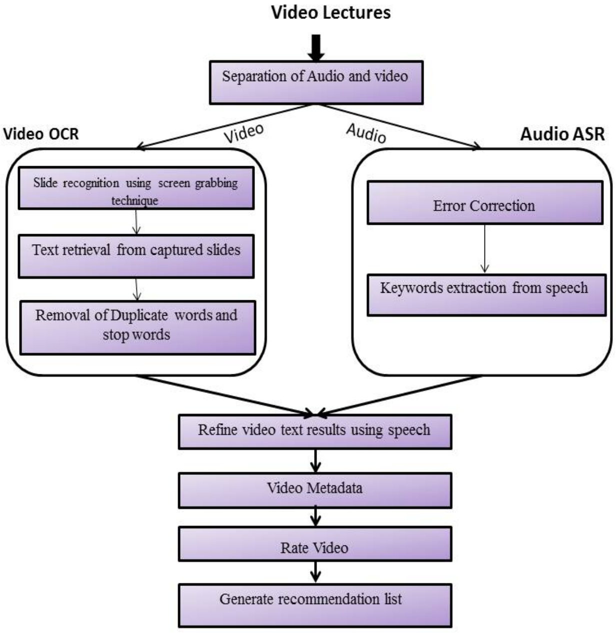

Video Lectures
Separation of Audio and video
Video OCR
Video
Audio
Audio ASR
Slide recognition using screen grabbing technique
Text retrieval from captured slides
Removal of Duplicate words and stop words
Error Correction
Keywords extraction from speech
Refine video text results using speech
Video Metadata
Rate Video
Generate recommendation list

Figure 1Architecture of system

Video OCR for Lecture Videos

Texts in the lecture slides are closely related to the lecture content can thus provide important information for the retrieval task. In this framework, a novel video OCR system for gathering video text is developed. In the detection stage, an edge-based multi-scale text detector is used to quickly localize candidate text regions with a low rejection rate. For the subsequent text area verification, an image entropy-based adaptive refinement algorithm not only serves to reject false positives that expose low edge density, but also further splits the most text and non-text-regions into separate blocks. Then Stroke Width Transform (SWT) based verification procedures are applied to remove the non-text blocks. Since the SWT verifier is not able to correctly identify special non-text patterns such as sphere, window blocks, garden fence, we adopted an additional SVM classifier to sort out these non-text patterns in order to further improve the detection accuracy. For text segmentation and recognition, image skeleton and edge maps are used to identify the text pixels. The proposed method consists of three main steps: text gradient direction analysis, seed pixel selection, and seed-region growing.

Slide Structure Analysis

Generally, in the lecture slide the content of title, subtitle and key point have more significance than the normal slide text, as they summarize each slide. Due to this fact, we classify the type of text lines recognized from slide frames by using geometrical information and stroke width feature. The lecture outline can be extracted using classified text lines, it can provide a fast overview of a lecture video and each outline item with the time stamp can in turn be adopted for video browsing.

Browsing with Lecture Key frames and Extracted Lecture Outline

By clicking on the outline items the video will jump to the corresponding time position. Furthermore, the user can use the standard text search function of a web browser to Fig. 8 shows the visualization of slide key frames in our lecture video portal. The segments are visualized in the form of a timeline within the video player. This feature is intended to give a fast hint for navigation. If the user wants to read the slide content clearly, a slide gallery has been provided underneath the video player. By clicking on the thumbnails or on the timeline-units, the video will navigate to the beginning of the segment.

Keyword Extraction and Video Search

The lecture content-based metadata can be gathered by using OCR and ASR tools. However the recognition results of automatic analysis engines are often error prone and a large amount of irrelevant words are also generated. Therefore keywords are extracted from the raw recognition results. Keywords can summarize a document and are widely used for information retrieval in digital libraries. In this work, only nouns and numbers are considered as keyword candidate. The top n words from them will be regarded as keyword. Segment-levels well as video-level keywords are extracted from different information resources such as OCR and ASR transcripts respectively. For extracting segment level keywords, consider each individual lecture video as a document corpus and each video segment as a single document, whereas for obtaining video-level keywords, all lecture videos in the database are processed, and each video is considered as a single document.

To extract segment-level keywords, arrange each ASR and OCR word to an appropriate video segment according to the time stamp. Then extract nouns from the transcripts by using the Stanford part-of-speech tagger and a stemming algorithm is subsequently utilized to capture nouns with variant forms. To remove the spelling mistakes resulted by the OCR engine, perform a dictionary-based filtering process.

Calculate the weighting factor for each remaining keyword by extending the standard TFIDF score. The ranked video-level keywords are used for the con-tent-based video search. in our lecture video portal. Segmented lecture slides, search hits and keyword weights are additionally provided to the user when hovering on the corresponding timeline-unit. Clicking on the unit the video segment will be played by a pop-up player. The video similarity can further be computed by using a vector space model and the cosine similarity measure.

IV. Results

To evaluate project, RMSE test is performed with existing and proposed recommendation algorithms. Root-mean-square error (RMSE) is a frequently used measure the differences between values predicted by a model and the values actually observed. For finding root mean square error, a set of samples is considered. And the recommendation list is generated for some configuration and using same configuration actual accurate results are found out. And finally RMSE is calculated by using the following formula. Less the value of RMSE, the model or algorithm is more accurate.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}|x_i - \hat{x}_i|}{N}}$$

Where:
$\{x_i\}$ is the actual observations,
$\{\hat{x}_i\}$ is the estimated values,
N is the number of data points.

Figure 2 shows RMSE test results. The test is executed first on pearson algorithm to find out Root Mean Square Error. A sample database of size 100 users with 100 items. Initially algorithm calculated similarity between users. Based upon that similarity algorithm predicted ratings for an items by a user under consideration. With the given size of data and other settings Root Mean Square Error is calculated which is found to be 29.525986.

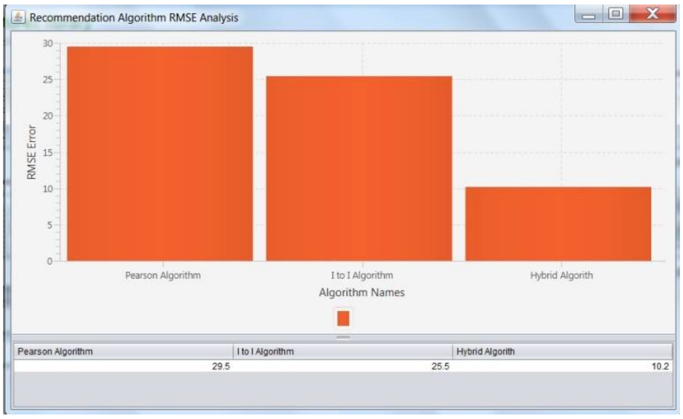
Table 1RMSE test Result

Model	RMSE
Pearson algorithm	29.525986.
I to I algorithm	25.466318
Proposed system	10.202349

Secondly Item based recommendation algorithm is evaluated by keeping other settings same. Item based algorithm works on similarity between items which user had selected with all other items. Based upon the result it predicted ratings for other items by the user under consideration. Using actual results and predicted results RMSE is calculated. With these parameters it was found to be 25.466318.

Finally the same test is executed on system where all the settings are kept same to get the right perspectives. In this algorithm predicted ratings for all items in the database for the user under consideration. The Root Mean Square Error result of that test was 10.202349.

After comparing the system with existing systems it is found that Root Mean Square Error is less than other system thus proving implemented system is much better in performance than other systems. A graph of results is plotted to represent our findings which is shown in figure 2 Recommendation algorithm RMSE Analysis.



Pearson Algorithm	I to I Algorithm	Hybrid Algorith
29.5	25.5	10.2

Figure 2Recommendation algorithm RMSE results

V. CONCLUSION

In this paper, a recommendation system for content based approach to retrieve textual data from the video archives is presented. Regardless the fact that, the textual data may have different size, colour, style and may have a plain or natural background. Similarly, audio keywords may have different tempo, volume, and any sort of noise mixed with it. Using this retrieved data, both textual and audio, are indexed and grouped into large video archives automatically. By this, the end user can search particular content in the video irrespective of video metadata given or not. This will be useful for users as, they won’t need to go through those long and boring videos. But they will get only the information they needed.

Future scope of this work can be optimizing search results to make it faster stable and scalable. Also it can be possible to use cloud to store video lectures and optimize current system in cloud architecture.

REFERENCES

[1] Haojin Yang, Christoph Meinel, “Content Based Lecture Video Retrieval Using Speech and Video Text Information” in IEEE Computer Society , Issue No.02 - April-June (2014 vol.7)

[2] Keechul Jung, KwangIn Kim, Anil K. Jain, “Text Information Extraction in Images and Video: A Survey”, in Pattern Recognition, Vol. 37, No. 5. (May 2004), pp. 977-997

[3] Boris Epshtein, Eyal Ofek, Yonatan Wexler, “Detecting Text in Natural Scenes with Stroke Width Transformation” for Microsoft Corporation.

[4] J. Fan, X. Zhu, J. Xiao, “Content-based Video Indexing and Retrieval”.\

[5] Y. Song, G.-J. Qi, X.-S. Hua, L.-R. Dai, and R.-H.Wang, “Video annotation by active learning and semi- supervised ensemble,” in Proc. IEEE Int. Conf. Multimedia Expo.

[6] Hyun JiJeong, Tak-Eun Kim, Myoung Ho Kim, “An Accurate Lecture Video Segmentation Method by Using SIFT And Adaptive Threshold”, in 10th Int. conf. on Advances in Mobile Computing and Multimedia.

[7] W. Barber, T. Cipolla, J. Mundy, “Optical Character Recognition”, General Electric Company.

[8] Y. Zhan, W. Wang, w. Gao, “A Robust Split-And- Merge Text Segmentation Approach For Images”, International Conference On Pattern Recognition, 06.

[9] S. Audithan, R. M. Chandrasekaran, “Document Extraction From Document Images Using Haar Discrete Wavelet Transform”, European Journal of Scientific Research

[10]H. Yang, B. Quehl, H. Sack, “A Framework for Improved Video Text Detection and Recognition”, for Multimedia tools and application.

[11] J. Einstein, R. Barzilay, R. Davis, “Turning Lectures into Comic Books Using Linguistically Salient Gestures”, Computer Science and AI laboratory, MIT Cambridge.

[12] J. Foote, “An Overview of Audio Information Retrieval”, Institute of Systems Science, Singapore,1999.

[13] L. Rabiner, “An Introduction to Hidden Markov Model”, AT&T Bell Laboratories.

[14] Van Thong, J.-M., Moreno, P.J., Logan, B. Fidler, B., Maffey, K., Moores, M., “Speechbot: An experimental speech-based search engine for multimedia content on the web”, In IEEE Transactions on Multimedia.

[15] Benjamin Chigier, “Automatic Speech Recognition”, for Purespeech Inc.

[16] Wolfgang Hurst, “A Qualitative Study Towards Using Large Vocabulary Automatic Speech Recognition to Index Recorded Presentations for Search and Access over the Web”, University of Freiburg, Germany.

[17] B. V. Patel, B. B. Meshram, “Content Based Video Retrieval Systems”.