International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

A Survey on Content based Video Retrieval Using Speech and Text information

Laxmikant S. Kate[1], M. M. Waghmare[2]

[1]ME Student, Department of Information Technology, Dattakala Group of Institute of Technology, University of Pune, Pune, Maharashtra, India

[2]Assistant Professor, Department of Information Technology, Dattakala Group of Institute of Technology, University of Pune, Pune, Maharashtra, 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

1. 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 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 and complex background make 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 utilised 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 is 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.

The important issue that researchers have to concentrate on is the generation of key words for the video, both textual and audio, as we do for the textual documents. The main issue to be researched in near future is, Video Database Indexing, because it is not feasible to extend traditional database indexing to suit videos[4]. The remaining paper can be sorted out as: Section 2 gives Video segmentation and the techniques that could be used for the same purpose. Later in the section, we have reviewed some text retrieval techniques, as well as audio retrieval techniques, like OCR and ASR. Finally, we have quickly examined some strategies for indexing and retrieval of video from large video archives. In section 3, is briefly reviewed conclusion.

2. Literature Review

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

2.1 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.

2.2 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 colour image was given by Y. Zhan[8]. The proposed algorithm uses the multiscale 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.

2.3 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 increased after considering audio signals rather than textual data from frames.

2.4 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].

3. Conclusion

In this report, we represented a content based approach to retrieve textual data as well as audio data, automatically from the videos over the www. 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, we can index and group large video archives automatically. By this, producers of video are no more in need to provide video related information manually, which in turn will be time consuming job. 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. Still, this paper does not claim that, we have solved all the issues related to content-based video indexing and retrieval, using text and audio. But we certainly have added some advanced steps towards our destination.

References

[1] Haojin Yang, Christoph Meinel, "Content Based Lecture Video Retrieval Using Speech and Video Text Information" for IEEE.
[2] Keechul Jung, KwangIn Kim, Anil K. Jain, "Text Information Extraction in Images and Video: A Survey".
[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".