Video Classification using Embedded Audio

Tejas Wadiwala
Student
Atharva College of Engineering
Mumbai University
Malad, Mumbai, India

Shweta Sharma
Assistant Professor
Atharva College of Engineering
Mumbai University
Malad, Mumbai, India

Suraj Yadav
Student
Atharva College of Engineering
Mumbai University
Malad, Mumbai, India

Anand Yadav
Student
Atharva College of Engineering
Mumbai University
Malad, Mumbai, India

Anish Sridhar
Student
Atharva College of Engineering
Mumbai University
Malad, Mumbai, India

Abstract— **In recent times people have excess to a stupendous amount of videos through digital television and largely through internet. With this there also arises the dubiety of what to view and what not to view. This has given rise to the need for classifying videos. Many of the video sharing websites now support the system of "tagging" the videos. These tags are given by the user when he uploads the video on the website or added by website manager later some time. These tags thus help in classification of the videos into various genre or categories such as sports, finance, movies, commercials, news etc. but this system of manual classification is not efficient as it is practically impossible to tag all the videos available on the internet. Our project provides automatic classification of videos.**

Keywords: Indexing, Searching, Classification, Automatic, Text, Video, Conversion

I. INTRODUCTION

Today tagging, though widely used, has its own drawbacks. Today there is need for videos to be searched based on its content. A system is needed which would search the videos for the occurrence of a given keyword and gives the audio-visual content, even specifying the correct position where the keyword was found in the videos. This can be achieved if we are aware of the content of the videos. The knowledge about the content of the videos comes from the metadata of the content. The metadata can be stored along with the videos as annotations. And as manual classification of videos does not give quality results, so will manual annotation of videos. This has given rise to the need of automatic and unsupervised classification and annotation of videos.

Whenever a video is given to the system, the video is first analyzed to extract its audio content, which is done in the audio extraction module. The extracted audio is worked upon to convert the speech in it into text. The text thus obtained s mined for words which can be probable keywords, when search is performed on the videos. The keywords thus found are stored in the database, mapping it with the playtime of the video where the keyword was spoken. This is video indexing, thus classifying videos and creating annotated videos.

II. LITERATURE REVIEW

Videos can be classified on the basis of the features drawn from the modalities-text, audio and visual. The text modality deals with the detection of the presence of an onscreen text i.e. indexing and searching is done on the basis of words found onscreen, like the sub titles or the embedded text in signage. The audio content can also be used for video classification i.e. indexing and searching is done on the basis of the word utterances of the subjects in the video. The visual modality deals with pattern matching and image mining performed on the frames extracted from the video. A combination of two or more modalities can also be used, as shown in Figure 1. Many principles of cinema can also be incorporated for video classification. For example, videos with horror are dimly lit while comedy videos are brightly lit. Videos of action movies or sports have more motion than videos of drama.[1]

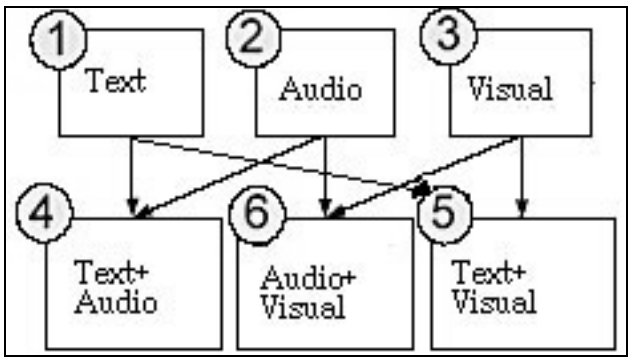

Figure 1: Modalities of video classification

Video classification is usually accompanied with video annotation which helps in retrieving the video archives. Video annotation is basically video metadata creation. Metadata stores content descriptions of the videos hence video annotation can be defined as process of creating metadata for video objects. Video annotation can be either manual annotation or automatic annotation.

Manual annotation uses a video annotation tool which takes descriptions from the user and stores the text as metadata.[3] Human errors are introduced in such an annotation method.
Automatic annotation can be done on the same three modalities on which video can be classified namely text on the screen, audio in the audio stream of the video and visual aspect of the video.
In our project we use the audio modality for video classification and annotation. Thus, videos can be classified based on the utterances of the words in the video. So, the content based retrieval of the videos would be based on its audio content rather than the visual content. Since most of the videos contain speech in some or the other form, we convert audio to text (A2T). The output text is mined to get the metadata for the video object.[4]

Given below is an in-depth study of the various APIs which could be used for our project for implementation purposes of the various modules. The various modules for which the study of APIs is carried out are as listed below:

Audio Extraction
Speech Recognition
Keyword Extraction
Indexing

III. PROBLEM STATEMENT

Within the last several years, object motion trajectory-based recognition has gained significant interest in diverse application areas including sign language gesture recognition, Car Navigation System (CNS), Global Positioning System (GPS), animal mobility experiments, sports video trajectory analysis and automatic video surveillance. Psychological studies show that human beings can easily discriminate and recognize an object's motion pattern even from large viewing distances or poor visibility conditions where other features of the object vanish. The development of accurate activity classification and recognition algorithms in multiple view situations is still an extremely challenging task. Object trajectories captured from different view-points lead to completely different representations, which can be modeled by affine transformation approximately. To get a view independent representation, the trajectory data is represented in an affine invariant feature space. With regard to that, a compact, robust view invariant representation due to camera motions is highly desirable.

IV. PROPOSED SYSTEM

We show a system that performs unsupervised, automated classification and annotation of videos based on its audio content which is embedded. The classified and annotated videos can be later searched and retrieved, when a user inputs a keyword. The basic architecture of the system is as shown in the figure 2.

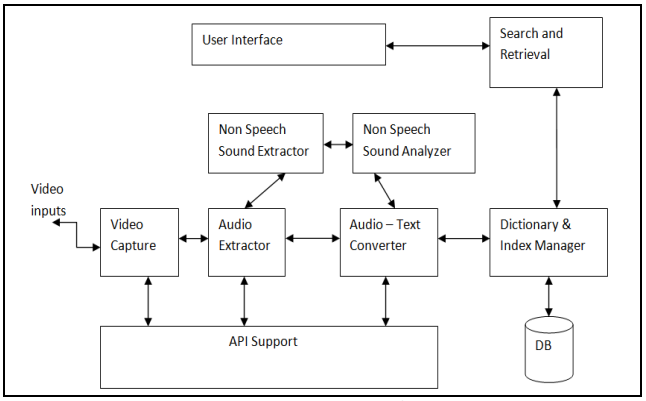

Figure 2 Basic Architecture of the system

The proposed system to be implemented is described briefly and can be depicted as in the above diagram. It illustrates a sketch of the outline architecture. This diagram serves the general purpose to explain the broad structure and working of the proposed system. The system consists of the following modules integrated within the framework.

Video Capture: It takes a series of video data frames or streams video content which acts as the basic input. For future retrievals a link to the source is maintained in the database.

Audio Extractor: This stage extracts the audio stream out of the composite audio-video stream by cleaning up any unwanted noise from the input.

Non-Speech Sound Extractor: Out of the composite audio-video stream this stage extracts the non-speech sounds.

Non-Speech Sound Analyzer: This stage analyzes non speech sound which has been extracted in the previous stage.

Audio to Text Conversion: Primary responsibility of this stage is to convert the audio stream into textual form. All common words mentioned in an exclusion list are removed from the text to realize the final set of interesting keywords.

Dictionary and Index Manager: The words which are encountered are maintained in a dictionary. We index these keywords to ensure efficient retrieval of the information which we want.

Search Engine: It is mainly responsible to search the keyword indexes for match as per user requirements.

Command Interface and UI manager: This module serves as an interface to the user. This module basically is used to accept query from user and to display results of the same.

V. CONCLUSION

In this paper we discussed our proposed method of detecting and extracting text from the video file. The system automates the manual process of extracting text from videos and hence is economical in terms of time and human efforts.

VI. ABBREVATION

SDK - Software Development Kit
API - Application Programmable Interface
JSGF - Java Speech Grammar Format
H/W – Hardware
S/W – Software
SAPI – Speech Application Programming Interface

VII. REFERENCES

[1] F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and TRECVid," in Proc. 8th ACM Int. Workshop on Multimedia InformationRetrieval (MIR) 2006. New York: ACM Press, 2006, pp. 321–330.
[2] Hauptmann, R. Yan, Y. Qi, R. Jin, M. Christel, M. Derthick, M.-Y. Chen, R. Baron, W.-H. Lin, and T. D. Ng, "Video classification and retrieval with the informedia digital video library system," presented at the Text Retrieval Conf. (TREC 2002), Gaithersburg, MD.
[3] Audio Denoising by Time-Frequency Block Thresholding Guoshen Yu*, St´ephaneMallat and Emmanuel Bacry Submitted to IEEE Transactions on signal processing.
[4] Hindawi Publishing Corporation EURASIP Journal on Advances in Signal Processing Volume 2007, Article ID 96384, 9 pages doi:10.1155/2007/96384
[5] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval,Addison Wesley, 1999.