An Automated Video Classification and Annotation Using Embedded Audio for Content Based Retrieval

Anil Kale
PVPP College of Engineering, Sion, Mumbai-22, India
anil5474@gmail.com

D. G. Wakde
P. R. Patil College of Engineering, Amravati, India
director_PRPCE@rediffmail.com

***Abstract*—Efficient and effective video classification and annotation demands automated unsupervised classification and annotation of videos based on its embedded video content as manual indexing is unfeasible. Audio is a rich source of information in the digital videos that can provide useful descriptor for indexing the video databases. Audio archives contrast with image or video archives in a number of important dimensions. First, they capture information from all directions and are largely robust to sensor position and orientation, allowing data collection without encumbering the user. Second, the nature of audio is distinct from video, making certain kinds of information (for example, what is said) more accessible, and other information (for example, the presence of nonspeaking individuals) unavailable. In general, processing the content of an audio archive could provide a wide range of useful information. As a first step the audio content of video is extracted and cleaned for further processing the next step converts audio into textual format .The text is processed upon to get the prime keywords in the video using text mining. The videos are classified and annotated on the keywords thus found. The annotated videos**

***Index Terms*—fContent-based access, video, multimedia, keyframes, automatic analysis component.**

I. INTRODUCTION

The advances in the digital and network technology have produced a flood of multimedia information. The people can easily access digital videos which is one of the major constituent of multimedia information.

The growing amount of digital video is driving the need for more effective methods for indexing, searching, and retrieving of videos based on its content. While recent advances in content analysis, feature extraction, and classification are improving capabilities for effectively searching and filtering digital video content, the process to reliably and efficiently index multimedia data is still a challenging issue.

Although this scenario provides ample choices and flexibility but at the same time it involves the serious issue of what to view and what not to view, obviously one of the application scenarios is a personalized digital video classification that is capable of segregation of videos to suit the interest of a particular user and filtering out others. To achieve this objective, automated video classification and annotation is essential.

Most of the video sharing web sites now support the system of "tagging" the videos. This tagging is manually provided by the user at the time of uploading or later on provided by web site manager which helps to classify the videos in different categories like sport, news, movies, commercial etc.

The manual tagging is unfeasible as it is practically impossible to tag all the videos available on the internet. There is also a growing need for videos to be search based on its content. A system is needed which would search the video for the occurrence of the given keywords and gives audio- visual content while specifying the correct position where the keyword was found in videos. This is possible with the help of metadata which can be stored along with the videos as annotation. As manual classification of videos does not provide quality result, so will manual annotation of videos. This warrants the urgent need of automatic and unsupervised classification and annotation of videos.

A. Aim

A video classification and annotation system based on the analysis of embedded audio content is to be implemented. Every video clip submitted should be analyzed and appropriately classified for any future search and retrieval operation based on the audio content buried in the video.

B. Objective

The primary objective is to create an automated system for video classification and annotation by getting rid of manual classification and annotation as it is practically highly unfeasible. To achieve this system must meet the following goals.

Manuscript received July 24, 2013; revised September 10, 2013.

doi: 10.12720/jiii.1.4.235-238

- To implement an unsupervised, automated system for video classification and annotation based on the audio content embedded within the video stream information.
- To perform video search and retrieve videos based on the utterances of the keyword within the video content.
- To provide a simple key word based search on the videos.
- To provide a near video match search on the video repository.

C. Scope and Limitations

Our proposed system provides a complete automation for video classification and annotation. The major limitation in this system is that the only modality i.e. embedded audio is focused where as other two modalities i.e. visual and texts are not emphasized.

II. Literature Survey

Today's advanced digital media technology has led to the explosive growth of multimedia data in scale that has never occurred before. The availability of such large-scale quantities of multimedia documents prompts the need for efficient algorithms to search and index multimedia files. Modern multimedia content is often characterized by having multiple varied forms, i.e. movies consisting of video-audio streams with text captions, web pages containing pictures, text, and songs. This heterogeneous multi-modal nature gives rise to challenging new research questions of how to best represent, classify, and effectively retrieve multimedia data. The tremendous potential of such aforementioned research in a wide array of applications has drawn considerable attention to the emerging field of multimedia information retrieval in recent years. Most previous work on multi-modal data retrieval and classification, e.g. [1], [2], assumes simplistically that different modalities of data are independent. Retrieval/classification is thus performed on each modality separately and the results are subsequently combined. Often, knowledge about one modality conveys a great deal of information about the others. Making use of such relations is expected to improve the performance on the retrieval and classification task. Furthermore, when one data type is missing, the correlation between data of different types allows for inference of features of the missing types from the observed types.

A large amount of multimedia content consists of digital videos, giving rise to an unprecedented high volume of data. The provision of an interactive access to this huge digital video information repository currently occupies researcher's minds in several fields. The visual information is traditionally used for video indexing. Here, we consider using embedded audio because it's a rich source of content-based information. Users of digital video are often interested in certain action sequences. While visual information may not yield useful "action" indexes, the audio information often directly reflects what is happening in the scenes and distinguishes the actions. Although image-based approaches are common, a few studies have also considered audio analysis. However, it remains an area of basic research. Since we must deal with mixed sound sources, existing speech recognition algorithms generally don't suit ordinary videos. Most studies on music and speech detection are aimed at improving speech recognition systems. The difficulty in handling mixed audio sources has, until now, hindered the use of audio information in handling video. Therefore, few have attempted to deal with the type of videos we come across in everyday situations [3].

Research in the area of developing automated classification has been going on for some time. Text retrieval conference (TREC) is dealing with several issues associated with this [4]. One of the many issues TRECVid deals with is classification of videos such as distinguishing videos based on indoor and outdoor location, identifying on screen face or text, identifying videos with speech or musical notes [5] [6].

Videos can be classified on the basis of features drawn from the modalities-text, audio and visual. The text modality deals with the detection of the presence of an on screen text i.e. the indexing and searching is done on the basis of words found on screen. The visual modality concerns with pattern matching and image mining performed on the frames extracted from the video. The embedded audio in the video can also be used for video classification in which the indexing and searching is done on the basis of the word utterances in the audio of the video [7] [8].

Video classification is usually accompanied with video annotation which helps in retrieving the video archives. Video annotation is about video metadata creation which can be manual or automatic. Automatic annotation can be done on same three modalities on which video can be classified i. e. text, audio and visual.[9]

The fundamental obstacle in automatic annotation is the semantic gap between the digital data and their semantic interpretation [10]. Progress is currently being made in known object retrieval [11] [12], while promising results are reported in object category discrimination [13], all based on the invariance paradigm of computer vision. Significant solutions to access the content and knowledge contained in audio/video documents are offered by StreamSage and Info media. While the field of content-based retrieval is very active by itself, much is to be achieved by combination of multiple modalities: data from multiple sources and media (video, images, text) can be connected in meaningful ways to give us deeper insights into the nature of objects and processes. With the rapid growth of multimedia application technology and network technology, processing and distribution of digital videos become much easier and faster. However, in searching through such large-scale video databases, indexation based on low-level features like color and texture, often fails to meet the user's need which is expressed through semantic concepts due to the "semantic gap"[14]. Consequently, how to establish the mapping between the low-level features and high-level semantic descriptions of

video content to bridge up the "semantic gap" efficiently, i.e., automatic annotation of video at the semantic level, is currently becoming an important topic in the multimedia research community. The overview of past literature survey can be given as, Ioannis Paraskevas and Edward Chilton [15] discussed a novel method for the automatic recognition of acoustic utterances is presented using acoustic images as the basis for the feature extraction. This method effectively employs the spectrogram, the Wigner- Ville distribution and eo-occurrence matrices. The high-level feature detection task has realized an important test bed for concept detection architectures that have proven to be an important performance enhancing component in video search systems [16, 17]. Alexander G. Hauptmann, Rong Jin, and Tobun D. Ng [18] provided an evaluation on the effects of different types of information used for video retrieval from a video collection. Marijn Huijbregts [19] reported on the setup and evaluation of robust speech recognition system parts. Ying Li and Chitra Dorai[20] used Support Vector Machine(SVM) for classifying audio for instructional video analysis. Roberto Vezzani [21] presented a general purpose system for a hierarchical structural segmentation and automatic annotation of video clips by means of standardized low level features. Block thresholding estimation procedure which adjusts all parameters adaptively to signal property by minimizing a Stein estimation of the risk was introduced by Guoshen Yu [22]. Padmapriya Srinivasan [23] was introduced FFMPEG tool for evaluation of audio and video quality for distance education. Milind R. Naphade and Thomas S. Huang [24] reviewed state-of-the-art multimedia understanding systems with particular emphasis on a system for semantic video indexing centered on multijects and multinets.

III. METHODOLOGY

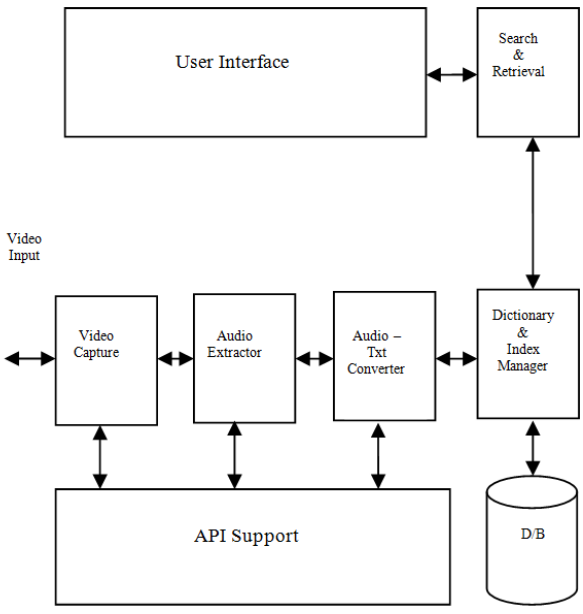

Figure.1. System Architecture

The system architecture to develop "Content Based Retrieval for video classification using embedded audio" is shown in Fig. 1. The system consists of the following modules integrated within the framework. *Video Capture* module captures a series of video data frames or streams video content which acts as the basic input. The *Audio Extractor* is responsible for cleaning up any unwanted noise from the input and extracts the audio stream out of the composite audio-video stream. *Audio to Text Conversion* phase is responsible to convert the audio stream into textual form. *Dictionary and Index Manager* maintains a dictionary of encountered keywords. *Search Engine* is to search the keyword indexes for match as per user requirements.

IV. CONCLUSION

This paper presents a model that provides automation of video classification and video annotation. How to satisfy the general user in searching videos of their interest and needs is an important and great challenge. This paper explores the audio modality in retrieving the video.

Future scope: The current research intended to improve the classification of videos by including automatically generated information such as automatic captioning stored in database.

REFERENCES

[1] T. Westerveld, T. Ianeva, L. Boldareva, A. de Vries, and D. Hiemstra, "Combining information sources for video retrieval," presented at TRECVID 2003 Workshop, 2004.
[2] P. Fraternali, M. Brambilla, and A. Bozzon, "Model-Driven design of audiovisual indexing processes for search-based applications," presented at Seventh IEEE International Workshop on Content-Based Multimedia, 2009.
[3] T. Amin, M. Zeytinoglu, and Ling, " Interactive video retrieval using embedded audio content," presented at IEEE ICASSP, 2004.
[4] F.Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and TRECVid," in *Proc. 8th ACM Int. Workshop on Multimedia Information Retrieval*, New York: ACM Press, 2006, pp. 321-330.
[5] A. 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.
[6] Ar. Amir and Ja. O. Argillander, *IBM Research TRECVID-2004 Video Retrieval System*.
[7] Cees G. M. Snoek and M. Worring, "Multimodal Video Indexing: A Review of the State-of-the-art," *Multimedia Tools and Applications*, vol. 25, pp. 5–35, 2005, Springer Science.
[8] Y. Li, S. Narayanan, and C.-C. Jay Kuo, "Content-Based movie analysis and indexing based on audiovisual cues," *IEEE Transactions on Circuits and systems for Video Technology*, vol. 14, no. 8, August 2004.
[9] G. IyengarP, H. J. Nock, and C. Neti. "Discriminative model fusion for semantic concept detection and annotation in videos" *MM'03*, Berkeley, California, USA, November 2–8, 2003.
[10] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content based image retrieval at the end of the early years," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, 2000.
[11] T. Gevers, A. Smeulders, "Color based object recognition," *Pattern Recognition*, vol. 32, pp. 453-464, 1999.
[12] K. Mikolajczyk, C. Schmid, "An affine invariant interest point detector," *Int. J. Comp. Vis*, vol. 60, pp. 63-86, 2004.

[13] R. Fergus, P. Perona, and A. Zissermann, "Object class recognition by unsupervised scale invariant learning," presented at IEEE Conf. on Computer Vision Pattern Recognition, 2003.
[14] A. Hauptmann, Yan, *et al*., "Filling the semantic gap in video retrieval: An exploration," *Semantic Multimedia and Ontologies*, pp. 253-278, 2008.
[15] I. Paraskevas and E. Chilton, "Audio classification using acoustic images for retrieval from multimedia database EC-VIP-MC," presented at 4th EURASIP Conference focused on Video Image Processing and Multimedia Communications. 2-5 July 2003, Zagreb, Croatia
[16] A. F. Smeaton, P. Over and W. Kraaij, "High-level feature detection from video in TRECVid: A 5-year retrospective of achievements," in *Multimedia Content Analysis, Signals and Communication Technology*, A. Divakaran Ed., pp. 151–174, 2009.
[17] Ar. Amir, M. Berg, J. R. Smith, B. Tseng, Y. Wu, and D. Q. Zhang, *IBM Research TRECVID-2003 Video Retrieval System*.
[18] A. G. Hauptmann, R. Jin, and T. D. Ng, "Video retrieval using speech and image information," presented at Electronic Imageing Conforance, Storage Retrieval for Multimedia Database, Santa Clara, CA, January 20-24,2003.
[19] M. Huijbregts, R. Ordelman, and F. de Jong, "Speech-based annotation of heterogeneous multimedia content using automatic speech recognition," *CTIT-technical Report*, version 1.0, May 2007
[20] Y. Li and C. Dorai "SVM-based audio classification for instructional video analysis," presented at IEEE ICASSP 2004.
[21] R. Vezzani, C. Grana, D. Bulgarelli, and R. Cucchiara "A semi-automatic video annotation tool with MPEG-7 content collections," presented at the Eighth IEEE International Symposium of Multimedia (ISM'06).
[22] G. S. Yu, S. Mallat and E. Bacry, "Audio denoising by time-frequency block thresholding," *IEEE Transactions on Signal Processing*, vol. 56, no. 5, May 2008
[23] P. Srinivasan, I. V. McLoughlin, and W. S. Lin "Joint audio video quality evaluation for distance or online education systems," presented at the IEEE 9th Malaysia International Conference on Communications, 15 -17 December 2009, Kuala Lumpur Malaysia.
[24] M. R. Naphade and T. S. Huang, " Extracting semantics from audiovisual content: The final frontier in multimedia retrieval," *IEEE Transactions on Neural Networks*, vol. 13, no. 4, July 2002.

Anil W. Kale was born in 1974 at Maharashtra state of India. He completed his graduation in Computer Engineering from Amravati University and post-graduation in Computer Engineering from Mumbai University. Presently he is working as an Assistant Professor at IT Engineering Department of PVPP's College of Engineering, Mumbai and Pursuing Ph.D. in Electronics Engineering from Amravati University. His research interests are Multimedia databases, Digital Signal Processing and Image processing.

G. Wakde was born in 1955 at Maharashtra state of India. He completed his Ph.D. in 1981 from Nagpur University, Maharashtra, India. Presently he is working as a Director of PRPCM, Amrawati. He is Chairman of Board of Studies in General Engineering, Applied Science & Humanities under the faculty of Engg. & Tech. at Amrawati University. He is having 27 years of Teaching and 11 years of Research experience. His research interests are Signal and Image processing, Wirless Sensor Networks, Network Traffic analysis.