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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

Multimedia information retrieval for lecture video using Automatic Speech Recognition and Optical Character Recognition

Surabhi Pagar, Gorakshanath Gagare

M.E. Student, Department of Computer Engineering, SVIT, Nasik, India

Assistant Professor, Department of Computer Engineering, SVIT, Nasik, India

ABSTRACT: Now-a-days increased development in multimedia technology allows the capturing and storing of video data with highly expensive computers. Users are not satisfied with video retrieval systems available because of analogue VCR functionality provided. This research represents an approach for automated audio ,video content retrieving over huge lecture video repositories. It apply/uses automated video segmentation and key-frame detection method .Subsequently, It extract/takes out textual meta-data by applying video Optical Character Recognition (OCR) technology algorithm on lecture video key-frames and Automatic Speech Recognition (ASR) on lecture audio tracks content of the video to retrieve audio from video and then convert that audio track into text details. Main purpose is to provide multimedia data such as image and video for extracted words. It provides relevant information i.e. text, audio, video to the user for more effectiveness.

KEYWORDS: Multimedia search, video segmentation, video retrieval OCR, ASR.

I. INTRODUCTION

Digital videos are becoming a very popular storage and exchange way of medium as there is fast development in recording technologies and so much highly improved video compression techniques and high-speed networks in last some years. That is why the visual-audio recordings are being used more and more frequently in e-lecturing system. There are many universities and research institutions which are taking advantage of this useful opportunity to record their lectures and then publish them online for students to access independent of time and location to provide mobility. So that they can study anywhere and anytime . Because of this reason there has been a huge increase in the amount of multimedia data on the Internet. Due to this for a user it is almost impossible to find desired videos without having a search function in video repositories. Even when the user has found related video data, it is still difficult most of the time to compare whether a video is useful or not just by only looking at the title and other global meta-data provided which are often brief and a high level document. Moreover, the requested information may be covered in only a few minutes, the user generally want to get only the piece of information he/she requires without viewing at the complete video. The problem then arises like how to retrieve and get the appropriate information in a large lecture video archive more efficiently and easily.Most of the video retrieval and video search systems such as You- Tube, Bing and Vimeo reply based on available textual meta-data such as title, genre, person, and detail description. Generally, this kind of meta-data has to be created by a human manually to ensure a high quality, but the creation step is rather time consuming as well as cost consuming. So the proposed system is an Multimedia based information retrieval approach for lecture video indexing based on automated video segmentation, OCR analysis and ASR technique. System uses optical character recognition to extract text from segmented images and key frames and automatic speech recognition to extract audio from segmented video so that we can extract audio tracks of video and then convert that audio tracks to text data again which we can again used as description for user. The system automatically determines keywords in video and search multimedia information should be added for textual answer by collecting data from web to enrich the collected information[1].

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

II. RELATED WORK

In this section, an overview of existing multimedia retrieval techniques for e-lecture videos are provided. The objective of this survey is to understand the limitations/disadvantages of existing schemes. Haojin Yang and Christoph Meinel ,mentioned in their paper that in the last years e-lecturing has become very popular. The amount of lecture video data on the Internet is increasing rapidly. So ,a more efficient method for video retrieval on internet or within large lecture video repositories is essential. This paper represents an approach for automated video indexing and video search in large lecture video repositories. First of all, they apply automatic video segmentation(ASR)algorithm and key frame detection technique to offer a visual guideline for the video content navigation. Subsequently, they extract textual metadata by applying video Optical Character Recognition(OCR)technology on key-frames and Automatic Speech Recognition (ASR) on lecture audio tracks. The OCR and ASR transcript as well as detected slide text line types are adopted for keyword extraction, by which both video and segment level keywords are extracted for content-based video browsing and search. The performance and the effectiveness of proposed indexing functionalities is proven by evaluation[1].E.Leeuwis, M.Federico and M. Cettolo, presented a new system for automatic transcription of lectures. This system combines number of novel features, including deep neural network acoustic models using multi-level adaptive networks to incorporate out-of-domain information, and factored recurrent neural network language models. It demonstrate that the system achieves large improvements on the TED lecture transcription task from the 2012 IWSLT evaluation. Results are currently showing an relative Word Error Rate reduction of more than 16 percent compared to the closest competing system from the evaluation[2].D. Lee and G. G. Lee, introduced a Korean spoken document retrieval system for lecture search .It automatically build a general inverted index table from spoken document transcriptions and extract additional information from textbooks or slide notes related to the lecture .It integrate these two sources for a search process. The speech corpus used in that system is from a high school mathematics lecture videos. Experimental results showed that the contents information is slightly beneficial for the lecture spoken document retrieval[3].W.Heurst ,T .Kreuzer and M.Wiesen hutter,done research by Recording lectures and putting them on the WWW for to access by students has become a general trend at various universities. To take advantage of the knowledge data base that is built by these documents elaborate search functionality has to be provided that should goes beyond search on meta-data level but performs a detailed analysis of the corresponding multimedia documents. In this paper, presented some experiments towards setting up a Web based search engine for audio recordings of presentations. Authors evaluate standard, state-of-the-art speech recognition software as well as achievable retrieval performance.In addition, They also compare the speech retrieval results with traditional ,text-based approach for searching to evaluate the value of speech processing for lecture retrieval[4].A. Haubold and J. R. Kender ,presented that there is also investigation of methods of segmenting, visualizing, and indexing presentation videos by separately considering audio and visual data.The audio track is segmented by speaker and augmented with key phrases which are extracted using an Automatic Speech Recognizer. The video clip is segmented by visual dissimilarities and augmented by representative key frames. An interactive user interface combines a visual representation of audio, video, text, and key frames, and allows the user to navigate a presentation video. Authors also explore clustering and labeling of speaker data and present preliminary results[5].James Glass, Timothy J. Hazen, Lee Hetherington and Chao Wang, presented in their paper are port on recent efforts to collect a corpus of spoken lecture material that will enable research directed towards fast, accurate and easy access to lecture content .Thus far, collected a corpus of 270 hours of speech from a variety of undergraduate courses and seminars. Authors report on an initial analysis of the spontaneous speech phenomena presenting these data and the vocabulary usage patterns across three courses. Finally, examine language model perplexities trained from written and spoken materials, and describe an initial recognition experiment on one course[6].G. Salton and C. Buckley describes the Term weighting approaches which we can use in automatic retrieval of text data[7].

III. PROPOSED ALGORITHM

A. *ARCHITECTURE:*

Following gives the description of blocks used in system architecture.

- **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 analyst is tasks such as

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

video OCR, visual concept detection, etc.In the next step, the entire slide video is analysed.It will try to capture every knowledge change between adjacent frames,for which it established an analysis interval of three seconds by taking both accuracy and efficiency into account.It means that segments with duration less than three seconds maybe discarded in this system. Since there are every few topic segments shorter than three seconds,this setting is therefore not critical. In next step each slides stores as image in the backend database for OCR algorithm processing.Where we can extract text from each and every image which is considered as key segment.

- **OCR Algorithm:** Text content given in thelecture slides are closely related to the lecture content ,which can thus provide important information for the retrieval task of information .In the proposed framework, we developed a novel video OCR system for gathering video text. This Propose system uses OCR algorithm to extract text from images and store extracted text in database which we can use further.
- **Information Pre-Processing:** From extracted text irrelevant keywords are remove by comparing predefine keyword list and stop words can also remove by using Stemming algorithm .These extracted keywords are used for multimedia query search process for matching.

_ Multimedia Query Search

_ Query Analysis

Query analysis helps to find the informative keyword for searching corresponding media data using multimedia search engines. The main objective of this process is to find the stem word which is considered as the informative keyword. System using an algorithm called stemming algorithm which removes stop-words make data efficient to use. Stemming algorithm is generally used to remove the stopwords which can be applied as follows:The first step is to consider the given query and initialize the empty variable of string data type.Split the query based on the space between them and pass them into array list of string type. Initialize for loop and remove the stop-words i.e., a, and, an, in, be, for and soon by passing the words in the array list. continue the process until length of the array list.Pass the remaining words into empty string variable initialized in step 1.Finally use the obtained words as informative keyword for search and vertically collect the media data.

- **Recommendation and Indexing:** The next component of our model is Recommendation and Indexing. In this module the given question is judged whether it requires any media data or it requires only textual answer. Here system will categorize mainly into four types such as text, text and image, text and video, text and image and video based on the given query. For those question which require media data such images and videos.

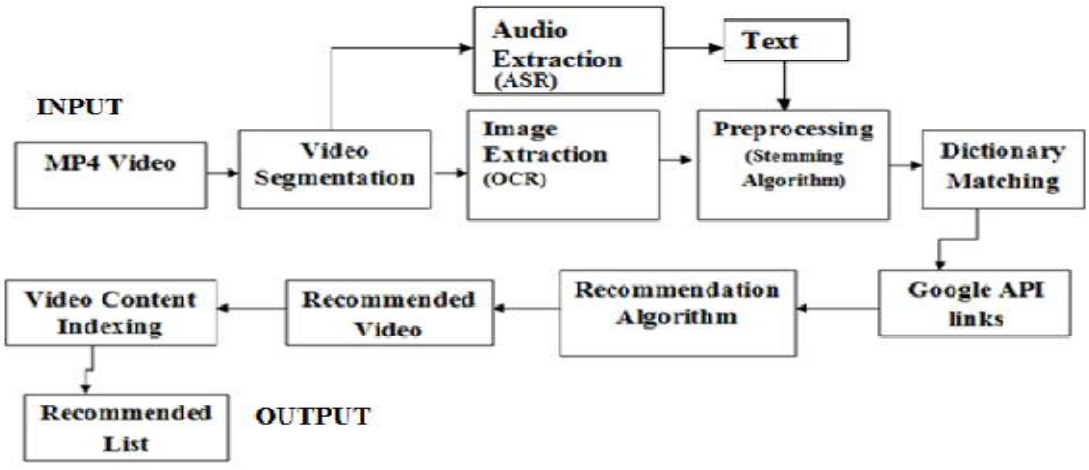

Figure 1. Overall Architecture of System

For this system will vertically collect media data by using multimedia search engines such as YouTube for videos and Google images API(Application programming interface).The overall Block diagram of the system is shown in the Figure 1.

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

IV. EXPERIMENTAL SETUP

This section gives the results of Video To Audio Details.

Experimental Setup: For the experiment, we take computer network(CN) videos as a input initially.We are considering this computer network videos for dataset.This dataset are freely available.All experimentation is performed using Pentium Dual Core processor and 1 GB RAM. The operating system is windows XP.We are using Java for programming.This will give the results of multimedia retrieval as follows.

TABLE 1
Video to Audio

File Name	Size	Extraction Time	Audio Size
CN-1	3.23mb	4278ms	1.24mb
CN-2	11.2mb	1577ms	4.37mb
CN-3	21.9mb	37740ms	11.7mb
CN-4	3.10mb	8720ms	1.58mb
CN-5	5.89mb	9721ms	3.15mb
CN-6	18.1mb	37128ms	13.1mb

Table 1 presents video files along with their size in MB and then give us the extracted audio files along with size in MB. It is done with the help of OCR algorithm.

TABLE 2
Audio to Text

Audio File Size	Audio to Text Time	Total Word Extracted
1.24mb	10ms	5
4.37mb	1024ms	64
11.7mb	4052ms	2857
1.58mb	4804ms	50
3.15mb	4502ms	810
13.1mb	5458ms	4299

Table 2 presents audio files along with their size in MB and then it convert that audio files to text using ASR algorithm. And gives the total time for audio to text conversion and also the total extracted words.

TABLE 3
Precision And Recall values

text precision	text recall	Thumbnail Creation	Total Images
0.8	0.9	4520ms	3
0.4	0.2	23615ms	10
0.86	0.9	59729ms	25
0.32	0.5	3668ms	3
0.78	0.88	19096ms	6
0.8	0.9	56287ms	29

We have calculated precision and recall from Table 1 and Table 2 on different computer network(CN) videos in Table 3.

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

V. SIMULATION RESULTS

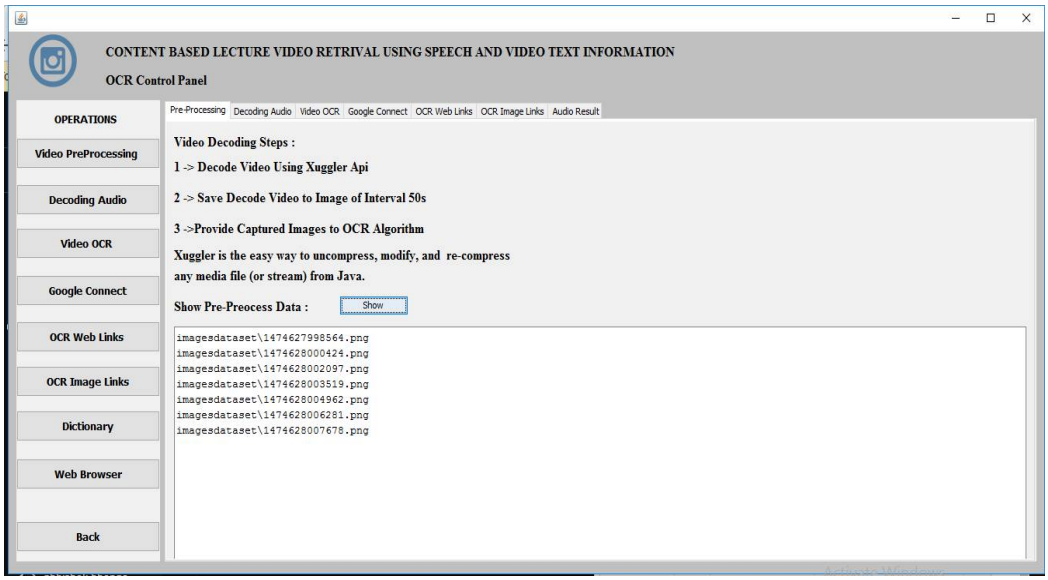

Figure 2. Preprocessing of video file

In the execution of system first we decode the video using Xuggler api.Then after interval of every 50 sec we save images from video and provide it to OCR algorithm.

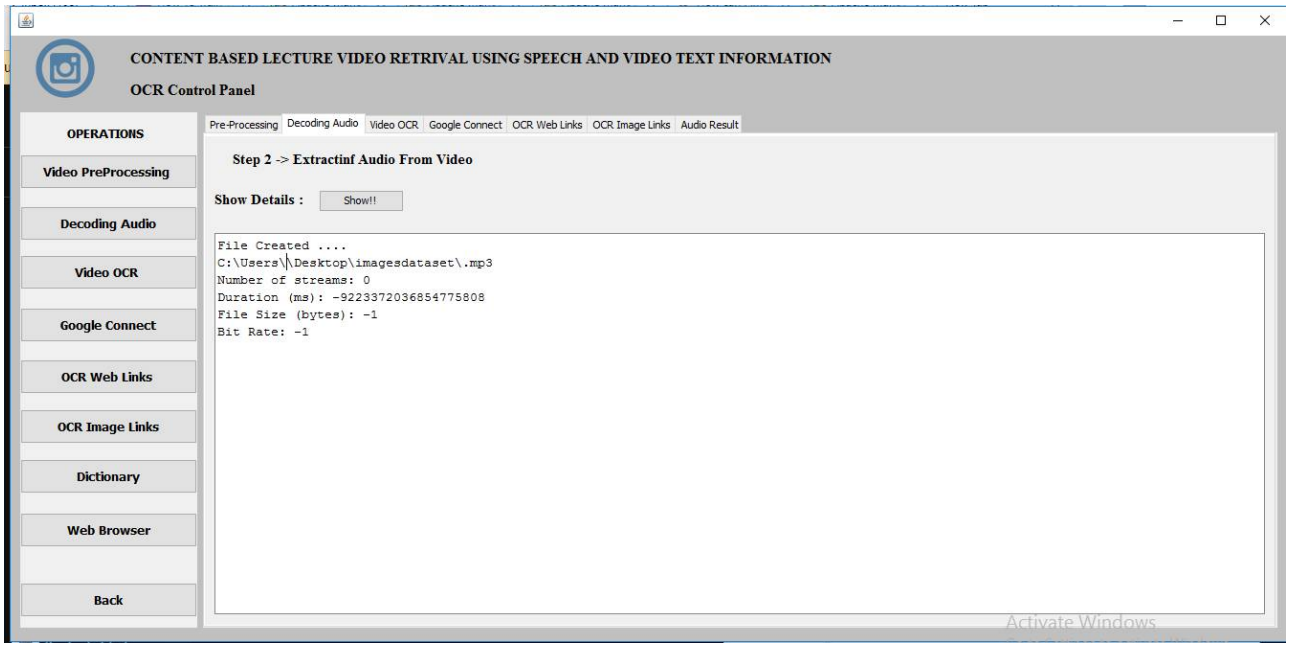

Figure 3. Audio extraction from video file.

Then we extract audio from video separately using ASR algorithm and create a audio file.

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

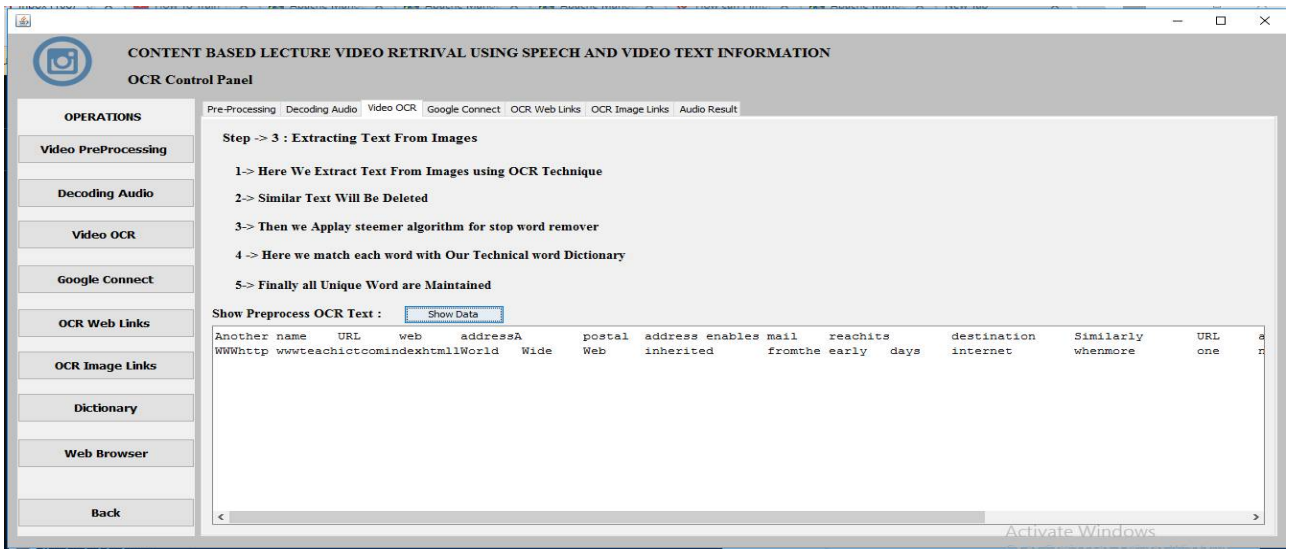

Figure 4. Extraction of text from images

Now we will extract text from images using OCR technique.we also removed duplicate words from the content and used porter stemming algorithm to remove stopwords .we maintained a dictionary of unique words.

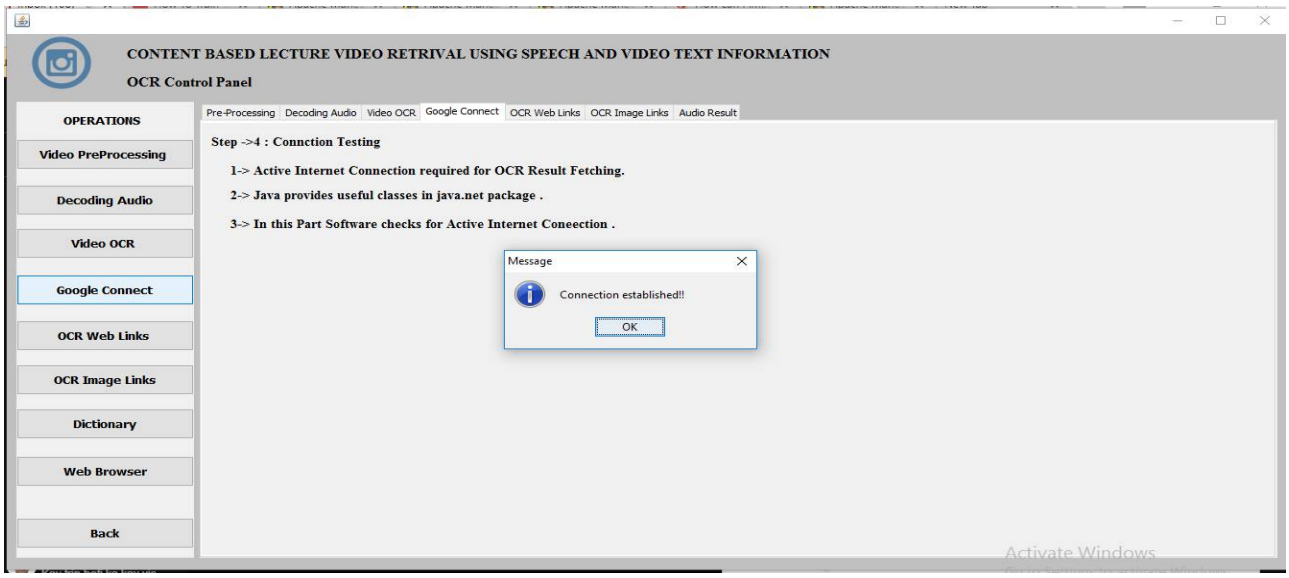

Figure 5. Connect google to get OCR web and images links.

We require google connection to get OCR web links and OCR image links.also system will can recommend different other videos and links and images related to intended video which are rated by other users of system which can be useful to current user to fulfil information requirement.

VI. CONCLUSION AND FUTURE WORK

In this system an algorithm is devised for content based retrieval of video to give effective search results to E-learners .The methodology used is more efficient than the existing one. The main ingredients of this process are detected audio from video using ASR algorithm within some time interval and accuracy of character set given as input

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

to OCR algorithm for text extraction. Both of these algorithms give highly accurate results within less computation time .In future we plan to make use of more efficient and usable algorithms to separate audio from video ,to extract text from audio and video segments and to divide video in multiple segments to choose key-frames among them.We can increase the number of nodes and analyze the performance.

REFERENCES

1. Haojin Yang and Christoph Meinel, "Content Based Lecture Video Retrieval Using Speech and Video Text Information",IEEE Transactions On Learning Technologies,Vol.7,No.2 April-June.
2. E. Leeuwis, M. Federico and M. Cettolo, "Language modeling and transcription of the ted corpus lectures",in Proc. IEEE Int.Conf. Acoust.,Speech Signal Process ,pp. 232235,2003.
3. D. Lee and G. G. Lee, "A korean spoken document retrieval system for lecture search",in Proc. ACM Special Interest Group Inf. Retrieval Searching Spontaneous Conversational Speech Workshop,2008.
4. W.Heurst,T.Kreuzer,and M.Wiesenhutter, "A qualitative study towards using large vocabulary automatic speech recognition to index recorded presentations for search and access over the web",in Proc. IADIS Int. Conf.WWW/Internet,pp.135143,2002.
5. A. Haubold and J. R. Kender, "Augmented segmentation and visualization for presentation videos",in Proc. 13th Annu. ACM Int. Conf. Multimedia,pp. 5160,2005.
6. James Glass,Timothy J. Hazen,Lee Hetherington and Chao Wang, "Analysis and Processing of Lecture Audio Data: Preliminary Investigations",in Proc. HLT-NAACL Workshop Interdisciplinary Approaches Speech Indexing Retrieval,pp.912,2004.
7. G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval",Inf. Process. Manage., vol. 24, no. 5,pp. 513 523,1988.
8. B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform",in Proc. Int. Conf. Comput. Vis. Pattern Recog.,pp. 29632970,2010.
9. C. Meinel, F. Moritz, and M. Siebert, "Community tagging in tele-teaching environments", in Proc. 2nd Int. Conf.e-Educ.,e- Bus.,e-Manage and E-Learn,2011.
10. M. Grcar, D. Mladenic, and P. Kese, "Semi-automatic categorization of videos on videolectures.net", in Proc. Eur. Conf. Mach. Learn. Knowl. Discovery Databases ,pp.730733,2009.
11. H. Yang, B. Quehl, and H. Sack, "A framework for improved video text detection and recognition" ,Multimedia Tools Appl.,pp.129,2012.
12. J. Eisenstein, R. Barzilay, and R. Davis, "Turning lectures into comic books using linguistically salient gestures" ,in Proc. 22nd Nat.Conf.Artif.Intell.,1,pp.877882,2007.

BIOGRAPHY

Miss. Surabhi Dilip Pagar has received B.E degree in Information Technology from NDMVP COE, Nasik. She is currently pursuing Master Degree in Computer Engineering at SVIT, Chincholi ,Nasik, India .Her research interest includes Data Mining ,DBMS.

Prof. Gorakshnath J. Gagare presently working as assistant professor in the Department of Computer Engineering at SVIT, Chincholi ,Nasik .He received B.E. degree in Computer Science And Engineering from Government College Of Engineering ,Aurangabad and M.Tech.in Computer Engineering from BVUCOE, Pune ,India. .