Jun. 30
International Journal of advanced studies in Computer Science and Engineering
IJASCSE, Volume 4, Issue 6, 2015

Holistic Approach for Multimodal Lecture Video Retrieval

Dipali Patil
Department of Computer Engineering
D.Y. Patil College of Engineering, Akurdi
Savitribai Phule Pune University, Pune, India

Mrs. M. A. Potey
Department of Computer Engineering
D.Y. Patil College of Engineering, Akurdi
Savitribai Phule Pune University, Pune, India

***Abstract*- Many organizations and universities provide distance learning by recording classroom lectures and making them available to students over the Internet. A repository generally contains hundreds of such lecture videos. Each lecture video is typically a more than hour's duration and is often huge. It is sometimes clumsy for students to search through an entire video, or across many videos, in order to find portions of their immediate interest. It is desirable to have a system that takes user-given keywords as a query and provides a results of relevant lecture videos and highlight the specific part within the video where that concept is explained. The lecture recordings having great variety of media or modalities such as video, audio, lecture media, presentation behavior and formats. This paper propose a technique to automatically generate text metadata and tags for Multimodal lecture videos. This is based on generating text transcripts automatically using a text recognition engine and automatic indexing and search of the transcripts. Proposed system takes keywords from users as a query and returns a list of videos which provide a link to video and concept of video using timestamps as the results.**

Keywords- Content based Video Retrieval, Multimodel Recordings, Video Segmentation, Optical Character Recougnision

I. INTRODUCTION *(HEADING 1)*

Since recording technologies have become increasingly robust and easier to use in the last few years, number of universities are taking the opportunity to record their lectures and publish them online in order to make them available for students. Main focus is on the lecture recordings which capture the frames projected onto the screen during the lecture through a frame grabber tool.

Text is a high-level semantic feature which has usually been used for content-based information retrieval. In lecture videos, visual texts from lecture slides serve as an outline for the lecture and are important concept for understanding. Therefore after segmenting a lecture video into a set of key frames, the text detection procedure will be applied and executed on each key frame, and the extracted text objects metadata will be further used in text detection, verification and slide structure analysis processes. Especially, the extracted text metadata can enable more flexible video browsing and video search functions.

Content-based retrieval includes gathering of textual data within video that has to be provided manually by the users. For this, techniques from common OCR- high-resolution scan of printed text documents have to be improved and adapted and it also applicable for video OCR. In video OCR, video key frames containing visual textual information have to be identified first. Then, the text on each key frame has to be separated from its background, and geometrical transformations have to be applied before common OCR algorithms can process the text successfully.

This paper presents an entire work-flow for the slide segmentation of lecture videos, the video OCR analysis, and the slide structure Analysis for Multimodal lecture video recordings. Approach is based on the analysis of fast connected component (CC) called CC differencing metrics and slide transition recognition on multimodal lecture videos. Text detection and text verification scheme using Stroke Width Transform (SWT) is used on text transcript. SWT-based analysis is a verification procedure to remove false alarms. To segment text from its heterogeneous background, a multi-hypotheses framework consisting of multiple text segments, common OCR analysis, spell checking, and result merging processes is implemented.

The major focus of this paper is to extract metadata from visual resources of lecture videos automatically by applying appropriate analysis techniques. For evaluation purposes, several automatic indexing functionalities are developed in a lecture video portal, which can guide both visually and text-oriented users to navigate within lecture video. For visual analysis, a new method for slide video segmentation is used and applied for video OCR to gather text metadata. Furthermore, lecture outline is extracted from OCR transcripts by using stroke width and geometric information. This can provide outline for lecture video browsing and search. By using this outline, user can watch any part of video where concept is explained.

The rest of the paper is organized as follows: Section II reviews related work in lecture video retrieval and content based video search domain. Section III describes video segmentation and key frame extraction process. A content based lecture video search engine using multimodal information resources is introduced in further sections. Finally, Section V concludes the paper with an outlook on future work.

II. RELATED WORK

A. *Existing Lecture Video Repositories*

NPTEL [24], freevideolectures.com [25] and MIT Open Courseware [26] are some of the existing lecture video repositories. We investigated support for search and browsing features available in those repositories explained in table I. Some repositories have manual transcriptions, subtitles for lecture videos but they are not providing use of them to provide search features, Dipali et al., [23], present survey on lecture video retrieval.

B. *Lecture Video Retrieval*

Tuna et al. [4] presented method for lecture video indexing and search. First, they segment lecture videos into representative's key frames by using frame differencing metrics. Then standard Optical character recognition engine is applied for gathering textual metadata from slide, in which they apply some image transformation methods to improve the OCR result. Jeong et al., [5] proposed a lecture video segmentation method based on Scale Invariant Feature Transform (SIFT) feature and the adaptive threshold. In their work, SIFT feature is mainly applied to measure slides with similar content. An adaptive threshold selection algorithm is used to detect slide transitions between video frames. In their evaluation, this approach achieved promising results for processing one scene lecture video.

Recently, collaborative tagging has become a popular and standard functionality in lecture video portals. Sack and Waitelonis [6] and Moritz et al. [7] uses tagging data mechanism for lecture video retrieval and search. Automatic Speech Recognition (ASR) offers speech-to-text information on audio lecture videos, which is thus well appropriate for content based lecture video retrieval. Leeuwis et al., [8] and Munteanu et al., [9] concentrate the research on English speech recognition for Technology Entertainment and Design (TED) lecture videos and webcasts. In this, the training corpus is created manually, which is thus difficult to be improved or optimized periodically. In this way, OCR and ASR are used to obtain text and speech transcript respectively.

C. *Content Based Retrieval*

Several content-based video search engines have been proposed in recent times. Adcock et al. [10] proposed a lecture webcast search system in which they applied a slide frame segmenter to extract lecture slide images. The system retrieved more than 36,000 lecture videos from different resources such as YouTube, Berkeley, etc.

TABLE I. LECTURE VIDEO REPOSITORIES COMPARISION

Repository	Search	Navigation Features
NPTEL	No	No
Freelecturevideos.com	Meta-data	No
Videolectures.com	Meta-data	Slide Synchronization
MIT Open course ware	Content	Speech-transript

The search indices are created based on the metadata obtained from the video hosting website and texts extracted from slide videos by using a standard OCR engine.

H. J. Jeong et al., [13] proposed accurate method for video segmentation using SIFT and an Adaptive threshold. Using SIFT, it is easily possible to compare two slides frames, having similar contents but different backgrounds. And then calculate Frame transition somewhat accurately by using Adaptive threshold. Haojin Yang and Christoph Meinel [16] presented an approach for video indexing and video search in large lecture video archives. First of all, they apply video segmentation and key-frame extraction to offer a visual guideline for the video content navigation and then keywords are extracted for video navigation within lecture video.

A vital approach to retrieve text from a color image was given by Y. Zhan et al., [14] the proposed algorithm uses the multiscale Wavelet features and the structural information to locate the text lines. Then a Support Vector Machine (SVM) classier was used to get the exact text from those previously located text lines. H.Yanget.al. [15] have developed a Skeleton-Based binarization method to separate and extract text from complex backgrounds. These can be processed by standard OCR software. Toni-Jan Keith Palma Monserrat et al. [17] developed and evaluated new Note Video system for identifying the conceptual objects of a blackboard-based video. Similarly Yen-Chia Hsu et al. [22] developed syntag system for labelling real time video which uses three types of tags Good, Question, and Disagree.

III. PROPOSED SYSTEM

The proposed system includes user interface through which user enter keyword to search, then slide video segmentation and key frame extraction done on each frame, retrieve the targeted slide as shown in fig 1.

A. *SVM Classification and Thresholding Selection*

Stefano Masneri and Oliver Schreer [18] presented classification system for video lectures and conferences based on Support Vector Machines (SVM). Using this, class-

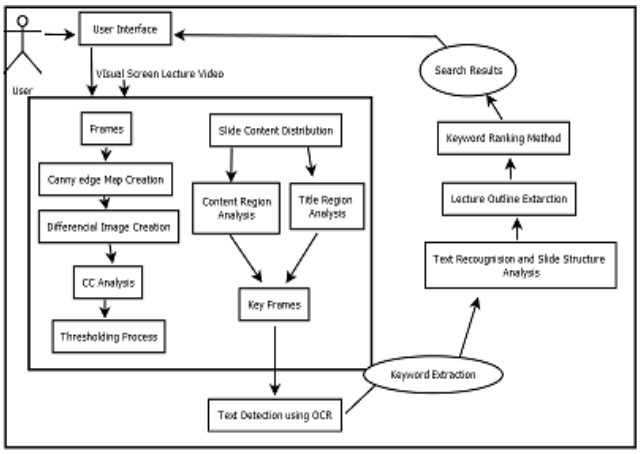

Figure 1. System Architecture Diagram

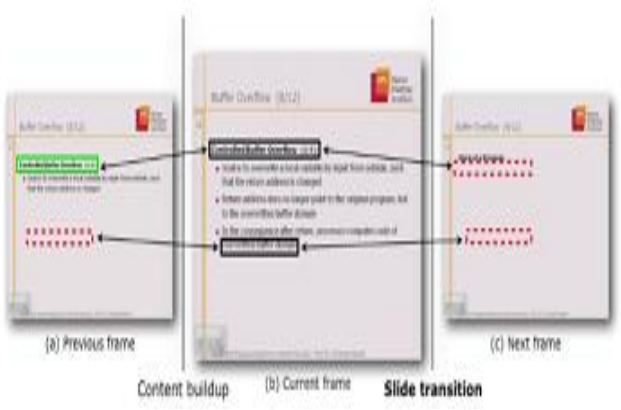

Figure 2. We detect the rst and the last object line in Rc vertically and perform the CC differencing metric on those line sequences from adjacent frames. In frame (a) and (b), a same text line (top) can be found; whereas in frame (b) and (c), the CC-based differences of both text lines exceed the threshold Ts2. A slide transition is thus found between frame (b) and (c)[16]

ification is performed on videos into four different classes (talk, presentation, blackboard, mix). On top of this, the system further analyses presentation segments to detect slide transitions, animations and dynamic content such as video inside the presentation. The system performs the classification on frame-by-frame basis. Thresholding selection algorithm is used for detecting slide frames having textual data from multimodal lecture videos [19] [20].

B. Video Segmentation and Key frame Extraction

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. To index the video, segmentation is performed on video, to obtain the representative key-frames. In the lecture video domain, the video sequence of an individual lecture topic or subtopic is often considered as a video segment. Many approaches [1], [2] make use of global pixel-level differencing metrics for capturing slide transitions. A drawback is that the salt a pepper noise of video signal can affect the segmentation accuracy. After observing the content of lecture slides, major content are present on slides as, e.g., text lines, gures, tables, etc., can be considered as Connected Components (CCs). Therefore use CC instead of pixel as the basis element for the differencing analysis.

Segmentation algorithm consists of two steps as shown in Fig 3. In the first step, the entire multimodal slide video is analyzed. For reasons of efficiency and accuracy, do not perform the analysis on every video frame; instead, established a time interval of three second and analyzed only one frame per second. Then create the differential edge map creation process of two adjacent frames and perform the CC analysis on this edge map. In order to ensure that each newly emerging knowledge change point or newly added figure within a slide can be detected, we have identified the segmentation threshold value T_{s1}.

This means that a new segment is captured if the number of CCs of a differential image exceeds T_{s1}. The result of the first segmentation step is too redundant for indexing, since there are many changes within the same slide. Hence, the segmentation process continues with the second step that aims to find the actual slide page transition based on the frames detected in the first step.

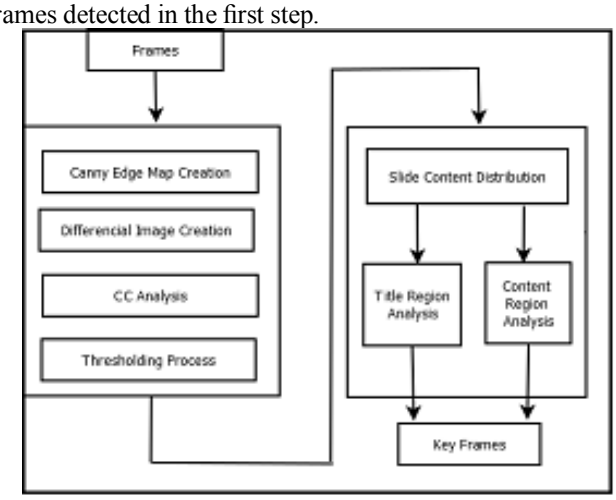

Figure 3. Slide Video Segmentation Workflow

After a statistical analysis of large amounts of slide videos, generally the content distribution of the most commonly used slide style (Fig. 2(b)). Let R_t and R_c denote the title and the content regions in Fig. 2(b) which account for 23% and 70% of the entire frame height respectively. If the amount of CCs in the differential edge image for R_t exceeds the threshold Ts2, a slide page transition is captured. Otherwise, in R_c, we detect the first text lines top-down and bottom-up respectively, and perform the corresponding CC differencing analysis on these text line images from two adjacent frame. If both of the two difference values exceed T_{s2}, a slide page transition is also captured. In this study, set the values of T_{s1} and T_{s2} to 20 and 5 respectively.

C. Text Detection using OCR

Texts in the lecture slides are closely related to the lecture content can thus provide important information for the retrieval task. Haojin Yang et al. [21] presented an approach of lecture video indexing using OCR technology. In this framework, a novel video OCR system is developed for gathering video text. For text detection, a localization verification scheme is used. 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. By applying the open-source print OCR engine tesseract-ocr, we achieved recognition of 96 percent.

D. Text Verification

Stroke Width Transform (SWT) [12] based verification procedures are applied to remove the non-text blocks. The

algorithm computes the width of the most likely stroke containing this pixel where each pixel contains the potential stroke width value of input image pixels. The output of SWT is a feature map Since the SWT verifier is not able to correctly identify special non-text patterns such as sphere, window blocks, garden fence, system adopted an additional SVM classifier to sort out these non-text patterns in order to further improve the detection accuracy. Then apply a verification method based on SWT analysis. Epshtein et al., [12] have proved in their work that the stroke width feature is robust enough to distinguish text from other complicated image elements. In order to use the method for this approach, there is need to extend it in following manners: First, Epshtein et al., [12] describe that the SWT analysis depends on the edge gradient direction. In order to accommodate both bright text on dark background and vice versa, they apply the algorithm twice for each gradient direction, this leads to a loss of performance.

However, two-stage analysis approach solves this issue: In the first stage, single-line text bounding boxes already are detected. Subsequently, a corresponding sub image is created for each bounding box, then apply SWT on those sub images in order to improve performance. Hence, it is relatively easy to determine the gradient direction for such bounding box images: First of all, convert the bounding box image to a grayscale image and binaries it subsequently using thresholding method.

- A box is discarded if its stroke width variance exceeds a threshold range *MinVar* and *MaxVar*.
- Its mean stroke width exceeds a threshold range *MinStroke* and *MaxStroke*.

For some boxes with a size larger than *MaxH* which may contain both text lines and other image objects, apply the detection algorithm proposed by Epshtein et al., [12] on them. In this way, the detection recall can in turn be increased during the verification process. In our study, *MinVar* and *MaxVar* have been set to 100 and 500, while *MinStroke* and *MaxStroke* have been set to 1 and 10, respectively. These values were also learned from our training data.

E. 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, system 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. Method further opens up the video content and enables the search engine to give more accurate and flexible search.

1) Title Identification: The title recognition procedure is intended to identify potential title text lines within the key frames. It is based on the detected text objects geometrical information. A potential title object is identied if:

- It is in the upper third of the image.
- Its width is larger than the minimal word width MinW.
- It is one of the three highest bounding boxes and has the uppermost position.

If a text objects satisfies the above conditions, it will be labeled as a title line. The process repeats these steps for the remaining text objects in order to find the next potential title lines; however, title lines within the same frame must have similar height and stroke width.

2) Text Object Classification: After the title identification, the remaining text objects are classied into three classes: subtitle/key point (bold texts), normal content and footline. The classification is based on the stroke width s_t and the height h_t of text objects. Let s_{mean} and h_{mean} denote the average stroke width and the average height of text objects within a slide frame respectively. The classification is processed as follows:

key-point if $s_t > s_{mean} \wedge h_t > h_{mean}$

footline if $s_t < s_{mean} \wedge h_t < h_{mean} \wedge y = y_{max}$

content text otherwise.

F. Video content browsing using keyword extraction

Keywords can summarize a document and are widely used for information retrieval in digital libraries. Segment-level as well as video-level keywords are extracted from OCR information resources. 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, first arrange each 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 [3] 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. Then calculate the weighting factor for each remaining keyword by extending the standard Tf-idf score [11]. Therefore, following formula is used for calculating Tf-idfscore:

$$tfidf_{seg-level}(kw) = \frac{1}{N}\left(tfidf_{ocr} \cdot \frac{1}{n_{type}} \sum_{1}^{n_{type}} w_i\right)$$

Where *kw* is the current keyword, $tfidf_{ocr}$ denote its Tf-idf score computed from OCR resource, *w* is the weighting factor for various resources, *ntype* denotes the number of various OCR text line types. N is the number of available information resources, in which the current keyword can be found, namely the corresponding Tf-idf score does not equal

0. Following is a formula for computing the video-level Tf-idf score, as shown below

$$tfidf_{video-level}(kw) = \frac{1}{N}(\sum_1^n tfidf_i w_i)$$

Where $tfidf_i$ and w_i denote the TFIDF score and the corresponding weighting factor for each information resource.

IV. EXPERIMENTAL METHODOLOGY AND RESULTS

A. Dataset

In proposed system, TED dataset [27] and VideoLectures Dataset [25] are used. In order to test slide video segmentation and key frame extraction for multimodal lecture videos, a subset of lecture videos has been used from both the dataset. The system is composed by 5 videos from the TED dataset and 4 videos from the VideoLectures dataset, for a total of around 4 hours of video content.

B. Results

To evaluate slide-video segmentation method, compile a lecture video data set, which consist of 5 randomly selected videos. Each unique slide with complete contents is regarded as a Key-frames. For our lecture video test set, the achieved segmentation accuracy is 96%. The classification of video frames is done by using SVM. But if some threshold is applied for detection of correct frames, it will give lower accuracy than SVM classifier. Table III provide classification results by using SVM Classification and Thresholding Selection.

TABLE II. COMPARISION OF CLASSIFICATION ACCURACY USING SVM AND THRESHOLDING SELECTION

Video No.	Video Name	SVM Classification Accuracy	Thresholding Selection Accuracy
1	Viola-Jones Rapid Object Detection	86.85%	82.35%
2	The Business Aspects of Software Enginering	91.75%	85.66%
3	Developing Requirements in Software Engineering	94.87%	87.00%

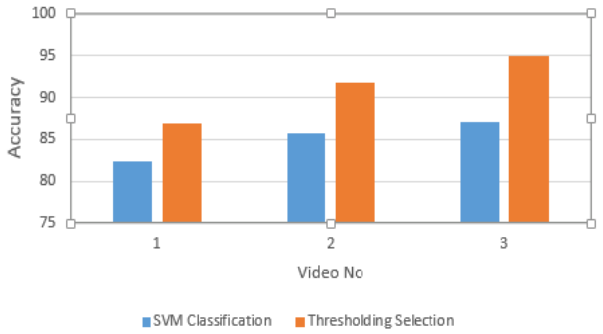

Figure 4. Comparision of SVM Classification Accuracy and Thresholding Selection

The comparison result of two features of video such title section and key point sections are illustrated in Table III. Using title section, we will get higher recall, precision results because title section contains more information than key point.

TABLE III. EVALUATION RESULTS OF LECTURE OUTLINE EXTRACTION

	Recall	Precision	F measure
Title	0.86	0.95	0.90
Key-point	0.61	0.77	0.68

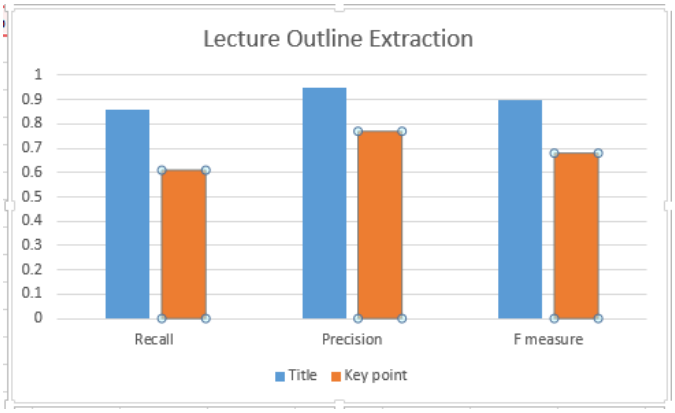

Figure 5. Evaluation Results of Lecture Outline Extraction

Table IV shows the evaluation results for different keyword in terms of recall as R, precision as P, F measure as F M, Word error rate as WER, word accuracy as WA and time period as T in seconds required to search that particular keyframe.

TABLE IV. EVALUATION RESU;LTS ON DIFFERENT QUERY

Keyword	R	P	F M.	WER	WA	T
Object Detection	1	1	1	0.0083	0.99	0.69
Integral Image	1	0.16	0.28	0.016	0.98	0.29
Domain Analysis	1	0.25	0.4	0.0083	0.99	0.304

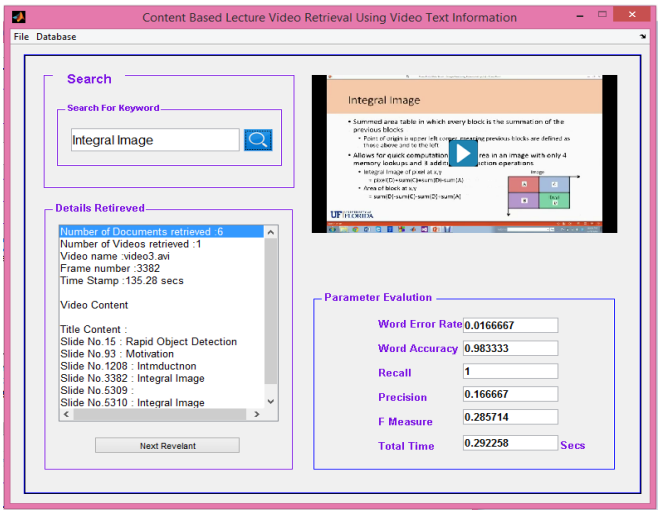

Figure 6. Schematic Diagram showing results

V. Conclusion and Future Work

In this paper, we have tried to present an effective, robust system for indexing and analyzing lecture videos. Slide video segmentation and verification scheme is presented and evaluated. Furthermore, this paper proposed a method for slide structure analysis using the geometrical information and SWT for Multimodal lecture videos.

As the future work, the usability study for the video search function in our lecture video portal will be conducted. Online processing of video in such a manner will be a challenge. Automated annotation for OCR and ASR results using Linked Open Data resources offers the opportunity to enhance the amount of linked educational resources significantly. Therefore more efficient search and recommendation method could be developed in lecture video archives.

Acknowledgment

We express our thanks to publishers, researchers for making their resource available & teachers for their guidance. We also thank the college authority for providing the required infrastructure and support. Last but not the least we would like to extend a heartfelt gratitude to friends and family members for their support.

References

[1] Tuna, T., Subhlok, J., Barker, L., Varghese, V., Johnson, O., and Shah, S. "Development and evaluation of indexed captioned searchable videos for STEM coursework", Proceedings of the 43rd ACM technical symposium on Computer Science Education. ACM, 2012.

[2] Adcock, J., Cooper, M., Denoue, L., Pirsiavash, H., and Rowe, L. A. "Talkminer: a lecture webcast search engine", Proceedings of the international conference on Multimedia. ACM, 2010.

[3] Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. "Feature-rich part-of-speech tagging with a cyclic dependency network", Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. Association for Computational Linguistics, 2003.

[4] Tuna, T., Subhlok, J., Barker, L., Varghese, V., Johnson, O., and Shah, S. "Development and evaluation of indexed captioned searchable videos for STEM coursework", Proceedings of the 43rd ACM technical symposium on Computer Science Education. ACM, 2012.

[5] Jeong, Hyun Ji, Tak-Eun Kim, and Myoung Ho Kim "An accurate lecture video segmentation method by using sift and adaptive threshold", Proceedings of the 10th International Conference on Advances in Mobile Computing and Multimedia. ACM, 2012.

[6] Sack, Harald, and Jrg Waitelonis. "Integrating social tagging and document annotation for content-based search in multimedia data.", Semantic Authoring and Annotation Workshop (SAAW). 2006.

[7] Moritz, F., M. Siebert, and C. Meinel "Community tagging in teleteaching environments", 2nd International Conference on eEducation, e-Business, e-Management and E-Learning (to appear). 2011.

[8] Leeuwis, Erwin, Marcello Federico, and Mauro Cettolo "Language modelling and transcription of the TED corpus lectures", Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP'03). 2003 IEEE International Conference on. Vol. 1. IEEE, 2003.

[9] Munteanu, C., Penn, G., Baecker, R., and Zhang, Y. "Automatic speech recognition for webcasts: how good is good enough and what to do when it isn't.", Proceedings of the 8th international conference on Multimodal interfaces. ACM, 2006.

[10] Adcock, J., Cooper, M., Denoue, L., Pirsiavash, H., and Rowe, L. A. "Talkminer: a lecture webcast search engine", Proceedings of the international conference on Multimedia. ACM, 2010.

[11] Salton, Gerard, and Christopher Buckley. "Term-weighting approaches in automatic text retrieval", Information processing and management 24.5 (1988): 513-523.

[12] Epshtein, Boris, Eyal Ofek, and Yonatan Wexler "Detecting text in natural scenes with stroke width transform.", Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on IEEE, 2010.

[13] Jeong, Hyun Ji, Tak-Eun Kim, and Myoung Ho Kim "An accurate lecture video segmentation method by using sift and adaptive threshold", Proceedings of the 10th International Conference on Advances in Mobile Computing and Multimedia. ACM, 2012.

[14] Zhan, Yaowen, Weiqiang Wang, and Wen Gao. "A robust split-andmerge text segmentation approach for images", PatternRecognition,2006.ICPR 2006. 18th International Conference on. Vol. 2. IEEE, 2006.

[15] Yang, Haojin, Bernhard Quehl, and Harald Sack. "A framework for improved video text detection and recognition", Multimedia Tools and Applications 69.1 (2014): 217-245.

[16] 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, APRILJUNE 2014

[17] Toni-Jan Keith Palma Monserrat, Shengdong Zhao, Kevin McGee, Anshul Vikram Pandey "NoteVideo: Facilitating Navigation of

Blackboardstyle Lecture Videos", Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 2013.

[18] Stefano Masneri, Oliver Schreer "SVM-based Video Segmentation and Annotation of Lectures and Conferences",

[19] Huizhong Chen, Matthew Cooper, Dhiraj Joshi, Bernd Girod "Multimodal Language Models for Lecture Video Retrieval ", Proceedings of the ACM International Conference on Multimedia. ACM, 2014'

[20] Gtl, C., Chang, V., Spnola, . M. F., Sampaio, P. N. M., and Kappe, F. "Indexing and retrieval of multimodal lecture recordings from open repositories for personalized access in modern learning settings."World Conference on Educational Multimedia, Hypermedia and Telecommunications. Vol. 2009. No. 1. 2009.

[21] Haojin Yang, Maria Siebert, Patrick Lhne, Harald Sack, Christoph Meinel "Lecture video indexing and analysis using video ocr technology." Signal-Image Technology and Internet-Based Systems (SITIS), 2011 Seventh International Conference on. IEEE, 2011.

[22] Yen-Chia Hsu, Tay-Sheng Jeng, Yang-Ting Shen, Po-Chun Che "SynTag: a web-based platform for labeling real-time video." Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work. ACM, 2012.

[23] Patil, Dipali, and Mrs MA Potey. "Survey of Content Based Lecture Video Retrieval."

[24] http://www.nptel.iitm.ac.in

[25] http://www.freevideolectures.com

[26] http://ocw.mit.edu/courses/audio-video-courses

[27] https://www.idiap.ch/dataset