ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Optical Character Recognition Based Video Retrieval Application for High End Application

Palla Madhava Reddy
PG scholar in Digital Systems & Computer Electronics,
Holy Mary Institute of Technology.
Keesara, Rangareddy, Hyderabad, Telangana.

M. Devaraju
HoD
Department of ECE
Holy Mary Institute of Technology.
Keesara, Rangareddy, Hyderabad, Telangana.

ABSTRACT:

Transcribing lectures is a challenging task, both in acoustic and in language modeling. In this work, we present our first results on the automatic transcription of lectures from the TED corpus, recently released by ELRA and LDC. In particular, we concentrated our effort on language modeling. Baseline acoustic and language models were developed using respectively 8 hours of TED transcripts and various types of texts: conference proceedings, lecture transcripts, and conversational speech transcripts. Then, adaptation of the language model to single speakers was investigated by exploiting different kinds of information: automatic transcripts of the talk, the title of the talk, the abstract and, finally, the paper. In the last case, a 39.2% WER was achieved. Proposed Method: In the last decade e-lecturing has become more and more popular. The amount of lecture video data on the World Wide Web (WWW) is growing rapidly. Therefore, a more efficient method for video retrieval in WWW or within large lecture video archives is urgently needed. This paper presents an approach for automated video indexing and video search in large lecture video archives. First of all, we apply automatic video segmentation and key-frame detection to offer a visual guideline for the video content navigation. Subsequently, we 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. Extension Method: As an extension method we adopt the OCR method for text extraction from a given video and Speaker recognition process as to identify that which speaker delivering this given video lecture on the same topic presentation. In proposed we didn't have any speaker verification its only audio extraction now we add speaker identification process as an extension and save this data in a template which is used to generate the relevant documents or power point presentation.

INTRODUCTION

DIGITAL video has become a popular storage and exchange medium due to the rapid development in recording technology, improved video compression techniques and high-speed networks in the last few years. Therefore audiovisual recordings are used more and more frequently in e-lecturing systems. A number of universities and research institutions are taking the opportunity to record their lectures and publish them online for students to access independent of time and location. As a result, there has been a huge increase in the amount of multimedia data on the Web. Therefore, for a user it is nearly impossible to find desired videos without a search function within a video archive. Even when the user has found related video data, it is still difficult most of the time for him to judge whether a video is useful by only glancing at the title and other global metadata which are often brief and high level. Moreover, the requested information may be covered in only a few minutes, the user might thus want to find the piece of information he requires without viewing the complete video. The problem becomes how to retrieve the appropriate information in a large lecture video archive more efficiently. Most of the video retrieval and video search systems such as YouTube, Bing and Vimeo reply based on available textual metadata such as title, genre, person, and brief description, etc. Generally, this kind of metadata has to be created by a human to ensure a high quality, but the creation step is rather time and cost consuming. Furthermore, the manually provided metadata is typically brief, high level and subjective. Therefore, beyond the current approaches, the next generation of video retrieval systems applies automatically generated metadata by using video analysis technologies. Much more content-based metadata can thus be generated which will lead to two research questions in the e-lecturing context: Can those metadata assist the learner in searching required lecture content more efficiently? If so, how can we extract the important metadata from lecture videos and provide hints to the user? According to the questions, we postulated the following hypothesis: Hypothesis 1. The relevant metadata can be automatically gathered from lecture videos by using appropriate analysis techniques. They can help a user to find and to understand lecture contents more efficiently, and the learning effectiveness can thus be improved. Traditional video retrieval based on visual feature extraction cannot be simply applied to lecture recordings because of the homogeneous scene composition of lecture videos. Fig. 1a shows an exemplary lecture video recorded using an outdated format produced by a single video camera. Varying factors may lower the quality of this format. For example, motion changes of the camera may affect the size, shape and the brightness of the slide; the slide can be partially obstructed when the speaker moves in front of the slide; any changes of camera focus (switching between the speaker view and the slide view) may also affect the further slide detection process. Nowadays people tend to produce lecture videos by using multi-scenes format (cf. Fig. 1b), by which the speaker and his presentation are displayed synchronously. This can be achieved either by displaying a single video of the speaker and a synchronized slide file, or by applying a state.

In this work, we present our first results on the automatic transcription of lectures from the TED corpus, recently released by ELRA and LDC. In particular, we concentrated our effort on language modeling. Baseline acoustic and language models were developed using respectively 8 hours of TED transcripts and various types of texts: conference proceedings, lecture transcripts, and conversational speech transcripts. Then, adaptation of the language model to single speakers was investigated by exploiting different kinds of information: automatic transcripts of the talk, the title of the talk, the abstract and, finally, the paper. In the last case, a 39.2% WER was achieved. D. Lee and G. G. Lee, (2008), we introduced a Korean spoken document retrieval system for lecture search. We automatically build a general inverted index table from spoken document transcriptions, and we extract additional information from textbooks or slide notes related to the lecture. We integrate these two sources for a search process. The speech corpus used in our 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. J. Glass, T. J. Hazen, L. Hetherington, and C. Wang, et al, (2004), we report on our 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, we have collected a corpus of 270 hours of speech from a variety of undergraduate courses and seminars. We report on an initial analysis of the spontaneous speech phenomena present in these data and the vocabulary usage patterns across three courses. Finally, we examine language model perplexities trained from written and spoken materials, and describe an initial recognition experiment on one course.

Fig. 1b illustrates an example of such a system which delivers two main parts of the lecture: the main scene of lecturers which is recorded by using a video camera and the second which captures the desktop of the speaker's computer (his presentations) during the lecture through a frame grabber tool. The key benefits of the latter one for a lecturer are the flexibility. For the indexing, no extra synchronization between video and slide files is required, and we do not need to take care of the slide format. The main drawback is that the video analysis methods may introduce errors. Our research work mainly focuses on those lecture videos produced by using the screen grabbing method. Since two videos are synchronized automatically during the recording process. Therefore, the temporal scope of a complete unique slide can be considered as a lecture segment. This way, segmenting two-scenes lecture videos can be achieved by only processing slide video streams, which contain most of the visual text metadata. The extracted slide frames can provide a visual guideline for video content navigation. Text is a high-level semantic feature which has often been used for content-based information retrieval. In lecture videos, texts from lecture slides serve as an outline for the lecture and are very important for understanding. Therefore after segmenting a video file into a set of key frames (all the unique slides with complete contents), the text detection procedure will be executed on each key frame, and the extracted text objects will be further used in text recognition and slide structure analysis processes. Especially, the extracted structural metadata can enable more flexible video browsing and video search functions. Speech is one of the most important carriers of information in video lectures. Therefore, it is of distinct advantage that this information can be applied for automatic lecture video indexing. Unfortunately, most of the existing lecture speech recognition systems in the reviewed work cannot achieve a sufficient recognition result, the Word Error Rates (WERs) having been reported from [1], [2], [3], [4], [5] and [6] are approximately 40–85 percent. The poor recognition results not only limit the usability of speech transcript, but also affect the efficiency of the further indexing process. In our research, we intended to continuously improve the ASR result for German lectures by building new speech training data based on the open-source ASR tool. However, in the open-source context, it lacks method for generating the German phonetic dictionary automatically, which is the one of the most important part of ASR software. Therefore, we developed an automated procedure in order to fill this gap. A large amount of textual metadata will be created by using OCR und ASR method, which opens up the content of lecture videos. To enable a reasonable access for the user, the representative keywords are further extracted from the OCR and ASR results. For content-based video search, the search indices are created from different information resources, including manual annotations, OCR and ASR keywords, global metadata, etc. Here the varying recognition accuracy of different analysis engines might result in solidity and consistency problems, which have not been considered in the most related work (cf. Section 2.2). Therefore, we propose a new method for ranking keywords extracted from various information resources by using the extended Term Frequency Inverse Document Frequency (TFIDF) score [7]. The ranked keywords from both segment and video-level can directly be used for video content browsing and video search. Furthermore, the video similarity can be calculated by using the Cosine Similarity Measure [8] based on extracted keywords. In summary, the major contributions of this paper are the following: We extract metadata from visual as well as audio resources of lecture videos automatically by applying appropriate analysis techniques. For evaluation purposes we developed several automatic indexing functionalities in a large lecture video portal, which can guide both visually- and text-oriented users to navigate within lecture video. We conducted a user study intended to verify the research hypothesis and to investigate the usability and the effectiveness of proposed video indexing features. For visual analysis, we propose a new method for slide video segmentation and apply video OCR to gather text metadata. Furthermore, lecture outline is extracted from OCR transcripts by using stroke Fig. 1. (a) An example of outdated lecture video format.(b) An exemplary lecture video. Video 1 shows the professor giving his lecture,

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

whereas his presentation is played in video 2. 1. Tele-TASK system was initially designed in 2002 at the University of Trier. Today, weekly 2000 people (unique visits) around the world visit the tele-TASK lecture video portal (www.tele-task.de) with more than 4,800 lectures and 14,000 podcasts free of charge via internet.

Content-Based Image Retrieval (CBIR), a technique which uses visual contents to search images from large scale image databases according to users' interests, has been an active and fast advancing research area since the 1990s. During the past decade, remarkable progress has been made in both theoretical research and system

Linear SVM

Given some training data $\mathcal{D}$, a set of n points of the form

$$\mathcal{D} = \left\{ (\mathbf{x}_i, y_i) \mid \mathbf{x}_i \in \mathbb{R}^p,\, y_i \in \{-1,1\} \right\}_{i=1}^{n}$$

where the y_i is either 1 or −1, indicating the class to which the point $\mathbf{x}_i$belongs. Each $\mathbf{x}_i$is a p-dimensional real vector. We want to find the maximum-margin hyperplane that divides the points having $y_i = 1$from those having $y_i = -1$. Any hyperplane can be written as the set of points $\mathbf{x}$ satisfying

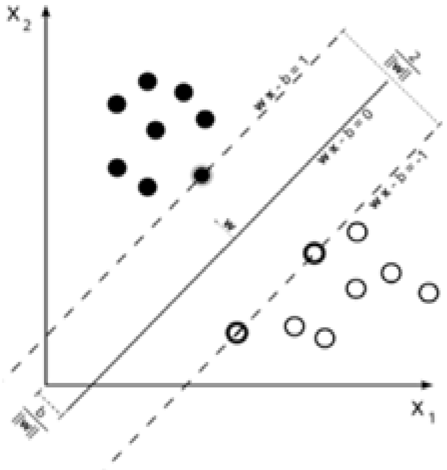


Figure 1.Linear SVM

Maximum-margin hyper plane and margins for an SVM trained with samples from two classes. Samples on the margin are called the support vectors.

$$\mathbf{w} \cdot \mathbf{x} - b = 0,$$

Where . denotes the dot product and $\mathbf{w}$the (not necessarily normalized) normal vector to the hyperplane. The parameter $\tfrac{b}{\|\mathbf{w}\|}$determines the offset of the hyperplane from the origin along the normal vector $\mathbf{w}$.

If the training data are linearly separable, we can select two hyperplanes in a way that they separate the data and there are no points between them, and then try to maximize their distance. The region bounded by them is called "the margin". These hyperplanes can be described by the equations

$$\mathbf{w} \cdot \mathbf{x} - b = 1$$

and

$$\mathbf{w} \cdot \mathbf{x} - b = -1.$$

By using geometry, we find the distance between these two hyperplanesis $\tfrac{2}{\|\mathbf{w}\|}$, so we want to minimize $\|\mathbf{w}\|$. As we also have to prevent data points from falling into the margin, we add the following constraint: for each $i$either

$$\mathbf{w} \cdot \mathbf{x}_i - b \geq 1 \qquad \text{for } \mathbf{x}_i \text{ of the first class}$$

or

$$\mathbf{w} \cdot \mathbf{x}_i - b \leq -1 \qquad \text{for } \mathbf{x}_i \text{ of the second.}$$

This can be rewritten as:

$$y_i(\mathbf{w} \cdot \mathbf{x}_i - b) \geq 1, \quad \text{for all } 1 \leq i \leq n. \qquad (1)$$

We can put this together to get the optimization problem:

Minimize (in $\mathbf{w}, b$)

$$\|\mathbf{w}\|$$

subject to (for any $i = 1, \dots, n$)

$$y_i(\mathbf{w} \cdot \mathbf{x_i} - b) \geq 1.$$

Primal form

The optimization problem presented in the preceding section is difficult to solve because it depends on ||**w**||, the norm of **w**, which involves a square root. Fortunately it is possible to alter the equation by substituting ||**w**|| with $\frac{1}{2}\|\mathbf{w}\|^2$(the factor of 1/2 being used for mathematical convenience) without changing the solution (the minimum of the original and the modified equation have the same **w** and b). This is a programming optimization problem. More clearly:

$$\arg\min_{(\mathbf{w},b)} \frac{1}{2}\|\mathbf{w}\|^2$$

subject to (for any $i = 1, \dots, n$)

$$y_i(\mathbf{w} \cdot \mathbf{x_i} - b) \geq 1.$$

By introducing Lagrange multipliers $\boldsymbol{\alpha}$, the previous constrained problem can be expressed as

$$\arg\min_{\mathbf{w},b} \max_{\boldsymbol{\alpha}\geq 0} \left\{ \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n} \alpha_i [y_i(\mathbf{w} \cdot \mathbf{x_i} - b) - 1] \right\}$$

that is we look for a saddle point. In doing so all the points which can be separated as $y_i(\mathbf{w} \cdot \mathbf{x_i} - b) - 1 > 0$do not matter since we must set the corresponding $\alpha_i$to zero.

This problem can now be solved by standard quadratic programming techniques and programs. The "stationary" Karush–Kuhn–Tucker condition implies that the solution can be expressed as a linear combination of the training vectors

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x_i}.$$

Only a few $\alpha_i$will be greater than zero. The corresponding $\mathbf{x_i}$are exactly the support vectors, which lie on the margin and satisfy $y_i(\mathbf{w} \cdot \mathbf{x_i} - b) = 1$. From this one can derive that the support vectors also satisfy

$$\mathbf{w} \cdot \mathbf{x_i} - b = 1/y_i = y_i \iff b = \mathbf{w} \cdot \mathbf{x_i} - y_i$$

which allows one to define the offset b. In practice, it is more robust to average over all N_{SV}support vectors:

$$b = \frac{1}{N_{SV}} \sum_{i=1}^{N_{SV}} (\mathbf{w} \cdot \mathbf{x_i} - y_i)$$

Dual form

Writing the classification rule in its unconstrained dual form reveals that the maximum-margin hyperplane and therefore the classification task is only a function of the support vectors, the subset of the training data that lie on the margin.

Using the fact that $\|\mathbf{w}\|^2 = \mathbf{w} \cdot \mathbf{w}$and substituting $\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x_i}$, one can show that the dual of the SVM reduces to the following optimization problem:

Maximize (in α_i)

$$\tilde{L}(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j)$$

subject to (for any $i = 1, \dots, n$)

$$\alpha_i \geq 0,$$

and to the constraint from the minimization in b

$$\sum_{i=1}^{n} \alpha_i y_i = 0.$$

Here the kernel is defined by $k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i \cdot \mathbf{x}_j$

W can be computed thanks to the α terms:

$$\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i.$$

Biased and unbiased hyperplanes

For simplicity reasons, sometimes it is required that the hyperplane pass through the origin of the coordinate system. Such hyperplanes are called unbiased, whereas general hyperplanes not necessarily passing through the origin are called biased. An unbiased hyperplane can be enforced by setting $b = 0$ in the primal optimization problem. The corresponding dual is identical to the dual given above without the equality constraint

$$\sum_{i=1}^{n} \alpha_i y_i = 0$$

RESULT AND DISCUSSION

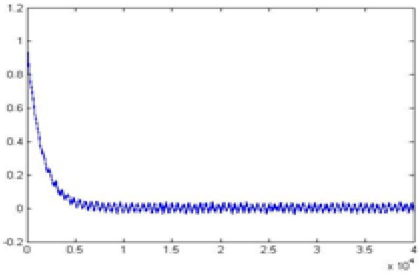
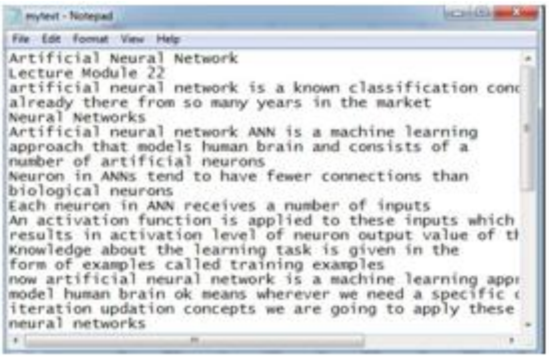


Figure 2. Output wave form



```
Artificial Neural Network
Lecture Module 22
artificial neural network is a known classification conc
already there from so many years in the market
Neural Networks
Artificial neural network ANN is a machine learning
approach that models human brain and consists of a
number of artificial neurons
Neuron in ANNs tend to have fewer connections than
biological neurons
Each neuron in ANN receives a number of inputs
An activation function is applied to these inputs which
results in activation level of neuron output value of tl
Knowledge about the learning task is given in the
form of examples called training examples
now artificial neural network is a machine learning appr
model human brain ok means wherever we need a specific c
iteration updation concepts we are going to apply these
neural networks
```

Figure 3.Data in Notepad

CONCLUSION AND SCOPE OF WORK

In this paper, we presented an approach for content-based lecture video indexing and retrieval in large lecture video archives. In order to verify the research hypothesis we apply visual as well as audio resource of lecture videos for extracting content-based metadata automatically. Several novel indexing features have been developed in a large lecture video portal by using those metadata and a user study has been conducted.

As the future work, the usability and utility study for the video search function in our lecture video portal will be conducted. 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.

REFERENCES

1. E. Leeuwis, M. Federico, and M. Cettolo, “Language modelling and transcription of the ted corpus lectures,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., 2003, pp. 232–235.

2. D. Lee and G. G. Lee, “A korean spoken document retrieval system for lecture search,” in Proc. ACM Special Interest Group Inf. RetrievalSearching Spontaneous Conversational Speech Workshop, 2008.

3. J. Glass, T. J. Hazen, L. Hetherington, and C. Wang, “Analysis and processing of lecture audio data: Preliminary investigations,” in Proc. HLT-NAACL Workshop Interdisciplinary Approaches Speech Indexing Retrieval, 2004, pp. 9–12.

4. A. Haubold and J. R. Kender, “Augmented segmentation and visualization for presentation videos,” in Proc. 13th Annu. ACM Int. Conf. Multimedia, 2005, pp. 51– 60.

5. W. H€urst, T. Kreuzer, and M. Wiesenh€utter, “A qualitative study towards using large vocabulary automatic speech recognition toindex recorded

presentations for search and access over the web," in Proc. IADIS Int. Conf. WWW/Internet, 2002, pp. 135–143.

6. C. Munteanu, G. Penn, R. Baecker, and Y. C. Zhang, "Automatic speech recognition for webcasts: How good is good enough andwhat to do when it isn't," in Proc. 8th Int. Conf. Multimodal Interfaces,2006.

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. G. Salton, A. Wong, and C. S. Yang. (Nov. 1975). A vector space model for automatic indexing, Commun. ACM, 18(11),pp. 613–620, [Online]. Available: http://doi.acm.org/10.1145/361219.361220

9. T.-C. Pong, F. Wang, and C.-W. Ngo, "Structuring lowquality videotaped lectures for cross-reference browsing by video textanalysis," J. Pattern Recog., vol. 41, no. 10, pp. 3257–3269, 2008.

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, 2009, pp. 730–733.

11. T. Tuna, J. Subhlok, L. Barker, V. Varghese, O. Johnson, and S. Shah. (2012), "Development and evaluation of indexed captioned searchable videos for stem coursework," in Proc. 43rd ACM Tech. Symp. Comput. Sci. Educ., pp. 129–134. [Online]. Available: http://doi.acm.org/10.1145/2157136.2157177.

12. H. J. Jeong, T.-E. Kim, and M. H. Kim.(2012), "An accurate lecture video segmentation method by using sift and adaptive threshold,"in Proc. 10th Int. Conf. Advances Mobile Comput., pp. 285–288.[Online]. Available: http://doi.acm.org/10.1145/2428955.2429011.

13. H. Sack and J. Waitelonis, "Integrating social tagging and document annotation for content-based search in multimedia data," inProc. 1st Semantic Authoring Annotation Workshop, 2006.

14. 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.

15. S. Repp, A. Gross, and C. Meinel, "Browsing within lecture videos based on the chain index of speech transcription," IEEE Trans.Learn. Technol., vol. 1, no. 3, pp. 145–156, Jul. 2008.

16. J. Eisenstein, R. Barzilay, and R. Davis. (2007). "Turning lectures into comic books using linguistically salient gestures," in Proc. 22nd Nat. Conf. Artif. Intell., 1, pp. 877–882. [Online]. Available: http://dl.acm.org/citation.cfm?id=1619645.1619786

17. J. Adcock, M. Cooper, L. Denoue, and H. Pirsiavash, "Talkminer: A lecture webcast search engine," in Proc. ACM Int. Conf. Multimedia, 2010, pp. 241–250.