THE FAU VIDEO LECTURE BROWSER SYSTEM

K. Riedhammer, M. Gropp, E. Nöth

Pattern Recognition Lab, University of Erlangen-Nuremberg, GERMANY
korbinian.riedhammer@informatik.uni-erlangen.de

ABSTRACT
A growing number of universities and other educational institutions provide recordings of lectures and seminars as an additional resource to the students. In contrast to educational films that are scripted, directed and often shot by film professionals, these plain recordings are typically not post-processed in an editorial sense. Thus, the videos often contain longer periods of inactivity or silence, unnecessary repetitions, or corrections of prior mistakes. This paper describes the FAU Video Lecture Browser system, a web-based platform for the interactive assessment of video lectures, that helps to close the gap between a plain recording and a useful e-learning resource by displaying automatically extracted and ranked key phrases on an augmented time line based on stream graphs. In a pilot study, users of the interface were able to complete a topic localization task about 29% faster than users provided with the video only while achieving about the same accuracy. The user interactions can be logged on the server to collect data to evaluate the quality of the phrases and rankings, and to train systems that produce customized phrase rankings.

Index Terms— automatic speech recognition, key phrase extraction, key phrase ranking, visualization, user interaction

1. INTRODUCTION
A growing number of universities and other educational institutions provide recordings of regularly scheduled lectures and seminars as an additional resource to the students. In contrast to educational films that are scripted, directed and often shot by film professionals, these plain recordings are typically not post-processed in an editorial sense. That is, the videos often contain longer periods of inactivity or silence, unnecessary repetitions, or even corrections of prior false statements or mistakes. Furthermore, they usually lack a summary or table of contents which would be highly valuable: the videos have a typical duration of 45 or 90 (double period) minutes but the viewer might only be interested in segments that cover certain topics. For example, consider a student preparing for an exam watching a lecture recording. Typical questions include: What was the lecture about in general, i.e., is it worth watching? When was a specific topic covered? What issues are related to certain topics? Watching the whole video will answer these questions but might end in frustration if there was nothing of interest to the viewer. On the other hand, when “skimming” the video by sampling the topics at more or less random positions the student is at risk missing relevant sections.

For textbooks, table of contents and indices are taken for granted, so why not also have these for video lectures? This paper describes the FAU Video Lecture Browser system, a web-based platform for the interactive assessment of video lectures that is intended to assist the student in getting an overview of the lecture and finding the relevant sections. The user is presented with an automatically extracted and ranked list of key phrases. Beside a simple phrase occurrence indicator that can be used for a single selected phrase, the user can simultaneously visualize a selection of phrases using an augmented time line that shows which phrase appears at what time and how dominant the phrase is at the given time.

The presented video lecture browser is distinct from other interfaces that are publicly accessible. Most platforms focus on the acquisition, organization, storage, and distribution of the videos as for example the OpenCast Matterhorn\(^1\) platform, the commercial Apple iTunes U\(^2\), or the HPI tele-Teaching Solution Kit (tele-TASK)\(^3\). The MIT OpenCourseWare (OCW)\(^4\) is an e-learning platform with extensive possibilities that joins syllabus, lecture notes, assignments, videos and further material; on top of the OCW video archive, the MIT CSAIL developed prototype of an interactive browser that allows the user to search within an automatic transcript. The Czech start-up superlectures\(^5\) developed a video browser that shows the presentation slides, automatic transcript, and a rudimentary table of contents based on slide transitions. The NTU Virtual Instructor\(^6\) features (unranked) key terms, audio-visual extractive summarization, and a surrogate table of contents based on the slides and the transition times. Despite the numerous features, the design and usability is however less sophisticated and intuitive than for example superlectures. The FAU Video Lecture Browser is distinct from the previously listed interfaces by listing ranked key phrases and, for a user-defined selection, shows their occurrence and dominance on an augmented time line. Furthermore, the user can interact with the interface by adding, deleting or re-ranking phrases, and by selecting or deselecting phrases for the visualization. The interactions are logged on a server to establish a database for evaluation and training of user-dependent ranking algorithms.

The organization of this paper follows the processing pipeline of the FAU Video Lecture Browser. Sec. 2 describes a newly acquired corpus of academic spoken English that is used to evaluate each step of the processing pipeline and the final interface. On the server side, the audio track of the video is processed by an automatic speech recognition system which is described in Sec. 3. Sec. 4 details the key phrase extraction from the automatic transcriptions and the subsequent unsupervised ranking. Sec. 5 describes the visualization of the key phrases and the possible interactions on the client side. The implementation details of both server and client are listed in Sec. 6. Sec. 7 describes a pilot study that shows the use and usability of the presented interface. The paper concludes with a summary and outlook on future work in Sec. 8.

The authors would like to emphasize that this paper is intended as a comprehensive system description that subsumes prior work on

\(^{1}\)http://opencast.org/matterhorn
\(^{2}\)http://www.apple.com/education/itunes-u
\(^{3}\)http://www.tele-task.de
\(^{4}\)http://ocw.mit.edu
\(^{5}\)http://www.superlectures.com
\(^{6}\)http://speech.ee.ntu.edu/RA/lecture
the acoustic model [1], key phrase extraction and visualization [2, 3]. The key contribution is that the validity of the methods on the one hand, and the use and usability of the interface on the other hand, are all evaluated on the same corpus of academic spoken English, the LME Lectures [4].

2. DATA

The LME Lectures corpus of academic spoken English [4] consists of recordings of two lecture series read in the summer term 2009 at the University of Erlangen-Nuremberg. The graduate level computer science lectures cover topics in pattern analysis (series PA, 18 recordings) and interventional medical image processing (series IMIP, 18 recordings). In contrast to other data sets such as the corpus of British Academic Spoken English (BASE) or Michigan Corpus of Academic Spoken English (MICASE) [5], the LME Lectures feature a single speaker that was recorded in a constant environment with the same chip-on close-talking microphone. In total, the corpus consists of about 39.5 hours of high-definition audio and video (about a single speaker that was recorded in a constant environment with the same chip-on close-talking microphone. In total, the corpus consists of about 39.5 hours of high-definition audio and video (about 29 hours of speech). The test (PA [06,08], IMIP [05,07]) and development set (PA [15,17], IMIP [13,17]) consist of about 1 800 turns (2 hours) each, the remaining 24 hours are used for training; the out-of-vocabulary rate is about 1 % for both development and test set. The manual transcription of the roughly 300 500 words results in a vocabulary of 5 838 words (excluding foreign words and word fragments). The individual lecture PA06 was further annotated with a ranked list of key phrases by five human subjects that have either attended the lecture or a similar lecture in a different term. In addition to the ranking, the annotators assigned grades to the phrases from 1 (“very relevant”) to 6 (“useless”).

3. AUTOMATIC SPEECH RECOGNITION


At the acoustic front-end, the first 13 MFCC coefficients are extracted using a frame size of 25 ms and a frame shift of 10 ms. For each frame, the neighboring nine consecutive MFCC vectors are concatenated and the dimension reduced to 40 using a HMM state alignment based LDA [9] and subsequent MLLT [10].

Two acoustic models are trained that mainly differ in the number of parameters and computational complexity: a two-level tree based multi-codebook semi-continuous system (2lvs) [1] using 1 000 codebooks, 10 000 Gaussians, and 7 000 distinct acoustic states, and a subspace Gaussian mixture model (sgmm) [11] using a codebook of 600 Gaussians, 25 000 sub-states, and 5 500 states.

The language model is derived from a large tri-gram model for the lecture domain covering about 60 000 words and about 11 000 000 bi- and trigrams each [12] that was reduced to the 5 838 words actually appearing in the lexicon. The pronunciations are taken from the CMU DICT, missing entries were generated using the RWTH Aachen g2p graphem-to-phoneme tool [13], also trained on the CMU dictionary.

The systems achieve a word error rate (WER) of 12.75 % using the 2lvs, and 11.03 % using the more complex sgmm model. The WER is competitive given the spontaneous nature of the data and the fact that the speaker is a non-native speaker of English. The majority of the errors is due to mix ups of singular and plural words, as well as to foreign words and word fragments. Thus, the automatic transcripts provide a solid base for the later key phrase extraction and ranking. Further details on the experimental setup including parameter settings and training procedure can be found in [14].

4. KEY PHRASE EXTRACTION AND RANKING

The transcripts are pre-processed to remove disfluencies, to identify stopwords, and to apply Porter-based stemming as a basic means of lemmatization. The key phrase candidates are extracted using a regular expression that matches the part-of-speech (POS) tags found using the CRFTagger [15]. The idea is to extract noun phrases based on the POS tags associated with their components:

\[ 2^{+} 1^{+} (3^{+} 2^{+} 1^{+})^{*} \]

where the digits represent the components of a basic noun phrase: proper nouns, foreign words, numbers and gerunds form the main component (1), adjectives, comparatives, superlatives and past participles (2) that may be placed before the main component. Finally, a basic noun phrase may be followed by other noun phrases after a determiner or preposition (3).

The extracted candidates are ranked based on a heuristic that combines the phrase frequency and length. The weight \( w(\varphi_i) \) is defined by

\[
w(\varphi_i) = \begin{cases} f_1 & n_1 = 1 \\ f_1 \cdot (n_1 + 1) & n_1 > 1 \end{cases}
\]

where \( f_1 \) is the raw frequency of phrase \( \varphi_i \). The linear re-weighting puts emphasis on longer phrases which is motivated from a teaching point of view: if a certain longer phrase is deliberately repeated, it may be a salient phrase. Similarly, if a long phrase occurs the same number of times as a short phrase, it is likely to be more salient than the shorter one.

For the lecture PA06, the automatic ranking is compared to the human rankings (see Sec. 2) using average precision (AP) [16] that evaluates how many relevant phrases are paced on the top of a list. Using binary relevance judgments \( \chi(\varphi) = \{0, 1\} \) of a phrase \( \varphi \), the precision at the position \( n \) of a key phrase list can be defined as the fraction of relevant phrases in the top \( n \) items of that list as

\[
p_n = \frac{\sum_{i=1}^{n} \chi(\varphi_i)}{n}.
\]

Considering the \( N \) top items of the candidate list, the Average Precision is computed as

\[
AP(N) = \frac{\sum_{n=1}^{N} p_n \chi(\varphi_n)}{\sum_{n=1}^{N} \chi(\varphi_n)}.
\]

The annotated grade interval (1 to 6) is split in half, assigning \( \chi(\varphi) = 1 \) if phrase \( \varphi \) was rated 1 to 3, and \( \chi(\varphi) = 0 \) otherwise. As with the regular precision, AP values range from 0 to 1, with 1 being the best possible value. The parameter \( N \) that controls how many items of the ranking are actually considered needs to be set depending on the target application. In case of multiple reference rankings, a single value can be computed by averaging the AP values of each pair-wise combination.

Fig. 1 shows the average precision of the automatic ranking compared to the average human performance. For rankings up to length eight, both human and automatic rankings are of similar quality, suggesting that the first five to ten phrases can be reliably detected and ranked. For longer lists, the automatic rankings show an even higher AP value than the average human. This means that the automatic ranking seems to match a somewhat general human ranking, or in other words is more objective than the individual human
The rather basic ranking \( w(\phi_i) \) is chosen to demonstrate the domain independence of the extraction and subsequent ranking. Further ranking strategies that include prior knowledge about the data and domain are discussed in \([2, 14]\).

5. VISUALIZATION AND INTERACTION

The speech recognition, key phrase extraction and ranking are processed off-line on the server once a video is added to the database. The visualization is dynamically rendered on the client’s web browser; an example view is shown in Fig. 2. The video is accompanied by a list of ranked key phrases (far right). The stream graph \([17]\) shows a tube for each selected phrase in the respective color (bottom left). The stream resembles a timeline (from left to right) where the diameter reflects the current dominance. The graph is constructed by stacking the splines that correspond to the number of phrase occurrences within each segments as a measure of dominance.

The user interactions are split in two groups, where the first relates to pure browser functionality and the second relates to data collection for future research. To assess the video, the user can

- alter the list of phrases that are displayed in the stream graph by using drag-and-drop: move a phrase from the right list into the left list for inclusion, and vice versa for exclusion.
- select a single phrase to update the occurrence indicator bar.

Fig. 3 shows an example log file. For all interactions, the time and id are stored together with the interaction parameters. For the stream graph clicks, both actual horizontal click position and the phrase occurrence driven correction are recorded.

A demo of the interface can be found at http://goo.gl/3Nrh1; note that the logging of the interactions is redirected to the JAVASCRIPT console.

6. IMPLEMENTATION

The FAU Video Lecture Browsing system is implemented based on open toolkits and standards. The speech recognition is driven by the KALDI toolkit which is implemented in c++ and licensed under GPL v.3. The key phrase extraction and ranking process is implemented within the Apache-licensed JAVA-based Unstructured Information Management Applications (UIMA) framework\(^8\) that supports the automatic parallelization of independent steps such as POS tagging and lemmatization. The Google Web Toolkit (GWT)\(^9\) is the joint between the server and the client. The GWT provides a framework for (a)synchronous remote procedure calls between the client and the server which are used to log the user interactions. On the client side, it uses HTML5 and CSS3 to provide the look and feel of regular desktop applications including drag and drop.

\(^7\) Also, the human average AP results from comparing each rater against each of the remaining four; the automatic rankings are compared against the five human rankings.

\(^8\) http://uima.apache.org

\(^9\) http://code.google.com/webtoolkit
Fig. 2. Screen shot of the FAU Video Lecture Browser interface showing the video (top left), the list of available (far right) and currently selected (center) key phrases, an occurrence indicator bar for the highlighted phrase (below the video; red dashes), the stream graph showing the occurrence and dominance of the selected phrases (bottom left), and a text field for the textual search for key phrases (bottom center).
Table 1. Task performance of the two groups for detecting the salient segments within the video. \( R, P \) and \( F \) are the recall, precision and F-measure with respect to the reference solution; the average time \( t \) required to solve the task is measured in minutes.

<table>
<thead>
<tr>
<th>group</th>
<th>( R )</th>
<th>( P )</th>
<th>( F )</th>
<th>( t )</th>
</tr>
</thead>
<tbody>
<tr>
<td>control</td>
<td>0.76</td>
<td>0.62</td>
<td>0.68</td>
<td>30</td>
</tr>
<tr>
<td>test</td>
<td>0.68</td>
<td>0.71</td>
<td>0.69</td>
<td>21</td>
</tr>
</tbody>
</table>

HTML5 further supports the native playback of video making the interface independent of third party video players. The stream graph is rendered by an extension to the Processing.js library, namely streamgraph.js\(^9\).

7. EVALUATION

The FAU Video Lecture Browser was evaluated in a pilot user study focusing on lecture PA06. The evaluation is split in two parts, namely an assessment task and a post-use questionnaire. The test subjects are divided on two groups. The control group performs the task with access to the video only, the test group is assigned the same task but uses the browser. Naturally, the post-use questionnaire is only given to members of the test group. A group of ten computer science graduate students were split in groups of five students; each student had either observed the lecture or attended a similar lecture in a different term.

For the assessment task, each subject was provided with a list of three briefly described topics that are covered in the lecture, and is asked to indicate those three minute segments on a time line that correspond to these topics. The task can be solved by watching the whole video and without actually understanding the context, thus the evaluation is unlikely to be skewed by personal differences. Also, a simple key word search, as possible with the browser, will lead to false hits, as phrases occur throughout the lecture but not necessarily imply that the related topic is covered at that moment. The task accuracy is measured with respect to a reference sheet that was assembled by an annotator familiar with the topic, lecture recording, and presentation slides. Additionally, the time required to complete the task is measured.

Tab. 1 shows the accuracy of the control and test group in detecting the segments that cover the three topics, measured in precision, recall and F-measure. Both groups have a very similar performance, however, the test group finishes the task on average 29\% faster using the interface. This indicates that the browser indeed helps to assess the video efficiently as demanded in the introduction of this paper.

The post-use questionnaire contains questions about the users’ impression of the browser that are answered on five point Likert scales that read “strongly agree, agree, neutral, disagree, strongly disagree.” The statements are selected from the USE [18] and PUTQ [19] catalogs, and slightly adapted to match the browser evaluation scenario.

Tab. 2 shows the questions and the corresponding average human ratings after using the interface. Although the users are rather neutral when asked if the browser increased their productivity and would rather not pay for the tool, most of the users find the tool to be easy and intuitive to use. Furthermore, most users like the interface, and find that it would enable them to find important facts quickly and make it easier to work with the video lectures. The key phrases are found to be helpful and accurate, and, together with the visualization, to give a good overview of the lecture.

8. SUMMARY AND OUTLOOK

The experiments and results presented in this paper confirm that the FAU Video Lecture Browser is a useful and intuitive tool to close the gap between ordinary recordings of lectures and edited educational films. The automatic speech recognition, the base of all subsequent analyses, has a WER of about 10\% which is sufficient for a reliable extraction and ranking of key phrases. In fact, the automatic key phrase rankings seem to be more objective than human rankings when compared to a panel of human annotators. The browser allows the user to interact with the visualization by selecting or deselecting certain phrases for display on the stream graph. Furthermore, the user can add, delete and re-rank phrases to give feedback about the quality. The user interactions are logged on the server to establish a database for the evaluation and improvement of both general and user-specific rankings.

Although the presented system is a working prototype, there are a number of aspects that need to be addressed to allow a general use in the context of a department or school. The speech recognition system vocabulary and language model needs to be extended to minimize the out-of-vocabulary events and to ensure a low WER. A possible solution is to extract text from the presentation slides and use textual resources to obtain data from the related domain; the previously used grapheme-to-phoneme converter can be used to synthesize missing pronunciations.

If the speaker and the room are known, the speech recognition system can be adapted to the speaker’s voice and language and to the room acoustics. This is especially promising for the sgmm system that already allows per-speaker transformations.

The user interface needs to be personalized to allow the data collection under a pseudonym. This is required to satisfy privacy concerns while still maintaining the possibility of producing user-specific rankings.

A large-scale evaluation of the video browser within a term and the subsequent exam instead of a task based evaluation will shed some light on how “genuine” students profit from the interface.

The FAU Video Lecture Browser is mainly based on GPL or Apache licensed toolkits which allows a future commercialization.

9. REFERENCES


---

\(^9\)https://github.com/jsundram/streamgraph.js
<table>
<thead>
<tr>
<th>Perceived Usefulness</th>
<th>1</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Using it in my studies would enable me to find important facts quickly.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Using it in my studies would increase my productivity.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Using it in my studies would enhance my effectiveness.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Using it would make it easier to work with the video lectures.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I would find it useful in my studies.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ease of Use</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>It is easy to use.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>It is intuitive to use.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I learned to use it quickly.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Satisfaction</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>I like the interface.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I would recommend it to a friend.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I would pay for it.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Browser Features</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>The key phrases are helpful.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>The key phrases are accurate.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>The phrase visualization is helpful.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>The phrase visualization is accurate.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phrases and visualization give a good overview of the video lecture.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Average human ratings for the questionnaire statements quantizing “strongly agree” through “strongly disagree” as 1 to 5.


