Multi-factor Segmentation for Topic Visualization and Recommendation: the MUST-VIS System

Chidansh Bhatt
Idiap Research Institute
Martigny, Switzerland
cbhatt@idiap.ch

Andrei Popescu-Belis
Idiap Research Institute
Martigny, Switzerland
apbelis@idiap.ch

Maryam Habibi
Idiap and EPFL
Martigny, Switzerland
mhabibi@idiap.ch

Sandy Ingram
Klewel SA
Martigny, Switzerland
sandy.ingram@klewel.com

Stefano Masneri
Heinrich Hertz Institute
Berlin, Germany
stefano.masneri@hhi.fraunhofer.de

Fergus McInnes
University of Edinburgh, UK
fergus.mcinnes@ed.ac.uk

Nikolaos Pappas
Idiap and EPFL
Martigny, Switzerland
npappas@idiap.ch

Oliver Schreer
Heinrich Hertz Institute
Berlin, Germany
oliver.schreer@hhi.fraunhofer.de

ABSTRACT

This paper presents the MUST-VIS system for the MediaMixer/VideoLectures.NET Temporal Segmentation and Annotation Grand Challenge. The system allows users to visualize a lecture as a series of segments represented by keyword clouds, with relations to other similar lectures and segments. Segmentation is performed using a multi-factor algorithm which takes advantage of the audio (through automatic speech recognition and word-based segmentation) and video (through the detection of actions such as writing on the blackboard). The similarity across segments and lectures is computed using a content-based recommendation algorithm. Overall, the graph-based representation of segment similarity appears to be a promising and cost-effective approach to navigating lecture databases.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Retrieval; H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems

Keywords

lecture segmentation, lecture recommendation

1. INTRODUCTION

The MUST-VIS system presented in this paper is our solution to the ACM Multimedia 2013 Grand Challenge on Temporal Segmentation and Annotation proposed by MediaMixer and VideoLectures.NET. MUST-VIS stands for multi-factor segmentation for topic visualization and recommendation, because the system first performs lecture segmentation using audio, text and visual information, and then uses techniques inspired from recommender systems to connect segments. The goal is to improve access to repositories of audio-visual recordings of lectures, through segmentation and recommendation of segments, unlike many existing methods that are limited to entire lectures only.

MUST-VIS introduces a multi-modal algorithm for lecture segmentation based on video and audio/text, and annotates segments using keyword clouds, which offer direct access to the information content, while taking into account the major speaker actions. While these annotations make segments searchable by users, an efficient way to access lecture content is via recommendations of segments and lectures. This technique is demonstrated in MUST-VIS with a prototype interface, as visualization is essential for exploring large amounts of data at various levels of granularity. The MUST-VIS processing algorithms are time-efficient and scale well to large repositories. Moreover, segments and recommendations can all be computed offline rather than at search time. We also sketch evaluation methods for the system.

The paper is organized as follows. In Section 2, we describe how users interact with the MUST-VIS system. In Section 3, we present the system components, and in Section 4 we present some evaluation results.

2. MUST-VIS: THE USERS' VIEW

The users of lecture databases such as VideoLectures.NET, YouTube.com/edu or KhanAcademy.org are faced with the challenge of efficient search and browsing, especially when searching for specific pieces of information, such as a fact, a proof or argument, or a reference. In addition, obtaining the gist of a lecture without entirely watching it is another challenge. We assume here that the priority of end-users is to explore the most semantically-relevant information contained within audio-visual lecture recordings.

The MUST-VIS system presents to its users a graphical user interface (GUI) shown in Fig. 1. The goal is to provide an insight into the content of each of the topical segments of lectures, using keyword clouds, which are magnified when hovering over them with the mouse. The segments of a lec-

MM'13, October 21–25, 2013, Barcelona, Spain.

http://dx.doi.org/10.1145/2502081.2508120.

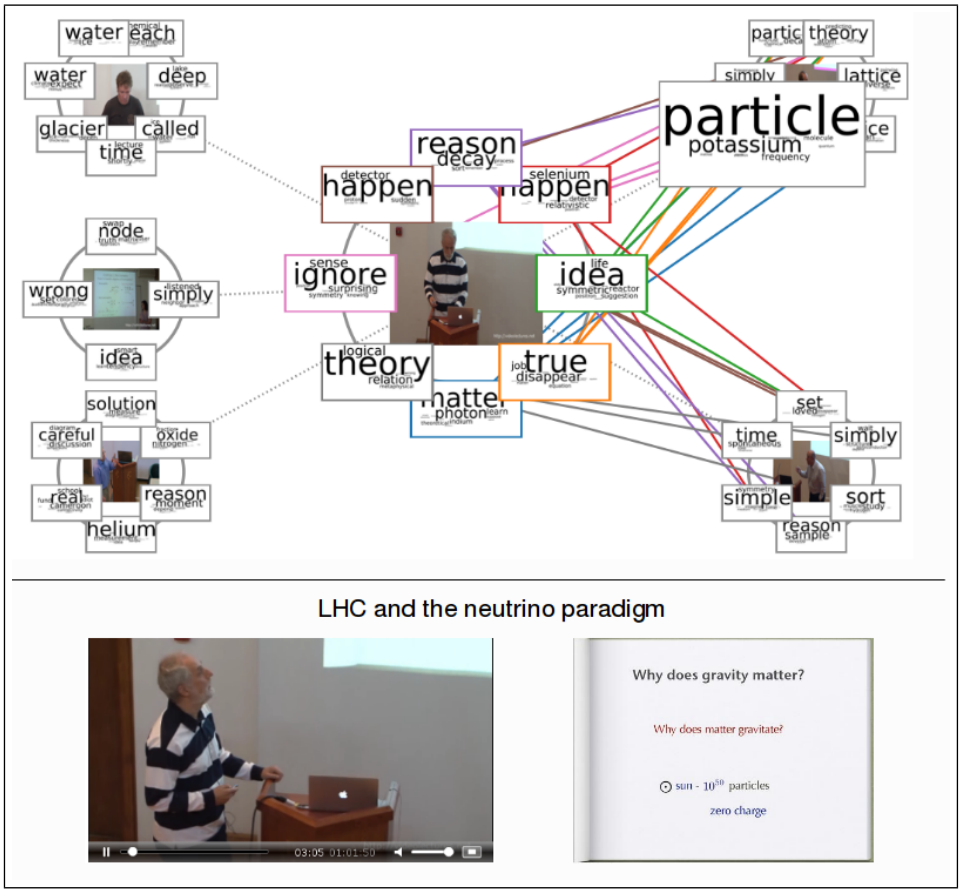

Figure 1: The MUST-VIS navigation graph (top) and video/slide player (bottom). Each lecture is represented with a keyframe and keyword clouds for each segment around it. The lecture in focus (center: LHC and the Neutrino Paradigm) is surrounded by lectures with related segments (e.g., on the right, Ultracold Atoms, and Bioinspired Nanostructured Materials).

ture are ordered by starting time in a clockwise circular manner around the keyframe provided in the lecture repository. The lecture featured at the center of the screen is considered to be in focus and can be played (audio/video plus slides) by clicking on its center or on a segment.

Using content-based recommendation techniques, each segment and each lecture are related to the most similar ones, forming a *navigation graph*, which for simplicity is limited to the five most related lectures. We assume for now that the lecture from which navigation starts is found either directly (by URL) or through keyword-based search. Depending on user preferences, the GUI displays either segment-to-segment similarity links, or lecture-to-lecture ones, or it can zoom from one type to the other. In Fig. 1, the two most similar lectures to the one in focus are shown with segment-to-segment links, while three additional lectures are shown only with lecture-to-lecture links. Each of the segments of the lecture in focus and its links have a unique color code, to facilitate tracing segment-to-segment links. A mouse click on one of the recommended lectures or segments brings it in focus (allowing the user to play it) and redraws the graph to show its own set of recommendations.

3. COMPONENTS OF THE SYSTEM

To compute the navigation graph, MUST-VIS makes use of state-of-the-art multimodal processing of audio, video and text. The fully-automatic components, shown in Fig. 2, are run offline for temporal segmentation and keyword extraction from lectures. Similarity across lectures and segments is then computed using a state-of-the-art content-based recommendation algorithm.

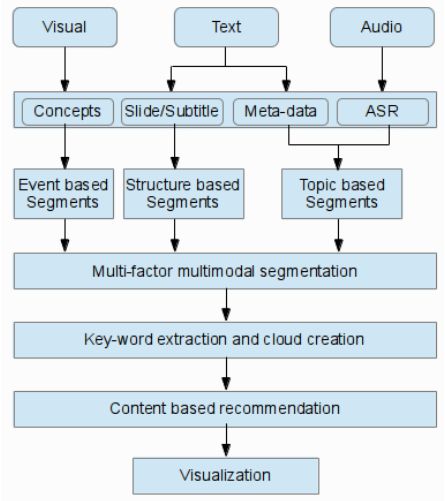

Figure 2: Lecture processing in MUST-VIS.

3.1 Video Processing for Segmentation

We implemented a classical frame-based temporal segmentation algorithm [9] with four possible classes: 'talk' (video representing the speaker), 'slide presentation' (only the slide show is visible in the video), 'mixed' (both speaker and screen are visible, including when the camera is moving from one to the other), and 'blackboard' (speaker writing on the blackboard). In MUST-VIS, we use the segmentation based on temporal boundaries derived from these four semantic-level action detection in the video, rather than traditional shot boundary detection. This is because shot boundary does not provide the same semantic information as the four classes above, and is not applicable to lectures that are recorded with a single shot.

An SVM classifier was trained using 51 features: the number of detected faces, the width of the largest face, its horizontal position, and the 48 values of 16-bin histograms on the three color channels. A training dataset was created manually by labeling one frame every second in the first ten minutes of each video. Classification is performed for each frame, but to avoid over-segmentation, we merge every 50 frames and choose the majority class. Two lectures from the 20 provided by the Grand Challenge organizers were annotated as test data ('geanakoplos_ lec18' and 'ekaykin_drilling'). The classification accuracy of our method was 79% for the first one and 87% for the second one, over respectively 432K and 323K frames. Two sources of error have been identified: the first and most important is the classification of 'talk' frames as 'mixed', which may happen if the detected face is small. The second one is due to small shifts in starting times, appearing when merging frames; being lower than 0.5 s, this does not impact MUST-VIS.

3.2 Audio Processing and Speech Recognition

Speech/non-speech segmentation and speaker diarization were performed using components from the AMIDA system [7], incorporating the programs 'shout_segment' and 'shout_ cluster' from the SHoUT toolkit [8]. Speech recognition was performed using a system [1] trained primarily over TED talks as used for the IWSLT 2012 ASR evaluation. The system has two passes of decoding, both using hybrid models in which HMM observation probabilities are computed using a deep neural network. The second pass incorporates speaker adaptation through a CMLLR transform. Both passes used a trigram language model (LM), and the final transcriptions were obtained by word lattice re-scoring with a 4-gram LM. The LMs were derived by interpolating in-domain models trained on TED talk transcripts with multiple out-of-domain models trained on Europarl, News Crawl and News Commentary data and the LM from the AMIDA system [7]. The performance of the system, as shown in [1] (Table 4, fifth line of results), is around 18% word error rate.

3.3 Text-Based Segmentation and Keywords

The automatic transcripts or the subtitles of each lecture of the dataset are topically segmented using the TextTiling algorithm implemented in the NLTK toolkit [2]. Then the words in the each topic segment are ranked using a recent diverse keyword extraction technique [6], which selects keywords so that they cover the maximum number of topics mentioned in each segment. Finally, word cloud representations are generated using WordCram (`www.wordcram.org`) for each segment and also for entire lectures. Words ranked higher become graphically emphasized in the word cloud.

3.4 Multimodal Segmentation

State-of-the-art video segment detection methods use multiple modalities only infrequently [4]. When cross-modal alignment is performed, it is generally at a lower-level of granularity than needed here. In MUST-VIS, the three available types of segmentation information (from words, video and slides) are combined using a novel segmentation method that is inspired from multimodal alignment. Let $\mathcal{V}$ be the set of 20 lectures provided for the Grand Challenge. One subset, noted $V_{slide} = \{V_1, \ldots, V_{10}\}$, is accompanied by slides, while the other, $V_{subtitles} = \{V_{11}, \ldots, V_{20}\}$, is accompanied by subtitles.

For a given video $V \in \mathcal{V}$, ASR-based segments are noted $S_{ASR} = \{SA_1, \ldots, SA_I\}$ where I depends on V. For a given video $V \in V_{slide}$, segments with text from each slide are noted $S_{slide} = \{SS_1, \ldots, SS_J\}$, where J depends on V. And, for a given video $V \in V_{subtitles}$, the text from the subtitles is mapped onto the temporal boundaries computed by video segmentation and noted $S_{subtitle+visual} = \{SV_1, \ldots, SV_K\}$ (K depends on V). In practice, $I < J$ and $I < K$. For each segment S, S^{start} is its start time, S^{end} its ending time, and S^{text} is the text contained within it.

The multimodal segmentation algorithm has parameters σ, ρ and Δt. Here, σ (initialized to 1) represents the starting index and similarly ρ (initialized to 1) represents the ending index of the segment in S_{slide} that matches the segment in S_{ASR}. The number of segments considered for computing the cosine similarity around the segment index $\tau \in S_{slide}$ is Δt. In practice, $\Delta t = 3$ for V_{slides} and $\Delta t = 5$ for $V_{subtitle+visual}$. The segments of S_{ASR} are considered as principal, due to size and availability.

Input: $S_{ASR}, S_{slide}, \Delta t$ Output: SAS
$\sigma, \rho \leftarrow 1$
for each segment SA_i in S_{ASR} **do**
  $\tau \leftarrow \arg\min_{j \in [\rho, J]} dist(SA_i^{end}, SS_j^{end})$
  $\rho \leftarrow \arg\max_{x \in [\tau - \Delta t, \tau + \Delta t]} sim(SA_i^{text}, SS_{[\sigma:x]}^{text})$
  $SAS_i^{text} \leftarrow SA_i^{text} \bigcup SS_{[\sigma:\rho]}^{text}$
  $SAS_i^{start} \leftarrow SS_\sigma^{start}, SAS_i^{end} \leftarrow SS_\rho^{end}$
  $\sigma \leftarrow \rho + 1$
end for

For each SA_i^{end}, the algorithm first identifies the segment index τ of the nearest segment SS_j^{end}. As the individual SA_i is usually larger, it aligns with multiple SS_j. The segment alignment boundary is selected after identifying the index ρ of the segment in SS_j which has the maximum cosine similarity between the combined text content of all the segments from $SS_{[\sigma:\rho]}$ and the text content of SA_i^{text}. The value of ρ is constrained to be within $[\tau - \Delta t, \tau + \Delta t]$ to avoid the situation where a smaller number of segments from S_{slide} is aligned to S_{ASR}, leaving behind many unallocated segments from S_{slide}. All the segments SA_i are thus aligned by order of start time to ordered groups of segments from SS_j.

Between 4 and 10 segments are found for each of the lectures in the dataset, a value that was aimed for so that segments are easily grasped through the GUI (with equal sizes for now). Depending on user requirements, finer-grained segments can be easily obtained.

3.5 Recommending Lectures and Segments

While keyword-based search in annotated lectures and segments is now well-understood, we propose here to recommend to viewers new segments and lectures, as an alternative to search which improves the accessibility of a lecture database. We use techniques from content-based rec-

ommender systems, and compute similarity between items based on their content descriptors, through standard vector space models based on TF-IDF weighting, which have been shown to perform well for multimedia recommendation [10]. In addition to the words from the above-mentioned features, we used all available meta-data such as titles, speaker names, descriptions, subtitles, and slide titles with start time. We thus generated for each segment a ranked list of the most similar other segments.

An additional goal is to generate segment-to-segment recommendations which, when integrated over all segments of a lecture, are coherent with the lecture-level recommendations obtained directly. To achieve this, we generated from the segment-to-segment similarity matrix a summation matrix considering all possible pairings of segments from two talks, which was normalized. Based on these scores, we generate the top recommendations for each lecture, then select only the best segment-to-segment links from them. In Section 4, we show experiments with various selection cut-offs to increase the coherence between the two levels of recommendations.

3.6 Visualization: GUI

The GUI presented in Section 2 was implemented using the D3 JavaScript library for manipulating documents [3], allowing interactive visualization and exploration using force-directed graphs. This allowed us to position the nodes of the graph in a two-dimensional space so that all the edges are of equal length and there are as few crossing edges as possible, by assigning forces among the set of edges and the set of nodes, and minimizing the potential energy.

4. RESULTS AND DISCUSSION

As no ground-truth annotations are available for the Grand Challenge dataset, and evaluation in use is a costly alternative, we provide the following quality indicators for the MUST-VIS system. The use of all multimodal features for recommendation, in addition to standard metadata, increases the vocabulary by about 35% and the average similarity scores for segments by about 0.1. This is unlike recent observations that multimodal features decreased performance on a hyperlinking task [5]. In our case, the average increase in similarity scores for segments with slides is higher than for those with subtitles, indicating that slides are (predictably) a more useful complement to ASR than subtitles.

In a pilot experiment, two of the authors have assigned ground-truth recommendations to each of the 20 lectures in the dataset. Table 1 compares the ground-truth rankings versus automated methods in several configurations, using Mean Reciprocal Rank (MRR) and Mean Average Precision (MAP). The highest agreement is between the two annotators, followed very closely by the agreement between lecture-to-lecture (LL) and averaged segment-to-segment similarities (SS). Therefore, the possibility of "smooth zooming" between LL and SS recommendations is validated. The lecture-level recommendation is slightly closer to the ground-truth than the averaged segment-level one, though both are at some distance with respect to inter-coder agreement. Such distance can be reduced with the help of hybrid recommendation systems, using collaborative filtering based on user-log information, when available, along with the proposed content-based method. The random recommendation performs poorly compared to other methods.

Table 1: Comparison of ground-truth recommendations (A1, A2) with automated ones: lecture-to-lecture (LL), segment-to-segment (SS) and random (R, as baseline, 500 draws), using MAP (a) and MRR (b) over 1–5 top recommendations.

	A1		**A2**		**LL**		**SS**		**R**	
	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)
A1	1.0	1.0	.91	.54	.33	.14	.28	.10	.13	.03
A2	-	-	1.0	1.0	.31	.15	.22	.10	.10	.02
LL	-	-	-	-	1.0	1.0	.91	.52	.29	.11
SS	-	-	-	-	-	-	1.0	1.0	.30	.11
R	-	-	-	-	-	-	-	-	1.0	1.0

5. CONCLUSION

The MUST-VIS system performs segmentation and annotation of lectures based on features from several modalities, and displays the results in a novel GUI, which enables navigation based on recommended lectures and segments. With respect to current recommender systems, MUST-VIS offers clearer justifications of its recommendations, through the number of links and the keyword clouds, and facilitates access to relevant parts within lectures. Both features should thus help answering the Grand Challenge on Temporal Segmentation and Annotation of Lectures.

6. ACKNOWLEDGMENTS

This work was supported by the Swiss National Science Foundation through the AROLES project n. 51NF40-144627 and by the European Union through the inEvent project FP7-ICT n. 287872 (`www.inevent-project.eu`).

7. REFERENCES

[1] P. Bell, P. Swietojanski, and S. Renals. Multi-level adaptive networks in tandem and hybrid ASR systems. In *Proc. ICASSP*, 2013.

[2] S. Bird. NLTK: the natural language toolkit. In *Proc. COLING-ACL*, pages 69–72, 2006.

[3] M. Bostock and al. D3 data-driven documents. *IEEE Tran. Vis. Comput. Gr.*, 17(12):2301–2309, 2011.

[4] M. Del Fabro and L. Boszormenyi. State-of-the-art and future challenges in video scene detection: a survey. *Multimedia Systems*, pages 1–28, 2013.

[5] M. Eskevich and al. Multimedia information seeking through search and hyperlinking. In *Proc. ICMR*, pages 287–294, 2013.

[6] M. Habibi and A. Popescu-Belis. Diverse keyword extraction from conversations. In *Proc. ACL*, 2013.

[7] T. Hain and al. Transcribing meetings with the AMIDA systems. *IEEE Tran. ASL*, 20(2), 2012.

[8] M. Huijbregts. *Segmentation, Diarization and Speech Transcription: Surprise Data Unraveled.* PhD thesis, Centre for Telematics and Information Technology, University of Twente, Enschede, Nov. 2008.

[9] I. Koprinska and al. Temporal video segmentation: A survey. *Sig. Proc.: Image Comm.*, 16(5), 2001.

[10] N. Pappas and A. Popescu-Belis. Combining content with user preferences for TED lecture recommendation. In *Proc. CBMI*, 2013.