Easy Navigation through Instructional Videos using Automatically Generated Table of Content

Ankit Gandhi, Arijit Biswas, Kundan Shrivastava, Ranjeet Kumar, Sahil Loomba, Om D Deshmukh
Xerox Research Centre India
Bangalore, India
Email: om.deshmukh@xerox.com

***Abstract*—The amount of instructional videos avaiable online, already in tens of thousands of hours, is growing steadily. A major bottleneck in their wide spread usage is the lack of tools for easy consumption of these videos. In this demonstration, we will exhibit MMToC: Multimodal Method for Table of Content, a technique that automatically generates a table of content for a given instructional video and enables text-book-like efficient navigation through the video. MMToC quantifies word saliency for visual words extracted from the slides and spoken words obtained from the lecture transcript. These saliency scores are combined using a dynamic programming based segmentation algorithm to identify likely points in the video where the topic has changed. MMToC is a web-based modular solution that can be used as a stand alone video navigation solution or can be integrated with any e-platform for multimedia content management. MMToC can be seen in action on sample videos at http://104.130.241.45:8080/TopicTransitionV2/index.html.**

I. Introduction

Massive Open Online Courses (MOOCs), a recent phenomenon, where leading educational institutions of international repute offer their courses online for anybody with access to internet, is seen as a game changer for democratising quality education world over. One significant impact of the MOOC phenomenon is that they have accelerated the widespread availability of quality education content. The educational content generated by MOOCs, for most part, is made freely available and has also encouraged other educational institutions and experts to make their content publicly available online. We refer to this content as the Open Educational Resources (OERs). Almost all of this material is in the form of instructional videos. Video content, by its very nature, does not provide for user friendly ways of consumption. For exmaple, consider a textbook: the text document provides for an efficient medium to skim through the document to estimate the temporal flow of concepts, skip from the left top of the current page to the right bottom and back up to the middle as needed. Textbooks, by design, provide two more means of efficient non-linear navigation: (a) A table-of-content at the beginning of the book that outlines the global flow and relative significance of each topic (in terms of number of pages assigned), and (b) end-of-the-book index which collects relevant concepts and pointers on where to find them in the book. Instructional videos on the other hand have no such navigational capabilities.

In this paper we demonstrate a technique, called MMToC (Multimodal Method for Table of Content), for automatic generation of table of content for a video and linking it with the video for easy navigation. Related work includes segmentation of text documents using topic modelling techniques [1], segmenting spoken lecture recordings by modeling the acoustic parametrization as a multi-way normalized cut problem [2], vision-based techniques for lecture segmentation which use temporal variations in the content density functions [3]. The proposed MMToC solution defines word saliency across the visual and the spoken dimension, proposes ways to combine these two modalities and formulates a dynamic programming based cost function to identify topic boundaries. The salient phrases in each topic segment are chosen as representative tiles for the segment. A detailed description of each of these steps and comparative analysis with other state of the art techniques is presented in [6].

MMToC also has several practical application. For example, in a MOOC setting, MMToC can help user revise through the video lectures in a more efficient way: A large scale study on the EdX platform [4] found that certificate earning students, on an average, spend only about 4.4 minutes on a 12-15 minute-long video while skipping about 22% of the content. MMToC can help students navigate to the right content more efficiently. MMToC is also beneficial in bandwidth-constrained situations where downloading long videos may not be smooth. For example, consider a situation where a user in a bandwidth-constraint situation (say surfing on a mobile device in a remote area) searches for 'SVM video tutorial' online and is shown a list of video tutorial. S/he would not want to open each video, play for a few minutes to understand the flow of the video and then choose a different one. Instead, with the help of MMToC, the user can preview the flow of different videos and play only the most relevant ones.

The rest of the paper is organized as follows: Section 2 describes the functionality of the MMToC solution, section 3 explains the system details, section 4 presents the performance details. We briefly describe the demonstration logistics in section 5 and a sketch of future directions in section 6.

II. MMToC Functioning

The MMToC fuctioning can be broadly categorized into three main components: (a) multimodal saliency detection, (b) topic segmentation and naming, and (c) video navigation. Each of these components are explained below. Figure 1 shows the main building blocks of the proposed MMToC method. The method takes a set of uniformly sampled video frames and the speech transcript of the educational video as inputs and generates a list of topics along with their beginning times and title keyphrases.

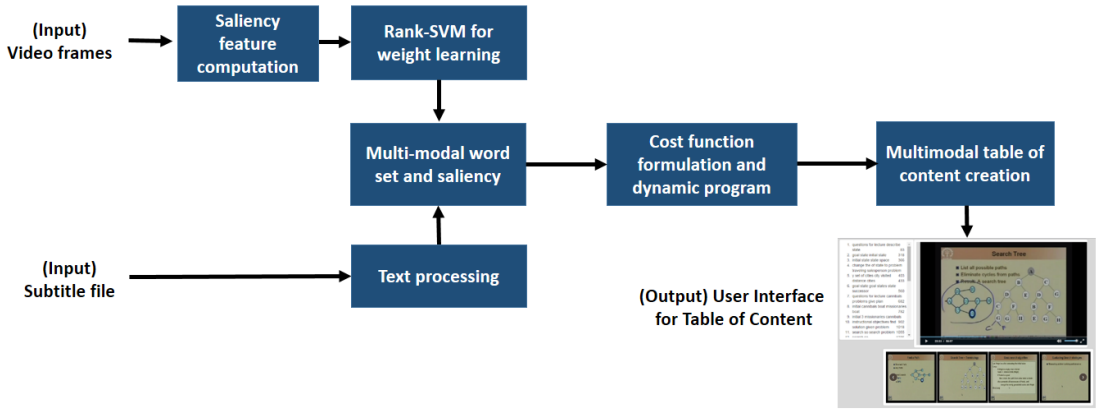

Fig. 1: Pipeline of the proposed method MMToC. MMToC takes the video frames and the speech-to-text transcript from an educational video as inputs and automatically creates a table of content.

Multimodal Saliency Detection: First, the methods used to determine the saliency of words in the slides and the text-to-speech transcript are described. We define and quantify saliency of words present on slides and use these saliency scores to improve topic partition. We establish that the identity of the words and the manner in which the words are rendered on the slides provide significant cues regarding the words' significance in topic change. For example, a word in bold and located towards the top left of the slide contributes more in the topic partition than a word located at the bottom right corner of a slide. To capture these visual characteristics, we propose seven novel mid-level features for the words present in educational videos. These features are called *underlineness, boldness, size, capitalization, isolation, padding,* and *location*. These features are combined using a weight vector to create a saliency score corresponding to every word in the video. The optimal weight vector is learnt using a novel formulation of the Rank-SVM algorithm [7] on human-annotated salient words. The weights determine the relative contribution of each visual feature to the overall saliency. The weights were learnt by collecting a training dataset from 10 users over 5 videos. 10 slides were randomly selected from each video (hence, total of 50 slides) to collect the training set. Each slide has been shown to 3 users and thus, a single user provides data for 15 unique slides. For each slide, the user was asked the following question - '*What are the salient words present in the slide that describe the overall content of the slide?*'. Generally, the number of selected salient words per slide vary between $2-12$ depending upon the user and the slide. To overcome inter-user subjectivity, a word is accepted as salient only if it is marked as salient by at least 2 users. Since in each slide users considered the selected words more salient than the words which were not selected, we consider them as pairwise preferences. These pairwise preferences can naturally be used in a Rank-SVM framework to learn the corresponding feature weights. We found that combining the visual features with equal weights often do not match the human provided ordering of salient words in a slide. Hence gathering training data from humans and using that in a discriminative learning framework to find the weights was required to accurately determine the saliency scores of words in a slide.

The Speech-to-text transcript of the lecture is also processed to estimate the saliency of the spoken words. We observe that in most cases, where the instructional videos are professionally created (e.g., a MOOC provider), the human annotated speech-to-text transcript is also provided. In cases where the manual transcript is not available we have an automatic speech recognition (ASR) setup built on the Kaldi framework [9]. The acoustic model for the ASR is trained on close to 50 hours of instructional video data from a variety of courses and instructor backgrounds. The language model is built using the text transcripts from close to 200 hundred hours of video lectures. A novel combination of graph-based and unsupervised text processing algorithms in conjunction with the visual saliency is used to generate a ranked list of multimodal salient words. These words along with their saliency scores are used to estimate the topic segmentation.

Topic Segmentation and Naming: The topic segmentation problem is formulated as an optimization problem and the optimization problem is solved using dynamic program. The cost function in the dynamic program is defined in such a way that it simultaneously minimizes a saliency-based metric of words common to two adjacent segments and maximizes the saliency-based metric of words unique to either of the two segments. The topic segments obtained using this approach are compared with ground truth topic segments on two different educational video datasets and we obtain an F-score of around 0.71 and 0.81 respectively on these datasets.

Given the topic segments, the next step is to automatically assign a representative name to each of the topics. The following steps summarize this process:

All the visually salient keywords in the given segment are identified. The spoken saliency score for these keywords, if any, is added to the visual saliency score. The keyword list is ranked based on this combined saliency score and the five most salient words are retained. The text transcript is then analysed to identify the salient words that are most co-occurring (within a window of ± 3 words). Up to three most common phrases of these salient words are chosen as the representative name for the topic.

The perfect algorithm for table of content creation should

select only one keyphrase for each segment. However, in real life, the best phrase selected by the algorithm may not be fully indicative of the actual topic of that segment. Thus, we choose multiple top key-phrases for each segment in the hope that the combination reflects the true topic in that segment. This also creates a more meaningful and complete table of content for the educational video.

Video Navigation: Although quantitatively our method is superior to other state-of-the-art approaches, we also evaluated how good MMToC is for non-linear navigation with real users. We create a user interface (UI) such that users can easily navigate through a video using MMToC. A screen-shot of the UI is displayed in Figure 2. There are two major components of the UI other than the actual video. First, the table of content generated by MMToC is displayed at the left of the video. Second, the beginning screen-grabs corresponding to each segment in the table of content are displayed below such that users can also scroll through them to find a topic. The slides were included in the UI to provide users the visual information such as figures, equations or text written in hand, which cannot be captured in the table of content. These components are hyperlinked to their corresponding time instances in the video for ease of navigation.

III. MMToC System Description

The MMToC platform is hosted on a cloud instance running Ubuntu 12.04 LTS OS. The user interface layer of is built using RubyOnRails MVC framework, bootstrap API and HTML5. A PostgreSQL server is used for database. The system is currently hosted as a web application on Tomcat web server and Apache application server. The architecture is built in a pure restful API manner where the client needs to get authentication code from the server and then the client can make a request for the MMToC API. In response to this client request, the server creates three essential components: the identified table of content with duration description, the time instances of the key frames with thumbnail details and the mapping between topic text and key frames.

The client application is built using RubyOnRails, HTML5, JQuery, CSS3 and bootstrap framework. We are using the YouTube iframe video player API to achieve this. Once the user authenticates with the authentication key then s/he can make a request to the server to get the topic transition data for a specific YouTube video ID, Table of content (i.e., topic transition) data creation is an offline process so if the user makes a request for the topic transition data for a video which is not preprocessed on server, then the system generates 'topic transition data not found' response to client and registers a request to generate topic transition data for that particular video ID. So when the user visits next time, then the data for the previous request can be made available to the user.

The UI design has three sections for a video: The video is played using the YouTube Iframe API. User's activity on the page and the player is tracked to perform approrpriate action on the video being played. The second section is the table of content which is located on the left side of the video where we populate topic and start-time pairs. The last section is the topic based key frames with corresponding times details. This is placed below the video. The user can chose any topic from the table of content by text search or s/he can visually identify key frames and use the visual cue for navigation to that particular topic. We are also provide a interlinks between table of content text and key frames video images, So if a user chooses a text form table of content then we automatically highlight the selected text and the key frame image with a different color bordering in the bottom slider for that particular topic and navigate the video to the beginning time of the topic. Similarly, if the user chooses an image from the key frame slider then we automatically add the color border around clicked image, highlight corresponding text in table of content section and navigate the video to the beginning of the corresponding topic.

Figure 2 shows a screen shot of the MMToC UI for an hour long instructional video. The table of content on the left side gives the global view of all the topics covered in the video along with the relative time spent on each topic. The slide-sorter at the bottom provides visual aids for the different topics. Three way hyperlinking among the textual table of content,topic specific slide sorter and the actual video help in efficient navigation.

IV. Performance Details

The MMToC setup was evaluated using two different ways: segmentation accuracy as compared with human-generated ground truth and with a few state-of-the-art baselines, and efficiency and effectiveness in user navigation.

Segmentation Accuracy: Human evaluators were asked to manually transcribe times of topic transitions for 20 different videos. These videos varied in duration from about 20 minutes to about 75 minutes. Each viedo was annotated by two human evaluators. Only those topic transition points where both the annotators agreed (within a tolerance of 5 ms) were used as the final ground truth transition points. On an average, we observe that the shorter videos (less than 30 minutes) have 5 ground truth partition points while the longer ones (more than an hour) have about 14 ground truth transition points. Baseline methods include (a) Latent Dirichlet Allocation (LDA) based method where the content between two adjacent unique slides is considered as a document and each document is assigned a topic ID. The time/ instance where LDA predicts a change of topic ID is the topic transition point. (b) proposed MMToC method with only the visual words, and (c) proposed MMToC method with only the spoken words. The topic transition point estimated by any of the automatic methods is marked as correct if it lies within ± 10 seconds of the actual ground truth. We also impose a one-to-one mapping between the ground truth marks and the estimated tranisiton points. The MMToC system using a combination of speech and visual information outperforms the three baselines. The improvement in F-Score using the proposed MMToC method as compared to the LDA method is 37%.

User Study: We conducted user studies on 9 participants using 3 hour-long videos. The users were shown either the MMToC navigation setup (Fig) or the youtube+transcript style navigation setup and asked to answer the following question: *where does the instructor start talking about topic X in the video?*, where *X* was one of the 5 topics randomly chosen from the ground truth. We observe that using the MMToC setup leads to, on an average, half the time as compared to the

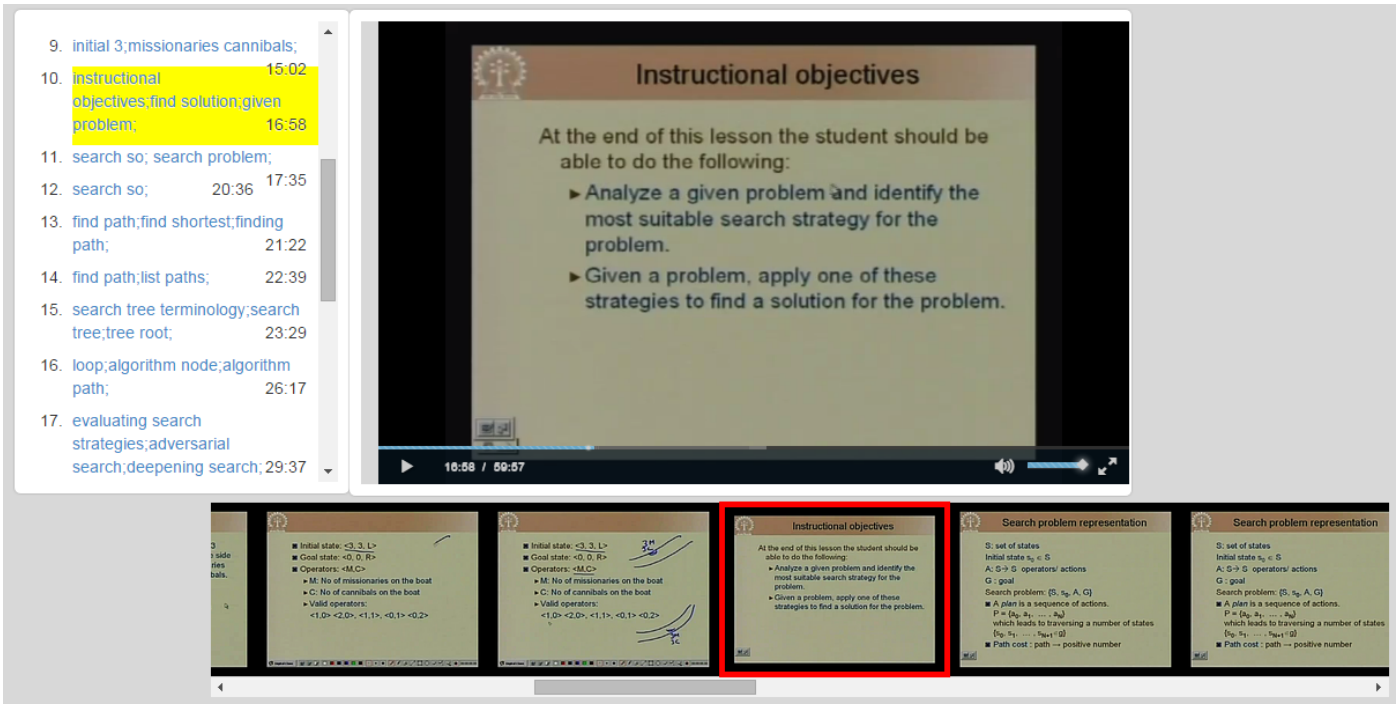

Fig. 2: Screen-shot of the user interface for displaying the table of content. Along with the generated table of content the interface also displays the corresponding slides, such that users can look into the figures, equations and hand-written content which are not included in the table of content.

time taken by the baseline youtube+transcript setup. Moreover, the percentages of correct answers (within ± 10 seconds) is also 10% higher using the proposed setup.

V. Demonstration Logistics

The proposed MMToC user interface is available online at http://104.130.241.45:8080/TopicTransitionV2/index.html for a sample video. We will be able to demonstrate the system in live action on videos of the conference attendees' choice. For these demonstrations, we will bring our laptops but would need access to power outlets. Access to a projector and/or internet connection is benefical but not needed.

VI. Conclusion and Future Work

In this work we propose a multimodal technique for automatic table of content generation for instructional videos and demonstrate the efficiency of the proposed system on a variety of videos. We are currently exploring ways to incorporate the speech modality in further improving the performance of the MMToC system. Specifically, spoken language literature mentions that the acoustic signature of the speech changes substantially when a current topic ends and a new topic begins (for example, [8]). Some of the current directions we are pursuing include combining the table-of-content view with the end-of-the-book index page component automatically generated for the same videos to provide a complete textbook like navigation interface. We had presented a multimodal approach for such index page generation in [5]. We look forward to discussing the MMToC way of navigating through instructional videos and getting inputs on other ways of providing non-linear efficient video navigation techniques.

References

[1] Du, Lan, Wray L. Buntine, and Mark Johnson. "Topic Segmentation with a Structured Topic Model." HLT-NAACL. 2013.

[2] Malioutov, Igor, and Regina Barzilay. "Minimum cut model for spoken lecture segmentation." Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2006.

[3] Phung, Dinh Q., Svetha Venkatesh, and Chitra Dorai. "High level segmentation of instructional videos based on content density." Proceedings of the tenth ACM international conference on Multimedia. ACM, 2002.

[4] Guo, Philip J., and Katharina Reinecke. "Demographic differences in how students navigate through MOOCs." Proceedings of the first ACM conference on Learning@ scale conference. ACM, 2014.

[5] Yadav, Kuldeep, et al. "Content-driven Multi-modal Techniques for Non-linear Video Navigation." Proceedings of the 20th International Conference on Intelligent User Interfaces. ACM, 2015.

[6] Biswas, Arijit, et. Al. MMToC: A Multimodal Method for Table of Content Creation in Educational Videos, to appear, Proceedings of the ACM Multimedia conference, 2015

[7] Chapelle, Olivier, and S. Sathiya Keerthi. "Efficient algorithms for ranking with SVMs." Information Retrieval 13.3 (2010): 201-215.

[8] Stolcke, Andreas, et al. "Combining words and speech prosody for automatic topic segmentation." Proceedings of the DARPA Broadcast News Workshop. 1999.

[9] Povey, Daniel, et al. "The Kaldi speech recognition toolkit." (2011).