Automated segmentation of MOOC lectures towards customized learning

Xiangrong Zhang
Xidian University, Xi'an, China
MIT Computer Science & Artificial Intelligence Lab
xrzhang@csail.mit.edu

Chen Li, Shang-Wen Li, Victor Zue
MIT Computer Science & Artificial Intelligence Lab
{cli, swli, zue}@csail.mit.edu

Abstract— **The sheer size of the student body for MOOC and the diversity of their learning styles and backgrounds demand that we develop alternatives to the one-size-fits-all pedagogy used in residential education. An important aspect of this endeavor is the segmentation of the video material, since it forms the omnipresent and central part of every course, and structuralized videos allow non-linear navigation as well as help learners with various needs find desired information efficiently. Here, we propose an automatic visual transition detection method to partition lecture videos into self-contained segments, which is the foundation to structuralize video and support non-linear navigation. Our method can be done at scale and has been proved being able to achieve reasonable quality.**

Keywords – MOOCs; customized learning; automatic lecture video segmentation; lecture video structrurization

I. Introduction

In contrast to residential education, the sheer size of the student body taking Massive open online courses (MOOCs) and the diversity in their backgrounds and learning styles pose new challenges to the conventional one-size-fits-all pedagogies [1]. Methods utilizing digitalized educational resources have been studied to provide tailored learning experience [2][3]. To enhance learning experience, we explore how learning resources could be organized and presented to provide customized learning for different learners.

MOOC resources include multiple modalities -- lecture videos and the audio transcriptions, slides, textbooks, forum discussions, and clickstream log data. Among them, lecture video is arguably the central and omnipresent component for knowledge transfer, to which other data modalities support. Thus, we focus on designing a method that can organize video resources to dynamically fit different learners.

The diverse learning styles and backgrounds suggest that learners may need nonlinear learning paths to suit their needs during lecture watching. These may include **searching, skipping**, and **reviewing**. These activities are commonly found in MOOCs based on the analysis of Kim et al. [4] on millions of log events. Therefore, we surmise that learners can benefit from video search/navigation. Quickly locating the demanded information in a video requires segmenting and annotating videos in a high, structural granularity.

The automatic segmentation of lecture video described in this paper is motivated by three premises. First, education research has shown that powerful learning gain can be achieved by presenting knowledge *structure* to learners [5]. To generate a structure of a lecture video, the first step is to divide a complete video into segments, in accordance with a certain knowledge granularity. Secondly, video shot segmentation is able to divide a video into relatively independent visual entities, which is *meaningful* for educational contents since visual changes reflect how a lecturer organizes the pieces of knowledge in a lecture video, and the organization greatly affects learners' watching behavior, e.g., visual changes usually coincide the change of topics (introducing new topic/concluding previous topic), change of slides, etc. The study [4] illustrates through interactive data that most re-watching peaks (61%) coincide with visual transitions in the video. These transitions could be meaningful segmentation boundaries to achieve reasonable information granularity. Thirdly, *automatic* video shot segmentation is required for massive online lecture videos. The segmentation does not only enable quickly locating and navigating, but also is the foundation of more high-level video analysis, e.g. video semantic analysis, summarization, etc.

In our case, being different from conventional scene segmentation (e.g., movies) which is based on abrupt scene changes, educational video shot changes are due to gradual topic shift. Similar scenes (e.g. slides, talking head in the same classroom) may occur across an entire video. Because of the correlation of a sequence of frames, visual transition detection can be regarded as a low-rank problem where the global structure of a video is explored as a whole. In addition, due to the similarity of adjacent frames, the local information is embedded. Exploring both global and local information, an automatic visual transition detection method has been specifically designed for lecture video segmentation, where some domain knowledge in education is applied. Our experiments use crowd-sourcing methods, e.g., Amazon Mechanical Turks (AMT), to establish the ground truth, and the resulting performance evaluation are presented. Based on the visual transition boundaries, we design a prototype of MOOC player guided by transition plots, with which learners will benefit from browsing at their own pace.

II. Lecture Video Transition Boundary Detection

Generally, the types of content in lecture videos include: talking head, slide, writing on board, illustration, etc. In the production of a lecture video, there are a lot of transitions between visual views, e.g. switching from a talking head view to a slide, or switching from a slide to an illustrative example. To realize non-linear navigation of educational videos, transition boundary detection is a good solution.

2161-377X/16 $31.00 © 2016 IEEE
DOI 10.1109/ICALT.2016.25

A. Representation of Visual Content of Lecture Videos

We first obtain a representation of visual content. For our application, the descriptive efficiency of hue, saturation, and value (HSV) histogram descriptor is exploited. HSV histogram represents a video frame as a feature vector describing color distribution in the frame. It works well in detecting the transition boundaries between a talking head shot and a slide. However, it cannot well distinguish visual transitions within the same scene – e.g., building up a slide one bullet at a time, because spatial distribution information is not considered. Besides HSV, a horizontal projection technique is applied in our study to identify slide changes. Specifically, for each frame with size of $m \times n$, we first project the content horizontally to one end, and then calculate the intensity sum of each row. As a result, each frame is represented by a vector with size $m \times 1$. We concatenated two types of descriptors into one vector to represent each frame.

B. Automatic Visual Transition Boundary Detection via Sparse and Low-rank Decomposition

Most existing shot segmentation methods concentrate on differences between consecutive frames [6]. It has been demonstrated that embedding additional information, e.g. contextual information, in shot boundary detection would effectively reduce the influence of various disturbances [7]. Thus, we focus on differences of all frames and propose a video shot segmentation based on low-rank and sparse matrix decomposition. Frames inside shots are approximately low rank due to visual similarity and continuity. Besides, neighboring frame information is embedded in low-rank component by total variation (TV) regularization [8] which assumes that neighboring frames are likely to be similar.

Given a matrix $X \in \Re^{d \times n}$ where each column is a feature vector of frame, the algorithm is to find a low-rank matrix $L \in \Re^{d \times n}$ and a sparse matrix $S \in \Re^{d \times n}$ such that $\mathbf{X} = L + S$ and $S \geq 0$. The non-negativity is more consistent with physical significance and convenient to compare differences of visual transitions. It can be formulated as:

$$\min_{L,S} \|L\|_* + \lambda \|S\|_1 + \beta \|CL^T\|_{2,1} \qquad (1)$$
$$s.t. X = L + S, S \geq 0$$

where $\|L\|_*$ is a nuclear norm of *L*. $\|S\|_1$ is the l_1-norm of *S*. $\|\bullet\|_{2,1}$ is the $l_{2,1}$-norm of a matrix. $\lambda > 0$ is a parameter that trades off between low-rank and sparse components. The third term measures the TV, where $C \in \Re^{n \times n}$ with the elements $c_{i,i} = -1$, $c_{i,i+1} = 1$, $i = 1 \cdots n-1$ and all the remaining elements being 0. We minimize the TV to guarantee the frames inside a shot to be similar. $\beta > 0$ controls the impact of the TV-regularization term in the optimization function.

Using the Augmented Lagrange Multiplier (ALM) method [9] to solve Eq. (2), we will finally get the low-rank component *L* and sparse component *S*. Each column of matrix *L* represents the correlation coefficients of each frame regarding to the whole data *X*. The norm of each column is calculated as the measure for comparison. High, medium, and low boundaries respectively correspond to prominent transitions between different views (e.g., transition between a talking head to a slide), the visual changes like transition from a slide to a new slide, and bullets changes on a slide.

III. Experimental Results and Discussions

A. The Course Material

We investigate and experiment on a complete MOOC course–6.00x: *Introduction to Computer Science and Programming Using Python,* offered by MIT from MITx on edX in spring, 2013. It contains 148 lectures totaling 1095 minutes of video, and 208 pages of slides. Removing 24 lectures of short introductions or demos less than 1 minute, we conduct experiments on the remaining 124 videos.

B. Ground Truth Generation through Crowdsourcing

To evaluate the effectiveness of the proposed method, we first collect an educational dataset with ground truth of visual transitions boundaries. We apply crowdsourcing techniques and recruit workers on AMT (https://www.mturk.com/mturk/welcome) for experiments, which is faster and more economical. Due to the existence of spammer among Turkers, quality control is necessary for the collected data. Studies have shown that Turkers can achieve similar quality to experts with quality control [10].

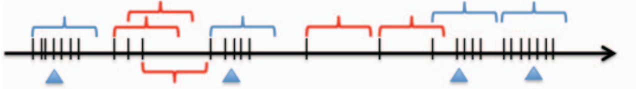

Figure 1. Example of the ground truth generation. If there are enough labels (5 in our experiment) in a window of 2 seconds, we will mark a time label for the ground truth and the location is the mean of values in the window; otherwise, we will move to the next window starting from the next time label. There are above 5 time labels in each blue bracket. The blue triangles mark the corresponding time position for the ground truth.

The purpose of our AMT experiment is to collect visual change points occurring in the lecture video on two levels (i.e. view changes and slide page changes) by crowdsourcing. We collect annotation from 10 Turkers for each of the 124 videos. In our case, each Turker's response to the visual changes is different, resulting in imperfect alignment of transitions marked by *all* Turkers. Therefore, we perform a majority voting in a sliding window of 2 seconds to establish our ground truth of video changes, which is described in detail in Figure 1. Furthermore, the majority voting works as a quality control mechanism which filters outliers.

We investigate the quality of each Turker's label by computing a kappa score between the Turker's result and a voting result of the other 9. The average kappa is 0.63, which implies a relatively good agreement. For the Turkers having significantly low agreement with others, we can generate the ground truth without those outliers with voting mechanism.

C. Comparison of Labeling Time and Computaional Time

To demonstrate the efficiency of the automatic video shot segmentation, we record the average computational time on the 124 lecture videos. For AMT labeling time, after removing outliers (labeling time more than 2 hours), the average labeling time of Turkers over all videos is 1002.0 seconds. The *baseline*, differences of consecutive frames based shot segmentation, costs 1.97 seconds averagely. Our method averagely costs 108.72 seconds over the 124 videos. Although the baseline is the fastest, both automatic approaches significantly outperform human and are immune to outliers.

D. Lecture Video Transition Boundary Detection Results

We evaluate the performance considering the transitions at 2 levels. One is view change (e.g., transition from a talking head to a slide; from a slide to a computer screen, etc.), and the other is a slide view changing to a new slide. The results are shown in Table 1, where our method outperforms the baseline method. This is due to the fact that the baseline method does not consider any contextual information, while the low-rank model well keeps a global structure of a video, and meanwhile a local similarity among frames is preserved in the matrix decomposition via the total variation regularization. Therefore, our method utilizes the global and local structural information of a lecture video.

TABLE 1. LECTURE VIDEO TRANSITION BOUNDARY (LARGE & MEDIUM) DETECTION RESULTS

Method	Recall	Precision	F-score
Baseline	0.584	0.390	0.468
Sparse and low-rank decomposition	0.565	0.439	0.494

IV. VISUAL TRANSITION GUIDED PLAYER

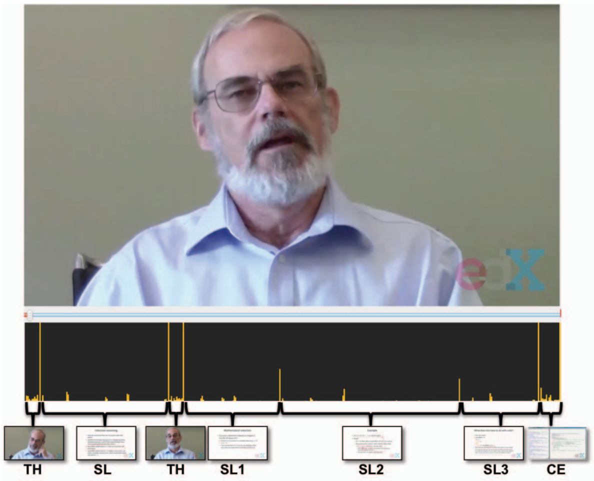

Figure 2. Visual transition guided player. TH, SL, and CE stand for talking head, slides, and code examples respectively. SL1, SL2, and SL3 are three consecutive slides during a particular shot. The height of bars depicts the difference of consecutive frames. High bars correspond to large changes, e.g. switching from a talking head to a slide. Medium bars correspond to medium changes, e.g. switching from SL1 to SL2. Low bars correspond to small changes, e.g. bullet changes. The thumbnails illustrate which part is talking head and which part is a slide etc.

We have conducted a prototype of visual transition guided MOOC video player, as shown in Figure 2. Guided by visual transitions, learners may watch a lecture video tailored to their own purpose and progress. For example, it is convenient in our player to skip to a new slide or go back to review a code demo, instead of watching a lecture video linearly. Non-linear navigation is impossible for traditional education. In a traditional classroom, all students have to follow the same progress, and could not switch to a new knowledge point until the instructor presents it. Besides, compared to search at random, visual transition is more likely to be a meaningful breakpoint (e.g. begin/end of a concept/example) to access/search materials in structure [5].

V. CONCLUSION

This paper describes the first step in our effort to provide students customized learning, to automatically partition lecture videos into self-contained learning segments with higher information granularities. The visual transition detection is regarded as a low-rank and sparse decomposition problem. The results suggest that the proposed automatic method is faster than the manual labeling of Turkers, which makes it capable for scale applications. Our method outperforms the baseline method in terms of precision and F-score.

On the basis of the results we obtained, the future work will firstly focus on semantically annotating video segments by exploiting multimodal data, i.e., both video and text information. This will open a new way of automatic linking of educational sources, video, slides, textbook, forum, etc., which will largely benefit customized learning and further exploit the great potential of MOOC.

ACKNOWLEDGMENT

The work is sponsored by Quanta Computer, Inc. under the Qmulus Project.

REFERENCES

[1] S.-W. Li and V. Zue, “Would linked mooc courseware enhance information search?” ICALT, 2015.

[2] A. Altun, and G. Kaya, “Development and evaluation of an ontology based navigation tool with learning objects for educational purposes,” The New Development of Technology Enhanced Learning (pp. 147-162): Springer Berlin Heidelberg, 2014.

[3] A. M. Souki, F. Paraskeva, A. Alexiou, K. A. Papanikolaou, “Developing personalised e-courses: tailoring students’ learning preferences to a model of self-regulated learning,” International Journal of Learning Technology 10 (3), 188-202, 2015.

[4] J. Kim, P. J. Guo, C. J. Cai, S.-W. (Daniel) Li, K. Z. Gajos, R. C. Miller, “Data-Driven Interaction Techniques for Improving Navigation of Educational Videos,” UIST 2014.

[5] J. Kim, P. T. Nguyen, S. Weir, P. J. Guo, R. C. Miller, and K. Z. Gajos, “Crowdsourcing step-by-step information extraction to enhance existing how-to videos,” CHI, 2014.

[6] J. Yuan, H. Wang, L. Xiao, W. Zheng, J. Li, F. Lin, and B. Zhang, A Formal Study of Shot Boundary Detection, IEEE Trans. Circuits and Systems for Video Technology, 17(2): 168-186, 2007.

[7] Y. Song, J. Vallmitjana, A. Stent, A. Jaimes, “Tvsum: Summarizing web videos using titles,” CVPR, 2015.

[8] Z. Harchaouia and C. Lvy-Leduca. “Multiple change-point estimation with a total variation penalty,” Journal of the American Statistical Association, 105:1480–1493, 2010.

[9] Z. Lin, R. Liu, and Z. Su, “Linearized alternating direction method with adaptive penalty for low-rank representation,” NIPS, 2011.

[10] M. Erdt, and C. Rensing, “Evaluating recommender algorithms for learning using crowdsourcing,” ICALT, 2014.