TRACE: Linguistic-based Approach for Automatic Lecture Video Segmentation Leveraging Wikipedia Texts

Rajiv Ratn Shah
School of Computing
National University of Singapore
Singapore
rajiv@comp.nus.edu.sg

Yi Yu
Digital Content and Media Sciences Research Division
National Institute of Informatics
Tokyo, Japan
yiyu@nii.ac.jp

Anwar Dilawar Shaikh
Department of Computer Engineering
Delhi Technological University
Delhi, India
anwardshaikh@gmail.com

Roger Zimmermann
School of Computing
National University of Singapore
and NUSRI, Suzhou, China
rogerz@comp.nus.edu.sg

***Abstract*—In multimedia-based e–learning systems, the accessibility and searchability of most lecture video content is still insufficient due to the unscripted and spontaneous speech of the speakers. Moreover, this problem becomes even more challenging when the quality of such lecture videos is not sufficiently high. To extract the structural knowledge of a multi-topic lecture video and thus make it easily accessible it is very desirable to divide each video into shorter clips by performing an automatic topic-wise video segmentation. To this end, this paper presents the TRACE system to automatically perform such a segmentation based on a linguistic approach using Wikipedia texts. TRACE has two main contributions: (i) the extraction of a novel linguistic-based Wikipedia feature to segment lecture videos efficiently, and (ii) the investigation of the late fusion of video segmentation results derived from state-of-the-art algorithms. Specifically for the late fusion, we combine confidence scores produced by the models constructed from visual, transcriptional, and Wikipedia features. According to our experiments on lecture videos from VideoLectures.NET and NPTEL[1], the proposed algorithm segments knowledge structures more accurately compared to existing state-of-the-art algorithms. The evaluation results are very encouraging and thus confirm the effectiveness of TRACE.**

***Keywords*-Lecture video segmentation; e–learning systems; multimodal analysis; Wikipedia; late fusion; linguistic features**

I. Introduction

A large volume of digital lecture videos has accumulated on the web due to the ubiquitous availability of cameras and affordable network infrastructures. However, a significant number of old (but important) videos with low visual quality from well known speakers are also commonly part of such databases. Because a specific topic of interest is often discussed in only a few minutes of a long video recording, it is essential to perform an efficient and fast topic boundary detection that also works robustly with lower quality videos. Moreover, such topic-wise segmentation of a lecture video into smaller cohesive intervals is advantageous to enable an easy search of the desired pieces of information. However, an automatic segmentation, indexing, and content-based retrieval of appropriate information from a large collection of lecture videos is very challenging because: (i) SRT (subtitle resource tracks) of lecture videos contain repetitions, mistakes, and rephrasings, (ii) the low visual quality of such videos may be challenging for topic boundary detection, and (iii) the camera may in many parts of a video focus on the speaker instead of the, *e.g.*, whiteboard.

State-of-the-art methods for automatic lecture video segmentation are based on the analysis of visual content, speech signals, and transcripts/SRT. However, none of the prior approaches consistently yields the best segmentation results for all lecture videos due to unclear topic boundaries, varying video qualities, and the subjectiveness inherent in transcripts. Since multimodal information augments multimedia based applications and services [6], [8], we postulate that a crowd-sourced knowledge base such as Wikipedia can be very helpful in the automatic lecture video segmentation because it provides several semantic contexts to analyze and segment videos more accurately. Empirical results in Section IV confirm our intuition. Thus, segment boundaries computed from SRT using state-of-the-art methods are further improved by refining these results using Wikipedia features. Our proposed TRACE system also works well for the detection of topic boundaries when only Wikipedia texts and SRT of lecture videos are available. Generally, the length of lecture videos ranges from 30 minutes to 2 hours, and computing the visual and audio features is a very time consuming process. Since TRACE is based on a linguistic approach, it does not require to compute such features from video content and

[1]National Prog. on Technology Enhanced Learning: http://nptel.ac.in/

978-1-5090-0379-2/15 $31.00 © 2015 IEEE
DOI 10.1109/ISM.2015.18

audio signals. Therefore, the TRACE system is scalable and executes fast. Interestingly, since the segment boundaries derived from the different modalities are highly correlated, it is desirable to investigate the idea of their late-fusion. Combining Wikipedia with other segmentation techniques also shows significant improvements in the *recall* measure.

Since the information requested by a user may be buried within a long video among many other topics, it is our goal to produce a semantically meaningful segmentation of lecture videos that is appropriate for information retrieval in e–learning systems. Specifically, we target the lecture videos whose video qualities are not sufficiently high to allow robust visual segmentation. To solve this problem, we propose the TRACE system which employs a linguistic-based approach for automatic lecture video segmentation using Wikipedia texts. We propose a novel approach to determine segment boundaries by matching blocks of SRT and Wikipedia texts of the topics of a lecture video. An overview of the method is as follows. First we create feature vectors for Wikipedia blocks (one block for one Wikipedia topic) and SRT blocks (120 words in one SRT block) based on noun phrases in the entire Wikipedia texts. Next we compute the similarity between a Wikipedia block and a SRT block using cosine similarity. Finally, the SRT block which has both the maximum cosine similarity and is above a similarity threshold δ is considered as a segment boundary corresponding to the Wikipedia block. We use a supervised learning technique on video content and linguistic features with SRT inspired by state-of-the-art methods to compute segment boundaries from video content and SRT, respectively. Next, we compare these results with segment boundaries derived from our proposed method.

II. Related Work

The rapid growth in the number of digital lecture videos makes distance learning very attainable [4]. Traditional video retrieval based on a feature extraction can not be efficiently applied to e–learning applications due to the unstructured and linear features of lecture videos [5]. For an effective content–based retrieval of the appropriate information in such e–learning applications, it is desirable to have a systematic indexing which can be achieved by an efficient video segmentation algorithm. The manual segmentation of a lecture video into smaller cohesive units is an accepted approach to find appropriate information [4]. However, it is not feasible due to the high cost of manual segmentation and rapid growth in the size of a large lecture video database.

Earlier approaches [2], [3] attempted to segment videos automatically by exploiting visual, audio, and linguistic features. Haubold and Kender [2] investigated methods of segmenting, visualizing, and indexing presentation videos by separately considering audio and visual data. Lin *et al.* [3] proposed a lecture video segmentation method based on natural language processing (NLP) techniques. N-gram

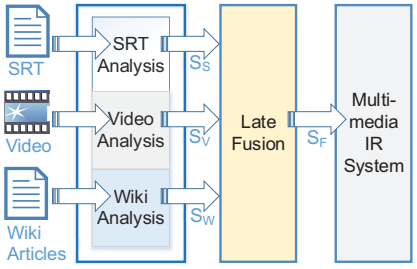

Figure 1. Architecture for the late fusion of the segment boundaries derived from different modalities such as SRT (S_S), video content (S_V), and Wikipedia texts (S_W).

based linguistic methods are also very useful in an effective retrieval of appropriate information [7], [9]. Most state-of-the-art methods on the lecture video segmentation by exploiting the visual content are based on color histogram. Zhang *et al.* [10] presented a video shot detection method using HMM with complementary features such as HSV color histogram difference and statistical corner change ratio (SCCR). However, not all features from a color space, such as RGB, HSV, or Lab from a particular color image are equally effective in describing the visual characteristics of segments. Therefore, Gao *et al.* [1] proposed a projective clustering algorithm to improve color image segmentation, which can be used for a better video segmentation.

III. System Overview

Figure 1 shows the system framework for the late fusion of segment boundaries derived form different modalities. First, the segment boundaries of a lecture video are computed from SRT (S_S) using the state-of-the-art work [3]. Second, they are predicted from the visual content (S_V) using the supervised learning method described in the state-of-the-art works [7]. Third, they are computed by leveraging the Wikipedia texts of the lecture video's subject (S_W) using our proposed method. Finally, the segment boundaries are derived from the previous steps are fused as described in the earlier work [7] to compute the fused segment boundaries.

SRT Segment Boundaries. We implemented the state-of-the-art work [3] based on NLP techniques to compute segment boundaries from a lecture video. They used content-based features such as noun phrases and discourse-based features such as cue phrases, and found that the noun phrase feature is salient. We used Reconcile and Stanford POS Tagger to compute noun phrases and part of speech (POS) tags from the available texts, respectively. We used Porter stemmer for stemming words. As suggested in earlier work [3], we used a block size of 120 words, shifted the window by 20 words every time, and computed the cosine similarity between feature vectors of adjacent windows by the standard formula $(\mathrm{A} \cdot \mathrm{B})/(||A|| + ||B||)$. A and B are the linguistic feature vectors for the adjacent SRT windows b_s. $||A||$ and $||B||$ are the magnitudes of the feature vectors.

Wikipedia Segment Boundaries. TRACE performs the temporal segmentation of a lecture video by leveraging SRT

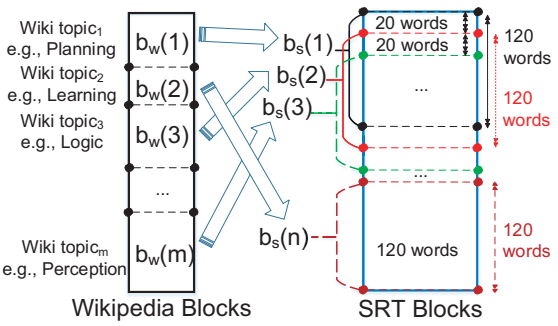

Figure 2. Architecture for segment boundary detection using Wikipedia.

and Wikipedia texts using linguistic features. Figure 2 shows the architecture of segment boundary detection from SRT using the proposed linguistic-based method which leverages the Wikipedia texts of subjects. We assume that the subject (*e.g.*, Artificial Intelligence) of the lecture video is known. Since a Wikipedia article consists of many topics, we parse the Wikipedia article to get texts of different topics. We refer to the block of texts corresponding to a Wikipedia topic as b_w. Next, we find a block b_s of 120 words from SRT which matches closely with the Wikipedia block b_w for each topic in Wikipedia texts. Specifically, first we create a Wikipedia feature vector f_w for each Wikipedia topic and a SRT feature vector f_s for each SRT block b_s of 120 words based on noun phrases in the entire Wikipedia texts. Next, we compute the cosine similarity $\alpha(b_w, b_s)$ between a Wikipedia block b_w and all SRT blocks b_s. The SRT block b_s which has both the maximum cosine similarity $\alpha(b_w, b_s)$ and is above a similarity threshold δ is considered as a segment boundary corresponding to the Wikipedia block b_w. Algorithm 1 describes the procedure of determining the segment boundaries using SRT and Wikipedia texts.

Video Segment Boundaries. A video is composed of several shots that are delimited by cuts and gradual transitions. In our earlier work [7], we leveraged the color histogram of keyframes of a lecture video to detect such boundaries by dividing each component of RGB into four bins. We used color histograms of keyframes to train a SVM^{hmm} model with three classes/events as follows [7]: (i) segments when only a slideshow is visible, (ii) segments when only speaker is visible, and (iii) segments when both speaker and slideshow are visible. However, we find that this model detects very few transitions for some lecture videos due to their low video quality. Thus, it is necessary to leverage other modalities as well for the accurate prediction of the segment boundaries.

Late Fusion of Segment Boundaries. Similar to earlier work [7], we fused the segment boundaries derived from the visual content, SRT, and Wikipedia texts by replacing two transitions less than thirty seconds apart by their average transitions time and keeping the remaining transitions as the final segment boundaries for the lecture video.

Algorithm 1 Computation of lecture video segments using SRT and Wikipedia texts

procedure WIKIPEDIABASEDTEMPORALSEGMENTATION
  INPUT: SRT $\mathcal{S}$ for a lecture video
  OUTPUT: A list of segment boundaries S_W
  P = getWikiPages($'Artificial\ Inteligence'$)
  T = getTopics(P) ▷ Get topics from Wikipedia pages
  B_W = getWikiBlocks(P, T) ▷ List of Wikipedia blocks
  B_S = getSRTBlocks($\mathcal{S}$, 120) ▷ List of SRT blocks
  for for each b_w in B_W **do** ▷ b_w is a Wikipedia block
    f_w = getWikiFeatVector(b_w) ▷ Wiki feature vector
    $\bar{\alpha}$ = 0 ▷ Initialize cosine similarity value
    for for each b_s in B_S **do** ▷ b_s is a SRT block
      f_s = getSRTFeatVector(b_s) ▷ SRT feature vector
      $\alpha(b_w, b_s)$ = getCosineSimilarity(f_w, f_s)
      if $(\alpha(b_w, b_s) > \bar{\alpha}) \wedge (\alpha(b_w, b_s) > \delta)$ **then** ▷ Select b_s from SRT which has the highest cosine similarity with b_w
        $\bar{\alpha} = \alpha(b_w, b_s)$ ▷ $\bar{\alpha}$ is max. cosine similarity
        $b_{s\prime} = b_s$ ▷ $b_{s\prime}$ is a boundary block in SRT
    t_s = getStartTime($b_{s\prime}$) ▷ Get start time of the block $b_{s\prime}$
    S_W = addToSegmentBoundaries(t_s) ▷ S_W is a list of derived segment boundaries by leveraging Wikipedia texts
  S_W = sortList(S_W) ▷ S_W is sorted in ascending order

IV. EVALUATION

Dataset and Experimental Settings. We used 133 videos[2] with several metadata annotations such as SRT, slides, transition details (ground truths), *etc*., from the VideoLectures.NET and NPTEL. Specifically, there are 65 videos V of different subjects from VideoLectures.NET and 68 videos V_T belong to the *Artificial Intelligence* course from NPTEL. Videos in V, whose transition details are known, are used for training different learning models, and videos in V_T are added to the testset. Most of the videos in V_T are old low quality videos since the target lecture videos for the TRACE system are mainly old lecture videos in low visual quality. We used the Wikipedia API[3] to determine texts for different courses and topics.

Results. Similar to earlier work [7], we computed precision, recall, and F-1 scores for each video in V_T to evaluate the effectiveness of our approach. For a few videos in V_T, these scores are very low if the quality of either the videos or the unscripted SRT is low. Therefore, it is desirable to leverage crowdsourced knowledge bases such as Wikipedia to overcome these issues. Moreover, we implemented state-of-the-art methods of lecture video segmentation based on SRT [3] and video content analysis [7], and investigated the fusion of segment boundaries derived from different modalities. Figure 3 shows first a few segment boundaries derived from different modalities for the lecture video[4]. We compared a predicted segment boundary c with a ground truth segment boundary g. Similar to earlier work [3], we

[2]Additional details of dataset: http://tinyurl.com/TRACE-Dataset-ISM15
[3]API URL: https://en.wikipedia.org/w/api.php
[4]Video URL: http://nptel.ac.in/courses/106105077/1

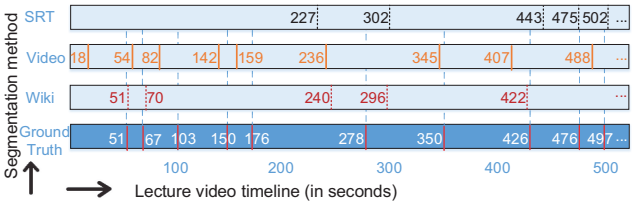

Figure 3. Segment boundaries derived from different modalities.

Table I
EVALUATION OF THE LECTURE VIDEO SEGMENTATION FOR THE LECTURE VIDEOS IN THE TEST SET V_T.

Segmentation Method	Avg. Precision	Avg. Recall	Avg. F-1 Score
Visual	0.360247	0.407794	0.322243
SRT	0.348466	0.630344	0.423925
Visual + SRT	0.372229	0.578942	0.423925
Wikipedia	**0.452257**	0.550133	**0.477073**
Visual + Wikipedia	0.396253	0.577951	0.436109
SRT + Wikipedia	0.388168	0.62403	0.455365
Visual + SRT + Wiki.	0.386877	**0.630717**	0.4391

considered a perfect match if c and g are at most 30 seconds apart, and partial match if c and g are at most 120 seconds apart. We computed the score for each (c, g) pair based on the time difference between them by employing a staircase function as follows:

$$score(c,g) = \begin{cases} 1.0, & \text{if } distance(c,g) \leq 30 \\ 0.5, & \text{else if } distance(c,g) \leq 120 \\ 0, & \text{otherwise.} \end{cases}$$

We use the following equations to compute precision and recall to evaluate the accuracy of the lecture video segmentation. Moreover, we compute F-1 score using the standard formula $(2 \times \text{precision} \times \text{recall})/(\text{precision} + \text{recall})$.

$$precision = \frac{\sum_{k=1}^{r} score(c,g)}{M}, \qquad recall = \frac{\sum_{k=1}^{r} score(c,g)}{N}$$

where M and N are the number of true and predicted transitions, respectively, and r is the number of (c, g) pairs. Table I shows these scores of the lecture video segmentation for TRACE, state-of-the-art works (for SRT and visual content), and their late fusion. Experimental results show that our proposed scheme to determine segment boundaries by leveraging Wikipedia texts results in the highest precision and F-1 scores, and performs well especially when the state-of-the-art methods based on the visual content and SRT fails to detect lecture video segmentations. Furthermore, when we performed the late fusion of all approaches then it results in the highest recall value.

V. CONCLUSIONS

The proposed TRACE system provides a novel way to automatically determine the segment boundaries of a lecture video by leveraging Wikipedia texts. To the best of our knowledge, our work is the first attempt to compute segment boundaries using crowdsourced knowledge base such as Wikipedia. We further investigated their fusion with the segment boundaries determined from the visual content and SRT of the lecture video using state-of-the-art works. Experimental results confirm that the TRACE system can effectively segment the lecture video to facilitate the accessibility and traceability within their content despite video quality is not sufficiently high. In the future, we plan to introduce a browsing tool for use and evaluation by students.

ACKNOWLEDGMENTS

This research was supported in part by the National Natural Science Foundation of China under Grant no. 61472266, the National University of Singapore (Suzhou) Research Institute, Suzhou Industrial Park, Jiang Su, China, and by JSPS KAKENHI Grant Number 15H06829.

REFERENCES

[1] S. Gao, C. Zhang, and W.-B. Chen. An Improvement of Color Image Segmentation Through Projective Clustering. In *International Conference on Information Reuse and Integration*, pages 152–158. IEEE, 2012.

[2] A. Haubold and J. R. Kender. Augmented Segmentation and Visualization for Presentation Videos. In *International Conference on Multimedia*, pages 51–60. ACM, 2005.

[3] M. Lin, M. Chau, J. Cao, and J. F. Nunamaker Jr. Automated video segmentation for lecture videos: A linguistics-based approach. *In IJTHI*, 1(2):27–45, 2005.

[4] C.-W. Ngo, F. Wang, and T.-C. Pong. Structuring lecture videos for distance learning applications. In *ISMSE*, pages 215–222. IEEE, 2003.

[5] S. Repp, A. Groß, and C. Meinel. Browsing within Lecture Videos based on the Chain Index of Speech Transcription. *In IEEE TLT*, 1(3):145–156, 2008.

[6] R. R. Shah, A. D. Shaikh, Y. Yu, W. Geng, R. Zimmermann, and G. Wu. EventBuilder: Real-time Multimedia Event Summarization by Visualizing Social Media. In *International Conference on Multimedia*, pages 185–188. ACM, 2015.

[7] R. R. Shah, Y. Yu, A. D. Shaikh, S. Tang, and R. Zimmermann. ATLAS: Automatic Temporal Segmentation and Annotation of Lecture Videos Based on Modelling Transition Time. In *International Conference on Multimedia*, pages 209–212. ACM, 2014.

[8] R. R. Shah, Y. Yu, and R. Zimmermann. ADVISOR: Personalized Video Soundtrack Recommendation by Late Fusion with Heuristic Rankings. In *International Conference on Multimedia*, pages 607–616. ACM, 2014.

[9] A. D. Shaikh, R. R. Shah, and R. Shaikh. SMS based FAQ Retrieval for Hindi, English and Malayalam. In *Forum on Information Retrieval Evaluation*, page 9. ACM, 2013.

[10] W. Zhang, J. Lin, X. Chen, Q. Huang, and Y. Liu. Video Shot Detection using Hidden Markov Models with Complementary Features. In *ICICIC*, pages 593–596. IEEE, 2006.