DynamicSlide: Reference-based Interaction Techniques for Slide-based Lecture Videos

Hyeungshik Jung
School of Computing, KAIST
hyeungshik.jung@kaist.ac.kr

Hijung Valentina Shin
Adobe Research
vshin@adobe.com

Juho Kim
School of Computing, KAIST
juhokim@kaist.ac.kr

ABSTRACT
Presentation slides play an important role in online lecture videos. Slides convey the main points of the lecture visually, while the instructor's narration adds detailed verbal explanations to each item in the slide. We call the link between a slide item and the corresponding part of the narration a *reference*. In order to assess the feasibility of reference-based interaction techniques for watching videos, we introduce DynamicSlide, a video processing system that automatically extracts references from slide-based lecture videos and a video player. The system incorporates a set of reference-based techniques: emphasizing the current item in the slide that is being explained, enabling item-based navigation, and enabling item-based note-taking. Our pipeline correctly finds 79% of the references in a set of five videos with 141 references. Results from the user study suggest that DynamicSlide's features improve the learner's video browsing and navigation experience.

Author Keywords
Educational videos; Visual navigation; Video learning.

CCS Concepts
•**Applied computing** → **E-learning;**

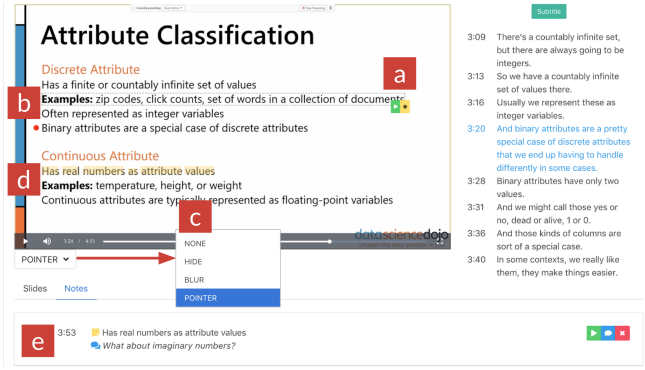

Figure 1. DynamicSlide recovers and uses the link information between a slide item and the corresponding verbal explanation in a lecture video to enable a set of reference-based interaction techniques. (a) Each text item in the slide has a play button and bookmark button that appear when hovered; (b) The current item explained by the instructor is emphasized by an indicator symbol (red dot); (c) Three different style of emphasis is supported; (d) Users can directly highlight text on a video frame; (e) Bookmarked text items are copied and collected in a separate pane with links to the relevant parts of the video.

INTRODUCTION
Among diverse styles of lecture videos, slide-based lectures are widely used for their familiarity, abundant pre-existing materials [11], and ease of sharing with students. While previous work has leveraged slides and narrations used in slide based-videos for effective browsing of videos [2, 10, 13], finding and utilizing the link between narrations and slide items for video learning is still an open problem.

In this work, we define *reference* as a pair of slide item and a relevant sentence in the narration. Leveraging references has the potential to improve the watching experience of slide-based lecture videos. References provide fine-grained connections between two complementary materials designed together—while slides provide a visual, structured summary of the lecture, narrations provide more detailed explanations about each topic represented in the slide. For example, using references, Tsujimura et al. [12] estimated the current explanation spot of a lecture and indicated it on the slide in the video. In addition to highlighting, references can support other common video watching tasks such as navigation and note-taking.

We present DynamicSlide, a video processing system that automatically finds references from slide-based videos and provides interaction techniques using these references. Our video processing pipeline extracts a set of slides from a video, finds text segments from the slides, and finds references between the text segments and parts of the narration. We focus on text items as they reveal more linguistic information that can be matched with narrations. Users can navigate to relevant points of a lecture by clicking items in the slide, make notes on slide items with pointers to the relevant explanation, and get support from the automatic highlighting of currently explained items.

In a preliminary study, 12 participants watched slide-based lecture videos using a baseline player and DynamicSlide player. Participants finished navigation tasks in less time using DynamicSlide, and responded that automatic highlighting and note-taking features were helpful. The result suggests that references can be automatically extracted and help common video watching tasks.

UIST '18 Adjunct October 14–17, 2018, Berlin, Germany

ACM ISBN 978-1-4503-5949-8/18/10.
DOI: https://doi.org/10.1145/3266037.3266089

DYNAMICSLIDE SYSTEM

Player Interface

The video player has three main features designed to help users learning with slide-based video.

Automatic emphasis of current item: The player informs the current part of slide being explained by the instructor to the user. Each item of a slide is emphasized when the sentence referencing the item starts. The system replicates three common practices of instructors for emphasizing the current item (Figure 1(c)): make the slide item visible simultaneous to its explanation, emphasize the slide item (e.g., brightening its color), or place an indicator (e.g., red dot) near the item, which is analogous to using a laser pointer (Figure 1(b)).

Item-based navigation: With item-based navigation, users can navigate the video by selecting a particular text item on the slide. Extending dynamic manipulation techniques for videos [1, 9], hovering over a text item in the slide reveals two action buttons next to it (Figure 1(a)): play and bookmark. By clicking the play button, users can navigate to the starting point of the sentence that is paired with the text item. This content-based navigation enables precise playback control at the item level, which provides a significant advantage over conventional linear navigation.

Item-based note-taking: Users can also make notes on text items in the slide. By clicking the bookmark button (Figure 1(a)) users can highlight a text item and copy highlighted items into a separate note pane (Figure 1(e)). They can add custom notes to any bookmarked item, which can be used for navigating to the relevant part of the lecture.

Video Processing pipeline

DynamicSlide's video processing pipeline takes a slide-based video as input and returns a set of unique slides, a set of text items in each slide, and matches between text items and the narration script as output. Its goal is to automatically generate the necessary data to power the reference-based interaction techniques. The pipeline consists of three main components as described below.

Stage 1: Slide Boundary Detection: The purpose of this stage is to extract a set of slides used in a lecture video and to segment the video based on the slide boundaries. To measure the difference between two consecutive frames, we calculate image difference using the method suggested by Zhao et al. [13] and text difference using Levenshtein distances between words found by OCR framework [6]. We regard two frames as different if both the image and text differences between two frames are higher than a predefined threshold.

Stage 2: Text Segmentation within Slides: The purpose of this stage is to group words in the slide into a set of semantic units, such as a phrase or a sentence. The main idea is to use the position information of words, inspired by the method used by Zhao et al. [13]. After grouping words into semantic units, we detect the slide title or headline, using a set of heuristics suggested by Che et al. [3]. We don't extract a reference for the title since titles often convey the overall theme of slides.

Stage 3: Text-to-Script Alignment: The goal of this final stage is to find an alignment between the text items in a slide (identified in Stage 2) and sentences from the narration script. The main idea is to find the most textually similar sentence in the script for each text item in the slide. We chose a sentence as a minimum information unit in the script since words or phrases are too fine-grained to be used in video navigation.

We represent each text item and script sentence as a bag-of-words vector, weighted by the TF-IDF score of each word. Then we calculate the cosine similarity between these vectors. Finally, for each text item in a slide, we match the most similar sentence from the script as its pair sentence to form a reference. Our pipeline correctly finds 79% of references in a set of five videos with 141 references. Videos are mainly composed of bullet-point text, produced by four lecturers, and they cover data science, statistics, biology, and journalism.

USER STUDY

We conducted a user study to assess how DynamicSlide's reference-based interaction techniques help users learn with slide-based videos. For comparison, we built a baseline player without the reference-based techniques but with navigation support using slide and transcript inspired by other research prototypes for lecture videos [8, 13]. In the study using within-subjects design, twelve participants watched two similar videos [5, 4] with baseline and DynamicSlide player for each video. After watching each video, they were asked to report self-measured cognitive load [7], and navigate to the relevant part of the video to answer given questions.

Results

Participants reported less cognitive load using DynamicSlide compared to the baseline (DynamicSlide/baseline overall: 3.1/3.2, mental: 3.2/3.5, temporal: 3.0/3.0, performance: 2.7/3.3, effort: 3.5/3.2, frustration: 2.6/3.1, out of 5), although the differences were not statistically significant using MANOVA. Users found information faster using DynamicSlide compared to the baseline when the target information was included in the slide (DynamicSlide/baseline: 21.5/26.5 sec). As expected, item-based navigation did not help when the information was only in the script (DynamicSlide/baseline: 32.6/27.7 sec). The performance between participants varied widely and the results were not statistically significant. User study in more realistic settings, such as watching a longer video or skimming a video for review will better evaluate the user experience of using DynamicSlide.

FUTURE WORK

We are improving DynamicSlide's video processing pipeline, specifically the text-to-script alignment. We also plan to run deeper analysis on the types of references in various slide-based lecture videos.

ACKNOWLEDGMENTS

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korean government (MSIT) (No.2017-0-01217, Korean Language CALL Platform Using Automatic Speech-writing Evaluation and Chatbot).

REFERENCES

1. Nikola Banovic, Tovi Grossman, Justin Matejka, and George Fitzmaurice. 2012. Waken. In *Proceedings of the 25th annual ACM symposium on User interface software and technology - UIST '12*. ACM Press, New York, New York, USA, 83. DOI: **http://dx.doi.org/10.1145/2380116.2380129**

2. Arijit Biswas, Ankit Gandhi, and Om Deshmukh. 2015. MMToC: A Multimodal Method for Table of Content Creation in Educational Videos. In *Proceedings of the 23rd ACM international conference on Multimedia - MM '15*. ACM Press, New York, New York, USA, 621–630. DOI: **http://dx.doi.org/10.1145/2733373.2806253**

3. Xiaoyin Che, Haojin Yang, and Christoph Meinel. 2013. Lecture video segmentation by automatically analyzing the synchronized slides. *Proceedings of the 21st ACM international conference on Multimedia - MM '13* (2013), 345–348. DOI: **http://dx.doi.org/10.1145/2502081.2508115**

4. Data Science Dojo. 2017a. Intro to Data Mining: Data Attributes(Part2). Video. (6 January 2017). Retrieved August 9, 2018 from **https://www.youtube.com/watch?v=xrFtN_UJhYc**.

5. Data Science Dojo. 2017b. Introduction to Data Mining: Data Attributes (Part 1). Video. (6 January 2017). Retrieved August 9, 2018 from **https://www.youtube.com/watch?v=hu7iGGnzq3Y**.

6. Google. 2018. Google Cloud Vision API. (2018). **https://cloud.google.com/vision**.

7. Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. *Advances in Psychology* 52, C (1988), 139–183. DOI: **http://dx.doi.org/10.1016/S0166-4115(08)62386-9**

8. Juho Kim, Philip J. Guo, Carrie J. Cai, Shang-Wen (Daniel) Li, Krzysztof Z. Gajos, and Robert C. Miller. 2014. Data-driven interaction techniques for improving navigation of educational videos. In *Proceedings of the 27th annual ACM symposium on User interface software and technology - UIST '14*. ACM Press, New York, New York, USA, 563–572. DOI: **http://dx.doi.org/10.1145/2642918.2647389**

9. Cuong Nguyen and Feng Liu. 2015. Making Software Tutorial Video Responsive. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*. ACM Press, New York, New York, USA, 1565–1568. DOI: **http://dx.doi.org/10.1145/2702123.2702209**

10. Amy Pavel, Colorado Reed, Björn Hartmann, and Maneesh Agrawala. 2014. Video digests. In *Proceedings of the 27th annual ACM symposium on User interface software and technology - UIST '14*. ACM Press, New York, New York, USA, 573–582. DOI: **http://dx.doi.org/10.1145/2642918.2647400**

11. Jasmine Rana, Henrike Besche, and Barbara Cockrill. 2017. Twelve tips for the production of digital chalk-talk videos. *Medical Teacher* 39, 6 (jun 2017), 653–659. DOI: **http://dx.doi.org/10.1080/0142159X.2017.1302081**

12. Shoko Tsujimura, Kazumasa Yamamoto, and Seiichi Nakagawa. 2017. Automatic Explanation Spot Estimation Method Targeted at Text and Figures in Lecture Slides. In *Interspeech 2017*, Vol. 2017-Augus. ISCA, ISCA, 2764–2768. DOI: **http://dx.doi.org/10.21437/Interspeech.2017-750**

13. Baoquan Zhao, Shujin Lin, Xiaonan Luo, Songhua Xu, and Ruomei Wang. 2017. A Novel System for Visual Navigation of Educational Videos Using Multimodal Cues. In *Proceedings of the 2017 ACM on Multimedia Conference - MM '17*. ACM Press, New York, New York, USA, 1680–1688. DOI: **http://dx.doi.org/10.1145/3123266.3123406**