FitVid: Towards Development of Responsive and Fluid Video Content Adaptation

Jeongyeon Kim, Juho Kim
School of Computing, KAIST
Daejeon, South Korea
{imurs34, juhokim}@kaist.ac.kr

Abstract

Video lecture content is increasingly consumed in mobile environments with varying screen sizes. However, most video content originally designed for desktop is not readable and digestible on small screens. We developed a computational pipeline that automatically adapts learning video content to a smaller screen by segmenting and resizing the in-video elements. We present FitVid, a video interface that provides both the pipeline-generated content adaptation and user-controlled direct manipulation of the in-video elements to fit their own needs. FitVid also provides customized content adaptation based on learners' manipulation log. In the user study (N=24) we find that FitVid significantly improves learning experience with increased concentration and readability. We further discuss design implications for responsive and customized video content adaptation.

1 Introduction

In online learning, video is a dominant medium with the educational benefits of media-rich content (Havice et al. 2010; O'Neill-Jones 2004; Liu, Liao, and Pratt 2009). In addition, the lockdowns and school closures caused by the global pandemic have created a spike in learners on video learning platforms such as MOOCs (e.g., edX, Coursera, Udacity, and FutureLearn) due to their openness and easy accessibility (Zhou et al. 2020; Alamri et al. 2020; Seale et al. 2020). MOOC learners are not bound to a desk and often access learning content via mobile devices at both a time and location suitable for them (DeWaard et al. 2011).

However, one of the key limitations of mobile learning is the small screen size, which deteriorates the learning experience and decreases the effectiveness of learning with too small font size, content-heavy lecture slides, and complex graphics to digest in a mobile environment. Most of the existing video learning content is originally designed for desktops with wide screens, resulting in a degraded learning experience when learners access learning material from a smartphone or a tablet.

Despite the need for content adaptation, the inflexible nature of video as a medium poses a challenge to fluid content adaptation compared to static content such as text and images. It remains time-consuming and tedious to edit and tailor video content, which often requires complex editing skills (Hua, Wang, and Li 2005; Long et al. 2004; Casares et al. 2002).

To address these challenges, we propose FitVid, an interactive video interface which supports fluid content adaptation in response to target devices and user needs. We first developed a computational pipeline which automatically adapts video learning content to mobile environments by trimming information and resizing content, based on the existing design guidelines for font size and amount of information (Inc. 2020 (accessed September 10, 2020; Pugsley 2010; Alred, Brusaw, and Walter; Larocque, Kenny, and McInnes 2015). The pipeline-generated content showed an improved compliance rate for the design guidelines, from 2% to 89% for the font size, and 67% to 87% for the word count with ten sampled videos. FitVid then provides the pipeline-generated content adaptation in response to the target devices.

Another main aspect of FitVid is the direct manipulation of the in-video elements, which enables content customization to fit individual learners' needs. While watching a video, FitVid enables users to directly manipulate the in-video content in real-time by repositioning and resizing objects shown on screen. This approach gives learners control over content adaptation instead of automating the entire adaptation process.

Furthermore, we attempt to improve the default content adaptation based on users' manipulation log. While users manipulate the design elements, the system captures the manipulation log and creates user profiles to reflect them to future content adaptation. We explore the possibility of improving adaptation iteratively, aiming to reduce the need for manual adjustments. While having direct control on design is important, making manipulations each time can be cumbersome and may even hinder learning.

In our user study (N=24), FitVid significantly improved participants' learning experience with increased readability and concentration. We also observed the patterns and strategies of using direct manipulation, and categorized them by the purpose of manipulations. The identified purposes include adjusting design and improving concentration. Participants used the direct manipulation to refine the automated content adaptation and increase their level of concentration.

In summary, the primary contributions of this work are:

- An automated computational pipeline that generates video content adaptation

- A system that provides learners with fluid and responsive video content adaptation
- An implementation and exploration of user-controlled customization of content adaptation
- An empirical evaluation with learners showing that FitVid improved video-based learning experience

2 Related Work

Our work is informed by previous work in the domains of learning content adaptation, direct manipulation interfaces for video, and learning content design.

Learning Content Adaptation

Previous work has investigated techniques for learning content adaptation in response to an increasing demand for ubiquitous learning. Researchers proposed web design adaptation for mobile learning by introducing the concept of responsive web design for mobile learning (Peng and Zhou 2015; Bhuttoo, Soman, and Sungkur 2017). Others suggested responsive content adaptation for information visualization (Hoffswell, Li, and Liu 2020; Wu et al. 2020). However, most approaches are limited to adjusting static content such as text and images. Our research extends the domain of learning content adaptation from static content to dynamic content—video—which has been challenging due to its inflexible nature, not allowing adjustments for in-video elements such as text and images once the video is recorded.

Direct Manipulation Interfaces for Video

Direct manipulation interaction coined by Ben Shneiderman is an interaction style in which users act on displayed objects of interest using physical, incremental, and reversible actions whose effects are immediately visible on the screen (Shneiderman 1997; Shneiderman and Maes 1997). A rich body of work attempted to support direct manipulation video navigation by enabling in-video object dragging along its motion trajectory (Dragicevic et al. 2008; Karrer et al. 2008; Karrer, Wittenhagen, and Borchers 2009). Another body of previous work introduced zoomable video interfaces to overcome the constraint of small screen size (Pang et al. 2011; Song et al. 2010; Quang Minh Khiem et al. 2010; Axel, Ravindra, and Tsang 2010; Carlier et al. 2011). Our system builds upon the existing work on video content adaptation with direct manipulation for in-video elements, enabling the direct manipulation for in-video elements to overcome the constraint of small screen size and to fit their own needs.

Learning Content Design

When creating mobile learning content, instructional designers and video engineers should consider the limited screen size of mobile devices. Design principles for mobile learning content suggest adjusting the amount of information displayed on the screen (Ally 2005). The conceptual framework of design issues and learning contexts in mobile learning is also discussed (Parsons, Ryu, and Cranshaw 2006; O'Malley et al. 2005; Stanton and Ophoff 2013; Kukulska-Hulme and Traxler 2007). Other research conducted content analysis of MOOC videos to explore ideal content distribution for course developers (Carlon, Keerativoranan, and Cross 2020).

3 Design Goals

The need for learning content adaptation for mobile devices led to the following design goals that informed the design of techniques for interactive content adaptation in response to mobile devices:

D1. Support responsive design of video content for mobile devices Existing content adaptation techniques lack adapting content at an in-video level mainly due to the inflexible nature of video as a medium. Content adaptation at an in-video level that enables fluid and responsive video content adaptation across various devices is needed. For this adaptation, we should enable flexible deconstruction and reconstruction of the in-video elements such as text and images.

D2. Enable user-controlled direct manipulation The *"one-size-fits-all"* approach applying a single design to all the learning contexts without tailoring to individual needs has limitations in two main ways. First, each user has their own preferences and constraints. Second, lecture content varies significantly depending on the subject and instructional design, thereby requiring different content designs according to the characteristics of each lecture. Thus, the blanket application of the suggested guidelines from the literature would not address the diversity of user needs and lecture content. This demand indicates the importance of customization and user control in content adaptation. One way to provide customization and user control is to enable direct manipulation of the in-video elements as needed. We design a direct manipulation technique based on the important factors in video content adaptation which are informed by the formative interviews with video production engineers.

D3. Support customized content adaptation While having direct control over the content design allows users to adjust the design to fit their needs, it might hinder learning if they need to manually adjust the content each time. To reduce the need for manual manipulation, we design techniques to personally calibrate the content adaptation by reflecting individual users' preferences.

4 System Overview

To accomplish the design goals, we present FitVid, a fluid and responsive video content adaptation tool. Before learners start watching a video, FitVid generates content adaptation automatically based on the design guidelines from literature.

Design Guidelines

We investigated the existing guidelines for key design factors in lecture material including text and image elements.

Font Size Inappropriate font sizes of learning material impose unnecessary cognitive load (Lewis 2016) and lower judgements of learning (JOLs) (Rhodes and Castel 2008; Halamish 2018). Apple's Human Interface Guidelines adopt 17 pt as a default body text size (Inc. 2020 (accessed September 10, 2020) and Google Material Design Guidelines suggest 16 pt as body text size (GoogleLLC 2020 (accessed September

10, 2020), whereas the guidelines for presentation slides encourage using font size above 24 or 26 pt in the body of the slide (Pugsley 2010; Holzl 1997; Cavanaugh and Cavanaugh 2000).

Number of Words An excessive amount of words is another factor that increases cognitive load (Sweller 1994; Sweller, Van Merrienboer, and Paas 1998; Lewis 2016) and information overload (Ally 2005). Using no more than 45 words per presentation slide is recommended (Alred, Brusaw, and Walter) and more strict guidelines advocate using less than 20 words per slide (Brock and Joglekar 2011). Another work suggests that the maximum number of words per slide should be 25 (Stein 2006).

Image Image elements can also increase cognitive load by splitting learners' attention (Lee, Plass, and Homer 2006; Lewis 2016). Existing work on lecture slide design for radiology recommends lecture slides to contain maximum two images in a single slide (Larocque, Kenny, and McInnes 2015; Christian Davidson and Wiggins 2003).

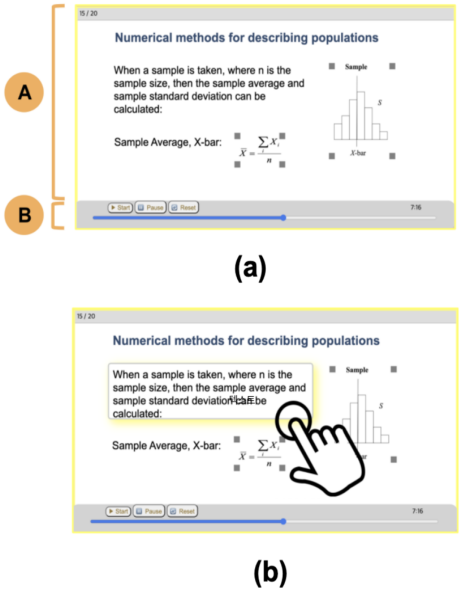
Before watching: automated content adaptation

FitVid applies automated content adaptation based on the design guidelines (addressing D1: Support responsive design of video content for mobile devices). The adaptation enlarges the text and trims the amount of information. The criteria of design features we considered are font size, the number of words, and the number of images. To generate content adaptation, our system deconstructs an original video into in-video visual elements such as text box and images, and reconstructs them to comply with the following design guidelines: (1) font size: enlarge the font size of text when it is smaller than 28.5 px; (2) number of words: segment a slide with more than 30 words into multiple slides; and (3) number of images: segment a slide with more than two images into multiple slides. The details of content adaptation technique are described in Section 4. After reconstructing the visual elements, we determine the time duration the reconstructed elements should be displayed. To this end, we establish correspondences between visual elements such as text and images in a video lecture and their descriptive speech text (i.e., instructor's current narration). We use a variant of slide-transcript matching algorithms suggested in previous work (Tsujimura, Yamamoto, and Nakagawa 2017; Xu et al. 2019; Zhao et al. 2019; Jung, Shin, and Kim 2018). We use BERT-based semantic similarity (Reimers and Gurevych 2019) to improve the performance of text similarity estimation.

During watching: in-video direct manipulation

While users are watching a video lecture with FitVid, they can directly manipulate in-video elements (addressing D2: Enable user-controlled direct manipulation). They can manipulate text and images using two types of interactions: reposition and resize. For example, learners can enlarge a complex image which is not visible on a small screen. They can also resize text as needed by dragging the edges of the text box.

We developed a web-based interactive video player allowing users to directly manipulate the in-video elements



(a)

(b)

Figure 1: (a) The FitVid video player consists of (A) the manipulatable video content area and (B) the video player control bar. (b) If the learner touches an in-video element, then edges of element are highlighted.

such as text boxes and images. FitVid is implemented using JavaScript, jQuery, and CSS media queries. Figure 1 (a) shows an interface of the video player. All elements in the video content area can be manipulated in real-time while watching a video. The video player control bar has three buttons; start, pause, and reset. The reset button initializes all the changes made in the current frame. Learners can resize elements by dragging the edges and reposition them by touch and drag interaction. Figure 1 (b) demonstrates that the edges of the in-video elements are highlighted when learners touch them.

After watching: customized content adaptation

We built a prototype which generates customized content adaptation based on users' manipulation log (addressing D3: Support customized content adaptation). In this work, we specifically focus on font sizes. The system calculates the average size of fonts to which a user manipulates the text. It then creates a user profile on their preferred font size. Based on the user profile, the system generates future content adaptation tailored to individual learners. For instance, if a certain learner watches a video resizing the font size as 20 pt on average, then the prototype displays the future content adaptation which adjusts the font sizes of text as 20 pt.

Computational Pipeline

To automatically apply content adaptation, we introduce a technical pipeline which includes: (1) shot boundary detection, (2) deconstruction into in-video elements, (3) text-to-script matching, and (4) adaptation generation. The pipeline is shown in Figure 2. In this section, we discuss the technical

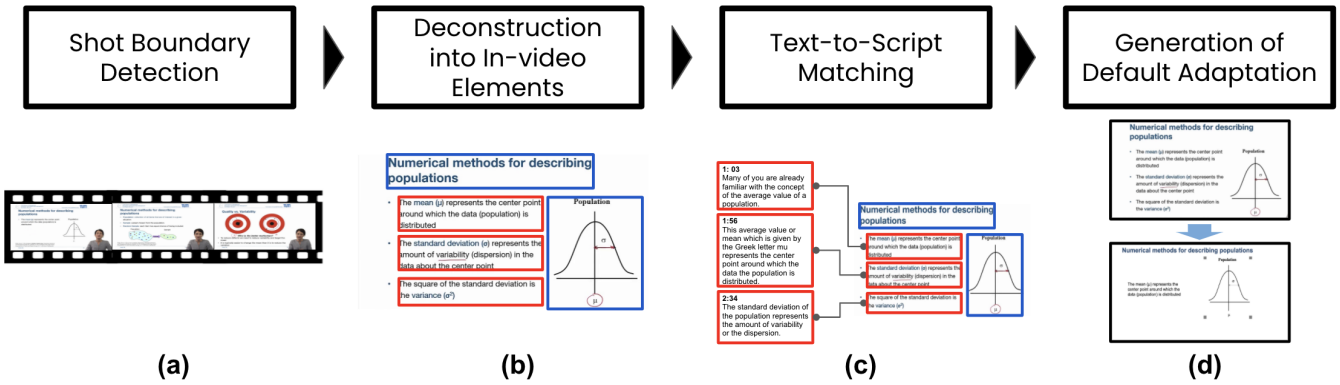


Figure 2: Our computational pipeline for generating content adaptation: (a) extraction of video frames using edge-based shot boundary detection algorithm, (b) deconstruction of original video content into in-video elements by detecting connected-components from the edge image, (c) matching in-video elements with transcript of instructor's current narration using BERT-based semantic similarity, (d) generation of adaptation through reconstructing the elements based on existing design guidelines.

details of the pipeline.

(a) Shot Boundary Detection We first extract the set of frames in a video lecture and segment the frames based on shot boundary detection. For shot boundary detection, we use a variant of methods suggested by Zhao et al. (Zhao et al. 2019) and Jung et al. (Jung, Shin, and Kim 2018). The shots extracted from the shot boundary detection correspond to unique lecture slides in most cases. We then select the first frame of each shot as a keyframe to represent a slide. We simply choose the first frame as a keyframe, since most frames in a shot with a lecture video only include trivial changes such as the movement of an instructor's head in a picture-in-picture video.

(b) Deconstruction into In-video Elements For each shot extracted using shot boundary detection, our pipeline deconstructs the shot into in-video elements such as text boxes and images. We use a variant of methods proposed by Zhao et al. (Zhao et al. 2019) and Jung et al. (Jung, Shin, and Kim 2018). The suggested methods first detect all connected-components from the edge image of the shot. For each identified component, they find the minimal bounding box which contains the component. All the text contained in each bounding box is then recognized using the Tesseract OCR engine (PythonSoftwareFoundation 2020 (accessed September 10, 2020).

(c) Text-to-Script Matching In this stage, the pipeline estimates the part of a slide that is currently verbally explained by the instructor. We match text elements in a frame with the transcript of the instructor's narration using Google's Speech-to-Text API (GoogleLLC 2020 (accessed September 10, 2020). There have been attempts to align text elements with a transcript for lecture retrieval (Lu et al. 2014; Zhao et al. 2019; Jung, Shin, and Kim 2018; Tsujimura, Yamamoto, and Nakagawa 2017). However, the challenges stem from a diversity of features of lectures such as linearity of instruction and interactivity between elements (Sweller 2010; Leahy and Sweller 2005). For example, if an instructor mentions multiple elements in a single narrated sentence, then text and transcript might not have a one-to-one correspondence. Due to such complexity of the task, we devise a rule-based matching algorithm enhanced by BERT-based text similarity. The matching algorithm aims to match sentences in a transcript with grouped text in a frame. The algorithm takes a two-step approach: (1) matching and (2) grouping.

(c)-1 Matching Stage: In the matching stage, we establish matches between the slide text and the transcript by using two factors: progression of in-video elements and text similarity. First, we consider the progression of each frame. We observe that the latest element added to a frame corresponds to the current explanation spot, which aligns with existing work (Zhao et al. 2019; Monserrat et al. 2013; Shin et al. 2015). Hence we detect the progression of elements and match them with the transcript. Secondly, we calculate text similarity of each sentence in a transcript with every text element in the same frame, using a BERT-based model presented by Reimers et al. (Reimers and Gurevych 2019). We build matches between text and transcript if they have a similarity score higher than a threshold. The threshold is determined based on the ratio of text elements to sentences in the transcript since each text element is explicitly mentioned by the instructor at least once. We first sort all the similarity scores in descending order. The i-th similarity score is selected as a threshold where i is calculated as follows:

$$i = \lfloor \frac{number\ of\ text\ boxes}{number\ of\ sentences\ in\ transcript} \rfloor$$

(c)-2 Grouping Stage: In the grouping stage, we merge the deconstructed text elements in stage (b) into groups. The goal of grouping is to combine text elements into units that need to be displayed to learners at once in a single frame. For example, three bullet points should be displayed at once in a single frame if an instructor explains them in a non-linear manner, such as referring to them back and forth. In this case,

Rule Number	Rule	Rationale
Rule 1	**IF** lecturing is non-linear **THEN** merge all non-linearly lectured text elements into a group	Non-linearly lectured elements should be displayed at once so that learners can refer to them back and forth.
Rule 2	**IF** text elements in a single frame has high cohesion **THEN** merge all text elements into a group	Text elements with high cohesion implying high interactivity between elements should be displayed at once so that learners can refer to them together.
Rule 3	**IF** any sentence in transcript within a single frame has multiple text elements with high text similarity score **THEN** merge all similar elements into a group	Multiple text elements mentioned in a single sentence in transcript should be displayed at once so that learners can refer to them in the sentence together.

Table 1: Rules and rationale for the rule-based method that groups text elements

the three bullet points should not be segmented into three frames, even if they are deconstructed into three text boxes in the previous step. Thus, we establish them as a single atomic unit. The sentences in a transcript matched with text elements in the matching stage are also grouped along with the grouped text elements. We devised a rule-based method to group text elements into atomic units. The defined rules and rationale for each rule are shown in Table 1.

For Rule 2, we utilize the text analysis technique suggested by Crossley et al. (Crossley, Kyle, and McNamara 2016), which estimates text cohesion indices. Text cohesion index calculates the amount of semantic overlap between adjacent sentences. The text elements with high cohesion imply that they have high interactivity between elements (Sweller 2010; Leahy and Sweller 2005). Hence we merge all text elements with high cohesion into a group so that learners can refer to them together in a single frame. We set the threshold as 0.3 in accordance with previous work (Karuna et al. 2018; Ward and Litman 2008).

For Rules 1 and 3, we identify the linearity of a lecture and concurrently mentioned text elements based on the established matches in the matching stage. For Rule 1, if an instructor does not mention text elements in a linear manner from top to bottom and from left to right, then we merge all non-linearly lectured text elements into a group since they should be displayed at once. For Rule 3, we merge the concurrently mentioned text elements by combining text elements that span multiple matches with a single sentence in the transcript. If an instructor mentions multiple text elements in a single sentence, then learners should be able to refer to all elements at once.

(d) Content Adaptation Based on Design Principles To set up basic design principles for content adaptation, we use the average of the existing design guidelines in Section 4. We use the averaged value instead of adopting a single guideline since there is no universal design guideline with a hard number. The averaged design principles we use are: (1) average font size of a frame should be above 21.4 pt, (2) number of words per frame should be below 30 words, and (3) a frame should contain maximum two images. Based on these design principles, we generate content adaptation. The rest of this section describes details of generating the content adaptation.

Font Size: If the average font size of a frame is smaller than 21.4 pt, then the fonts are resized to meet the guidelines. In case when the screen space is not enough to enlarge the fonts, the system enlarges the font to the extent to which there is no overlap between content.

Number of Words: The amount of text is adjusted if a frame contains more than 30 words. For example, if a frame consists of two text elements with 25 words and 20 words each, then we segment the frame into two frames, each having one text element with less than 30 words. We do not segment a single text box even if it violates the guidelines. Another exception is the text element which is grouped as an atomic unit in the grouping stage. We do not segment the atomic unit regardless of the number of words they contain.

Image and Headlines: We do not adjust image content during the adaptation process for two reasons. It was because of the difficulty of extracting semantic information from images (Xu et al. 2019; Tsujimura, Yamamoto, and Nakagawa 2017). Hence, the images in a single frame are not segmented into multiple frames.

Layout: The final stage is to compose the layout of a frame. We use the original positions of the elements in the video as we focus on the four design features (font size, amount of words, image size, amount of image), not including other design features such as an optimal layout or line spacing. Another main aspect of the generation of content adaptation is a fluid and responsive design based on relative units and media queries. The sizes of font and image are in relative units (e.g., percentage (%), em) instead of absolute units (e.g., px, pt), so that they are resized in response to the viewport size of devices.

5 User Study

The goals of evaluation were to investigate learner satisfaction and task performance using the content adaptation and direct manipulation features of of FitVid. We conducted a controlled user study that compares video players with and without the content adaptation and manipulation features. Our hypotheses are as follows:

- H1. FitVid with automated content adaptation increases the readability of content and perceived usefulness.
- H2. FitVid improves perceived learning experience and concentration.
- H3. FitVid with direct manipulation and customized content adaptation enhances interactivity, allowing learners to

refine and complement the automated content adaptation.

The study was a within-subjects design, where each learner used two different video players: (1) baseline interface and (2) content adaptation + direct manipulation. To maintain uniformity in look and feel for our comparative study, the baseline condition had the same layout and interface design as our system. We selected two videos each from three different courses (C1: Quantum Mechanics for Everyone[1], C2: Introduction to Financial[2], C3: Understanding Political Concepts[3]) and six videos in total. Each video has similar length (C1: 4:54, 5:53, C2: 6:22, 4:52, C3: 6:27, 7:51).

Participants

We recruited 24 participants [P1-P24] (12 male and 12 female) through online social media posting. Most of the participants were college students. They received $15 for up to 70 minutes of participation.

Procedure

Participants were first required to watch two video lectures using two different video players. They were randomly assigned to watch two videos from one of the three courses. After watching the video, they were asked about their perception of each video player and the reasons behind the real-time manipulations. They then completed a questionnaire on the usability, learning experience, cognitive load, and mind wandering for each interface. The questionnaire also includes scoring four design features (font size, amount of words, image size, amount of image) in each condition.

Results

We summarize the results (Figure 3) and describe the main findings focused on the three hypotheses, patterns of tool usage, and usability and usefulness of our system.

H1. FitVid with automated content adaptation increases the readability of content and perceived usefulness

In response to two 7-point Likert scale questions (1: strongly disagree, 7: strongly agree) about the readability of the video content, pair-wise Wilcoxon signed rank tests revealed that the lecture with adapted content is significantly more readable (Question 1: $Z = 45.5$, $p < 0.005$, Question 2: $Z = 29.0$, $p = 0.0025$). To measure the perceived usefulness of content adaptation, we asked participants to evaluate the design features of adapted content if they are significantly more appropriate than the ones of baseline interface (size of elements: $Z = 6.0$, $p = 0.0001$, amount of elements: $Z = 29.5$, $p < 0.01$). The scores on design features of the adapted content (Size of Text: 4.27 / 5, Amount of Text: 4.275 / 5, Readability: 4.18 / 5) were higher compared to the design features of the baseline interface (Size of Text: 1.82 / 5, Amount of Text: 1.68 / 5, Readability: 1.57 / 5). Thus, H1 is supported (Figure 3).

In accordance with the survey result about the readability of the video content, most participants noted that the adapted content is more readable and legible during the interview. *"The lecture material with less text and larger font size is more readable. The learning material with dense text and small fonts makes me not want to read them all."* (P1). *"The video content with less amount of information causes less fatigue."* (P23). Some participants mentioned that it's great that there's no need to find the spot where the instructor is currently mentioning in the slide, which makes it easier to understand the learning material. With the adapted content, all of the participants (24/24) responded that they are willing to watch a video lecture with a poor design that they would not have watched without the adaptation.

H2. FitVid improves perceived learning experience and concentration

FitVid improves the perceived learning experience with a significant difference ($Z = 33.5$, $p < 0.05$). We also found a significant difference in the levels of concentration and attention ($Z = 51.0$, $p < 0.05$), confirming that H2 is supported (Figure 3).

Participants explained that the adapted content with a large font size helps learn the content. *"I could get the main concept of content at a glance without additional effort to find it."* (P18). They also mentioned that the adapted content increases the levels of concentration and attention with less distraction.

H3. FitVid with direct manipulation feature enhances interactivity allowing learners to refine and complement the automated content adaptation

Most of the participants used the manipulation feature (22/24), while their goals differed. Common goals reported were to refine the incomplete default adaptation (20/24) and to improve concentration (18/24). We conceptually identified two high-level categories: 'Adjusting design' and 'Improving concentration' based on the purpose of manipulations explained in the interviews. We then classified the observed manipulation interactions into four different types. Within the 'Adjusting design' category: resizing image, resizing text, and large repositioning. For the 'Improving concentration' category: small repositioning and touching (highlighting). We describe the types of manipulations based on our observation and the reasons behind them explained in the interviews.

Adjusting design (20/24 participants) Participants explained they used the manipulation feature to make the content more readable. They noted that resizable elements compensate for the limitations of small screen size, enabling them to enlarge the content as needed. Another reason for resizing was to selectively view the elements. *"To efficiently use the limited screen space, I filled the whole screen with what I'm focusing on, covering other elements that I finished watching."* (P12). Some participants changed the alignment of elements, to be center aligned in most cases, based on their preferences. On the other hand, participants who have not used the direct

[1] https://courses.edx.org/courses/course-v1:GeorgetownX+PHYX-008-01x+1T2017/course/

[2] https://www.coursera.org/learn/wharton-accounting

[3] https://courses.edx.org/courses/course-v1:FedericaX+Fed.X-21+1T2020/course/

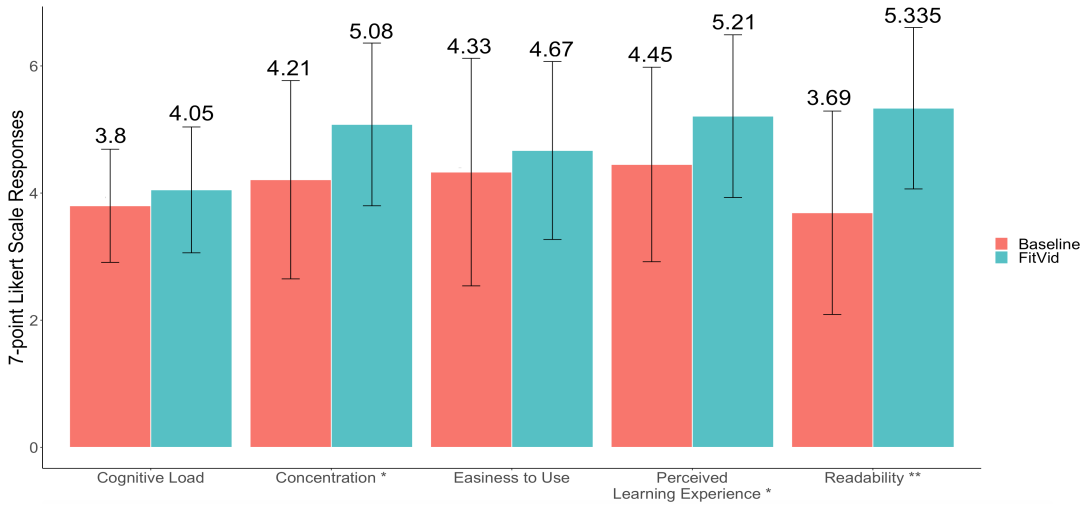


Figure 3: 7-point Likert scale responses for video learning experience using the baseline and FitVid. (*: $p < 0.05$, **: $p < 0.01$)

manipulation feature explained they did not feel the need for it since the adapted content was readable enough. *"There was no huge need for additional resizing or repositioning. The design (of the adapted content) was sufficient for me to read and understand the lecture content on my smartphone screen."* (P11).

Improving concentration (18/24 participants) Notable feedback from participants was that enabling interactions while watching a video itself helps them concentrate on the lecture. They slightly repositioned the elements or merely touched the element with the edges highlighted. *"I like that the elements are highlighted when I touch them. Reactions to my touch interactions make me recognize what content I'm focusing on and processing."* (P16). When asked about the reasons behind the manipulations they made while watching a video, the participants elaborated that they unconsciously touch the elements they are focusing on simply for the feeling of interaction. Hence, the qualitative feedback from the users supports H3.

We also found that using the manipulation feature in real-time while watching a video does not pose additional cognitive load to participants ($Z = 99.0$, $p = 0.37$), showing no significant difference in the easiness of use compared to the baseline interface with the use of the new feature ($Z = 68.5$, $p = 0.5$). On the other hand, the negative feedback for the manipulation feature includes unintended interactions. One participant noted that *"It was confusing when unwanted elements are touched."* (P18).

Customized Content Adaptation The participants all expressed positive feedback on the customized content adaptation of the font sizes, focusing mainly on the advantages of having tailored design settings based on their preferences. *"If I enlarged font sizes, then it implies that the original font size was too small for me. I like that it automatically captures my preference."* (P15). On the other hand, some participants pointed out the importance of consideration for the context and the need for user control. *"My preference would differ depending on the types of lecture content. It would be nice if the system can also consider such context."* (P10). *"Although the automated content adaptation is convenient, I hope I can change the font size manually if needed."* (P19).

6 Discussion and Limitations

We discuss findings, limitations, and possible extensions of this work.

Responsive Video Content Adaptation

The result of the user study showed that responsive content adaptation enhanced the video-based learning experience with improved readability and levels of concentration. Most of the participants evaluated the automated content adaptation as sufficiently readable. While the current work focused on text and images as the most basic and major objects as an initial step for the automated content adaptation, future work can adjust video content more flexibly by applying a responsive and flexible layout such as Apple's AutoLayout (Sadun 2013) or CSS FlexBox (W3Schools 2020 (accessed September 10, 2020), and even adapting to different orientations (portrait and landscape). We also plan to further investigate design features such as line spacing, font styles, and font colors.

Extended Content Customization by Direct Manipulation

The user study result revealed clear needs for content customization. Some participants did not manipulate the adapted content since the automated content adaptation is sufficient for them, while other participants manipulated them to suit their preference. It is important to give users control over machine-generated results, instead of fully automating the content adaptation process that lacks tailoring and customization. This control enabled them to refine the incomplete results provided by FitVid. Future interaction mechanisms can

be extended to a hybrid workflow between humans and machines.

A possible extension of the direct manipulation in FitVid includes add and delete interactions through which learners can adjust and select the amount of elements to be displayed. The added interaction can allow learners to interactively access the original content before adaptation. On the other hand, some participants mentioned the need for a lock feature which disables touch interactions including resizing and repositioning, since unintended interactions occurred when they touch the screen for other purposes such as video navigation.

We also identified notable patterns and strategies for using the direct manipulation feature. One of the unexpected benefits of the manipulation was that the interaction with the content itself without pragmatic goals helps learners increase the levels of attention. They sometimes touched or moved the in-video elements without a specific manipulation intent. This behavior can be seen as analogous to text highlighting. Highlighting while reading text is a common activity and is known to improve comprehension of content (Fowler and Barker 1974; Gowases, Bednarik, and Tukiainen 2011). In this regard, future work can explore additional interactions such as enabling haptic pen writing on a video or highlighting the region of interest, which can improve cognition and attention.

AI-based Content Adaptation

In this research, we developed a prototype which generates customized content adaptation for font sizes. The system improved the default content adaptation by using the tailored font sizes based on the users' manipulation log. All the participants expressed positive feedback on the customized content adaptation, however they also pointed out the need for consideration of various contexts such as characteristics of lecture content and learning environment. The participants expected AI to generate customized content designs that reflect their preferences, while they want to control the design settings via manual adjustments. Future work can apply a mixed-initiative approach (Horvitz 1999) to iteratively improve content adaptation depending on individual learners' context, which involves learners in-the-loop around AI. We envision that the tailored context-aware content adaptation could provide enhanced learning experiences by minimizing the extraneous cognitive load caused by device environments and capacities of designers.

7 Conclusion

In this paper, we have introduced FitVid, a video interface that enables automated content adaptation and direct manipulation for mobile learning environments. We develop an automated pipeline which adapts learning video content to mobile devices by segmenting and resizing the in-video elements. The pipeline-generated content shows the improved compliance rate for the design guidelines, from 2% to 89% for the font size, and 67% to 87% for the word count. The user evaluation illustrated that FitVid improves the learning experience with increased readability and levels of concentration. We expect to apply the techniques suggested in this paper to support the video-based learning ecosystem which encompasses instructors, engineers, and learners.

8 Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2020R1C1C1007587).

References

Alamri, A.; Sun, Z.; Cristea, A. I.; Senthilnathan, G.; Shi, L.; and Stewart, C. 2020. Is MOOC Learning Different for Dropouts? A Visually-Driven, Multi-granularity Explanatory ML Approach. In *International Conference on Intelligent Tutoring Systems*, 353–363. Springer.

Ally, M. 2005. Using learning theories to design instruction for mobile learning devices. *Mobile learning anytime everywhere* 5–8.

Alred, G. J.; Brusaw, C. T.; and Walter, E. O. ???? Handbook of Technical Writing. Bedford/St. Martins, Boston, MA, USA, 2006. *paperback), 0-312-35267-0 (hardcover). xxiv* .

Axel, C.; Ravindra, G.; and Tsang, O. W. 2010. Towards characterizing users' interaction with zoomable video. In *Proceedings of the 2010 ACM workshop on Social, adaptive and personalized multimedia interaction and access*, 21–24.

Bhuttoo, V.; Soman, K.; and Sungkur, R. K. 2017. Responsive design and content adaptation for e-learning on mobile devices. In *2017 1st International Conference on Next Generation Computing Applications (NextComp)*, 163–168. IEEE.

Brock, S.; and Joglekar, Y. 2011. Empowering PowerPoint: Slides and teaching effectiveness. *Interdisciplinary Journal of Information, Knowledge, and Management* 6(1): 85–94.

Carlier, A.; Ravindra, G.; Charvillat, V.; and Ooi, W. T. 2011. Combining content-based analysis and crowdsourcing to improve user interaction with zoomable video. In *Proceedings of the 19th ACM international conference on Multimedia*, 43–52.

Carlon, M. K. J.; Keerativoranan, N.; and Cross, J. S. 2020. Content Type Distribution and Readability of MOOCs. In *Proceedings of the Seventh ACM Conference on Learning@ Scale*, 401–404.

Casares, J.; Long, A. C.; Myers, B. A.; Bhatnagar, R.; Stevens, S. M.; Dabbish, L.; Yocum, D.; and Corbett, A. 2002. Simplifying video editing using metadata. In *Proceedings of the 4th conference on Designing interactive systems: processes, practices, methods, and techniques*, 157–166.

Cavanaugh, T.; and Cavanaugh, C. 2000. Interactive PowerPoint for teachers and students. In *Society for Information Technology & Teacher Education International Conference*, 496–499. Association for the Advancement of Computing in Education (AACE).

Christian Davidson, H.; and Wiggins, R. H. 2003. Radiology teaching presentation tools. In *Seminars in ultrasound, CT, and MR*, volume 24, 420–427.

Crossley, S. A.; Kyle, K.; and McNamara, D. S. 2016. The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior research methods* 48(4): 1227–1237.

DeWaard, I.; Abajian, S.; Gallagher, M. S.; Hogue, R.; Keskin, N.; Koutropoulos, A.; and Rodriguez, O. C. 2011. Using mLearning and MOOCs to understand chaos, emergence, and complexity in education. *International Review of Research in Open and Distributed Learning* 12(7): 94–115.

Dragicevic, P.; Ramos, G.; Bibliowitcz, J.; Nowrouzezahrai, D.; Balakrishnan, R.; and Singh, K. 2008. Video browsing by direct manipulation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 237–246.

Fowler, R. L.; and Barker, A. S. 1974. Effectiveness of highlighting for retention of text material. *Journal of Applied Psychology* 59(3): 358.

GoogleLLC. 2020 (accessed September 10, 2020)a. *Google Speech-To-Text*. URL https://cloud.google.com/speech-to-text/.

GoogleLLC. 2020 (accessed September 10, 2020)b. *The type system (Material Design)*. URL https://material.io/design/typography/the-type-system.html#type-scale.

Gowases, T.; Bednarik, R.; and Tukiainen, M. 2011. Text highlighting improves user experience for reading with magnified displays. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*, 1891–1896.

Halamish, V. 2018. Can very small font size enhance memory? *Memory & cognition* 46(6): 979–993.

Havice, P. A.; Davis, T. T.; Foxx, K. W.; and Havice, W. L. 2010. The impact of rich media presentations on a distributed learning environment: Engagement and satisfaction of undergraduate students. *Quarterly Review of Distance Education* 11(1): 53.

Hoffswell, J.; Li, W.; and Liu, Z. 2020. Techniques for Flexible Responsive Visualization Design. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13.

Holzl, J. 1997. Twelve tips for effective PowerPoint presentations for the technologically challenged. *Medical Teacher* 19(3): 175–179.

Horvitz, E. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, 159–166.

Hua, X.-S.; Wang, Z.; and Li, S. 2005. LazyCut: content-aware template-based video authoring. In *Proceedings of the 13th annual ACM international conference on Multimedia*, 792–793.

Inc., A. 2020 (accessed September 10, 2020). *Typography (Human Interface Guidelines)*. URL https://developer.apple.com/design/human-interface-guidelines/ios/visual-design/typography/.

Jung, H.; Shin, H. V.; and Kim, J. 2018. DynamicSlide: Exploring the Design Space of Reference-based Interaction Techniques for Slide-based Lecture Videos. In *Proceedings of the 2018 Workshop on Multimedia for Accessible Human Computer Interface*, 33–41.

Karrer, T.; Weiss, M.; Lee, E.; and Borchers, J. 2008. DRAGON: a direct manipulation interface for frame-accurate in-scene video navigation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 247–250.

Karrer, T.; Wittenhagen, M.; and Borchers, J. 2009. Pocketdragon: a direct manipulation video navigation interface for mobile devices. In *Proceedings of the 11th International Conference on Human-Computer Interaction with Mobile Devices and Services*, 1–3.

Karuna, P.; Purohit, H.; Uzuner, O.; Jajodia, S.; and Ganesan, R. 2018. Enhancing Cohesion and Coherence of Fake Text to Improve Believability for Deceiving Cyber Attackers. In *Proceedings of the First International Workshop on Language Cognition and Computational Models*, 31–40.

Kukulska-Hulme, A.; and Traxler, J. 2007. Designing for mobile and wireless learning. *Rethinking pedagogy for a digital age: Designing and delivering e-learning* 180–192.

Larocque, N.; Kenny, S.; and McInnes, M. D. 2015. Medical school radiology lectures: what are determinants of lecture satisfaction? *American Journal of Roentgenology* 204(5): 913–918.

Leahy, W.; and Sweller, J. 2005. Interactions among the imagination, expertise reversal, and element interactivity effects. *Journal of Experimental Psychology: Applied* 11(4): 266.

Lee, H.; Plass, J. L.; and Homer, B. D. 2006. Optimizing cognitive load for learning from computer-based science simulations. *Journal of educational psychology* 98(4): 902.

Lewis, P. J. 2016. Brain friendly teaching—reducing learner's cognitive load. *Academic radiology* 23(7): 877–880.

Liu, S.-H.; Liao, H.-L.; and Pratt, J. A. 2009. Impact of media richness and flow on e-learning technology acceptance. *Computers & Education* 52(3): 599–607.

Long, A. C.; Myers, B.; Casares, J.; Stevens, S.; and Corbett, A. 2004. Video Editing Using Lenses and Semantic Zooming .

Lu, H.; Shen, S.-s.; Shiang, S.-R.; Lee, H.-y.; and Lee, L.-s. 2014. Alignment of spoken utterances with slide content for easier learning with recorded lectures using structured support vector machine (SVM). In *Fifteenth Annual Conference of the International Speech Communication Association*.

Monserrat, T.-J. K. P.; Zhao, S.; McGee, K.; and Pandey, A. V. 2013. NoteVideo: facilitating navigation of blackboard-style lecture videos. In *Proceedings of the SIGCHI conference on human factors in computing systems*, 1139–1148.

O'Malley, C.; Vavoula, G.; Glew, J.; Taylor, J.; Sharples, M.; Lefrere, P.; Lonsdale, P.; Naismith, L.; and Waycott, J. 2005. Guidelines for learning/teaching/tutoring in a mobile environment .

O'Neill-Jones, P. 2004. Bringing media rich content to online learning. In *E-Learn: World Conference on E-Learning*

in Corporate, Government, Healthcare, and Higher Education, 155–158. Association for the Advancement of Computing in Education (AACE).

Pang, D.; Halawa, S.; Cheung, N.-M.; and Girod, B. 2011. Mobile interactive region-of-interest video streaming with crowd-driven prefetching. In *Proceedings of the 2011 international ACM workshop on Interactive multimedia on mobile and portable devices*, 7–12.

Parsons, D.; Ryu, H.; and Cranshaw, M. 2006. A study of design requirements for mobile learning environments. In *Sixth IEEE International Conference on Advanced Learning Technologies (ICALT'06)*, 96–100. IEEE.

Peng, W.; and Zhou, Y. 2015. The design and research of responsive web supporting mobile learning devices. In *2015 International Symposium on Educational Technology (ISET)*, 163–167. IEEE.

Pugsley, L. 2010. How To... Design an effective power point presentation. *Education for Primary Care* 21(1): 51–53.

PythonSoftwareFoundation. 2020 (accessed September 10, 2020). *pytesseract 0.3.6*. URL https://pypi.org/project/pytesseract/.

Quang Minh Khiem, N.; Ravindra, G.; Carlier, A.; and Ooi, W. T. 2010. Supporting zoomable video streams with dynamic region-of-interest cropping. In *Proceedings of the first annual ACM SIGMM conference on Multimedia systems*, 259–270.

Reimers, N.; and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* .

Rhodes, M. G.; and Castel, A. D. 2008. Memory predictions are influenced by perceptual information: evidence for metacognitive illusions. *Journal of experimental psychology: General* 137(4): 615.

Sadun, E. 2013. *iOS Auto Layout Demystified*. Addison-Wesley Professional.

Seale, A. C.; Ibeto, M.; Gallo, J.; de Waroux, O. l. P.; Glynn, J. R.; and Fogarty, J. 2020. Learning from each other in the COVID-19 pandemic. *Wellcome Open Research* 5(105): 105.

Shin, H. V.; Berthouzoz, F.; Li, W.; and Durand, F. 2015. Visual transcripts: lecture notes from blackboard-style lecture videos. *ACM Transactions on Graphics (TOG)* 34(6): 1–10.

Shneiderman, B. 1997. Direct manipulation for comprehensible, predictable and controllable user interfaces. In *Proceedings of the 2nd international conference on Intelligent user interfaces*, 33–39.

Shneiderman, B.; and Maes, P. 1997. Direct manipulation vs. interface agents. *interactions* 4(6): 42–61.

Song, W.; Tjondronegoro, D. W.; Wang, S.-H.; and Docherty, M. J. 2010. Impact of zooming and enhancing region of interests for optimizing user experience on mobile sports video. In *Proceedings of the 18th ACM international conference on Multimedia*, 321–330.

Stanton, G.; and Ophoff, J. 2013. Towards a method for mobile learning design. In *Proceedings of the informing science and information Technology education conference*, 501–523. Informing Science Institute.

Stein, K. 2006. The dos and don'ts of PowerPoint presentations. *Journal of the American Dietetic Association* 106(11): 1745–1748.

Sweller, J. 1994. Cognitive load theory, learning difficulty, and instructional design. *Learning and instruction* 4(4): 295–312.

Sweller, J. 2010. Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational psychology review* 22(2): 123–138.

Sweller, J.; Van Merrienboer, J. J.; and Paas, F. G. 1998. Cognitive architecture and instructional design. *Educational psychology review* 10(3): 251–296.

Tsujimura, S.; Yamamoto, K.; and Nakagawa, S. 2017. Automatic Explanation Spot Estimation Method Targeted at Text and Figures in Lecture Slides. In *INTERSPEECH*, 2764–2768.

W3Schools. 2020 (accessed September 10, 2020). *CSS Flexbox*. URL https://www.w3schools.com/css/css3_flexbox.asp.

Ward, A.; and Litman, D. 2008. Semantic cohesion and learning. In *International Conference on Intelligent Tutoring Systems*, 459–469. Springer.

Wu, A.; Tong, W.; Dwyer, T.; Lee, B.; Isenberg, P.; and Qu, H. 2020. MobileVisFixer: Tailoring Web Visualizations for Mobile Phones Leveraging an Explainable Reinforcement Learning Framework. *arXiv preprint arXiv:2008.06678* .

Xu, C.; Wang, R.; Lin, S.; Luo, X.; Zhao, B.; Shao, L.; and Hu, M. 2019. Lecture2Note: Automatic Generation of Lecture Notes from Slide-Based Educational Videos. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, 898–903. IEEE.

Zhao, B.; Xu, S.; Lin, S.; Wang, R.; and Luo, X. 2019. A New Visual Interface for Searching and Navigating Slide-Based Lecture Videos. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, 928–933. IEEE.

Zhou, T.; Huang, S.; Cheng, J.; and Xiao, Y. 2020. The Distance Teaching Practice of Combined Mode of Massive Open Online Course Micro-Video for Interns in Emergency Department During the COVID-19 Epidemic Period. *Telemedicine and e-Health* 26(5): 584–588.