STUDY ON AUTOMATIC GENERATION OF LECTURE VIDEOS BASED ON CONTENT ANALYSIS OF LECTURE SLIDES

Kota MIZUKAMI, Daisuke DEGUCHI, Tomoki TODA, Hiroshi MURASE
Nagoya University

Haruya KYUTOKU
Aichi University of Technology

Tsubasa MINEMATSU
Kyushu Institute of Technology

ABSTRACT

With the spread of online education, the demand for lecture videos in which instructors read aloud from slide decks has been increasing. However, creating such videos entails significant labor and cost. This study proposes a method that automatically generates lecture videos—including narration audio and highlighted text—using the slides themselves as input. First, the slide content is analyzed with large language models and optical character-recognition. Based on the extracted information, narration speech is synthesized and key text within the slides is highlighted to enhance learners' understanding. Experiments using real presentation slides confirmed that the proposed method can automatically generate lecture videos.

KEYWORDS

Automatic Lecture Video Generation, Text Highlighting, Large Language Model

1. INTRODUCTION

Online and hybrid learning have greatly increased the need for video-based teaching materials.
Lecture videos—where instructors read aloud from slides—can enhance learning by combining verbal and visual information (Mayer and Anderson, 1991). However, producing such content requires significant time for recording and editing. To reduce this burden, we propose a method that automatically generates lecture videos from slide decks. The system analyzes slide content using language and image processing, then outputs a narrated slide show with synchronized highlights. This automation benefits both academic and corporate education. Our method consists of: (i) generating narration text, (ii) synthesizing narration audio, and (iii) highlighting the narrated slide content. Section 2 reviews related work; Section 3 explains our method; Section 4 presents experiments; Section 5 concludes.

2. RELATED WORK

Ando and Ueno (2011) showed that synchronizing pointers with narration and visual elements enhances learner comprehension by directing attention to relevant screen areas. This supports the use of pointer-based emphasis in AI-generated lecture materials.
Xu et al. (2024) combined OCR, LLMs, and TTS to construct a pipeline that produces lecture videos in which an avatar delivers synthetic speech based on slide content. A user study comparing these AI-generated lectures to conventional recordings suggested improved memory retention in English language learning. However, their system lacks the mechanisms for highlighting slide elements.
Optical character recognition (OCR) extracts textual content from images. Modern approaches like CRAFT (Baek et al., 2019) apply deep learning to enhance detection accuracy.

Text embedding represents text as numerical vectors in a semantic space, allowing similarity comparisons. Sentence-BERT (Reimers & Gurevych, 2019) is widely used for this purpose.

Large language models (LLMs) are pretrained neural networks capable of handling diverse natural-language tasks. Models like ChatGPT also support multimodal inputs (OpenAI, 2022).

Text-to-speech (TTS) systems synthesize speech from text. Modern TTS services such as Google Text-to-Speech (gTTS) use deep learning for natural-sounding output (Google, 2025).

3. PROPOSED METHOD

We propose a pipeline that takes a slide deck as input and automatically (i) generates narration text and speech, (ii) highlights relevant text on each slide, and (iii) produces a lecture video combining these elements. As shown in Figure 1, each slide is rasterized into an image, denoted I_i, and processed using OCR. Narration is generated via ChatGPT, speech via TTS, and highlights are selected by comparing OCR text with the narration using Sentence-BERT embeddings. Details of each stage are described in Sections 3.1–3.3.

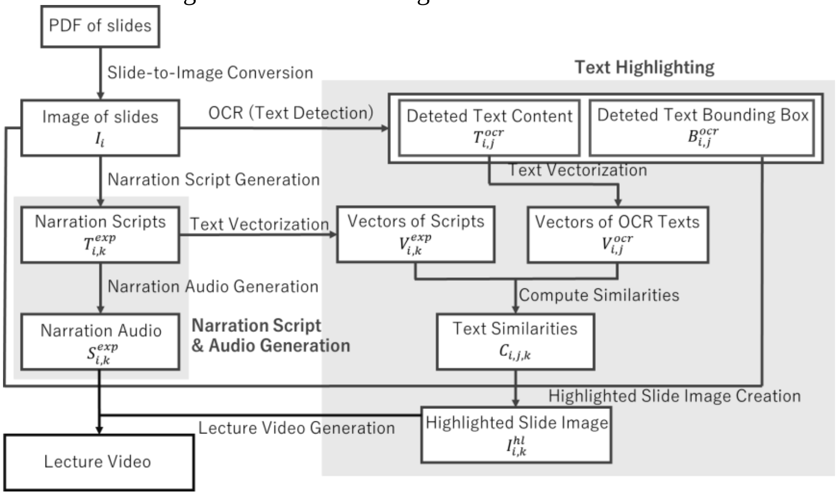

PDF of slides
Slide-to-Image Conversion
Text Highlighting
Image of slides I_i
OCR (Text Detection)
Deteted Text Content $T^{ocr}_{i,j}$
Deteted Text Bounding Box $B^{ocr}_{i,j}$
Text Vectorization
Narration Script Generation
Narration Scripts $T^{exp}_{i,k}$
Text Vectorization
Vectors of Scripts $V^{exp}_{i,k}$
Vectors of OCR Texts $V^{ocr}_{i,j}$
Narration Audio Generation
Compute Similarities
Narration Audio $S^{exp}_{i,k}$
Narration Script & Audio Generation
Text Similarities $C_{i,j,k}$
Highlighted Slide Image Creation
Lecture Video Generation
Highlighted Slide Image $I^{hl}_{i,k}$
Lecture Video

Figure 1. The workflow of proposed method.

3.1 Generation of Narration Script and Audio

To obtain the script spoken in the lecture video, each image of slides is supplied to ChatGPT. ChatGPT is prompted with the following three-stage sequence:

- **Precondition confirmation and overall task instructions** – a prompt provides ChatGPT with information essential for narration generation and an overview of the entire task.
- **Grasping the entire slide deck** – all slide images are then presented so that ChatGPT can understand the structure and progression of the presentation and devise a strategy that avoids biased or redundant narration.
- **Generating narration for each page** – finally, the slide images are fed one by one, and ChatGPT outputs the narration text for each page.

Table 1 lists the full prompt sequence used to instruct ChatGPT in the narration-generation process.

Table 1. The full prompt sequence.

Stage	Prompt Text	Associated Input
Precondition confirmation and overall task	You are a university professor. Please think about what to say in the class using the slides I will give you. Please meet the following conditions.	No input

instructions	1. Return only the content to be spoken 2. Do not include control characters such as titles and line breaks 3. Combine the amount of information in the slides with the amount of information in the text. For example, do not speak for a long time on title slides or scene transition slides. 4. In the closing slide, summarize what has been said so far. 5. Be aware that research and other materials in the document may not be the presenter's achievements. 6. Pay special attention to numbers and letters in formulas and tables. 7. Do not speak about things that are not written on the slide. 8. Do not read the page number of the slide.	
Grasping the entire slide deck	First, I will give you all the pages of the slides. Please make a plan for what to say while minimizing what you say.	Entire slide PDF
Generating narration for each page	This is page {page_num}. Please give a presentation while looking at this page.	Image of each page

As a result, executing these steps for the i-th slide image I_i yields the page-level narration text T_i^{exp}. And then, T_i^{exp} is split at punctuation; each fragment $T_{i,k}^{exp}$ is converted to speech via TTS, yielding $S_{i,k}^{exp}$.

3.2 Text Highlighting

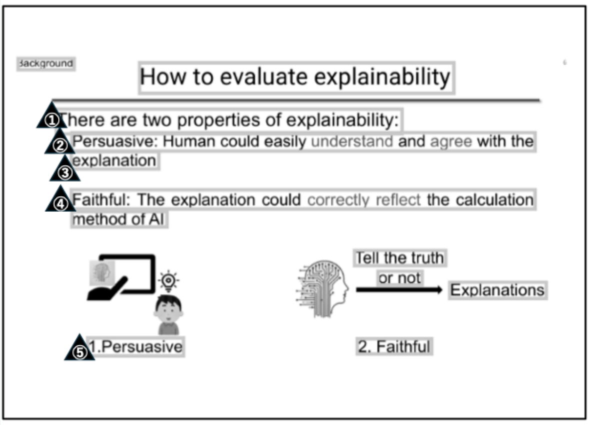
First, OCR detects all text regions on I_i, obtaining both the bounding boxes $B_{i,j}^{ocr}$ and their corresponding strings $T_{i,j}^{ocr}$. Next, Sentence-BERT embeds $T_{i,k}^{exp}$ and $T_{i,j}^{ocr}$ to produce vectors $V_{i,k}^{exp}$ and $V_{i,j}^{ocr}$. Then, for each narration fragment $T_{i,k}^{exp}$ on i-th slide, we compute cosine similarities $C_{i,j,k} = \cos(V_{i,k}^{exp}, V_{i,j}^{ocr})$ between its embedding $V_{i,k}^{exp}$ and the embeddings $V_{i,j}^{ocr}$ of all OCR-detected strings on the same slide. Among all candidate strings, only the one with the highest cosine similarity is selected as the highlighted text. Formally, we define $j_{i,k}^{*} = \arg\max_{j} C_{i,j,k}$: the index of the single text region with the highest similarity. Finally, the bounding box B_{i,j^*}^{ocr} chosen in the previous step is drawn onto the original slide image I_i to produce the highlighted frame $I_{i,k}^{hl}$.

3.3 Video Generation

All speech segments $\mathrm{S}_{\mathrm{i,k}}^{\mathrm{exp}}$ and highlighted images $\mathrm{I}_{\mathrm{i,k}}^{\mathrm{hl}}$ are concatenated to produce the final lecture video that synchronizes narration with visual emphasis.

4. EXPERIMENTS

To evaluate the practical effectiveness of our method, we conducted a focused case study using a representative English-language slide. We extracted text regions from the slide using optical character recognition (OCR) and assigned each region to a unique bounding box (BBox) number. For each narration fragment generated by ChatGPT, our method selected the most semantically similar OCR region based on Sentence-BERT cosine similarity. Figure 2 shows the slide image with numbered bounding boxes, along with a table listing each narration fragment and its corresponding BBox number. We subjectively evaluated whether highlighted regions matched the narration. In general, they were judged contextually appropriate. However, in cases where no appropriate pointing target exists on the slide for a given narration fragment, the system still selects a text region to highlight. For example, as shown in Fig.2. short narration text “First,” was incorrectly assigned to Slide component (5). This behavior may distract learners by drawing attention to irrelevant content, revealing a limitation of the proposed method. Therefore, future improvements should consider allowing the system to refrain from highlighting when no semantically appropriate region is found.



Background

How to evaluate explainability

① There are two properties of explainability:

② Persuasive: Human could easily understand and agree with the ③ explanation

④ Faithful: The explanation could correctly reflect the calculation method of AI

Narration Fragment	Highlighted OCR Text
Evaluating explainability involves two key properties.	①
First,	⑤
persuadability ensures that humans can easily understand and agree with AI explanations.	②
Second,	③
Faithfulness checks if the explanations accurately reflect the AI's actual computational methods.	④
Both aspects are essential for building trust in AI systems.	④

Figure 2. Slide image annotated with OCR-detected text regions (BBox numbers) and their corresponding narration fragments generated by ChatGPT. The table on the right shows each narration fragment and the matched bounding box ID selected by the proposed method.

5. CONCLUSION

We proposed a method that takes lecture slides as input, analyzes their content with an LLM and OCR, generates narrated speech and text highlights, and automatically outputs a lecture video. Experiments confirmed that the proposed method can generate lecture videos automatically. Future work includes extending the highlight-generation process to support not only text but also figures and tables. Additionally, we aim to develop a method that incorporates temporal constraints—such as avoiding repeated pointing to content that has already been explained—to improve the coherence and instructional quality of the generated video. Furthermore, we plan to conduct user studies to evaluate the perceived clarity, usefulness, and educational effectiveness of the generated videos.

ACKNOWLEDGEMENT

This research was partially supported by CREST, JPMJCR22D1.

REFERENCES

Ando M. and Ueno M., 2011. Effect analysis of pointer presentations on multimedia e-learning materials based on dual channel model. *Educational Technology Research*, 34(1-2), pp.59-73.

Baek Y. et al., 2019. Character Region Awareness for Text Detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, USA, pp. 9365–9374.

Google, 2025. Google Cloud Text-to-Speech. Available at: https://cloud.google.com/text-to-speech (Accessed 16 February 2025).

Mayer R.E. and Anderson R.B., 1991. Animations need narrations: An experimental test of a dual-coding hypothesis. *Journal of Educational Psychology*, 83 (4), pp. 484–490.

OpenAI, 2022. ChatGPT: Optimizing Language Models for Dialogue. Available at: https://openai.com/blog/chatgpt (Accessed 16 February 2025).

Reimers N. and Gurevych I., 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China, pp. 3982–3992.

Xu T. et al., 2024. From recorded to AI-generated instructional videos: A comparison of learning performance and experience. *British Journal of Educational Technology (Early View, published 28 October 2024)*. DOI: 10.1111/bjet.13530.