Concept and Initial Learning Log Analysis for Lecture Archive Summarization Platform

Shinobu Hasegawa
Center for Innovative Distance Education and Research, Japan Advanced Institute of Science and Technology,
Ishikawa, Japan
hasegawa@jaist.ac.jp

Xiaoting Liu
Division of Advanced Science and Technology, Japan Advanced Institute of Science and Technology
Ishikawa, Japan
liuxiaoting@jaist.ac.jp

Wen Gu
Center for Innovative Distance Education and Research, Japan Advanced Institute of Science and Technology,
Ishikawa, Japan
wgu@jaist.ac.jp

Koich Ota
Center for Innovative Distance Education and Research, Japan Advanced Institute of Science and Technology,
Ishikawa, Japan
ota@jaist.ac.jp

***Abstract*— The final objective of this research project is to develop a lecture archive summarization platform that can extend learners' experience by automatically estimating and providing temporal and spatial ROI (Regions of Interest) according to multimodal features and learners' learning logs in lecture archives that record face-to-face lectures. To develop this platform, we (a) establish a method for extracting spatiotemporal multimodal features of lecture archives and (b) construct a method for estimating a learning style model based on the learning logs when watching the archives with the extracted features. Furthermore, to maximize the learning effect, we will (c) develop a prototype system to adaptively control the spatiotemporal ROI at the terminal side according to the learning style model as an adaptive summarization. This article describes the concept of the proposed platform and the initial analysis of learners' learning logs.**

Keywords—lecture archive summarization, multimodal features, learning log analysis, learning style model, adaptation

I. INTRODUCTION

The spread of COVID-19 was a great challenge to educational institutions to transfer conventional face-to-face (F2F) lectures to synchronous or asynchronous online lectures and accommodate hybrid lectures that learners can selectively attend according to their situations [1]. Such a flexible learning environment can be an essential foundation for future higher education with a perspective to the after-COVID-19. In particular, lecture archives, in which F2F lectures are recorded by a fixed and high-resolution camera and microphone delivered without editing, are relatively easy to introduce and deploy regarding content production costs [2].

However, it is difficult for learners to learn lecture archives with high engagement compared to the lecture videos that are created or edited in consideration of the following regions of interest (ROI) of learners [3].

- Temporal ROIs: Scenes in the archives that learners should/want to watch.
- Spatial ROIs: Objects such as instructors, blackboards, slides, etc., at a specific scene.

Therefore, developing an automated lecture archive summarization platform that considers such ROIs is essential to utilize recorded and stored archives effectively.

For the temporal ROIs, one of the previous studies proposed a lecture archive summarization approach using natural language processing (NLP) based on an attention-based recurrent neural network (RNN) [4]. We have also developed an estimation method for the temporal ROIs in lecture archives by using the instructor's actions, voice intensity, and slide differences as multimodal features [5]. On the other hand, for the perspective of the spatial ROIs in lecture archives, a recording system has been proposed to generate content by switching the camera to pre-defined ROIs based on detecting the instructor's pose during the lecture [6]. Furthermore, we have defined the spatial ROIs based on the learner's gaze coordinates during learning and estimated the ROIs in the archives based on the combination of U-Net and ResNet [7].

However, lecture archives' ROIs differ due to differences in learners' learning styles (e.g., learning purposes, habits, and trends). In other words, it is difficult to satisfy the needs of each learner by delivering pre-summarized lecture archives. Resolving this issue is one of the important applications of the Next Generation of Affective Computing (NGAC).

The final objective of this research is to develop a platform that summarizes lecture archives and enhances the learner's learning experience by estimating the spatiotemporal ROIs based on individual learning logs for the archives. As the first output of this research project, this article surveys the related study in Section 2, then explains the concept of the proposed platform in Section 3, describes the initial analysis from learning logs in Section 4, and discusses future directions in Section 5.

II. RELATED STUDY

A. Lecture Archives

In this article, lecture archives are collections of video and audio recordings of F2F lectures held mainly in lecture rooms. Various approaches/styles have been proposed to realize lecture archives [8]. We have also installed a lecture

archiving system with a fixed camera and microphone at our institution since 2007. Considering cost, quality, and technical development, we currently operate a system that can automatically record and distribute video, audio, and PC screens, as shown in Fig. 1. More than 2,000 archives are distributed annually. While students respond that the system is adequate for reviewing F2F lectures, there are some problems, such as difficulty in reviewing essential parts of the lecture and maintaining engagement for learning because the lecture time (100 minutes) is too long. This research aims to address these issues with limited human and budget resources.

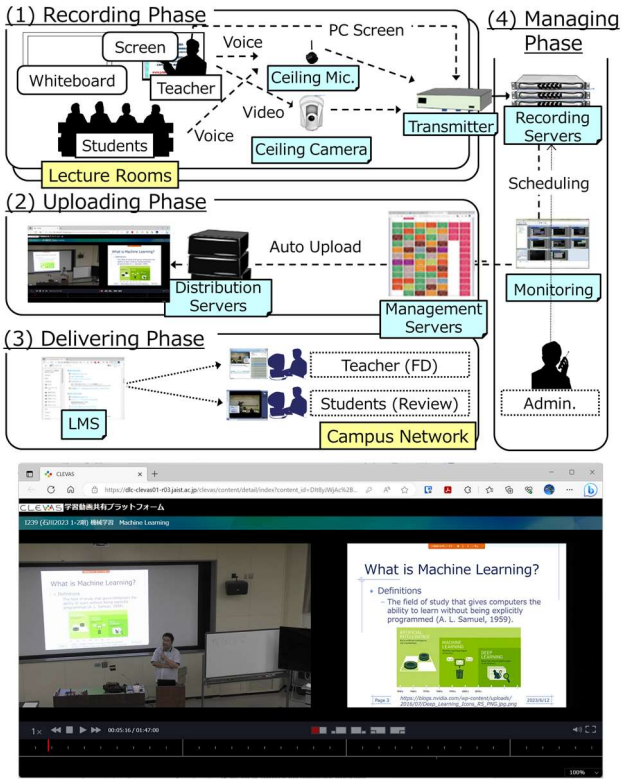

Fig. 1. Architecture of Lecture Archives and its Example

B. Related Study for Temporal ROI

Andra et al. proposed an attention-based RNN architecture with linguistic features from lecture transcripts to generate lecture archive summaries [4]. Kobayashi et al. also proposed a search function in presentation archives based on presentation structure and transcript [9]. These studies added textual indexes/metadata to archive information to identify the temporal ROIs. With the recent development of large-scale language models (LLMs), these methods may be effective for various subjects if good-quality lecture transcripts are available. However, recording with a fixed ceiling microphone in real-world situations is challenging to apply these approaches because of the various noises generated by the lecture room environment.

In our previous study, Sheng et al. proposed a deep neural network architecture for detecting the temporal ROIs from learning logs of online students from lecture archives. We divided lecture archives into 1-minute segments, and the number of times students accessed each segment from the learning management system (LMS) was counted as label data to define the ROIs [5]. Then, to improve detection reliability, we demonstrated a deep neural network architecture that combined feature maps for instructor behavior, voice intensity, and slide differences, as shown in Fig. 2. This made detecting specific ROIs with a small amount of computation possible without using semantic features. However, many issues remain to be solved to estimate the ROIs considering individual differences among learners.

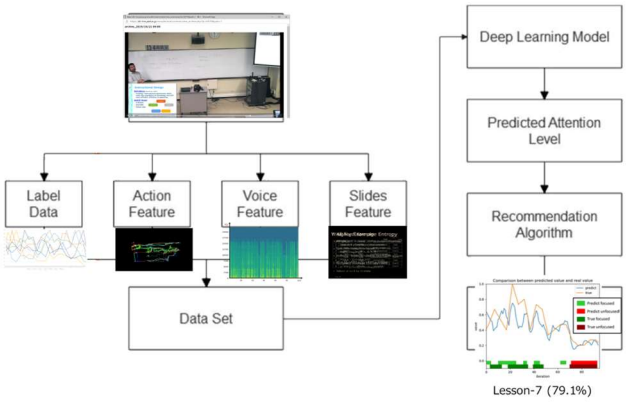

Fig. 2. Architecture of our Previous Temporal ROI Detection

C. Related Study for Spatial ROI

Previous studies have mainly focused on the spatial ROIs based on the instructor's location in the archives and applied virtual camera techniques to zoom in [10-12]. However, Zhang et al. pointed out that learners also pay attention to the blackboard or slides that the instructor indicates when they watch the lecture archives [13]. Hulens et al. employed special hardware integrated multiple cameras [6]. They switched among different focus areas, such as the instructor, blackboard, and screen, depending on the instructor's pose using a pre-trained neural network for pose recognition.

In our previous work, Yang et al. introduced an automatic method to predict the spatial ROIs from the instructor's behaviors in archived lectures using a deep neural network for smaller screens such as smart devices. They first obtained the spatial ROIs from the eye-tracking data of learners who watched the lecture archives for one-second video segments. Then, they extracted the feature maps of the instructor's behaviors from the segment, using frame difference, Optical Flow, OpenPose, and temporal features. They also designed an Encoder-Decoder architecture that integrated U-Net and ResNet with these behavioral features as input for ROI prediction, as shown in Fig. 3, and demonstrated its positive potential.

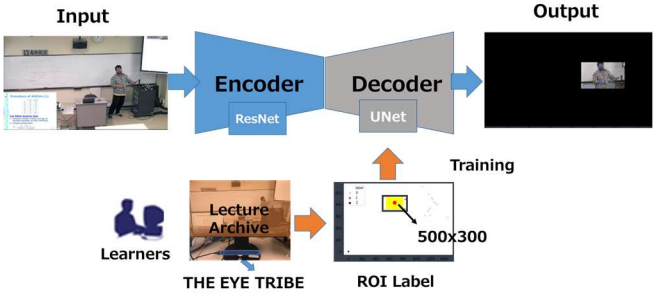

Fig. 3. Architecture of our Previous Spatial ROI Detection

III. Proposed Platform

In this study, we aim to develop a learning platform that extends our previous findings and introduces an adaptive

summarization for lecture archives at the terminal side by using estimated spatiotemporal ROIs based on the learning styles of individual learners. In particular, we hypothesize that each learner has different ROIs in lecture archives which will be determined from the learning logs, such as the learning timing and seeking operation, in addition to the multimodal features of the archives. This ongoing project proceeds in the following steps, as shown in Figure 4.

A. *Multimodal Feature Extraction*

(1) To extend the primary (F1) features proposed in the previous study [5], which consist of the instructor's actions, voice intensity, and slide differences, we are constructing a dataset to estimate the temporal ROIs in the archives, especially from the voice style perspectives [14].

(2) We are developing a multi-task deep learning architecture to estimate the instructor's poses (such as talking, writing, pointing, waiting, and walking), emotions (happiness, sadness, fear, anger, surprise, and disgust) or intentions (important, normal, and unimportant), lecture topics, etc., for a certain segment in the archives from these multimodal features and establish a framework to utilize the estimated results as secondary (F2) features.

B. *Learning Tendency Estimation*

(1) We are implementing an archive player that can record learners' operations, such as seeking, zooming in/out, etc., during watching the archives using HTML Living Standard and video.js. on a Moodle LMS and learning record store (LRS).

(2) To develop a method to estimate learning styles, we will propose a combination of machine/deep learning methods based on the bias of the distribution of individual learners' learning log data against the total logs and the multimodal features in (a).

C. *Adaptive Summarization Platform*

To maximize the learning effect of individual learners, we will implement a prototype system with an adaptive summarization mechanism that controls the spatiotemporal ROIs at the player side according to the estimation results of the learner's past learning style model (b) when watching new lecture archives and can update the style model based on operations during learning.

D. *Overall Characteristics*

The characteristics of this project are to integrate temporal and spatial ROIs and extract general-purpose multimodal features from lecture archives recorded with a fixed camera and microphone in a real environment (easy to record, but insufficient quality for speech recognition). Most previous studies on ROI estimation in lecture archives realize automatic summarization based only on the features extracted from the archives. However, our platform tries to combine learning log features obtained from learners and model their learning styles.

In addition, many conventional machine/deep learning methods assume that training and test data have the same distribution. Therefore, it is easy to reflect features common to all learners but difficult to reflect characteristics specific to each learner. In this study, we extend the idea of importance-weighted SVM[15], compensating for the bias in the distribution of training and test data on the kernel by covariate shift, to other machine/deep learning to realize personalized models.

IV. Initial Learning Log Analysis

An initial analysis is conducted to identify learners' learning styles from the learning logs stored in the platform. The target data are the access logs and grades of two archives, namely Lecture A and B, recorded F2F lectures on the "Machine Learning" course (held in Japanese, but materials in English) by one of the authors. The lecture archives were opened to the students around the evening of the day of the F2F lectures. 127 students attended the course, and 76 (40 Japanese and 36 international) students accessed the archives between the end of the F2F lectures and the exam. Because they were expected to attend the F2F lectures in principle, and the archives were a supplemental learning environment for reflection on the lectures. Although detailed data on the Japanese language proficiency of international students is unavailable, approximately half of them answered the examinations in English. The logs recorded the access date,

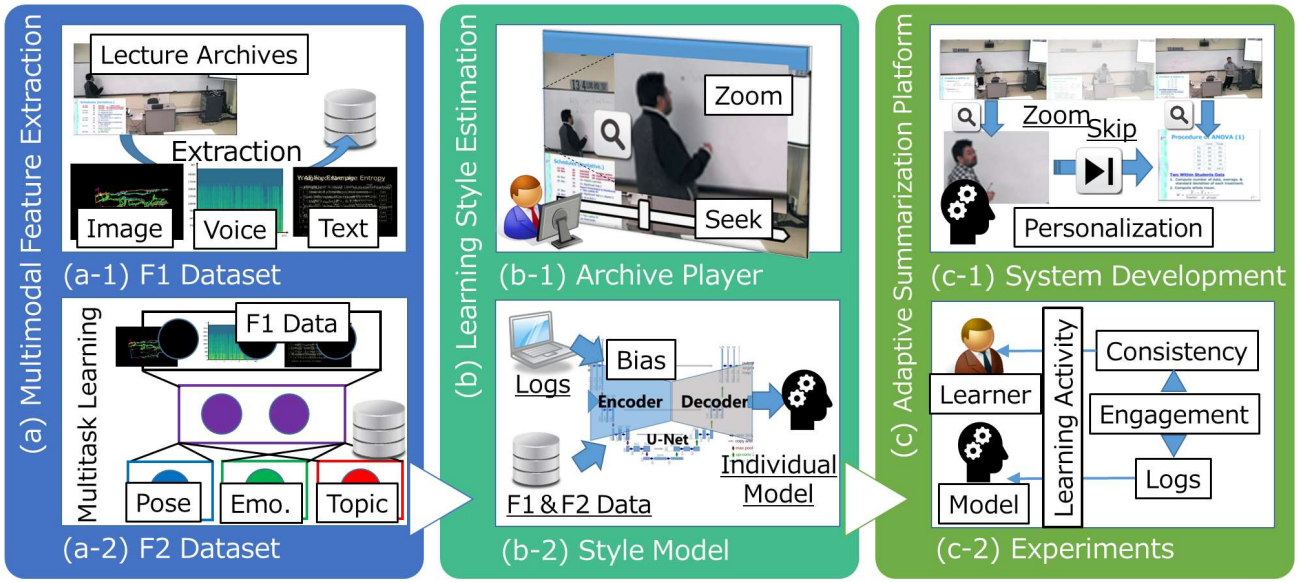

Fig.4. Research Procedure for Summarization Platform

time, and timeline watched in the archives.

A. Access to Lecture Archive Timeline

Fig. 5 and 6 show the access to the timeline of each lecture archive during the period. The horizontal axis shows the timeline in the archives. One lecture at our university is 100 minutes. However, the archive records from two minutes before the lecture starts to five minutes after the lecture ends, so the archives contain around 107 minutes. The vertical axis indicates the number of accesses to the timeline (upper graph) and the number of unique students who watched the timeline (lower graph). These results indicate that the distribution of accesses to the timeline of the archives was not significantly characterized. Therefore, we next attempted an analysis from the students' playback time perspective.

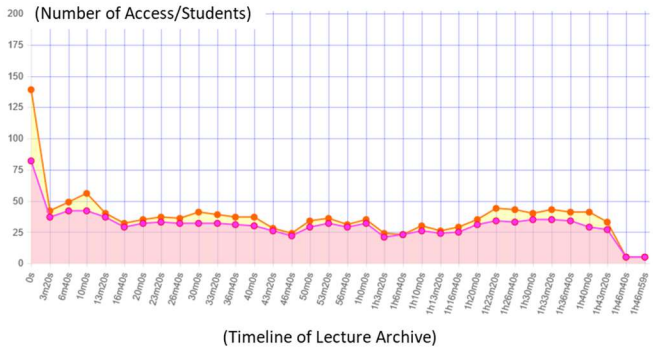

Fig.5. Number of Accesses/Students for Lecture A Timeline

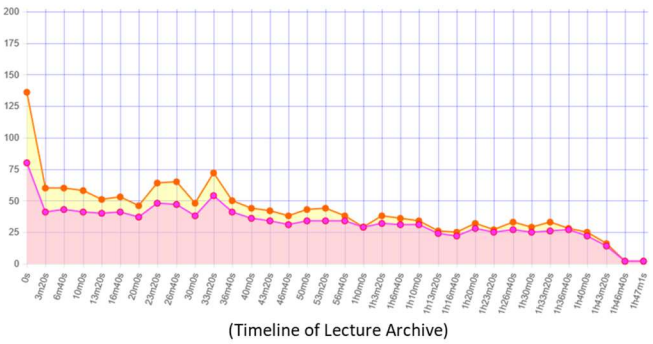

Fig.6. Number of Accesses/Students for Lecture B Timeline

B. Trends in Reflection from Playback Time/Rate

Fig. 7 and 8 show the correspondence between the playback time and rate for each lecture archive and the exam results (maximum score is 10, respectively). The horizontal axis indicates the playback rate, where 1 means the entire archive (around 107 mins.) was watched. The vertical axis is the playback time, which indicates the total watched/learning time in minutes. In other words, the data above the straight line in the figures show a trend of watching the same part of the archive multiple times, while the data below the straight line shows a tendency to watch the lecture by double-speed playback, etc.

The dots plotted in Fig. 7 and 8 indicate the exam scores corresponding to each lecture. The scores are categorized as 5 or less (low/red), 6 to 8 (medium/yellow), and 9 or more (high/blue) for Japanese (circle) and international (square) students. The results indicate that the trend of reflection using the lecture archives for both lectures is divided into two groups: those who watched less than 50% of the archives (short: only a part of the archives) and those who watched more than 50% of the archives (long: most of the archives).

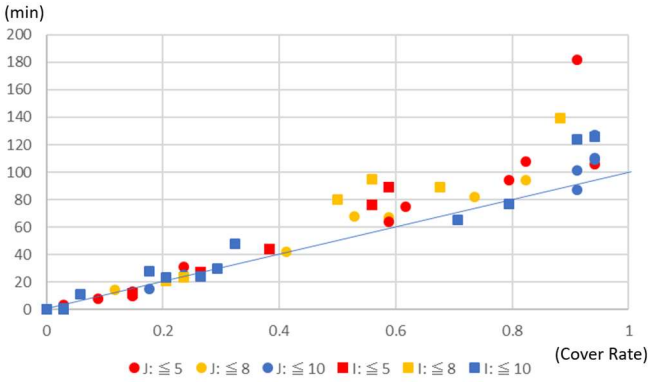

Fig.7. Playback Time/Rate and Exam Results for Lecture A

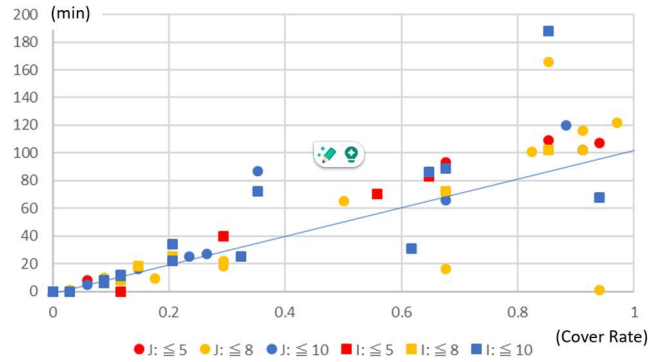

Fig.8. Playback Time/Rate and Exam Results for Lecture B

These results point to a relatively minor discrepancy between playback time and rate for the short playback rate group. In addition, Japanese students tend to have low and medium scores, while international students have medium and high scores, as summarized in TABLE I and II. On the other hand, there was no significant difference between Japanese and international students in the long playback rate group, but about half of them watched longer than their playback rate, and only a few students watched shorter, indicating that they watched the archives repeatedly. However, about 70% of students in this group received low or medium scores.

TABLE I. SUMMARY OF PLAYBACK RATE AND EXAM SCORES FOR LECTURE A

Rate	Time/ Rate	Japanese			International			# of Students
		L	M	H	L	M	H	
Short	< 120	8	4	2	4	8	9	35
	> 120	1	0	0	0	0	3	4
Long	< 120	3	0	3	0	4	2	12
	> 120	3	2	1	4	1	2	13
# of Students		15	6	6	8	13	16	64

In summary, student trends are divided into two groups with a playback rate of around 0.5 as a boundary. The group with a low playback rate includes students who focused on reviewing essential points in the archives. The overall

distribution of the exam scores was similar to the group with the high playback group. However, more than half of the students in the high playback rate group had a long playback time relative to the playback rate, and more than 70% had low or medium scores. These students were likely to watch most of the archives and repeat some parts of the lectures, possibly due to a considerable lack of understanding (or absence of F2F lectures).

TABLE II. SUMMARY OF PLAYBACK RATE AND EXAM SCORES FOR LECTURE B

Rate	Time/ Rate	Japanese			International			# of Students
		L	M	H	L	M	H	
Short	< 120	0	11	10	2	1	14	38
	> 120	1	0	0	1	2	1	5
Long	< 120	2	2	1	0	2	2	9
	> 120	2	5	1	2	0	2	12
# of Students		5	18	12	5	5	19	64

C. *Trends in Reflection from Playback Timing*

Fig. 9 and 10 show the date of the first access to each lecture archive on the horizontal axis and the playback time on the vertical axis. The meanings of the dots are the same as in Fig. 7 and 8. The earliest students accessed the archives immediately after their release. Access timing was also polarized, with one group accessing before the next lecture and another group accessing near the exam. However, the exam scores were similar between the two groups regarding access timing. However, when access timing and playback time were combined, we can find a high-scoring cluster that watched the archives briefly just before the exam, and low and medium-scoring groups watched the archives for around 100 minutes relatively early after the lectures. These characteristics help generate adaptive archive summaries.

Finally, the trends of repeated watching between Japanese and international students were similar. However, the international students performed relatively better than the Japanese students in the low playback group. Since the overall distribution of scores was also better for the international students than for the Japanese students, a more detailed analysis is needed to determine whether this effect is due to the watching behavior of the lecture archives.

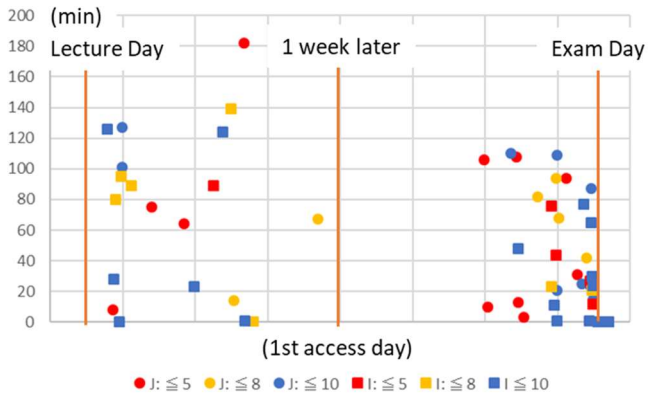

Fig.9. Access Day/Playback Time and Exam Results for Lecture A

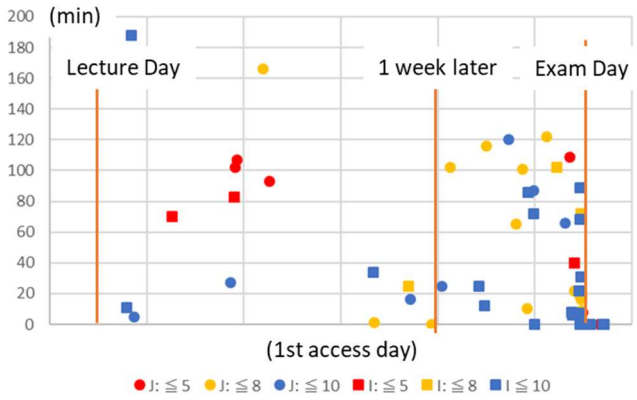

Fig.10. Access Day/Playback Time and Exam Results for Lecture B

V. CONCLUSION

This article describes the concept and initial analysis of the proposed summarization platform for lecture archives. In the analysis, the number of students whose learning styles changed over the two archives was 15, meaning that more than half of the learners' learning styles were consistent. Therefore, ROIs in the archives should reflect individual differences. The realization of learning experience extension that adaptively supports the attention to the spatiotemporal ROIs that differ from learner to learner on the terminal side is a noteworthy study from the perspective of a sustainable learning support system.

In this research, the raw data of the instructor's behavior during the lecture is used as F1 features, and the estimation results of the instructor's pose, emotion, and topics, which are necessary to explain his/her intention, are used as F2 features. By embedding these multimodal features as a context in each scene of the lecture, we could improve the explanation of the summarization method.

Future work will require an analysis of the differences in watching timelines in the archives by low, medium, and high-score groups and a comparison with the overall trend, including students who have not watched the archives. We are also considering analyzing the relationship between the submission of reports, another method of evaluating learning outcomes. Furthermore, we plan to conduct a similar analysis in the same course in the new academic year to analyze the trends/styles in general.

ACKNOWLEDGMENT

This work is supported by the Japan Science and Technology Agency and the JSPS KAKENHI Grant Number 23H03506.

REFERENCES

[1] S., Andreas, The Impact of COVID-19 on Education: Insights from "Education at a Glance 2020," OECD Publishing, 2020.

[2] S. Hasegawa, Y. Tajima, M. Matou, M. Futatsudera, and T. Ando, Case studies for self-directed learning environment using lecture archives, in Proc. of The Sixth IASTED International Conference on Web-basedEducation (WBE 2007), Citeseer, pp. 299-304, 2007.

[3] P. J. Guo, J. Kim, and R., Rubin, How video production affects student engagement: An empirical study of mooc videos, in Proceedings of the first ACM conference on Learning@ scale conference, pp. 41-50, 2014.

[4] M. B. Andra and T. Usagawa, Automatic lecture video content summarization with attention-based recurrent neural network, in Proceedings of 2019 International Conference of Artificial Intelligence and Information Technology, IEEE, pp. 54-59, 2019.

[5] R. Sheng, K. Ota, and S. Hasegawa, An Automatic Focal Period Detection Architecture for Lecture Archives, In Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners' and Doctoral Consortium. AIED 2022. Lecture Notes in Computer Science, vol.13356, Springer, Cham, 2022.

[6] D. Hulens, B. Aerts, P. Chakravarty, A. Diba, T. Goedeme, T. Roussel, J. Zegers, T. Tuytelaars, L. V. Eycken, L. V. Gool, H. V. Hamme and J. Vennekens, The CAMETRON Lecture Recording System: High Quality Video Recording and Editing with Minimal Human Supervision. In MultiMedia Modeling. MMM 2018. Lecture Notes in Computer Science, vol.10704, Springer, Cham, 2018.

[7] Y. Yang, K. Ota, W. Gu, and S. Hasegawa, Automatic Region of Interest Prediction from Instructor's Behaviors in Lecture Archives, 2022 14th International Conference on Knowledge and Systems Engineering (KSE), Nha Trang, Vietnam, pp. 1-6, 2022.

[8] R. C. Choe, Z. Scuric, E. Eshkol, A. Arndt, R. Cox, S. P. Toma, C. Shapiro, M. Levis-Fitzgerald, G. Barnes, R. H. Crosbie, Student Satisfaction and Learning Outcomes in Asynchronous Online Lecture Videos, CBE—Life Sciences EducationVol. 18, No. 4, 2019.

[9] T. Kobayashi, W. Nakano, H. Yokota, K. Shinoda, and S. Furui, Presentation Scene Retrieval Exploiting Features in Videos Including Pointing and Speech Information, Proc. Symposium on Large-Scale Knowledge Resources(LKR2007), pp. 95-100, 2007.

[10] X. Sun, J. Foote, D. Kimber, and B. S. Manjunath, Region of interest extraction and virtual camera control based on panoramic video capturing, IEEE Transactions on Multimedia, 7(5), pp.981-990, 2005.

[11] P. E. Dickson, W. R. Adrion, and A. R. Hanson, Automatic creation of indexed presentations from classroom lectures, Proceedings of the 13th annual conference on Innovation and technology in computer science education, pp.12-16, 2008.

[12] A. Mavlankar, P. Agrawal, D. Pang, S. Halawa, N. M. Cheung, and B. Girod, An interactive region-of-interest video streaming system for online lecture viewing, 18th International Packet Video Workshop, pp. 64-71, 2010.

[13] J. Zhang, G. Venture, and M. L. Bourguet, The effects of video instructor's body language on students' distribution of visual attention: an eye-tracking study, Proceedings of the 32nd International BCS Human Computer Interaction Conference 32, pp.1–5, 2018.

[14] X. Liu, W. Gu, K. Ota, and S. Hasegawa, Design for Voice Style Detection of Lecture Archives, Proc of the IEEE Region 10 Technical Conference 2023 (submitted 2023).

[15] M. Sugiyama, Learning under non-stationarity: Covariate shift adaptation by importance weighting, In: Handbook of computational statistics. Springer, pp. 927–952, 2012.