Demo: Real-Time Attention State Visualization of Online Classes

Taeckyung Lee
KAIST
taeckyung@kaist.ac.kr

Hye-Young Chung
Hanyang University
hy2315@hanyang.ac.kr

Sooyoung Park
KAIST
sypark0614@kaist.ac.kr

Dongwhi Kim
KAIST
dhkim09@kaist.ac.kr

Sung-Ju Lee
KAIST
profsj@kaist.ac.kr

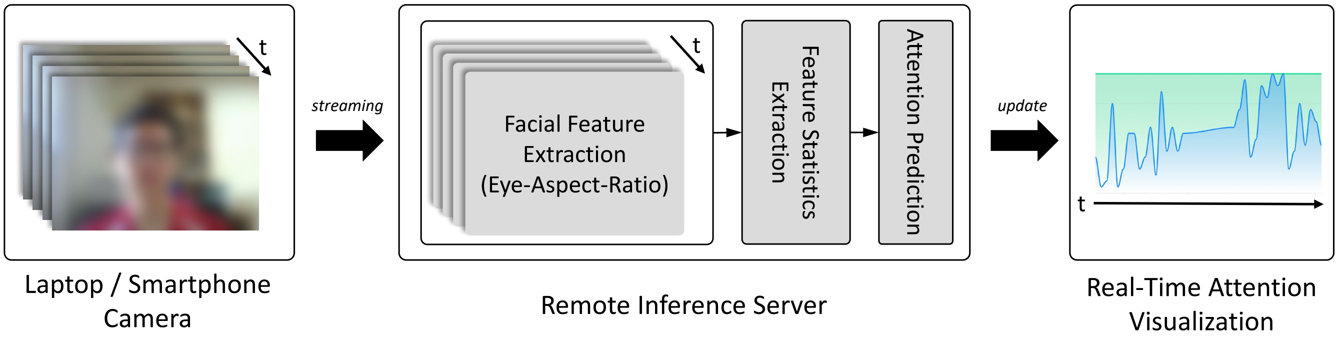
t
Laptop / Smartphone Camera
streaming
t
Facial Feature Extraction (Eye-Aspect-Ratio)
Feature Statistics Extraction
Attention Prediction
Remote Inference Server
update
t
Real-Time Attention Visualization

Figure 1: End-to-end attentional state prediction and visualization pipeline.

ABSTRACT

We propose the design of real-time end-to-end attentional state prediction system that utilizes only a webcam from a student's mobile device to provide a graph visualization.

CCS CONCEPTS

• **Human-centered computing → Ubiquitous and mobile computing systems and tools**.

ACM Reference Format:
Taeckyung Lee, Hye-Young Chung, Sooyoung Park, Dongwhi Kim, and Sung-Ju Lee. 2022. Demo: Real-Time Attention State Visualization of Online Classes. In *The 20th Annual International Conference on Mobile Systems, Applications and Services (MobiSys '22), June 25-July 1, 2022, Portland, OR, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3498361.3538655

1 INTRODUCTION

Getting students' attention is important in education. Therefore, many lecturers monitor students' attention states, try to grab their attention, and possibly adapt the learning contents based on such information. However, unlike in-person classes, online lecturers face challenges in understanding students' attention states as they have limited access to students' status. Moreover, students are more likely to lose attention during online, computer-based learning [1, 3].

Recent work showed that attentional states are predictable with specialized hardware such as gaze trackers, electroencephalography (EEG) sensors, electrodermal activity (EDA) sensors, and functional magnetic resonance imaging (fMRI) sensors. Utilizing these sensors, existing work estimates student attention states to improve student learning performance in real-time [2].

However, existing attentional state prediction systems have limitations when applied to live online lectures. Particular devices such as eye trackers, EEG, EDA, and fMRI are not easily deployable in the real world due to the high cost, requires physical contacts, and accuracy drops when replaced with real-world devices (e.g., eye-trackers [5]). On the other hand, webcams are already equipped in most student devices taking the lectures, such as smartphones, tablets, or laptops.

We present an end-to-end attention state prediction and visualization system (Figure 1), which enables lecturers to observe real-time attention states of the class. Our system captures facial recordings from the webcam of a student's mobile device and streams them to our inference server. The server extracts the facial features related to physiological representations of attention (e.g., emotion, drowsiness, and eye gaze), extracts feature statistics with a sliding window, and predicts the attention with the pre-trained XGBoost model. Inference results are updated in the remote database and connected to a live front-end webpage for visualization.

Our system visualizes the real-time attentional state of online learners without requiring any specialized or expensive hardware; it only uses webcams.

MobiSys '22, June 25-July 1, 2022, Portland, OR, USA
© 2022 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9185-6/22/06.
https://doi.org/10.1145/3498361.3538655

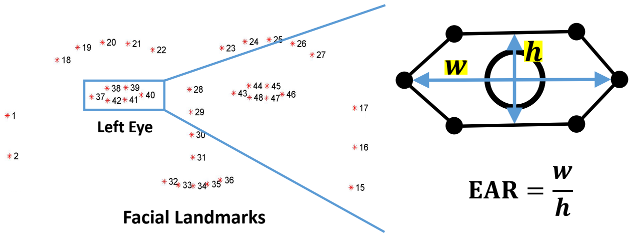

Figure 2: The eye-aspect-ratio (EAR) represents the ratio of width over height, which is the key representative of attentional states during online learning [4].

2 SYSTEM

We present the end-to-end design of a real-time attentional state visualization system in online lectures. We also provide the lightweight attentional state prediction to support the real-time system.

2.1 End-To-End System

First, we develop an application to capture and stream real-time facial videos during online learning. Each video frame is captured in 30FPS, up to 640x360 resolution; which are easily achievable by real-world laptop webcams or smartphone cameras. Each frame is then compressed with a JPEG codec to reduce the network bandwidth and sent to the remote server via a TCP connection.

The remote inference server runs the TCP server to receive video frames. The server distributes the image to multiple parallel processes, where each process

(1) decodes the JPEG image into raw RGB frames,
(2) performs face bounding box detection, facial landmark extraction, and eye-aspect-ratio calculation,
(3) calculates the statistics of facial features within a sliding window,
(4) predicts the attentional states by feeding the feature statistics to the pre-trained model, and
(5) passes the inference result to the merging process.

The merging process is unique and collects the data from distributed inference processes and uploads the predicted attentional states to a remote database.

Finally, we implement a real-time graph visualization webpage. Our webpage consists of a graph of attentional state ratio under sliding windows. The graph supports the real-time updates by attaching to the remote database updates. We expect our system to be further utilized for the attentional state visualization in real-world for students (to be aware of own's state) and lecturers (to be aware of multiple students' state).

2.2 Real-Time Attentional State Prediction

We build the attentional state prediction pipeline based on our previous work on data collection and attentional state prediction [4]. The previous work collects 1100 attentional state probings and corresponding webcam recordings of 15 students during online lecture viewing. Based on the dataset, four attentional state-related physiological features are extracted: eye-aspect-ratio, head movement, gaze, and emotion. Eye-aspect-ratio (EAR) is the ratio between eye width over eye height, calculated from facial landmark positions, as in Figure 2. Previous work [4] showed that eye-aspect-ratio is the critical indicator of attentional states during online education and could be successfully utilized for prediction models. Also, the EAR is invariant to possible (translation-, rotation-, scale-) variances in the real world, which is applicable in real-world scenarios.

Instead of utilizing the full 13 statistical features (e.g., EAR, emotion, gaze, and head translation), we only use eye-aspect-ratio statistics (mean, SD, 1q, 2q, 3q, MAD) in the current system. By removing other features, we eliminate the end users' gaze-calibration overhead. We also support real-time inference by avoiding the calculation delay of executing the few-shot gaze model and deep emotion model. We use XGBoost for the classification as XGBoost showed the best classification accuracy for full-feature prediction. We trained XGBoost only with eye-aspect-ratio and achieved AUROC=0.60, comparable to the previous full-feature-based approach (AUROC=0.67) [4].

3 DEMONSTRATION

Our demo will use webcam-attached devices (e.g., laptops, tablets, smartphones) to capture the facial videos and show the current user's (and aggregated class') real-time attentional state. We provide an example lecture series for the demonstration. During the demonstration, the participant will first watch the video for 20 seconds for initial data collection, and the attentional state graphs will appear after then. Then, our demo device will live-stream the facial video to the external server. At the server, the data will only be processed in memory and not be manually observed nor stored in the file at any physical storage.

ACKNOWLEDGEMENTS

This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No.NRF-2020R1A2C1004062) and the Institute of Information & communications Technology Planning & evaluation (IITP) grant funded by the Korea government (MSIT) (No.2022-0-00064, Development of Human Digital Twin Technologies for Prediction and Management of Emotion Workers' Mental Health Risks).

REFERENCES

[1] Rosa María Guadalupe García-Castelán, Andres González-Nucamendi, Víctor Robledo-Rella, Luis Neri, and Julieta Noguez. 2021. Face-to-face vs. Online learning in Engineering Courses. In *2021 IEEE Frontiers in Education Conference (FIE)*. 1–8. https://doi.org/10.1109/FIE49875.2021.9637177

[2] Stephen Hutt, Kristina Krasich, James R. Brockmole, and Sidney K. D'Mello. 2021. Breaking out of the Lab: Mitigating Mind Wandering with Gaze-Based Attention-Aware Technology in Classrooms. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 52, 14 pages. https://doi.org/10.1145/3411764.3445269

[3] Scott A. Jensen. 2011. In-Class Versus Online Video Lectures: Similar Learning Outcomes, but a Preference for In-Class. *Teaching of Psychology* 38, 4 (2011), 298–302. https://doi.org/10.1177/0098628311421336

[4] Taeckyung Lee, Dain Kim, Sooyoung Park, Dongwhi Kim, and Sung-Ju Lee. 2022. Predicting Mind-Wandering with Facial Videos in Online Lectures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

[5] Xucong Zhang, Yusuke Sugano, and Andreas Bulling. 2019. Evaluation of Appearance-Based Methods and Implications for Gaze-Based Applications. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300646