Concentration Estimation in Online Video Lecture Using Multimodal Sensors

Noriyuki Tanaka
Osaka Metropolitan University
Osaka, Japan
sd24263r@st.omu.ac.jp

Ko Watanabe
RPTU Kaiserslautern-Landau & DFKI
Kaiserslautern, Germany
ko.watanabe@dfki.de

Shoya Ishimaru
Osaka Metropolitan University
Osaka, Japan
ishimaru@omu.ac.jp

Andreas Dengel
RPTU Kaiserslautern-Landau & DFKI
Kaiserslautern, Germany
andreas.dengel@dfki.de

Shingo Ata
Osaka Metropolitan University
Osaka, Japan
ata@omu.ac.jp

Manato Fujimoto
Osaka Metropolitan University
Osaka, Japan
manato@omu.ac.jp

ABSTRACT

Online lecture is one of the technology-wise challenges in the education field. It provides the advantage of encouraging anyone to join from worldwide. However, understanding students' concentration in remote is one of the difficulties. In this paper, we evaluate multimodal sensors for estimating students' concentration levels during online video lectures. We collect multimodal sensor data such as accelerometers, gyroscopes, heart rates, facial orientations, and eye gazes. We conducted experiments with 13 university students in Japan. The results of our study, with an average accuracy rate of 74.4% for user-dependent cross-validation and 66.3% for user-independent cross-validation, have significant implications for understanding and improving student engagement in online learning environments. Most interestingly, we found that facial orientations are significant for user-dependent and eye gazes for user-independent classification.

CCS CONCEPTS

• **Human-centered computing** → **Ubiquitous and mobile computing**; • **Applied computing** → **E-learning**; **Distance learning**; • **Hardware** → Sensor applications and deployments.

KEYWORDS

Wearable sensor, multimodal sensing, concentration detection, online learning, machine learning

ACM Reference Format:
Noriyuki Tanaka, Ko Watanabe, Shoya Ishimaru, Andreas Dengel, Shingo Ata, and Manato Fujimoto. 2024. Concentration Estimation in Online Video Lecture Using Multimodal Sensors. In *Companion of the 2024 ACM International Joint Conference on Pervasive and Ubiquitous Computing Pervasive and Ubiquitous Computing (UbiComp Companion '24), October 5–9, 2024, Melbourne, VIC, Australia.*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3675094.3677587

UbiComp Companion '24, October 5–9, 2024, Melbourne, VIC, Australia.

ACM ISBN 979-8-4007-1058-2/24/10
https://doi.org/10.1145/3675094.3677587

1 INTRODUCTION

Since COVID-19, education has entered an era of diverse learning environments. According to a survey by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) in Japan, many universities adopted online lectures, especially in 2020 [5]. Currently, 896 out of 1069 universities in Japan offer both onsite and online lectures. Although the pandemic has subsided and face-to-face classes have resumed, some universities and institutions continue to provide online education. Therefore, offline and online education must be the scope of the education environment.

Online education presents challenges compared to onsite learning, particularly in understanding students' cognitive status. This issue arises due to the lack of capturing non-verbal information from students [9, 10]. As a potential solution, sensing offers a promising approach to visualizing students' cognitive states [6, 8, 11]. Different educational institutions have different devices and sensors, comparisons among sensors would be beneficial. From these points of view, we aim to collect multiple sensors and compare their performance in measuring cognitive state.

In ubiquitous computing, "concentration" is one of the important cognitive states to measure in education. Uema and Inoue has proposed an approach to measure concentration using JINS MEME glasses to measure concentration levels [8]. Regarding the findings, we encourage collecting gaze information and comparing it with the other sensor features.

This research aims to compare multimodal sensors for estimating concentration levels. To do so, we synchronously collect and compare data from cameras and wearable sensors and compare sensors against each other to verify which is the appropriate sensor. Specifically, we embark on a complex process of collecting sensor data such as accelerometers, gyroscopes, heart rates, facial orientations, and eye gazes. Collecting multimodal sensor data, we leverage to compare which sensors effectively estimate binary concentration levels. This paper presents two contributions,

C1 Concentration estimation with multimodal sensors.
C2 Comparison of multimodal sensors to assess valid features.

2 RELATED WORK

Betto et al. [2] propose a method for detecting student distraction state in e-learning lectures based on student face and posture information collected from webcams. A binary classification model for the user-dependent Random Forest model achieved 90 % recall.

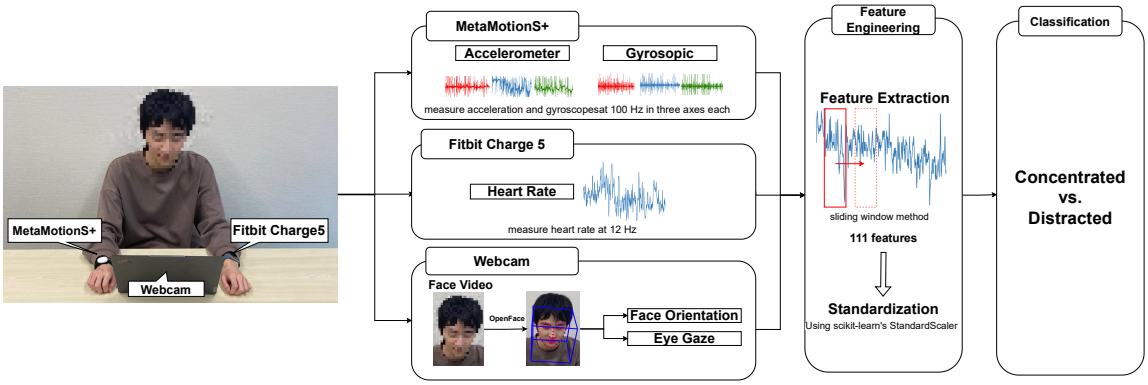

Figure 1: Experimental Overview.

Sevil et al. [7] proposed an algorithm to detect and characterize physical activity and acute psychological stress from heart rate and acceleration data obtained from wristbands. As a result, physical activity and acute psychological stress were detected and classified with an accuracy of 99.3% and 92.7%, respectively.

Kimura et al. [4] proposed a method to estimate intellectual concentration using pupil diameter and heart rate variability. Experiments were conducted on 31 participants, and the state of concentration was estimated with an average accuracy of 57.3%. The simultaneous use of such features may enable us to estimate the concentration level more accurately. In addition, the use of devices that are burdensome to students may impede their concentration.

In our study, we aim to realize a sensor-based concentration estimation system that does not depend on the student's classroom environment and does not burden the wearer.

3 METHODOLOGY

In this section, we describe a methodology for estimating concentration levels. First, we explain the three sensors we use in this study. Then, we present an approach to the feature engineering preprocessing procedure before applying it to the classifier.

3.1 Multimodal Sensors

In this section, we describe the sensors used in our study. There are various sensors that are considered to be effective for concentration estimation. In this study, we use five types of sensor data: accelerometers, gyroscopes, heart rates, facial orientations, and eye gazes. As shown in Figure 1, we use three types of devices to measure the five types of data. The following section describes the sensors used in detail.

3.1.1 Accelerometer and Gyrosopic Sensing. In this study, we use MetamotionS+[1] for measuring acceleration and gyroscopic. MetamotionS+ is a sensor that can measure various data such as acceleration, gyroscopes, magnetic force, temperature, and air pressure, and can be worn on the wrist using a special band. In this study, MetamotionS+ is worn on the participant's dominant arm and measures acceleration and gyroscopes at 100 Hz in three axes each.

3.1.2 Heart Rate Sensing. In this study, we use Fitbit Charge 5[2] for measuring heart rate. The Fitbit Charge 5 is a wristwatch-type wearable device that can measure heart rate, sleep data, and exercise data by wearing it on the participant's arm. In this study, Fitbit Charge 5 is worn on the opposite arm of the participant's dominant hand. Fitbit Charge 5 is basically set to measure heart rate at 12 Hz, but since measurement may fail in rare cases, the number of data tends to be smaller than other data such as acceleration and gyroscopes.

3.1.3 Face Orientation and Eye Gaze Sensing. In this section, we describe a sensor that measures face orientation and eye gaze sensing. In this study, we use *ThinkPad X1 Yoga Gen 5* as a notebook PC. Using a built-in web camera with a 30 frames per second, we capture participant face images. This laptop is the same as the one used for watching the online lecture. We record the participant's face during the experiment, and after the experiment is over, we apply OpenFace [1] to extract facial orientation data from the video. We use the angle of rotation of the face and the line of sight.

3.2 Feature Engineering

The sliding window method is used for feature extraction. In this study, window sizes of 2.5, 5, 10, and 15 seconds are used. Data from the 30 seconds immediately before the annotation was used. This is because we judged that the time away from the annotation timing is less reliable. The ten extracted features are shown in Table 1. The signal magnitude area (SMA) was not extracted from the eye gaze and heart rate data because data from all three axes are required. Therefore, a total of 111 features are extracted: 28 from the acceleration data, 28 from the gyro data, nine from the heart rate data, 28 from the facial orientation data, and 18 from the eye gaze data.

In addition, as mentioned in Section 3.1.2, heart rate is basically measured at 12 Hz, so there are cases where no value is measured at all within the window. In such cases, it is impossible to extract the features, so in this study, they are replaced by 0. In addition to the four window sizes described above, four overlaps (0, 0.25, 0.5, and 0.75) are employed, and a total of 16 combinations of these overlaps are used to extract features. After feature extraction, scaling is

[1]https://mbientlab.com/store/metamotions-p

[2]https://www.fitbit.com/global/us/products/trackers/charge5

Table 1: Feature lists applied while applying sliding window.

Function	Definition	Algorithm
$mean(s)$	Mean	$\bar{s} = \frac{1}{N}\sum_{i=1}^{N} s_i$
$max(s)$	Maximum	$\max_i(s_i)$
$min(s)$	Minimum	$\min_i(s_i)$
$std(s)$	Standard Deviation	$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(s_i - \bar{s})^2}$
$mad(s)$	Median Absolute Deviation	$\text{median}_i\left(\lvert s_i - \text{median}_j(s_j)\rvert\right)$
$energy(s)$	Mean Square	$\sum_{i=1}^{N} s_i^2$
$sma(s_1, s_2, s_3)$	Signal Magnitude Area	$\frac{1}{3}\sum_{i=1}^{3}\sum_{j=1}^{N} \lvert s_{i,j}\rvert$
$iqr(s)$	Interquartile Range	$Q_3(s) - Q_1(s)$
$range(s)$	Range	$\max_i(s_i) - \min_i(s_i)$
$rms(s)$	Root Mean Square	$\sqrt{\frac{1}{N}\sum_{i=1}^{N} s_i^2}$

performed using standardization, and the training model is used for verification.

4 DATA COLLECTION

This section describes the details of the evaluation experiments conducted based on the proposed method.

4.1 Participants

In this study, we recruited thirteen students (12 males and one female) from a university in Japan. The participant is between 22 and 24 (M=23.2) years old. The experiment conductor provides clear and detailed instructions to the participants, ensuring they understand the procedure and what kind of data to collect. This fosters a sense of confidence and knowledge. The use of the data is also explained, and once they confirm, participants fill out the consent form. During the experiment, participants are instructed to opt out of the experiment anytime they want.

4.2 Procedure

In this section, we describe the experimental procedure. An experiment was conducted to evaluate the accuracy of concentration estimation during a video class using the three types of sensors described in Section 3.1. To control the experiment settings, we instruct participants not to leave their seats, take notes, or fall asleep during the experiment.

The experiment is shown in Figure 1. Our data collection process is meticulously designed. Participants were first asked to wear the MetamotionS+ on their dominant wrist (all participants were right-handed) and the Fitbit Charge 5 on the opposite arm. Then, participants sit in front of a laptop computer with a webcam recording. We ask participants to watch two YouTube videos for the online video lecture. One lecture is about explaining Git (a version control system)[3]. Another video is about SQL (Structured Query Language)[4]. The video has a total of 90 minutes.

In this study, the vibration function and switches of MetamotionS+ were used for annotation. The shorter the annotation interval, the more labels can be obtained, but the annotation should be as unobtrusive as possible to the participant's awareness to reproduce the actual classroom situation as much as possible. D'Mello and Mills, which tracked participants' emotional states while writing an essay, reported that 4.8% of participants reported feeling very frustrated when asked to rate 11 different emotional states and their intensity every 90 seconds. Referring to the results of this study, if the participants were only asked to answer whether they were concentrating or not, it would not be a significant burden to require them to take annotations every 90 seconds. During the 90 minutes, 60 labels were obtained per participant.

[3]https://youtu.be/WHwuNP4kalU
[4]https://youtu.be/v-Mb2voyTbc

Table 2: Balance of the collected concentration levels.

Participant	Concentration	Non-Concentration
P1	35	25
P2	38	22
P3	38	22
P4	27	33
P5	36	24
P6	34	26
P7	25	35
P8	35	25
P9	37	23
P10	39	21
P11	30	30
P12	47	13
P13	58	2
Total	479	301

5 DATA ANALYSIS

In this study, we applied gradient boosting, decision tree, logistic regression, random forests, and SVM for the binary concentration level classification. For the performance measurement, we use accuracy, which is the percentage of correct predictions. In this research, we verify two types of evaluations. User-dependent and independent analysis.

User-dependent validation is performed by applying participant-self data for training and testing. As explained in Section 4.2, we collect 60 labels for each participant. We select six labels as a chunk of data and divide them into ten segments. We apply one segment as test data and nine segments as training data. We continue this procedure for ten times.

User-independent analysis is performed by applying leave-one-participant-out cross-validation. We apply one participant's data as the test data, and the remaining twelve participants' data as the training data. We continue this procedure for thirteen times until all participant data is set as test data.

6 RESULT

In this section, we describe both user-dependent and independent concentration estimation results.

6.1 User-Dependent Performance

In this section, we discuss the results of the user-dependent analysis. For user-dependent analysis, we observed that the highest accuracy rate was scored with a window size of 15 and an overlap of 0.25, using a random forest. The following are discussed based on these best-performed results.

Table 3: Concentration estimation result for user-dependent (UD) and user-independent (UI) cross validation.

	Accelerometer		Gyroscope		Face Orientation		Eye Gaze		Heart Rate		All Sensors	
Participant	UD	UI	UD	UI	UD	UI	UD	UI	UD	UI	UD	UI
P1	0.900	0.559	0.733	0.556	0.850	0.601	0.808	0.662	0.642	0.583	0.917	0.731
P2	0.592	0.653	0.600	0.629	0.725	0.733	0.667	0.698	0.500	0.633	0.783	0.717
P3	0.683	0.586	0.717	0.657	0.667	0.647	0.592	0.659	0.500	0.633	0.750	0.651
P4	0.583	0.529	0.533	0.479	0.733	0.536	0.708	0.563	0.500	0.450	0.758	0.596
P5	0.658	0.611	0.708	0.613	0.708	0.672	0.675	0.639	0.600	0.600	0.711	0.696
P6	0.525	0.560	0.483	0.570	0.583	0.618	0.700	0.640	0.450	0.567	0.661	0.646
P7	0.450	0.418	0.533	0.402	0.808	0.661	0.783	0.640	0.608	0.417	0.842	0.650
P8	0.625	0.642	0.575	0.616	0.567	0.589	0.558	0.616	0.592	0.583	0.667	0.660
P9	0.617	0.622	0.617	0.618	0.525	0.627	0.608	0.653	0.517	0.617	0.625	0.607
P10	0.542	0.634	0.600	0.656	0.558	0.584	0.533	0.656	0.483	0.650	0.600	0.661
P11	0.300	0.468	0.325	0.506	0.475	0.519	0.433	0.514	0.625	0.500	0.606	0.564
P12	0.667	0.772	0.733	0.770	0.767	0.700	0.750	0.738	0.717	0.783	0.783	0.718
P13	0.967	0.827	0.967	0.913	0.967	0.809	0.958	0.852	0.950	0.967	0.967	0.726
Mean	0.624	0.606	0.625	0.614	**0.687**	0.638	0.675	**0.656**	0.591	0.614	**0.744**	**0.663**

UD column in Table 3 shows the accuracy for each participants. These results underscore the crucial role of participant variability in determining accuracy in user-dependent analysis. Participants like P1, P7, and P13 scored higher accuracy than others. Upon closer examination, we found that participant P13 exceptional performance can be attributed to the presence of unbalanced labels, as evidenced in Table 2. However, P1 and P7 are performing significantly well.

Comparing different sensor results, we discovered that face orientation performed with the highest accuracy of 0.687 and heart rate at the lowest level of 0.591 for user-dependent concentration estimation. This result states that facial orientation has user-dependent unique movement related to concentration.

6.2 User-Independent Performance

In this section, we discuss the results of the leave-one-participant-out cross-validation. For user-independent analysis, we observed that the highest accuracy rate scored with a window size was 10, and the overlap was 0, using random forest. The following are discussed based on these best-performed results.

UI column in Table 3 shows the results of the leave-one-participant-out cross-validation. The average concentration level estimation performs 0.663 for all sensors as input. The highest performed participant was 0.731 with P1, and the lowest was 0.564 with P11. The result did not perform well with leave-one-participant-out cross-validation.

Comparing different sensors, we observed that eye gaze performed best with an accuracy of 0.656 and lowest with an accuracy of 0.606 for the accelerometer. This result states that eye gaze is a generalized feature that can be used as a feature to estimate participant concentration compared to the other four sensors.

7 LIMITATIONS AND FUTURE WORK

Our work compared sensors to measure student concentration levels during online video lectures. We visualize which sensor performs well for each user-dependent and independent scenario. However, we have yet to discover data fusion between two, three, or four. In this study, we only apply all sensor data as input.

Another area for improvement is the need for more variety in participant backgrounds. We collected participants from Japan and did not consider the experiment on people from other countries. Human behavior and cultural differences have yet to be considered and must be addressed.

Lastly, our experiment collects self-report binary concentration levels. However, concentration cannot be measured in binary and needs to be in regression form. We need to consider measuring concentration levels in regression form.

For future works, we can check which combination can be the best pattern for classifying concentration levels. Furthermore, it is crucial to emphasize the necessity of collecting participants from diverse backgrounds, which will generalize the results of our research. We also consider implementing a real-time student concentration level estimation application while taking online video lectures.

8 CONCLUSION

In this study, we collect students' concentration levels and multimodal sensor data while taking online video lectures. We collect sensor data such as accelerometers, gyroscopes, heart rates, facial orientations, and eye gazes. We conducted experiments with 13 university students in Japan. Using all sensor data as input and applying random forest, we achieved an accuracy rate of 74.4% for user-dependent cross-validation and 66.3% for leave-one-participant-out cross-validation. Notably, face orientation performs the highest accuracy for user-dependent classification and eye gaze for user-independent analysis. This study shows how multimodal sensors measure concentration levels and which sensor is useful in user-dependent and independent scenarios.

ACKNOWLEDGMENTS

This work was supported in part by the Japan Society for the Promotion of Science, Grants-in-Aid for Scientific Research number JP24K02934.

REFERENCES

[1] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 1–10.

[2] Iku Betto, Ryo Hatano, and Hiroyuki Nishiyama. 2023. Distraction detection of lectures in e-learning using machine learning based on human facial features and postural information. *Artificial Life and Robotics* 28, 1 (2023), 166–174.

[3] Sidney D'Mello and Caitlin Mills. 2014. Emotions while writing about emotional and non-emotional topics. *Motivation and Emotion* 38 (2014), 140–156.

[4] Kaku Kimura, Shutaro Kunimasa, You Kusakabe, Hirotake Ishii, and Hiroshi Shimoda. 2018. Estimation of Intellectual Concentration States using Pupil Diameter and Heart Rate Variability.. In *CHIRA*. 62–69.

[5] Ministry of Education, Culture, Sports, Science and Technology (MEXT). 2020. New Coronavirus and Distance Learning Measures. https://www.mext.go.jp/content/20200717-mxt_kouhou01-000004520_2.pdf Accessed: 2024-05-22.

[6] Akshay Palimar Pai, Jayasankar Santhosh, and Shoya Ishimaru. 2023. Real-Time Feedback on Reader's Engagement and Emotion Estimated by Eye-Tracking and Physiological Sensing. In *Adjunct Proceedings of the 2022 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2022 ACM International Symposium on Wearable Computers* (Cambridge, United Kingdom) *(UbiComp/ISWC '22 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 97–98. https://doi.org/10.1145/3544793.3560329

[7] Mert Sevil, Mudassir Rashid, Mohammad Reza Askari, Zacharie Maloney, Iman Hajizadeh, and Ali Cinar. 2020. Detection and characterization of physical activity and psychological stress from wristband data. *Signals* 1, 2 (2020), 188–208.

[8] Yuji Uema and Kazutaka Inoue. 2017. JINS MEME algorithm for estimation and tracking of concentration of users. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers* (Maui, Hawaii) *(UbiComp '17)*. Association for Computing Machinery, New York, NY, USA, 297–300. https://doi.org/10.1145/3123024.3123189

[9] Ko Watanabe, Tanuja Sathyanarayana, Andreas Dengel, and Shoya Ishimaru. 2023. EnGauge: Engagement Gauge of Meeting Participants Estimated by Facial Expression and Deep Neural Network. *IEEE Access* 11 (2023), 52886–52898. https://doi.org/10.1109/ACCESS.2023.3279428

[10] Ko Watanabe, Yusuke Soneda, Yuki Matsuda, Yugo Nakamura, Yutaka Arakawa, Andreas Dengel, and Shoya Ishimaru. 2021. DisCaaS: Micro Behavior Analysis on Discussion by Camera as a Sensor. *Sensors* 21, 17 (2021). https://doi.org/10.3390/s21175719

[11] Hiroki Yoshikawa, Akira Uchiyama, Yuki Nishikawa, and Teruo Higashino. 2019. Combining a thermal camera and a wristband sensor for thermal comfort estimation. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers* (London, United Kingdom) *(UbiComp/ISWC '19 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 238–241. https://doi.org/10.1145/3341162.3343813

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009