Estimation of Wakefulness in Video-based Lectures based on Multimodal Data Fusion

Ryosuke Kawamura
k.ryosuke@fujitsu.com
Fujitsu Laboratory Ltd., Japan

Shizuka Shirai
shirai@ime.cmc.osaka-u.ac.jp
Osaka University, Japan

Mehrasa Aizadeh
mehrasa@cmc.osaka-u.ac.jp
Osaka University, Japan

Noriko Takemura
takemura@ids.osaka-u.ac.jp
Osaka University, Japan

Hajime Nagahara
nagahara@ids.osaka-u.ac.jp
Osaka University, Japan

ABSTRACT

In distance learning contexts, drowsiness is a major factor which disturbs learning. However, it is not easy for instructors to monitor students' wakefulness. In order to improve learning efficacy, accurate estimation of wakefulness is needed. In this study, we propose a multimodal wakefulness estimation method based on face and body movement information. We utilize web-cameras to obtain facial and head (face-head) movements and pressure mats for body movements, the latter of which can record the distribution of upper body pressure while watching video lectures. To confirm the effectiveness of multimodal data for wakefulness estimation, we conducted an experiment to collect data from students as they engaged in e-learning and their level of wakefulness was annotated in one-second windows. We extracted 45 features from face-head movements, and 80 features from seat pressure data. Two types of fusion methods, early and decision level fusion were applied, and the late fusion approach achieved an average F1-macro score of 0.70 in three levels of wakefulness estimation, which is higher than the unimodal approach. This result indicates that fusion of facial images and seat pressure features can be effective for learner wakefulness estimation.

CCS CONCEPTS

• **Applied computing** → **E-learning**; *Computer-assisted instruction*.

KEYWORDS

wakefulness, e-Learning, facial expression, pressure sensor

ACM Reference Format:

Ryosuke Kawamura, Shizuka Shirai, Mehrasa Aizadeh, Noriko Takemura, and Hajime Nagahara. 2020. Estimation of Wakefulness in Video-based Lectures based on Multimodal Data Fusion. In *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers (UbiComp/ISWC '20 Adjunct)*, September 12–16, 2020, Virtual Event, Mexico. ACM, New York, NY, USA, 4 pages. <https://doi.org/10.1145/3410530.3414386>

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

UbiComp/ISWC '20 Adjunct, September 12–16, 2020, Virtual Event, Mexico

© 2020 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-8076-8/20/09.

<https://doi.org/10.1145/3410530.3414386>

1 INTRODUCTION

Video-based learning has been paid attention to in recent years since anyone can learn with videos regardless of time and location. However, few learners manage to complete all online coursework requirements successfully [11]. One of the major factors that decreases learning efficiency is drowsiness. In video-based learning, it is not easy to grasp students' drowsiness compared to face-to-face settings. To approach this problem, Intelligent Tutoring Systems (ITS) which monitor learners' states and provide suitable feedback automatically have gained popularity. If an ITS can detect students' wakefulness accurately, it can provide them with appropriate feedback and make their video-based learning experience more effective. Nevertheless, accurate wakefulness estimation is still a core challenge. Researchers have utilized facial expressions to estimate students' internal states since video-based unimodal data is much easier to collect. For instance, [13] and [7] have analyzed facial image data including facial expressions and head poses in cognitive and educational task and built internal state (such as engagement and frustration) estimation models. Terai et al. [12] also indicated the effectiveness of facial expressions and head movements for wakefulness estimation in video-based learning. In addition, recent studies show the relationship between body movements and internal states in e-learning contexts [5, 9]. Those studies utilize body movement information obtained from pressure sensors or web camera to estimate learners' engagement.

In this study, we propose a new approach to estimate students' wakefulness based on facial images and body movements. We have deployed web cameras to monitor the changes of facial expressions and pressure detection sensors attached to seats to capture body movements. Web cameras obtain facial expressions and head movements, and pressure sensors can capture the distribution of upper body pressure. Our model is constructed based on these multimodal sensor data and estimates wakefulness in video-based learning. We have evaluated the model to investigate the effectiveness of multimodal sensor data for wakefulness estimation.

2 RELATED WORK

Many studies have focused on facial expression information to estimate internal states such as concentration in e-Learning settings. For example, Whitehill et al. [13] utilized action units and head pose with some machine learning method to estimate the level of learner engagement. Gupta et al. [7] constructed a dataset of facial images during video lectures and built a baseline model

Figure 1: The steps in feature extraction and two types fusion method; Early (red part) and Decision Level fusion (blue part)

for estimation of four kinds of internal states (concentration, frustration, boredom, and confusion). In the context of writing tasks, Monkarese et al. [8] deployed a video-based estimation approach for learners' engagement estimation. They analyzed facial expression data as captured by Microsoft Kinect Face Tracker and heart rate measures from video-based sensing, using photoplethysmography. For wakefulness estimation, Ghoddoosian et al. [6] constructed a realistic database including learning situation and desk work. They proposed a baseline model with a hierarchical LSTM based on blinking information. Terai et al. [12] focused on action units and head pose information obtained from OpenFace2.0 [3]. They estimated annotated wakefulness level with a personalized model, and the results showed effectiveness of facial expressions for wakefulness estimation, in spite of the fact that their model was constructed for each examined subject.

Moreover, some researches have paid attention to body movements for estimation of internal states. D'Mello et al. [5] utilized back and seat pressure sensors to investigate the possibility of differentiating engagement from disengagement. Nomura et al. [9] investigated engagement in online lectures using a smart chair with a pressure sensor and 2D keypoints of upper body. Although their model was user-dependent, an average accuracy of 79.3% for student engagement measure was reported. From these studies, it is expected that body movements can potentially be useful for wakefulness estimation.

3 WAKEFULNESS ESTIMATION METHOD BASED ON MULTIMODAL DATA

The steps to extract feature values and build the model are described in Fig. 1.

3.1 Recording multimodal data

We used web-cameras to capture facial expressions and pressure sensors to obtain seat pressure data. Facial expression data were recorded with a software program called Bandicam (30 fps). Seat pressure data were captured with pressure sensors (SR soft vision [1]) at 5 fps. This pressure mat has 256 pressure sensors with 16 rows x 16 columns (Fig. 2 shows a sample of the pressure data).

Figure 2: An example of pressure data

Figure 3: Facial feature points around the eyelid

3.2 Feature extraction

3.2.1 Features in facial expression and head pose. Blinks are considered to be important for wakefulness estimation since they vary with changes in wakefulness. In e-learning contexts, eye data is not always fully captured due to head motion for checking documents, taking notes, or nodding caused by drowsiness. Therefore, we also focused on features related to head motion. In order to extract features related to blinks, Action Unit 45 (AU45) was computed using OpenFace 2.0 [3]. In addition, Eye Opening Rate (EOR) of both left and right eyes were calculated using facial feature points around the eyelid (p_1 p_6 in Fig. 3) based on the following equation.

$$EOR = \frac{\|p_2 - p_6\| + \|p_3 - p_5\|}{\|p_1 - p_4\|} \quad (1)$$

We calculated EOR from both left and right eyes. In order to extract features related to head motion, rotation and transition of head around x, y, z axis were extracted by using OpenFace 2.0. We also extracted head rotation (yaw, pitch, roll) and transition. Finally, we calculated feature values as displayed in Table 2, with f_t^{face} as the features from facial images in time t.

Figure 4: 16 grids of seat pressure data

Table 1: Feature values from facial images and seat pressure

Feature type	Features
Facial images	mean, median, standard deviation, zero crossing rate of forward differences, range, sum of forward differences of AU45, EOR, Head Rotation, Transition and in one second
Seat pressure	mean, median, standard deviation, range, zero crossing rate of forward differences, sum of forward differences of pressure data in each area in one second

3.2.2 Seat pressure feature. Initially, 16 rows x 16 columns in seat pressure sensor data were divided into 4 rows x 4 columns grids (Fig. 4), and average pressure of each grid was calculated. Secondly, we calculated feature values described in Table 2 in every one second, with f_t^{seat} as the features from seat pressure data in time t . In total, we extracted 45 features from facial images and 80 features from seat pressure data.

3.3 Normalization

After calculating feature values f_t^{face} and f_t^{seat} , we standardized them by using mean and standard deviation in one fifth of the data at highest level of wakefulness state. The same approach was adopted in [6], and it assumes that the recorded data at the beginning is the data in wakeful state. $\overline{f_t^{face}}$ and $\overline{f_t^{seat}}$ denote standardized feature values of both face and seat pressure features.

3.4 Modeling for wakefulness estimation

We used Catboost classifier [10] to build our wakefulness estimation model. In this model, $\overline{f_t^{face}}$, $\overline{f_t^{seat}}$ or both feature values are used as input and the model estimates wakefulness state in time t . In order to deal with imbalanced data, for the Catboost classifier, we used SMOTE [4] and random undersampling to reduce any imbalance between the level of wakefulness. Hyper parameters of Catboost classifier are tuned by using Optuna [2].

4 EXPERIMENT AND RESULTS

We collected multimodal data from learners during e-learning and evaluated the accuracy of our model to confirm the effectiveness of multimodal data fusion for wakefulness estimation.

Figure 5: Screenshot of a slide in a video lecture

Figure 6: Distribution of each wakefulness state

4.1 Collection of multimodal data

We collected data from 53 undergraduate students as they were watching video lectures on informatics and selected 4 students (aged 18 to 19, SD = 0.5) who had fallen asleep during the lectures for analysis. In each session, the participants watched one to three videos that were approximately 10 minutes long, and each video was created as a slideshow with audio narration. The selected students watched videos from three to five sessions, with the total length of collected data for analysis as 18,035 seconds.

Fig. 5 shows a sample slide in a video lecture. While watching the videos, facial expressions, screen capture, and seat pressure were recorded. The experiment was run on laptops (17.3-inch screen). After the experiment, students' wakefulness level was coded on a scale of 0 to 2 (0: Asleep, 1: Drowsy, 2: Awake) through human observations. Following the majority rule, three observers independently coded the video data based on the coding scheme in Table 2 using one-second windows. The distribution of each wakefulness state is shown in Fig. 6.

Table 2: Coding scheme of wakefulness state

Level	Scheme
Asleep	-Eyes are closed for over 1 second
Drowsy	-Eyes are not always open -Pupils do not move -Body movements and head poses are uncontrolled
Awake	-Eyes are closed for less than 1 second -Eyes are wide open -Pupils move to the right and left -Body movements and head poses are under control

Table 3: average F1-macro scores (standard error) of each classification, feature and fusion type

Feature types	Asleep vs Others	Awake vs Others	3 levels
Chance rate	0.49	0.49	0.33
<i>Face</i>	0.88 (0.03)	0.64 (0.05)	0.67 (0.04)
<i>Seat</i>	0.55 (0.02)	0.72 (0.05)	0.45 (0.03)
<i>Early Fusion</i>	0.88 (0.03)	0.71 (0.03)	0.70 (0.05)
<i>Decision Fusion</i>	0.89 (0.03)	0.72 (0.05)	0.69 (0.03)

4.2 Evaluation method

In order to evaluate the accuracy of wakefulness estimation based on multimodal sensor data, we employed leave-one-out cross validation. The data from one person was used for testing, and the remaining data was used for train models. This method repeats until all persons' data are tested. We employed average F1-macro score as our evaluation index since the distribution of three levels of wakefulness is skewed. In order to offer tailored feedback to students, we evaluated the performance of three classification conditions: one three-class classification and two types of two-class classifications (asleep vs drowsy or awake, asleep or drowsy vs awake).

4.3 Comparison of multimodal fusion

The collected data contains two types of feature values, namely features from facial images and seat pressure. To confirm the effectiveness of multimodal data for wakefulness estimation, we made a comparison across four cases of modeling (1) inputting only features from facial images (*Face*), (2) inputting seat pressure data (*Seat*), (3) inputting fusion of these features (*Early Fusion*) and (4) fusion of the results from *Face* and *Seat* (*Decision Fusion*). In *Decision Fusion*, probability for each class ($P_{Decision}$) is calculated by using the equation 2.

$$P_{Decision} = \alpha P_{Face} + (1 - \alpha) P_{Seat} \quad (2)$$

where P_{Face} and P_{Seat} are predicted probabilities for each class from both models of *Face* and *Seat*, and $\alpha \in [0, 1]$ shows the importance of these features. Chance rate is computed by averaging thousand times of random prediction based on the distribution of the data.

4.4 Results

Table 3 shows the average F1-macro scores in each feature, fusion and classification types. In asleep vs drowsy or awake classification, *Face*, *Early Fusion* and *Decision Fusion* outperform *Seat*. In contrast to asleep vs drowsy or awake classification, *Seat* and *Fusion* yield better scores compared to *Face* in asleep or drowsy vs awake classification. Based on the results of two types of two-class classification, we can conclude that *Face* and *Seat* are effective for different levels of wakefulness classification.

In three-class classification, both *Early Fusion* and *Decision Fusion* mark the highest score (0.70, 0.69), and they are higher than *Face* and *Seat*. These results indicate that multimodal sensor fusion of facial images and seat pressure data is capable of capturing effective features for both asleep vs drowsy or awake and asleep or drowsy vs awake classifications. Thus, multimodal fusion can be considered as an effective means of improving wakefulness estimation.

5 CONCLUSION

In this study, we propose a new approach to estimate students' wakefulness based on facial images and body movements. We collected facial images and seat pressure data from students studying with video-based lectures. The data was annotated at three levels of wakefulness, and we evaluated the performance of wakefulness estimation models based on features from facial images, seat pressure data, and their fusion. The estimation models were built with Catboost classifier, and we found that multimodal fusion features yield higher scores than single mode features. It is thus concluded that multimodal fusion can improve the performance of wakefulness estimation. Our data collection is limited to university students, and it should be noted that facial expressions and postures can be different across varying age groups such as elementary school students or middle-aged workers. As future work, we plan to collect data from a larger number of participants to confirm the generalizability of our method and make an interactive system which can provide feedback to learners.

ACKNOWLEDGEMENTS

This research is partially supported by Innovation Platform for Society 5.0 from Japan Ministry of Education, Culture, Sports, Science and Technology (MEXT)

REFERENCES

- [1] [n.d.]. <https://www.sumitomoriko.co.jp/product/health/srsv/> (Accessed on 26/05/2020).
- [2] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2623–2631.
- [3] Tadas Baltrušaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 59–66.
- [4] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.
- [5] Sidney S D'Mello, Patrick Chipman, and Art Graesser. 2007. Posture as a predictor of learner's affective engagement. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 29.
- [6] Reza Ghoddoosian, Marnim Galib, and Vassilis Athitsos. 2019. A Realistic Dataset and Baseline Temporal Model for Early Drowsiness Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 0–0.
- [7] Abhay Gupta, Arjun D'Cunha, Kamal Awasthi, and Vineeth Balasubramanian. 2016. Daisee: Towards user engagement recognition in the wild. *arXiv preprint arXiv:1609.01885* (2016).
- [8] Hamed Monkaresi, Nigel Bosch, Rafael A Calvo, and Sidney K D'Mello. 2016. Automated detection of engagement using video-based estimation of facial expressions and heart rate. *IEEE Transactions on Affective Computing* 8, 1 (2016), 15–28.
- [9] Kazuaki Nomura, Motoi Iwata, Olivier Augereau, and Koichi Kise. 2019. Estimation of student's engagement based on the posture. 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*. 164–167.
- [10] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. 2018. CatBoost: unbiased boosting with categorical features. In *Advances in neural information processing systems*. 6638–6648.
- [11] Mingjie Tan and Peiji Shao. 2015. Prediction of Student Dropout in E-Learning Program Through the Use of Machine Learning Method. *International Journal of Emerging Technologies in Learning* 10, 1 (2015).
- [12] Shogo Terai, Shizuka Shirai, Mehrasa Alizadeh, Ryosuke Kawamura, Noriko Take-mura, Yuki Uranishi, Haruo Takemura, and Hajime Nagahara. 2020. Detecting Learner Drowsiness Based on Facial Expressions and Head Movements in Online Courses. In *Proceedings of the 25th International Conference on Intelligent User Interfaces Companion*. 124–125.
- [13] Jacob Whitehill, Zewelanjji Serpell, Yi-Ching Lin, Aysha Foster, and Javier R Movellan. 2014. The faces of engagement: Automatic recognition of student engagement from facial expressions. *IEEE Transactions on Affective Computing* 5, 1 (2014), 86–98.