Prediction of Image Preferences from Spontaneous Facial Expressions

doi: 10.4036/iis.2022.A.02

Interdisciplinary Information Sciences Vol. 28, No. 1 (2022) 55–62

ISSN 1340-9050 print/1347-6157 online
DOI 10.4036/iis.2022.A.05

Body Posture Analysis for the Classification of Classroom Scenes

Yasuhiro HATORI[1,*], Taira NAKAJIMA[2] and Shinichi WATABE[2,3]

[1]*Research Institute of Electrical Communication, Tohoku University, Sendai 980-8577, Japan*
[2]*Graduate School of Education, Tohoku University, Sendai 980-8576, Japan*
[3]*Advanced Institute for Yotta Informatics, Tohoku University, Sendai 980-8577, Japan*

Student feedback is useful for teachers to improve their teaching. Although it is common to receive student ratings in universities, the low frequency of such feedback reduces the utility of the information. Using methods that do not rely on ratings can increase the frequency of feedback. We investigated whether the body posture of students can be used as an indicator of classroom engagement. In this paper, we estimated body posture from videos taken of students in the audience during a presentation and classified the scenes based on the postural similarity. The obtained clusters showed that body posture changed over time and did not return to the original state. A comparison between clusters at the beginning and end of the presentation showed that the standard deviation of head direction becomes large at the end, suggesting that body posture might reflect the degree of distraction. We discussed how body posture information facilitates teachers' reflection.

KEYWORDS: reflection, clustering, body posture, engagement

1. Introduction

Student feedback is effective for helping teachers improve their teaching. One of the most commonly used types of feedback in universities is student ratings [1]. However, the frequency at which students rate their teachers is low (e.g., once every six months), so there is not much information to help teachers improve their teaching. As a result, it can be difficult to grasp which part of the class the students are having difficulty with. The rating support system PF-NOTE enables students to provide feedback and record the timing of evaluations by pressing a button on a keypad [2]. PF-NOTE records a video of the teacher in order to understand the relationship between the ratings and teaching. However, PF-NOTE provides information only when a student makes a response. For example, no feedback is obtained if a student sleeps and does not respond. Therefore, even if the teaching needs to be improved, there is no feedback from the system to help the teacher. If it were possible to estimate the students' feedback by using a method that does not rely on responses, such a system could provide rich information about teaching.

Mental states are related to physiological indices. For example, heart rate variability can be used to estimate stress levels. The low-frequency component of heart rate variability relates to blood pressure variability, which reflects both sympathetic and parasympathetic nervous system activity. The high-frequency component relates to respiratory variability, which reflects parasympathetic nervous system activity [3]. Because the sympathetic nervous system becomes dominant when stressed, the ratio of low-frequency to high-frequency components becomes large. Although contact devices are generally used to measure physiological indices, it is not practical to use such equipment in a classroom setting. Therefore, the need for methods to estimate psychological states by using non-contact devices has attracted increasing attention [4–9]. With the development of computer vision technology, it has become possible to extract important information from images and videos [8–10]. The most common method is to analyze video images of students' facial expression or body posture to estimate their mental state. Delgado *et al.* captured facial images of students while they were solving a math problem that was displayed on a monitor. They succeeded in estimating the students' level of engagement from the direction of their heads [4].

In previous studies, generally, only one or two people are recorded to focus on their face or body in greater detail. Although it is possible to record each person in an experimental setting, this is not so easy in an actual classroom. In addition, it is crucial for teachers to understand the level of engagement of the majority of the students in order to improve their teaching. In this study, we estimated body posture from videos taken of more than ten students during seminar presentations. We classified the scenes based on the similarity of their body postures. Comparing the clusters at the beginning and end of the presentation showed that the standard deviation of head direction increased at the

Received November 10, 2021; Accepted March 11, 2022
2010 Mathematics Subject Classification: Primary 91C20, Secondary 68T10.
This study was supported by Advanced Institute of Yotta Informatics by MEXT, Japan.
*Corresponding author. E-mail: hatori@tohoku.ac.jp

end. This result suggests that the audience was distracted and focused their attention somewhere other than the presentation.

2. Materials & Methods

2.1 Videos

Two videos taken of the audience listening to classmate's seminar presentations were used for analysis. Faculty and students of the Graduate school of Education in Tohoku University participated in the seminar. Participants were 19 (video 1) and 26 (video 2), respectively. The speaker talked about their research using a screen located at the front of the room to show their presentation slides. The audience received a handout with a summary of the presentation. One camera was placed at the front of the room and recorded the audience at a frame rate of 29.97 fps. The video size was 3840×2160 pixels, but this was downscaled to 640×360 pixels for the sake of computational efficiency. The inter-frame averages of detected persons were 16.9 ± 1.7 for video 1 and 20.2 ± 1.6 for video 2 (mean $\pm$ standard deviation). On average, we detected about 80% of the audience even at this image resolution. Person detection failed when the person in the back was occluded by the person in the front. The analysis was conducted only on the scenes during the presentation. The question and answer session was excluded from the analysis. The length of the two videos was 8 min 45 s and 8 min 40 s, respectively.

2.2 Clustering

The clustering of classroom scenes was performed through the following procedures: (1) keypoint detection, (2) calculation of posture indices, and (3) classification of video frames.

2.2.1 Keypoint detection

The term "keypoint" refers to a characteristic point that expresses the posture of a given person. The keypoints of the audience in the videos were estimated using open-source software called MMPose [11]. First, the position of each person in each video frame was detected. Then, the coordinates of 17 keypoints were estimated for each person in the scene (Fig. 1, left). The 11 keypoints used in this study were eyes, nose, ears, shoulders, elbows, and wrists (Fig. 1, right). The six keypoints of the lower body (pelvis, knees, and ankles) were excluded because they were occluded by the tables, thereby lowering the estimation accuracy.

2.2.2 Calculation of posture indices

The estimated keypoints are obtained as coordinates with the upper left corner of an image as the origin. Since the size of the person depends on the distance from the camera, the coordinate of the keypoint can be changed due to two factors: the distance from the camera and the posture. Therefore, we calculated distance-independent indices from the coordinates of the keypoints. We obtained four angles from three keypoints for each person: the yaw and pitch angle of the head (Fig. 2) and the angle of both elbows. The angle θ ($0 \leq \theta < \pi$ [radian]) formed by three keypoints is defined as below:

$$\theta = \arccos\left(\frac{(x_1 - x_0)(x_2 - x_0) + (y_1 - y_0)(y_2 - y_0)}{\sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2}\sqrt{(x_2 - x_0)^2 + (y_2 - y_0)^2}}\right), \tag{1}$$

where (x_0, y_0), (x_1, y_1), and (x_2, y_2) are the coordinates of the three keypoints p_0, p_1, and p_2, respectively, with (x_0, y_0) as the origin. The three keypoints used for calculating the four angles are listed in Table 1.

The angle obtained from Eq. (1) is unable to describe the posture of a person. For example, the hand may be raised or lowered when the elbow angle is π (i.e., straight arm). We considered the relative positions of two keypoints and

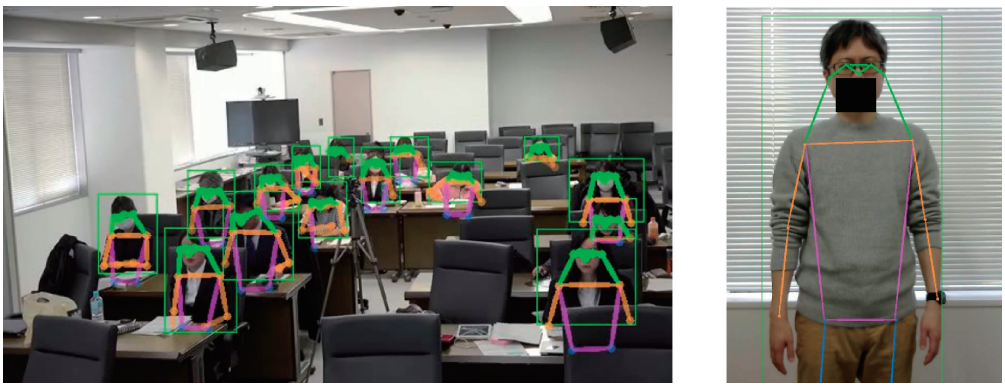

Fig. 1. Example MMpose output. (Left) The green rectangles show the locations of the detected persons. The dots indicate the estimated keypoints for each person. (Right) The 11 keypoints in the upper body were used to calculate body posture.

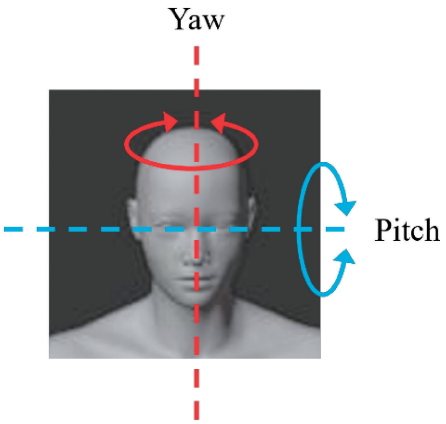

Fig. 2. The yaw angle of the head is rotation around the midline (red dotted line), and the pitch angle is rotation around the axis through the left and right ears (blue dotted line).

Table 1. Three keypoints used for the calculation of each angle.

	p_0	p_1	p_2
Head yaw angle	Nose	Right eye	Right shoulder
Head pitch angle	Nose	Left shoulder	Right shoulder
Left arm	Left elbow	Left shoulder	Left wrist
Right arm	Right elbow	Right shoulder	Right wrist

converted each angle to an index that expresses body posture. An index describing the horizontal direction of the head (H_x) was calculated as follows:

$$H_x = \frac{\theta}{\pi}. \tag{2}$$

An index describing the vertical direction of the head (H_y) was calculated as follows:

$$H_y = \begin{cases} (|\theta - \pi| + \pi)/2\pi & if\ p_0(y) > (p_1(y) + p_2(y))/2 \\ \theta/2\pi & otherwise \end{cases}. \tag{3}$$

An index of arm posture (E) was calculated as follows:

$$E = \begin{cases} (\theta + \pi)/2\pi & if\ p_2(y) > p_1(y) \\ (-\theta + \pi)/2\pi & otherwise \end{cases}. \tag{4}$$

Following the above three equations, the range of all indices was constrained from zero to one. The meanings of each index are listed in Table 2.

2.2.3 Classification of video frames

For each video frame, a histogram (bin width: 0.1) of the four posture indices of the audience constitutes a feature vector (Fig. 3). The feature vector has 40 dimensions (4 [posture indices] × 10 [the number of bins]). Principal component analysis was applied to the feature vectors to increase the computational efficiency, and up to the 20th principal component was used for further analysis. About 93 and 96% of the data variance for the two videos, respectively, could be explained by up to the 20th principal component.

We performed cluster analysis of the scenes by using a Gaussian mixture model with variational inference [12]. This model assumes that (1) a discrete probability distribution determines a cluster, and (2) the mixture of Gaussian distributions generates the data. We chose the Dirichlet process (DP) as a discrete probability distribution. Although DP

Table 2. Meanings of the values for each index.

Index	Minimum value (0)	Maximum value (1)
H_x	Facing left	Facing right
H_y	Facing down	Facing up
E	Lowering hand with a straight arm	Raising hand with a straight arm

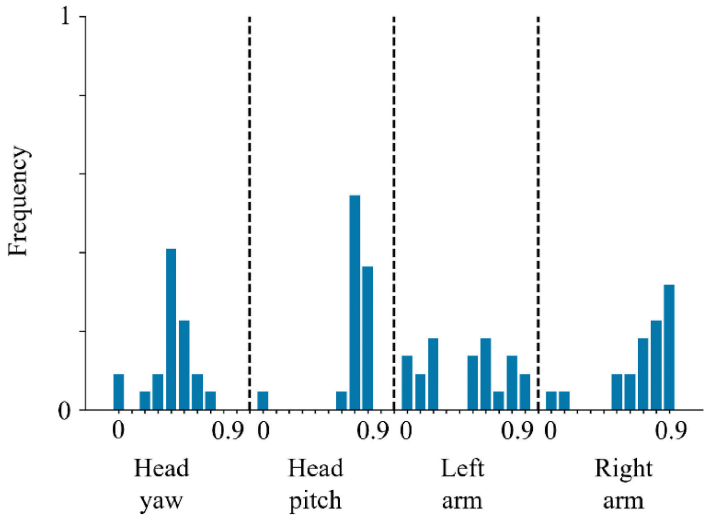

Fig. 3. Example histogram of the posture indices. The histogram of the four posture indices (bin width 0.1) was used as a feature vector for characterizing each video frame.

can theoretically handle infinite dimensions, in practice, a maximum number of clusters was defined. Because the number of clusters for a given set of data was unknown, we changed the maximum number of clusters from 1 to 20 and sought the optimal number based on the Bayesian information criterion (BIC; see Sect. 2.3). The parameters of Gaussian distributions were estimated by variational inference to maximize approximated posterior probability. We used the scikit-learn library for Python to run the model. The default values of all parameters were used except for the maximum number of clusters.

2.3 Determination of the number of clusters

The greater the number of clusters is, the smaller the sum of the distances from the center of each cluster to the data points is. In other words, the goodness of fit always improves when the number of clusters increases. In the most extreme case, each data point constitutes a cluster, where the sum of the distances between the cluster centers and data points is zero. To avoid such a trivial solution, we evaluated the number of clusters based on the BIC [13] as follows:

$$b = -2 \log L + k \log n, \tag{5}$$

where L is the likelihood, k is the number of clusters, and n is the number of data. The first term on the right-hand side of Eq. (5) represents the goodness of fit of the model; the second term evaluates the number of clusters as a penalty. The value of the second term is proportional to k because n is a constant. In this study, the maximum number of clusters was changed from 1 to 20. The appropriate number of clusters was obtained by finding a value k that minimizes b.

3. Results

3.1 Number of clusters

Figure 4 shows BIC plotted as a function of the number of clusters. In video 1, BIC rapidly decreased to 6 clusters and reached a minimum of around 7–12 clusters. In video 2, BIC decreased rapidly until the number of clusters reached 4, after which it decreased more slowly. Although BIC decreased as the number of clusters increased, the data were concentrated in several clusters. For this reason, the number of clusters was set to 8 for the subsequent analysis.

3.2 Clustering

Figure 5 shows the clustering results for the two videos. Different colors indicate different clusters. If the clusters changed within 1 s, the cluster of the previous time was considered to be continuous. It should be noted that the same color in videos 1 and 2 does not mean that they belong to the same cluster because clustering was performed independently for each video. In video 1, cyan and red appeared first, but only red appeared again, and then only briefly. In video 2, brown and orange appeared until around the middle but did not appear again in the second half. These results suggest that the body posture of the audience changed over time and did not return to its original state.

3.3 Comparison of features

The audience may look at the screen at the front of the room at first, but as time passes, they tend to look elsewhere. Therefore, we compared the standard deviation of the head yaw angle. Figure 6 shows the average standard deviation

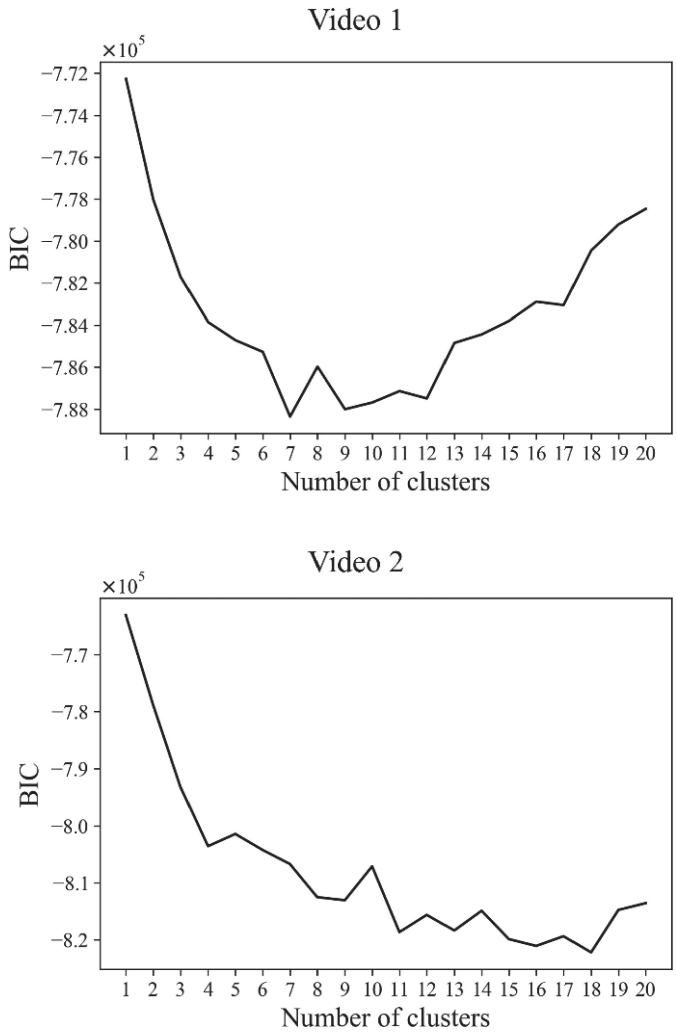

Fig. 4. BIC is plotted as the function of the number of clusters (top: video 1, bottom: video 2).

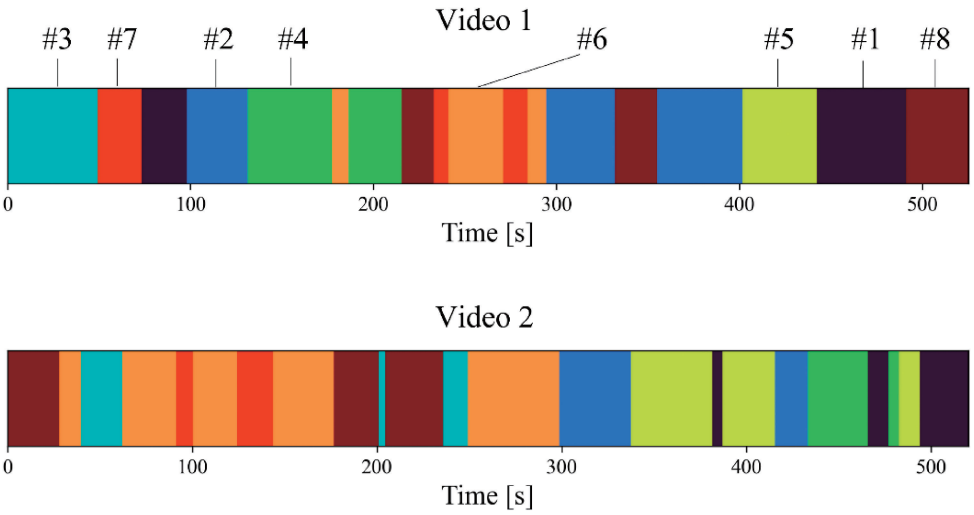

Fig. 5. Clustering results for the two videos. The horizontal axis represents time, and the colors represent clusters. Because clustering was performed independently for each video, the same color in videos 1 and 2 does not mean that the clusters have the same properties. Correspondence between the colors and the cluster numbers is as follows: 1 (dark blue), 2 (blue), 3 (cyan), 4 (green), 5 (yellow), 6 (orange), 7 (red), and 8 (brown).

of the head yaw angle for each cluster. We compared the clusters shortly after the start of the presentation (video 1: cluster 3, cyan in the upper panel of Fig. 5; video 2: cluster 8, brown in the lower panel of Fig. 5) with the clusters just before the end of the presentation (video 1: cluster 8, brown in the upper part of Fig. 5; video 2: cluster 1, dark blue in

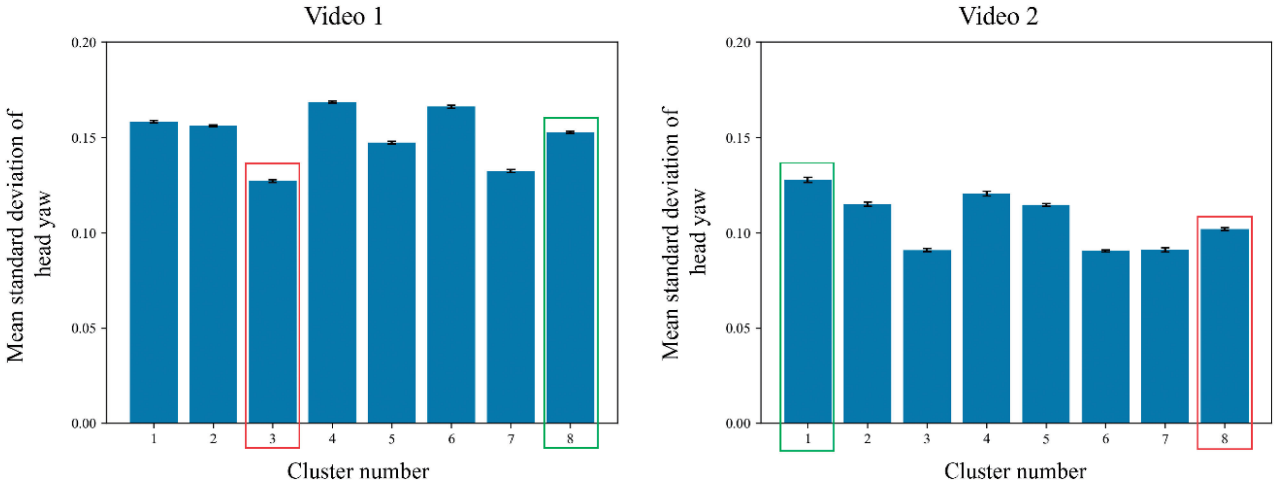

Fig. 6. Average standard deviation of head yaw angle for each cluster. The red box shows the cluster just after the start of the presentation, and the green box shows the cluster just before the end. The error bars are the standard errors of the mean.

the lower part of Fig. 5). The standard deviation tended to be smaller in the clusters shortly after the start of the presentation compared with the clusters just before the end. A comparison using an unpaired t-test showed that the standard deviation of the head yaw angle was significantly larger in the cluster shortly after the start of the presentation for both videos (video 1: $t(3685) = -28.91$, $p < 0.01$; video 2: $t(3784) = -17.92$, $p < 0.01$). This result reflects the tendency of the audience to look at different locations as time passes, suggesting that head orientation can be an indicator of distraction.

4. Discussion

We classified the scenes of the presentation into eight clusters based on the body posture of the audience. The clustering results suggested that body posture changed over time and did not return to its original state. More specifically, the standard deviation of head yaw orientation became larger at the end of the presentation compared with the beginning, which may reflect to the audience's waning attention. In the following, we discuss the future directions of this study and compare ours and other methods.

4.1 Number of clusters

In this study, we chose eight clusters based on BIC. The videos used for analysis contained only presentation scenes. The actual classroom scene includes a wide variety of situations: discussion, group work, and so on. The number of clusters could easily differ depending on the type of class, and the characteristics of each cluster could change as well. Collecting videos of various classroom scenes would lead to further understanding of what body postures are helpful for estimating mental states.

4.2 Comparison to other studies

There are many approaches for human pose recognition [14–17]. Some studies used a depth camera (e.g., Microsoft Kinect) to estimate human pose ([14, 18]). Although depth information is useful for representing three-dimensional poses, a single Kinect sensor is unreliable for human pose estimation [18]. The previous study suggested that the fusion of three Kinect sensors improves human pose estimation for a single person [18], implying that enormous Kinect sensors are necessary to cover the entire classroom.

Advances in convolutional neural networks enabled the estimation of human poses only from video images [19, 20]. Lin and colleagues proposed the model to classify activities observed in classroom scenes (e.g., raising hands; [19]). Their model estimated joint positions from video images using OpenPose [21], a well-known software for keypoint detection. Since OpenPose estimates keypoints without person detection, keypoints of multiple persons may be incorrectly connected (Fig. 7). In the previous study [19], person detection was performed after keypoint estimation to remove incorrectly joined keypoints. In that way, keypoints for more than two persons could be lost. In this study, keypoint estimation was performed on the detected persons, which can accurately estimate the pose information of many persons.

Gaze position is another key indicator of one's attention. Bixler and D'Mello used an eye-tracking device to detect mind wandering [22]. The disadvantage of this method is that each student needs to wear an eye-tracking device. Another way to estimate eye movements is to use video images of a face captured with high resolution [15]. Although high-resolution video images are useful for estimating gaze position, setting up a camera for each student in a

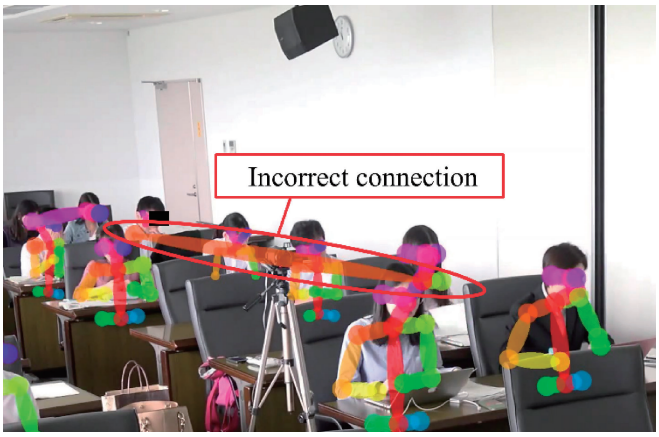

Fig. 7. Example OpenPose output. Detected keypoints from several persons are connected incorrectly, as indicated by the red circle.

classroom is not practical. One solution to such a problem is the coordination between the head and the eyes [23, 24]. It has been reported that the head and the eyes move in a coordinated manner to achieve efficient gaze shift, indicating that one can estimate eye position by observing head orientation. Gaze position would be estimated accurately by head orientation obtained from video images with the modeling of eye-head coordination.

4.3 Class reflection

We showed that classroom scenes can be classified by body posture and that some of the clusters differed in their characteristics, including the standard deviation of head direction. When the students look somewhere other than the screen at the front of the classroom, the standard deviation should be large. Such a situation may be related to the waning attention of the students. Teachers can obtain information useful for reflecting on their teaching by analyzing the body postures of their students. For example, it is possible to examine whether a teacher successfully attracts students' attention from the deviation of head orientation with a certain threshold. The teacher can know when the distraction began by examining the temporal changes in students' head orientations. In this case, clustering analysis is not mandatory because the teacher can follow the temporal changes of head orientations. It would be interesting to investigate whether student ratings can be improved by incorporating this type of indirect feedback based on body posture information.

4.4 Assessment of engagement

A previous study has reported the correlation between motivation for learning and the frequency of paying attention and listening behaviors [25]. In the previous study, the frequency of such behaviors was not quantitatively evaluated because questionnaires were used for the assessment [25]. We showed that the deviation of head orientations might correlate with distraction. Head orientation would enable teachers to assess paying attention and listening behaviors quantitatively. However, the estimation accuracy of head orientation in the pitch direction could be worse because cameras must be placed near the ceiling in an actual classroom. A way to deal with this problem is to use multiple cameras [17]. Head movements in the pitch direction can be obtained accurately with the videos taken from the top of the side of a room. It would be interesting to develop a method to deal with multiple cameras for accurate head orientation estimation.

5. Conclusion

We analyzed scenes during a presentation to obtain information that can contribute to improving teaching. We used a Gaussian mixture model with variational inference for classification. We found that the standard deviation of the head yaw angle was larger just before the end of the presentation compared with that at the start. Variation in head orientation may reflect the waning attention of the audience near the end of the presentation. Body posture analysis can provide information related to the mental state of students in terms of distraction, enabling teachers to improve their teaching.

Acknowledgments

This study was supported by Yotta Informatics Project by MEXT, Japan.

REFERENCES

[1] Cohen, P. A., "Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies," *Review of Educational Research*, **51(3)**: 281–309 (1981).

[2] Nakajima, T., "EduReflex: A Light Weight Class Reflection Tool for Teaching Improvement through Video-Recording with "Clickers"," *Proceedings of the IMSA 11th IASTED International Conference*, 18–23 (2007).

[3] Pomeranz, B., Macaulay, R. J., Caudill, M. A., Kutz, I., Adam, D., Gordon, D., Kilborn, K. M., Barger, A. C., Shannon, D. C., Cohen, R. J., and Herbert, B., "Assessment of autonomic function in humans by heart rate spectral analysis," *American Journal of Physiology*, **248(1 Pt 2)**: H151-3 (1985).

[4] Delgado, K., Origgi, J. M., Hasanpoor, T., Yu, H., Allessio, D., Arroyo, I., Lee, W., Betke, M., Woolf, B., and Bargal, S. A., "Student Engagement Dataset," *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 3621–3629 (2021).

[5] Komori, M., and Nagaoka, C., "The relationship between seating locations and instructor-student entrainment in a classroom," *Kansei Engineering International Journal*, **11(4)**: 179–182 (2012).

[6] Madsen, J., Júlio, S. U., Gucik, P. J., Steinberg, R., and Parra, L. C., "Synchronized eye movements predict test scores in online video education," *Proceedings of the National Academy of Sciences of the United States of America*, **118(5)**: 1–9 (2021).

[7] Whitehill, J., Serpell, Z., Lin, Y. C., Foster, A., and Movellan, J. R., "The faces of engagement: Automatic recognition of student engagement from facial expressions," *IEEE Transactions on Affective Computing*, **5(1)**: 86–98 (2014).

[8] Sato, Y., Horaguchi, Y., Vanel, L., and Shioiri, S., "Prediction of image preferences from spontaneous facial expressions," *Interdisciplinary Information Sciences*, in press.

[9] Shioiri, S., Sato, Y., Horaguchi, Y., Muraoka, H., and Nihei, M., "Quali-Informatics in the Society with Yotta Scale Data," *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*, 1–4 (2021).

[10] Oodaira, K., Miyazaki, T., Sugara, Y., and Omachi, S., "Importance estimation for scene texts using visual features," *Interdisciplinary Information Sciences*, in press.

[11] MMPose: https://github.com/open-mmlab/mmpose.

[12] Blei, D. M., and Jordan, M. I., "Variational inference for Dirichlet process mixtures," *Bayesian Analysis*, **1(1)**: 121–143 (2006).

[13] Schwarz, G., "Estimating the dimension of a model," *Annals of Statistics*, **6**: 461–464 (1978).

[14] Cippitelli E., Gasparrini S., Gambi, E., and Spinsante, S., "A human activity recognition system using skeleton data from RGBD sensors," *Computational Intelligence and Neuroscience*, **2016**: 1–14 (2016).

[15] Matsumoto, Y., Ogasawara, T., and Zelinsky, A., "Behavior Recognition Based on Head Pose and Gaze Direction Measurement," *Proceedings of the 2000 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2127–2132 (2000).

[16] Ba, S. O., and Odobez, J. M., "Recognizing visual focus of attention from head pose in natural meetings," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, **39(1)**: 16–33 (2009).

[17] Voit, M., and Stiefelhagen, R., "Tracking Head Pose and Focus of Attention with Multiple Far-field Cameras," *Proceedings of the 8th International Conference on Multimodal Interfaces*, 281–286 (2006).

[18] Ryselis, K., Petkus, T., Blažauskas, T., Maskeliūnas, R., and Damaševičius, R., "Multiple kinect based system to monitor and analyze key performance indicators of physical training," *Human-centric Computing and Information Sciences*, **10**: 51 (2020).

[19] Lin, F. C., Ngo, H. H., Dow, C. R., Lam, K. H., and Le, H. L., "Student behavior recognition system for the classroom environment based on skeleton pose estimation and person detection," *Sensors*, **21(16)**: 5314 (2021).

[20] Xu, X., and Teng, X., "Classroom attention analysis based on multiple euler angles constraint and head pose estimation," *Lecture Notes in Computer Science*, **11961**: 329–340 (2020).

[21] Wei, S. E., Ramakrishna, V., Kanade, T., and Sheikh, Y., "Convolutional Pose Machines," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4724–4732 (2016).

[22] Bixler, R., and D'Mello, S., "Automatic gaze-based user-independent detection of mind wandering during computerized reading," *User Modeling and User-Adapted Interaction*, **26**: 33–68 (2016).

[23] Nakashima, R., Fang, Y., Hatori, Y., Hiratani, A., Matsumiya, K., Kuriki, I., and Shioiri, S., "Saliency-based gaze prediction based on head direction," *Vision Research*, **117**: 59–66 (2015).

[24] Fang, Y., Emoto, M., Nakashima, R., Matsumiya, K., Kuriki, I., and Shioiri, S., "Eye-position distribution depending on head orientation when observing movies on ultrahigh-definition," *ITE Transactions on Media Technology and Applications (MTA)*, **3(2)**: 149–154 (2015).

[25] Fuse, M., Kodaira, H., and Ando, F., "Positive class participation by elementary school pupils: Motivation and differences in grade and gender," *Japanese Journal of Educational Psychology*, **54**: 534–545 (2006).