2025 IEEE/SICE International Symposium on System Integration (SII)
January 21-24, 2025, Munich, Germany

Behavior Monitoring System Leveraging Human Pose Estimation

Ryo Mitoma[1], Takayuki Mukaeda[1,2], Keisuke Shima[1], Haruto Kai[1], Masayuki Suzuki[3] and Keiji Kato[3]

Abstract— **In recent years, human-centric computer vision technologies, such as human pose estimation and action recognition, have garnered significant attention. This study, therefore, proposes the use of a classroom behavior monitoring system that integrates off-the-shelf human pose estimators to compute behavioral features from the video data. The extracted features include continuous values (body movements, head angles, etc.) and discrete features (timing of hand-raising), which were not leveraged in prior studies. The system can operate in real-time on consumer-grade GPUs. This makes it useful for real-time feedback during classes and post-class analysis of video recordings. The proposed system was experimentally validated using a motion capture dataset featuring scenes with occlusions. We demonstrated its accuracy and real-time performance, along with the challenges of capturing full-body poses in complex environments. Additionally, we applied this system to a novel learning analytics task: estimating cognitive and non-cognitive abilities from videos using real-world data. As part of our analysis, we processed a video from an elementary school classroom with our system and linked the computed metrics to self-reported skills. These experiments demonstrated promising results, highlighting the potential of evaluating student skills solely based on visual cues.**

I. INTRODUCTION

Recently, human-centric computer vision technologies, including human pose estimation and action recognition have attracted considerable interest. These technologies are important in the field of learning analytics, where they are used to analyze student behavior and activities during class, thereby aiding educational guidance. One of the primary objectives of video-based learning analytics is attention estimation, which employs techniques such as object recognition [1] and head pose estimation [2]. Another crucial objective is to identify the actions that occur within the classroom settings including reading, writing, and the use of mobile devices to analyze student behaviors [3]. Yu et al. [4] proposed an LLM-based improvement plan generation method based on action recognition and underscoring potential advanced AI to enhance educational outcomes.

However, previous studies have primarily focused on the attention or specific actions, thus failing to capture the detailed information necessary to evaluate students' concentration, noncognitive abilities, and interest in classes. To address this issue, our method takes full advantage of human pose-estimation capabilities, thereby enabling better integration within the learning analytics framework. In this paper, we propose a system that utilizes off-the-shelf human pose estimators to analyze videos and extract behavioral features from pose data. These features include continuous values, such as body movement and head angles, as well as discrete features, such as hand-raising timings. They encompass attention and action estimation while enabling more detailed behavioral analysis. The system operates in real-time on consumer-grade GPUs. Therefore, it is useful for providing real-time feedback during class and conducting post-class analyses. The accuracy and real-time performance were validated in challenging environments in which capturing an individual's full body is difficult. This method was also applied to a novel learning analytics task: estimation of cognitive and non-cognitive abilities from videos.

The key contributions of this study are as follows:

- Using only off-the-shelf machine-learning methods, the system computes valuable features for human behavioral analysis.
- The accuracy and real-time performance of the proposed system was demonstrated both quantitatively and qualitatively, highlighting its potential to address real-world problems.
- The application of the proposed system to real-world problems was demonstrated through a specific example: estimating cognitive and non-cognitive abilities from videos.

II. RELATED WORKS

A. Estimating Student Behavior and Attitudes from Visual Data

Attention estimation and action recognition are critical tasks for interpreting student behaviors and attitudes from media data.

Attention is an indicator of whether students are actively trying to learn in class and is related to academic performance [5]. Previous studies used features such as gaze [2], head movement [6] extracted from classroom videos and facial movements obtained from Kinect [7]. In addition, spatial features, such as the optical flow computed from the RGB cameras were employed [8].

Action recognition determines the behaviors that students exhibit during class. Many previous studies used end-to-end machine-learning methods to estimate actions, such as reading, writing, and using a phone, from an input video. Some of these methods employ SlowFast-based models [9] or MotionBERT [4].

[1]Ryo Mitoma, Takayuki Mukaeda, Keisuke Shima and Haruto Kai are with Graduate School of Environment and Information Sciences Yokohama National University, Kanagawa 240-8501, Japan mitoma-ryo-ph@ynu.jp

[2]Takayuki Mukaeda is also with Kanagawa Institute of Industrial Science and Technology (KISTEC), Kanagawa 243-0435 705-1, Japan

[3]Masayuki Suzuki and Keiji Kato are College of Education, Yokohama National University, Kanagawa 240-8501, Japan

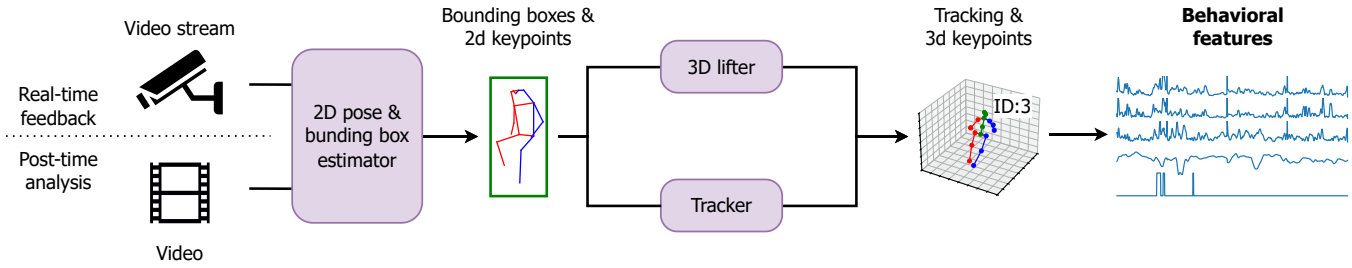


Fig. 1. System overview. An input single-view video passes through a one-stage 2D pose estimator and ID tracking. Subsequently, the 2D poses are transformed into 3D pose sequences by a pose lifter for each person in the video, with behavioral features computed from 3D poses.

While most of these methods aim to analyze whether students are concentrating in class, this study aims to explore and extract the characteristics of student behavior beyond the scope of concentration.

B. Deep-Learning-Based Real-Time Human Pose Estimation

Recent advances in machine learning have led to the development of real-time methods to detect or locate people using images and videos.

Two-dimensional (2D) human pose estimation is a task that involves identifying keypoints, such as joints and body parts, in images or videos. The main objective of this task is to estimate the keypoints for multiple people in a single image. Existing methods are generally classified into two categories, namely two-stage methods [10], [11], that first identifies human regions using an object detector and then estimates poses for each region, and one-stage methods [12], [13], that obtains keypoints for multiple people directly from an input. One-stage methods can process an image in constant time, regardless of the number of people in it. For example, RTMO [12] is a state-of-the-art one-stage model, that has been shown to outperform the two-stage RTMPose [10] in inference time with the number of people exceeding three to four [12].

Three-dimensional (3D) human pose estimation is a task that estimates the 3D coordinates of people in images or videos. This task is often solved by lifting the results of the 2D pose estimation to 3D. It can be configured in various ways, including monocular and multiview settings. While multiview settings are inherently more accurate than monocular settings, some studies have shown that monocular inputs yield accurate estimates. For example, MotionBERT [14] and P-STMO [15] use a masking strategy for input 2D poses that improves robustness to noisy inputs.

Object tracking involves tracking each object in a video and assigning the same ID to the same instance. Recent methods solve this task by tracking bounding boxes [16], [17] which are the results of the object detection. This approach, known as tracking-by-detection, does not rely on direct image input and balances accuracy and speed more effectively than end-to-end methods.

C. Cognitive and Non-Cognitive Abilities

Cognitive abilities (cognitive skills) included language, memory, and reasoning [18], which can be measured by academic or IQ tests. By contrast, non-cognitive abilities (non-cognitive skills) such as perseverance and self-control cannot be measured using these tests [19]. Non-cognitive abilities are believed to contribute to the development of cognitive abilities [20] and are considered critical to life outcomes [21]. To the best of our knowledge, no attempts have been made to estimate abilities using videos.

III. Proposed Method

In this section, we describe the structure of the proposed pose estimation module and the process of computing behavioral features. The pose estimation module employed off-the-shelf machine learning models to track the 3D poses of individuals in a video. Using these 3D poses, metrics that focus on the features related to keypoint movements can be computed, in conjunction with those covered in previous research, thereby allowing for comprehensive and detailed behavioral analysis.

A. 3D Pose Estimation and Tracking

Figure 1 shows an overview of the system. A 2D pose estimator was applied to the input video to obtain the bounding boxes and 2D keypoints. The RTMO [12] was used for 2D pose estimation, which allows for real-time estimation, even in crowded scenes. For each bounding box, the IDs were assigned using the tracking model ByteTrack [17], which supports real-time tracking. The 2D keypoints were processed using 3D pose-estimation model. MotionBERT [14] is used to obtain 3D keypoints owing to its high accuracy and robustness. The 3D pose estimator processed buffered 2D pose sequences equal in length to those of the model sequence that enables semi-real-time inference.

B. Behavioral Features Calculation

Once the 3D keypoints for each individual were obtained, the behavioral features were calculated using the 3D keypoints. Our method calculates discrete features such as hand-raising and continuous features such as the head, leg, and full body movements and head angles. Coordinates of the joint j at frame t is denoted by $(x_{j,t},\ y_{j,t},\ z_{j,t}) = \mathbf{p}_{j,t} \in \mathbb{R}^3$, where z coordinates represent the vertical axes. The movement features were calculated by taking the difference from the previous frames, which indicates the activity level of a person. The hand-raising feature in frame t is defined as

TABLE I

RESULTS FOR THE PANOPTIC SUB-DATASET

	# of ID reassignments		MPJPE (2D, px)		MPJPE (3D, mm)	
View	0	1	0	1	0	1
171026_cello3	0	0	44.6	25.2	113.3	119.8
161029_piano1	1	7	30.9	22.1	100.2	136.0
160906_band4	0	22	71.0	25.6	167.9	211.4
160906_ian2	7	6	54.3	22.5	123.7	116.8
170915_office1	4	3	35.1	36.8	143.7	161.2
170407_office2	5	4	59.7	26.1	129.5	160.6
161029_build1	34	5	51.0	20.7	168.6	149.5
161029_car1	0	0	25.8	15.6	155.0	112.9
Average	6.4	5.9	46.5	24.3	137.7	146.0

$$F_{\text{hand},t} = \begin{cases} 1 & \text{if } \begin{cases} z_{\text{wrist},t} > z_{\text{shoulder},t} & \text{and} \\ \|\mathbf{p}_{\text{wrist},t} - \mathbf{p}_{\text{opp\_shoulder},t}\|_2 > \tau \end{cases} \\ 0 & \text{otherwise} \end{cases}$$

where $\mathbf{p}_{\text{opp\_shoulder},t}$ is the shoulder keypoint opposite $\mathbf{p}_{\text{wrist},t}$ and τ is the threshold. This formulation can distinguish hand-raising from similar actions such as face touching. In later experiments, we set τ to the shoulder width. The pitch angle of the head at frame t is calculated based on the spine and neck keypoints. Formally,

$$F_{\text{head},t} = \angle(\mathbf{v}_{\text{spine},t}, \mathbf{v}_{\text{neck},t}) \,,$$

where

$$\mathbf{v}_{\text{spine},t} = \mathbf{p}_{\text{neck},t} - \mathbf{p}_{\text{chest},t} \,,$$
$$\mathbf{v}_{\text{neck},t} = \mathbf{p}_{\text{head},t} - \mathbf{p}_{\text{neck},t} \,.$$

The hand-raising feature aids in analyzing student engagement, while the pitch angle provides insight into a student's focus, such as whether they are looking at the blackboard or reading. This complements the engagement and attention analyses conducted in prior studies. Additionally, the analysis of body movements—an aspect often overlooked in previous research, enables a deeper and more comprehensive analysis of video data.

When calculating behavioral features from the estimated 3D poses, adjustments are required for the size variations caused by depth ambiguity. To address this, the average distance between the shoulders and neck was calculated for each frame, and all keypoints were normalized using this value. Preliminary experiments confirmed that this approach resulted in fewer posture collapses compared with normalization using other keypoint distances, such as shoulder width.

IV. EXPERIMENTS

A. Evaluation of the Proposed System

In the experiment, the proposed system was evaluated using a benchmark dataset, which provided ground-truth poses. In situations with numerous people and objects in the classroom, the accuracy of tracking and pose estimation may decrease due to occlusions. To verify the accuracy of the proposed system in these situations, we created a sub-dataset from a 3D motion-capture dataset containing scenes with occlusions caused by objects.

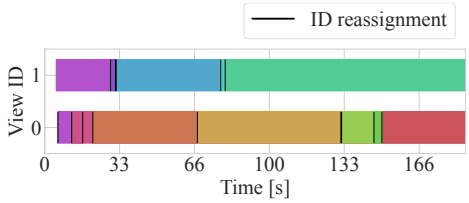


Fig. 2. Timeline showing the occurrence of ID reassignments for the same person in "170915_office1", indicating that tracking is disrupted momentarily. This frequently happens when an individual is nearly out of the video frame.

1) Setup: The Panoptic dataset [22] was used for 3D pose estimation using various scenes captured from multiple camera views. For this experiment, we extracted the following sequences from the Panoptic dataset: "171026_cello3," "161029_piano1," "160906_band4," "160906_ian2," "170915_office1," "170407_office2," "161029_build1," and "161029_car1". The total duration of these sequences was approximately 30 min. Occlusions occurred more frequently in these sequences, owing to the presence of musical instruments, office supplies, and toys. In this experiment, view IDs 0 and 1 were used for the HD cameras.

We used the number of ID reassignments as the evaluation metric in person tracking (which occurs when tracking is interrupted and redetected), mean per joint pose error (MPJPE) for 2D pose estimation, and P-MPJPE (MPJPE with rigid transformation) for 3D pose estimation. The RMSE was calculated to evaluate the behavioral features by equally distributing the output values over ten levels (that is, 0 for the exact estimation and 4.06, chance). This index allows the evaluation of the the relative accuracy of the predicted values while reducing the influence of outliers. In all the evaluations, keypoints with a confidence value of -1 (that is, invalid, or not detected) were excluded.

The experiment was conducted using a machine with an Intel Core i7-10700K CPU and NVIDIA GeForce RTX 3090 GPU. We used the same hyperparameters as those used in the original studies for all the models.

2) Results: Table I presents the experimental results. Person tracking was successful. However, ID reassignments occur in many cases. Figure 2 represents the timeline of ID reassignments for one person in the sequence "170915_office1." We observed that even when ID reassignment occurs in one view, it does not occur in another. This finding suggests that stable tracking can be achieved by integrating tracking results from multiple views, which is a topic for future research.

For the pose estimation results, the values of the evaluation metrics were higher than those obtained in previous studies on other datasets (e.g., [14]). Table II shows that the results

TABLE II

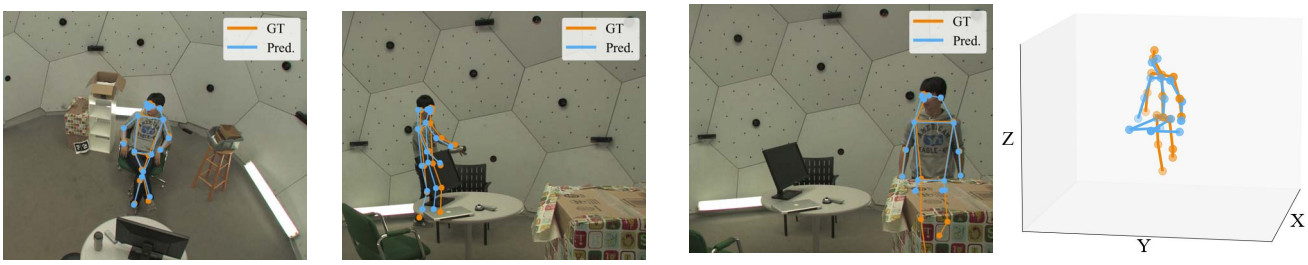
POSE ESTIMATION RESULTS PER BODY PART

Body part	Head	Arms	Lower body
MPJPE (2D, px)	22.6	18.6	68.0

TABLE III

RMSE EVALUATION OF BEHAVIORAL FEATURES IN 10 LEVELS

Full-body movement	Foots movement	Head movement	Head angle
2.88	3.55	3.10	3.05



(a) Success without occlusion (b) Success with occlusion (c) Failure with occlusion

Fig. 3. Qualitative results for the Panoptic sub-dataset

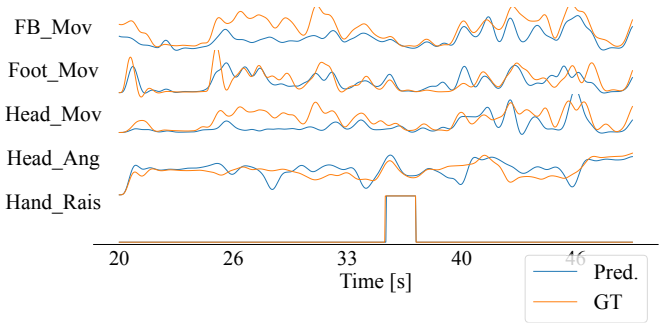


Fig. 4. Calculated and ground-truth behavioral features for a person. FB, _Mov, _Ang and _Rais denotes full-body, movement, angle and raising, respectively.

Fig. 5. Layout of the camera system

exhibit significant differences in 2D-pose estimation accuracy for each body part. This is possibly due to less occlusion in the upper body in an indoor environment. As shown in Table III, features computed from the legs, which have more occlusions, have low accuracy and are close to a chance rate of 4.06. In contrast, features computed for the entire body and the head were obtained with high accuracy. Therefore, focusing on behavioral features of the head, hands, and other parts of the upper body, which exhibit low errors, is expected to improve the accuracy of applications.

For a qualitative evaluation, Figure 3 shows examples of the estimated keypoints, and Figure 4 illustrates the calculated behavioral features. For pose estimation, scenes with occlusions contained more pose-estimation errors than those without. In particular, when half of the body is hidden, errors are caused by fitting all keypoints within the bounding box (Figure 3c). A possible solution to this problem is to explicitly account for keypoint visibility when estimating human poses or computing behavioral features. For behavioral feature calculations, an approximate agreement was observed between the predicted and ground truth.

By measuring the operating speed of the proposed system, we observed that 2D pose estimation component, which is a bottleneck, runs at 64.93 fps. Other aspects such as 3D pose estimation and tracking exceeded 1000 fps. From these results, we conclude that the proposed system can run on a single machine equipped with a consumer-grade GPU and that real-time feedback can be provided in the classroom.

B. Application Showcase

Finally, as a real-world application of the proposed system, we attempted to estimate students' cognitive and noncognitive abilities based on their behavior.

1) Setup: Data were collected from elementary schools in Yokohama, Japan. The recording was conducted in HD resolution at 30 fps using four Canon VB-S32D cameras, and Figure 5 shows their layout. Due to an inability to capture all the students simultaneously, some students were obscured by screen edges. A one-hour lesson was extracted from these recordings and used as the dataset for experimental analysis. Students in this classroom had been previously assessed for cognitive and non-cognitive abilities through achievement tests and questionnaires, and these normalized scores were used as estimation targets. Behavioral features were calculated for each student using the proposed system. The following analysis used data from nine students, excluding those with low detection reliability. The calculated feature values were averaged over time for each student and standardized across students. The dataset was then constructed by manually linking the behavioral

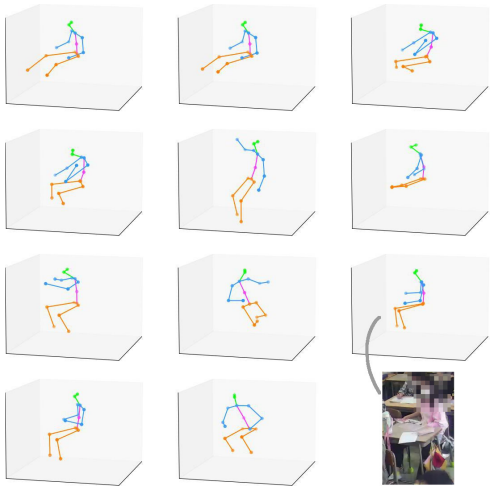


Fig. 6. Estimated 3D poses

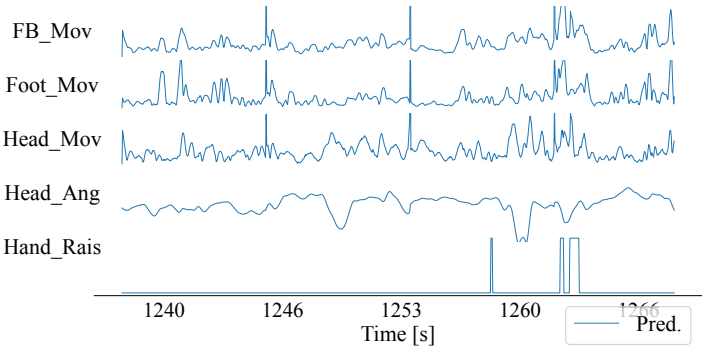


Fig. 7. Calculated behavioral features for a student

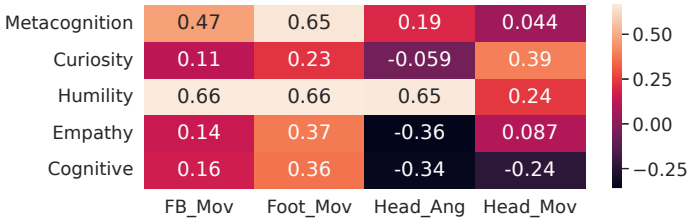


Fig. 8. Result of correlation analysis

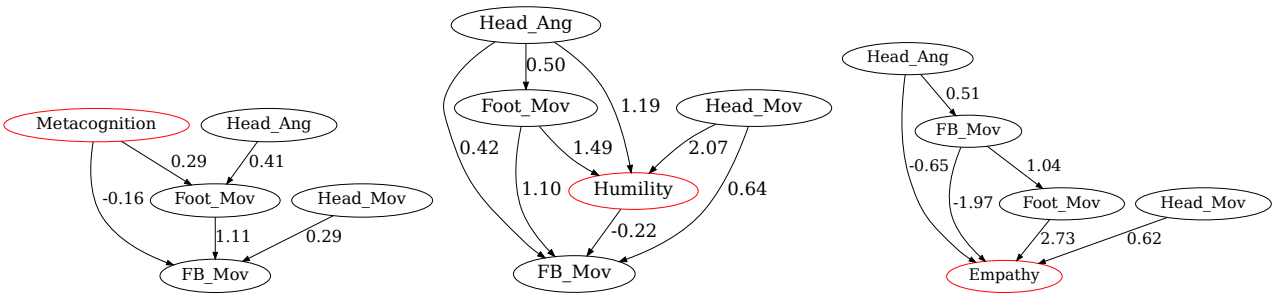


Fig. 9. Results of causal analysis

features with students' cognitive and non-cognitive ability scores. ID reassignment and broken tracking were corrected using the temporal medians of bounding box coordinates, assuming that the seat position did not change during the recording. Although mapping was performed manually in this experiment, it can be automated using methods such as feature vectors of faces [2] if each student's facial data is available. After obtaining these results, two types of analyses were performed: causal analysis using DirectLiNGAM [23] and correlation analysis.

2) Results: Figures 6 and 7 show the results of the 3D pose estimation and timeline of the estimated behavioral features, respectively. We obtained plausible results for both 3d pose and behavioral features. For detection and tracking, we found that one camera view covered about half of the classroom with moderate confidence scores, and when checking 10-minute spans, almost all IDs were reassigned at once due to teachers or students moving around. This indicates the need for multi-view camera setups to improve tracking accuracy and reduce the reliance on manual refinements. Figure 8 shows the Pearson correlation coefficients between the estimated behavioral features and students' abilities, indicating that some of the calculated features have weak correlations with the abilities. Figure 9 shows the results of the causal discovery. No causality was found for behavioral features that are not listed here. Humility and empathy were the only two factors identified in cases where causality from behavioral features to ability was confirmed.

Assuming the results are correct, the positive coefficient of causality from head movement (`Head_Mov`) to humility and empathy indicates that behaviors such as nodding and looking around the environment, are likely to increase humility and empathy along with head movements. Conversely, the negative causality from humility to full-body movement (`FB_Mov`) suggests that higher humility is expressed as less body movement in the classroom. These insights can be applied to estimate the students' abilities. Although the accuracy of ability estimation cannot be guaranteed due to the limited number of subjects and issues regarding questionnaire reliability, this experiment suggests that the

proposed system can be applied to the analysis of students' abilities from visual cues only.

V. Conclusion

In this paper, we propose a behavior monitoring system that utilizes off-the-shelf pose-estimation models. After evaluating the system using the motion-capture dataset, we demonstrated its application in estimating correlations and causal relationships between the cognitive and non-cognitive abilities of students, using real-world elementary school data. Other possible applications of this system include providing visual assistance for medical diagnosis, welfare care, and enhancing the functionality of smart security cameras. Future work in this area will focus on improving the tracking accuracy by enabling the system to switch between single-view and multiview modes depending on the environment as well as enhancing pose estimation accuracy through additional learning data that accounts for half-body occlusions.

Ethical Statement

The Yokohama National University Research Ethics Committee approved this study.

Acknowledgment

This study was supported by Yokohama City's Non-Cognitive Abilities (Social Emotional Competencies) research project.

References

[1] M. M. A. Parambil, L. Ali, F. Alnajjar, and M. Gochoo, "Smart classroom: A deep learning approach towards attention assessment through class behavior detection," in *2022 Advances in Science and Engineering Technology International Conferences (ASET)*, Feb. 2022, pp. 1–6.

[2] B. Ngoc Anh, N. Tung Son, P. Truong Lam, L. Phuong Chi, N. Huu Tuan, N. Cong Dat, N. Huu Trung, M. Umar Aftab, and T. Van Dinh, "A Computer-Vision Based Application for Student Behavior Monitoring in Classroom," *Applied Sciences*, vol. 9, no. 22, p. 4729, Nov. 2019.

[3] Z. Shou, M. Yan, H. Wen, J. Liu, J. Mo, and H. Zhang, "Research on students' action behavior recognition method based on classroom time-series images," *Applied Sciences*, vol. 13, no. 18, p. 10426, Sept. 2023.

[4] Z. Yu, M. Xie, J. Gao, T. Liu, and Y. Fu, "From Raw Video to Pedagogical Insights: A Unified Framework for Student Behavior Analysis," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 21, pp. 23 241–23 249, Mar. 2024.

[5] R. Steinmayr, M. Ziegler, and B. Träuble, "Do intelligence and sustained attention interact in predicting academic achievement?" *Learning and Individual Differences*, vol. 20, no. 1, pp. 14–18, Feb. 2010.

[6] M. Raca, "Camera-based estimation of student's attention in class," Ph.D. dissertation, EPFL, Lausanne, Switzerland, 2015.

[7] J. Zaletelj and A. Košir, "Predicting students' attention in the classroom from Kinect facial and body features," *EURASIP Journal on Image and Video Processing*, vol. 2017, no. 1, p. 80, Dec. 2017.

[8] M. Raca and P. Dillenbourg, "System for assessing classroom attention," in *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, Leuven Belgium, Apr. 2013, pp. 265–269.

[9] F. Yang, "A Spatio-Temporal Attention-Based Method for Detecting Student Classroom Behaviors," Oct. 2023, arXiv:2310.02523.

[10] T. Jiang, P. Lu, L. Zhang, N. Ma, R. Han, C. Lyu, Y. Li, and K. Chen, "RTMPose: Real-Time Multi-Person Pose Estimation based on MMPose," July 2023, arXiv:2303.07399.

[11] Z. Yang, A. Zeng, C. Yuan, and Y. Li, "Effective whole-body pose estimation with two-stages distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, Oct. 2023, pp. 4212–4222.

[12] P. Lu, T. Jiang, Y. Li, X. Li, K. Chen, and W. Yang, "RTMO: Towards High-Performance One-Stage Real-Time Multi-Person Pose Estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 1491–1500.

[13] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172–186, Jan. 2021.

[14] W. Zhu, X. Ma, Z. Liu, L. Liu, W. Wu, and Y. Wang, "MotionBERT: A unified perspective on learning human motion representations." in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2023, pp. 15 039–15 053.

[15] W. Shan, Z. Liu, X. Zhang, S. Wang, S. Ma, and W. Gao, "P-STMO: Pre-trained Spatial Temporal Many-to-One Model for 3D Human Pose Estimation," in *Computer Vision – ECCV 2022*, vol. 13665, Oct. 2022, pp. 461–478.

[16] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *2016 IEEE International Conference on Image Processing (ICIP)*, Sept. 2016, pp. 3464–3468.

[17] Y. Zhang, P. Sun, Y. Jiang, D. Yu, F. Weng, Z. Yuan, P. Luo, W. Liu, and X. Wang, "ByteTrack: Multi-object tracking by associating every detection box," in *Computer Vision – ECCV 2022*, ser. Lecture Notes in Computer Science, vol. 13682, 2022, pp. 1–21.

[18] J. B. Carroll, *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*. Cambridge University Press, 1993.

[19] T. Kautz, J. J. Heckman, R. Diris, B. ter Weel, and L. Borghans, "Fostering and measuring skills: Improving cognitive and non-cognitive skills to promote lifetime success," Institute for the Study of Labor (IZA), Bonn, IZA Discussion Papers 8696, 2014.

[20] F. Cunha and J. J. Heckman, "Formulating, identifying and estimating the technology of cognitive and noncognitive skill formation," *Journal of Human Resources*, vol. 43, no. 4, p. 738, Oct. 2008.

[21] J. J. Heckman, J. Stixrud, and S. Urzua, "The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior," National Bureau of Economic Research, Working Paper 12006, Feb. 2006.

[22] H. Joo, H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh, "Panoptic Studio: A Massively Multiview System for Social Motion Capture," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, Dec. 2015, pp. 3334–3342.

[23] S. Shimizu, T. Inazumi, Y. Sogawa, A. Hyvärinen, Y. Kawahara, T. Washio, P. O. Hoyer, and K. Bollen, "DirectLiNGAM: A Direct Method for Learning a Linear Non-Gaussian Structural Equation Model." *J. Mach. Learn. Res.*, vol. 12, pp. 1225–1248, 2011.