Student Engagement Assessment in Classrooms Using a Novel 3D Eye-gaze Estimation and Evaluation Algorithm

Matthew Korban[1], Jonathan McGee[1], Peter Youngs[2], and Scott T. Acton[1]

[1]C.L. Brown Dept. of Electrical and Computer Engineering, University of Virginia, Charlottesville, Virginia 22903
[2]Department of Curriculum, Instruction, and Special Education, University of Virginia, Charlottesville, Virginia 22903
Emails: acw6ze@virginia.edu, tkg5kq@virginia.edu, pay2n@virginia.edu, acton@virginia.edu

2024 58th Asilomar Conference on Signals, Systems, and Computers | 979-8-3503-5405-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/IEEECONF60004.2024.10942809

***Abstract*—Student attentiveness within the classroom can be assessed by observing student attention toward the teacher or whiteboard, which may be inferred through eye-gaze direction. This paper introduces a novel technique for evaluating student attentiveness by analyzing the direction of their eye gaze derived from their 3D skeletal pose in a reconstructed 3D environment. As for the contributions, the paper suggests a novel 3D head pose estimation algorithm that, unlike other works, does not need frontal face information. As a result, the method is highly effective in uncontrolled environments such as classrooms, where frontal face data is often unavailable. Moreover, a new algorithm was developed to evaluate student attentiveness based on 3D eye gaze information interpreted from the 3D head pose. The proposed method has been validated using a set of instructional videos collected at the University of Virginia.**

I. Introduction

In today's era, artificial intelligence (AI) is seamlessly integrated into various aspects of human life, including education, where it plays a crucial role in enhancing quality and accessibility [1]. One significant application of AI in education is the automation of teacher performance assessments [2]. Traditionally, these assessments have been conducted manually, a process that is labor-intensive and prone to errors [3]. In recent years, considerable research efforts have focused on developing automated, AI-driven systems for teacher evaluation [4], [5], [6], [7]. However, much of this research, such as instructional activity recognition, has emphasized teacher behavior while neglecting the critical role of students in the evaluation process [8].

A student-centered approach to assessing teacher performance can provide deeper insights into classroom dynamics [8]. One intuitive method is through analyzing students' eye gaze and head direction, as engaged students are more likely to focus their attention on their teacher [9], [10]. To facilitate this gaze analysis, estimating 3D head poses is a viable strategy [11]. However, traditional methods of 3D head pose estimation rely on frontal facial features [12], which are often ineffective in classroom settings.

This ineffectiveness arises for several reasons. First, classroom videos are typically captured with the camera focusing on the teacher, leaving students' faces partially or completely out of frame. Second, in densely packed classrooms, students are often occluded by peers or objects. Third, even for students whose faces are visible, the wide-angle focus of classroom cameras frequently results in low-resolution or blurry facial features.

To address these challenges, this paper proposes a novel method for estimating 3D head poses in classroom environments using 3D skeletal poses instead of facial features. Unlike faces, the body provides a more reliable and informative representation under varying camera angles, occlusions, and low-resolution conditions, making the proposed algorithm more effective in classroom settings.

Furthermore, we introduce an innovative algorithm to evaluate student attentiveness based on their estimated 3D head poses. Our pipeline is designed to operate on single images, making it highly generalizable for real-world applications where videos are often captured using standard cameras.

II. Related Work

Research in 3D head pose estimation has progressed significantly in recent years. Early methods focused on multi-stage techniques for real-time estimation, predicting Euler angles with data-driven regressors and optimizing face detection for efficiency [13]. Subsequent advancements utilized attention mechanisms to enhance pose estimation from single RGB images, addressing issues such as complex pose variations [14]. Later approaches incorporated heteroscedastic neural networks to improve robustness and provide uncertainty estimates, enabling more reliable predictions [15]. Collaborative multi-task learning frameworks combining RGB and sparse depth further reduced errors on benchmark datasets. However, all of the aforementioned methods require frontal face features for proper 3D head pose estimation [16].

Advancements in classroom activity monitoring have evolved from basic methods like eye status and head orientation analysis [10], which struggled with occlusions and nonstandard angles, to more integrated approaches. [9] combined emotion, gaze, and head movement data, addressing some limitations but still hampered by resolution issues. [17], improved robustness by integrating action units with head pose and gaze, while [18] leveraged deep learning to handle the diversity in input effectively. However, the problem of finding an effective student-centered method to work effectively in classroom settings with persistent challenges such as occlusion, various camera angles, and low resolution remained unsolved.

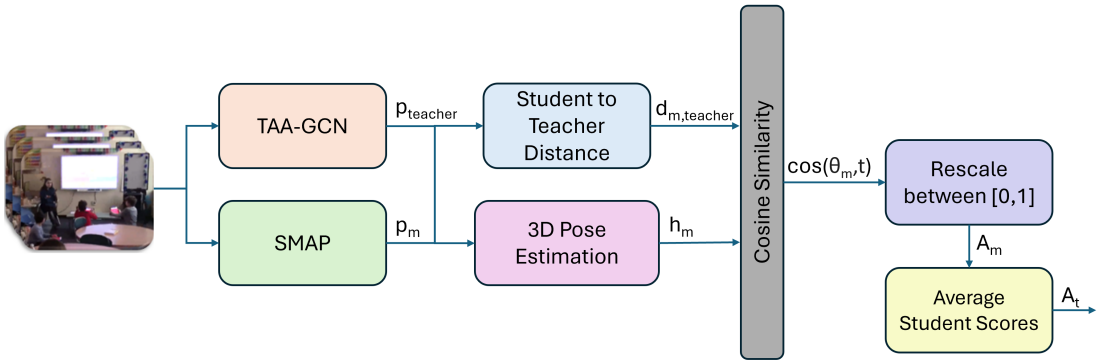

Fig. 1: Pipeline for attentiveness calculation for a video clip

III. Methods

A. Overview

Fig. 1 shows the overview of the proposed pipeline. The teacher locations are provided through annotated bounding boxes during the training phase. The video clips during the testing phase are first fed into the TAA-GCN and SMAP blocks. The TAA-GCN block [19] determines the teacher location, $p_{teacher}$, utilizing semantic information, including the skeletal structure and clothing style, to predict the age of the subjects. Then, the SMAP [20] block identifies the 3D skeletal structure of each student, p_m. Each element of the p_m list is then utilized to calculate the expected direction of attentiveness, $d_{m,teacher}$ in respect to $p_{teacher}$, and the 3D head pose estimation of the student, h_m.

Next, based on the estimated head pose vectors, the attentiveness scores are then calculated. The difference of the two vectors is compared using cosine similarity and then rescaled to a scale from 0 to 1. The measure of attentiveness will be measured from 0, no attention being given to the teacher, to 1, fully attentive to the teacher. These scores are then averaged over all the students in the video to find the average attentiveness for the clip.

In the following sections: Section III-B explains the 3D head pose block methodology and Section III-C details the comparison of the expected student to teacher vector and the estimated student engagement vector to find the total attentiveness for the students.

B. 3D Head Pose Estimation

We use SMAP [20] to extract 3D skeletal poses from classroom videos. SMAP provides a unique approach that can simultaneously reconstruct 3D head poses and the relative 3D positions of individuals in the scene. This distinctive feature is crucial for our study, as it enables us to capture both the 3D head poses and the relative spatial locations of students to the teacher, which are essential for accurately determining students' attentiveness scores.

The 3D head poses are then estimated based on the extracted 3D skeletal pose. We consider the normal vector of the 3D plane intersecting the head, right and left shoulder and an estimate for the 3D pose as shown in Fig. 2.

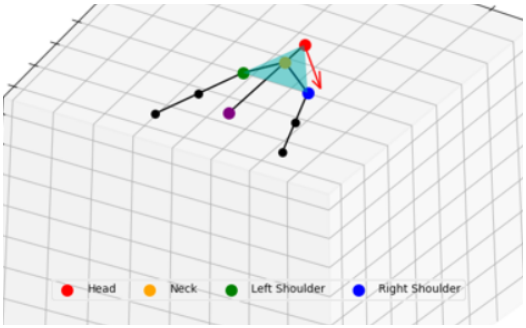

Fig. 2: 3D headpose estimation based on 3D skeletal pose.

C. Student Attentiveness Evaluation Algorithm

To evaluate student attentiveness, the alignment of the students' head directions concerning the teacher is computed. The attentiveness score is obtained for each frame by averaging it across M students. Finally, for each video clip, the attentiveness score is calculated by averaging across T frames. The alignment of the students' head direction is computed using cosine similarity, $\cos(\theta_{m,t})$ and accordingly, the attentiveness score $A_{m,t}$ for student m and frame t as follows:

$$\cos(\theta_{m,t}) = \frac{\mathbf{h}_{m,t} \cdot \mathbf{d}_{m,teacher}}{\|\mathbf{h}_{m,t}\| \|\mathbf{d}_{m,teacher}\|} \tag{1}$$

$$A_{m,t} = \frac{1 + \cos(\theta_{m,t})}{2} \tag{2}$$

where $\mathbf{h}_{m,t} \in \mathbb{R}^3$ are 3D head pose vectors for student m at frame t and the relative direction vector based on the position of teacher and students: $\mathbf{d}_{m,teacher} = \mathbf{p}_{teacher} - \mathbf{p}_m$. The above procedure is summarized as Algorithm 1.

IV. Experimental Results

An elementary school dataset was collected to analyze instructional activities. We used 10 hours of instructional activity videos annotated by a team of nine professional annotators at the University of Virginia. Each video is around 30 minutes long.

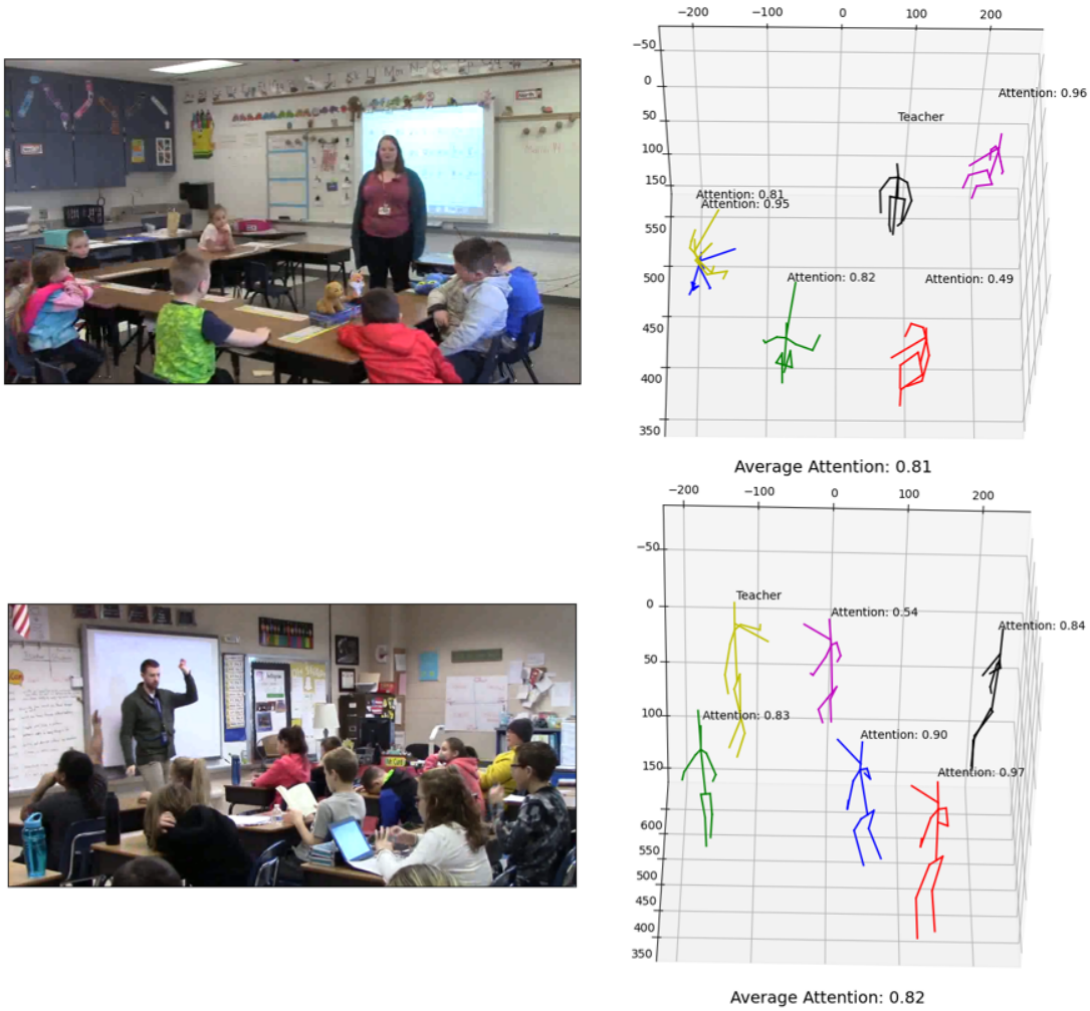

Fig. 3: Visualization of student attentiveness scores for two examples of classroom video frames.

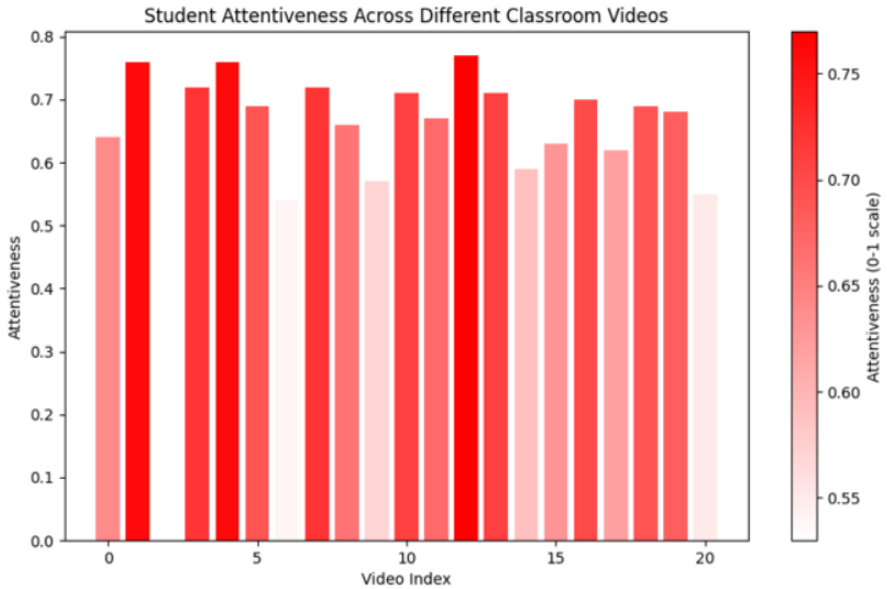

Fig. 4: Attentiveness scores for different classroom video samples.

Algorithm 1 Overall Attentiveness Score Calculation based on 3D Head Pose for Multiple Students

Require: M ▷ Number of students
Require: T ▷ Number of frames
Require: $\mathbf{h}_{m,t} \in \mathbb{R}^3$ ▷ 3D head pose vector for student m at frame t
Require: $\mathbf{p}_{teacher} \in \mathbb{R}^3$ ▷ Position of the teacher
Require: $\mathbf{p}_m \in \mathbb{R}^3$ ▷ Position of student m
$A_{overall} \leftarrow 0$ ▷ Initialize the overall attentiveness score
for $t = 1$ to T **do** ▷ Loop over all frames
  $A_t \leftarrow 0$ ▷ Initialize the attentiveness score for frame t
  for $m = 1$ to M **do** ▷ Loop over all students at frame t
    Compute relative direction vector $\mathbf{d}_{m,teacher} = \mathbf{p}_{teacher} - \mathbf{p}_m$
    Compute cosine similarity: $\cos(\theta_{m,t}) = \frac{\mathbf{h}_{m,t} \cdot \mathbf{d}_{m,teacher}}{\|\mathbf{h}_{m,t}\| \|\mathbf{d}_{m,teacher}\|}$
    Attentiveness score for student m at frame t: $A_{m,t} = \frac{1+\cos(\theta_{m,t})}{2}$
    Add student's attentiveness score to frame score: $A_t \leftarrow A_t + A_{m,t}$
  end for
  Compute average attentiveness for frame t: $A_t \leftarrow \frac{A_t}{M}$
  Add frame's attentiveness score to the overall score: $A_{overall} \leftarrow A_{overall} + A_t$
end for
Compute final overall attentiveness score across all frames: $A_{overall} \leftarrow \frac{A_{overall}}{T}$
Ensure: $A_{overall}$ ▷ Return the final overall attentiveness score

Fig. 3 visualizes the student attentiveness scores for two classroom video examples. In the top example, there is a small group activity in which, ideally, students pay full attention to the teacher by aligning their head direction to the teacher's position. However, the results show that the attentiveness score varies for different individuals. The same goes for the whole class activity Fig. 3 - bottom.

Fig. 4 shows the average attentiveness score for 21 instructional video clips of our instructional videos. The results show varying attentiveness scores for different videos but ranged generally between 0.65 and 0.75.

The video clips of our instructional videos range from no attention on the teacher to fully attentive students. Index 2 in Fig. 4, displays an attentiveness of nearly 0. This video clip was of students engaging in group activities and individual notebook work where their attention was not supposed to be on the teacher at that moment. A score of 0 attentiveness gives a strong indicator that attentiveness is not teacher-centered and instead, in this case, is focused on group and individual work.

Another example, index 20 shows an attentiveness score of about 0.5 indicating that only half the subjects in the video are attentive. This video clip shows a teacher sitting at a desk with one student next to her sitting facing the same direction and another student across from her facing her. This score reflects the activity in the clip indicating that the activity is less teacher centered.

The final example of index 12, shows the highest level of attentiveness of the example videos. With a score of about 0.75 index 12 indicates that the students are very attentive to the teacher and engaged in teacher-centered learning. The score of this clip is higher because the class is engaging in teacher-centered instruction where the teacher is lecturing on a math lesson and the students are engaging by watching, listening, and occasionally going to the board to write. The latter is most likely the reason that the score is not closer to 1 because the students in the clip are not always placing their attention on the teacher when they are writing on the board. The score of 0.75 gives a strong descriptor of the students attentiveness showing that the activity they are engaging in is more teacher-centered.

V. Conclusion

This study presents a novel framework for analyzing classroom interactions by integrating a new 3D head pose estimation with spatial context to assess student attentiveness. By combining precise 3D positional data and head orientation metrics, our approach offers a comprehensive perspective on student-teacher dynamics. This study establishes a foundation for future research exploring scalable, automated methods for evaluating learning environments. The effectiveness of the proposed method has been demonstrated using a collection of instructional videos recorded at the University of Virginia.

References

[1] S. Zheng and M. Han, "The impact of ai enablement on students' personalized learning and countermeasures—a dialectical approach to thinking," *Journal of Infrastructure, Policy and Development*, vol. 8, no. 14, p. 10274, 2024.

[2] L. Chen, P. Chen, and Z. Lin, "Artificial intelligence in education: A review," *Ieee Access*, vol. 8, pp. 75 264–75 278, 2020.

[3] H. Li, Z. Wang, J. Tang, W. Ding, and Z. Liu, "Siamese neural networks for class activity detection," in *Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part II 21*. Springer, 2020, pp. 162–167.

[4] H. Li, Y. Kang, W. Ding, S. Yang, S. Yang, G. Y. Huang, and Z. Liu, "Multimodal learning for classroom activity detection," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 9234–9238.

[5] P. J. Donnelly, N. Blanchard, B. Samei, A. M. Olney, X. Sun, B. Ward, S. Kelly, M. Nystrand, and S. K. D'Mello, "Multi-sensor modeling of teacher instructional segments in live classrooms," in *Proceedings of the 18th ACM international conference on multimodal interaction*, 2016, pp. 177–184.

[6] H. Ren and G. Xu, "Human action recognition in smart classroom," in *Proceedings of fifth IEEE international conference on automatic face gesture recognition*. IEEE, 2002, pp. 417–422.

[7] F.-C. Lin, H.-H. Ngo, C.-R. Dow, K.-H. Lam, and H. L. Le, "Student behavior recognition system for the classroom environment based on skeleton pose estimation and person detection," *Sensors*, vol. 21, no. 16, p. 5314, 2021.

[8] H. Niemi, R. D. Pea, and Y. Lu, *AI in learning: designing the future*. Springer Nature, 2023.

[9] A. Huang, M. J. C. Samonte, X. Huang, and Z. Zhang, "Application research of emotion analysis, eye tracking, and head movement monitoring based on facial recognition algorithms in esl student engagement assessment," in *2024 5th International Conference on Information Science, Parallel and Distributed Systems (ISPDS)*. IEEE, 2024, pp. 409–412.

[10] Y. Harshalatha, M. Varun, S. Thrinesh, S. Shilpashree, and S. D. Biradar, "Student attention monitoring: An automated approach," in *2024 International Conference on Smart Systems for applications in Electrical Sciences (ICSSES)*. IEEE, 2024, pp. 1–5.

[11] H. Yu, A. Gupta, W. Lee, I. Arroyo, M. Betke, D. Allesio, T. Murray, J. Magee, and B. P. Woolf, "Measuring and integrating facial expressions and head pose as indicators of engagement and affect in tutoring systems," in *International Conference on Human-Computer Interaction*. Springer, 2021, pp. 219–233.

[12] K. Khan, R. U. Khan, R. Leonardi, P. Migliorati, and S. Benini, "Head pose estimation: A survey of the last ten years," *Signal Processing: Image Communication*, vol. 99, p. 116479, 2021.

[13] X. Zhang, D. Zhang, J. Ge, K. Hu, L. Yang, and P. Chen, "Multi-stage real-time human head pose estimation," in *2019 6th International Conference on Systems and Informatics (ICSAI)*. IEEE, 2019, pp. 563–567.

[14] A. Behera, Z. Wharton, P. Hewage, and S. Kumar, "Rotation axis focused attention network (rafa-net) for estimating head pose," in *Proceedings of the Asian Conference on Computer Vision*, 2020.

[15] G. Cantarini, F. F. Tomenotti, N. Noceti, and F. Odone, "Hhp-net: A light heteroscedastic neural network for head pose estimation with uncertainty," in *Proceedings of the IEEE/CVF Winter Conference on applications of computer vision*, 2022, pp. 3521–3530.

[16] J. Zhang and H. Yu, "Collaborative 3d face alignment and head pose estimation with frontal face constraint based on rgb and sparse depth," *Electronics Letters*, vol. 58, no. 21, pp. 801–803, 2022.

[17] A. Anand, A. Mittal, L. Dhawan, J. Krishnamurthy, M. Ramesh, N. Lal, A. Verma, P. Bhuyan, R. R. Shah, R. Zimmermann *et al.*, "Exceda: Unlocking attention paradigms in extended duration e-classrooms by leveraging attention-mechanism models," in *2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE, 2024, pp. 301–307.

[18] V. Warankar, N. Jain, B. Patil, M. Faizaan, B. Jagdale, and S. Sugave, "Analysis of attention span of students using deep learning," in *2024 MIT Art, Design and Technology School of Computing International Conference (MITADTSoCiCon)*. IEEE, 2024, pp. 1–7.

[19] M. Korban, P. Youngs, and S. T. Acton, "Taa-gcn: A temporally aware adaptive graph convolutional network for age estimation," *Pattern Recognition*, vol. 134, p. 109066, 2023.

[20] J. Zhen, Q. Fang, J. Sun, W. Liu, W. Jiang, H. Bao, and X. Zhou, "Smap: Single-shot multi-person absolute 3d pose estimation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*. Springer, 2020, pp. 550–566.