Pose+Context: A Model for Recognizing Non-verbal Teaching Behavior of Normal College Student

Yonghe Zhang, Bing Li(✉), and Xiaoming Cao

Normal College, Shenzhen University, Shenzhen 518060, China
libingice@szu.edu.cn

Abstract. Normal college students practice their teaching skills to prepare for their teacher career. However, due to the complexity of teaching skills and the overburdening of instructors, their needs for instructional feedback are often not met. To server automatic feedbacks to normal college students, this paper proposes a deep learning model, called Pose+Contex, to recognize three types of non-verbal teaching behaviors (*NVTB*). The model includes three parts: (1) context detection, (2) pose estimation, and (3) behavior recognition. Our model is featured by the context detection component, and experiments show that it performs better than a similar model without this component.

Keywords: deep learning · computer vision · pose estimation · classroom behavior analysis

1 Introduction

Teacher training system in China has produced a large number of teachers to K12-education, and the training of teacher's teaching skills is highly valued by the education department. Before teaching career, normal college student (also called normal students) practice their teaching skills mainly in campus. Among various training activities for teaching skills, post-class reflection is a critical part for the self-improvement of normal students, which relies on feedbacks from the instructors. Due to the complexity of teaching skills and the overburdening of instructors, the needs for instructional feedback are often not met, and instruction in non-verbal teaching behavior is particularly lacking.

Compared with verbal expressions, teachers' nonverbal behavior is a kind of "silent language", including the position in classroom, gesture, gaze and facial expressions, etc. [1]. Many researchers believe that teachers' nonverbal behaviors in the classroom have a non-negligible impact on student learning [1–6], such as the improvement of teacher-student relations, the enhancement of classroom teaching effects and student learning efficiency [1]. However, normal students' nonverbal behavior problems are often exposed during the teaching practice, due to the lack of practical experience, relevant training and self-confidence of normal students, as well as overburdening of instructors.

R. Hou et al. (Eds.): BDTA 2021/2022, LNICST 480, pp. 257–264, 2023.
https://doi.org/10.1007/978-3-031-33614-0_17

Although the related research on teacher reflection by video analysis has been studied for a few decades [7], the analysis of non-verbal behavior still needs to be further expanded in the field of classroom teaching analysis. With the new development of big data and artificial intelligence, Computer Vision (CV) has reached an unprecedented state. In the field of CV, human pose estimation is one of the major breakthroughs have been made. It is promising to develop video analysis systems based on this technology to diagnosis non-verbal behavior problems for normal students just in time. Therefore, this article explores the extension and application and of pose estimation technology in helping normal college students' training.

2 Related Work

2.1 Category of Teachers' Non-verbal Behaviors

The classification of NVTB is the prerequisite for detecting and analyzing normal college students' teaching behavior pattern. Prof. Shen classified NVTB into seven categories based on the level of teachers' support for students, including enthusiasm-support, acceptance-help, elaboration-guidance, intermediate, avoid-falter, neglect, and disagree [14]. Cui and Wang divide NVTB into two categories, including body language and scene according to the relationship between the communicator and the context [15]. Tang classified NVTB into static behaviors, such as people's spatial position and body posture, and dynamic behaviors, such as sign language and head Sentimental language, gaze and facial expressions [16]. Liu's extend Tang's theory to add interpersonal communication as a new type of NVTB, such as interpersonal distance, body orientation, etc. [17].

2.2 Human Pose Estimation and OpenPose

In computer vision research, pose estimation refers to the process of recovering human joint points from a given image or a video, which is a very challenging research problem [19]. The key points of human skeleton are of great significance for describing human posture and predicting human behavior, such as behavior recognition, task tracking, and gait recognition [18]. Human pose estimation algorithms can be divided into two methods, based on traditional methods and methods based on deep learning. The traditional pose estimation algorithm is mainly based on the image structure model [20, 21], which is not robust caused by data factors such as shape, angle, and occlusion. The pose estimation algorithm based on deep learning has been developed rapidly in recent years [21]. So far, the methods of using deep learning for multi-person pose estimation are roughly divided into two types: top-down method and bottom-up method. Top-down (top-down), that is, the human body is detected first, and then the posture of a single person is estimated [22, 25–27]. Bottom-up (Bottom-up) is to first detect the human body joint points, and then connect the human body skeleton according to the detected joint points [23, 24, 29]. Generally speaking, the top-down pose estimation method has higher accuracy, but its processing speed is lower; the bottom-up method has a slightly lower accuracy than the former, but it runs faster than the former and can achieve real-time Detection [29].

OpenPose is currently a commonly used gesture recognition model. Using a bottom-up method, it can quickly and accurately obtain the skeleton key point coordinates of the characters in the image, which is suitable for single and multi-person recognition [28]. OpenPose can realize the recognition and estimation of the key points of the skeleton of multiple people's bodies, faces and hands. It has good robustness, so it can be used for application research related to gesture recognition [28–30].

To summarize the related works of this paper, we argue that the current research on non-verbal behavior analysis and its application in normal college students' training is still in the preliminary stage. There are particularly few studies aiming at recognizing normal college students' attention-related non-verbal behavior, like focusing on the audience, podium and the classroom screen. These behaviors may have influenced their teaching performance [14], and statistics analysis of this data would be inspiring for moderating allocation during normal college students' teaching practice.

3 The Pose+Context Model

The aim of this work is to recognize normal college students' attention-related non-verbal behavior, including:

- C_a: a normal student focus on the audience in classroom.
- C_p: a normal student focus on the podium (teacher's desk).
- C_s: a normal student focus on the classroom screen.

The proposed model is called the Pose+Context model as shown in Fig. 1. It has three components: (1) Context detection: locate the area of the podium through the object detection algorithm in each image. (2) Pose estimation: extract the skeleton key points of normal college students, and drop the unreliable key points according to confidence scores. (3) Behavior recognition: Based on the results of previous steps, posture recognition is performed to recognize normal college students' attention-related non-verbal behavior.

The context detection and pose estimation components provide input vectors for the multi-class classifier which predicts the type of behavior. The context detection components can be implemented by the architecture of object detection which trained to detect podiums and applied to provide context data. we use YOLOv5 [31] to implement this component in our experiments. As for pose estimation, a pre-trained Openpose model [28] is used to extract pose data in our experiments. The results of the first two components' 2D tensors are flatten and joint into a vector as the input of behavior recognition. For behavior recognition, a multi-class classifier is typically a deep learning model. In the following experiments, we build the model with a nine-layers perceptron with 32 to 128 nodes in each hidden layer. Batch normalization and ReLU activation are also set up after each hidden layer.

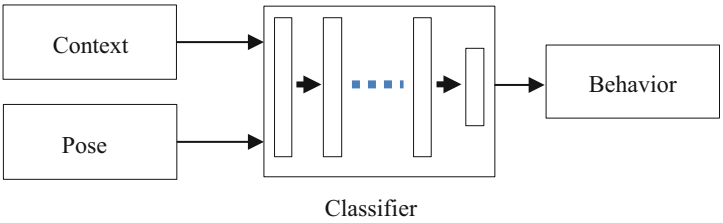

Fig. 1. The Pose+Context Model.

4 Experiment

4.1 Dataset Construction

This experiment was completed in micro-teaching classrooms in a normal college. In each micro-teaching classroom, there is a podium and a screen on wall in the frontal area, and audience desks and chairs in the rest area. Videos are captured by two cameras to produce views of teacher and the whole classroom (Table 1).

Table 1. Examples of normal college students' attention-related non-verbal behavior

			
C_a	1	0	0
C_p	0	1	0
C_s	0	0	1

In the experiment, videos of teaching practice were collected by a teaching-assisted system. A total of 23 normal students participated in this experiment, each of them completed 6 teaching practices. After data preprocessing, we constructed a labeled dataset of 1638 images. Labels of a sample include behavior type and a box marking the podium in it. Then we divided the dataset in cross-subject manner into training, validation and test subset by the ratio of 6:1:3 approximately.

4.2 Evaluation Metric

Our experiment compares the two cases: with context (Pose+Context) and without context (Pose only) as input for the behavior recognition model. Our hypothesis is that the Pose+Context model would enhance the recognition by the effect of context. We repeated the entire evaluation process (training, verification, and testing) for the two methods 40 times in order to avoid interference factors caused by data. For each time

of the process, we set 300 epochs for training and use the same train-test division of the dataset. The evaluation metric is micro-averaged F1 score shown as Eqs. (1–3).

$$Micro\_F1 = \frac{2 \times Micro\_P \times Micro\_R}{Micro\_P + Micro\_R} \tag{1}$$

$$Micro\_P = \frac{\sum_i TP_i}{\sum_i TP_i + \sum_i FP_i} \tag{2}$$

$$Micro\_R = \frac{\sum_i TP_i}{\sum_i TP_i + \sum_i FN_i} \tag{3}$$

where TP_i is the total number of true positive predictions of *i-th* class, FP_i is the total number of false positive predictions of *i-th* class, FN_i is the total number of false negative predictions of *i-th* class, and $i \in (C_a, C_p, C_s)$.

4.3 Experiment Results

The experimental processing was performed on a high-performance computer with 32GB memory, Intel i7 2.9GHz CPU, and NVIDIA GeForce GTX 3090 graphics. The results are shown in the Table 2.

Table 2. Micro-averaged F-1 score of the Pose+Context model and the Pose only model.

	N	mean	std	max
Pose+Context	40	**0.84**	0.05	**0.90**
Pose only	40	0.81	0.05	0.87

We performed independent samples T-test for the two groups of results, and the outcome $p = 0.008 < 0.05$ indicates that the two sets of means have very significant differences. The Pose+Context model has a higher mean of F-1 scores than the Pose only model. One of the recognition instances of each class by the Pose+Context model is shown in Table 3.

Table 3. Metrics of the Pose+Context model

	precision	recall	f1-score	support
C_a	0.81	0.69	0.75	101
C_p	0.93	0.93	0.93	168
C_s	0.91	0.98	0.94	220

The recall value of C_a is relatively low. We look into the error samples to explore the difficulties in the recognition process of our model. Two typical false cases shown

in Fig. 2 where context and pose marks are shown. From our observation, the model may have overused the distance between teacher and podium or the direction between them for inference. Hence, the context input component of the model has side effects in certain degree leading to the low recall score for C_a.

Case 1: Truth=C_a, Predition=C_p Case 2: Truth=C_a, Predition= C_s

Fig. 2. Error recognition instances of the Pose+Context model

5 Application

To demonstrate applications of the proposed model, we related the NVTB predations of the model and instructors' scores for teaching practices. Each teaching practice is marked with a score which is ranged from 1 to 5, and higher scores means better teaching performance. Ratios of predicted labels are also counted for each teaching practice. Figure 3 shows the ratio of C_a (focus of audience) increases and the one of C_p (focus of podium) decreases when scores go higher. This application shows the NVTB prediction by our model can be play as indices for normal students' teaching performance, and it is promising to server as diagnosis feedback for them to improve their teaching skill.

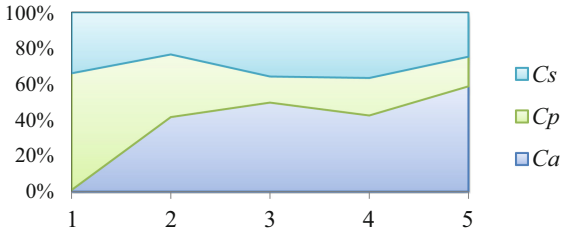

Fig. 3. An application to show NVTB labels distributions among difference scores of teaching.

6 Conclusions and Future Work

The goal of the non-verbal behavior analysis of normal college students' teaching practice is to provide feedbacks in time for teaching skill improvement. The Pose+Context model proposed in this paper has been proven to be effective to recognize three types of attention-related non-verbal behaviors with overall F1-score 0.84.

The experimental process and results of this research are constrained by a variety of factors, and further research has the following ideas: First, identify more context factors in the teaching process, and perform correlation analysis with teacher students' postures to improve the accuracy of teacher students' non-verbal behavior recognition; Second, combine speech behavior analysis methods to achieve multi-modal behavior analysis, and comprehensively describe the quantitative characteristics of the teaching process; third, improve the hardware environment, collect high-resolution eye image data, and promote eye tracking technologies to improve the outcomes.

Acknowledgements. This research was supported by the Ministry of Education in China (project: Research on the Identification Mechanism of Learning Participation Based on Multimodal Fusion in a Smart Learning Environment, 20YJA880001) and the Collaborative Innovation Research Institute of Sports Psychology Education of Shenzhen University Normal College (project: Recognition of Learning Engagement Based on Multi-view Stereo Vision).

References

1. Jinling, T.: A review of the research on teachers' classroom nonverbal behaviors. Zhejiang Educ. Sci. **1**, 10–12 (2010)
2. Cooper, P.J.: Speech communication for the classroom teacher. Gorsuch Scarisbbrick (1988)
3. Huang, L.: The impact of non-verbal behavior psychological semantics on classroom teaching. China Elect. Power Educ. **35**, 81–82 (2011)
4. Cheng, Y.: The significance and methods of using non-verbal communication strategies in college music classrooms. Northwestern Med. Educ. **18**(03), 566–569 (2010)
5. Zhou, M.: The influence of several typical nonverbal behaviors of teachers on classroom atmosphere in multimedia teaching. J. Hunan Med. Univ. (Soc. Sci. Ed.) **12**(04), 184–185 (2010)
6. Zhou, P.: A Brief Discussion on Teachers' Non-verbal Behavior Research. Nationalities Publishing House, Beijing (2006)
7. Tian, L., Zhang, Z., Chen, Y.: Research on the effect of video promoting micro-teaching reflection of normal students. Mod. Educ. Technol. **25**(10), 54–60 (2015)
8. Flanders, N.A.: Analyzing Teaching Behavior, p. 34. Addision-Wesley Publishing Company, New York (1970)
9. Yang, P., Liu, J., Luo, P.: Discussion on the teaching behavior classification system VICS. Teacher **6**, 3–4 (2009)
10. Fu, D., Zhang, H., Liu, Q.: Educational Information Processing (2nd ed). Beijing Normal University Press, Beijing, vol. 92, pp. 98–100 (2011)
11. Mu, S., Zuo, P.: Research on the analysis method of classroom teaching behavior under the information teaching environment. Audio-Vis. Educ. Res. **36**(09), 62–69 (2015)
12. Cheng, Y., Liu, Q., Wang, Y., et al.: Research on the construction and application of the cloud model of classroom teaching behavior analysis. J. Dist. Educ. **2**, 36–42 (2017)

13. Jin, J., Gu, X.: Analysis and research on classroom teaching behavior in information technology environment. China Audio-Vis. Educ. (9), 82–86 (2010)
14. Shen, L.: Non-verbal behavior in the classroom. Foreign Element. Second. Educ. **6**, 33–34 (1983)
15. Cui, Y., Wang, J.: Interpretation of non-verbal behaviors in classroom teaching from the perspective of cognitive pragmatics. Educ. Teach. Forum **42**, 98–100 (2014)
16. Tan, J.: A review of the research on nonverbal behavior of teachers in classroom. Zhejiang Educ. Sci. **1**, 10–12 (2010)
17. Liu, Y.: A Study on Nonverbal Behavior of Junior Middle School Music Novice Teachers. Master's Degree Thesis of Nanning Normal University (2019)
18. Iqbal, U., Milan, A., Gall, J.: PoseTrack: joint multi-person pose estimation and tracking. In: Computer Vision and Pattern Recognition, pp. 4654–4663, IEEE (2017)
19. Johnson, S., Everingham, M.: Learning effective human pose estimation from inaccurate annotation. CVPR, IEEE (2011)
20. Fischler, M.A., Elschlager, R.A.: The representation and matching of pictorial structures. IEEE Trans. Comput. **1**, 67–92 (1973)
21. Deng, Y., Luo, J., Jin, F.: Overview of human pose estimation methods based on deep learning. Comput. Eng. Appl. **55**(19), 22–42 (2019)
22. Huang, S., Gong, M., Tao, D.: A coarse-fine network for keypoint localization. IEEE International Conference on Computer Vision, pp3047–3056, IEEE (2017)
23. Cao, Z., Simon, T., Wei, S.E., et al.: Realtime multi-person 2D pose estimation using part affinity fields. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society (2017)
24. He, K., Zhang, X., Ren, S., et al.: Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. **37**(9), 1904–1916 (2015)
25. Papandreou, G., Zhu, T., Kanazawa, N., et al.: Towards accurate multi-person pose estimation in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4903–4911, IEEE (2017)
26. Chen, Y., Wang, Z., Peng, Y., et al.: Cascaded pyramid network for multi-person pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7103–7112, IEEE (2018)
27. Fang, H.S., Xie, S., Tai, Y.W., et al.: Rmpe:RegionaI multi-person pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2334–2343, IEEE (2017)
28. Yang, M., Li, J., Guo, R., Tang, X.: Realization and research of human sleeping posture recognition based on OpenPose. Phys. Exp. **39**(8), 4549 (2019)
29. Gong, W.: Design and implementation of student learning behavior recognition system based on bone key point detection. Jilin University (2019)
30. Zheng, Y.: An evaluation method of teacher's teaching behavior based on gesture recognition. Softw. Eng. **4**, 6–9 (2021)
31. YOLOv5 Homepage: https://github.com/ultralytics/yolov5. Last accessed 29 Aug. 2021