Student postures and gestures recognition system for adaptive learning improvement

Quang Trung, Nguyen
Vietnam Youth Academy
Hanoi, Vietnam
trungnq@vya.edu.vn

Hoang Tieu Binh
Hanoi National University of Education
Hanoi, Vietnam
htbinh@hnue.edu.vn

The Duy, Bui
University of Engineering and Technology
Hanoi, Vietnam
theduybui@gmail.com

Phuong Dung, N.T.
Thuy Loi University
Hanoi, Vietnam
dungntp@tlu.edu.vn

***Abstract*— Adaptive learning is an educational method that uses computers to coordinate interaction with student and provide customized learning resources and activities to meet each student's unique needs that are a trend. In order to adjust the content and teaching methods appropriate for each learner, it is necessary to track the attitude of students with the content and methods being implemented. One of the ways of detecting student's attitudes is through the gestures and posture of student in the classroom.**

In this study, we propose to develop a model to monitor and identify students' postures and gestures during class time to help assess the level of student participation in the content and teaching methods of teachers. From there, it is possible to give advice to trainers and training managers in adjusting the subject content, teaching methods suitable for each specific object.

Keywords*— Posture recognition, Gesture recognition, emotion recognition, adaptive learning, transfer learning.

I. INTRODUCTION

In recent years, detecting student emotion and actions in the classroom is a key topic for researchers who focus on intelligent training systems. Together with the development of advanced techniques and the power of computing ability, many solutions for these problems are presented and gained some positive results. Knowing exactly whether a student is interested in the lesson or not is very important for designing an adaptive learning system and intelligent tutoring systems. Besides this, detecting student engagement supports teachers or tutoring systems in tracking learner attention. The success of a traditional course or online learning course depends on the outcomes of students, and that results are related to the engagement of students. Some authors show that engagement detection makes opportunities for improving learning and teaching [1] [2] and even for issuing targeted interventions of the monitors or educational manager [3].

In the firm of engagement detection, Dewan [4]divided engagement detection into their categories, such as Automatic, Semi-Automatic and Manual. Toward computer vision-based point of view, he mentioned that Facial Expression, Gesture, and posture, Eye movement are three facets in automatic detection.

Facial expression recognition receives many concerns from researchers. Many researches concentrate on student's face to guess the emotions of learners. Nezam [5] used a convolution neural network to classify emotions of students, based on the pre-trained result from the FER-2013 dataset [6]. Whitehill [3] divided student engagement into four levels from 1 to 4 and used the SVM classifier to detect. Bosch [7] focused on mind wandering a type of cognitive disengagement. He concentrated on extracting low-level and high-level features from video clips for the face recognition process.

Eye movement is interested in recent years, Raina [8] proposed an eye-tracking-based model which used eye-tracker equipment to compare engagement between two student groups to reduce content skipping when they study on the online system. Krithika [9] built a system for student emotion recognition that can recognize and track emotions of the student in an e-learning system. This system provides a real-time feedback mechanism to enhance the e-learning aids for better content delivery. The system can track head movement and eyes for estimating student engagement level.

Gesture and posture recognition system is also interested by many authors. In [10] Mitra divided the gestures arising from different body parts into three classes are hand and arm gestures, head and face gestures, and body gestures. Wan [11] used the probabilistic distribution of the arm trajectory to classify 30 different hand gestures. These results show that the recognition system can obtain high accuracy of "Rain" and "Round" gestures. However, nine markers should be attached to indicate the feature points of a student's body. This assumption is difficult to implement in a classroom. Fang [12] recognize six popular gesture of student in classroom which are "Raising the left hand", "Raising the right hand", "Raising two hands", "Standing up", "Lying prone", and "Normal posture" to help the teacher notice some behaviors of the students in learning time. Grafsgaard [13] measured the postural quantity of motion of hands and head to find the relationship between nonverbal behaviors and dialogue and its effects in learning. Potnis [14] captured videos and detected hand gesture in the classroom to convey it to the lecturer. Sathyanarayana [15] built a dataset called SDMATH, then used HOG and SVM to localize deictic gestures in during lesson time. Zaletelj [16] measured level attention of students in classroom using Kinect

978-1-7281-5163-2/19/$31.00 ©2019 IEEE

device. They combined upper body posture, face gaze points, and facial features to make seven classes for classification, then compared results between these methods. Rich [17] used gesture and speech as one channel to follow up the user engagement during human-robot interactions. Klein [18] built a system called Wits Intelligent Teaching System to track engagement of students by monitoring their real-time feedback. They created an Interesting Map which pointed out students who are engaged or disengaged in the classroom in the lecture time.

In this paper, we propose a model deal with recognizing gesture and posture of students during learning. In our model, students in a classroom are supposed to sit through a lesson, thus student gestures constitute a motion space consisting of the upper body, face, and hands. The system is setup with one camera in front of the classroom, or two cameras in left and right corners in front of the classroom. We also build a dataset to identify student gesture and posture in the classroom to classify whether a student is interested or not in learning lessons. In our experiments, we use some pre-trained network model to train our dataset that is extracted by YOLOv3 [19] from video capturing by three cameras setting up in the left corner, the right corner and in front of a class.

The paper is organized as follows. After the introduction transfer learning is described in section II, the proposed model is presented in section III. Then is the experiment and result in Section IV. Finally, the paper is concluded in Section V.

II. Transfer Learning

Transfer learning is a machine learning method where a model trained on one task is re-purposed on a second related task. This provides a faster and better solution with less effort to recollect the needed training data and rebuild the model [20].

We denote D = $\{\mathcal{X}, P(X)\}$ is a domain where $\mathcal{X}$ is a feature space $\mathcal{X}$ and $P(X)$ is a marginal probability distribution $P(X)$, in which $X = \{x_1, \dots, x_n\} \in \mathcal{X}$. Given a specific domain D, a task $T=\{\mathcal{Y}, f(\bullet)\}$ where $\mathcal{Y}$ is a label space and $f(\bullet)$ is an objective predictive function, which can be learned from the training data pairs $\{x_i, y_i\}$, where $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$. The function $f(\bullet)$ can be used to predict the label of a new instance x, which can be rewritten by the probabilistic form of conditional probability distribution $P(Y|X)$. Then, the Transfer learning is defined as the following:

Given a source domain D_S and learning task T_S, and a target domain D_T and learning task T_T, Transfer learning aims to help improve the learning of the target predictive function f(•) in D_T using the knowledge in D_S and T_S, where $D_S \neq D_T$, or $T_S \neq T_T$ [21].

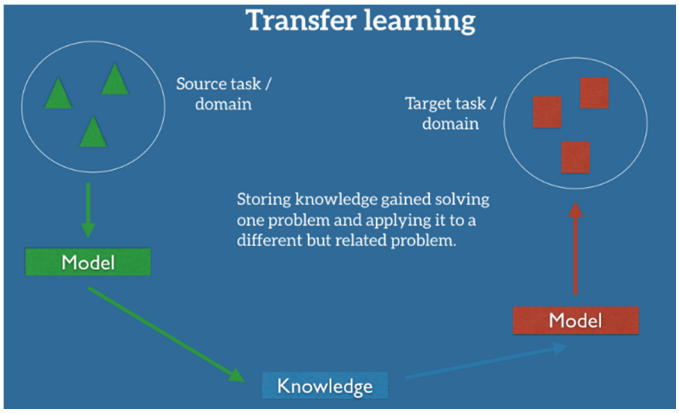

Fig. 1. Illustration of transfer learning method [22]

III. Proposal Model

This paper proposes a student gesture and posture recognition model employed in a classroom with one camera is set up at the front of the classroom or two cameras is set up at the left and right corner in front of the classroom to capture video sequences. Figure 3 shows example of picture from one camera is set in front of the classroom, while figure 4 is example of image capture from the right corner and figure 5 is example of image capture from the left corner at the front of the classroom.

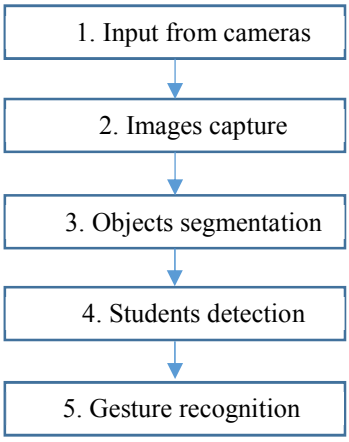

Fig. 2. Student gesture recognition model

Students are assumed to sit stationary in a position for the duration of a lesson, thus student gestures constitute a motion space consisting of the upper body, face, and hands.

Fig. 3. Example of an input sequence from front-camera

Fig. 4. Example of an input sequence from right-camera

Fig. 5. Example of an input sequence from left-camera

Firstly, input sequences from cameras are extracted for a certain period (In this paper we use 30 seconds). We do not use real-time sequences because the student's gesture is stable for a certain during. Secondly, images are extracted from cameras are take into object detection system. We propose using YOLOv3 [19] to extract object. Each object is bounded in a box. Then, we extract all student from the origin image for each object where that type is person. Each small picture of a student then is put into gesture recognition. Figure 6 illustrates object detection from YOLOv3. Each object is bounded in a box. Yellow boxes are students, while orange boxes are chairs.

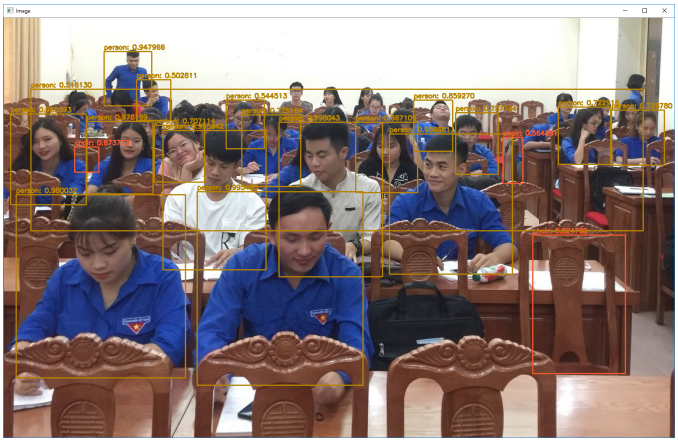

Fig. 6. Example of object detection from a frame

IV. Experiment and Result

A. Dataset

Related to this paper, we built a dataset from several classes in the lecture time. This database is built by extracting images from videos that were captured by some fixed angle cameras in the classroom. According to popular postures of a student in the lecture time, several posture and gesture can be considered as main actions and they were labeled. After collecting, frames are extracted from videos at 30 seconds period. Then student image extracted from the frame was processed by cropping, scaling and then labeling manually.

Fig. 7. Examples of student guesture in classroom

The dataset includes 1102 files which are divided into 8 classes are "Writing", "Reading", "Raising hand", "Concentrating", "Using telephone", "Looking around", "Standing", "Head down". The dataset was divided into two sets, train set and test set with the rate of 80% and 20% respectively. Thus the "Writing" class has 83 training samples and 20 testing samples, the "Reading" class has 94 training samples and 24 testing samples, the "Raising hand" class has

81 training samples and 20 testing samples, the "Concentrating" class has 105 training samples and 26 testing samples, the "Using telephone" class has 130 training samples and 32 testing samples, the "Standing" has 60 training samples and 15 testing samples, the "Looking around" class has 240 training samples and 60 testing samples, and the "Head down" class has 90 training samples and 22 testing samples.

B. Experiment

In recent years, deep learning shows some advantages compare to other learning methods. Especially, Convolution neural network is one of the deep structure with archive state-of-the-art result in various domains. The main components of this network include some convolution layers, Max-pooling layers, Fully-connected layers and a soft-max layer to classify the number of classes. In this paper, we construct a convolution neural network based on some pre-trained network model with the input is a series of 256x256 images which are resized from the origin.

First, we re-use deep convolution network based on two pre-trained models that are VGG16 and VGG19 [23]. These are trained on the ImageNet [24] dataset which contains 1.2 million images [25]. Then, we apply transfer learning [26] with some augmentation techniques such as re-scaling, flipping, cropping, rotating and shifting [27] to improve overall accuracy and avoid over-fitting. The learning rate is set to 0.0001 and momentum of the network is 0.9 in our experiments.

Secondly, these trained model is pop out by freezing some first layers for the next classification process.

Next steps, we use regularization technique with dropout rate is 0.2.

Finally, eight classes which are equivalent to eight gesture and posture are classified by a soft-max activation function. Our proposed procedure was mentioned as below:

Step 1: Create a new model which inherited the pre-trained model.

Step 2: Freeze seven first layers of the pre-trained model.

Step 3: Use weights matrix from the pre-trained model.

Step 4: Use data augmentation on new dataset.

Step 5: Re-compile and re-train model with target dataset.

With this configuration, the VGG19 has 54,108,744 parameters with 53,848,584 trainable parameters and 260,160 non-trainable parameters and the VGG16 has 48,799,048 parameters with 48,538,888 trainable parameters and 260,160 non-trainable parameters.

Our experiment is conducted on an Intel Xeon based system with an 8Gb memory GPU card GeForce GTX 1070 and CUDA 9.2 library. Keras framework with TensorFlow backend is chosen for developing our convolution network.

C. Result

Although the comparison to other researches is not appropriate because of the different datasets and methods, we just show the accuracy of the other similar papers to prove that the achievement is acceptable when applying a small dataset. compare result on some approaches.

TABLE I. COMPARISION ACCURACY OF TWO MODELS

Method	Accuracy
Applied transfer learning with VGG16	88.9%
Applied transfer learning with VGG19	87.3%

In our experiment, our accuracy gained 88.9% in VGG16 and 87.3 in VGG19 compared to the Fang experiment [12] gained 72% and 89.7% of Klein [18] result with 20.000 training epochs and over 200.000 images.

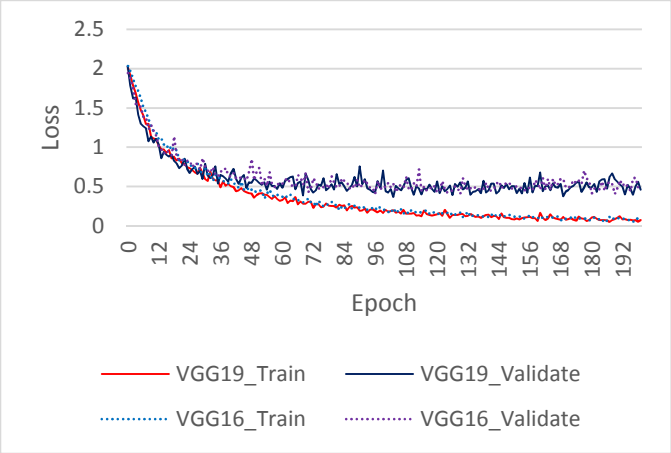

Fig. 8. Loss curves of training and validating process

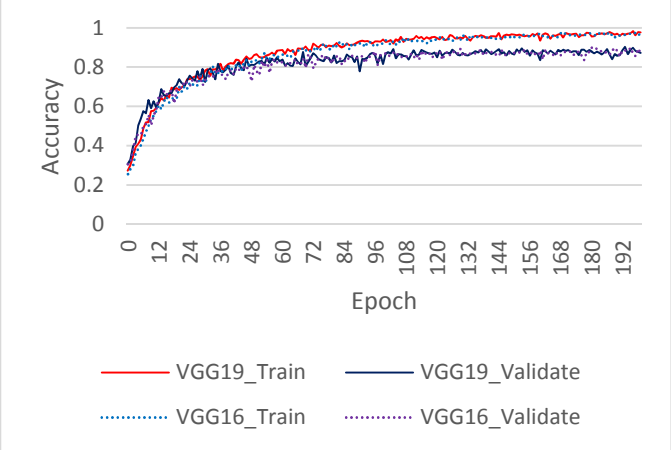

Fig. 9. Accuracy curves of training and validating process

Figure 8 and figure 9 show loss curves and accuracy curves, respectively. After approximate 200 iterations, the validating loss/accuracy does not change. The result shows that VGG16 has higher accuracy than VGG19 and both results have lower

accuracy compare to Klein experiment. The reasons why the result is not high enough compared to Klein experiment [18] because our model uses the smaller dataset and the quality of image extract from cameras is still low.

V. CONCLUSION

Postures and gestures of students in the classroom reflect the learner's interest in the content and teaching methods. The ability of analysis attitude of student may help to support teachers and managers in improving quality of training services in an adaptive learning model.

In this paper, we have proposed a model for recognizing student's posture and gesture in the classroom by using YOLOv3 to extract students from frames in camera's videos and use transfer learning method on some pre-train model that success in image classification. The experiment shows that applying transfer learning and convolution neural network can classify effectively student gesture and posture in the classroom based on predefined action labels. The result also shows that this model can be applied in an adaptive learning system to improve quality of training services.

In the future, to improve the overall accuracy of the classification process by collecting more training data samples. Besides this, we can optimize YOLOv3 library to improve the quality of object detection.

Finally, this model is only the first step of an adaptive training system. To support teachers in the classroom or some of the e-learning system, we need to build more components such as domain knowledge model, student trait model, human-machine interface and other necessary modules for a complete adaptive training system.

REFERENCES

[1] J. Smallwood, D. J. Fishman, and J. W. Schooler, "Counting the cost of an absent mind: Mind wandering as an underrecognized influence on educational performance," *Psychonomic Bulletin & Review*, vol. 14, pp. 230–236, Apr. 2007.

[2] S. D'Mello, "A selective meta-analysis on the relative incidence of discrete affective states during learning with technology.," *Journal of Educational Psychology*, vol. 105, no. 4, pp. 1082–1099, 2013.

[3] J. Whitehill, Z. Serpell, Y. C, Lin, A. Foster, and J. R. Movellan, "TRANSACTIONS ON AFFECTIVE COMPUTING The Faces of Engagement: Automatic Recognition of Student Engagement from Facial Expressions," *IEEE Transactions on Affective Computing*, vol. 5, no. 1, pp. 86 - 98, Mar. 2014.

[4] M. A. A. Dewan, M. Murshed, and F. Lin, "Engagement detection in online learning: a review," *Smart Learning Environments*, vol. 6, 2019.

[5] M. Nezami, H. Len, R. Deborah, and D. Mark, "Deep Learning for Domain Adaption: Engagement Recognition," *Computer Vision and Pattern Recognition*, 2018.

[6] I. J. Goodfellow, D. Erhan, and P. L. Carrier, "Challenges in representation learning: A report on three machine learning contests," *Neural Networks*, vol. 64, pp. 59–63, 2015.

[7] N. Bosch, "Detecting Student Engagement : Human Versus Machine," *24th Conf. User Model. Adapt. Pers.*, pp. 317–320, 2016.

[8] S. Raina, L. Bernard, B. Taylor, and S. Kaza, "Using eye-tracking to investigate content skipping: A study on learning modules in cybersecurity," in *IEEE Conference on Intelligence and Security Informatics (ISI)*, 2016.

[9] L.B. Krithika and P.G. Lakshmi, "Student emotion recognition system (SERS) for e-learning improvement based on learner concentration metric," *Procedia Computer Science*, vol. 85, pp. 767–776, 2016.

[10] S. Mitra and T. Acharya, "Gesture Recognition: A Survey," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 37, no. 3, pp. 311-324, 2007.

[11] K. Wan and H. Sawada, "Dynamic Gesture Recognition Based on the Probabilistic Distribution of Arm Trajectory," in *Proceedings of International Conference on Mechatronics and Automation, Takamatsu*, 2008, pp. 426-431.

[12] C.Y. Fang, M.H. Kuo, G.C. Lee, and S.W. Chen, "Student gesture recognition system in classroom 2.0," in *Computers and Advanced Technology in Education*, 2011.

[13] J.F. Grafsgaard, J.B. Wiggins, K.E. Boyer, E.N. Wiebe, and J.C. Lester, "Embodied affect in tutorial dialogue: Student gesture and posture," in *Lecture Notes in Computer Science - Springer*, Berlin Heidelberg, 2013, pp. 1–10.

[14] S.S. Potnis and A.S. Jahagirdar, "Real time hand gesture recognition for smart classroom environment," *International Journal of Computer Trends and Technology*, vol. 17, pp. 78–83, 2014.

[15] S. Sathayanarayana et al., "Towards automated understanding of student-tutor interactions using visual deictic gestures," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014.

[16] J. Zaletelj and A. Koˇsir, "Predicting students' attention in the classroom from kinect facial and body features," *EURASIP Journal on Image and Video Processing*, vol. 2017, 2017.

[17] C. Rich, B. Ponsleur, A. Holroyd, and C.L. Sidner, "Recognizing engagement in human-robot interaction," in *Proceeding of the 5th ACM/IEEE international conference on Human-robot interaction -HRI '10*, 2010, p. 375.

[18] R. Klein and T. Celik, "The wits intelligent teaching system: Detecting student engagement during lectures using convolutional neural networks," in *2017 IEEE International Conference on Image Processing (ICIP), IEEE*, 2017.

[19] Redmon Joseph and Farhadi Ali, "YOLOv3: An Incremental Improvement," *Computer Vision and Pattern Recognition*, vol. arXiv:1804.02767, 2018.

[20] S. J. Pan and Q. A. Yang, "A survey on transfer learning," in *IEEE Trans. Knowledge and Data*, 2010, pp. 345–1359.

[21] Yuan-Pin Lin and Tzyy-Ping Jung, "Improving EEG-Based Emotion Classification Using Conditional Transfer Learning," in *Frontiers in Human Neuroscience.11*, 2017.

[22] (2017, Mar.) http://ruder.io/transfer-learning/.

[23] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *Computer Vision and Pattern Recognition*, 2014.

[24] J.D.J. Deng, W.D.W. Dong, and R. Socher, "ImageNet: A Large-Scale Hierarchical Image Database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 2–9.

[25] A. Krizhevsky, I. Sutskever, and G.E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances In Neural Information Processing Systems*, 2012, pp. 1–9.

[26] L. Torrey and J. Shavlik, "Transfer Learning," in *Encyclopedia of the Sciences of Learning*, 2012, pp. 3337–3337.

[27] L. Wang et al., "Temporal segment networks: Towards good practices for deep action recognition," in *Computer Vision – ECCV* , 2016, pp. 20–36.