A Video Analytic In-Class Student Concentration Monitoring System

Mu-Chun Su, *Senior Member, IEEE*, Chun-Ting Cheng, *Member, IEEE*, Ming-Ching Chang, *Senior Member, IEEE*, and Yi-Zeng Hsieh, *Member, IEEE*

***Abstract*—Automatic learning feedback monitoring and analysis are becoming essential in modern education. We present a video analytic system capable of monitoring in-class student's learning behaviors and providing feedback to the instructor. It is a common practice nowadays for students to take electronic notes or browse online using laptops and cellphones in class. However the use of technology can also impact student concentration and affect learning behaviors, which can seriously hinder their learning progress if not controlled properly. In this pioneering study, we propose a non-intrusive deep-learning based computer vision system to monitor student concentration by extracting and inferring high-level visual behavior cues, including their facial expressions, gestures and activities. Our system can automatically assist instructors with situational awareness in real time. We assume only RGB color images as input and runable system on edge devices for easy deployment. We propose two video analytic components for student behavior analysis: (1) The facial analysis component operates based on Dlib face detection and facial landmark tracking to localize each student and analyze their face orientations, eye blinking, gazes, and facial expressions. (2) The activity detection and recognition component operates based on OpenPose and COCO object detection can identify eight types of in-class gestures and behaviors including raising-hand, typing, phone-answering, crooked-head, desk napping, *etc.* Experiments are performed on a newly collected real-world In-Class Student Activity Dataset (ICSAD), where we achieved nearly 80% activity detection rate. Our system is view-independent in handling facial and pose orientations with average angular error < 10°. The source code of this work is at : https://github.com/YiZengHsieh/ICSAD.**

***Index Terms*—Video analytics, deep learning, student concentration, behavior cues, face detection, landmark tracking, facial orientation, facial expression, eye blinking, yawning, posture analysis, object detection, activity detection.**

Manuscript received May 21, 2021; revised October 14, 2021; accepted November 6, 2021. Date of publication November 10, 2021; date of current version December 23, 2021. This work was supported in part by the Ministry of Science and Technology, Taiwan, under Grant MOST 110-2221-E-019-051, Grant MOST 109-2622-E-019-010, Grant MOST 109-2221-E-019-057, Grant MOST 109-2622-E-008-018-CC2, Grant MOST 109-2221-E-008-059-MY3, Grant MOST 110-2634-F-019-001, Grant MOST 110-2634-F-008-005, Grant MOST 109-2221-E-008-059-MY3, Grant MOST 107-2221-E-008-084-MY2, Grant MOST 109-2634-F-008-007, Grant MOST 109-2221-E-019-057, Grant MOST 107-2221-E-019-039-MY2, and Grant MOST 109-2634-F-019-001, and in part by LSH-NCU Joint Research Foundation under Grant NCU-LSH-109-B-010, Grant NCU-LSH-108-A-007, and Grant NCU-LSH-108-A-009. *(Corresponding author: Yi-Zeng Hsieh.)*

Mu-Chun Su and Chun-Ting Cheng are with the Department of Computer Science and Information Engineering, National Central University, Taoyuan 320317, Taiwan.

Ming-Ching Chang is with the Computer Science Department, University at Albany, State University of New York, Albany, NY 12222 USA.

Yi-Zeng Hsieh is with the Department of Electrical Engineering and the Center of Excellence for Ocean Engineering, National Taiwan Ocean University, Keelung 20224, Taiwan (e-mail: yzhsieh@mail.ntou.edu.tw).

Digital Object Identifier 10.1109/TCE.2021.3126877

I. Introduction

THE RAISE of mobile technologies is imperative on impacting our everyday life. However the globally increasing popularity of smart phones are seen as mixed blessing especially for the teens. Recent researches suggest that in education and learning, student concentration can decrease with the use of smartphones and laptops [2], [25]. The question of whether allowing or regulating in-class technology usage is becoming a major debate. The concern is how best to leverage the power of technology in learning, while keeping learners from unnecessary distractions. Regularization and policies on digital distractions can have long-term impact to the study behaviors, attendance, and learning outcome. Student distractions caused by technology usage are also important from the instructor points of view. Immediate actions can be taken to improve student concentration, if prompt awareness of student distractions is feasible. In this study, we propose a non-intrusive computer vision system for student learning behavior and concentration monitoring. By detection and recognizing the student's facial attentions and body poses, our video analytic algorithm can recognize student behaviors and oversees the in-class concentration status.

The proposed video analytic system consists of a camera and an edge computer set up in a typical classroom environment. The camera captures RGB color videos of students in a lecturing classroom. Our system aims to recognize student fatigues and in-class behaviors and activities to estimate overall concentration status. The system can automatically handle view variations by estimating the face and body pose orientations in each student's coordinate system.

We developed two video analytic components for student behavior and concentration monitoring. Our system leverages popular open-source deep neural network (DNN) packages for effective visual detection and tracking, see Fig. 1: (1) A **facial analysis** component operates based on Dlib [11] face detection facial landmark tracking that can localize each student and analyze their face orientations, eye linking, gazes, and facial expressions. (2) An **activity detection and recognition** component operates based on OpenPose [5] can analyze body gestures for activity detection. We developed a DNN model to detect and recognize up to 8 types of student behaviors that are relevant for concentration and distraction analysis. These include note writing, typing, phone texting, raising hand, crooked head, desk napping, one-handed head, and phone talking. Activity detection accuracy can be improved

1558-4127 © 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: JIANGSU NORMAL UNIVERSITY. Downloaded on July 20,2023 at 17:05:25 UTC from IEEE Xplore. Restrictions apply.

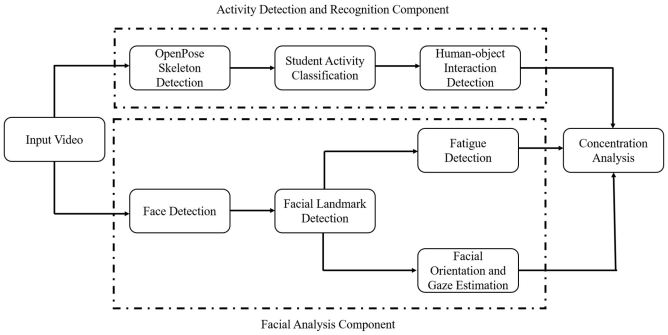


Fig. 1. Overview of the proposed in-class student concentration analysis pipeline.

by better recognizing human-object interactions, via the detection of commonly appearing classroom objects (such as cell phones, note books, calculators) using the YOLO visual object detector [20].

Contributions of this paper is summarized in the following.

1) We propose an automatic, non-intrusive video analytic system for in-class student concentration monitoring with extensive experiments and evaluations.
2) The proposed facial analysis pipeline consists of face detection and landmark tracking for a line of comprehensive facial expression analyses, including the detection of per-frame face orientation, eye blinking, gaze, and facial expression analysis of each student. These information can be used to reliably infer closed or blinking eyes, yawning, and gaze for attention and concentration analysis.
3) The proposed student activity analysis pipeline is based on the detection of student skeletal gestures and the recognition of common classroom objects. The system can recognize up to 8 types of student activities via a view-point invariant analysis.
4) Experiments are performed on a newly collected real-world **In-Class Student Activity Dataset (ICSAD)**. Results in confusion matrices show that our system can achieve nearly 80% activity detection rate, with average angular error $< 10°$. The TCSAD dataset consists of 10 student subjects in 400 video clips, 8 classroom activities in 3 viewing angles, which will be released upon paper publication. Evaluations on facial orientation and expression (yawn detection) are also performed on the YawDD [1] dataset.
5) The proposed analytic algorithms can run with standard RGB cameras on lightweight edge devices in real-time, without the need to store or transmit visual data to a remote cloud. This on-line, edge-computing design ensures easy deployment and can largely alleviate privacy concerns.

This paper is organized as follows. Section II reviews background knowledge and literature survey. Section III describes our facial analysis pipeline, which can extract student attention and potential fatigue cues. Section IV describes our pose and activity analysis pipeline, where in-class student behaviors, gestures and actions can be directly monitored to infer the concentration status. Section V describes experimental validation, evaluation metric and results performed on open datasets and our newly collected in-class student activity dataset. Section VI concludes this work.

II. Background

A. Review of Concentration Estimation

Concentration is often defined as receiving external information with sustained attention [10]. In this work, we view **learning concentration** as maintaining long-term attention in class. Since student learning concentration is closely related to their learning outcomes, abundant works in education research were investigated on how best to assess and retain learning concentration. Classic studies of human attention mainly relied on *intrusive* sensor-based methods, such as the Electroencephalogram (EEG) sensing. In [6] and [14], a Support Vector Machine (SVM) model is trained to recognize EEG brain waves for assessing students attention and concentration. These sensor-based approach is difficult to implement in real-world classroom settings.

With the recent advancement of deep learning and artificial intelligence, video sensing approaches are becoming versatile. Such video analytic methods are *non-intrusive* and can now reliably recognize rich visual signals for behavior and sentiment analysis. However, privacy and ethics issues arise along with these video-based methods [9]. In [19], three cameras are used to capture student views in a classroom environment, where the student line of slight and body movements are only roughly estimated. Due to the insufficient visual information, this method cannot reliably estimate student concentration. In [26], student concentration is estimated at five levels. With Kinect-based action detection and manual estimation of student concentration levels at every second of the video, and machine learning models are developed for student concentration classification.

The major differences between our model and the model proposed in [26] are as follows: (1) Our approach works on a single RGB camera (with ease of deployment), while their method requires a Kinect sensor to work. (2) Our system estimates detailed student attention measures (including fatigue status, eye blinking, facial orientation, eye gaze, and eight types of in-class gestures and behaviors) of students, while [26] only estimates a three-level scale (high, medium, low) of student attention. Our system provides rich contextual information and explainable cues for the reported student concentration status. (3) As for detecting student activities involving body movements, we detect eight types of in-class student activities (raising-hand, typing, phone-answering, crooked-head, desk napping, etc.), while [26] detects only four student activities (observing slides, writing note, lean forward, and lean back). Our monitoring system (at its current state) is implemented in a classroom setting for providing analytical results for the instructors to review after classes. We plan to implement our system in an on-line video conference setting as a future work, such that the instructor can receive live attention information from the students.

B. Review of Head/Body Pose Orientation Estimation

Head pose or body orientation can be represented in Euler angles in terms of *yaw* (rotation along the y-axis), *pitch*

Authorized licensed use limited to: JIANGSU NORMAL UNIVERSITY. Downloaded on July 20,2023 at 17:05:25 UTC from IEEE Xplore. Restrictions apply.

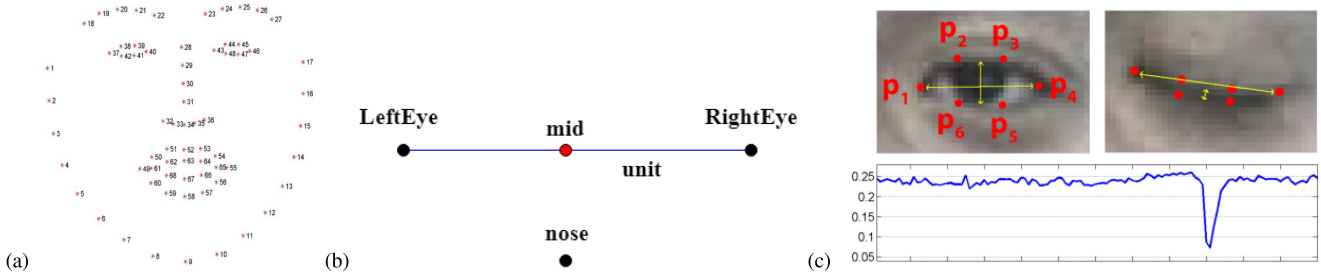


Fig. 2. (a) The 68 facial landmarks detected by FAN from [3]. (b) Facial features normalized per unit length. (c) Feature points of the open and closed eyes adopted from [22].

(rotation along the x-axis) and *roll* (rotation along the z-axis), respectively. Head pose orientation estimation methods in the literature can be divided into two categories: *learning-based* and *model-based* [17] methods.

Learning-based approaches use data-driven machine learning to determine the correspondence between the face(s) in the image and the underlie orientation angles in the 3D world. In [8], K-clusters Regression Forest (KRF) is used to estimate the face orientation. A flexible node splitting algorithm using a regression tree is proposed for head pose estimation. In [27], a Supervised Descriptor Learning (SDL) algorithm establishes a discriminative and compact feature representation for multi-output head pose regression.

Model-based approaches estimate the unknown head pose by inferring the projection matrix that maps a 3D face model into the 2D image view. In [15], the POSIT algorithm estimates such projection matrix for driver's head pose estimation. A cascade regressor is used in [23] to predict the facial feature coordinates in 3D space from the input 2D images and calculate face orientation estimations. In [12] paper, Faster R-CNN method has been developed to address the infant monitoring problem. However, the proposed system is person-dependent, it is difficult to model a generalized way for daily usage.

III. Facial Cue Analysis

We focus on two kinds of facial cues that are relevant to student concentration. The first is the **gaze attention** of the students in their facial orientations (e.g., focusing on or listening to the instructor). The second is **fatigue** that is reflected by the *blinking eyes*, *closed eyes*, or *yawning*. Both cases can be reliably detected via detecting of facial landmark points including the eyes and month.

Our facial analysis pipeline starts with facial landmark detection and feature extraction. Facial features then are scaled against the eye distances for normalization that compensates the viewpoint, orientation, and scale variations. After such normalization, all feature calculations can be performed *w.r.t.* each individual's own coordinate to achieve feature invariance (Section III-A). In Section III-B, the identified eye open/close states are used to detect closed eyes and eye blinking. In Section III-C, the identified mouth landmarks are used to detect yawning as an indication of fatigue. In Section III-D, the identified facial landmarks are used to estimate the facial orientation and eye gaze directions, which provides an indicator for attention and learning concentration.

Fig. 1 overviews the proposed in-class student concentration analysis architecture, where the process consists of two pipelines. The first pipeline starts with face detection and localization. Facial features are extracted and normalized to detect possible fatigue behaviors such as closed eyes, eye blinking and yawning. Facial and gaze angles are determined for concentration estimation. The second pipeline starts with person localization and pose estimation. Space-time skeleton features are extracted for activity and gesture classification.

A. Facial Landmark Detection and Feature Normalization

Facial Landmark Detection: We use the open-source Dlib [11] CNN model for face detection. For each detected face, we use the Face Alignment Network (FAN) [4] to accurately detect the 68 facial landmarks as shown in Fig. 2a. This enables accurate localization of the eyes, nose and mouth in the image. The FAN network structure consists of 4 stacked Hourglass (HG) networks and a residual module.

Facial feature normalization is required to deal with multiple subjects in the camera review, where each subject (student) sitting in the classroom can appear at arbitrary distances, scales, and viewing angles. Their facial features will also differ with respect to head rotations. In order to remove these variations, it is necessary to normalize the facial feature representation. We perform two types of face landmark normalization *w.r.t.* geometrical translation and scaling as shown in Fig. 2b.

For **translation normalization**, we calculate the midpoint between the two eyes as the new origin to offset the translations. Specifically, for the facial landmarks depicted in Fig. 2a and at time t, let $p_{LeftEye}(t)$ denote the left outer landmark of the left eye (p_{46}), and $p_{RightEye}(t)$ denote and the right outer landmark of the right eye (p_{46}). We offset all landmark points by the midpoint $p_m(t)$ between the two eyes as the following:

$$p_m(t) = \frac{p_{LeftEye}(t) + p_{RightEye}(t)}{2}, \tag{1}$$

$$\hat{p}_x(t) = p_x(t) - p_m(t), \tag{2}$$

where $p_x(t)$ and $\hat{p}_x(t)$ denote each feature point x in the original and translated coordinates at time t, respectively.

For **scaling normalization**, we calculate the normalized unit distance $d_u(t)$ as the distance between $p_{LeftEye}(t)$ and $p_{RightEye}(t)$, and scale all translated landmark points $\hat{p}_x(t)$ by $d_u(t)$ as the following:

$$d_u(t) = \left|p_{LeftEye}(t) - p_{RightEye}(t)\right|, \tag{3}$$

Authorized licensed use limited to: JIANGSU NORMAL UNIVERSITY. Downloaded on July 20,2023 at 17:05:25 UTC from IEEE Xplore. Restrictions apply.

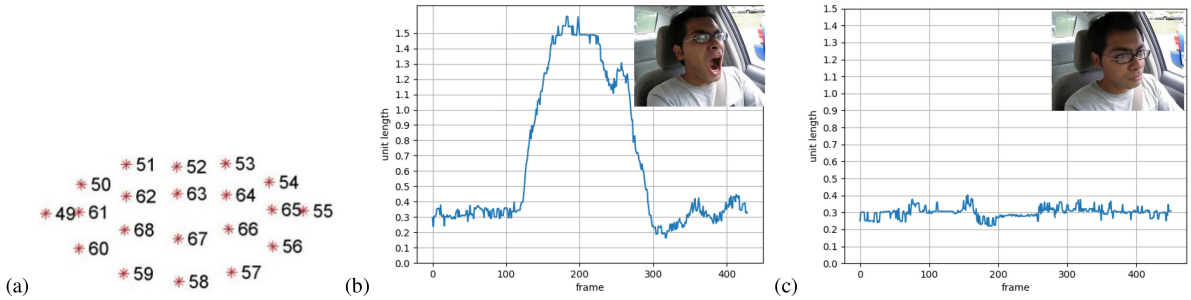


Fig. 3. **Yawning detection:** (a) Zoom-in to the mouth landmark points from Fig. 2a from [3]. (b) shows a yawning case, and (c) shows a typical closed mouth, where y-axis depicts l_{Mouth} in Eq. (6) and x-axis is video frame (time).

$$\tilde{p}_x(t) = \frac{\hat{p}_x(t)}{d_u(t)}, \tag{4}$$

where $\tilde{p}_x(t)$ denotes each feature point x after the translation and scaling normalization at time t.

B. Closed Eyes and Blinking Detection

We adopt the fast method in [22] to detect closed eyes based on the 6 eye landmark points ($\mathtt{p}_1$ to $\mathtt{p}_6$) in Fig. 2c to calculate the Eye Aspect Ratio (*EAR*):

$$EAR = \frac{|\mathtt{p}_2 - \mathtt{p}_6| + |\mathtt{p}_3 - \mathtt{p}_5|}{2|\mathtt{p}_1 - \mathtt{p}_4|}. \tag{5}$$

The lowest *EAR* value is used to determine closed eyes after thresholding. Specifically, if $EAR < 0.2$ over a preset period of 150ms, a *closed* eye is detected. If the *EAR* value oscillates within the time period, a *blinking* eye is detected.

C. Yawning Detection

The inhalation and exhalation of breath when yawning cause an open mouth for a period of time. Yawning can thus be robustly recognized by detecting the mouth opening and closing as in Fig. 3. We calculate the vertical distance d_{Mouth} between the midpoints of the upper and lower lips (p_{52} and p_{58} in Fig. 3a, respectively) to determine the mouth opening action unit:

$$d_{Mouth} = |p_{52} - p_{58}|. \tag{6}$$

In our experiments, for non-yawning cases, typically d_{Mouth} is about 0.2 to 0.4 units after landmark normalization. When yawning, d_{Mouth} increases to about 0.5 units for 100 frames.

We calculate the beginning of yawning when d_{Mouth} exceeds a threshold of 0.5 the width of shoulder. If this yawning status exceeds a time period with $d_{Mouth} >$ the width of shoulder, the yawning action unit is detected.

D. Facial Orientation and Gaze Estimation

We estimate face and gaze orientations using Euler angles. Specifically, as in Fig. 4a, the *yaw* angle variance is calculated as the distance between nose tip (p_{34} in Fig. 2a) and the two eye corners (p_{37} and p_{46}). In Fig. 4b, the *pitch* angle variance related to the depth information is calculated as the distance between the nose tip (p_{34}) to the two month corners (p_{49} and p_{55}).

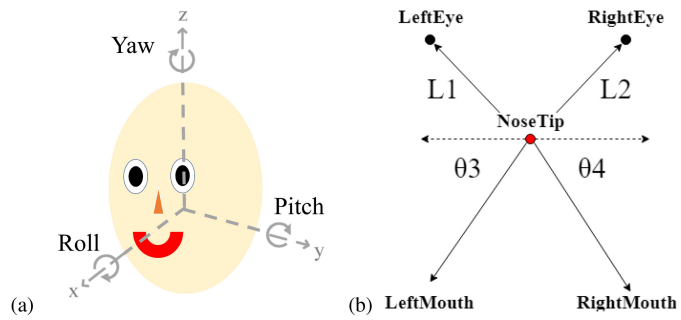


Fig. 4. (a) The yaw, pitch, and raw for head pose and face orientation estimation (figure adopted from [24, Fig. 5]). (b) Yaw represents the orientation angle. Facial orientation angle estimation characteristics schematic.

The facial orientation are calculated via the *roll* angle variance using the above five facial feature points again (left eye corner p_{46}, right eye corner p_{37}, nose tip p_{34}, left mouth corner p_{55}, right mouth corner p_{49}). We next describe the facial orientation estimation process in details.

Distance variance of Facial Landmark Points: Yaw and Pitch angle changes affect the distance between the left and right eyes and the tip of the nose, as well as the distance between the left and right corners of the mouth and the tip of the nose. This study thus adopted the corners of the eyes and mouth as four feature points to calculate the distance between them and the tip of the nose. The calculation process is expressed as:

$$d_x = \left|p_x - p_{NoseTip}\right|, \tag{7}$$

where d_x is the distance between the feature points and the tip of the nose; p_x is the feature points, and $p_{NoseTip}$ is the tip of the nose; and L_1 and L_2 are the distances between the left and right eye and the tip of the nose, as shown in Fig. 4.

Point Features of Angular Change: Angle changes in Yaw and Pitch are identified by changes in the distances between the eyes and the tip of the nose, in the horizontal angle of the mouth and around the tip of the nose, while Roll angle changes with the left and right corner to a horizontal angle about the tip of the nose. This study used the angles related to the eyes to determine facial orientation. The four feature points around the mouth are the tip of the horizontal angle as characterized in the determination, derived using Equations (8) and (9):

$$\vec{u} = p_x - p_{NoseTip}, \tag{8}$$

Authorized licensed use limited to: JIANGSU NORMAL UNIVERSITY. Downloaded on July 20,2023 at 17:05:25 UTC from IEEE Xplore. Restrictions apply.

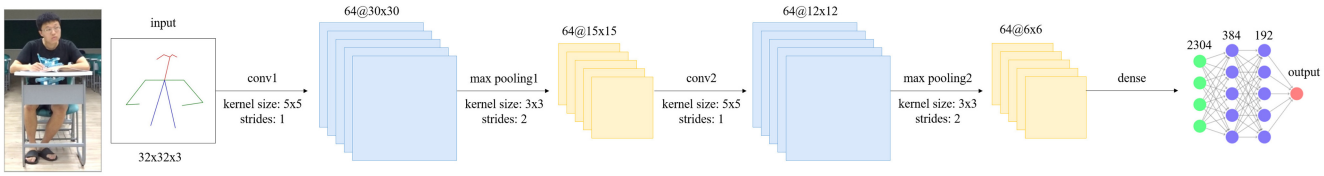


Fig. 5. The input image and student pose skeleton are fed into the proposed CNN for in-class activity recognition.

TABLE I
STUDENT FACIAL ORIENTATIONS CALCULATION

Student focus	Student pitch	x Student yaw
Left blackboard	0° ∼ 30°	-45° ∼ -15°
Center blackboard	0° ∼ 30°	-15° ∼ 15°
Right blackboard	0° ∼ 30°	15° ∼ 45°
Student's desktop	-30° ∼ 0°	-45° ∼ 45°

$$\theta_x = \cos^{-1} \frac{\vec{u} \cdot \vec{v}}{|\vec{u}||\vec{v}|}, \tag{9}$$

where $\vec{u}$ is the feature vector of the nose, $\vec{v}$ is a horizontal vector, and θ_x is the horizontal angle between nose tip and feature point. Taking the sides of the mouth as an example in Fig. 4, the dotted line is the horizontal vector, and θ_3 and θ_4 are the angles between the nose tip and left/right sides of the mouth. Once the feature vectors are estimated, students' facial orientation can be calculated. Table I lists the student facial orientations calculated using the proposed method.

IV. POSE AND ACTIVITY ANALYSIS

The degree of a student's focus can be determined by their physical behavior. For example, sitting with their head supported by one hand may indicate that a student is feeling bored. This study adopts **OpenPose** to extract a skeleton feature vector, which is then used as an input for the proposed neural network to evaluate a student's behavior. This method involves several processes, described below.

We use Openpose to extract skeletal information. Openpose is an open source skeleton pose estimation method [5]. After the skeletal information is extracted, it is normalized, and then used as the input of a neural network in order to identify body positions and movements. A sliding window is adopted to increase the stability of the system. Finally, the posture information is combined with motion detection, distraction behavior detection and face orientation in order to analyze a subject's concentration level.

A. Human Pose Estimation and Pose Feature Normalization

Skeleton size is related to the position of the camera, and will vary accordingly based on how far a subject is seated from the camera. This, in addition to the fact that everyone's skeleton differs, means that skeleton information must be normalized to avoid errors. Normalization consists of data shift, and data scaling.

In data shift, the neck is the origin and every point position is subtracted from the neck position to evaluate a new position, as expressed in Equation (10):

$$j'_x(t) = j_x(t) - j_{neck}(t), \tag{10}$$

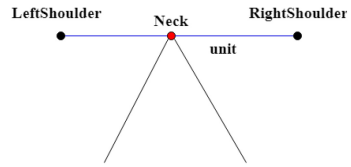


Fig. 6. Skeleton information normalized per unit length.

where $j_{neck}(t)$ is the neck coordinates at time t; $j'_x(t)$ is the new coordinates obtained by translating feature point x at time t; $j_x(t)$ are the new coordinates of the feature point x at the time t.

In data scaling, the shoulder length is the unit length. The original coordinates are divided by the unit length, yielding new coordinates, using Equations (11) and (12):

$$L_{unit}(t) = \left| j_{leftShoulder}(t) - j_{rightShoulder} \right|, \tag{11}$$

$$j''_x(t) = \frac{j'_x(t)}{j_{unit}(t)}, \tag{12}$$

where $L_{unit}(t)$ is the unit length at time t; $j''_x(t)$ is the new coordinates after translating and scaling. In Fig. 6, the blue line is the shoulder length.

B. Student Activity Classification

Using an image as the input of a convolution neural network results in some errors caused by background interference or other objects. The false recognition in [16] may be due to this reason. In [18], the dataset is divided into four classes, but this obviously cannot fully present the situation in the classroom. Our proposed method adopts the extracted skeleton feature from eight common gestures in classroom as the convolution neural network input to improve performance. The kinds of skeleton images are shown in Fig. 11.

This study's motion scenario is a sitting situation, so the skeleton of the lower body is ignored. The normalized skeleton is divided into three parts. The first part is the head, including features like the left ear, right ear, left eye, right eye and nose. The second part is the upper body, including features like the left wrist, left elbow, left shoulder, neck, right shoulder, right elbow, and right wrist. The third part is the lower body, including the left hip and right hip. Fig. 5 shows the skeleton mapping to the human body.

CNN Architecture: The architecture of the proposed convolutional neural network (CNN) is shown in Fig. 5. The input image is the skeleton image resized to 32x32. The CNN consists of two convolutional layers, two pooling layers and a fully connected layer. The kernel size is 5x5, and stride size is 1. The pooling size is 3x3, and the stride size is 2. Zero-padding is adopted. The fully connected layer has 2 hidden layers and

Authorized licensed use limited to: JIANGSU NORMAL UNIVERSITY. Downloaded on July 20,2023 at 17:05:25 UTC from IEEE Xplore. Restrictions apply.

1 output layer. The first hidden layer consists of 384 neurons, and the second hidden layer consists of 192 neurons. The output layer is determined by the number of gestures made.

C. Detecting Human-Object Interactions

Since some actions of the skeleton image are difficult to distinguish, additional information is required to determine those actions. The most common scenario in this study is students using objects such as books, notebooks, smartphones, etc. The skeleton is recognized based on which gesture is made; the gesture is then confirmed by classifying some objects. Therefore, this study proposes two object classification methods to improve gesture identification accuracy.

The first method is image classification, in which the region of the object is searched, and the object is then classified according to a classifier; 8 gestures are recognized in this study's classroom scenario. This study classifies note writing using notebooks and smartphone texting by identifying the use of two hands to perform the gesture with the object. A gesture in which one hand supports the head and the other holds a smartphone falls into the category of one hand using an object. This study introduces two object segment region categories as follows.

The first is the two hands on the table category. The center of the two hands is also the center of the object, as shown in equation (13):

$$o_{mid}(t) = \frac{j_{leftWrist}(t) + j_{rightWrist}(t)}{2}, \tag{13}$$

where $o_{mid}(t)$ is the center of the object at time t; $j_{LeftWrist}(t)$ is the center coordinates of the left wrist at time t; and $j_{RightWrist}(t)$ is the center coordinates of the right wrist at time t.

Then, in order to cover the range of the object, this study uses the center of the shoulder length to extend the range, as shown in Equations (14), (15) and (16):

$$L_{shoulder}(t) = \left|j_{leftShoulder}(t) - j_{rightShoulder}(t)\right|, \tag{14}$$

$$o_{topleft}(t) = o_{mid}(t) - L_{shoulder}(t), \tag{15}$$

$$o_{bottomRight}(t) = o_{mid}(t) + L_{shoulder}(t), \tag{16}$$

where $L_{shoulder}(t)$ is the shoulder length at time t; $j_{leftShoulder}(t)$ and $j_{rightShoulder}(t)$ are the left and right shoulder coordinates at time t, respectively; and $o_{topLeft}(t)$ and $o_{bottomRight}(t)$ are the upper-left corner and right-bottom corner coordinates at time t, respectively.

The second category is that of one hand using an object. When using an object, the reference point is usually the wrist. However, in this case, because the range of the object is at the top of the wrist, the reference is shifted up by half of the shoulder length as the center of object. Distinction between right or left hands is made by the y coordinate of the hand. Equations (17) and (18) are used for this category:

$$o_{ref}(t) = \begin{cases} j_{leftWrist}(t), & j^y_{leftWrist}(t) < j^y_{rightWrist}(t) \\ j_{rightWrist}(t), & j^y_{rightWrist}(t) < j^y_{leftWrist}(t) \end{cases}, \tag{17}$$

$$L_{shoulder}(t) = \left|j_{leftShoulder}(t) - j_{rightShoulder}(t)\right|, \tag{18}$$

$$o_{mid}(t) = \left(o^x_{ref}(t), o^y_{ref}(t) - \frac{l_{shoulder}(t)}{2}\right), \tag{19}$$

TABLE II
TWO CATEGORIES OF OBJECT REGION SEARCH

motion	Writing note	Using notebook	Texting smartphone
object	note	notebook	smartphone
result			
motion	One hand head	Answering smartphone	
object	none	smartphone	
result			

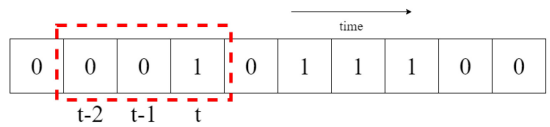


Fig. 7. Sliding window.

where $o_ref(t)$ is the reference coordinate at time t; and $j^y_{leftWrist}(t)$ and $j^y_{rightWrist}(t)$ are the y-axis of the left and right wrists at time t, respectively. Because the object scale is small, this study extends the range by half of the shoulder length from the object center, as shown in Equations (20) and (21):

$$o_{topLeft}(t) = o_{mid}(t) - \frac{L_{Shoulder}(t)}{2}, \tag{20}$$

$$o_{bottomRight}(t) = o_{mid}(t) + \frac{L_{Shoulder}(t)}{2}, \tag{21}$$

Table II shows the two object region search categories.

The classifier adopted in this study is VGGNet [21], which won first prize in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). This study adopts architecture D from the table, and the input image size is 224 × 224. There are 16 convolutional layers and 4 pooling layers, connected with fully connected layers composed of 2 hidden layers and one output layer. Architecture D is also called VGG16.

The second object classification method consists of object detection and locating the position of the object, so the You Only Look Once (YOLO) [20] method is adopted. Its advantage is its faster processing while still maintaining good accuracy. YOLO's main characteristic is its adoption of the CNN model to train the whole image. After predicting the whole image, the image is segmented into many regions, and the center of each region can determine many bounding boxes, which each have confidence scores and categories. Every bounding box has 5 values, including center coordinates, height, width, and confidence score.

Sliding-Window Post Processing: This study identifies every action in each frame of the camera input. In order to ensure system stability, and avoid noise interference, a sliding window and the 'winner take all' method were adopted to achieve sequential motion detection. Using the sliding window, each action in 3 frames can be identified. Frames $t-2$, $t-1$ and t are recognized as 0, 0 and 1, respectively. The 'winner take all' method is then used to identify the action as 0.

Authorized licensed use limited to: JIANGSU NORMAL UNIVERSITY. Downloaded on July 20,2023 at 17:05:25 UTC from IEEE Xplore. Restrictions apply.

TABLE III
FATIGUE SITUATION

Situational description	Fatigue behavior	Facial orientation	Motion recognition
Non-class distraction for a long time		Desktop facing	Laptop usage, phone texting
			Call answering
Insufficient motivation, bored by class content	Closed eyes, blinking, yawning		
			Sleepy, one-handed head

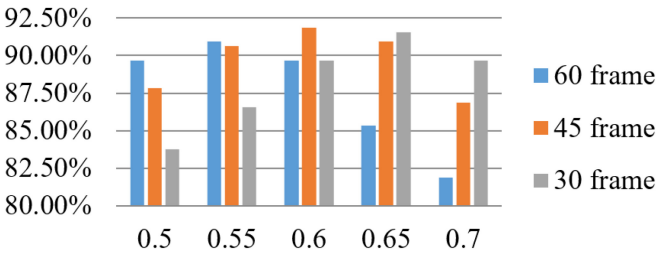


Fig. 8. Yawn detection results.

D. Concentration Analysis

This study divides concentration into two levels: normal and fatigue. First, it is assumed that a student's initial concentration is normal. After classifying facial information and determining motion categories, if the system detects that a subject is fatigued, his/her status is changed from a normal concentration level to a fatigue concentration level. Table III describes the fatigue situation.

V. EXPERIMENTAL VALIDATION AND RESULTS

A. Yawn Detection Evaluations

The proposed method can detect yawning behavior by the yawn feature extraction. The experiment shows that the proposed method detects yawn behavior accurately, and uses open data to validate the proposed method in a real world situation.

The YawDD open dataset [1] includes two kinds of video under different light environments. The main scenario is one of a person driving. The first dataset consists of footage captured by a camera placed under the rear-view mirror; every tester is tested by three different camera views: normal, talking and yawning. There are 320 films in total. The second dataset consists of footage captured by a camera placed on the dashboard, with a total of 29 films.

Yawn Detection Results: The yawn detection experiment determines if yawning behavior occurs in the footage, and is conducted using different time periods. The yawning film was tested by analyzing the distance and time thresholds. The best distance was 40 pixels; and there were 45 frames at different time thresholds. The 45 frames were each about 1.5 seconds long. The failure cases occurred when the mouth scale size was too small, or the yawning behavior was covered by a hand, as shown in Fig. 8.

B. Facial Orientation Evaluations

We use the **Pointing 04 dataset** from [7] for experiments on the training and testing of our facial orientation estimator network. The dataset includes 15 testers; every tester is recorded in two different facial orientation videos, in which the yaw angle is between −90 degrees and 90 degrees, and the pitch angle is between −90 degrees and 90 degrees. There are 2940 videos in total.

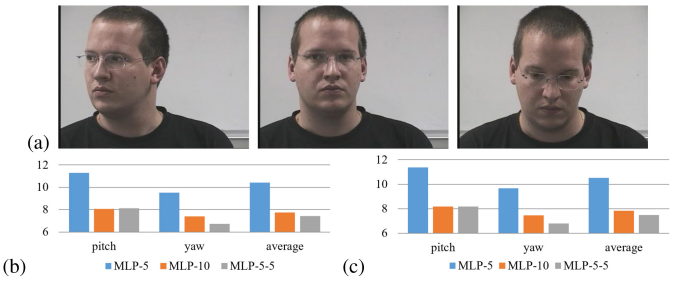
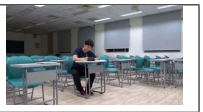
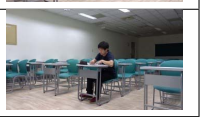


Fig. 9. (a) Facial orientations. (b) Training error of each angle type. (c) Testing error of each angle type.

TABLE IV
CLASSROOM VIEWS

Shot angle	View	Shot angle	View
Front-Rear		Left 45°	
Front-Far		Right 45°	
Overlook			

This study adopted 5 facial features (left eye corner, right eye corner, nose tip, left mouth corner, and right mouth corner), and resized the angle range as experiment data, such that the yaw angle was between −45 degrees and 45 degrees, and the pitch angle was between −30 degrees and 30 degrees. The angles were changed by 15 degree intervals, with a total of 1050 videos. 5-fold cross validation was adopted, and the model was trained by the average of the five results. The multi-layer neural networks had two-hidden layer architectures: 5,10 and 5,5. The best performance was achieved using the 5,5 architecture, but the error of the pitch was greater than that of the yaw. The features involved in looking up and looking down were not obvious; however, the error was less than 5-10 degrees, as shown in Figs. 9b,c.

C. Activity Detection and Recognition Evaluations

This study constructed 5 different scenarios involving two distances (far and near), two horizontal angles (−45 degrees and 45 degrees) and two vertical angles (looking up and down), as shown in Table IV.

The In-Class Student Activity Dataset (ICSAD): A total of ten subjects participated in the experiment, six males and four females, as shown in Table V. This experiment summarizes the eight common actions in the classroom context: writing notes, using a laptop, texting on a mobile phone, holding something

Authorized licensed use limited to: JIANGSU NORMAL UNIVERSITY. Downloaded on July 20,2023 at 17:05:25 UTC from IEEE Xplore. Restrictions apply.

TABLE V
CLASSROOM SUBJECTS

#	Gender	Height	#	Gender	Height
1	M	170	6	F	170
2	M	172	7	F	162
3	M	175	8	M	172
4	F	162	9	M	170
5	F	157	10	M	170

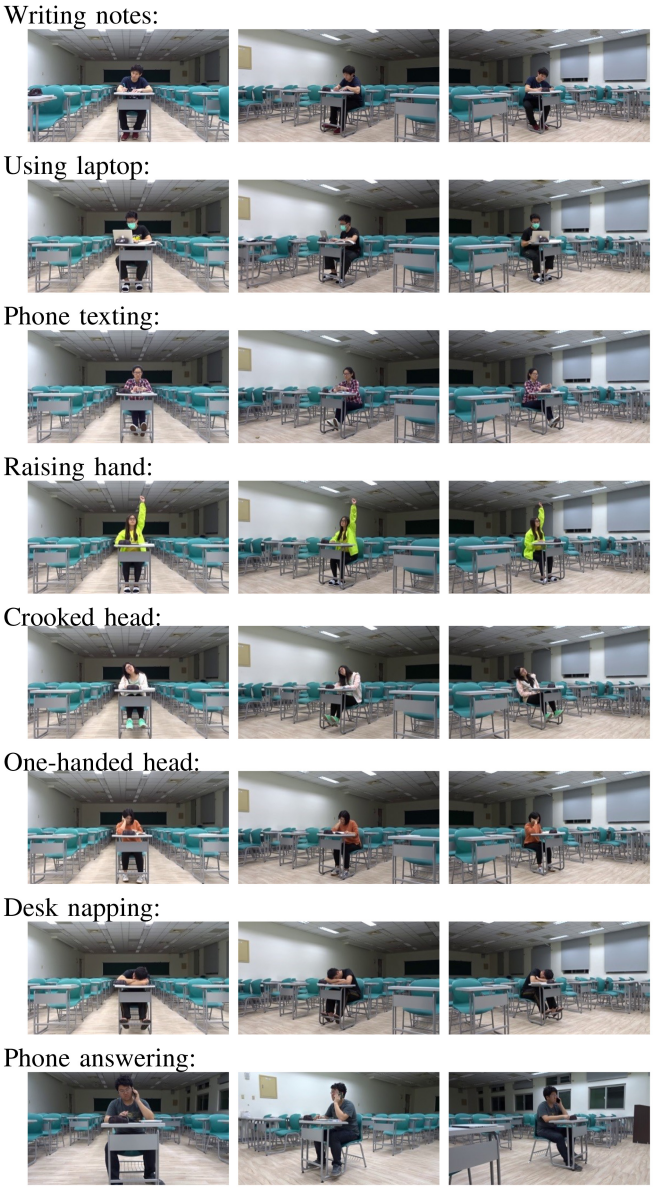


Fig. 10. Classroom activities.

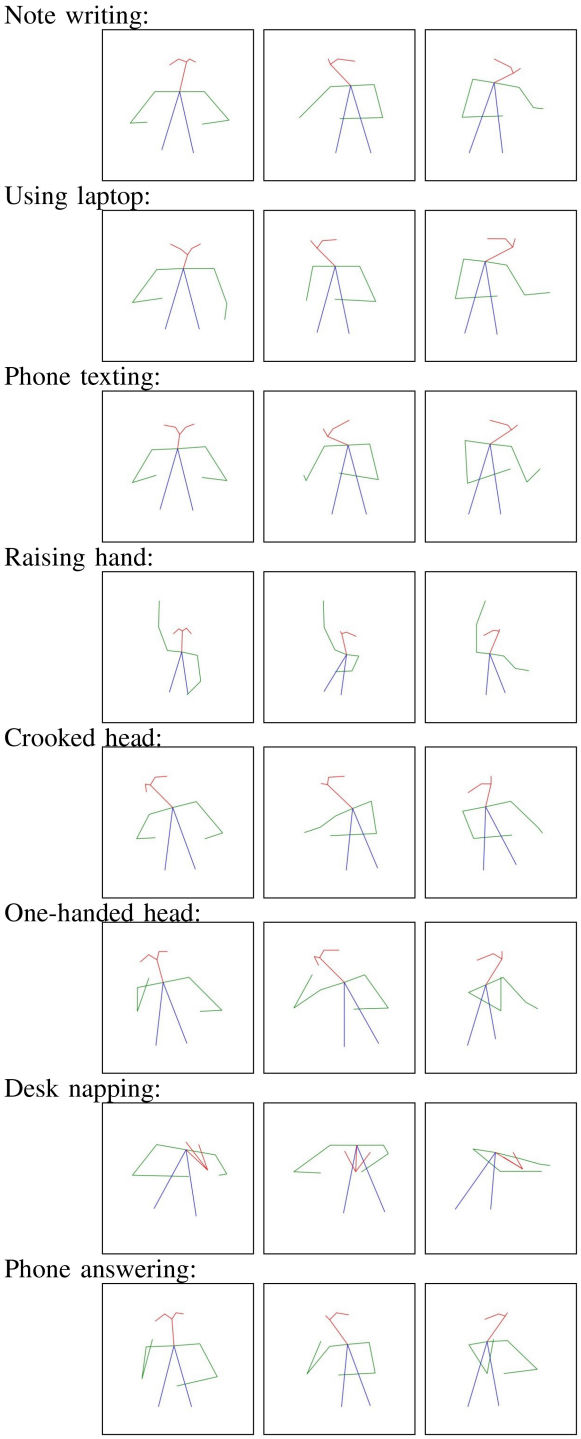


Fig. 11. Classroom activities - OpenPose human body skeleton detection results: first column: 0°, 2nd column: right 45°, 3rd column: left 45°.

in one hand, tilted head, supporting head with one hand, sleeping, and answering a phone. The method was recorded 5 times, and the actions are shown in Fig. 10. A total of 400 films were recorded for the experimental study.

Three camera viewing (0, −45, and 45 degrees) skeleton images were extracted using the Openpose method, and the skeleton images were normalized for input into the CNN in order to classify each motion. The kinds of skeleton images are shown in Fig. 11.

Since it is impossible to pre-record the action data for each student in the actual application and then use the neural network for training, the pre-trained data are particularly important in order to ensure that the pre-trained neural network model can be applied to different people in actual situations. This experiment therefore tested the features of the paper presented in different real world scenarios. The experiment also added the ability to promote the original image and skeleton features in different people.

Authorized licensed use limited to: JIANGSU NORMAL UNIVERSITY. Downloaded on July 20,2023 at 17:05:25 UTC from IEEE Xplore. Restrictions apply.

TABLE VI
CONFUSION MATRIX FOR CNN

		Prediction classification							
		Note writing	Using laptop	phone texting	Raising hand	Crooked head	Desk napping	One-handed head	Phone talking
Ground truth	Note writing	**47.35**	15.64	14.26	0.23	5.74	12.19	3.57	1.02
	Typing	17.08	**59.59**	12.89	0.86	6.45	0.00	0.44	2.69
	Phone texting	21.66	12.51	**51.98**	0.00	4.32	6.51	2.25	0.77
	Raising hand	0.03	1.21	0.00	**96.00**	1.20	0.04	0.28	1.24
	Crooked head	7.14	7.39	12.00	0.81	**56.71**	4.04	11.22	0.69
	Desk napping	2.16	0.62	0.50	0.79	1.95	**93.01**	0.53	0.44
	One-handed head	5.14	0.14	2.28	0.09	9.24	1.70	**65.30**	16.11
	Phone talking	1.49	4.63	2.21	0.06	1.01	0.00	15.68	**74.92**

In this experiment, the data of different people were not particularly mixed, so the subjects used in training and testing were different people, and the 5-fold cross-validation was divided into 5 equal parts for two people.

The feature comparison used the original image and the skeleton image. The framework using the original image was AlexNet, while the framework using the skeleton image was the proposed CNN. The two kinds of images yielded recognition rates of almost 100% in training, but in the test process, the recognition rate for the original image is less than 60%, and the recognition rate for the skeleton image was improved by about 10% over that of the original image. This result proved that the skeleton feature is more popular among different people, and is better than the original image, but achieves a less than 70% recognition rate, meaning the system is less stable. It is necessary to identify the problem of the action category from the test results in the confusion matrix.

The testing confusion matrix is the result of the proposed CNN shown in Table VI. Errors occur in identifying the writing notes, using a notebook, and texting on a smartphone behaviors because the skeleton images in these actions are all very similar: all involve using an object on the table. Confusion also occurs between resting one's head on one hand, and answering a phone, because the two skeleton are also very similar. Object detection is thus necessary to prevent such confusion and improve recognition accuracy.

D. Human-Object Interaction Evaluations

The above experiment results show the need for object recognition in order to improve recognition results, so the proposed object recognition method helps distinguish similar skeleton actions. First, sub-skeletons are defined as similar actions in the same category, and then object recognition is used to distinguish between them. This study used the YOLO object detection method, and the COCO data set [13] was used as training data. Three objects are defined: books, notebook computers and mobile phones.

The motion detection combined with the object detection method improved the recognition rate, as shown in Table VII. The accuracy rate increased by 10%, and the testing accuracy rate increased from 68% to 77%.

The performance of writing notes, using notebook and one hand supporting head actions improved, but remained poor for typing on a smartphone and one hand answering phone actions because the region of a smartphone is small, resulting in occlusion by the hand holding it. The identification accuracy for the leaning head motion did not improve after object recognition because everybody tilted their heads in different directions, as shown in Table VIII.

TABLE VII
ACCURACY RATE FOR DIFFERENT INDIVIDUALS

Motion recognition class	Testing accuracy
Skeleton feature	68.10%
Skeleton + object detection	77.89%

TABLE VIII
CONFUSION MATRIX OF SKELETON IMAGES COMBINED WITH OBJECT DETECTION

		Prediction classification							
		Note writing	Using laptop	phone texting	Raising hand	Crooked head	Desk napping	One-handed head	Phone talking
Ground truth	Note writing	**76.67**	9.85	0.00	0.06	4.73	5.82	2.74	0.12
	Typing	2.00	**90.25**	0.00	0.28	3.75	0.76	1.72	1.26
	Phone texting	20.85	4.98	**64.11**	0.12	2.11	4.34	2.10	1.40
	Raising hand	0.56	0.08	0.17	**90.84**	7.06	0.90	0.32	0.07
	Crooked head	27.62	0.83	4.59	1.91	**54.03**	6.33	4.12	0.57
	Desk napping	2.48	1.90	0.00	0.59	0.51	**93.13**	0.45	0.94
	One-handed head	3.78	0.36	0.52	0.32	9.45	3.21	**82.36**	0.00
	Phone talking	8.44	0.22	0.00	1.91	2.14	2.75	12.76	**71.78**

TABLE IX
ANGLE IMPROVEMENT RESULTS

Camera angle	Training accuracy	Testing accuracy
Front rear	94.43%	85.64%
Right 45°	94.04%	69.65%
Left 45°	94.49%	78.69%
Front-far	95.59%	81.77%
Front-overlook	93.91%	79.38%
Average	94.49%	79.03%

Angle Promotion Test: Due to different classroom environments, the camera setup may not be able to capture the student's movements from the front, and students are not guaranteed to face the camera. Therefore, this experiment tested whether the recognition could be successfully performed under different camera angles. When the data record was collected, every participant was recorded from 5 different angles: from the front (near, far, overlooking) and sides (left 45 degrees and right 45 degrees), respectively. This experiment and the data were mixed together, and then divided into 5 equal parts for 5-fold cross-validation based on the 5 angles.

Table IX shows the five angles and average individual test recognition rate. Respectively, each angle was used for the test data, and the remaining four angles were used as training data. The results showed that the right 45 degree and left 45 degree angles yielded poor recognition rates because side and front views of each skeleton image differ, and thus are difficult to classify correctly.

Sliding Window Test: This study proposed using the sliding window approach to increase identification stability. This experimental network architecture uses a CNN, and the test dataset recorded three simulation scenarios consisting of eight actions, with each action about 5 seconds in duration.

Authorized licensed use limited to: JIANGSU NORMAL UNIVERSITY. Downloaded on July 20,2023 at 17:05:25 UTC from IEEE Xplore. Restrictions apply.

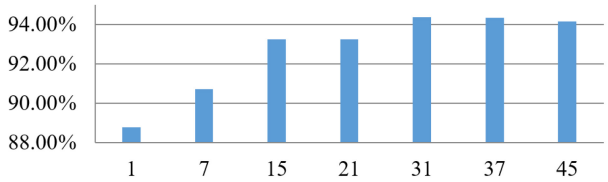


Fig. 12. Identification accuracy using of different sliding window sizes.

Fig. 13. Two views of real-world in-class test examples from our ICSAD dataset.

TABLE X
CONFUSION MATRIX OF REAL WORLD TEST SCENARIOS (UNIT: PERCENTAGE)

		Prediction classification							
		Note writing	Using laptop	phone texting	Raising hand	Crooked head	Desk napping	One-handed head	Phone talking
Ground truth	Note writing	**89.33**	2.22	0.00	0.00	0.00	8.44	0.00	0.00
	Typing	0.00	**100.00**	0.00	0.00	0.00	0.00	0.00	0.00
	Phone texting	13.33	0.00	**78.57**	0.00	8.10	0.00	0.00	0.00
	Raising hand	0.00	0.00	0.00	**97.12**	0.00	2.88	0.00	0.00
	Crooked head	3.32	0.00	0.00	0.00	**69.10**	25.57	0.00	0.00
	Desk napping	1.00	0.00	0.00	0.00	0.00	**99.00**	0.00	0.00
	One-handed head	0.00	0.00	0.00	0.00	0.00	0.00	**83.27**	16.73
	Phone talking	0.00	0.00	0.00	0.00	0.00	0.00	1.44	**98.56**

Since the eight actions were static motions, and did not change over a short time, this experiment window test size was relatively large in order to compare the performance of windows consisting of 7, 15, 21, 31, 37, 45 and 1 frames (without using a sliding window), as shown in Fig. 12. The lowest recognition rate was achieved when no sliding window was used, while the highest recognition rate was achieved with a window of 31 frames. Therefore, the 31 frame sliding window most accurately recognized the actions defined in this study; 31 frames are equivalent to 0.5 seconds.

Real-World Scenario Test: The above experiments are based on limited data for training and testing. However, in order to test the recognition performance of the proposed method in the actual context of the classification results, this study recorded a few videos in a real classroom, and identified actions defined in this paper, as shown in Fig. 13, where the red box indicates a student using a notebook, and the blue box indicates a student using one hand to support his head. In Fig. 13b, a red box indicates a student writing notes. The network architecture of this experiment is the proposed CNN, and the sliding window size is 31.

The confusion matrix in Table X shows that the average recognition rate is 89.37%. The tilted head action is confused with the sleeping with head on the table action because the tilted head action made the skeleton appear smaller. The typing on a smartphone action resulted in errors because the smartphone object was too small.

VI. CONCLUSION

This paper proposed a deep learning method for determining the depth of focus of student learning. The study analyzed the facial information and movement behavior of students, focusing on the degree of reference, in order to inform teachers of a student's current learning status. This information can be used to adjust teaching schedules and methods to improve students' concentration in the classroom.

Using facial information, this paper designed the detection of two kinds of fatigue behaviors: blinking and yawning, making it possible to recognize when students are in a state of mental discomfort. In addition, the proposed method estimated the facial orientation angle, thus identifying when students are distracted. This study also defined eight common movements made by students, four of which are movements in using items. In terms of using electronic devices, the proposed method is able to use the facial orientation data to identify when students are querying relevant materials in the classroom, or staring at their screens as a result of being distracted. General actions can be used to determine whether students actively participate in the classroom.

Three experiments were designed to test the reliability of the facial and skeleton features. The accuracy of identifying yawning behavior in the film was over 90%. The facial orientation estimation exhibited errors at 10 or less degrees. The identification operation at different angles and different human recognition rates achieved nearly 80% accuracy, and finally, combining the sliding window mechanism, the recognition rate for a real situation test was nearly 90%. The above experiment results prove that the proposed method can provide teachers with accurate student learning status information.

This study used facial information and action behavior as the basis for the analysis of concentration. Facial orientation estimation can be used to train a CNN to convert 2D facial feature points into corresponding 3D models. The more accurate estimate face toward the corner and the movement behavior, although the record of 8 kinds of action already contains most of the student movement; inevitably there will be actions other than these, and then we will need to determine the undefined actions.

REFERENCES

[1] S. Abtahi, M. Omidyeganeh, S. Shirmohammadi, and B. Hariri, "YawDD: Yawning detection dataset," in *Proc. ACM Multimedia Syst. Conf. (MMsys)*, 2014, pp. 24–28. [Online]. Available: http://dx.doi.org/10.21227/e1qm-hb90

Authorized licensed use limited to: JIANGSU NORMAL UNIVERSITY. Downloaded on July 20,2023 at 17:05:25 UTC from IEEE Xplore. Restrictions apply.

[2] N. A. Attia, L. Baig, Y. I. Marzouk, and A. Khan, "The potential effect of technology and distractions on undergraduate students' concentration," *Pakistan J. Med. Sci.*, vol. 33, no. 4, pp. 860–865, 2017.

[3] A. Bulat and G. Tzimiropoulos, "Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 3726–3734.

[4] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks)," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 1021–1030.

[5] Z. Cao, G. H. Martinez, T. Simon, S. Wei, and Y. A. 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.

[6] C.-M. Chen, J.-Y. Wang, and C.-M. Yu, "Assessing the attention levels of students by using a novel attention aware system based on brainwave signals," *Brit. J. Educ. Technol.*, vol. 48, no. 2, pp. 348–369, 2015.

[7] N. Gourier, D. Hall, and J. L. Crowley, "Estimating face orientation from robust detection of salient facial features," in *Proc. Int. Conf. Pattern Recognit. (ICPR) Int. Workshop Visual Observ. Deictic Gestures*, 2004, pp. 1–9.

[8] K. Hara and R. Chellappa, "Growing regression forests by classification: Applications to object pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 552–567.

[9] L. P. Hartman, "Technology and ethics: Privacy in the workplace," *Bus. Soc. Rev.*, vol. 106, no. 1, pp. 1–27, 2004.

[10] W. James, *The Principles of Psychology*. New York, NY, USA: Henry Holt, 1890.

[11] D. E. King, "Dlib-ml: A machine learning toolkit," *J. Mach. Learn. Res.*, vol. 10, pp. 1755–1758, Dec. 2009.

[12] C. Li, A. Pourtaherian, L. van Onzenoort, W. E. T. A. Ten, and P. H. N. de With., "Infant monitoring system for real-time and remote discomfort detection," *IEEE Trans. Consum. Electron.*, vol. 66, no. 4, pp. 336–345, Nov. 2020.

[13] T. Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 740–755.

[14] N.-H. Liu, C.-Y. Chiang, and H.-C. Chu, "Recognizing the degree of human attention using EEG signals from mobile sensors," *Sensors*, vol. 13, no. 8, pp. 10273–10286, 2013.

[15] R. O. Mbouna, S. G. Kong, and M.-G. Chun, "Visual analysis of eye state and head pose for driver alertness monitoring," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 3, pp. 1462–1469, Sep. 2013.

[16] L. Mothwa, J. Tapamo, and T. Mapati, "Conceptual model of the smart attendance monitoring system using computer vision," in *Proc. Int. Conf. Signal-Image Technol. Internet-Based Syst. (SITIS)*, 2018, pp. 229–234.

[17] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 4, pp. 607–626, Apr. 2009.

[18] B. N. Anh *et al.*, "A computer-vision based application for student behavior monitoring in classroom," *Appl. Sci.*, vol. 9, no. 22, p. 4729, 2019.

[19] M. Raca and P. Dillenbourg, "System for assessing classroom attention," in *Proc. Int. Conf. Learn. Anal. Knowl. (LAK)*, 2013, pp. 265–269.

[20] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2014, pp. 1–14.

[22] T. Soukupová and J. Cech, "Real-time eye blink detection using facial landmarks," in *Proc. 21st Comput. Vis. Winter Workshop*, 2016, pp. 1–8.

[23] S. Tulyakov and N. Sebe, "Regressing a 3D face shape from a single image," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 3748–3755.

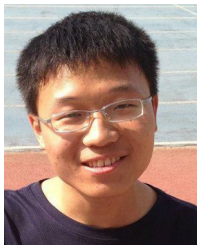
[24] Y. Wang *et al.*, "Continuous driver's gaze zone estimation using RGB-D camera," *Sensors*, vol. 19, no. 6, p. 1287, 2019.

[25] H. H. Wilmer, L. E. Sherman, and J. M. Chein, "Smartphones and cognition: A review of research exploring the links between mobile technology habits and cognitive functioning," *Front. Psychol.*, vol. 8, p. 605, Apr. 2017.

[26] J. Zaletelj and A. Kosir, "Predicting students' attention in the classroom from kinect facial and body features," *EURASIP J. Image Video Process.*, vol. 2017, p. 80, Dec. 2017.

[27] X. Zhen, Z. Wang, M. Yu, and S. Li, "Supervised descriptor learning for multi-output regression," in *Proc. IEEE CVPR*, 2015, pp. 1211–1218.

Mu-Chun Su (Senior Member, IEEE) received the B.S. degree in electronics engineering from National Chiao Tung University, Taiwan, in 1986, and the M.S. and Ph.D. degrees in electrical engineering from the University of Maryland, College Park, MD, USA, in 1990 and 1993, respectively. He is currently a Professor of Computer Science and Information Engineering with National Central University, Taiwan. He has authored more than 100 journal and refereed conference papers. His current research interests include computational intelligence, neural networks, fuzzy systems, swarm intelligence, affective computing, human–computer interaction, robotics, pattern recognition, biomedical signal processing, image processing, and rehabilitation technology. He was the IEEE Franklin V. Taylor Award recipient in 1991. He serves as an associate editor for many journals. He is a Senior Member of the IEEE Computational Intelligence Society and Systems, Man, and Cybernetics Society. He is also an IET Fellow.



Chun-Ting Cheng (Member, IEEE) received the M.S. degree in computer science and information engineering from National Central University, Taiwan, in 2016. His research interests include machine learning and deep learning.

Ming-Ching Chang (Senior Member, IEEE) received the B.S. degree in civil engineering and the M.S. degree in computer science and information engineering from National Taiwan University in 1996 and 1998, respectively, and the Ph.D. degree from the Laboratory for Engineering Man/Machine Systems, School of Engineering, Brown University in 2008. He is an Assistant Professor with the Department of Computer Science, College of Engineering and Applied Sciences, University at Albany, State University of New York. He was with the Department of Electrical and Computer Engineering from 2016 to 2018. He was an Adjunct Professor with the Computer Science Department from 2012 to 2016. From 2008 to 2016, he was a Computer Scientist with GE Global Research Center. He was an Assistant Researcher with the Mechanical Industry Research Labs, Industrial Technology Research Institute, Taiwan, from 1996 to 1998. His research projects are funded by GE Global Research, IARPA, DARPA, NIJ, VA, and UAlbany. He has authored more than 85 peer-reviewed journal and conference publications, seven U.S. patents and 15 disclosures. His expertise includes video analytics, computer vision, image processing, and artificial intelligence. He is the recipient of the IEEE Advanced Video and Signal-based Surveillance (AVSS) 2011 Best Paper Award—Runner-Up, the IEEE Workshop on the Applications of Computer Vision 2012 Best Student Paper Award, the GE Belief—Stay Lean and Go Fast Management Award in 2015, and the IEEE Smart World NVIDIA AI City Challenge 2017 Honorary Mention Award. He serves as the Co-Chair of the annual AI City Challenge CVPR 2018–2021 Workshop, the Co-Chair of the IEEE Lower Power Computer Vision Annual Contest and Workshop 2019–2021, the Program Chair of the IEEE AVSS 2019, the Co-Chair of the IWT4S 2017–2019, the Area Chair of IEEE ICIP (2017 and 2019–2021) and ICME (2021).

Yi-Zeng Hsieh (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in computer science and information engineering from National Central University, Taoyuan, Taiwan, in 2004, 2006, and 2012, respectively. He is currently an Associate Professor of Electrical Engineering with National Taiwan Ocean University. His current research interests include neural networks, pattern recognition, image processing, machine learning, and deep learning.

Authorized licensed use limited to: JIANGSU NORMAL UNIVERSITY. Downloaded on July 20,2023 at 17:05:25 UTC from IEEE Xplore. Restrictions apply.