Power System Technology
Volume 49, No. 1, 2025
ISSN: 1000-3673

Advanced Behavioural Analysis for Classroom Instruction: A Computer Vision Approach

Liam Wang, Xander Wu

[1]College of Foreign Languages, Northeast Forestry University, Harbin 150040, China

[2]College of Electrical and Mechanical Engineering, Northeast Forestry University, Harbin 150040, China

Abstract

Teachers primarily engage with students through classroom observation and questioning in traditional teaching methods, which will inevitably lead to one-sidedness and a lagging of information transmission and feedback because of things like a lack of personal energy. In contrast, artificial intelligence brings computer vision recognition into the classroom, which unquestionably increases teaching efficiency. The final teaching effect of the close relationship between the effective teaching strategies of the teachers and the good classroom behavior of the students can be enhanced by the positive classroom behavior of the students. Through the use of deep learning-based target recognition techniques, classroom camera visual analysis may be utilized to gather information on the teaching behaviors of students. This includes the identification and analysis of voice, posture, facial, physiological signal, and other data, to extract and define the distinctive behaviors of students. Lastly, the study demonstrates how computer vision applied to the teaching classroom can precisely identify and analyze students' emotional states and learning status, enabling teachers to adapt their instruction to the students' emotional states and learning status. In order to enhance teaching effectiveness and student progress, it assists educators in refining and optimizing their pedagogical approaches in accordance with the students' current circumstances. Consequently, there are numerous benefits to using computer vision recognition in the classroom.

Keywords: Visual recognition, classroom teaching, teaching methods, data collection.

1. Introduction

The application of visual behavior has many applications due to the advancements in science and technology; cameras are necessary in all spheres of life. Teaching visual behavior in the classroom offers a wealth of application opportunities. Teacher lectures are typically structured into three time domain stages: first, a new concept is introduced and briefly explained; second, students are divided into small groups and a representative is sent to present the group's conclusions; and third, the teacher summarizes based on the student's speech. Artificial intelligence can help with the effective development of intelligent evaluation by focusing on the dynamic changes in students' emotional information, restoring the accuracy of real classroom data collection, capturing the behaviors of both teachers and students in the classroom, quantifying the teaching and listening behaviors of teachers and students using deep learning algorithms, creating classroom behavior curves and a list of the best and worst teaching solutions, and providing feedback to the teacher as the instructor. The teachers use the results as a starting point to optimize the lesson plan, which is then confirmed by the teachers' team's optimized lesson plan examples and student and teacher feedback. The findings demonstrate that, while each

instructor has optimized their lesson plans to a different extent, overall classroom focus and instructional effectiveness have increased.

2. Classroom Behaviour and Identification

Classroom behavior, sometimes referred to as classroom engagement, is the physiological and psychological state in which students are involved in learning and academic-related activities in the classroom. It encompasses all visible and audible behaviors, identifies the objects, scenes, and activities involved, and analyzes them based on the various facial expressions that students display (e.g., fidgeting, listening attentively to lectures, chatting, sleeping, playing with mobile phones, etc.). This technology can create a more individualized and intelligent learning environment for kids by assisting teachers in better understanding their learning status and requirements.

In order to foster a supportive learning environment and enable efficient teaching and learning, effective classroom behaviors are essential. The way teachers teach and how attentive students are in the classroom have the biggest direct effects on how successful teaching and learning are, even if there are many other elements that also play a role. To encourage positive classroom behavior, teachers frequently employ a range of tactics and approaches, including setting clear expectations, offering constructive criticism, modeling acceptable behavior, and providing positive reinforcement.

Shao et al. [1] attempted to improve the way that "Information Technology" was taught in Natural Science classes. She compiled a list of four teachers' classroom behavior curves and sorted the best and worst lesson plans for the class. The four teachers have fully affirmed the value of quantitative analysis for teaching reform. In the practice of teaching reform, the teachers optimized their lesson plans based on the sequence of the best lesson plan, the sequence of the worst lesson plan, the classroom behavior curves, and the feedback from the students on the lesson plans. The optimized lesson plans improved the overall concentration in the classroom to varying degrees.

In an effort to reform the way energy technology courses are taught, Shao et al. [2] summarized the concentration curves of the students in the classrooms of six teachers. The teaching process was repeated numerous times, resulting in the iteration of both the optimal characteristic curves and the sequence of optimal teaching styles through the collaborative efforts of all the teachers. When it comes to teaching reform, the optimal characteristic curve's guidance can help teachers' groups try fewer different teaching methods with less effort. If teachers' groups follow the optimal teaching methods' sequence, students' classroom concentration curves may even exceed the optimal characteristic curve, providing the best possible effect on students' concentration in the classroom.

Incorporating computer vision into educational software can give teachers real-time access to student performance data and enable them to modify their tactics as needed to ensure that all students receive adequate instruction. Additionally, students will be able to view their performance in class following the lesson. AI-based classroom evaluation can enhance feedback between teachers and students while expediting the evaluation process when compared to conventional feedback techniques like exams, questionnaires, or interviews. This approach successfully fosters the improvement of classroom instruction evaluation, increasing its accuracy and efficiency. Table 1 lists some common classroom behaviors from children along with activities that serve as representations.

Table 1 Typical student classroom behaviours and representational actions.

Student Behaviour in the Classroom	expressive action
listen in class	Sit upright with your eyes flat in front of you
read books	Small bow of the head, flipping through the textbook
write characters	Small bow of the head, writing with the pen in the hand
whisper to each other's ear	Face deflected, lips slightly moving to talk to people around them
play with mobile phones	Lower your head considerably and place your hands on the table or under the table
asleep	Head down, support with one hand or hands on desktop

be stunned	Dull eyes, staring at something for a long time
talk over	Students face each other with slight or no facial smiles
chats	The student's eyes are averted and he is always aware of the teacher's position.
raise a hand	Sit up straight with one arm extended upwards

The system architecture of classroom teaching evaluation under AI is designed to include the object layer, data layer, technology layer, and application layer from top to bottom[3], as shown in Figure 1, based on the current state of AI technology development and the changing needs of classroom teaching evaluation.

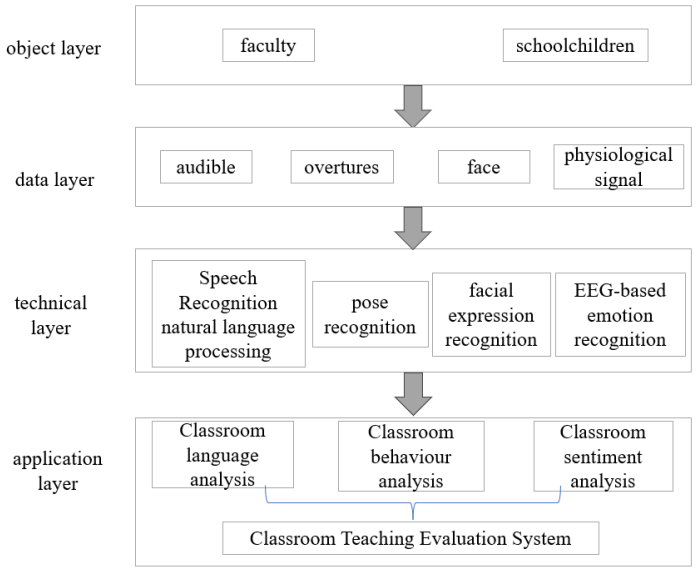


Figure 1 Classroom teaching evaluation system architecture under artificial intelligence

Classroom teaching evaluation under artificial intelligence focuses on two types of evaluation objects: teachers and students. Through the collection of their voices, postures, facial and physiological signals, recognition and analysis algorithms are designed to achieve the purpose of improving the efficiency of teaching. The two main components of classroom teaching activities are teachers' teaching and students' learning.

3. Data Analysis Based on Facial Recognition

Expressions are the feelings that come to light during the communication and expression process. The goal of facial expression recognition is to record and interpret a person's expressions on their face so that a machine can decipher the intentions and signals they convey. Facial expression recognition processing can be divided into two basic categories. The first is local feature recognition. Identifying regional characteristics is the first step. face muscle movements can be recognized and utilized to categorize face expressions by detecting and analyzing important features such eyebrows, eyes, noses, lips, etc. The identification of general traits comes in second. The feature face approach and the elastic matching method are two frequently used techniques for recognition based on the general facial qualities [4]. These techniques help identify and analyze the face as a whole as well as distinguish the facial features of distinct expressions. Classifier algorithms, databases of frequently used facial expressions, and techniques for extracting facial features are closely related to facial expression recognition. Four general categories can be used to group facial feature extraction methods: those based on edge information (like linear edge maps); those based on texture information (like wavelet transform); those based on both global and local information (like principal component analysis); and those based on geometrical information (like local curve wavelet transform). There are two types of facial expression classifiers: deep learning models and conventional machine learning techniques. Despite the fact that facial expression recognition research and applications have been conducted extensively outside the nation in recent years, there are only a few publicly accessible expression databases, and the definitions of these databases vary.

3.1 Facial recognition based algorithms

3.1.1 Feature face algorithm

An ideal orthogonal transform for picture compression is the Karhunen-Loeve (KL) transform. KL transforms the high-dimensional picture space to produce a new set of orthogonal bases while keeping the significant orthogonal bases that are needed to build the low-dimensional linear space. The fundamental principle behind the feature face technique is that the projections of the face in these low-dimensional linear spaces can be employed as feature vectors for recognition if it is considered that these projections are differentiable. Let the face image be represented as a n*1 dimensional vector X in face recognition, where n is the product of the image width and image height. Assume that there are N training samples, and use the training sample set's overall scatter matrix as the generating matrix, that is:

$$C = E[XX^T] \approx \frac{1}{N}\sum_{k=1}^{N} X_k X_k^T \tag{1}$$

Using an n*N matrix to represent the N face vectors, $X = [X_1, X_2, X_3, \cdots X_N]$, then C can be expressed as:

$$C \approx \frac{1}{N} XX^T \tag{2}$$

The generating matrix for the KL transformation was the interclass scatter matrix of the training sample set, that is to say:

$$S = \sum_{i=0}^{p-1} p(\omega_i)(m_i - m)(m_i - m)^T \tag{3}$$

where m_i is the average image vector of the i th person in the training sample set, and p is the total number of people in the training sample set. Obviously, compared to the overall scatter matrix C, the number of face images used to generate features is reduced from N to p without any effect on the recognition rate [5].

The orthogonal bases of all the subspaces can be placed in an image array in such a way that they resemble a human face. For this reason, these orthogonal bases are sometimes referred to as feature faces, which is how this recognition method gets its name. The selection of orthogonal bases to build the subspace involves several factors. It is also understood that the low-frequency components are represented by principal components, and the high-frequency components are represented by subcomponents. Some orthogonal bases corresponding to larger eigenvalues (also known as principal components) are able to express the general shape of the face, while the details need to be complemented by eigenvectors corresponding to smaller eigenvalues (also known as subcomponents). Principal Component Analysis (PCA) is the process of creating a new orthogonal space by employing one principal component as its orthogonal basis. Since faces generally have similar shapes and structures, some individuals also employ sub-components as the orthogonal basis. This is because the high-frequency components that those sub-components convey are what actually separate distinct faces from one another.

After obtaining a series of eigenfaces, for the face to be recognised, it is projected to a new dimensional face space, and a projection vector is obtained, which represents the face to be recognised, and then the face recognition problem is transformed into a classification problem of coordinate system vectors in m -dimensional space. Usually, the eigenvectors corresponding to the largest m eigenvalues are chosen as orthogonal bases, and the following approximate expressions are obtained:

$$\hat{X} = \sum_{i=1}^{m} k_i \mu_i \tag{4}$$

In the specific calculation of feature vectors, assuming that the width and height of the image are 128 pixels, then the dimension of C is $128^2 * 128^2$, and presumably $2.7 * 10^8$, it is difficult to perform the calculation,

for this reason, the singular value decomposition [6] algorithm can be used to reduce the amount of computation.

3.1.2 Elastic matching algorithm

The elastic matching method uses an attribute topology graph to represent the face (Figure 2 uses a regular 2D mesh, but other mesh shapes can be used in practice). Any vertex of the topology graph contains a feature vector to record the information of the face in the vicinity of the vertex position, as shown in Figure 3. This method defines a distance in 2D space that is invariant to the usual face deformations.

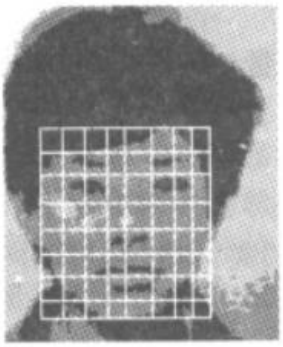


Figure 2 Two-bit topology localised on the human face

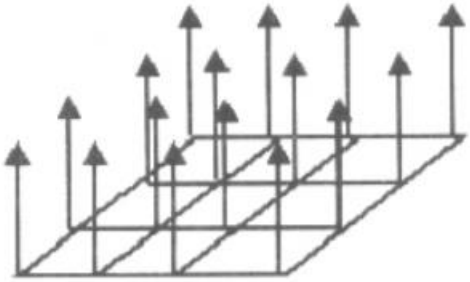


Figure 3 2D vector field expressing face features

Assuming that S_1 is a 2D grid of face templates defined on a face image from a known library, the information near node i on this face image grid can be represented by a feature vector C_i . C_i can be chosen in several ways, the most common being Gabor features: define the members of $G = [g_1, g_2, g_3, \cdots g_m]^T$ to be 2D Gabor filters with different centre frequencies, different bandwidths and different orientations, then C_i is the value at node i of the convolution of G and the face image. Similarly, a vector field on a two-dimensional grid is defined on the face image to be recognised. Where X_i is a feature vector of the same type as C_i , except that it is defined on a larger and denser two-dimensional grid S .In elastic matching, the matching between the faces in the library and the faces to be recognised is transformed into the matching between S_1 and S , that is, finding the best matching node in S for each node in S_1 . The optimal matching should take into account both feature matching and local geometric position matching [7].

3.2 Facial feature based detection method

One key way to judge the quality of teaching is to look at how engaged and active the students are in the classroom. You can measure this by looking at the way students move their heads and make facial expressions, among other nonverbal cues. Finding knowledge on the essential elements of head postures and expressions is

necessary in order to capture pupils' postures and expressions. By pre-training the model with the official DLIB feature extractor, the important features of the face are acquired [8]. The DLIB C++ library comprises machine learning techniques and tools. It generates a model by using a training set of 68 keypoint-labeled face photos. The model is then used to estimate the feature point locations for newly obtained images. To increase the accuracy of the keypoints that are recognized, it is required to first normalize each frame of the intercepted image due to the changeable conditions of the classroom, such as light occlusion [9]. The following is the key point acquisition procedure: Every video frame has face detection; every face that is intercepted has key point labeling; and every key point coordinate is sent to a CSV file. Figure 4 displays the key point distribution of the 68 faces in the DLIB as well as the impact of facial profile recognition.

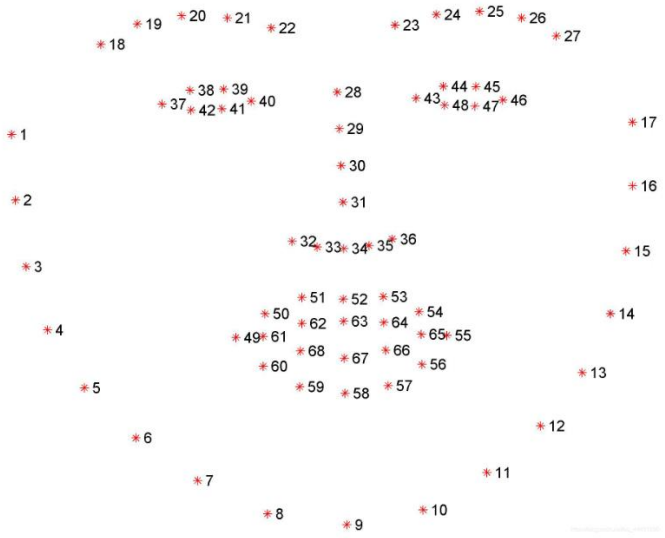


Figure 4 Distribution of key points of 68 faces in DLIB, effect of face contour recognition

The face image is represented by Harr-like features in the AdaBoost algorithm, which is based on statistical theory. An "integral map" is introduced to represent the image, which shortens the time required to calculate the feature values [10]. The teacher and kids in the classroom video are not fixed, and their faces are not always turned toward the camera. These features make the face detection method—which makes use of OpenCV's own classifiers—unique. Consequently, this study's methodology is to select the side face detection classifier rather than the front face classifier based on these features, with the intention of employing the classifier to develop a face detection program [11].

By way of comparison, it is discovered that the majority of the images depicting the behavioral state of the students are more complicated and have more contour lines, whereas the images depicting the behavior of the teacher are simpler and have clearer outlines. As a result, one of the variables in determining the final subject behavioral feature can be the contour detection results. By removing the smaller contours, the quantity and size of qualifying contours are determined and recorded for this study.

(1) Overview of Contour Detection The most crucial information that is gathered when target detection is applied to a picture is the image contour. In order to speculate on the properties of the subject target object in the image, contour detection essentially involves hollowing out the inside pixels of the image while maintaining the closed contour shape process. This is done by sequentially traversing the closed contour of the contour points, access to coordinate points, contour area, the number of contours, and other data.

(2) Overview of the contour detection technique Extracting the contour of a given image, removing small areas from the contour, obtaining the contour with a reference value, and storing the pertinent data are the objectives of contour detection[12]. Figure 5 depicts the fundamental stages of image contour detection.

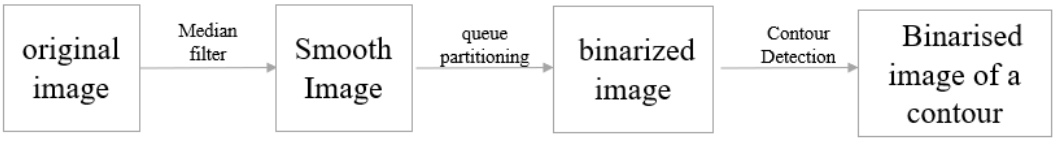


Figure 5 Basic steps of image contour detection

4. Data Analysis Based on Action Behaviour

4.1 Algorithms for human movement recognition

When an image of a person sleeping is provided to the Open Pose algorithm model, it can recognize the person's sleeping position by calculating the coordinates of their joint points and the skeleton diagram of the human body.

For detection and localization, Open Pose offers 18 or 25 body joints, 70 facial joints, and 21 hand joints. The primary distinction between 18 and 25 human joints is that more joints from the left and right foot are identified in the case of 25 joints [13]. This paper selects 18 human joint point data as the input of the sleeping posture identification algorithm because the joint point information of the foot in the sleeping posture recognition is not particularly distinct for the feature extraction of the sleeping posture [14]. Furthermore, since hand and face joint detection significantly affects the image's computing rate and hand and face joint information is not a significant aspect of sleeping posture, enabling it will significantly lower the computing rate, making it challenging to recognize a person's sleeping posture in real time. In conclusion, the sleeping posture identification algorithm uses only 18 human joints as inputs, which can increase computation rate while maintaining sleeping posture recognition accuracy. OpenPose outputs the skeleton of the human body with 18 joints. According to Figure 6

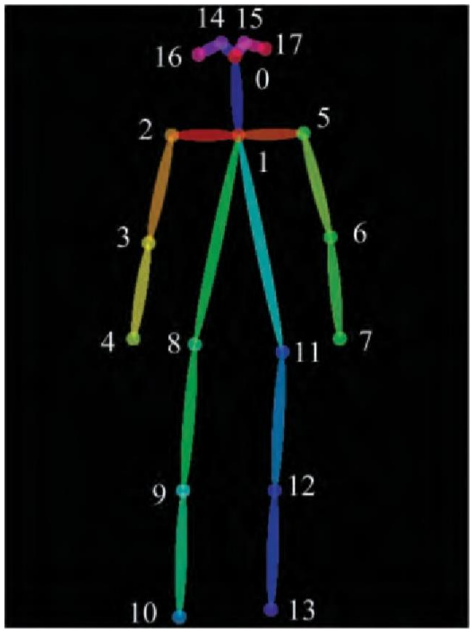


Figure 6 Open Pose output skeleton map (COCO dataset)

In order to offer a gesture recognition algorithm based on the spatial distribution features of the gesture, the gesture recognition algorithm must first achieve gesture recognition in the complicated background based on the regional shape characteristics of the gesture image.

The human hand is an articulated complicated malformation made up of five neighboring fingers, each of which has a joint and a finger segment, as well as the palm. The gesture is a jointed structure as a whole, and the hand's

shape varies as the joints move. The state-space modifications of the finger segments and joints can be used to characterize the various hand gestures.

The spatial distribution of hand gesture pixels can be used to describe the various hand motion shapes. As illustrated in Figure 7, the segmented binary image of gestures consists of background pixel points (white pixel points) and gesture skin color pixel points that interleave to form a variety of gestures. Therefore, for different images of the same gesture, whose skin color spatial distribution information is similar, the spatial regional distribution information of skin color pixel points is crucial, and the density distribution feature of gestures can be extracted [15], i.e., the distribution of the skin color pixel points of gestures in various spatial regions as one of the foundation for gesture recognition.

Given the rotation, translation, and scaling invariance characteristics of the gesture, the center of gravity is extracted using the moment depictor [16]. Using the center of gravity as the circle center, the region of the largest outer circle of the current gesture is divided into equidistant regions from the inside to the outside. The density distribution feature vectors are then obtained by counting the relative density of the target pixel points in each sub-image region.

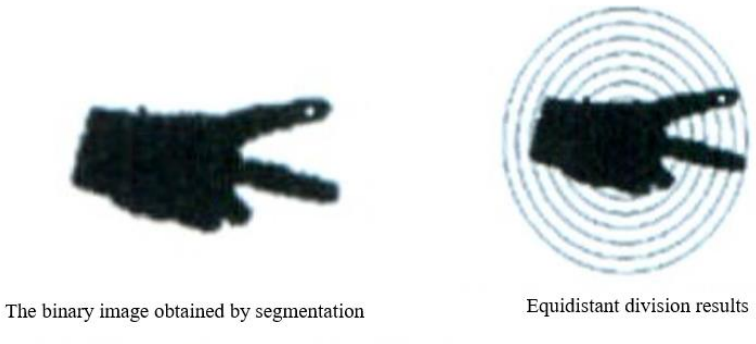


Figure 7 Gesture area and equidistant division

4.2 Detection methods for human movement characteristics

Student target testing is often based on examinations of the face, head, and skeleton of the body.

4.2.1 Differential method

The inter-frame difference method and the backdrop difference approach are examples of difference methods. In the former, the pixel values in the target region are dramatically changed when there is a moving target because the pixel widths of successive neighboring frames are compared using a difference operation. Inter-frame differencing is easier to compute and performs better in real-time; however, it requires manual selection of the appropriate differencing interval for moving targets with varying speeds because the pixel differences are primarily on both sides of the target motion direction, resulting in the emergence of internal voids. In order to acquire the moving target using the background difference approach, an adaptive background model must be established beforehand. This is done by comparing the pixel values of the detected frame image with the background image. The technique is straightforward and real-time, but it is susceptible to outside interference from things like shifting light, shifting shadows, shifting environmental conditions, etc. It works well in scenes where the background is static or barely changes.

4.2.2 Optical flow method

The inter-frame difference method and the backdrop difference approach are examples of difference methods. In the former, the pixel values in the target region are dramatically changed when there is a moving target because the pixel widths of successive neighboring frames are compared using a difference operation. Inter-frame differencing is easier to compute and performs better in real-time; however, it requires manual selection of the appropriate differencing interval for moving targets with varying speeds because the pixel differences are primarily on both sides of the target motion direction, resulting in the emergence of internal voids. In order to acquire the moving target using the background difference approach, an adaptive background model must be

established beforehand. This is done by comparing the pixel values of the detected frame image with the background image. The technique is straightforward and real-time, but it is susceptible to outside interference from things like shifting light, shifting shadows, shifting environmental conditions, etc. It works well in scenes where the background is static or barely changes.

4.2.3 Viola-Jones algorithm

To achieve adaptive enhanced learning, the Viola-Jones algorithm combines the Ada boost algorithm, the Cascade cascade structure, and Haar-like features. After a few training iterations, multiple weak classifiers are obtained, which are then integrated to form a strong classifier that meets the expected minimum error rate. In each round of iteration, the samples that are incorrectly realized are given higher weights, and the samples that are correctly realized are given lower weights. A new weak classifier is added at the same time. The Viola-Jones method is capable of classifying face images and is frequently utilized in face recognition. Nevertheless, side face detection is not achievable, and in settings with bright backgrounds, the detection performance is often good.

5. Deep Learning Based Target Detection Methods

5.1 SSD algorithm

Wei Liu proposed the object detection algorithm known as SSD (Single Shot Multi-Box Detector) in 2016. The method can identify several items in a single image in real time and is based on a deep convolutional neural network. The SSD algorithm's primary benefit over more conventional object identification techniques, like rapid R-CNN, is that it performs region suggestion and classification using a single network, which leads to notable speed gains. The SSD algorithm predicts whether items with varying dimensions and aspect ratios will be present in each of the grid of fixed-size boxes that it creates from the input image. In addition, the method forecasts each box's offset value to increase object localization precision. A set of bounding boxes and the associated class probabilities of the discovered objects are the algorithm's final outputs. Many computer vision applications, including robots, autonomous driving, and security surveillance, have embraced the SSD algorithm. Additionally, it has been expanded and enhanced in later studies, such as SSD with MobileNet architecture, which enhances the algorithm's accuracy and speed even more [17].

5.2 YOLO algorithm

Joseph Redmon et al. proposed the YOLO (You Only Look Once) algorithm for object detection in 2016.The primary benefit of the Yolo method is its high accuracy real-time object detection of many items in one image. The input image is divided into cells, and the bounding box and category probabilities are predicted for each cell. The end result of the technique, which makes use of a deep convolutional neural network for prediction, is a set of bounding boxes and the associated category probabilities for the items that were spotted. The YOLO approach employs a single neural network for both object localization and classification, in contrast to existing object detection algorithms that rely on area suggestion techniques like R-CNN. Because it only needs to process each image once, the YOLO technique is much faster than other traditional object detection algorithms. Numerous computer vision applications, including robots, surveillance, and autonomous driving, have made extensive use of the YOLO technique. Subsequent research has further refined and expanded it, leading to the development of the YOLOv3 algorithm, which enhances the method's accuracy and resilience.

Classroom state was analyzed using the YOLOX deep learning network by Wang et al. [18] Equation (2) represents the deep learning network's loss function, which is utilized to analyze students' classroom status recognition, whereas Equation (1) can yield the confidence score.

$$C_{\mathrm{j}}^{i} = P_{i,j} \times IOU_{pred}^{truth} \tag{5}$$

$$Loss = \mathrm{loss}_{\mathrm{Reg}} + loss_{Obj} + loss_{Cls} \tag{6}$$

The loss function is divided into 3 parts: bounding box regression loss $loss_{Reg}$, confidence loss $loss_{Obj}$ and category loss $loss_{Cls}$. During the training process, the loss function decreases and converges to a minimum value. During this process, the loss function decreases and converges to the mean value. The deep training network for the classroom state recognition method is deemed trained and the exercise is finished if the loss function either does not drop or decreases below a predetermined threshold multiple times. Using a YOLOX-based deep learning network to analyze the classroom allows teachers to rapidly and precisely assess students' learning situations, make timely adjustments and improvements to their techniques, and enhance both the quality and efficiency of their instruction.

5.3 R-CNN algorithm

In 2014, Ross Girshick et al. introduced the region-based convolutional neural network, or R-CNN, as an object detection system. It is the first algorithm to apply deep learning to object detection and makes use of deep learning techniques. On a number of popular datasets, it has demonstrated state-of-the-art performance. Table 2 displays the R-CNN architecture. It operates by segmenting the input image into regions of interest (RoIs) and employing a convolutional neural network that has been trained to extract information from each RoI. To categorize whether objects are present in each region of interest or not, the characteristics from each region of interest are then fed into a collection of class-specific linear SVMs. Lastly, the bounding box of each ROI is predicted using a regression model. The R-CNN algorithm's primary benefit over more conventional object detection techniques (like HOG and SIFT) is its ability to extract more accurate features from the data and achieve higher accuracy. The R-CNN algorithm does, however, have many drawbacks, such as its large memory use and sluggish processing speed. Since R-CNN's debut, a number of other object detection methods, including Fast R-CNN, Faster R-CNN, and Masked R-CNN, have been developed and built upon its foundation. By improving various components of the system, these methods enhance the speed and accuracy of the original R-CNN algorithm [19].

Table 2 R-CNN framework.

Region proposal(Selective Search)	
Feature extraction(CNN)	
Classification(SVM)	Bounding-box regression (regression)

6. Evaluating Effectiveness Based on Facial Expressions and Behavioural Traits

A comprehensive and methodical technique for evaluating the effectiveness of instruction was developed by integrating traditional cognitive behavior with the head position and facial expression patterns of the students. The system assesses the overall teaching impact of the classroom as well as the individual students.

6.1 An study of each student's head posture

For head state tracking, the head's three-dimensional angle, α, is introduced. A single student's head posture changes are monitored at time $t < 2$ min, as illustrated in Figure 8.

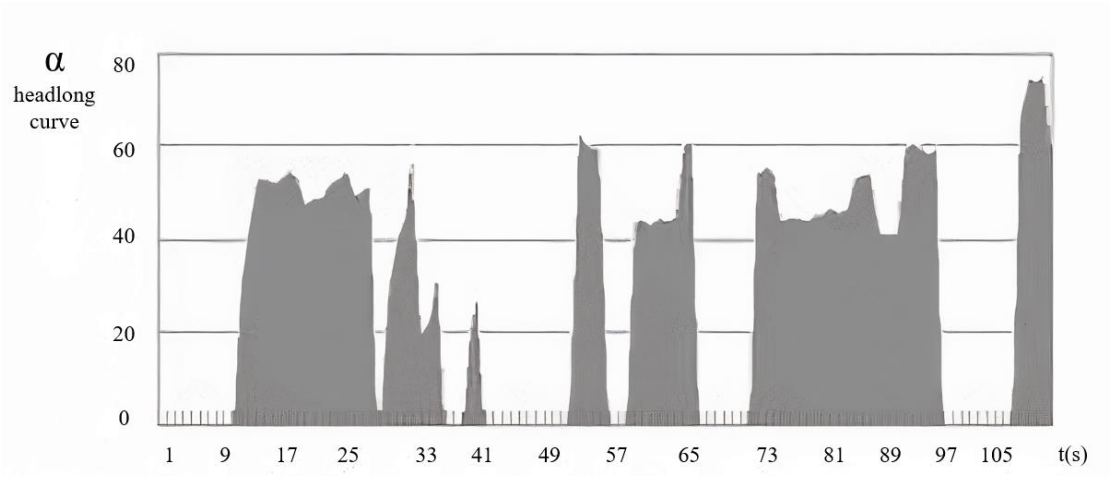


Figure 8 Dynamic tracking plot of individual student's head posture

6.2 Analyzing each student's facial expression individually

In this study, we present the angle β between the ends of the eyebrows and the center point of the eyes for the analysis of the brow: β <= 120° indicates the state of the brow stretching, and β > 120° indicates the state of the brow frowning (Figure 9). The characteristic line of the mouth corner's positive and negative deviations from the center line can be used to determine the lip analysis (Figure 10).

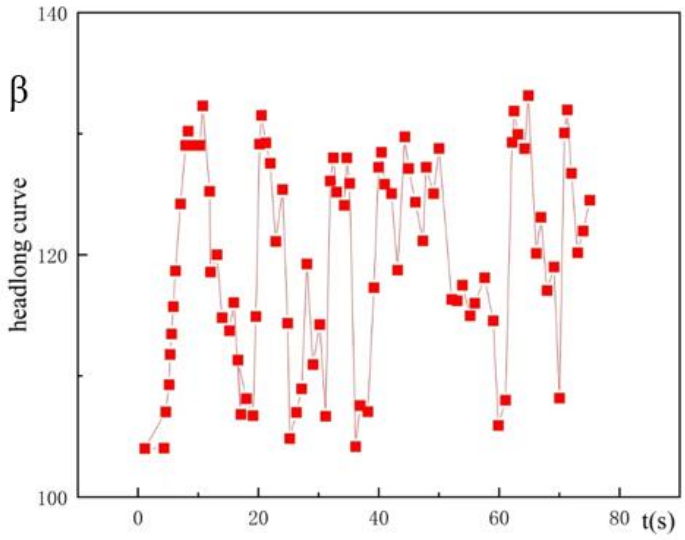


Figure 9 Individual student browbeat status map

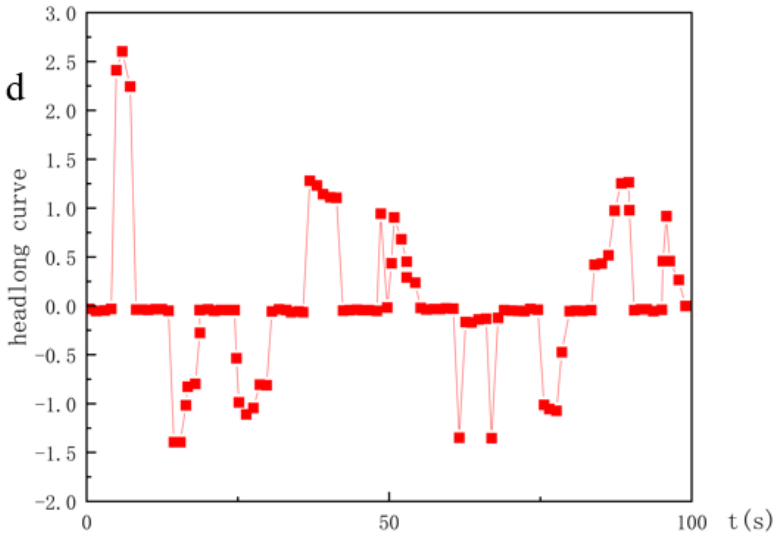
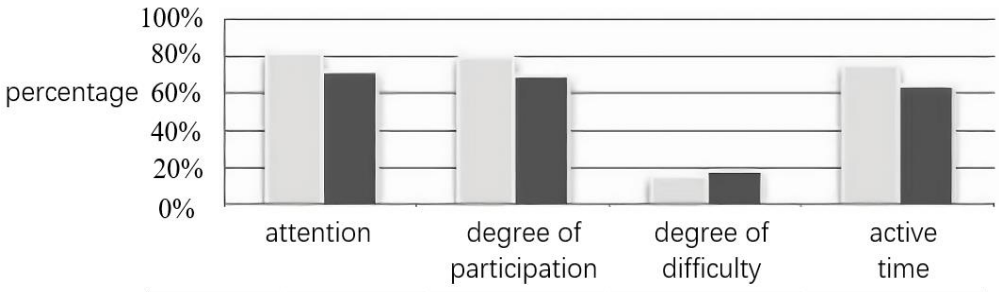


Figure 10 Lip status map of individual students

Figure 9 illustrates how the students' eyebrows are typically "furrowed" during this time. When combined with the head posture, this suggests that although the pupils are mostly psychologically "puzzled," they are currently engaged in "active listening." The condition of the student's lips is depicted in Figure 10, which supports the notion that the student finds it challenging to comprehend the course material.

The attention span, engagement, active time, and difficulty level of the students were all taken into account when evaluating the overall success of the classroom. Through analysis of the discretized images into different t-moment intervals, one may determine the level of difficulty, attentiveness, and involvement of the pupils.

Finally, by comparing the manual statistics with the system detection, the study confirms the system's accuracy in detecting the overall attention, participation, difficulty, and active time in the classroom. The experimental results are displayed in Figure 11. The accuracy rates of the attention, participation, difficulty, and active time are 88%, 87%, 80%, and 85%, respectively, which are all above 80%. This suggests that the system can be used to teach in classrooms and that more precise sentiment data can be obtained [20]. In the figure, the black represents the system detection values and the white represents the actual values.



reality	81%	89%	15%	75%
inspection	71%	79%	18%	64%

Figure 11 Experimental results

7. Results and Discussion

Although the use of computer vision recognition in classroom behavior aids educators in understanding students' learning status in real time and enhances the quality of instruction, many challenges remain that require immediate attention.

It is necessary to increase the precision of target behavior recognition in authentic settings. The latter research project is made more difficult by the fact that in actual classroom settings, items like desks can obscure student targets, variations in lighting will also affect how student data is gathered, and the camera's viewing angle has certain restrictions. The picture information, optical flow information, and audio information acquired in the video may all be fully integrated together using the approach of fusion of target information from different viewpoints. A deep neural network is then built for feature extraction, thereby improving the detection effect.

The dimensions should be further increased by capturing classroom behavior. To obtain more individualized and refined individual behavioral features, collect physiological data on teachers and pupils, such as body temperature, blood pressure, pulse rate, and changes in facial expression. Further investigation is necessary into the fundamental relationship that exists between the classroom behavior of students and the teaching style of teachers. With practice, it is discovered that the two will support one another, and that there is an innate and more sophisticated relationship between teachers' teaching behavior and students' behavior. Teachers' performance will also rapidly improve when students' concentration reaches a particular level. To enhance the diversity of teaching programs, more educators should be involved in classroom behavior modeling school reform research.

In order to improve student learning and teacher lectures, the relationship between teachers and students should be promptly fed back into the background data analysis. Additionally, information security should be prioritized, and teachers' and students' privacy should be further reinforced.

By recognizing and analyzing classroom teaching events in several dimensions from a fresh angle, computer vision brings artificial intelligence into the classroom. This will make it possible to quantify and analyze the ways in which teachers instruct using actual data, which will serve as a foundation for the creation of rational and scientific teaching strategies. Future AI technology is expected to be more integrated into education and to offer substantial backing for instructional improvement.

References

[1] Shao Yichuan, Li Changdi, Cao Yong, et al. Intelligent analysis of classroom behaviour in teaching reform. Teaching and Management, 2021, (15): 29-33.

[2] Shao Yichuan, Li Changdi, Zhao Qian, et al. Artificial intelligence analysis of classroom behaviour characteristics to help teaching reform. Heilongjiang Animal Husbandry and Veterinary Medicine, 2020, (17): 153-158+172-174.

[3] Wu Libao, CAO Yannan, CAO Yiming. Framework Construction of Artificial Intelligence-Enabled Classroom Teaching Evaluation Reform and Technology Implementation. China Electronic Education, 2021, (05): 94-101.

[4] Pantic M., Rothkrantz, L. J. M. Facial Action Recognition for Facial Expression Analysis from Static Face Images. IEEE transactions on systems, man, and cybernetics, Part B. Cybernetics, 2004, 34(3): 1449-1461.

[5] Peng Hui, Zhang Changshui, Rong Gang, et al. Automatic face recognition method based on K-L transform. Journal of Tsinghua University (Natural Science Edition), 1997, (03): 68-71.

[6] L Sirovich, M Kirby. Low-dimensional procedure for the characteri-zation of human face. J. Opt. Soc. Amer., 1987; 4: 519-524.

[7] Ding Rong, Su Guangda, Lin Xinggang. Comparison of feature face and elastic matching face recognition algorithms. Computer Engineering and Applications, 2002(07): 1-2+19.

[8] King D E. Dlib-ml: A machine learning toolkit. Journal of Machine Learning Research, 2009, (10): 1755-1758.

[9] Jia Oriyu, Zhang Zhaohui, Zhao Xiaoyan, et al. Classroom student state analysis based on artificial intelligence video processing. Modern Education Technology, 2019, 29(12): 82-88.

[10] Feng Mantang, Ma Qingyu, Wang Ruijie. Research on intelligent network teaching system based on face expression recognition. Computer Technology and Development, 2011, 21(06): 193-196.

[11] Ziyuan Liu, Chengzhi Jiang. Image number detection based on OpenCV and Haar feature classifier. Journal of Liaoning University of Science and Technology, 2011, (4): 384-388.

[12] Zhou Pengxiao, Deng Wei, Guo Cultivation, et al. Research on intelligent recognition of S-T behaviours in classroom teaching videos. Modern Education Technology, 2018, 28(06): 54-59.

[13] Open Pose Demo-Output. https://github.com/CMU-Perceptual-Computing-Lab/openpose/blob/master/doc/output.md,2018.11.7

[14] Castleman K R. Digital Image Processing, Zhu Zhigang, Lin Xueyan, Shi Dingji, Translation. Beijing: Electronic Industry Press, 1998.

[15] Huang Chunmu, Zhou Lili. Density distribution feature and application in binary image retrieval. Journal of Image and Graphics, 2008, 13(2): 307-311.
[16] Ruan Qiuqi. Digital image processing. Beijing: Publishing House of Electronics Industry 2000: 401-403.
[17] Chiu Y C, Tsai C Y, Ruan M D, et al. Mobilenet-SSDv2: An improved object detection model for embedded systems. 2020 International conference on system science and engineering (ICSSE). IEEE, 2020: 1-5.
[18] Wang Guohui, Zhang Xuan, Zheng Hao. Analysis of students' classroom learning status based on deep learning. Journal of Higher Education, 2022, 8(31): 1-5.
[19] Rosati R, Romeo L, Silvestri S, et al. Faster R-CNN approach for detection and quantification of DNA damage in comet assay images. Computers in Biology and Medicine, 2020, 123: 103912.
[20] Han Li, Li Yang, Zhou Zijia, et al. Analysis of teaching effect based on facial expression in classroom environment. Modern Distance Education Research, 2017, (04): 97-103+112.