Sensors and Materials, Vol. 32, No. 10 (2020) 3243–3259
MYU Tokyo

S & M 2336

Continuous Facial Emotion Recognition Method Based on Deep Learning of Academic Emotions

Szu-Yin Lin,[1*] Chao-Ming Wu,[2] Shih-Lun Chen,[3] Ting-Lan Lin,[4] and Yi-Wen Tseng[2]

[1]Department of Computer Science and Information Engineering, National Ilan University,
No. 1, Section 1, Shennong Road, Yilan City, Yilan County 26047, Taiwan
[2]Department of Information Management, Chung Yuan Christian University,
No. 200, Chung Pei Road, Chung Li District, Taoyuan City 32023, Taiwan
[3]Department of Electronic Engineering, Chung Yuan Christian University,
No. 200, Chung Pei Road, Chung Li District, Taoyuan City 32023, Taiwan
[4]Department of Electronic Engineering, National Taipei University of Technology,
No. 1, Section 3, Zhongxiao East Road, Taipei City 10608, Taiwan

(Received March 15, 2020; accepted June 23, 2020)

Keywords: academic emotions, face emotion recognition, deep learning, convolutional neural networks, long short-term memory networks

It is important to comprehend students' academic emotions in interactive teaching environments. Academic emotions refer to facial expressions that students display along with their academic performance in a learning process. By noting students' academic emotions, teachers can provide the most suitable teaching material according to the emotions to improve their academic performance and motivation. The results can also be subsequently applied to adaptive learning. Recently, some researchers have attempted to study academic emotions with the aid of facial and emotion recognition technologies. However, most studies focused on the analysis and recognition of a single image. It was not considered that academic emotions are a continuous expression in response to the learning situation over a period of time. To address this problem, a continuous facial emotional pattern recognition method based on deep learning is proposed in this study to analyze academic emotions. This method combines the convolutional neural network (CNN) and the long short-term memory (LSTM) network for deep learning to recognize and analyze the continuous facial academic emotional pattern of students and thus recognize academic emotions. Through this method, the e-learning system can understand the learning progress of students quickly and accurately, and offer the students appropriate teaching materials to enhance their academic performance and motivation. The experimental results showed that the recognition accuracies of the CNN model and CNN plus LSTM were 72.47 and 84.33%, respectively. The combination of two neural networks improved the accuracy by approximately 12% compared with that for the CNN alone.

1. Introduction

Academic emotions refer to facial expressions that students display in a learning process along with their academic performance and evaluation results. Teachers can adjust the teaching

*Corresponding author: e-mail: szuyin@niu.edu.tw
https://doi.org/10.18494/SAM.2020.2863

ISSN 0914-4935

content according to students' academic emotions. The purpose is to improve students' academic performance and motivation, and ensure that learning content and evaluation questions are challenges suitable for the levels of the students. Analogous to individualized instructions, such a learning system of dynamic adjustment can provide each student with a unique learning experience and customized learning methods. The learning needs of each student can be satisfied through real-time interaction with computers or people.(1) However, the most difficult task in academic emotion studies is recognizing students' emotions in class, evaluating their level of understanding of the course content, and providing evaluation questions accordingly. Conventional methods record the status of learning and answering for each student through examinations, and the students' status and level of understanding are learned from the examination scores. Subsequently, through the adaptive learning system, courses and exercises for each student can be appropriately adjusted and each student can be guided in different ways or to different degrees. However, this method has one drawback, which is the insufficient immediacy of the adaptive presentation: the learning situation of each student cannot be known immediately at any time during the learning process. Further adjustment and modification can only be implemented after the examination.(2) Therefore, to achieve an adaptive learning system with full immediacy of response, better solutions are required. Understanding academic emotions in real time may be one of the solutions.

In recent years, the world has witnessed the third boom in the development of artificial intelligence (AI). Hardware equipment in the previous boom failed to keep up with the technological progress in AI. Furthermore, obtaining extensive, meaningful, and relevant training data was difficult. Additionally, machine learning algorithms had inherent limitations, hindering the implementation of AI despite the theoretical basis. Recently, following advancements in computational power, the construction of extensive intelligent sensors, progress in machine learning algorithms, and the prevalence of open data concepts, AI technology has begun to enter people's lives and has been formally applied in many fields. The breakthroughs and progress in the field of computer vision are important in AI technology. Computer vision technology aims to replace human eyes with cameras for image recognition and tracking, and to conduct further image analysis and processing via computers. Many researchers continue to focus on affective computing in the field of computer vision. One branch of affective computing takes advantage of images, sounds, words, or videos as training datasets to enable computers to identify human emotions, which is called emotion recognition.(3)

An effective resolution to the problem of delayed feedback in adaptive learning is to recognize students' real-time facial emotional response through images, and make accurate predictions and judgments on their level of understanding and learning status. Facial emotion recognition is a method of processing images of real human faces and extracting from them the corresponding emotions through computer vision and image recognition technology. Important facial features can be captured through the camera in real time, and the changes in these features can be analyzed as the basis of classification for judging various human expressions.(4) However, previous research on facial expression and emotion recognition used in e-learning mainly focused on the analysis and recognition of a single image. These studies did not consider academic emotions as continuous expressions in a time interval, instead of a single

facial expression at some point, for example, in the application of facial expression analysis to e-learning. Therefore, in this study, we aim to understand students' academic emotions through the analysis and recognition of continuous facial academic emotional patterns. Students who are confused in class can be identified through the recognition of facial expressions and emotions, and subsequently, they can be provided with feedback and help to resolve their learning difficulties.

Recently, following the evolution of computer hardware and deep learning algorithms, many recognition and prediction functions have shown good performance in both image processing and data analysis. The number of research studies on related applications is also increasing. Many scholars previously proposed the application of a convolutional neural network (CNN) to image recognition, such as facial recognition or detection using a deep CNN.[5] Most studies were based on a single static image for training, and few were focused on continuous facial emotions. From the above research motivations, in this study, we propose a continuous facial emotional pattern recognition method based on deep learning that can be applied to identify learners' academic emotions. From the analysis and recognition of continuous facial emotional patterns, feedback on students' learning status required in the adaptive learning process can be provided to the system. The two main goals of this study are to (1) establish a continuous facial emotional pattern recognition model and (2) identify the academic emotions of students using a continuous facial emotional pattern model. If the system accurately recognizes the continuous facial emotional patterns of students, it can provide immediate responses in adaptive learning.

2. Related Works

2.1 Academic emotions

Academic emotions are emotions that students develop during learning, which are assessed by researchers.[6] Pekrun proposed two definitions of academic emotions: a broad sense and a narrow sense. Academic emotions in a broad sense refer to the activity-related emotions that students develop from external factor evaluation. Academic emotions in a narrow sense refer to emotions in the process of learning. In 2000, Pekrun also divided emotions into positive and negative categories. Relative to time, academic emotions can be categorized into past, present, and future academic emotions.[7] Academic emotions are one of the branches of emotion classification. There are relatively few studies on academic emotions, thus increasing the difficulty in research as compared with those on general emotions. Before 2000, studies on academic emotions were not fully developed, and hence it was difficult to measure academic emotions. Therefore, in 2000, Pekrun combined achievement-based emotions and attribution theory related to students, and explained that the major academic emotions of students come from value and control. This resulted in three different evaluation criteria, namely, self-performance, situational results, and behavior outcomes.[7]

Additionally, academic emotions are defined by students' academic motivation, learning strategies, cognitive resources, and self-adjustment. Different academic emotions have different influencing mechanisms. Pekrun classified these mechanisms into activation and deactivation.

Activation refers to behavior that aids learning activities, while deactivation refers to behavior that hinders learning activities. By combining the two mechanisms of classification with the positive and negative directions, academic emotions can be classified as positive activation, positive deactivation, negative activation, or negative deactivation. There are three ways to measure academic emotions: curriculum-related, learning-related, and examination-related. Academic emotions can be measured in different environments. Additionally, the measurement methods can deliver different results at different times. A rating scale numerically qualifies students' emotions in a learning environment. Therefore, in 2002, Pekrun developed a situation-specific method to measure more academic emotions.[8]

2.2 Deep learning

Deep learning is a branch of machine learning. Deep learning can process complex data structures via multilevel frameworks, such as speech recognition, image recognition, prediction, and analysis models. The principle is to use a large quantity of data for autonomous learning analogous to that conducted by the complex nervous system of the human brain. Recently, the capability of hardware equipment to manage the computing load required in deep learning has been improved. This accounts for the rapid advances in deep learning. Several algorithms from deep learning have been used to address traditional AI problems[9] in data science, such as speech recognition, image recognition, and data retrieval. These techniques aim at data abstraction through multiple nonlinear transformations.[10] Compared with traditional shallow neural network methods, deep learning is efficient in terms of the accuracy and error rate of the model.[11] If the model has more layers, its accuracy and error rate are better. However, it is difficult to train a deep neural network. The difficulty lies in maintaining the parameter relationships between the layers as the number of deep learning layers increases. Unqualified training results in vanishing and exploding gradients. There are many models of deep learning. In this study, we employed several long short-term memory (LSTM) and CNN models, and applied the trained models to the adaptive learning system. The CNN was used for image recognition, and the processed output value of the CNN was trained by the LSTM. Finally, the relationship between the LSTM output and learners' emotions was analyzed.

2.3 LSTM

The LSTM model mainly replaces the structure of hidden layers in the recurrent neural network (RNN). The LSTM uses gates and memory cells to solve the problems of vanishing and exploding gradients in the RNN. In the control valve, a forget gate, an input gate, and an output gate are used to control the input and output values. There is a special framework in the LSTM model that enables the LSTM to forget unnecessary messages. The sigmoid and tanh functions are used as switches to control whether the LSTM needs to forget messages. The LSTM remembers all messages or forgets the information entered in a layer if the sigmoid function value is 1 or 0, respectively. By adjusting the values of the sigmoid and tanh functions, the input and output values can be controlled, and consequently, the problems of vanishing and exploding gradients during RNN training can be solved.[12]

2.4 CNNs

The CNN was proposed in 1960. It is a unique network structure based on the neurons used for direction selection in the cat cortex and can effectively reduce the complexity of a feedback neural network. However, advances in CNNs were hindered by the inability of computer hardware to manage the complexity of neural networks at that time. Owing to technological advances in recent years, computers now have the power required to manage neural networks. Consequently, CNNs have brought about another boom in machine learning. In particular, CNNs, which are typical deep neural networks in deep learning, have excellent accuracy performance in image recognition and object detection.[13,14] A typical CNN structure is composed of multiple convolutional, pooling, and fully connected layers.[9] The convolutional and pooling layers extract the features of the target image and consider these features as parameters of the input image (weights and bias), and the fully connected layers update the parameters of the neural network. Parameter adjustment is used to find the most suitable training model; after sufficient training iterations, the training can be stopped.[9] Finally, the output is classified by the softmax function. Recently, CNNs have been widely used in areas such as image recognition and face detection in computer vision. Most image-based studies related to neural networks employ CNNs. However, studies such as face detection focus on a single static image.[15] Few studies such as the video dataset[16] and video recognition using two independent CNNs[17] focus on the CNN-based dynamic video analysis model.

2.5 Facial emotion recognition

Face detection is the first step of facial emotion recognition. Face detection is mainly used to detect and locate a human face in an image and obtain the feature point coordinates of the human face. Face recognition extracts the identity features contained in each human face and compares them with known human faces to determine the identity of each face. Following developments in deep learning, face detection and recognition methods based on deep learning techniques have been realized.[14] For example, a multitask cascaded convolutional network (MTCNN) is a face recognition model based on the CNN. Its main framework is divided into three types: P-NET, O-NET, and R-NET. Three models are used to filter images. Finally, the faces in the image are saved to achieve the effect of face recognition.[18] Face recognition is based on the Facial Action Coding System (FACS) proposed by Ekman and Friesen in 1978. FACS defines action units (AUs) for the human face. Using a combination of AUs, the emotional expression of a human face can be defined.[19] Facial emotion recognition can be realized through human expressions such as happiness, anxiety, and sadness. Human psychological states are often reflected in facial expressions; thus, a person's internal emotions can be learnt from changes in facial expressions. In previous studies, computer vision was used to detect human facial emotional expressions, such as by using the AdaBoost algorithm proposed by Freund and Schapire in 1997.[20] In recent years, with the rapid developments in neural networks, emotional recognition has become an important research topic. Emotional recognition through neural networks is mainly carried out using a CNN. For example, in 2014,

Ouellet proposed a computer-vision-related method to extract facial expression features for facial emotional recognition. However, most methods were based on a single static image rather than a continuous pattern of facial expressions.[4]

3. Methodology

3.1 Deep learning model

In this study, we propose a model and method of continuous facial emotional pattern recognition that combines a CNN, LSTM, and facial emotion recognition. We first trained a CNN model to recognize facial emotions. Subsequently, the classification results from the CNN model were input into the LSTM for time series tagging, and a continuous facial emotion model was developed. Finally, we used the facial emotion model described above to build the prediction method for students' emotions.

(1) First stage: establishment of continuous facial academic emotion recognition model based on CNN

The first stage of this method establishes a continuous facial emotion recognition model based on the CNN to recognize academic emotions. The first step involves three steps. The first step is the data preprocessing of the emotion dataset, where the face in the image is extracted, then the AUs of the face are extracted, and finally, the academic emotion is defined according to the FACS emotion definition. The second step is to input the processed academic emotion dataset into the CNN for training to obtain the academic emotion recognition model. The final step is to classify academic emotions based on dataset recognition. The first stage corresponds to Phase 1 of Fig. 1.

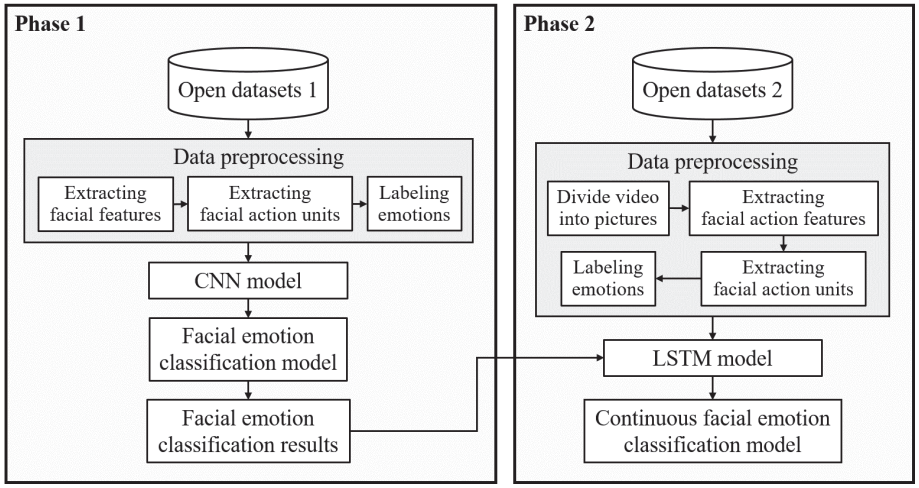

Fig. 1. System architecture.

• Data preprocessing

Data preprocessing ensures the availability of data in the system, keeps the data format consistent, and provides clear information for training the CNN. After a face position is detected, the emotional AUs of the face are extracted. Subsequently, as per the definition of FACS emotions, academic emotions are defined and inputted into the CNN model for training. Finally, the preprocessed emotional images are classified into training and test datasets. The training dataset is input into the CNN model for learning, and the test dataset is used to verify the accuracy of the trained model.

• Definition of academic emotions

Previous studies on expression recognition were based on seven general emotions defined by FACS: neutral, surprised, fearful, disgusted, happy, sad, and angry. Academic emotions are the emotions of students while using a learning system. In this study, we used action features with larger changes in facial features to define the academic emotion in order to increase the recognition accuracy of the model. Both academic and general emotions were defined relative to the basic AUs of FACS; thus, they have similarities in facial expression features. Therefore, if the academic emotion dataset is insufficient, we can apply the transfer learning method in deep learning to enlarge the training dataset used to train the emotion recognition model.

• Data training on AlexNet network model

Studies indicate that the model accuracy increases with the number of hidden layers in the CNN framework. However, the training time also increases. Considering the time consumption and the model, in this study, we used the AlexNet model for training. AlexNet is an eight-layer CNN model, including five convolutional layers and three fully connected layers. The nonlinear activation function was used to accelerate the convergence. The method of model training was supervised learning. The input data contained the tag data of the image, and the model accurately predicted the academic emotions through the repeated training of the neural network. In image recognition, AlexNet demonstrates good performance. Although it is shallower than other CNNs, a high recognition accuracy is guaranteed. Therefore, AlexNet is a highly suitable deep learning framework for this study.

• Facial emotion classification

The CNN model was trained with the training dataset, and the first-stage prediction model was generated after repeated iterative training and parameter adjustment. Then, the test dataset was used to verify the first-stage prediction model and obtain the first-stage prediction results. The final prediction results of the first stage can be divided into two cases:

(a) Accurate recognition of facial emotions: accurate classification of facial emotions in images.

(b) False recognition of facial emotions: inaccurate classification of facial emotions in images.

Finally, in the final prediction results of the first stage, the probability of correct recognition of facial emotions was used as the input dataset of the second stage.

(2) Second stage: training of the LSTM model

In the second stage, the LSTM model was used for model training. The LSTM is efficient in processing time series datasets such as continuous images or videos. In this study, the CNN output of academic emotion classification was input into the LSTM for training to make the classification of academic emotions a time series. Thus, we can determine the emotional changes of students over a period of time and realize continuous facial emotion recognition. The second stage corresponds to Phase 2 of Fig. 1.

3.2 Establishment of classification method

From the CNN and LSTM proposed in Sect. 3.1, a deep learning model of prediction and classification was established. The model framework is shown in Fig. 2. The AlexNet framework, which included five convolutional layers and three fully connected layers, was used for the CNN. The SoftMax layer was the final output layer, and the results were classified into seven categories. The classification results were input into the LSTM model for training to obtain the final output. The CNN model parameters referenced the settings of AlexNet, as shown in Table 1.

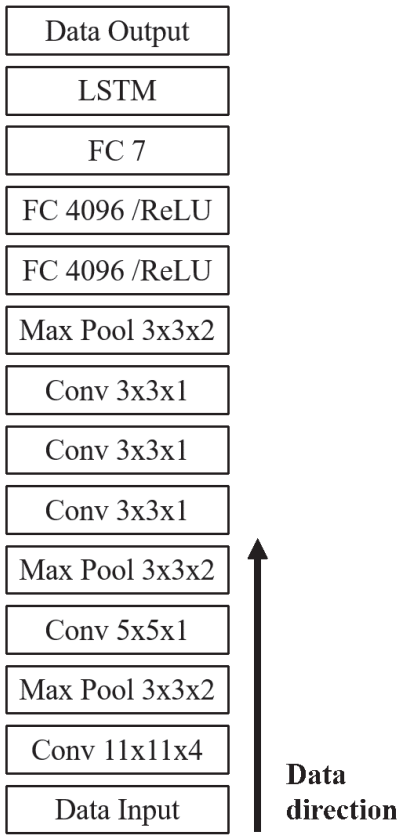

Fig. 2. Network layers.

Table 1
Setting parameters.

	Kernel size	Convolutional kernel size	Convolutional kernel number
Conv 1	11 × 11	96	11
Max Pooling 1	3 × 3	—	—
Conv 2	5 × 5	256	5
Max Pooling 2	3 × 3	—	—
Conv 3	3 × 3	384	3
Conv 4	3 × 3	—	—
Conv 5	3 × 3	384	3
Max Pooling 3	3 × 3	256	3
LSTM	—	—	—
SoftMax	Classifier	—	—

- Data input layer: The data input layer was mainly for the image input. After preprocessing, each image used for expression recognition was tagged. 80% of the dataset was used as the training dataset and the remaining 20% was used as the test dataset. The input images were gray-scale images with 48 × 48 pixels.
- Convolutional layer: The convolutional layer is the core of the CNN. The CNN can use the convolutional layer to obtain the features of the image and produce many different feature images that can be input to the fully connected layer. There were five convolutional layers in this experiment, and the parameter adjustment was conducted relative to AlexNet.
- Pooling layer: In this experiment, the pooling method was max-Pooling. The pooling layer reduces the number of parameters and dimensions of features, with a focus on reducing the training time for the same training effect. In this experiment, there were three pooling layers, which were added after the first, second, and fifth convolutional layers.
- ReLU: In the training of the deep learning model, all parameters were updated from their slopes. The ReLU endows models with nonlinear characteristics to approximate the real data.
- FC: The fully connected layer transforms the two-dimensional vector into a one-dimensional vector. In this experiment, the expression recognition results were classified into seven categories, and thus the fully connected layer had seven output parameter nodes.
- SoftMax layer: The SoftMax layer assigns weights to the outputs of the fully connected layer through the SoftMax function. The sum of these results was 1, and the result with the highest classification value was the model prediction.

3.3 Collection of academic emotion data

The collection of academic emotions was based on the Achievement Emotions Questionnaire (AEQ) manual proposed by Pekrun in 2005. The AEQ rating scale reveals students' academic emotions through questions that were designed according to the criteria of the five-point Likert scale. There are three ways to measure the AEQ rating scale: curriculum-related, learning-related, and examination-related. In our experiment, we selected the curriculum-related questions and compiled them into questions on academic emotions. There were 16 questions in the rating scale that corresponded to eight types of academic emotions, namely, enjoyment, hope, pride, anxiety, shame, anger, hopelessness, and boredom. In the experiment, the collected academic emotions were classified into positive and negative emotions to accurately distinguish them.

4. Experiments

This section is divided into two parts. The first part deals with the evaluation and design of the experiment, and the second part is the presentation and discussion of each model and experimental data.

4.1 Experimental design

4.1.1 Dataset for emotion classification

We used two types of public datasets in this experiment: an open dataset, the Facial Expression Recognition Challenge 2013 (FER2013), and an academic emotion dataset, the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). FER2013 is the 2013 champion image collection of facial expression recognition in the Kaggle community, with about 40000 examples of 48 × 48 grayscale images. Many studies have been carried out using this image dataset, which includes images with seven emotional categories: neutral, surprise, fear, disgust, happy, sad, and angry. The RAVDESS emotional dataset contains 828 videos of 5 s duration divided into five types (neutral, happiness, sadness, anger, and fear), each portrayed by 23 professional actors. The numbers of videos and images in each emotion category are shown in Table 2. In the experiment, each video was stopped at intervals of 1 s, which were saved as image files. In other words, a 5 s video was saved as 5 images, giving a total of 4140 images showing emotions.

4.1.2 Dataset for academic emotion classification

In this experiment, 21 college students were tested using the predesigned AEQ questions. The measurement table was divided into three units. The questions in the three units were the same, and the steps of the test are shown in Fig. 3. After watching the clip of each unit, the corresponding measurement table was filled, and after watching all three clips, the test was completed.

Table 2
Numbers of videos and images in each emotion category.

	Neutral	Happiness	Sadness	Anger	Fear	Total
Videos	92	184	184	184	184	828
Images	460	920	920	920	920	4140

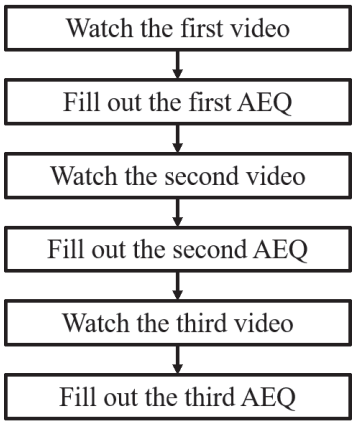

Fig. 3. Test steps.

The scoring method of the rating scale adds the scores of questions on the same academic emotion, and the emotion with the highest score was considered the academic emotion of the subject after watching the clip. Scores of 1, 2, 3, 4, and 5 points were assigned to "strongly disagree", "disagree", "neutral", "agree", and "strongly agree", respectively. Enjoyment, hope, and pride were classified as positive academic emotions, while anxiety, shame, anger, hopelessness, and boredom were classified as negative academic emotions. As there were three positive academic emotions and five negative academic emotions, this led to an uneven distribution of scores in the rating scale. Therefore, score weightage adjustment is required. Finally, the academic emotions derived from the rating scale were matched to the corresponding clips; the process is shown in Fig. 4. Each of the 21 subjects watched three clips, making a total of 63 clips. After the statistical analysis, a total of 55 clips of positive academic emotions and 8 clips of negative academic emotions were obtained.

4.1.3 Data preprocessing

The data preprocessing method used in this study was face detection, in which the facial features of unprocessed images were extracted and then the images were saved as new image files. In the data preprocessing, we conducted face extraction for the emotions taken from the RAVDESS dataset. In total, 4140 images of emotions and 63 clips of academic emotions were extracted from the 828 5 s clips.

4.1.4 Evaluation metric

We used the accuracy of the model as the metric for the experimental evaluation. The definition of accuracy is as follows.

$$Accuracy = (TP + TN) / (TP + TN + FP + FN), \quad (1)$$

where TP = true positive; FP = false positive; TN = true negative; FN = false negative.

As shown in Fig. 5, this experiment involved three comparisons, as follows:

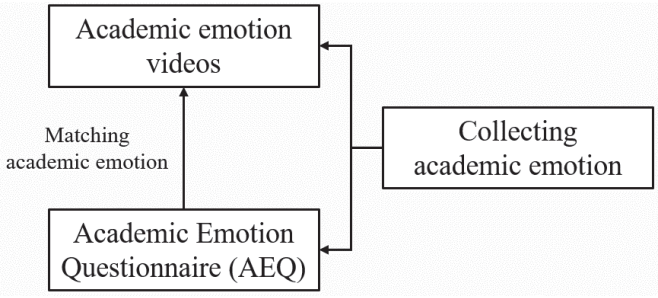

Fig. 4. Academic emotion matching.

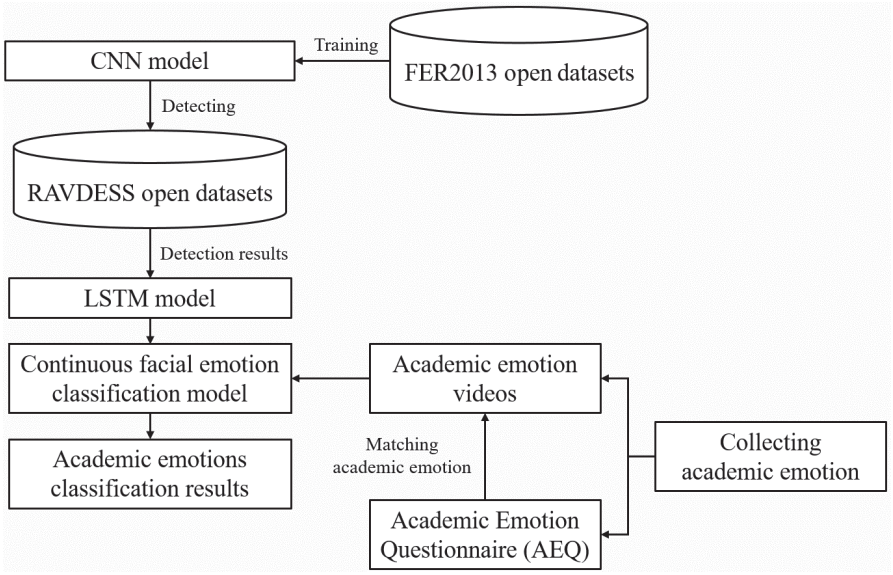

Fig. 5. Experimental process.

(1) Comparison of data preprocessing methods

We compared images with and without face detection to determine the recognition accuracy of the model. In the experiment, face detection was performed on the 4140 images of emotions to obtain their corresponding preprocessed images. Finally, a comparison was made between the images of emotions with and without preprocessing to determine the recognition accuracy of the model.

(2) Comparison of neural network models

We compared the facial emotion recognition accuracy between the model with only the CNN and the model with both the CNN and LSTM. We input the same 4140 preprocessed images of emotions into the two models to compare their recognition accuracies.

(3) Comparison of subjective and objective academic emotions

We compared the difference between the academic emotions measured using the rating scale and those tagged by researchers.

4.2 Experimental results

The experiment in this section was divided into three cases for discussion. Case 1 compared data preprocessing methods, Case 2 compared neural network models, and Case 3 compared subjective and objective academic emotions.

- **Case 1: Comparison between data preprocessing methods**

In this section, the experiment compared the recognition accuracies of images of emotions with and without face detection. The training dataset was the public FER2013 dataset, and

the test dataset was composed of images of faces with and without face detection. The goal of accuracy evaluation was to accurately classify the five facial emotions of neutral, happiness, sadness, anger, and fear, as well as positive and negative emotions. First, the classification of the five facial emotions was conducted. The first step considered the 4140 original images without face detection as the test dataset, which were directly input into the AlexNet CNN for recognition. As shown in Table 3, experimental results showed that the accuracy was 12.76%. In the second step, the 4140 images with face detection were used as the test dataset, which were input into the AlexNet CNN for recognition. The accuracy was 43.17%. Next, we examined the classification of positive and negative emotions. In the classification, happiness was classified as a positive emotion, whereas sadness, anger, and fear were classified as negative emotions. The recognition accuracy of positive and negative emotions was 72.47%, as shown in Table 3.

In Case 1, the experiment determined the difference between the methods with and without face detection in data preprocessing. The accuracy of face detection with data preprocessing was 43.17%, and that without face detection was 12.76%. From the above experimental charts, it is observed that the difference in accuracy was significant. Without face detection, the model nearly always failed to correctly judge the emotion of images. Although the images were captured within the same second, it is observed that the final results depend on whether face detection was conducted. However, the accuracy was 43.17% for the CNN, although the accuracy was increased with the aid of face detection. Then, emotions were classified into positive and negative classes. The accuracy increased to 72.47% through the classification into positive and negative emotions. Although this accuracy is not perfect for the CNN, it is much greater than that of 12.76% without any data preprocessing.

- **Case 2: Comparison of neural network models**

(1) CNN model

The experiment on the CNN model in Case 2 reused the results of Case 1. The same 4140 images of emotions obtained from the RAVDESS dataset were used. In this section, a comparison was conducted relative to the five emotion classifications and positive/negative emotion classifications. The five emotion classifications were neutral, happiness, sadness, anger, and fear. Emotions were also divided into positive and negative ones.

(2) CNN model + LSTM model

In the experiment on the CNN with the LSTM, the dataset also consisted of clips of emotions taken from the RAVDESS dataset. Out of the 828 videos of emotions in total, 108 videos were used as the test dataset and 720 videos were used as the training dataset. As before, five images were obtained from videos at 1 s intervals. Thereafter, the training dataset was input into

Table 3
Recognition results of five emotions and positive and negative emotions (with and without face detection).

Method (CNN model used)	Accuracy (%)
Five emotions without face detection	12.76
Five emotions with face detection	43.17
Positive and negative emotions with face detection	72.47

the CNN model for classification, and then the SoftMax classification result obtained by the CNN was used as the input of the LSTM. Finally, a continuous emotion classification model was trained. Then, the test dataset with 108 videos was input into the continuous emotion classification model for testing. The results are shown in Table 4. The recognition accuracy of the five emotions was 74.07% and that of the positive and negative emotions was 84.33%.

The experiment in Case 2 showed that the recognition accuracy of 84.33% obtained using both the CNN and the LSTM was higher than that obtained using only the CNN (72.47%). Additionally, since 84.33% was the average emotion recognition accuracy, the model was particularly accurate in classifying certain emotions.

- **Case 3: Comparison of subjective and objective academic emotions**

(1) Subjective recognition of academic emotions

The subjective recognition of academic emotions refers to emotions that the human labelers recognize as positive/negative academic emotions in the images. In the experiment, the subjectively recognized academic emotions were regarded as the training dataset, and the academic emotions obtained from the rating scale were regarded as the test dataset. There were 828 videos in the training dataset, including 63 videos of academic emotions. The experimental procedure followed that of Case 2. The 828 videos of emotions were used as the training dataset to train a continuous emotion classification model. Finally, 63 videos of academic emotions were used as the test dataset, and 50 videos in the test dataset were accurately recognized, giving an accuracy of 79.3%, as shown in Table 5.

(2) Objective recognition of academic emotions

The objective recognition of academic emotions refers to the positive/negative academic emotions obtained from subjects assessing emotions on the academic emotion rating scale. The dataset contained 63 videos used in the objective recognition of academic emotions, with 12 videos used as the test dataset and 51 videos used as the training dataset. The experiment in this section followed the same training method as the continuous emotion recognition model used in Case 2. Fifty-one videos were used as the training dataset and 12 videos were used as the test dataset. Nine videos were accurately recognized, giving an accuracy of 75%.

There was little difference between the recognition results for subjective and objective academic emotions, which were close to the value of 84.33% obtained for Case 2. The objective recognition of academic emotions was hindered owing to the small size of the dataset, reducing the accuracy. However, the subjective recognition of academic emotions was closer to the result obtained in Case 2. This shows that the combination of the CNN and LSTM improves recognition accuracy, as shown in Table 5.

Table 4
Recognition results of CNN + LSTM model.

Method (CNN+LSTM model)	Accuracy (%)
Five emotions	74.07
Positive and negative emotions	84.33

Table 5
Recognition results of subjective and objective academic emotions.

Method	Accuracy (%)
Subjective academic emotions	79.3
Objective academic emotions	75

5. Conclusions

In this study, a continuous facial emotion pattern recognition method based on deep learning is proposed and applied to academic emotions. This method combines both the CNN and the LSTM for deep learning to identify the continuous facial academic emotional pattern of students and recognize academic emotions. By this method, future learning systems can understand students' learning situation quickly and accurately, and offer immediate responses to improve students' academic performance and motivation. We experimentally showed that the recognition accuracy was 72.47% when only the CNN was used and 84.33% when both the CNN and the LSTM were used, i.e., combining the two neural networks improved the accuracy by approximately 12%, indicating that the method proposed in this study can detect students' continuous academic emotions more accurately. However, a drawback is the difficulty in actual implementation because the changes in the expressions of students in an e-learning environment are often insignificant. Therefore, in this study, we showed three distinct types of instructional videos to the subjects, so that the subjects were more likely to see different expressions, thereby increasing recognition accuracy. Experiments were used to determine whether face detection used in data preprocessing has a profound impact on recognition accuracy. To maintain the accuracy of academic emotion recognition, it is necessary to detect the face first and then carry out continuous expression recognition. In future research using sufficiently detailed academic emotion datasets (such as eight classifications of academic emotions) for facial emotion recognition, more detailed academic emotion classification training can be conducted. Regarding model selection, with sufficient resources, a more advanced CNN model is recommended to improve the model accuracy. In future works, this model can be used in situations of adaptive learning. By identifying the continuous facial academic emotions of students in class, we can then recognize students that may have difficulties in learning and determine suitable teaching modes for these students.

Acknowledgments

This research was supported by the Ministry of Science and Technology, Taiwan, under Grant MOST 108-2410-H-197-004.

References

1 M. J. Tyre and E. v. Hippel: Organ. Sci. **8** (1997) 71. http://dx.doi.org/10.1287/orsc.8.1.71
2 P. Brusilovsky: Künstliche Intelligenz **13** (1999) 19. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.363.9506
3 S. E. Kahou, X. Bouthillier, P. Lamblin, C. Gulcehre, V. Michalski, K. Konda, S. Jean, P. Froumenty, Y. Dauphin, N. Boulanger-Lewandowski, R. Chandias Ferrari, M. Mirza, D. Warde-Farley, A. Courville, P. Vincent, R. Memisevic, C. Pal, and Y. Bengio: J. Multimodal User Interfaces **10** (2016) 99. http://dx.doi.org/10.1007/s12193-015-0195-2
4 S. Ouellet: ArXiv 1408.3750 (2014) 1. https://arxiv.org/pdf/1408.3750.pdf
5 Y. Wen, K. Zhang, Z. Li, and Y. Qiao: A Discriminative Feature Learning Approach for Deep Face Recognition (LNCS, 2016) p. 499. https://doi.org/10.1007/978-3-319-46478-7_31
6 R. Pekrun: Analyzing and Enhancing Students' Emotions: A Neglected Topic of Educational Research (Psychologie in Erziehung und Unterricht, 1998) p. 230.

7 R. Pekrun: A Social-Cognitive, Control-Value Theory of Achievement Emotions: In Advances in Psychology (Elsevier Science, North-Holland, 2000) p. 143. https://doi.org/10.1016/S0166-4115(00)80010-2
8 R. Pekrun, T. Goetz, W. Titz, and R. P. Perry: Educ Psychol **37** (2002) 91. https://doi.org/10.1207/S15326985EP3702_4
9 Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew: Neurocomputing **187** (2016) 27. https:/doi.org/10.1016/j.neucom.2015.09.116
10 B. Chandra and R. K. Sharma: Neurocomputing **171** (2016) 1205. https://doi.org/10.1016/j.neucom.2015.07.093
11 F. Seide, G. Li, and D. Yu: Proc. Interspeech (2011) 437–440.
12 S. Hochreiter and J. Schmidhuber: Neural Comput. **9** (1997) 1735. https://doi.org/10.1162/neco.1997.9.8.1735
13 R. Girshick, J. Donahue, T. Darrell, and J. Malik: 2014 IEEE Conf. Computer Vision and Pattern Recognition. (IEEE, 2014) 580–587. https://doi.org/10.1109/CVPR.2014.81
14 A. Krizhevsky, I. Sutskever, and G. Hinton: Neural Inf. Process. Syst. **25** (2012). https://doi.org/10.1145/3065386
15 H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua: 2015 IEEE Conf. Computer Vision and Pattern Recognition (CVPR) (IEEE, 2015) 5325–5334. https://doi.org/10.1109/CVPR.2015.7299170
16 A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and F. F. Li: 2014 IEEE Conf Computer Vision and Pattern Recognition (CVPR) (IEEE, 2014) 1725–1732. https://doi.org/10.1109/CVPR.2014.223
17 K. Simonyan and A. Zisserman: Proc. 27th Int. Conf. Neural Information Processing Systems 1 (NIPS, 2014) 568-576. https://dl.acm.org/doi/10.5555/2968826.2968890
18 K. Zhang, Z. Zhang, Z. Li, and Y. Qiao: IEEE Signal Process. Lett. **23** (2016) 1499. https://doi.org/10.1109/LSP.2016.2603342
19 P. Ekman and W. V. Friesen: Facial Action Coding System: A Technique for the Measurement of Facial Movement (Consulting Psychologists Press, Palo Alto, Calif., 1978).
20 Y. Freund and R. E. Schapire: J. Comput. Syst. Sci. **55** (1997) 119. https://doi.org/10.1006/jcss.1997.1504

About the Authors

Szu-Yin Lin received his M.S. and Ph.D. degrees from National Chiao-Tung University, Taiwan, in 2012. He was a visiting scholar with Coventry University, UK, in 2012. He was an assistant professor from 2013 to 2017 and an associate professor from 2017 to 2019 with the Department of Information Management, Chung Yuan Christian University in Taiwan. Since 2019, he has been an associate professor with the Department of Computer Science and Information Engineering, National Ilan University in Taiwan. His research interests include the areas of intelligent data analysis, applied artificial intelligence, deep learning, management information systems, service-oriented computing, and multi-agent distributed computing.
(szuyin@niu.edu.tw)

Chao-Ming Wu received his B.S. and M.S degrees from Chung Yuan Christian University (CYCU), Taiwan (R.O.C.), in 1987 and 1989, respectively, and his Ph.D. degree from Central University, Taiwan, in 1999. From 1999 to 2010, he was an associate professor at CYCU, Taiwan. Since 2010, he has been a professor at CYCU. His research interests are in data mining, smart city, e-marketing, and e-government. (mislighter@gmail.com)

Shih-Lun Chen received his B.S., M.S., and Ph.D. degrees from National Cheng Kung University, Tainan, Taiwan, in 2002, 2004, and 2011, respectively, all in electrical engineering. He was an assistant professor from 2011 to 2014 and an associate professor from 2014 to 2017 with the Department of Electronic Engineering, Chung Yuan Christian University in Taiwan, where he has been a professor with the same department since 2017. His current research interests include VLSI chip design, image processing, wireless body sensor network, Internet of Things, wearable devices, data compression, fuzzy logic control, bio-medical signal processing, and reconfigurable architecture. Dr. Chen was a recipient of the Outstanding Teaching Award from Chung Yuan Christian University in 2014 and 2019. (chrischen@cycu.edu.tw)

Ting-Lan Lin received his B.S. and M.S. degrees in Electronic Engineering from Chung Yuan Christian University, Zhongli, Taoyuan, Taiwan, in 2001 and 2003, respectively, and his Ph.D. degree in electrical and computer engineering from the University of California, San Diego, La Jolla, CA, USA, in 2010. In 2008, he interned in the Display System group at Qualcomm, San Diego, CA, USA. Since 2011, he was an assistant professor with the Department of Electronic Engineering, Chung Yuan Christian University, Taiwan, and since August 2015, he was an associate professor with the Department of Electronic Engineering, Chung Yuan Christian University, Taiwan. Since August 2018, he has been an associate professor with the Department of Electronic Engineering, National Taipei University of Technology, Taipei, Taiwan. His current research interests include (multiview) video compression, pixel estimation, and 3D point cloud video compression, with the techniques of mathematical optimization. (dtxion@gmail.com)

Yi-Wen Tseng received his B.S. and M.S. degrees from Chung Yuan Christian University, Taiwan, in 2017 and 2019, respectively. His research interests are in intelligent data analysis, machine learning, deep learning, and computer vision. (ian543850204@gmail.com)