IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 24 August 2023, accepted 3 October 2023, date of publication 6 October 2023, date of current version 16 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3322454

SURVEY

Facial Expression Recognition in Educational Research From the Perspective of Machine Learning: A Systematic Review

BEI FANG, XIAN LI, GUANGXIN HAN, AND JUHOU HE
Key Laboratory of Modern Teaching Technology, Ministry of Education, Shaanxi Normal University, Xi'an, Shaanxi 710062, China
Corresponding author: Juhou He (juhouh@snnu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 62107027 and Grant 62177032, and in part by the China Postdoctoral Science Foundation under Grant 2021M692006.

ABSTRACT Facial expression analysis aims to understand human emotions by analyzing visual face information and is a popular topic in the computer vision community. In educational research, the analyzed students' affect states can be used by faculty members as feedback to improve their teaching style and strategy so that the learning rate of all the students present can be enhanced. Facial expression analysis has attracted much attention in educational research, and a few reviews on this topic have emerged. However, previous reviews on facial expression recognition methods in educational research focus mostly on summarizing the existing literature on emotion models from a theoretical perspective, neglecting technical summaries of facial expression recognition. In order to advance the development of facial expression analysis in educational research, this paper outlines the tasks, progress, challenges, and future trends related to facial expression analysis. First, facial expression recognition methods in educational research lack an overall framework. Second, studies based on the latest machine learning methods are not mentioned in previous reviews. Finally, some key challenges have not been fully explored. Therefore, unlike previous reviews, this systematic review summarizes two kinds of educational research methods based on facial expression recognition and their application scenarios. Then, an overall framework is proposed, along with various kinds of machine learning methods and published datasets. Finally, the key challenges of face occlusion and the expression uncertainty problem are presented. This study aims to capture the full picture of facial expression recognition methods in educational research from a machine learning perspective.

INDEX TERMS Facial expression recognition, educational research, machine learning, survey.

I. INTRODUCTION

Emotions strongly influence human behavior in individual and social context. Emotion recognition plays an important role in human-computer interaction and can be used for digital advertising, customer feedback evaluation, and healthcare. Emotion recognition also provides support in a variety of important educational applications, including the recognition of learners' cognitive states in E-learning [1], the detection of students' emotional engagement [2], the assessment of students' mental health, and the assessment of teachers [3], etc. The analyzed students' affect state can be used by faculty members as feedback to improve their teaching style and strategy so that the learning rate of all the students present can be enhanced. Therefore, the application of emotion recognition in the field of education has been extensively studied by researchers.

Specifically, one study [4] stated that emotions can have a negative or positive effect on motivation and therefore emotions should be taken into consideration during learning. Pekrun et al. [5] suggested that some positive emotions are positively associated with intrinsic motivation, effort, self-regulation, and more complex learning strategies, whereas negative emotions such as anger are associated with anxiety

The associate editor coordinating the review of this manuscript and approving it for publication was Gang Mei.

and boredom, reduced effort, poor performance, increased external regulation and reduced self-regulatory strategies [6], [7], [8]. Generally speaking, on the one hand, learners' emotional states can reflect their preferences concerning the teaching content, teaching media, and teaching environment, which is helpful for understanding cognitive styles and learning interests. On the other hand, it can reflect the influence of a learner's knowledge level, cognitive structure, and learning motivation on the learner's subjective learning experience, which can be helpful for analyzing the learning mechanism. Therefore, it is necessary to study emotion recognition in education [9], [10], [11], [12], [13], [14].

In traditional classroom learning, teachers can adjust their teaching strategies by observing students' facial expressions and body movements. However, due to the large number of students in classroom or E-learning setting, teachers cannot estimate the emotions of each student, so it is difficult to help teachers change their teaching strategies according to the emotions of students. In recent years, with the rapid development of computer technology and artificial intelligence, automatic emotion recognition has become an effective recognition method. The students' emotions can be analyzed using intrusive and non-intrusive techniques [15]. Compared to the use of intrusive techniques, such as physiological sensors, to acquire emotional characteristics, with the advent of input devices such as cameras, samples of learners' facial expressions can be easily collected to use as a baseline for emotion recognition [16]. Facial expressions can be recognized from static images or a sequence of images or videos. Facial expressions are one of the most powerful, natural, and universal signals that human beings can use to convey their emotional states and intentions, especially in a learning environment [17]. Mehrabian [18] claimed that 93% of the emotional meaning is transmitted as follows: 7% come from verbal expression, 38% come from vocal expression, and 55% come from facial expression. A study by the psychologist Paul Ekman showed an accuracy rate of 88% for mapping facial expressions to a single specific emotional state [19]. It is evident that facial expressions play a key role in the expression of learners' emotions.

In recent years, exhaustive surveys on the use of machine learning techniques in computer vision for automatic facial expression analysis have been published [20], [21], [22]. However, reviews of the literature on the use of facial expression recognition for emotion detection in a learning environment are relatively scarce and focus mostly on summarizing the existing literature on emotion models from a theoretical perspective, while neglecting technical summaries of facial expression recognition. The following four representative literature reviews were found [23], [24], [25], [26].

Specifically, some reviews [23], [24] analyzed affective computing in educational research through a review of journal publications. However, there are fewer analyses and summaries of the algorithms or systems used in affective computing. Another review [25] analyzed the main applications programmable interfaces (APIs) and tools available today for emotion detection and discussed their main features. However, it focused on APIs, and the algorithmic techniques have rarely been reviewed from a machine learning perspective. Recently, a review of emotion recognition methods applied to E-learning was provided in [26];it reviews the development of emotion recognition from its beginnings to 2020 and comparatively analyzes the applicability of algorithms in E-learning systems. However, the key problems of automatic facial expression recognition in the traditional classroom environment are not summarized.

Although previous reviews have involved considerable research work and proposed some solutions, there exist some common limitations. First, different types of facial expression recognition methods in educational research are not summarized and classified. Second, the framework and technological process of facial expression recognition in educational research are not analyzed. Third, with the development of machine learning techniques, an increasing number of studies on facial expression recognition have emerged. Some other key techniques and issues have not been fully explored in educational research, such as the occlusion and variant pose problems in real classroom settings and expression uncertainty problem. Therefore, unlike past reviews, one objective of this review is to overcome these limitations.

The main contributions of the present review are the followings: (1) On the basis of previous studies, this paper summarizes two kinds of educational research methods based on facial expression recognition and their application scenarios. (2) The overall framework of facial expression recognition in educational research from a machine learning perspective is proposed; this framework is not mentioned in previous reviews. (3) Various kinds of machine learning methods are introduced, and the characteristics of published datasets are discussed. (4) Some key challenging problems, including the face occlusion and pose problems in an uncontrolled environment and the expression uncertainty problem, are discussed in this review.

In short, this study conducts more specific and detailed research on both static and dynamic facial expression recognition tasks in educational settings up to 2023. Our aim is to provide newcomers to the field with an overview of the system framework and key skills for facial expression recognition, and to establish a standard set of algorithmic pipelines for facial expression recognition in educational settings.

II. REVIEW METHOD AND SELECTION OF ARTICLES

In this study, we considered articles from journals, conferences, and workshops published in the English language from January 2012 to March 2023. The key terms used to perform the search are ''facial expression recognition/sentiment analysis/emotion recognition in education/learning''. Articles are selected according to the keywords for the whole text. The electronic bibliographic databases used include the

IEEE Xplore Digital Library, ACM Digital Library, Elsevier (ScienceDirect), Wiley Online Library, Springer (SpringerLink), Taylor & Francis, Google Scholar, and Web of Science.

Using the results of the literature search, we determined that this review article should consist of two main parts. One research stream focuses on interpreting or exploring pedagogical issues through the results of sentiment recognition. Furthermore, another research stream focuses on constructing sentiment datasets in educational settings and using techniques such as machine learning and artificial intelligence to improve the sentiment recognition accuracy based on these datasets. We only focus on the research literature in the second part.

III. FRAMEWORK OF FACIAL EXPRESSION RECOGNITION IN EDUCATIONAL RESEARCH

The purpose of facial expression recognition is to classify facial expressions using specific expression labels. In this review, we propose a systematic framework and technological process for facial expression recognition in educational research. Due to the differences in the application of facial expression recognition algorithms in educational research, the diversity of the data features to be recognized and the detailed analyzing the data to obtaining the classification results need to be clearly represented. Therefore, we focus on the following research questions: (1) How do researchers choose different types of facial expression recognition methods according to the needs of educational research? (2) Which machine learning methods have been applied to student facial expression recognition? (3) What are the main challenges in the current study? The general framework of the above research questions is shown in Figure1.

Specifically, it is easy to record students' facial expression data using a camera, both for online and offline teaching, which enables facial expression recognition research. First, the review summarizes two types of facial recognition methods from most of the current facial recognition studies and provides recommendations for their use. Second, traditional machine learning and deep learning methods used for student facial expression recognition are summarized. Finally, some key challenges of student facial expression recognition are presented.

IV. DIFFERENT TYPES OF FACIAL EXPRESSION RECOGNITION METHODS

Based on the summary of previous studies, we categorize the facial expression recognition methods in educational research into two major categories, namely, manual labeling-based methods and machine learning-based automatic facial expression recognition methods.

The manual labeling-based approach can be divided into self-report methods and observer-annotation methods. The self-report methods simply ask individuals to describe their nature and emotions [27], [28]. Although the use of the self-report method is easy and efficient, it is biased by the participants. According to [29], it has been found that the traditional methods have poor efficiency in many studies. One such study [30] demonstrated that students' self-reported emotions could be automatically inferred from the physiological data streamed to the tutoring software in real educational settings. Observer annotation is another manual labeling-based method in which the observer labels the subject. A previous study [31] showed that the average rating obtained from multiple observers improves the performance of the recognizer. However, annotating facial expressions is a complex task. Annotating facial expressions can be very time-consuming and challenging for psychologists. Although some databases use crowdsourcing for annotation [32], it is very difficult, expensive, and time-consuming to annotate facial expressions with a large dataset, because determining expression labels requires specific expertise and takes months of learning and refinement [33].

With the aim of correcting the deficiencies of the manual labeling-based approach, automatic recognition methods based on machine learning were developed. There are two main categories of machine learning-based automatic recognition methods in educational research. The first category contains algorithms designed specifically for the data to be classified. This class of algorithms based on the analysis of data features can achieve the desired recognition results but requires the researcher to have a certain technical background. The second category is recognition research using application programmable interfaces (APIs). Although the API approach does not require researchers to have as much of a technical background, there are obvious disadvantages. For example, one study [34] pointed out that when APIs are used for the facial expression recognition of students in a real classroom setting, factors such as the ambient lighting, camera image quality, occlusions, and variant poses cause problems, and there are some images that cannot be analysed. For a systematic review of the use of APIs for emotion detection in learning environments using facial expression recognition, please refer to [25].

From the above analysis, we can see that both types of facial expression recognition methods have their advantages and disadvantages. Which method is appropriate depends on the application. For example, when the user's perception and interpretation of the emotion being felt is important in the study, a self-report method or observer-annotation method should be selected [26]. However, the manual approach has limitations in terms of subjectivity and human cost, and it cannot be implemented on a large scale. Meanwhile, the automatic student sentiment recognition approach is an important learning analytics technique based on sentiment computing that has emerged in recent years and has a variety of important applications. In the next section, we will provide a detailed summary of the machine learning-based approach. In addition, it is advisable for the researcher to consider several methods to obtain a comprehensive representation of the user's emotional state, to ensure the consistency and accuracy of the collected data, and to increase the reliability of inferences [35].

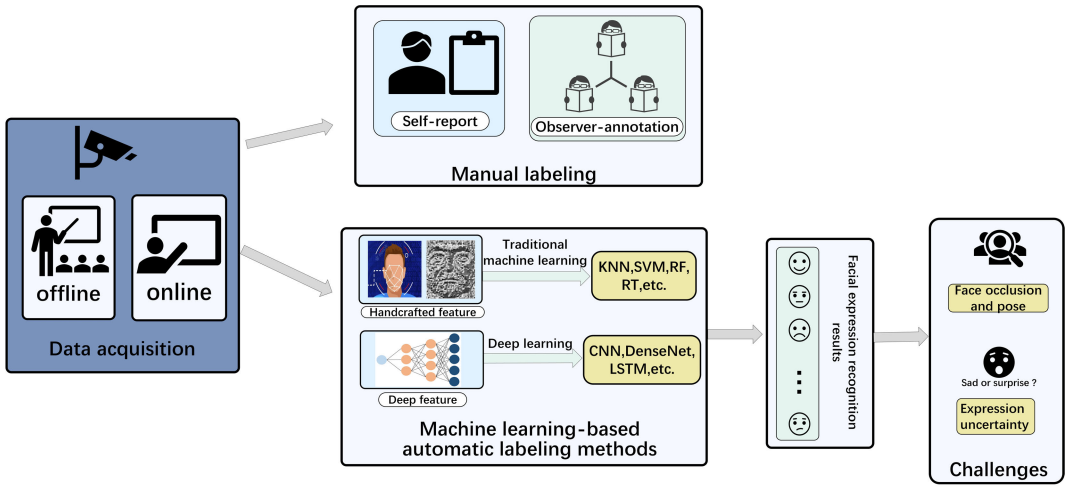


FIGURE 1. The overall framework of facial expression recognition in educational research.

V. MACHINE LEARNING-BASED METHODS

The machine learning-based facial expression recognition framework generally consists of three steps: the acquisition of faces, the extraction of facial expression features, and the output of facial expression classification results, as shown in Figure2.

Face acquisition is the first step in all face analysis-related tasks and typically involves face detection, face alignment, and cropping. It is worth noting that this step is particularly critical when studying facial expression recognition in real classroom settings. Traditional face detectors include the Viola-Jones detector [36] and the histogram of oriented gradient (HOG) [37] detector. As the handcrafted features used in traditional methods are not robust to different illumination conditions, occlusions, and poses, deep learning-based face detectors are more commonly used, such as the multi-task cascaded convolutional network (MTCNN) detector [37], [38], single-shot multi-boxes detector (SSD) [39], and you-only-look-once (YOLO) detector [40]. In order to eliminate the interference of planar rotation and translation, the faces are aligned uniformly using key point information and a pre-defined template and then cropped out. After obtaining the face region, the face image is converted into a feature vector through feature extraction using a handcrafted-based feature extractor [41] or a deep learning-based feature extractor [42]. Once the feature vectors are obtained, they are classified using a classifier such as the support vector machine (SVM) or the Softmax classifier.

From the literature selected for this review, without considering face acquisition strategies, we classify facial expression recognition methods into handcrafted-feature-based recognition methods and deep learning-based recognition methods according to the way in which facial expression features are acquired. The following subsections focus on reviewing traditional and recent advances in facial expression recognition methods. The handcrafted-feature-based recognition methods and deep learning-based recognition methods that have been successfully applied in educational research are described in Table1 to Table 6.

A. HANDCRAFTED FEATURE-BASED RECOGNITION METHODS

1) APPEARANCE-FEATURES EXTRACTION

For handcrafted features, appearance-based features and geometric-based features can be used. Appearance features generally reflect the texture structure of the human face. Appearance features include local binary patterns (LBP) features, Gabor features, pyramid gradient histogram features, scale-invariant feature transform (SIFT) features, etc. For example, appearance-based features are used in [43] and [51]. Specifically, Akputu et al. [43] used Gabor filtering for feature extraction. Following the Gabor feature extraction, multiple kernel learning was introduced for facial expression recognition in E-learning environments. Musa [44] used a grey-level co-occurrence matrix (GLCM) to extract texture description features and applied them to support vector regression (SVR) to recognize the facial emotion of a student during a learning session. In general,these methods utilize various filters and image processing techniques to extract relevant features. In an educational context, appearance-based methods are valuable for their simplicity and real-time processing capabilities. However, they may struggle to capture subtle nuances in facial expressions and might be sensitive to lighting and pose variations.

2) GEOMETRIC-FEATURES EXTRACTION

Geometric-based features describe the face using shape and distance information, which can reflect the expression information to a certain extent. For example, some studies [45], [46], [64] used the active appearance model (AAM) [65] to identify the landmark points or the feature points.

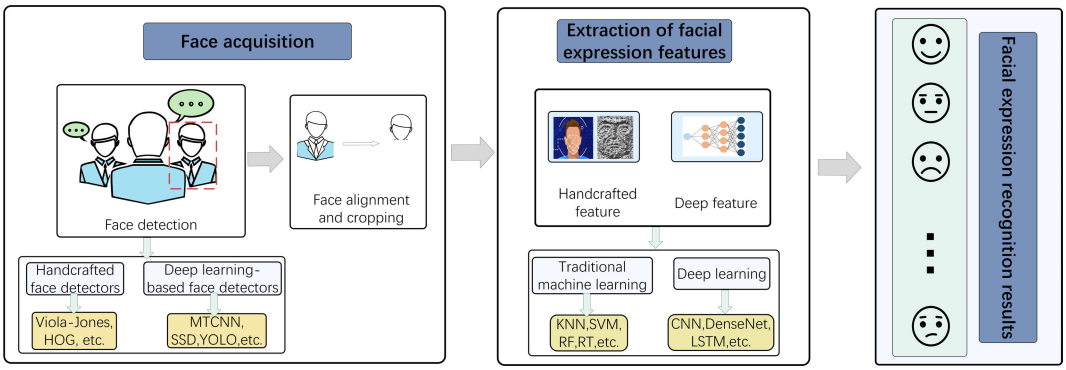


FIGURE 2. The standard framework for machine learning-based facial expression recognition.

The facial shape is defined by a shape vector containing the coordinates of landmarks, including some key distances and measurements, such as the distance between eyelids, the distance to the mouth, and the width of the mouth. Then, the AAM is used to create a facial emotion model for classification. Another studies [47], [66] used FaceTracker [67] to extract geometric-based features. FaceTracker gathers facial information based on 33 mapped feature points: eight points around the mouth, six on each eye, three on each eyebrow and the chin, two on the nostrils, and two outlining the lateral extremities of the face. In summary,these methods provide interpretable insights into facial muscle movements. In education, geometric-based approaches can offer detailed information about students' expressions, aiding in the assessment of their engagement. Nevertheless, they may require precise alignment and struggle with occluded facial areas. Different facial expression recognition methods based on appearance-features and geometric-features in educational research are shown in Table 1.

3) MULTI-FEATURES EXTRACTION

In addition to the two types of features described above, some methods use a combination of these features. For example, multiple texture features were implemented by, among others, [48], [49], [51], and [68]. Specifically, in [48] and [49], smile recognition in classroom was carried out by extracting multiple texture features, namely Gabor, LBP, and image intensity features, from patches in the mouth and eye regions, as a component of multimodal concentration detection. Another study [51] employed fused features based on Gabor and LBP features that were shown to be very robust to illumination changes and misalignment in the classroom environment. In short, these methods attempt to mitigate the limitations of individual approaches, providing a more holistic understanding of students' emotional states in educational scenarios. However, the complexity of combining multiple features might lead to increased computational costs and require substantial training data. Different facial expression recognition methods based on multi-features in educational research are shown in Table 2.

4) CLASSIFIER

Typically, after extracting the characteristics of these designs, an efficient feature classifier is required to generalize emotion classes to reflect the feature diversity; patterns are often used. Classification methods such as the K nearest neighbor (KNN) method, support vector machine (SVM), and multilayer perceptron (MLP) have been utilized extensively. Ayvaz et al. [69] developed a facial emotion recognition system (FERS), which recognizes the emotional states and motivation levels of students in video conference-type E-learning settings. The system uses four machine learning algorithms (SVM, KNN, random forest and classification, and regression trees), and the best accuracy rates were obtained using the KNN and SVM algorithms. Another studies [47], [70] used ensemble classifiers to enhance the precision of emotion classification. Multiple kernel learning (MKL) was shown to boost the classification performance by using multiple kernels rather than a single fixed kernel. The authors of [43] and [50] proposed using an MKL decision tree with weighted kernel alignment (WFA) for facial emotion recognition for educational learning.

In studies such as those referenced, a combination of these classification algorithms was employed to bolster facial emotion recognition in educational settings. The choice of classifier depends on factors such as data complexity, computational resources, and the desired balance between accuracy and efficiency. Furthermore, the trend towards ensemble methods and the use of MKL underscores the inclination to harness multiple classifiers or kernels to achieve heightened accuracy and address individual classifier limitations. These approaches, however, necessitate careful parameter tuning and resource considerations to attain optimal results.

The choice of handcrafted feature-based recognition method should be guided by the specific educational context and research goals. Each method category presents distinct advantages and limitations, making a comprehensive and context-aware approach essential for successful implementation in educational settings. Further research can explore hybrid approaches that leverage the strengths of multiple methods to enhance the accuracy and applicability of facial expression recognition in education.

TABLE 1. Different facial expression recognition methods based on appearance-features and geometric-features in educational research.

Literature	Machine learning-based method	Algorithm	Aim and scope	Learning environment
[43]	Appearance-feature extraction	Gabor filter+Multiple Kernel Learning	To explore the potential of utilizing affect or emotion recognition research in Adaptive Educational Hypermedia E-learning models.	E-learning
[44]	Texture description feature extraction	LGCM+SVR	To identify the facial expressions of university students in the classroom during the class session using the static frontal face image.	Classroom
[45]	Geometric-features extraction	AAM+ SVM	The system recommends an improvised teaching strategy for a group of eLearning students, which makes eLearning a better educational system.	E-learning
[46]	Geometric-features extraction	AAM+ Manually coding	To estimate the learners' facial expressions, which were then used to reveal the relationships between the social interactive process and learning performance while using a tutoring system.	None specified
[47]	Geometric-features extraction	FaceTracker+ Ensemble classifiers	Using emotion recognition to assess simulation-based learning.	None specified

TABLE 2. Different facial expression recognition methods based on multi-features in educational research.

Literature	Machine learning-based method	Algorithm	Aim and scope	Learning environment
[48]	Multi-features extraction	Gabor and LBP + K-means based voting	To detect student's interest and make them more engaged in the learning process for productive learning.	Classroom
[49]	Multi-features extraction	Gabor and LBP + K-means based voting	A 3D learning interest model designed from the perspective of educational psychology describes students' interests in the learning environment.	Classroom
[50]	Multi-features extraction	PCA+LDA+Gabor and MKL	The emotion classification performance on this study has shown good merits and prospects for future affective educational learning frameworks.	None specified
[51]	Multi-features extraction	Gabor and LBP+KNN	In order to solve the problem of high cost and low efficiency caused by employing human analysts to observe classroom teaching effect.	Classroom

B. DEEP LEARNING-BASED RECOGNITION METHODS

"Good features" are a vital part of an emotion recognition system and deep learning [71] has proved to be a very effective feature extraction method. Deep learning-based recognition methods are typically combined to perform the end-to-end training of feature representations and classifiers (usually a full connection layer + Softmax) at the back end of a given task objective. In 2011, Turing Award winners

Hinton et al. used the deep belief network (DBN) and gated Markov random field (MRF) for feature learning and classification with an SVM, and they achieved the best results on the CohnKanade (CK) database [72], [73]. Later, given that the convolutional neural network (CNN) greatly outperformed traditional feature design methods in the ImageNet competition [74], deep learning was widely applied in facial expression recognition and educational research.

1) CNN AND DENSENET

Since 2016, CNNs [75], [76], [77], [78], [79], [80], [81], [82] and DenseNet [63] have been applied to facial emotion recognition in the field of education. For example, one study [52] introduced a three-dimensional DenseNet self-attention neural network (3D DenseAttNet) for the automatic detection of students' engagement in E-learning platforms. The self-attention module helps to extract only the relevant high-level intra-inter-frame-dependency features of videos obtained from the 3D DenseNet block. Recently, transformers have been successfully and widely used in natural language processing and computer vision. The transformer has an encoder-decoder architecture with a self-attention mechanism, and each part of the data is weighted differently according to its importance. With the recent success of transformers in natural language processing, another study [53] focused on the development of transformers to replace traditional CNN architectures in facial expression recognition in education. In educational settings, these methods offer high accuracy in recognizing facial expressions, which can aid educators in assessing students' emotional responses. However, CNNs often require substantial amounts of labeled data for training, and their interpretability can be limited. Different facial expression recognition methods based on CNN and DenseNet in educational research are shown in Table 3.

2) RNN

Compared with the CNN, the recurrent neural network (RNN) is more suitable for time series data because it is able to connect previous information to the current state. However, a difficulty exists in the training of the RNN; there are vanishing and exploding gradient problems, which hamper the learning of long data sequences. Long short-term memory (LSTM) and gated recurrent unit (GRU) networks are special RNNs that are capable of learning long-term dependencies. Thus, they are commonly used to model long-term sequence data, and they have also been applied in video-based facial expression analysis in educational research [2], [54], [83]. For example, Huang et al. [2] proposed a deep engagement recognition network (DERN) that combines temporal convolution, a bidirectional LSTM, and an attention mechanism to identify the degree of engagement based on facial features. In education, RNNs can help track students' emotional trajectories over time. However, they may struggle with long-range dependencies and suffer from vanishing gradient problems during training. Different facial expression recognition methods based on RNN in educational research are shown in Table 4.

3) HYBRID NETWORK MODEL

In addition to the above network models, some hybrid network models have also been proposed. These models can be divided into deep learning-based multi-network hybrid models [55], [59] and manual-feature-and-depth-feature-combination hybrid networks [1], [56], [58], [84]. One study [55] proposed a model for continuous facial emotional pattern recognition that combines a CNN, LSTM, and facial emotion recognition. The CNN and LSTM are combined to perform deep learning to recognize and analyze the continuous facial emotional patterns of students and thus recognize emotions. Another study [57] proposed an effective technique for detecting students' facial expressions that combines discrete Chebyshev wavelet transformations (DCHWTs) with mathematical methods and a convolutional neural network (ChWCNN). For educational applications, hybrid models offer a holistic understanding of students' emotional responses, but designing and training such models can be complex and computationally intensive. Different facial expression recognition methods based on hybrid network model in educational research are shown in Table 5.

4) EFFICIENT NEURAL NETWORK

In order to ensure that network models can quickly and accurately evaluate students' emotions in practical applications, an efficient network model and efficient training strategies are essential. The existing literature mainly studies the following two important strategies: the use of lightweight network models and the use of transfer learning methods.

Researchers have designed several lightweight neural network architectures, such as the Inception [85], ShuffleNet [86], and MobileNet [87] architectures, which have the potential to create highly efficient deep networks with fewer calculations and parameters; they have been applied to facial emotion recognition (FER) in educational research in recent years [60], [61], [62]. For example, one study [62] proposed an optimized lightweight convolutional neural network (CNN) model for engagement recognition using facial expressions within a distance-learning setup; this model can easily be adapted for mobile platforms and deliver an outstanding performance. In education, efficient models enable facial expression recognition in resource-constrained environments. However, there might be a trade-off between efficiency and recognition accuracy.

The transfer learning method is an effective model training method [88]. Fine-tuning is a general method used in transfer learning. In FER, researchers generally pre-train the network for expression recognition by using publicly available facial expression datasets. These networks can then be fine-tuned based on any other FER dataset to accurately predict emotions in educational research. One study [63]

TABLE 3. Different facial expression recognition methods based on CNN and DenseNet in educational research.

Literature	Machine learning-based method	Algorithm	Aim and scope	Learning environment
[63]	Deep features	CNN+Transfer learning	The system identifying learners' facial emotions can provide feedback that teachers can understand students' learning situation and provide help or improve teaching strategy.	None specified
[56]	Deep features	CNN and fuzzy system	This approach should be further tested and evaluated on a larger number of individuals in several learning domains.	E-Learning
[59]	Deep features	CNN	This architecture uses the students' facial expressions, hand gestures and body postures for analyzing their affective states.	Classroom
[52]	Deep features	DenseNet+ Self-Attention	To identify and evaluate student participation in modern and traditional educational programs.	None specified
[53]	Deep features	Transformer	In order to improve human and computer interfaces, and enhance the feedback mechanism actions taken by computers from the users.	None specified

TABLE 4. Different facial expression recognition methods based on RNN in educational research.

Literature	Machine learning-based method	Algorithm	Aim and scope	Learning environment
[2]	Deep features	Temporal convolution, bidirectional LSTM and attention mechanism	By identifying the degree of learners' engagement in online learning environment, it can help instructors get feedback in time and help the online learning platform recommend appropriate learning resources.	Online learning
[55]	Deep features	CNN and LSTM	By noting students' academic emotions, teachers can provide the most suitable teaching material according to the emotions to improve their academic performance and motivation.	None specified
[54]	Deep features	Multilayer Bi-LSTM	This work is the first very stage for reliable automatic engagement estimation in online learning with taking into account the sequential characteristic of engagement state.	Online learning

used transfer learning to build a learning emotion recognition model and verify the recognition accuracy. The effectiveness of efficient neural networks, especially when using transfer learning, heavily relies on the availability of high-quality labeled data. In educational research, obtaining large and diverse datasets for training can be a challenge, potentially limiting the full potential of these models. Different facial expression recognition methods based on efficient neural network in educational research are shown in Table 6.

In conclusion, efficient neural networks, when paired with strategies like lightweight models and transfer learning, offer significant advantages for real-time emotion recognition in educational applications. They are efficient, quick, and

TABLE 5. Different facial expression recognition methods based on hybrid network model in educational research.

Literature	Machine learning-based method	Algorithm	Aim and scope	Learning environment
[55]	Deep features	CNN and LSTM	By noting students' academic emotions, teachers can provide the most suitable teaching material according to the emotions to improve their academic performance and motivation.	None specified
[56]	Deep features	CNN and fuzzy system	This approach should be further tested and evaluated on a larger number of individuals in several learning domains.	E-Learning
[59]	Deep features	CNN	This architecture uses the students' facial expressions, hand gestures and body postures for analyzing their affective states.	Classroom
[57]	Appearance-features+Deep features	DCHWT+CNN	Predict student' conditions or degree of contentment by identifying student' emotions while learning from instructors.	Classroom
[58]	handcrafted-features+Deep features	HOG+CNN	This method can be performed in real-time so that this can be used in any learning environment.	Smart classroom

adaptable to various settings. However, their performance may come at the cost of slightly reduced accuracy compared to more complex models. Balancing efficiency and accuracy, along with addressing data limitations, remains a key challenge in implementing efficient neural networks effectively for educational research.

Deep learning-based facial expression recognition has revolutionized the field of educational research, offering profound insights into students' emotional states and engagement. The selection of a suitable deep learning approach should be guided by the specific educational context and research objectives. Each method category presents distinct advantages and challenges, underscoring the need for a well-informed and context-sensitive approach to implementation. Future research directions could explore the customization of deep learning models for specific educational scenarios, striking a balance between accuracy, efficiency, and interpretability.

C. DATASET FOR FACIAL EXPRESSION RECOGNITION IN EDUCATION LEARNING

Due to privacy considerations, many algorithms are tested on non-public datasets [51], [52], [89], [90], [91]. Publicly available databases of students' emotions are limited. Table3 summarizes the currently available facial expression datasets based on educational research.

In 2014, Whitehill et al. introduced the HBCU dataset [92] to determine student engagement from facial expressions. In 2016, Kamath et al. [93] constructed a dataset of learner engagement in massive open online courses (MOOCs). Subsequently, the dataset for affective states in E-environments (DAiSEE) was introduced [75]; this dataset captures the real-world challenges of recognising user engagement in natural settings. Due to the advent of deep learning frameworks, larger databases representing diverse settings are required. Kaur et al. [94] introduced the EngageWild dataset for common benchmarking and the development of engagement assessments in diverse 'in-the-wild' conditions. This dataset was also presented as a sub-challenge in EmotiW2019 [98] and EmotiW2020 [99]. Due to the high cost of validation by psychological experts, the frames in the DAiSEE and EngageWild datasets were labelled based on crowdsourcing, whereas in the HBCU dataset, the frames were labelled by human experts. In both cases, ambiguity in labelling frequently occurs due to the lack of clear guidelines for mapping facial indicators to different affective states or engagement levels of the online learners [100].

There are very few datasets based solely on students' facial expressions. For example, Bian et al. [95] established an online learning spontaneous facial expression database (OL-SFED) for emotion inference in education. In educational research, researchers usually collect multimodal data for follow-up research due to the collection cost; however, most of these multimodal datasets contain a database of only facial expressions. For example, Sun et al. presented BNU-LSVED 1.0 [101], a spontaneous multimodal database developed in an actual learning environment. However, the emotions were stimulated by videos instead

TABLE 6. Different facial expression recognition methods based on efficient neural network in educational research.

Literature	Machine learning-based method	Algorithm	Aim and scope	Learning environment
[60]	Deep features	Inception v3	Provide the students' affect feedback of each classroom to the expert faculty member to improve the teaching style and strategy of the newly joined instructor or the teaching assistant.	Smart classroom
[61]	Deep features	EfficientNet and MobileNet	This framework can be integrated into existing E-learning tools for fast and accurate assessment of students' emotions and comprehension.	Online learning
[62]	Deep features	ShuffleNet v2	This model is suitable for student engagement recognition for distance learning on mobile platforms.	E-learning
[63]	Deep features	CNN+Transfer learning	The system identifying learners' facial emotions can provide feedback that teachers can understand students' learning situation and provide help or improve teaching strategy.	None specified

TABLE 7. Some available facial expression datasets in educational research.

Dataset	Video/image	Number of subjects	Emotions
HBCU [92]	120 Videos	34 (9 male and 25 female)	Not engaged, Nominally engaged, Engaged, Very engaged
WACV2016 [93]	4,408 images	23	Not engaged, Nominally engaged, Very engaged
DAiSEE [75]	9,068 Videos	112 (80 male and 32 female)	Engaged, Frustration, Boredom, Confusion
EngageWild [94]	264 Videos	91 (64 male and 27 female)	Disengaged, Barely engaged, Normally engaged, Highly engaged
OL-SFED [95]	1,274 Videos /30,184 images	82 (29 male and 53 female)	Enjoyment, Confusion, Fatigue, Distraction, Neutral
BNU-LSVED2.0 [96]	2,117 Videos	81 (31 male and 50 female)	Bored, Tired, Confused, Concentrated, Interested, Happy, Thoughtful, Relaxed, Assentient, In a daze
Affective [97]	72,000 images	350	Happiness, Surprise, Delight, Neutral,Sadness, Fear,Disgust,Engaged,Sleepy, Boredom, Frustrated, Confused
RFAU [33]	3,325 Videos /256,220 images	1,796	6-level intensity of 12 action units.

of by the actual speech of teachers conducting classes. Subsequently, Wei et al. proposed BNU-LSVED2.0 [96], which is the first large-scale spontaneous and multimodal student affect database created in a classroom environment. On the basis of BNU-LSVED2.0, BNU-SDED was created. BNU-SDED [102] was created by performing keyframe image selection in BNU-LSVED 2.0. The study [102] analysed the BNU-LSVED2.0 and BNU-SDED datasets. In 2020, Ashwin and Guddeti [97] proposed a new affective database for both E-learning and classroom environments that includes students' facial expressions, hand gestures, and body postures. It contains a database of only facial expressions.

Hu et al. [33] presented a manually annotated facial action unit database (RFAU) of the expressions of students in real classrooms. This dataset describes the students' facial expressions with action units and detailed intensities.

VI. CHALLENGES AND FUTURE DIRECTIONS

In terms of the practical accuracy of facial expression recognition in an uncontrolled environment, facial expression recognition is far from reaching the accuracy needed in educational research at present. In this paper, it is found that the main problems lie in the face occlusion and pose problems in uncontrolled environments and the uncertainty of the expressions shown in the captured images. Each of these problems and potential solutions are described below.

A. FACE OCCLUSION AND POSE PROBLEMS IN UNCONTROLLED ENVIRONMENT

Occlusions in a real environment are usually people's hands, hair, eyeglasses, and other objects with irregular shapes and indeterminate colours. The RFAU dataset demonstrates that the problem of occlusion is universal [33]. The RFAU dataset marks occlusions and faces with eyeglasses. In the classroom environment, the recognition of students' facial expressions is greatly affected by face occlusions and posture. The traditional method for studying the occlusion problem of facial expression recognition is to use the artificial setting of local black blocks [103], which is not necessarily suitable for the real occlusion situation, which may involve wearing goggles, wearing masks, etc. Since the emotion analysis of occluded faces is a field in its infancy, there are few studies that specifically focus on occluded face emotion recognition in the field of educational research.

In order to mitigate the influence of occlusions and posture on facial expression recognition, the use of local face information is generally recognized as an effective strategy. Li et al. [104] and Wang et al. [105] used the local block attention mechanism at the feature level and image level, respectively, to improve the robustness of the model to occlusion and attitude. Another possible strategy is to use large-scale face recognition data to first learn a robust face recognition model for posture and occlusion, and then fine-tune it for facial expression recognition. Ding et al. [106] used the VGGFace2 database to pre-train their facial recognition model, and the final facial expression recognition performance was about 2% better than that of models pre-trained using other databases, indicating that it is feasible to first learn a face recognition model that is robust to different postures and occlusions.

B. EXPRESSION UNCERTAINTY PROBLEM

For basic expression recognition, ambiguous expressions, low-quality expression images, and the labeller's emotional subjectivity mean that the expression category is not uniquely determined in many cases, that is, the expression is uncertain [107]. It is pointed out in [100] that since some of the published datasets are labelled by experts and some are labelled by crowdsourcing, there is no clear criterion that can be used to map facial indicators to the different emotional states of online learners, so there are often ambiguities in the labels. However, there are few studies on this issue in educational studies based on students' facial expressions.

In order to alleviate the expression uncertainty, Zeng et al. [108] and Wang et al. [107] conducted preliminary explorations, respectively. Zeng et al. [108] used the prediction results of the deep learning model on multiple databases to assist in training on potential correct tags to improve the robustness of feature learning. Wang et al. [107] used a self-attention mechanism and a relabelling scheme in each batch to suppress a part of the facial uncertainty. One study [100] pointed out that in order to alleviate the expression uncertainty problem, frames with fuzzy marks are usually deleted during the experiment, which will eventually reduce the size of the dataset and eliminate the diversity of the information in the dataset. In addition, visual cues, user activities, self-evaluation, and transfer learning in the learning environment can be further studied to solve the above problems.

C. PARAMETER TUNING AND OVERFITTING ISSUES

Parameter tuning and overfitting issues are common technical challenges in algorithm implementation. In this section, we will address the topic of hyperparameter tuning, including approaches to avoid overfitting issues and a summary of relevant research methods. Overfitting is a common problem in machine learning where a model performs well on the training data but fails to generalize well on new, unseen data. Due to the limited availability of datasets in the education domain, and the relatively smaller number of publicly available datasets, the quantity of datasets that can be utilized for research is considerably lower compared to other domains. Given this constraint, many researchers in the field of education facial expression recognition have encountered overfitting issues when working with limited datasets [52], [60], [62]. Here are some common methods to detect overfitting in facial expression recognition in educational research:

Train-Test Split: One common approach is to divide the dataset into two parts: a training set and a test set. The model is trained on the training set, and its performance is then evaluated on the test set. If the performance is significantly worse on the test set compared to the training set, it might indicate overfitting [52].

Validation Set Performance Monitoring: During the training process, monitoring the performance of the model on a separate validation set can help detect overfitting. If the validation performance plateaus or starts to decline, it may indicate overfitting [60].

Cross-Validation: Cross-validation is a technique used to assess how well a model generalizes to new data. It involves dividing the data into multiple subsets (folds), and iteratively training the model on different subsets while validating

it on the remaining data. Consistent performance discrepancies between training and validation sets can indicate overfitting [62]. While cross-validation is a valuable model validation technique, we also acknowledge that it may not be sufficient in all cases. Therefore, we recommend employing a combination of multiple validation methods to thoroughly assess the model's performance and ensure its reliability.

The proposal technique for solving the overfitting problem as following:

Regularization: Applying regularization techniques, such as L1 or L2 regularization, can help prevent overfitting by penalizing complex models [61].

Feature Selection: Overfitting can occur when the model learns noise or irrelevant features. Conducting feature selection or engineering can help prevent this [47], [51].

Ensemble Methods: Using ensemble techniques like bagging or boosting can also help reduce overfitting by combining multiple models' predictions [47].

By implementing these techniques, it is possible to improve the generalization ability of the model and ensure its effectiveness in educational facial expression recognition. In summary, these methods can serve as technical guidance for future research on facial expression recognition in the field of education.

VII. DISCUSSION, CONCLUSION, AND FUTURE WORK

The systematic review presented in this paper aimed to provide a comprehensive overview of facial expression recognition methods in educational research from a machine learning perspective. We addressed the gaps in existing literature by outlining tasks, progress, challenges, and future trends related to facial expression analysis.

Our review identified two primary educational research methods based on facial expression recognition: manual labeling method and machine learning-based automatic labeling method. Manual labeling method focuses on the user's perception and interpretation of the emotion, which is mainly divided into self-report method and observer annotation method, while automatic annotation methods based on machine learning emphasize effective automatic recognition on large data sets. By categorizing these methods and detailing their application scenarios, we aimed to provide educators and researchers with a clear understanding of how facial expression analysis can be integrated into educational settings.

Furthermore, our review highlighted the absence of a comprehensive framework for facial expression recognition in educational research. This observation underscores the need for researchers to develop a unified framework that can guide future studies in this field. The lack of a structured framework may hinder the scalability and reproducibility of research efforts.

We also addressed the gap in the literature regarding the utilization of the latest machine learning methods in facial expression analysis for educational research. Our review introduced various machine learning techniques and discussed their suitability for this context. This inclusion is pivotal, as it provides researchers with insights into cutting-edge methodologies that can enhance the accuracy and efficiency of facial expression recognition systems in educational settings.

The implications of this systematic review are twofold. Firstly, it offers educators valuable insights into how facial expression analysis can be leveraged to improve teaching styles and strategies. By understanding students' affect states, faculty members can adapt their teaching methods to cater to the emotional needs of their students, ultimately enhancing the learning experience. This has the potential to positively impact the learning rates of all students present in a classroom.

Secondly, our proposed overall framework for facial expression recognition in educational research has the potential to standardize the methodology employed in this field. A standardized framework could lead to greater consistency in research practices, making it easier to compare and replicate studies. This is essential for the continued growth and advancement of the field.

In conclusion, this paper has contributed to the field of facial expression analysis in educational research by providing a comprehensive overview, proposing an overall framework, introducing machine learning methods, and highlighting key challenges. By addressing these aspects, we aim to inspire future research efforts that will further advance our understanding of how facial expression analysis can be effectively applied in educational settings. This, in turn, can lead to improved teaching strategies and ultimately benefit the educational experience of students worldwide.

In future research, there is a vast space worth exploring. At present, although facial expression analysis in uncontrolled natural environments has developed rapidly, many problems and challenges remain to be solved. Facial expression analysis is a practical task. The future development of this field should not only consider the accuracy of the method but also pay attention to the time and storage consumption of the method. Because facial expressions are usually complex and diverse, it is difficult to describe them with a single label, so the problem of multi-label facial expressions should be given more attention in the future.

REFERENCES

[1] K. P. Rao and M. Rao, "Recognition of learners' cognitive states using facial expressions in e-learning environments," *J. Univ. Shanghai Sci. Technol.*, pp. 93–103, 2020.

[2] T. Huang, Y. Mei, H. Zhang, S. Liu, and H. Yang, "Fine-grained engagement recognition in online learning environment," in *Proc. IEEE 9th Int. Conf. Electron. Inf. Emergency Commun. (ICEIEC)*, Jul. 2019, pp. 338–341.

[3] M. Li, L. Chen, M. Wu, W. Pedrycz, and K. Hirota, "Dynamic expression recognition-based quantitative evaluation of teaching validity using valence-arousal emotion space," in *Proc. 13th Asian Control Conf. (ASCC)*, May 2022, pp. 1079–1083.

[4] R. J. Wlodkowski and M. B. Ginsberg, *Enhancing Adult Motivation to Learn: A Comprehensive Guide for Teaching all Adults*. Hoboken, NJ, USA: Wiley, 2017.

[5] R. Pekrun, T. Goetz, A. C. Frenzel, P. Barchfeld, and R. P. Perry, "Measuring emotions in students' learning and performance: The achievement emotions questionnaire (AEQ)," *Contemp. Educ. Psychol.*, vol. 36, no. 1, pp. 36–48, Jan. 2011.

[6] A. R. Artino, "Think, feel, act: Motivational and emotional influences on military students' online academic success," *J. Comput. Higher Educ.*, vol. 21, no. 2, pp. 146–166, Aug. 2009.

[7] L. M. Daniels, R. H. Stupnisky, R. Pekrun, T. L. Haynes, R. P. Perry, and N. E. Newall, "A longitudinal analysis of achievement goals: From affective antecedents to emotional effects and achievement outcomes," *J. Educ. Psychol.*, vol. 101, no. 4, pp. 948–963, Nov. 2009.

[8] R. Pekrun, A. J. Elliot, and M. A. Maier, "Achievement goals and achievement emotions: Testing a model of their joint relations with academic performance," *J. Educ. Psychol.*, vol. 101, no. 1, pp. 115–135, Feb. 2009.

[9] D. Singh, M. Kaur, M. Y. Jabarulla, V. Kumar, and H.-N. Lee, "Evolving fusion-based visibility restoration model for hazy remote sensing images using dynamic differential evolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 1002214.

[10] Z. Sabir, M. A. Z. Raja, S. E. Alhazmi, M. Gupta, A. Arbi, and I. A. Baba, "Applications of artificial neural network to solve the nonlinear COVID-19 mathematical model based on the dynamics of SIQ," *J. Taibah Univ. Sci.*, vol. 16, no. 1, pp. 874–884, Dec. 2022.

[11] A. Arbi, J. Cao, M. Es-saiydy, M. Zarhouni, and M. Zitane, "Dynamics of delayed cellular neural networks in the stepanov pseudo almost automorphic space," *Discrete Continuous Dyn. Syst. S*, vol. 15, no. 11, p. 3097, 2022.

[12] A. Arbi and N. Tahri, "Almost anti-periodic solution of inertial neural networks model on time scales," in *Proc. MATEC Web Conf.*, vol. 355, 2022, p. 02006.

[13] A. Arbi and N. Tahri, "Stability analysis of inertial neural networks: A case of almost anti-periodic environment," *Math. Methods Appl. Sci.*, vol. 45, no. 16, pp. 10476–10490, Nov. 2022.

[14] Z. Sabir, S. Saoud, M. A. Z. Raja, H. A. Wahab, and A. Arbi, "Heuristic computing technique for numerical solutions of nonlinear fourth order Emden–fowler equation," *Math. Comput. Simul.*, vol. 178, pp. 534–548, Dec. 2020.

[15] T. Tabassum, A. A. Allen, and P. De, "Non-intrusive identification of student attentiveness and finding their correlation with detectable facial emotions," in *Proc. ACM Southeast Conf.*, Apr. 2020, pp. 127–134.

[16] S. K. D'Mello, S. D. Craig, A. Witherspoon, B. McDaniel, and A. Graesser, "Automatic detection of learner's affect from conversational cues," *User Model. User-Adapted Interact.*, vol. 18, nos. 1–2, pp. 45–80, Feb. 2008.

[17] J. Chen, N. Luo, Y. Liu, L. Liu, K. Zhang, and J. Kolodziej, "A hybrid intelligence-aided approach to affect-sensitive e-learning," *Computing*, vol. 98, nos. 1–2, pp. 215–233, Jan. 2016.

[18] A. Mehrabian, *Communication Without Words*, vol. 2, no. 4. New York, NY, USA: Psychology Today, 1968.

[19] E. Friesen and P. Ekman, "Facial action coding system: A technique for the measurement of facial movement," *Palo Alto*, vol. 3, no. 2, p. 5, 1978.

[20] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 6, pp. 1113–1133, Jun. 2015.

[21] M. Pantic and L. J. M. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1424–1445, Dec. 2000.

[22] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1195–1215, Sep. 2022.

[23] E. Yadegaridehkordi, N. F. B. M. Noor, M. N. B. Ayub, H. B. Affal, and N. B. Hussin, "Affective computing in education: A systematic review and future research," *Comput. Educ.*, vol. 142, Dec. 2019, Art. no. 103649.

[24] J. Zhou and J.-M. Ye, "Sentiment analysis in education research: A review of journal publications," *Interact. Learn. Environments*, vol. 31, no. 3, pp. 1–13, 2020.

[25] M. Bustos-López, N. Cruz-Ramírez, A. Guerra-Hernández, L. N. Sánchez-Morales, and G. Alor-Hernández, "Emotion detection in learning environments using facial expressions: A brief review," in *Handbook on Decision Making*, 2023, pp. 349–372.

[26] M. Imani and G. A. Montazer, "A survey of emotion recognition methods with emphasis on e-learning environments," *J. Netw. Comput. Appl.*, vol. 147, Dec. 2019, Art. no. 102423.

[27] I. Lopatovska and I. Arapakis, "Theories, methods and current research on emotions in library and information science, information retrieval and human–computer interaction," *Inf. Process. Manage.*, vol. 47, no. 4, pp. 575–592, Jul. 2011.

[28] C. Lacave, J. Á. Velázquez-Iturbide, M. Paredes-Velasco, and A. I. Molina, "Analyzing the influence of a visualization system on students' emotions: An empirical case study," *Comput. Educ.*, vol. 149, May 2020, Art. no. 103817.

[29] F. Camelia and T. L. J. Ferris, "Validation studies of a questionnaire developed to measure students' engagement with systems thinking," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 48, no. 4, pp. 574–585, Apr. 2018.

[30] I. Arroyo, D. G. Cooper, W. Burleson, B. P. Woolf, K. Muldner, and R. Christopherson, "Emotion sensors go to school," in *Artificial Intelligence in Education*. Amsterdam, The Netherlands: IOS Press, 2009, pp. 17–24.

[31] K. P. Truong, M. A. Neerincx, and D. A. Van Leeuwen, "Assessing agreement of observer-and self-annotations in spontaneous multimodal emotion data," Tech. Rep., 2008.

[32] S. Li and W. Deng, "Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 356–370, Jan. 2019.

[33] Q. Hu, C. Mei, F. Jiang, R. Shen, Y. Zhang, C. Wang, and J. Zhang, "RFAU: A database for facial action unit analysis in real classrooms," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1452–1465, Jul. 2022.

[34] G. Tonguç and B. O. Ozkara, "Automatic recognition of student emotions from facial expressions during a lecture," *Comput. Educ.*, vol. 148, Apr. 2020, Art. no. 103797.

[35] J. J. Vogel-Walcutt, L. Fiorella, T. Carper, and S. Schatz, "The definition, assessment, and mitigation of state boredom within educational settings: A comprehensive review," *Educ. Psychol. Rev.*, vol. 24, no. 1, pp. 89–111, Mar. 2012.

[36] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, May 2004.

[37] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jul. 2005, pp. 886–893.

[38] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

[39] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 21–37.

[40] 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.*, Jun. 2016, pp. 779–788.

[41] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image Vis. Comput.*, vol. 27, no. 6, pp. 803–816, May 2009.

[42] S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, Ç. Gülçehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, and R. C. Ferrari, "Combining modality specific deep neural networks for emotion recognition in video," in *Proc. 15th ACM Int. Conf. Multimodal Interact.*, 2013, pp. 543–550.

[43] O. K. Akputu, K. P. Seng, Y. Lee, and L.-M. Ang, "Emotion recognition using multiple kernel learning toward E-learning applications," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 14, no. 1, pp. 1–20, Feb. 2018.

[44] N. H. B. Musa, "Facial emotion detection for educational purpose using image processing technique," Tech. Rep., 2020.

[45] T. Ashwin, J. Jose, G. Raghu, and G. R. M. Reddy, "An e-learning system with multifacial emotion recognition using supervised machine learning," in *Proc. IEEE 7th Int. Conf. Technol. Educ. (TE)*, Sep. 2015, pp. 23–26.

[46] Y. Hayashi, "Detecting collaborative learning through emotions: An investigation using facial expression recognition," in *Proc. Int. Conf. Intell. Tutoring Syst.* Cham, Switzerland: Springer, 2019, pp. 89–98.

[47] L. Y. Mano, A. Mazzo, J. R. T. Neto, M. H. G. Meska, G. T. Giancristofaro, J. Ueyama, and G. A. P. Júnior, "Using emotion recognition to assess simulation-based learning," *Nurse Educ. Pract.*, vol. 36, pp. 13–19, Mar. 2019.

[48] Z. Luo, L. Liu, J. Chen, Y. Liu, and Z. Su, "Spontaneous smile recognition for interest detection," in *Proc. Chin. Conf. Pattern Recognit.* Cham, Switzerland: Springer, 2016, pp. 119–130.

[49] Z. Luo, C. Jingying, W. Guangshuai, and L. Mengyi, "A three-dimensional model of student interest during learning using multimodal fusion with natural sensing technology," *Interact. Learn. Environments*, vol. 30, no. 6, pp. 1117–1130, Jul. 2022.

[50] A. O. Kingsley, U. G. Inyang, O. Msugh, F. T. Mughal, and A. Usoro, "Recognizing facial emotions for educational learning settings," *IAES Int. J. Robot. Autom. (IJRA)*, vol. 11, no. 1, p. 21, Mar. 2022.

[51] C. Tang, P. Xu, Z. Luo, G. Zhao, and T. Zou, "Automatic facial expression analysis of students in teaching environments," in *Proc. Chin. Conf. Biometric Recognit.* Cham, Switzerland: Springer, 2015, pp. 439–447.

[52] N. K. Mehta, S. S. Prasad, S. Saurav, R. Saini, and S. Singh, "Three-dimensional DenseNet self-attention neural network for automatic detection of student's engagement," *Appl. Intell.*, vol. 52, pp. 13803–13823, Mar. 2022.

[53] A. Qayyum, I. Razzak, M. Tanveer, and M. Mazher, "Spontaneous facial behavior analysis using deep transformer based framework for child–computer interaction," *ACM Trans. Multimedia Comput., Commun., Appl. (TOMM)*, 2022.

[54] S. Nur Karimah, T. Unoki, and S. Hasegawa, "Implementation of long short-term memory (LSTM) models for engagement estimation in online learning," in *Proc. IEEE Int. Conf. Eng., Technol. Educ. (TALE)*, Dec. 2021, pp. 283–289.

[55] S.-Y. Lin, C.-M. Wu, S.-L. Chen, T.-L. Lin, and Y.-W. Tseng, "Continuous facial emotion recognition method based on deep learning of academic emotions," *Sensors Mater.*, vol. 32, no. 10, pp. 3243–3259, 2020.

[56] M. Megahed and A. Mohammed, "Modeling adaptive e-learning environment using facial expressions and fuzzy logic," *Exp. Syst. Appl.*, vol. 157, Nov. 2020, Art. no. 113460.

[57] S. A. Mohammed, A. A. Abdulrahman, and F. S. Tahir, "Emotions students' faces recognition using hybrid deep learning and discrete Chebyshev wavelet transformations," *Comput. Sci.*, vol. 17, no. 3, pp. 1–13, 2022.

[58] S. Fakhar, J. Baber, S. U. Bazai, S. Marjan, M. Jasinski, E. Jasinska, M. U. Chaudhry, Z. Leonowicz, and S. Hussain, "Smart classroom monitoring using novel real-time facial expression recognition system," *Appl. Sci.*, vol. 12, no. 23, p. 12134, Nov. 2022.

[59] T. S. Ashwin and R. M. R. Guddeti, "Automatic detection of students' affective states in classroom environment using hybrid convolutional neural networks," *Educ. Inf. Technol.*, vol. 25, no. 2, pp. 1387–1415, Mar. 2020.

[60] S. K. Gupta, T. S. Ashwin, and R. M. R. Guddeti, "Students' affective content analysis in smart classroom environment using deep learning techniques," *Multimedia Tools Appl.*, vol. 78, no. 18, pp. 25321–25348, Sep. 2019.

[61] A. V. Savchenko, L. V. Savchenko, and I. Makarov, "Classifying emotions and engagement in online learning based on a single facial expression recognition neural network," *IEEE Trans. Affect. Comput.*, vol. 13, no. 4, pp. 2132–2143, Oct. 2022.

[62] Y. Hu, Z. Jiang, and K. Zhu, "An optimized CNN model for engagement recognition in an e-learning environment," *Appl. Sci.*, vol. 12, no. 16, p. 8007, Aug. 2022.

[63] J. C. Hung, K.-C. Lin, and N.-X. Lai, "Recognizing learning emotion based on convolutional neural networks and transfer learning," *Appl. Soft Comput.*, vol. 84, Nov. 2019, Art. no. 105724.

[64] S. Vairale, P. Moon, and P. Patil, "Student facial expression recognition for lecture review," *Int. Res. J. Eng. Technol. (IRJET)*, 2018.

[65] H. V. Kuilenburg, M. Wiering, and M. D. Uyl, "A model based method for automatic facial expression recognition," in *Proc. Eur. Conf. Mach. Learn.* Cham, Switzerland: Springer, 2005, pp. 194–205.

[66] K. Bahreini, R. Nadolski, and W. Westera, "Towards multimodal emotion recognition in e-learning environments," *Interact. Learn. Environments*, vol. 24, no. 3, pp. 590–605, Apr. 2016.

[67] J. M. Saragih, S. Lucey, and J. F. Cohn, "Deformable model fitting by regularized landmark mean-shift," *Int. J. Comput. Vis.*, vol. 91, no. 2, pp. 200–215, Jan. 2011.

[68] Y. Cui, S. Wang, and R. Zhao, "Machine learning-based student emotion recognition for business English class," *Int. J. Emerg. Technol. Learn. (iJET)*, vol. 16, no. 12, pp. 94–107, 2021.

[69] U. Ayvaz, H. Gürüler, and M. O. Devrim, "Use of facial emotion recognition in e-learning systems," *Inf. Technol. Learn. Tools*, vol. 60, no. 4, p. 95, Sep. 2017.

[70] L. Y. Mano, G. T. Giancristofaro, B. S. Faiçal, G. L. Libralon, G. Pessin, P. H. Gomes, and J. Ueyama, "Exploiting the use of ensemble classifiers to enhance the precision of user's emotion classification," in *Proc. 16th Int. Conf. Eng. Appl. Neural Netw. (INNS)*, 2015, pp. 1–7.

[71] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[72] T. Kanade, J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *Proc. 4th IEEE Int. Conf. Autom. Face Gesture Recognit.*, Sep. 2000, pp. 46–53.

[73] M. Ranzato, J. Susskind, V. Mnih, and G. Hinton, "On deep generative models with applications to recognition," in *Proc. CVPR*, 2011, pp. 2857–2864.

[74] A. Kirzhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 1097–1105.

[75] A. Gupta, A. D'Cunha, K. Awasthi, and V. Balasubramanian, "DAiSEE: Towards user engagement recognition in the wild," 2016, *arXiv:1609.01885*.

[76] A. Sun, Y.-J. Li, Y.-M. Huang, and Q. Li, "Using facial expression to detect emotion in e-learning system: A deep learning method," in *Proc. Int. Symp. Emerg. Technol. Educ.* Cham, Switzerland: Springer, 2017, pp. 446–455.

[77] D. Yang, A. Alsadoon, P. W. C. Prasad, A. K. Singh, and A. Elchouemi, "An emotion recognition model based on facial recognition in virtual learning environment," *Proc. Comput. Sci.*, vol. 125, pp. 2–10, 2018.

[78] O. El Hammoumi, F. Benmarrakchi, N. Ouherrou, J. El Kafi, and A. El Hore, "Emotion recognition in e-learning systems," in *Proc. 6th Int. Conf. multimedia Comput. Syst. (ICMCS)*, 2018, pp. 1–6.

[79] I. Lasri, A. R. Solh, and M. E. Belkacemi, "Facial emotion recognition of students using convolutional neural network," in *Proc. 3rd Int. Conf. Intell. Comput. Data Sci. (ICDS)*, Oct. 2019, pp. 1–6.

[80] W. Wang, K. Xu, H. Niu, and X. Miao, "Emotion recognition of students based on facial expressions in online education based on the perspective of computer simulation," *Complexity*, vol. 2020, pp. 1–9, Sep. 2020.

[81] J. Liao, Y. Liang, and J. Pan, "Deep facial spatiotemporal network for engagement prediction in online learning," *Int. J. Speech Technol.*, vol. 51, no. 10, pp. 6609–6621, Oct. 2021.

[82] Y. Guo, J. Huang, M. Xiong, Z. Wang, X. Hu, J. Wang, and M. Hijji, "Facial expressions recognition with multi-region divided attention networks for smart education cloud applications," *Neurocomputing*, vol. 493, pp. 119–128, Jul. 2022.

[83] A. Abedi and S. Khan, "Affect-driven engagement measurement from videos," *Computer*, vol. 11, p. 12, Aug. 2021.

[84] S. Wang, "Online learning behavior analysis based on image emotion recognition," *Traitement du Signal*, vol. 38, no. 3, pp. 865–873, Jun. 2021.

[85] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 31, no. 1, 2017, p. 11231.

[86] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6848–6856.

[87] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.

[88] H.-W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler, "Deep learning for emotion recognition on small datasets using transfer learning," in *Proc. ACM Int. Conf. Multimodal Interact.*, Nov. 2015, pp. 443–449.

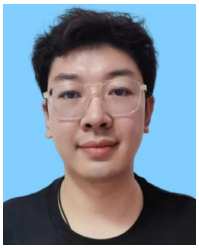
[89] R. Xu, J. Chen, J. Han, L. Tan, and L. Xu, "Towards emotion-sensitive learning cognitive state analysis of big data in education: Deep learning-based facial expression analysis using ordinal information," *Computing*, vol. 102, no. 3, pp. 765–780, Mar. 2020.

[90] J. M. Harley, F. Bouchet, M. S. Hussain, R. Azevedo, and R. Calvo, "A multi-componential analysis of emotions during complex learning with an intelligent multi-agent system," *Comput. Hum. Behav.*, vol. 48, pp. 615–625, Jul. 2015.

[91] S. Gupta, P. Kumar, and R. K. Tekchandani, "Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models," *Multimedia Tools Appl.*, vol. 82, no. 8, pp. 11365–11394, Mar. 2023.
[92] J. Whitehill, Z. Serpell, Y.-C. Lin, A. Foster, and J. R. Movellan, "The faces of engagement: Automatic recognition of student engagementfrom facial expressions," *IEEE Trans. Affect. Comput.*, vol. 5, no. 1, pp. 86–98, Jan. 2014.
[93] A. Kamath, A. Biswas, and V. Balasubramanian, "A crowdsourced approach to student engagement recognition in e-learning environments," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–9.
[94] A. Kaur, A. Mustafa, L. Mehta, and A. Dhall, "Prediction and localization of student engagement in the wild," in *Proc. Digit. Image Comput., Techn. Appl. (DICTA)*, 2018, pp. 1–8.
[95] C. Bian, Y. Zhang, F. Yang, W. Bi, and W. Lu, "Spontaneous facial expression database for academic emotion inference in online learning," *IET Comput. Vis.*, vol. 13, no. 3, pp. 329–337, Apr. 2019.
[96] Q. Wei, B. Sun, J. He, and L. Yu, "BNU-LSVED 2.0: Spontaneous multimodal student affect database with multi-dimensional labels," *Signal Process., Image Commun.*, vol. 59, pp. 168–181, Nov. 2017.
[97] T. S. Ashwin and R. M. R. Guddeti, "Affective database for e-learning and classroom environments using Indian students' faces, hand gestures and body postures," *Future Gener. Comput. Syst.*, vol. 108, pp. 334–348, Jul. 2020.
[98] A. Dhall, "EmotiW 2019: Automatic emotion, engagement and cohesion prediction tasks," in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2019, pp. 546–550.
[99] A. Dhall, G. Sharma, R. Goecke, and T. Gedeon, "EmotiW 2020: Driver gaze, group emotion, student engagement and physiological signal based challenges," in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2020, pp. 784–789.
[100] M. A. A. Dewan, M. Murshed, and F. Lin, "Engagement detection in online learning: A review," *Smart Learn. Environ.*, vol. 6, no. 1, pp. 1–20, 2019.
[101] B. Sun, Q. Wei, J. He, L. Yu, and X. Zhu, "BNU-LSVED: A multimodal spontaneous expression database in educational environment," in *Proc. SPIE*, vol. 9970, 2016, pp. 256–262.
[102] B. Sun, S. Lu, Y. Wen, J. He, and L. Yu, "Analyses and benchmark of a spontaneous student affect database," in *Proc. IEEE Int. Symp. Multimedia (ISM)*, Jul. 2021, pp. 206–209.
[103] I. Kotsia, I. Buciu, and I. Pitas, "An analysis of facial expression recognition under partial facial image occlusion," *Image Vis. Comput.*, vol. 26, no. 7, pp. 1052–1067, Jul. 2008.
[104] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expression recognition using CNN with attention mechanism," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2439–2450, May 2019.
[105] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region attention networks for pose and occlusion robust facial expression recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 4057–4069, 2020.
[106] H. Ding, P. Zhou, and R. Chellappa, "Occlusion-adaptive deep network for robust facial expression recognition," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Sep. 2020, pp. 1–9.
[107] K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao, "Suppressing uncertainties for large-scale facial expression recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jul. 2020, pp. 6897–6906.
[108] J. Zeng, S. Shan, and X. Chen, "Facial expression recognition with inconsistently annotated datasets," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 222–237.

BEI FANG received the Ph.D. degree from the School of Computer Science, Northwestern Polytechnical University, Xi'an, China, in 2019. Currently, she is a Postdoctoral Research Associate with the Key Laboratory of Modern Teaching Technology, Ministry of Education, Shaanxi Normal University, Xi'an. Her current research interests include multimedia computing, facial expression recognition, and visual reasoning.

XIAN LI is currently pursuing the M.S. degree with the Key Laboratory of Modern Teaching Technology, Ministry of Education, Shaanxi Normal University. Her current research interests include multimedia analysis, computer vision, and deep learning.



GUANGXIN HAN received the M.S. degree from the Key Laboratory of Modern Teaching Technology, Ministry of Education, Shaanxi Normal University, where he is currently pursuing the Ph.D. degree. His current research interests include multimedia analysis, computer vision, and deep learning.

JUHOU HE received the Ph.D. degree from the School of Computer Science, Northwestern Polytechnical University, in 2005. He is currently a Professor with the Key Laboratory of Modern Teaching Technology, Ministry of Education, Shaanxi Normal University. His current research interests include image processing, computation intelligence, and signal processing.