An Investigation of Student Facial Expression Recognition Based on Machine Learning Algorithms

Yanyi Chen

Northeast Yucai School, Shenyang, 110179, China

zhaoyue1@xhd.cn

Abstract. Facial Expression Recognition (FER) refers to the classification of facial expressions based on facial expressions in images by using classification algorithms, which are usually classified into 6 categories. Given the large number of students in a classroom, researchers have employed facial expression recognition technology to enhance the recognition of students' expressions, thereby ensuring the quality of teaching. Teachers can effectively complete classroom teaching by judging students' learning state in class through students' expression recognition technology. By providing a comprehensive review of recent literature on student facial expression recognition, this paper aims to provide valuable insights into the application of this technology in education. The first section of this paper provides an introduction to data preparation and preprocessing methods of students' facial expression recognition methods, including the construction of datasets and feature extraction, then summarizes the models used in facial expression methods and the verification effects of the models including Support Vector Machine (SVM) and Deep Learning-based methods, and final section summarizes the full paper.

Keywords: Student Facial Expression, Deep Learning, Learning State.

1. Introduction

Facial expression recognition technology has found widespread applications across diverse fields e.g. education, medicine, transportation safety, and agricultural pest control. Student facial expressions are important indicators of the effectiveness of a course, reflecting their interest and attention to the subject matter. Teachers can reflect on teaching processes and make adjustments to the curriculum through student facial expressions. Additionally, teachers can develop personalized training programs based on students' facial expressions. In conventional teaching methods, teachers manually observe and record students' facial expressions, which can be time-consuming. With the continuous development of modern teaching technology, video equipment can automatically collect facial data from students in the classroom, providing convenience for subsequent data analysis. Therefore, the use of facial expression recognition technology holds the potential to enhance teaching effectiveness while concurrently alleviating the workload burden on teachers.

B. H. Ahmad (ed.), *Proceedings of the 2023 International Conference on Data Science, Advanced Algorithm and Intelligent Computing (DAI 2023)*, Advances in Intelligent Systems Research 180,
https://doi.org/10.2991/978-94-6463-370-2_60

Facial expression recognition is an application of image recognition research and a popular research topic in the field of computer vision. Students' facial expressions are divided into different categories based on traditional facial expressions according to the research content. For instance, Zhao divided students' expressions into six categories: understanding, listening, curiosity, doubt, exhaustion, and distraction [1]. Wang and Lai classified students' online listening state into positive emotions, negative emotions, and neutral emotions [2]. Sujit et al. divided students' mood based on emotions into high positive affect, low positive affect, high negative affect, and low negative affect [3]. In the current research of students' facial expression recognition, the process of facial expression recognition is divided into five parts: data preparation, data preprocessing, feature extraction, classification model construction, and model evaluation. In data preparation, the datasets usually come from two sources: collections from classes and public datasets. The collections from classes are from videos of online classes or classes in schools. For example, Zhao built a database of primary school students' online learning expressions through online classroom videos [2], while Sujit et al. gathered the dataset from the learning videos of 95-110 students [3]. The public datasets mainly include FER-2013 [2], RAF-DB [4], CAER-S [4], and AffectNet-7 [4] datasets. In feature extraction, Convolutional Neural Networks (CNN), Histogram of Oriented Gradients (HOG) [5], Scale-invariant feature transform (SIFT), and other techniques are commonly used in the literature. In classification model construction, SVM, CNN, and Visual Geometry Group (VGG) [6] are commonly used in the literature. Model evaluation employs various classification model evaluation metrics such as confusion matrix, accuracy rate, recall rate, F-score.

This review summarizes research methods in literature for students' facial expression recognition. The methods are presented in order of data preparation, processing, feature extraction, classification model construction, and model evaluation to introduce the methods respectively. These selected articles are collected from the previous three years, including conference, journal, and dissertation articles.

The rest of the review is as follows: Section 2 presents the methods used for preprocessing including data preparation, data processing, and feature extraction. Section 3 concludes the proposed classification models and model evaluation in the literature. The last section services as a conclusion to the paper.

2. Data Introduction and Preprocessing

Data preprocessing is a crucial step before building the model. This section focuses on introducing some widely used students' face recognition dataset, data processing, and feature extraction methods.

2.1 Data preparation and processing

The process of selecting and constructing datasets serves as the foundation for studying student facial expression recognition. The researchers primarily utilize public datasets and self-built datasets for studying facial expression recognition. The

public datasets commonly employed in the literature for this paper mainly include: FER2013 [2,7-12], CK+ [2, 11, 13], RAF-DB [4, 11], CAER-S [4], AffectNet [4], WIDER [11], BAUM-1 [14], DAiSEE [14], YawDD [14], and LIRIS-CSE [7]. The description of these public datasets and the details of expression classification are shown in Table 1.

Table 1. The details of expression classification on these public datasets.

Dataset	Number of emotions	Number of images
FER2013	7	35,877
CK+	8	593
RAF-DB	8	29,672
CAER-S	7	65,983
AffectNet	8	450,000
WIDER	6	32,203
BAUM-1	8	1,502
DAiSEE	4	9,068
YawDD	2	351
LIRIS-CSE	6	208

The self-built dataset is created based on video recordings from classrooms [1, 3, 10-11,15-17]. The video duration and the number of pictures collected by each dataset are different. For example, Zhao collected data from the class videos of 115 primary school students, and each student recorded for no more than 5 minutes [1]. Sujit et al. collected data from 95-110 Indian Students, and each student collected four different states with four videos [3]. Wang et al. collected data from the lecture video in Superstar Learning Pass app, including three courses, College Chinese, advanced mathematics, and college English [15].

To enhance the accuracy and robustness of expression recognition, researchers employ data enhancement methods such as image flipping and cropping to augment the original data.

The keras library of Python, imagedatagenerator module can transform the image [1,12-14]. The OpenCV library is a commonly used tool in the field of computer vision due to its implementation of various common image processing methods. For instance, Wang et al. uses OpenCV to crop, normalize, scale, and flip the original image horizontally [2]. Sumalakshmi and Vasuki employ the Viola-Jones method to locate the face region, mouth, and eyes in the input face image, and use Fuzzy Weighted Histogram Equalization (FWHE) to enhance the input image [7]. Sujit et al. extract frames from the video and resize them to fixed size (800 by 600 pixels) for pre-processing purposes [3]. Feng et al. apply random rotations, expansions, contractions, and translations on the original images [16]. Sun et al. use the mosaic method to enhance data and compensate for the problem of insufficient sample size [8]. Fang et al. randomly flip and crop images to 48 by 48 pixels using OpenCV [10]. Abdullah and AlKan use data enhancement methods in MATLAB that increases the original image by 4 times [17].

In addition, some researchers directly complete the data processing work in the model [10, 11].

2.2 Feature Extraction

The purpose of image feature extraction is to extract useful information for image recognition and classification from the original image. The extracted features are used in the subsequent classification model to make the classification results more stable and reliable. The main feature extraction methods used in the research of students' expression recognition include traditional feature extraction methods SIFT, HOG, Local Binary Patterns (LBP), and feature extraction methods based on deep learning. In some studies, the author also uses a hybrid method of traditional methods and deep learning methods to realize feature extraction and fuse features.

During the feature extraction process, researchers use different methods to extract features at different locations on an image [1,2,4,7,13]. Zhao uses CNN to extract deep global features, HOG to extract local texture features, and SIFT to extract local features [1]. Next, Zhao fused the results of the three normalized features, and ultimately obtained the fused features of the expression image [1]. Mu et al. used the local feature extractor and channel space adjuster in the EfficientFace model to extract local and global features of the image [4]. Sumalakshmi and Vasuki used the KFDA-(Compound LBP) CLBP-(Active Appearance Model) AAM method to extract appearance features and geometric features using a geometric deformation model [7]. Fan proposed the feature extraction method that includes three steps: 1) Extracting features using the LBP algorithm in OpenCV; 2) Extracting the depth features of grayscale images using AlexNet and the features extracted using LBP algorithm, and normalizing the two obtained features to form fused features; 3) Using Principal Component Analysis (PCA) method to perform dimensionality reduction on the fused feature set [13]. Attention mechanism is used in feature extraction to improve the accuracy of feature extraction, Wang and Lai extract detail texture features using a shallow network, add an attention mechanism to the shallow network, and extract global expression features using a deep neural network [2]. The extracted features are then fused to form a unified representation of the expression.

Researchers have also used face detection methods in the feature extraction process [3,15]. Sujit et al. proposed the Max-Margin Face Detection (MMFD) method for face detection [3]. Wang et al. used the CNN to recognize facial action units (AUs), and then took AUs as the input value of the classification model [15].

In other research methods in this paper, either a traditional CNN model or a CNN-based deep learning model is used for feature extraction [8-12,14,16-17].

3. Classification Models and Performance Evaluation

The classification models used in the study of students' expression recognition mainly include SVM [1,13] and CNN-based models [2-4,7-12,14-17]. This section introduces the models designed for this research and summarizes their evaluation results.

3.1 Models

In the research of students' expression recognition, most tasks involve multi-class classification. SVM is a classical classification model that supports both linear and nonlinear classification. It is used after image feature processing to classify students' facial expressions. For example, Zhao and Fan both use SVM to classify students' facial expressions after fusing image features [1, 13].

The CNN model is widely used in the field of image recognition. Wang and Lai proposed a DS-EfficientNet model based on CNN that uses softmax function to classify expressions and employs Weighted Cross Entropy Loss (WLoss) as the loss function [2]. Sumalakshmi and Vasuki used Long Short-Term Memory network with Attention Mechanism (ALSTM) network with attention mechanism to classify expressions [7], while Sujit et al. used M-Inception-V3 model for expression classification [3]. Fang et al. proposed a CNN-based model with 7 convolutional layers, 6 max pooling layers, 1 dropout layer, and 1 global average pooling layer [9]. Abdullah and Alkan developed a system for automatically recognizing students' expressions in online teaching [17]. Abdullah and Alkan employeed several pre-training deep learning models, including AlexNet, MobileNetV2, GoogleNet, ResNet18, ResNet50, and VGG16, for transfer learning, and used K-fold validation for expression recognition [17]. The You Only Look Once (YOLO) model is a classification model that treats the problem as a regression problem, which is different from the CNN-based model. Sun et al. used the improved YOLO5 model, YG-YOLOv5s, to classify students' expressions by using the regression approach [8]. These modifications include the use of Soft Non-Maximum Suppression (NMS) instead of NMS, add Coordinate Attention (CA), and change the loss function to Edge Intersection Over Union (EIOU).

Some researchers have explored the classification of students' attention state or engagement states according to the classification results of students' expression representation, and apply them in the classroom [10-12,14-15]. In addition, Mu et al. used the tag distribution generator in the EfficientFace model to achieve expression classification, and combined it with a head detection model to locate the head in an image and complete expression recognition [4].

3.2 Model Evaluation

In the study of students' expression recognition, the main metrics used in model evaluation are Accuracy, Recall, Precision, F-Score, AUC curve and confusion matrix. In addition, ablation experiment is used to test the effect of fusion features.

In the research of students' expression recognition using SVM as classification model, Zhao compared the performance of SVM classifiers before and after transfer learning [1]. From the average accuracy results, it can be seen that after transfer learning (77.67%) is higher than before transfer learning (55.72%). In the experiment, it compares the average accuracy rates of three single features and fused features. The results show that the accuracy rate of CNN is higher when using single classification, and the average accuracy rate after fusing features (81.85%) is higher than that obtained using three single features. Fan uses the accuracy to verify the results, and it can be seen that the average accuracy of deep fusion features (92.25%) is about 10%

higher than that of single LBP texture features (80.2%) [13].

The practical applications of many of the results from student expression recognition studies have been demonstrated in the classroom. For example, the accuracy of EfficientFace model on RAF-DB, CAER-S and AffectNet-7 was 88.36%, 85.87%, and 63.7%, respectively [4]. Mu et al. applied the improved EfficientFace model to classroom teaching [4]. Wang et al. evaluated the accuracy of the model through the feedback results in class, and the accuracy reached 82.5% [15]. Pabba and Kumar proposed model validated on 19 classroom lecture video recordings and achieved 73.68% correctness [14]. Sun et al. verified the accuracy of the improved model (YG-YOLOv5s) and YOLO v5 model on the Fer2013 and self-built dataset respectively, and the results showed that YG-YOLOv5s had higher accuracy [8]. The MS-ResNet model proposed by Feng et al. has an average accuracy of 93.6% on self-built dataset [16]. The model proposed by Fang et al. has an accuracy rate of 73% on the FER2013, and has been applied to the teaching environment [9]. Gupta et al. evaluated the Inception-V3, VGG19, and ResNet-50 as models for engagement detection in real-time [11]. After experimentation, the accuracy for ResNet-50 is best. Abdullah and Alkan used several evaluation metrics to evaluate six models, and ResNet18 was better than other models [17]. The ResNet18 is chosen as the transfer learning model, and the F1-score and AUC curves of the proposed model reach 99% and 100%, respectively. The accuracy of L_VGG model proposed by Fang et al. on FER2013 is 0.6761 [12]. The concentration is calculated according to the results of L_VGG, and the model is effectively applied to online teaching.

In the field of student expression recognition, most researchers utilize deep learning as the classification model. Furthermore, several models [2-3,7,10] are compared in research to determine the most effective one. Table 2 summarizes these researches, including evaluation metrics of the model, the models compared in the experiment and the two models with the best effect.

Table 2. Evaluation metrics and effectiveness of expression recognition models.

Works	Evaluation Metrics	Compared Models	The Two Best Models (dataset)
[2]	Accuracy, F1-Score	VGG16, ResNet50, Xception, InceptionV3, MobileNetv2, EffcientNet, and DS-EfficientNet	DS-EfficientNet, MobileNetv2 (FER2013) DS-EfficientNet, EffcientNet (CK+)
[3]	Accuracy, mAP	CNN-KNN, Ensemble of 8 CNN, CNN with VGG, Human Accuracy, HOG+SVM, and Fused deep learning	Fused deep learning, CNN-KNN (FER2013)

[7]	Accuracy	AlexNet, VGGNet, ResNet, Inception v-3, M-Inception v-3	M-Inception v-3, Inception v-3 (self-built dataset)
[10]	Accuracy, F1-Score, AUC	HOG+SVM, CNN, VGGNet, ENGAGEMENT	ENGAGEMENT, VGGNet (self-built dataset)

In addition, the accuracy of the model proposed by Sumalakshmi and Vasukip on LIRIS data set is 88.24%, which is higher than that of VGGNet-CNN method (75%) [7].

4. Conclusion

Student expression recognition technology has been widely adopted in classroom teaching and has achieved remarkable results both online and offline. This paper summarizes the process of model construction and evaluation in student expression recognition technology based on research achievements over the past three years, providing a reference for researchers in this field. The model construction section highlights the datasets and feature extraction methods used to build the model, while the model evaluation section compares the performance of various models to assess their effectiveness.

In summary, the field of student expression recognition has made significant strides in recent years. However, there are still several areas that require further research and development. One such area is accounting for variations in expressions among students of different age groups. Another area that requires further research is expanding online learning expression databases to include a larger sample size and diverse expression categories. This will enable researchers to develop more accurate and reliable models for recognizing student expressions in various contexts. Furthermore, refining research on student microexpressions is crucial for enhancing their detection and analysis. Microexpressions are brief facial expressions that occur involuntarily and can reveal a person's true emotions. By improving researchers' ability to recognize and analyze microexpressions, they can gain valuable insights into students' emotional states and well-being. Finally, it is essential to prioritize protecting students' privacy when collecting and utilizing their expression data in the classroom setting. This includes obtaining informed consent from students and their parents, ensuring data confidentiality, and minimizing the risk of misuse or unauthorized access to the data.

References

1. Zhao, X. X.: Study on Online Learning Expression Recognition of Pupils Based on Feature Fusion (in Chinese) (2020).

2. Wang, L., Lai, M. L.: Analysis of Students' Concentration in Online Classroom Based on Facial Expression Recognition (in Chinese), Computer Systems & Applications, 55-62. (2023).
3. Sujit, K.G., et al.: Students' affective content analysis in smart classroom environment using deep learning techniques, Multimedia Tools and Applications, 25321-25348 (2019).
4. Mou, Y. R., et al.: On the Application of Students' Classroom Facial Expression Recognition Technology Based in Deep Learning (in Chinese), Journal of Hunan Radio and Television University, 17-24.(2023).
5. Lin, K. Z., et al.: Research on HOG Feature Extraction Algorithm Weighted by Information Entropy (in Chinese), Computer Engineering and Applications,147-152.(2020).
6. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition, ICLR 2015 (2015).
7. Sumalakshmi, C. H., et al.: Fused deep learning based Facial Expression Recognition of students in online learning mode, Concurrency and Computation: Practice and Experience. (2022).
8. Sun, X. Y., et al.: Student facial expression recognition based on improved YOLOv5, Journal of Qilu University of Technology, 28-35. (2023).
9. Fang, J., Yuewu., Cheng, L.: A lightweight convolutional neural network student learning behavior analysis based on facial expression recognition, EIECS 2021, 593-596 (2021).
10. Mohamad, N. O., et al.: Automatic Recognition of Student Engagement Using Deep Learning and Facial Expression, ECML PKDD 2019, 273-289 (2020).
11. Gupta, S., Kumar, P., Tekchandani, R. K.: Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models, Multimedia Tools and Applications, 1165-11394 (2023).
12. Fang, B.: Research on online learning Expression Recognition based on Convolutional Neural Network (in Chinese) (2023).
13. Fan, L. Y.: Learning Expression Recognition Based on Convolutional Neural Network and Depth Feature Fusion (in Chinese), scientific and technological innovation, 85-88 (2022).
14. Pabba, C., Kumar, P.: An intelligent system for monitoring students' engagement in large classroom teaching through facial expression recognition, Expert Systems (2022).
15. Wang, M., Jing, C., Huang, Z., Tan, X.: A Method of Students' Online Learning Status Analysis Based on Facial Expression, ICCSE 2022 (2023).
16. Feng, H.-X., Wang, L., Feng, F.-J., Tan, M., Peng, L.-Y., Zhao, Y.-Q., Zhang, X.-W.: Student Facial Expression Recognition Based on Multi-scale Residual Network, CISAI 2021, 207-211 (2021).
17. Abdullah, M.U., Alkan, A.: A comparative approach for facial expression recognition in higher education using hybrid-deep learning from students' facial images, Traitement du Signal, 1929-1941 (2022).