Multimedia Systems (2022) 28:469–478
https://doi.org/10.1007/s00530-021-00854-x

SPECIAL ISSUE PAPER

Assessing learning engagement based on facial expression recognition in MOOC's scenario

Junge Shen[1] · **Haopeng Yang**[1] · **Jiawei Li**[1] · **Zhiyong Cheng**[2]

Received: 9 November 2020 / Accepted: 27 September 2021 / Published online: 19 October 2021

Abstract
Online learning has become one of the most important learning styles, yet with the need of supervisors to consistently keep the learners motivated and on-task. Some learners could be supervised by outer factors, and distance learners have to be motivated by themselves. However, online learning engagement is hardly to be assessed by supervisors in real time. With the rapid development of information technology, it is able to remedy the above problem by using intelligent video surveillance techniques. In this paper, we propose a novel framework of learning engagement assessment which introduces facial expression recognition to timely acquire the emotional changes of the learners. Moreover, a new facial expression recognition method is proposed based on domain adaptation, which is suitable for the MOOC scenario. The experiments show the effectiveness of our proposed framework on assessing learners' learning engagement. The comparisons with the state-of-the-art methods also demonstrate the superiority of our proposed facial emotion recognition method.

Keywords Online learning assessment · Facial expression recognition · Domain adaptation · CNN

1 Introduction

With the development of the Internet and communication technology, the MOOCs education has become a mature plan which enable learners to learn anytime and anywhere. Especially, since the outbreak of novel coronavirus pneumonia in Wuhan in December 2019, this disaster enforces people to stay at home and students to study online. However, in the case of MOOCs, there is no teacher to keep learners motivated, whereas it is required to track the studying levels that how well the learners have mastered the courses. Consequently, it is expected to assess the learning engagement of MOOC learns automatically instead of relying on human supervisors.

There have been many methods proposed to assess the learning engagement [1], and those methods can be mainly divided into two categories. The first one is to use the physiological signals of human to achieve, including brainwave [2], muscle electricity [3], blood pressure [4], and so on. These methods are limited in real applications because of the dependence on the wearable collection equipment. The other one relies on human behaviors, such as action recognition [5–7], gesture recognition [8–10], facial expression recognition [11]. The perception of human behavior is spontaneous, so that it is more suitable to perceive learning engagement in real MOOC scenario by computer vision techniques.

The web-camera is a typical device for perceptual signal collection to assess learning engagement. In this scenario, the postures of learners cannot be completely photographed by the web-camera, and the learners may adopt any comfortable posture in any scene, which is different from the classroom scenario that has certain constraints for learners' sitting postures [12, 13]. Instead, the most relevant factor to learn engagement is the facial expression which could be consistently and easily captured by the web-camera during online learning. This inspires us to develop a novel framework for learning engagement assessment spontaneously via facial expression recognition in the MOOC scenario.

In this paper, we treat the learning engagement assessment as a classification problem to recognize basic facial expressions [14, 15]. It only needs to detect the facial images

✉ Junge Shen
shenjunge@nwpu.edu.cn

[1] Unmanned System Research Institute, Northwestern Polytechnical University, Xi'an, China

[2] Shandong Artificial Intelligence Institute, Qilu University of Technology (Shandong Academy of Sciences), Jinan, China

of learners returned by the camera, so as to perceive the learners' learning engagement in real time. Specifically, four types of facial expressions related to learning engagement are adopted, including neutral, understand, disgust, and doubt. The precision of facial expression recognition is determined by the representation of facial expression. The traditional hand-crafted feature extraction methods cannot extract appropriate features adaptively. Instead, deep learning provides a powerful tool to extract high-level features by using deep convolutional neural networks (CNNs) [16, 17], which becomes the main-stream framework for facial expression. To improve the performance of facial expression recognition, the architecture of the convolutional neural network is becoming more and more complicated, resulting in the increasing number of model parameters, the requirement of huge hardware computing power, and the risk of overfitting. However, In the MOOC scenario, it is necessary to understand the learners' class situation in real-time, and to recognize micro-expressions [18, 19]. Meanwhile, in the field of facial expression recognition, the datasets are generally small and insufficient for training complex network models. Therefore, it is crucial to a lightweight convolutional neural network is expected.

Notice that web-cameras produce the images with complicated backgrounds, different illuminations, and various resolutions. Hence, to develop a reliable facial expression recognition model, it is important to selectively amplify the influence of valuable features. This inspires us to employ an attention mechanism [20, 21], named the squeeze-and-excitation (SE) blocks [22], to the proposed facial expression recognition algorithm. In this way, we cannot only magnify the effects of valuable features from global information, but also inhibit the useless ones.

Considering that a large amount of labeled data are required for training the deep convolutional neural network whereas the accessible online data are limited, we propose to use the public labeled data as auxiliary data. This brings us the problem that different datasets of facial images have different distributions caused by the variant background, illumination, and resolution [23]. Hence, the auxiliary dataset cannot be directly used in training the recognition model. Targeting at this problem, in the absence of labeled online data, we adopt the domain adaptation technique [24–26] to solve the problem of distribution difference, such that the labeled data can be used to assist the identification of online facial expressions. Specifically, to improve the generalization ability of the model in the case of complex background, illumination, and varied expressions, we adopt the large face expression database with labels as the auxiliary domain, and transfer the common expression features to the target domain.

In summary, we develop a novel framework to access learning engagement based on facial expression recognition in the MOOC scenario. The facial images of learners are obtained based on web-cameras, and an SE-CNN based on domain adaptation is designed to detect facial expression to access learning engagement. The analysis results can be used to improve the effects of MOOC learning.

The main contributions of our paper are summarized as following:

1. We propose a novel framework to assess the learning engagement of online learners in the MOOC scenario. Our system can effectively perceive learners' learning status by timely recognizing their facial expressions.
2. To fulfill the real-time requirement in the scenario of complicated background, different illuminations and resolutions, we design a CNN model based on domain adaptation for facial expression recognition, which can exploit the auxiliary data to alleviate the lack of labeled data.
3. To evaluate the engagement of the learning behavior, a strategy is proposed to calculate students' learning status and the experiments are conducted to show the effectiveness of our proposed framework.

The reminder of this paper is organized as follows. In Sect. 2, the framework of spontaneous assisted learning in MOOC scenario is introduced, and the proposed method is described in detail. Section 3 presents the details of our experiments and results. Finally, conclusions are drawn in Sect. 4.

2 Learning engagement assessment via face expression recognition

Before describing our proposed framework, we would like to brief the assisted learning in the MOOC education. Specifically, a learner watches the MOOC instructional videos to learn knowledge without teachers' monitoring, so assessing learning engagement [27] is required to evaluate the engagement level of learners. The engagement is a bridge between the learner and the learning resource, where emotion could be employed to evaluate the engagement.

Formally, let l be a learner and V be the instructional video when studying.

The learning engagement assessment system composed of face acquisition, facial expression recognition, and learning engagement assessment in Fig. 1. This system collects the student expressions when they are learning MOOC and are supervised by the front camera. In this process, the captured student facial images are preserved by the front camera, which are represented as $I = \{i_1, i_2, \ldots, i_n\}$. For each facial image i_k, there would be a corresponding emotion e_k.

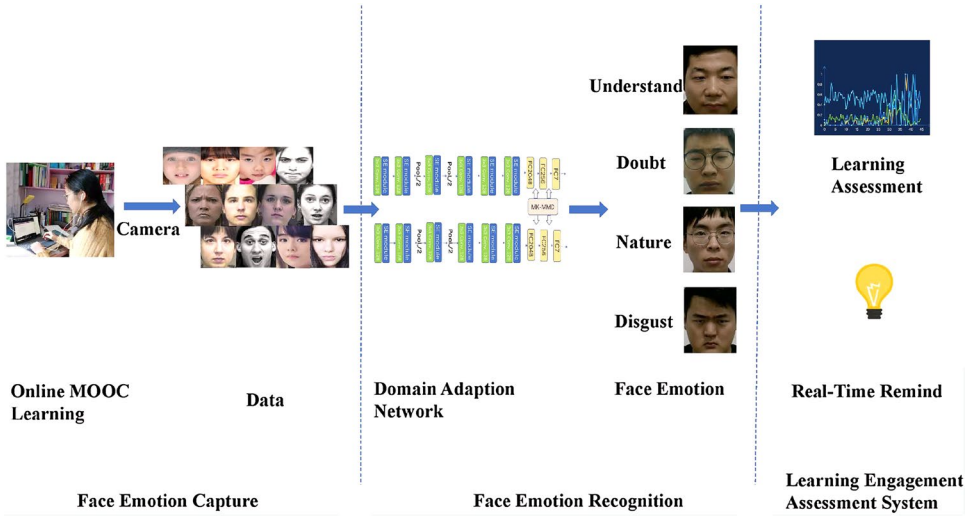

Fig. 1 The framework of spontaneous learning assessment based on facial emotion recognition

The emotion is recognized by the proposed lightweight classifier based on the domain adaptation strategy, which can recognize the students facial expressions in real time. To reduce the influence by the variations of resolution, definition, and complex backgrounds in the captured images, we introduce the attention mechanism SE module.

The recognized emotions are then used to judge the student's real-time concentration and to evaluate the corresponding learning effect level. When the student's concentration is judged to be low based on facial expression recognition, real-time reminding can be conducted to improve the student situation.

At the same time, with this captured emotions, teachers can also master the real-time concentration report and the input analysis report of each student, which can help students understand the classroom knowledge better and pertinently.

In the facial expression recognition model, we recognize four most common expressions in the MOOC scenario, which are understanding, neutral, disgust, and doubt, based on the images captured by the camera in the learning process. Based on the four expressions, we propose an evaluation system of the student learning effect. Psychological research shows that positive emotions in the learning process can improve the learning effect, while negative emotions can reduce the learning efficiency. Combined with MOOC's learning environment and learners' psychological analysis, we quantify four kinds of expressions, i.e., assigning understanding with 1 point, doubt with 0.7 point, neutral with 0.5 point, and disgust with 0.1 point. The final score is calculated as

$$score = \frac{1 \times e_1 + 0.7 \times e_2 + 0.5 \times e_3 + 0.1 \times e_4}{(e_1 + e_2 + e_3 + e_4)}.$$

where e_1, e_2, e_3 and e_4 denote the times of occurring understanding, neutral, disgust, and doubt, respectively. Finally, according to the score calculation, the learning engagement is evaluated by Table 1.

3 Face expression recognition based on domain-adaptive CNN

In this section, we present the proposed a lightweight attentional convolutional network for face expression recognition. To address the issue of insufficient training data, we extend the proposed model using a domain adaption technique to explore the additional facial images, which facilitates the training process in the case of limited data.

In the MOOC scenario, the photo captured by the webcam is often under a complex background. Since we only need to focus on the facial region to interpret the expression, we need to reduce the impact of the complex background. To achieve the goal, a soft attention module is adopted in our network. The whole architecture of the proposed network is illustrated in Fig. 2, which consists of two parts: feature extraction and classification.

Figure 2 indicates that the proposed model accepts two inputs, where the first input is the source data sampled from the auxiliary training dataset with labels, and the second input is the target data sampled from the application data without labels. The objective is to explore the features of the

Table 1 Evaluation of learning

Rank	Score	Evaluation and explanation
Level 1	0.7–1	Great: high-level concentration
Level 2	0.5–0.7	Not bad: middle-level concentration
Level 3	0–0.5	Not so well: low-level concentration

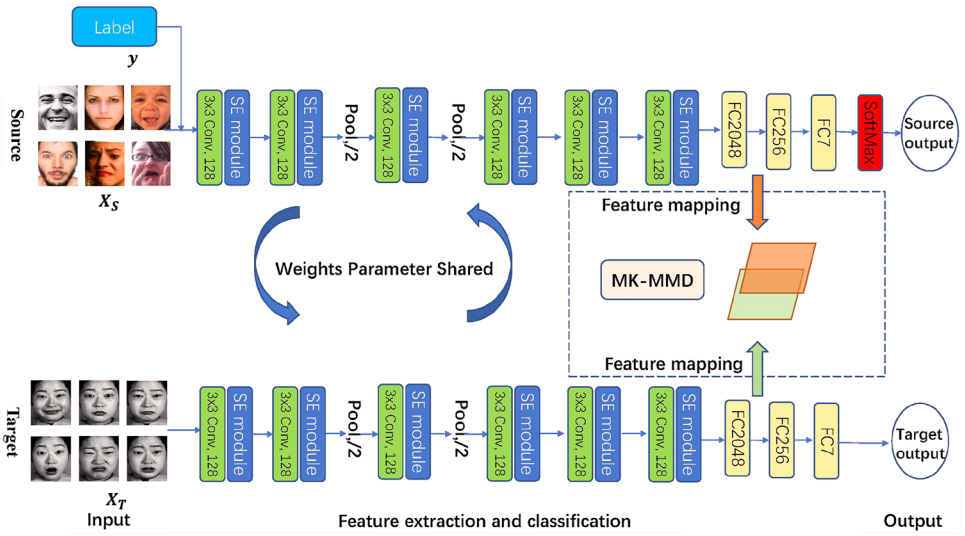

Fig. 2 The algorithm of facial expression recognition based on attentional convolutional network

source data which are used to assist the identification of the target data. We setup a parameter-sharing strategy, that is, the feature extraction network of source domain shares all the weight parameters with the target domain. The parameters are trained using a domain adaptation constraint, as discussed in Sect. 3.2.

We employ the same network for the data of both source and target domains. In training process, the feature vectors A and B of X_S and X_T are extracted from the source domain data X_s and the target domain data X_T, respectively, which are input to the network in the domain adaptation layer. Because of the parameter sharing strategy, the distribution difference of the feature vectors in the adaptation layer could reflect the difference between original source data and target data. That is, the feature vectors A and B can characterize the relations between the input data. Based on this, the Multi-Kernel Maximum Mean Discrepancies (MK-MMD) [28, 29] is used to calculate the distribution distance between the extracted features, where the distance is regarded as the distribution difference between the two domains. When the parameter-sharing convolutional neural network is updated, the loss generated by X_S and y can be propagated to the target domain through MK-MMD. Hence, the label of X_T can be inferred, yielding the category prediction of the facial expression of the target domain input.

The details of all layer settings are presented in Table 2. Specifically, since the webcams used by different learners have different resolutions, we set the size of the network input to 56×56. After two convolution stages and a max pooling stage with the stride of 2 for the dimensionality reduction, we build up deeper stages using the convolution layers with 256 filters and the convolution layers with 128 filters, which could hierarchically extract the features, producing high-level semantics. All the 6 convolutional blocks consist of the feature extraction part,

Table 2 Network structure and parameter information

Type	Filters	Size	Output	Parameters
Convolution	64	3×3	$56 \times 56 \times 64$	2048
SE module	–	–	$56 \times 56 \times 64$	580
Convolution	64	3×3	$56 \times 56 \times 64$	37,184
SE module	–	–	$56 \times 56 \times 64$	580
Max pooling	–	$3 \times 3/2$	$28 \times 28 \times 64$	0
Convolution	256	3×3	$28 \times 28 \times 256$	148,736
SE module	–	–	$28 \times 28 \times 256$	8464
Max pooling	–	$3 \times 3/2$	$14 \times 14 \times 256$	0
Convolution	256	3×3	$14 \times 14 \times 256$	591,104
SE module	–	–	$14 \times 14 \times 256$	8464
Convolution	128	3×3	$14 \times 14 \times 128$	295,552
SE module	–	–	$14 \times 14 \times 128$	2184
Convolution	128	3×3	$14 \times 14 \times 128$	148,096
SE module	–	–	$14 \times 14 \times 128$	2184
Max pooling	–	$3 \times 3/2$	$7 \times 7 \times 128$	0
Full connection	–	–	$1 \times 1 \times 2048$	12,855,296
Full connection	–	–	$1 \times 1 \times 256$	525,568
Full connection	–	–	$1 \times 1 \times 7$	1799

and each one composed of a convolutional layer, a batch normalization (BN) layer [30], and rectified linear unit (ReLU) layer [31]. BN adopts a learnable transformation and reconstruction method, so that the data distribution of the middle layer does not change in the training process, which not only speeds up the convergence speed of the network but also yields a higher initial learning rate. We apply the ReLU function as the activation function, which helps the network converge more rapidly and avoids the gradient disappearance problem caused by other activation functions. The $7 \times 7 \times 128$ feature map generated by the last convolutional stage is then input to the fully

connected layer with 2048 and 256 neurons which is called the domain adaptation layer.

Clearly, the domain adaption layer contains highly abstracted features with a high dimension, and the fully connected layer can reflect the difference of feature distribution extracted for different datasets. As such, the domain adaptation strategy can be implemented in this fully connection layer to compute the correlations between different inputs. Besides, we add a fully connected layer with 256 neurons to characterize the nonlinear property of the features, and the third fully connected layer with 7 neurons corresponding to the 7 expression categories. The SoftMax function is used to classify the expression vector, which converts the expression vector into the category probabilities.

3.1 Attention-based feature extraction

To reduce the influence of complex background on expression recognition in MOOC scene. The convolutional layers with squeeze and excitation (SE) [22] are employed to extract features in consideration of the attentional mechanism. The interdependence of the features between different channels can be established by the SE module without introducing new spatial dimensions for feature fusion. During the network training process, the importance of different feature channels can be automatically judged by learning, which can choose the most effective features from the maps.

Formally, it is assumed that the output of a convolution stage is $Z \in \mathrm{R}(H \times W \times C)$ with the spatial dimension $(H \times W)$ and the channel dimension C. The SE module produces a C-dimensional weight vector which is channel-wisely multiplicated with Z, resulting in a stronger feature representation $\widetilde{Z} \in R(H \times W \times C)$.

The detailed architecture of the SE module is illustrated in Fig. 3 As seen, the feature compression is performed in the spatial dimension, where the $W \times H \times C$ dimensional feature map Z is compressed into $1 \times 1 \times C$ dimensional vector through global average pooling. The compressed vector has a global receptive field, and the value C represents the amount of information of the original feature map Z.

After compression, we get the global description of the feature map, and then excite it using two fully connection layers. The size of the first fully connected layer is C/r, which is mainly to reduce the complexity of the model and improve the generalization ability. It reduces the channel dimension to $1/r$ of the input and recompact the feature information. The second fully connected layer promotes the $(1 \times 1 \times C)/r$ dimension back to $1 \times 1 \times C$ dimension, which learns the importance levels of different channels through nonlinear mapping. Between two FC layers, we use the ReLU function to reduce the calculation and prevent the gradient disappearance problem. Through a Sigmoid function, the SE module outputs a weight vector which is normalized to $0 \sim 1$. The normalized weight vector is multiplicated with the input features in a channel-wise manner. In this way, the features of different channels have different contributions on facial expression, resulting in efficient feature extraction abilities with less network parameters.

In the MOOC scenario, the SE module can help the model focus on the important facial features even in complex backgrounds. The network will automatically enhance the features which are more useful for the final objective. Although it increases the parameter number of the model and calculation to a certain extent, the cost is within an acceptable range.

3.2 Classification based on domain adaptation

Domain adaptation is a transfer learning method based on feature mapping, which projects the data features from the source feature space to the target feature space and tries to minimize the distribution difference between domains. Using this, the target domain data can be classified using the knowledge of the source domain.

In the MOOC scenario, the available data of the online users are very rare, and the label information is generally absent. Hence, we propose to use the existing large number

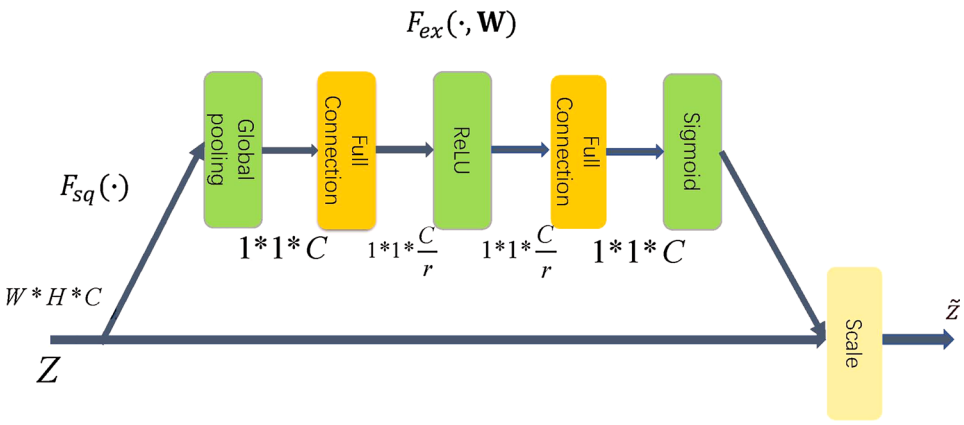

Fig. 3 The SE module

of labeled public data as the source domain datasets, and adapt the source information to the target domain, i.e., the data of online learners. Targeting at this, we next introduce the MK-MMD method in detail.

MK-MMD is a multiple kernel variant of the maximum mean discrepancies, which minimizes the distribution difference of the source and target domains in the reproducing kernel Hilbert space (RKHS). This method constructs a total kernel function by introducing multiple kernel functions, yielding improved feature characterization capabilities. The formulation of MK-MMD is written as

$$MMD_{MK}\left[H,p,q\right] = \left\| E_p\left[\Phi(X_s)\right] - E_p\left[\Phi(X_t)\right] \right\|_H^2,$$

where E_p is the expectation with respect to the distribution p of the source data, and E_q is the expectation with respect to the distribution q of the target data. H is the assemblage of the space feature mapping functions, which is characterized by the kernel function $k(X^s, X^t) = \Phi(X^s), \Phi(X^t)$. By involving multiple kernels, we have a combined kernel expression via the convex combination, which is

$$K := \left\{ k = \sum_{u=1}^{d} \beta_u k_u, \sum_{u=1}^{d} \beta_u = 1, \beta_u \geq 0, \forall u \right\},$$

where d is the number of kernels, β_u represents the weights of different kernels. The kernel k can be chosen by the optimal kernel selection strategy. For example, the Gaussian kernel function is denoted as

$$k(X^s, X^t) = exp(-\left\|X_s - X_t\right\|/2\sigma^2),$$

where σ is the bandwidth of the Gaussian kernel obtained by the median algorithm.

While the source domain and the target domain share the same parameters for feature extraction, the features generated in the source domain cannot be directly adapted to the target domain based on fine-tuning only with limited supervision of the source data. Instead, the distribution difference between the source and target features is minimized by multi-kernel-based MMD on the fully connected layers fc7–fc8. The MK-MMD-based multi-layer adaptation regularizer is added to the loss function, which can be represented as

$$L = L_c(y_p, y) + \lambda MMD_{MK}^2\left[H,p,q\right],$$

where $L_c(y_p, y)$ is the classification loss function, y_p is the label predicted by the network, y is the ground-truth label, λ is a penalty parameters. Note that the multi-layer adaptation regularizer is imposed on both the fc7 layer and the fc8 layer. Using MK-MMD in the learning process, the distributions of the source domain and the target domain can be similar in the resultant feature space.

3.3 Training

The training procedure is briefly discussed. We trained source and target data on the Intel(R) Core(TM)i9 CPU and two Nvidia Tesla 2080Ti GPUs. We use the SGD optimizer and the momentum is 0.9. The learning rates should be tuned to further optimize the network. The initial optimizer of the learning rate is 0.008, and the learning rate decreases by 0.0001 every one iteration. Each model is trained for 500 epochs from scratch. Data augmentation is used for the images in the training sets to train the model on a larger number of images, and make the trained model for invariant on small transformations.

4 Experimental results

In this section, we present the experimental details including the datasets, the evaluation metrics, and the results.

4.1 Dataset and data preprocessing

We use three datasets to train the network: JAFFE, ck+, and RAF-DB, which include seven kinds of expressions: Anger, Disgust, Fear, Happy, Sad, Surprise, and Neutral.

The RAF-DB dataset contains 15,339 images, which is used as the source domain training data. Due to the sample imbalance of different categories in RAD-DB, the number of samples of the happy and normal expressions is reduced by undersampling method, and the number of samples of the angry, disgusted, and afraid expressions is increased by data enhancement methods such as random rotation and horizontal flip. The balanced data set contains 14,535 images. We use the JAFFE and ck+ datasets as the target domain datasets.

JAFFE contains the expression samples of 10 Japanese women collected in the laboratory environment, with a total of 213 images. Each person has 3 to 4 images, and the distribution of categories is relatively balanced. ck+ is acquired by Patrick Lucey of Carnegie Mellon University on the basis of CK (Cohn-Kanade dataset). ck+ includes 593 expression sequences of 123 subjects, of which 1191 are selected.

To evaluate our proposed system, the dataset consists of videos of multiple learners watching MOOC materials. The average duration of each video is 10 min. After preliminary screening, 162 images are selected as the target domain for testing. The videos of learners are assessed in diverse scenarios, such as in the room, on the playground, and in the lab. The tools, such as webcam, computer camera, mobile phone camera, are employed to capture data at any time of a day, which may result in different resolutions and illumination. The resultant dataset contains 98 videos, which are considered as the target data. The facial images are extracted

from the videos automatically every 2 s. Four kinds of emotions are considered for recognition, including "understanding", "doubt", "neutral", and "disgust".

4.2 Evaluation metric

Since facial expression recognition in the MOOC scenario is considered as a classification problem, confusion matrix and accuracy are utilized for evaluating the performance of our method. Here are a few concepts to explain first: TP (True Positive) is the case of the ground-truth is 1 and the prediction is 1; FN (False Negative) is the case of the ground-truth is 0 and the prediction is 0; FP (False Positive) is the case of the ground-truth is 0 and the prediction is 1; TN (True Negative) is the case of the ground-truth is 1 and the prediction is 0. Classification accuracy is adopted as the evaluation metric, which is defined as

$$Accuracy = \frac{TP + FN}{TP + FN + FP + TN}.$$

Classification error is defined as

$$Error = \frac{TN + FP}{TP + FN + FP + TN}.$$

Confusion matrix indicates which two categories are easily confused to the model. In a multi-class case, the accuracy rate is computed as

$$Accuracy = \sum_{i=1}^{7} N_i / N,$$

where i is the $i-th$ category, N_i is the number of the correctly classified samples in $i-th$ category, and N is the total number of the target images.

4.3 Results and analysis

4.3.1 Performance of the model for facial expression recognition

Before reporting the performance of our learning-assisted system, we first show the performance of the proposed face expression recognition algorithm. The intuition of our proposed network (named as SE-DAN) is to tackle with the issue of insufficient labeled samples in target domain. To evaluate the performance of our algorithm, we use the source data with labels and the target data without labels to train the model, where public datasets are employed, such as JAFFE [32], ck+ [33] and RAF-DB [34]. Since the RAF-DB dataset contains more than ten thousand images, it is selected as the source dataset. The JAFFE and ck+ datasets are utilized as the target datasets. We compare our algorithm with three competitive baselines, including (1) Alexnet [17], (2) VGG-16 [35], (3) SE-CNN [22], (4) DAN [24]. From Table 3, we observe that our algorithm not only has better generalization ability compared with the competitors, but also can extract the facial expression features more effectively. The accuracy of our algorithm is higher than the others, validating that our designed network is more suitable

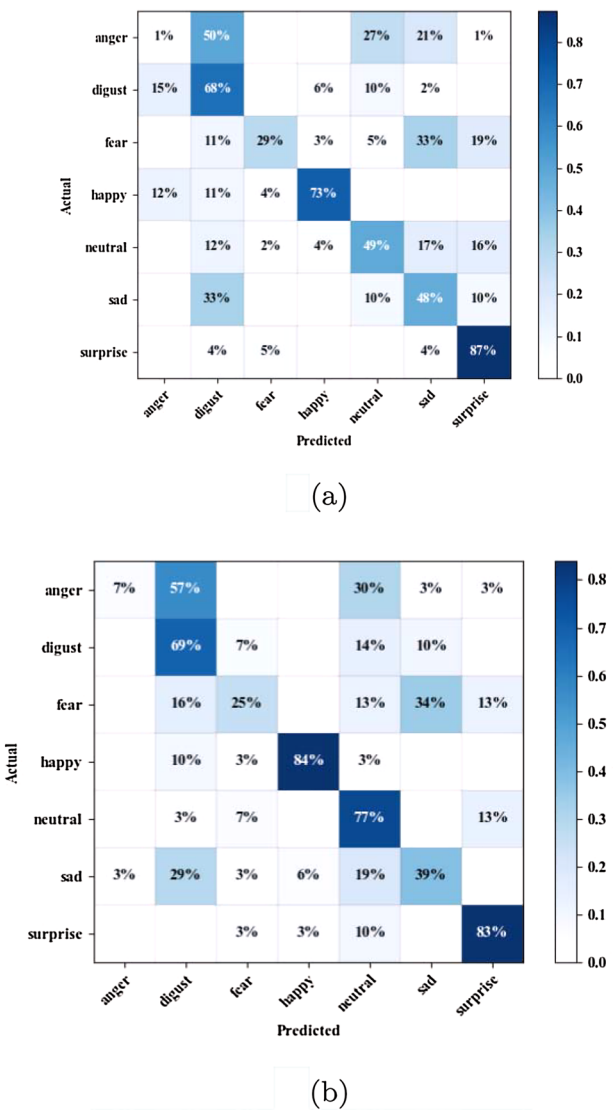

Fig. 4 The confusion matrixes of different target datasets

Table 3 The accuracy of different methods

Method	Accuracy (JAFFE)	Accuracy (ck+)
Alexnet [17]	0.44	0.45
VGG-16 [35]	0.46	0.46
SE-CNN [22]	0.48	0.53
DAN [24]	0.49	0.53
Ours	0.51	0.54

Table 4 The discussion of hyper-parameter r in SE block (JAFFE dataset)

r	Accuracy
2	0.4513
4	0.4618
8	0.4635
16	0.4772
32	0.4543

Table 5 The discussion of hyper-parameter r in SE block (ck+ dataset)

r	Accuracy
2	0.5102
4	0.5172
8	0.5200
16	0.5330
32	0.5156

Table 6 The accuracy of facial expression recognition with varies of λ in JAFFE dataset

λ	Accuracy
0.2	0.4818
0.4	0.48987
0.6	0.50881
0.8	0.49539
1	0.46395

Table 7 The accuracy of facial expression recognition with varies of λ in ck+ dataset

λ	Accuracy
0.2	0.5096
0.4	0.5205
0.6	0.5467
0.8	0.5305
1	0.5263

for facial expression recognition. The confusion matrixes are shown in Fig. 4, where the JAFFE dataset is the target domain in Fig. 4a, and ck+ is the target domain in Fig. 4b. The data distributions of "angry" and "fear" in the target domain are significantly different from that of the source domain, so the recognition accuracy is rather low. "Disgust", "happy", and "surprise", which have similar distributions between the source and target domains, yield high recognition accuracy.

4.3.2 Ablation experiments

Hyper-parameter r in SE block The hyper-parameter r denotes the reduction ratio, which not only determines the capacity of feature representation in the SE block, but also influences the computational cost. Thus, the experiments are conducted to investigate the appropriate value of r with the leverage of capacity and computational cost. Tables 4 and 5 show the performance of SE-CNN without involving the domain adaptation strategy in the datasets of JAFFE and ck+, respectively. The selected values of r include 2, 4, 6, 8, 16, and 32. From the comparison, the accuracy of facial expression recognition increases firstly and then decreases with the increase of r. As a result, we could achieve the best performance when r is set to 16. Consequently, $r = 16$ is utilized for all experiments.

Fine-tuning with domain adaptation regularization The regularization hyper-parameter λ affects the loss of the domain adaptation regularizer. If λ is set too low, it will cause minimal impact on the MMD regularizer. On the contrary, if λ is chosen with a high value, it would regularize heavily with a degenerate representation. Tables 6 and 7 show the accuracy of facial expression recognition with different λ, where λ is set as 0.2, 0.4, 0.6, 0.8, and 1. Tables 6 and 7 correspond to the results on JAFFE and ck+, respectively. These results indicate that when $\lambda = 0.6$, the source domain and target domain can be leveraged.

4.3.3 Evaluation on spontaneous learning-assisted system

For source domain dataset, 12,916 images from the RAF-DB dataset are used for training. 162 images from our dataset are selected as the target data in training and are also used to test different methods. The comparison between the proposed learning-assisted system in MOOC scenario with the previous works are provided in Table 8. As shown, we achieve the highest accuracy which is much better than the others. Visual examples are shown in Fig. 5, which indicates that the learner's expression can be detected at a fine level. Figure 6 gives the confusion matrix of our algorithm.

5 Conclusions

In this paper, we present a novel system for assessing learning engagement based on face expression recognition. To tackle the problem of the images obtained by web-camera with complicated background, illumination and resolutions, an attentional mechanism is embedded in the network for feature extraction. A lightweight architecture is designed for real-time application, resulting in the SE-CNN for classification. Moreover, domain adaptation is also employed to address the issue caused by the lack of labeled data and finally, SE-DAN is proposed for face expression recognition to assess the learning status in the MOOC scenario. Experiments show that our proposed method can recognize the emotions with high accuracy and with limited labeled data.

The following directions for future research are listed below: (1) the facial expression data involved in this paper are all static images, but dynamic video sequences

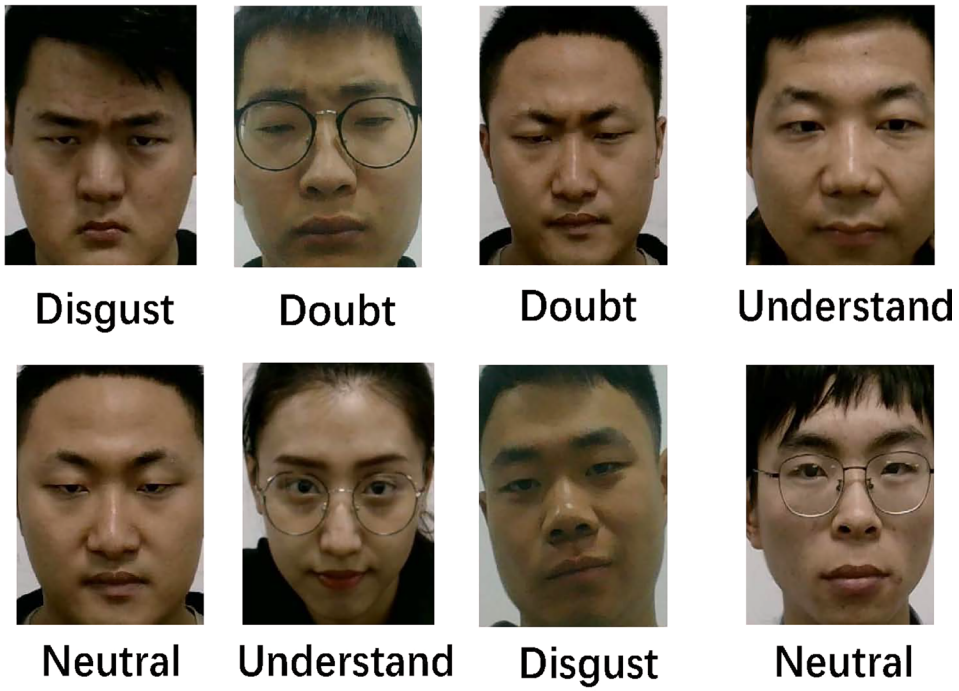

Fig. 5 The visual examples of face expression recognition

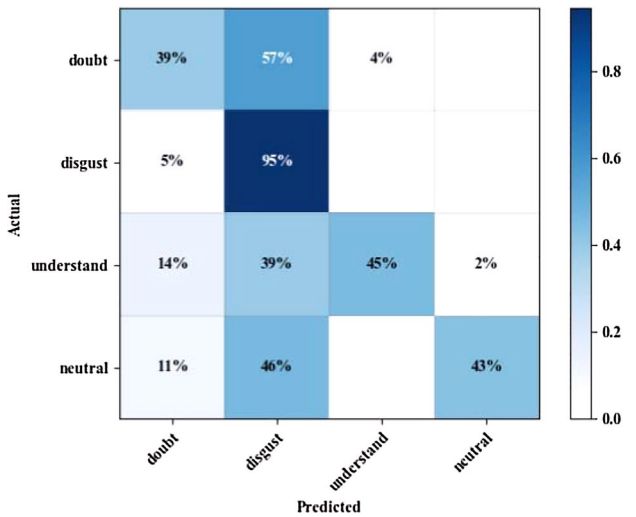

Fig. 6 The confusion matrix of our proposed method for learning assessment

Table 8 The accuracy of face expression recognition in our proposed learning-assisted system

Method	Accuracy
Alexnet [17]	0.46
VGG-16 [35]	0.47
SE-CNN [22]	0.49
DAN [24]	0.48
Ours	0.56

can contain more subtle information about the changes in facial expressions. Therefore, in future, the study of video sequences can evaluate the learning effect better. (2) The lightweight network needs to be designed for real-time detection in different hardware devices.

References

1. Brom, C., Šisler, V., Slavík, R.: Implementing digital game-based learning in schools: augmented learning environment of Europe 2045. Multimed. Syst. **16**(1), 23–41 (2009). https://doi.org/10.1007/s00530-009-0174-0
2. Murat, Z.H., et al.: EEG analysis for brainwave balancing index (BBI), second international conference on computational intelligence, communication systems and networks, pp. 28–30 (2010)
3. Nishiwaki, G.A., Urabe, Y., Tanaka, K.: EMG analysis of lower extremity muscles in three different squat exercises. J. Jpn. Phys. Ther. Assoc. **9**(1), 21–26 (2006)
4. Penzel, T.: Blood pressure analysis. J. Sleep Res. **4**(S1), 15–20 (1995)
5. Wang, L., Qiao, Y., Tang, X.: Action recognition with trajectory-pooled deep-convolutional descriptors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, pp. 4305–4314 (2015)
6. Fang, M., Bai, X., Zhao, J., et al.: Integrating Gaussian mixture model and dilated residual network for action recognition in videos. Multimed. Syst. **26**, 715–725 (2020). https://doi.org/10.1007/s00530-020-00683-4
7. Ji, S., et al.: 3D convolutional neural networks for human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **35**(1), 221–231 (2012)

8. Mitra, S., Acharya, T.: Gesture recognition: a survey. IEEE Trans. Syst. Man Cybern. Part C **37**, 311–324 (2007)
9. Singha, J., Laskar, R.H.: Hand gesture recognition using two-level speed normalization, feature selection and classifier fusion. Multimed. Syst. **23**, 499–514 (2017). https://doi.org/10.1007/s00530-016-0510-0
10. Liu, H., Wang, L.: Gesture recognition for human–robot collaboration: a review. Int. J. Ind. Ergon. **68**, 355–367 (2018)
11. Kotsia, I., Pitas, I.: Facial expression recognition in image sequences using geometric deformation features and support vector machines. IEEE Trans. Image Process. Publ. IEEE Signal Process. Soc. **16**, 172–187 (2007)
12. Dhall, A., Kaur, A., Goecke, R., Gedeon, T.: Emotiw 2018: Audio-video, student engagement and group-level affect prediction. In: Proceedings of the 20th ACM International Conference on Multimodal Interaction, pp. 653–656 (2018)
13. Newell, A., Kaiyu, Y., Jia, D.: Stacked hourglass networks for human pose estimation. In: Proceedings of Springer European Conference on Computer Vision (ECCV) (2016)
14. Li, S., Deng, W.: Deep facial expression recognition: a survey. IEEE Trans. Affect. Comput. **99**, 1 (2020)
15. Siddiqi, M.H., Ali, R., Khan, A.M., et al.: Facial expression recognition using active contour-based face detection, facial movement-based feature extraction, and non-linear feature selection. Multimed. Syst. **21**(6), 541–555 (2015). https://doi.org/10.1007/s00530-014-0400-2
16. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. The MIT Press, New York (2016)
17. Krizhevsky, A., Sutskever, I., Hinton, G.: Imagenet classification with deep convolutional neural networks. Adv. Neural. Inf. Process. Syst. **25**(2), 1097–1105 (2012)
18. Takalkar, M.A., Xu, M., Chaczko, Z.: Manifold feature integration for micro-expression recognition. Multimed. Syst. **26**, 535–551 (2020)
19. Pfister, T., et al.: Recognising spontaneous facial micro-expressions. In: Proceedings of the IEEE 2011 International Conference on Computer Vision (ICCV) (2011)
20. Jaderberg, M., Simonyan, K., Zisserman, A.: Spatial transformer networks. In: Advances in Neural Information Processing Systems, pp. 2017–2025 (2015)
21. Chorowski, J.K., et al.: Attention-based models for speech recognition. Adv. Neural. Inf. Process. Syst. **10**(4), 429–439 (2015)
22. Hu, J., Shen, L., Sun, G.: Squeeze- and -excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7132–7141 (2018)
23. Georghiades, A.S., Belhumeur, P.N., Kriegman, D.J.: From few to many: illumination cone models for face recognition under variable lighting and pose. IEEE Trans. Pattern Anal. Mach. Intell. **23**(6), pp. 643–660 (2001)
24. Long, M., Cao, Y., Wang, J., et al.: Learning transferable features with deep adaptation networks. In: International Conference on Machine Learning (ICML), vol 37, pp. 97–105 (2015)
25. Ganin, Y., Victor, L.: Unsupervised domain adaptation by backpropagation. In: International Conference on Machine Learning, pp. 1180–1189 (2015)
26. Tzeng, E., et al.: Adversarial discriminative domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
27. Otero, V., Steven, P., Noah, F.: A physics department's role in preparing physics teachers: the Colorado learning assistant model. Am. J. Phys. **78**(11), 1218–1224 (2010)
28. Sejdinovic, D., Sriperumbudur, B., Gretton, A., et al.: Equivalence of distance-based and RKHS-based statistics in hypothesis testing. Ann. Stat. **20**, 2263–2291 (2013)
29. Muandet, K., Fukumizu, K., Sriperumbudur, B., Schölkopf, B.: Kernel mean embedding of distributions: a review and beyond (2016). arXiv:1605.09522
30. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning (ICML) (2015)
31. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: International Conference on Machine Learning (ICML) (2010)
32. Lyons, M.J., Akamatsu, S., Kamachi, M., et al.: Coding facial expressions with gabor wavelets. In: The 3rd IEEE International Conference on Automatic Face and Gesture Recognition, pp. 200–205 (1998)
33. Lucey, P., Cohn, J.F., Kanade, T., et al.: The extended Cohn–Kanade dataset (ck+): a complete dataset for action unit and emotion-specified expression. In: Proceedings of the IEEE Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 94–101 (2010)
34. Li, S., Deng, W., Du, J.P.: Reliable crowd sourcing and deep locality-preserving learning for expression recognition in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2584–2593 (2017)
35. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (ICLR), pp. 1768–1776. MIT Press, Cambridge (2015)