Mathematical Biosciences and Engineering

MBE, 20(5): 8208–8225.
DOI: 10.3934/mbe.2023357
Received: 30 December 2022
Revised: 05 February 2023
Accepted: 09 February 2023
Published: 27 February 2023

http://www.aimspress.com/journal/MBE

Research article

Facial expression recognition using lightweight deep learning modeling

Mubashir Ahmad[1,2,*,†], Saira[2,†], Omar Alfandi[3], Asad Masood Khattak[3,*], Syed Furqan Qadri[4,*], Iftikhar Ahmed Saeed[2], Salabat Khan[5], Bashir Hayat[6] and Arshad Ahmad[7]

[1] Department of Computer Science, COMSATS University Islamabad, Abbottabad Campus, Tobe Camp, Abbottabad-22060, Pakistan
[2] Department of Computer Science, the University of Lahore, Sargodha Campus 40100, Pakistan
[3] College of Technological Innovation at Zayed University in Abu Dhabi, UAE
[4] Research Center for Healthcare Data Science, Zhejiang Lab, Hangzhou 311121, China
[5] College of Computer Science & Software Engineering, Shenzhen University, Shenzhen 518060, China
[6] Department of Computer Science, Institute of Management Sciences, Peshawar, Pakistan
[7] Department of IT & CS, Pak-Austria Fachhochschule: Institute of Applied Sciences and Technology (PAF-IAST), Haripur 22620, Pakistan

Correspondence: Email: mubashir_bit@yahoo.com, asad.khattak@zu.ac.ae, furqangillani79@gmail.com.

† These two authors contributed equally.

Abstract: Facial expression is a type of communication and is useful in many areas of computer vision, including intelligent visual surveillance, human-robot interaction and human behavior analysis. A deep learning approach is presented to classify happy, sad, angry, fearful, contemptuous, surprised and disgusted expressions. Accurate detection and classification of human facial expression is a critical task in image processing due to the inconsistencies amid the complexity, including change in illumination, occlusion, noise and the over-fitting problem. A stacked sparse auto-encoder for facial expression recognition (SSAE-FER) is used for unsupervised pre-training and supervised fine-tuning. SSAE-FER automatically extracts features from input images, and the softmax classifier is used to classify the expressions. Our method achieved an accuracy of 92.50% on the JAFFE dataset and 99.30% on the CK+ dataset. SSAE-FER performs well compared to the other comparative methods in the same domain.

Keywords: classification; deep learning; facial expression recognition; machine learning; stacked sparse auto-encoder

1. Introduction

Currently, FER is a major communication source which is the most important tool to interact with others without knowing their gender, race and national borders. FER comes under the category of non-verbal communication, which delivers a person's feelings in the form of gestures and body language [1] that allow people to infer others' feelings. Many of us can understand facial expressions from key emotions, but others cannot [2]. A robust emotion classification relies heavily on effective facial representations, so it is quite a challenging task to identify significant discriminative facial features that might demonstrate the appearances of each emotion because of the limitations and variability of facial expressions. In addition, the FER system has many challenges, such as difficulties occurring when the face is not in an accurate position, lighting problems while capturing the image, obstruction, noise and over-fitting [3]. Although there are several challenges in the existing FER system, still, it is a fascinating area that could help in different fields of life, such as online education, health management, audience analysis in the market and entertainment. It also helps in security, driver drowsiness detection, etc. S. H. Ma et al. [4] studied facial expressions and classified them into seven categories, which are happy, sad, angry, fearful, surprised, contemptuous and disgusted. During the last five years, different researchers have adopted different methods while working on FER. Some of them considered different types of data used as input to expression recognition systems. In light of perceptions, we observed that facial expressions are typically associated with emotions. For example, lifting the eyebrows indicates a sudden change that means the expression of fear or surprise. The image of human facial expressions is the standard and promising type of information, which gives many unexpected results, applying different algorithms on it such as convolutional neural network (CNN), artificial neural network (ANN), generative adversarial network (GAN), etc. These algorithms give results according to our desired expressions. However, it does not matter how an image is captured or which device is utilized for capturing the image, such as a digital camera, camcorder or DSLR. Human emotions can be perceived from numerous points of view, such as physiological and psychological signs. The FER might be a part of a telepresence bundle which provides information about the patient's mental condition to his doctor or counselor. When something is discovered to be wrong, the professional will choose to change his or her approach to patient care. According to [5], human facial expressions, visual contact, body language, voice tone and personal space are included in nonverbal communication. Since nonverbal communication is 55 to 94% of actual communication [6], positive nonverbal communication can help to boost interpersonal relationships and emphasize emotional bonds [7] proposed a deep learning method for facial recognition for distance learning that extracts features from the images and produced better results. Some recent deep learning methods have also been used for recognition in different fields like medical image classification and segmentation [8–17].

Over-fitting is the problem with existing FER systems which occurs when the model training has no issues, but they do not predict the indicated results at the time of testing. Many research authors are going to minimize the drawbacks of the prevailing systems, which tend to offer quick-expression classification in the same way. In the contemporary era, computer analysis of the face and facial expressions is the most commonly used issue among specialists. Facial expressions are one of the more significant aspects of human communication, with the face being responsible for communicating, not only for considerations but also for emotions. Over-fitting, due to a lack of training data, remains a major challenge that must be addressed by all deep learning FER systems to achieve high accuracy. The present study is conducted on the SSAE-FER framework for facial expression recognition. The main contributions are as follows.

i. The SSAE-FER model is used for the recognition of seven basic facial expressions. All the images are the same size for input into the model, which could give us better results than other comparison methods.

ii. Our model is a two-layered stacked auto-encoder (SAE) which is lightweight, and the training and testing time is very much lower due to its simplicity.

iii. The performance of the proposed system was assessed on two different datasets, JAFFE and CK+, and both datasets are publically available.

iv. Enhanced and adequate results were attained in terms of accuracy, specificity and sensitivity.

The remaining portion of this paper is organized as follows: Section 2 discusses related work. Material and methods are discussed in Section 3, Section 4 presents the results and discussion, and the Section 5 concludes the paper.

2. Related work

There are many machine learning and deep learning algorithms which have a great impact on facial expression recognition (FER) systems. Deep learning has different types of algorithms which play a vital role in FER, like generative adversarial networks (GAN), deep belief networks (DBN), stacked sparse auto-encoder (SSAE), conventional neural networks (CNN), recurrent neural networks (CNN-RNN) architecture [18], etc. Similarly, machine learning has different methods and classifiers which are used for classification, like support vector machine (SVM), k-nearest neighbors (KNN), decision tree (DT), weighted hierarchical adaptive voting ensemble (WHAVE), logistic regression (LR), which are used for FER, analyzing a social interaction, intelligent transportation system, fruit identification and anomaly detection [19–28]. Some feature extraction techniques have been used in the past few years, such as the active shape model (ASM) [29], used to extract features based on expression contours. A deep network was used [30] which has two further different models ,where the first model is a deep temporal appearance network (DTAN), and the other is a deep temporal geometry network (DTGN). Thus, DTAN extracts features based on temporal appearance, and another extracts deep temporal geometry network features using facial landmark points. The models are combined through a novel integration method to achieve the best accuracy for expression recognition. This network is known as the deep temporal appearance-geometry network (DTAGN). The selection of pairwise features and their classification is discussed in [31], and [32] introduced a peak-piloted deep network in which peak and non-peak expressions are involved. In this work, peak expressions supervise the recognition of non-peak expressions, but it can only distinguish the same expressions of the same subject. The process of non-peak-to-peak expression is indirectly inserted into the network to get the invariance to expression intensities. A back-propagation method, peak gradient suppression (PGS), is utilized for training the network.

Automatically recognizing facial expressions [33] is an interesting and important part of human-machine interaction, where [34] introduced the CNN and landmark feature technique for 3D facial expression recognition. [35] proposed a novel model for facial expression recognition (FER) using the color scheme and deep information through the Kinect sensor. The proposed system extracts the different features of facial expression and utilizes captured sensor information; it emphasizes vectors by face tracing algorithm and perceives the six facial emotions using the random forest (RF) algorithm. The implementation of RF is utilized for facial expression recognition execution for real-time scenarios. A novel deep-learning method for facial expression recognition is introduced in [36]. The training images were divided into seven groups, due to seven expressions, to train the sparse autoencoder network. Interestingly, a graph convolution neural network (GCNN) successfully recognized the object, text classification and human activities, so GCNN is used to represent the features. Euclidean distance is used to find out the shortest path between edges and joints; a convolutional neural network (CNN) deals with Euclidean data and performs further work on it. A spatial domain convolution kernel is stretched out to graph convolutional kernels to process the number of items over neighbor nodes. The application of pooling is used in the hidden layers of neighbor nodes to complete the data structure on incomplete nodes. The balance cuts and heavy edge matching (HEM) techniques are used for graph pooling. A graph may contain excess or blurring edges, so it utilizes the mechanism to notice the critical node [37].

A convolutional neural network (CNN) combined with bag of words is used. It was successful in object detection, while for further improvement in its results, supervised and unsupervised methods are used. It has been observed that there are different objects in an image that define feature descriptors that are used for forming histograms. By creating histograms of different images that form a bag of words (BoW) that can be learned by a classifier. For the new experimentation of the model, the features were extracted from images using CNN, and then spatial pyramid matching (SPM) was applied to this information to localize the objects. They used CaffeNet, which is similar to Alex-Net, where pooling is done before normalization. It used the t-distribution stochastic neighbor embedding (t-SNE) algorithm to visualize different obtained features from the last layer of CaffeNet in a high-dimensional histogram for each image, which also allows clustering. Although t-SNE is an unsupervised method to cluster data, it is used to see how it classifies suggested data by applying the K-means clustering algorithm on top of t-SNE. Therefore, this is useful in human actions recognition and accomplished the best outcomes using contrasting, where various classification algorithms are used, like k-nearest neighbor (KNN), support vector machine (SVM), relational neighbor (RN) [38] and particle swarm optimization (PSO) [39]. Different other methods have been used in a very efficient way using optical coherence tomography (OCT) in vivo imaging [40–43].

In video-based action recognition the understanding of actions, pose, estimation, and retrieval of images from different perspectives [44]. Many types of research on the FER system have been directed at both posed and natural expressions under various imaging conditions that include a few head poses, imaging resolutions, illumination factors and occlusion [45]. Conditional generative adversarial networks (cGANs) [46] are applied to gain the images from the neutral face. However, the model has many intermediate layers in which filters remain unchanged, and different layers of the same size are concatenated and combined with the last and fully connected layer for the classification of expression that includes a display of happy, angry, fearful, natural, sad, surprise and disgust [47].

Convolutional neural networks (CNN) are used for feature extraction and classification purposes. A method known as amalgam fusion consists of two levels; the prior level is a feature and the post level is a decision. Both are implemented in a way to pool the features in one place and observe its decision at different stages. As CNN model has trained with a different voice sample of the Ryerson Audio-Visual Dataset of Emotional Speech and Song (RAVDESS) dataset [48] and then joined with an output of an image classifier by utilizing the fusion results. The attained results through decision-level which further proceed towards the final decision [49]. In this study [50] the author proposed a novel technique of color channel-wise recurrent learning which obtained an accuracy rate of 85.74% on facial expression. All the above studies formed the basis of the significant results of their own proposed methods of facial expression, but there is still a need for a simple and attractive piece of facial expressions recognition. We are going to introduce the new technique for the FER framework using SSAE-FER, which will grab the user's attention and will help them to solve their challenges regarding facial expression recognition systems.

3. Material and methods

In this section, we present our proposed method, which consists of pre-processing, training, and testing. A flow diagram of our proposed method is shown in Figure 1.

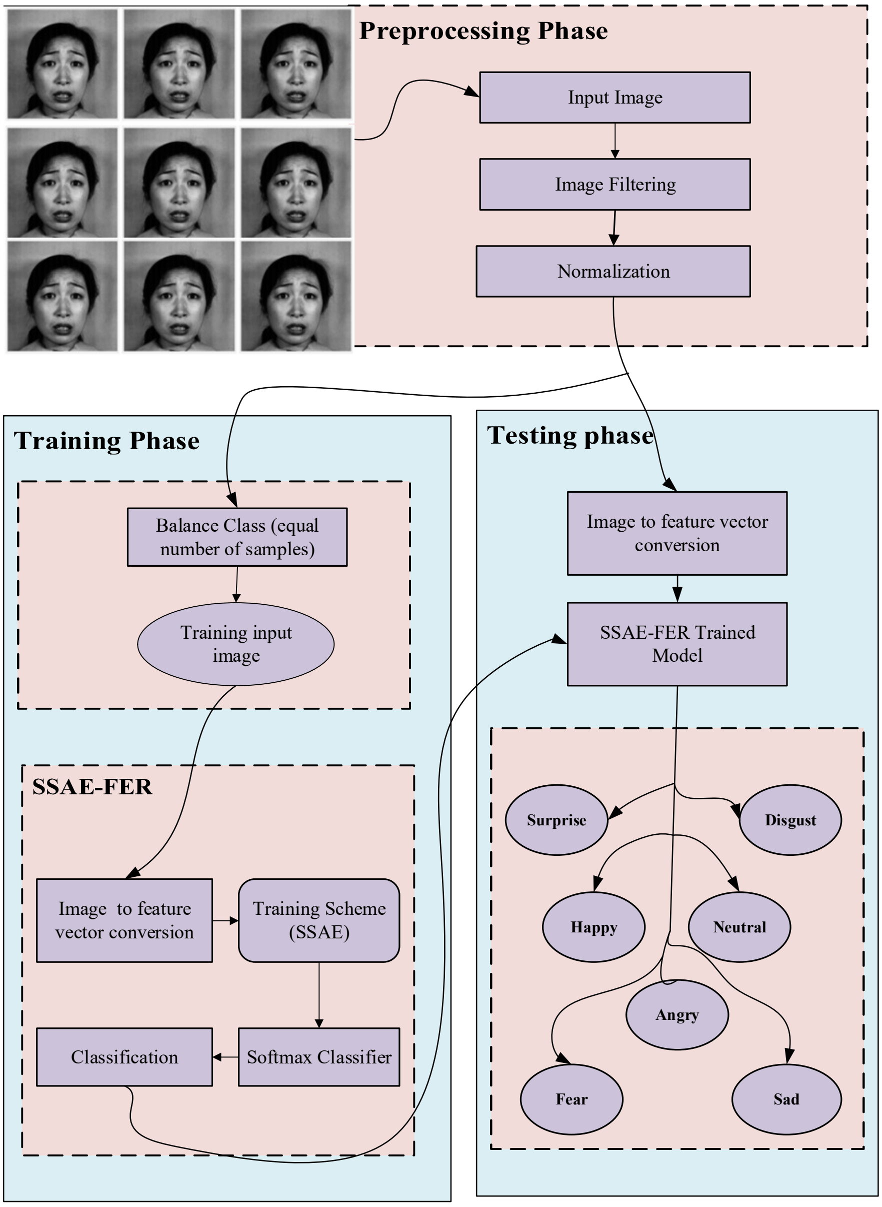


Figure 1. Block diagram of the proposed method.

3.1. Preprocessing

Image preprocessing is performed on the input dataset on a 2D facial image. The purpose of preprocessing is to enhance the images, where 2 datasets are used in this experiment. The same preprocessing is applied to both datasets. In the first step, we converted the images into grayscale, and a Gaussian smoothing filter is applied due to the noise in the original image. We set the sigma value of $\sigma = 0.5$, which is helpful for noise removal. The results of the preprocessing step are illustrated in Figure 2.

Normalizing the whole image is a good idea rather than normalizing some specific parts of the face such as the eye, lips, nose and eyebrows [51]. Normalization has a great influence on images because of different intensities, so we normalize the data using zero mean unit variance. After the normalization, we adjust the images to 0 and 1 values [52]. All the images are normalized, and we performed the class balancing method, where all the classes have the same images in the training set. This is an important step to cure overfitting [53], and our method gained effective results to prepare the input images for the training phase. Due to the preprocessing techniques, our images are more enhanced and prepared for training on the SSAE-FER model.

Figure 2. Application of Gaussian filter.

3.2. Training the SSAE-FER model

After the preprocessing step, we are ready to train our data on SSAE-FER. An SSAE is based on two phases, encoder and decoder, to extract high-level feature learning in an unsupervised manner in the first step. The original sizes of the input images were 48 × 48 and 256 × 256 pixels for the 2 individually trained datasets, and we flattened the original images to 2304 and 65,536 for further steps. There are 2 main steps in the training, which are pre-training and fine-tuning. In the pre-training step, we provided the data without ground truth, and in the second step, we trained the whole dataset with ground truths. The SSAE-FER comprises many layers of sparse autoencoder, where the output of each hidden layer is connected to the input of the successive hidden layer. The hidden layers are trained in an unsupervised algorithm and then are fine-tuned in a supervised fashion by using the stochastic gradient descent (SGD) algorithm. We train the autoencoder using input data of 48 × 48 size or 256 × 256 and acquire the learned data from it. The learned data from the previous layer is utilized as input for the next layer, and this continues until the training is completed. Once all the hidden layers are trained, the model uses the backpropagation between hidden layers by using SGD. In the proposed method, we worked on two SSAE layers. The working of SSAE in our work is shown in Figure 3.

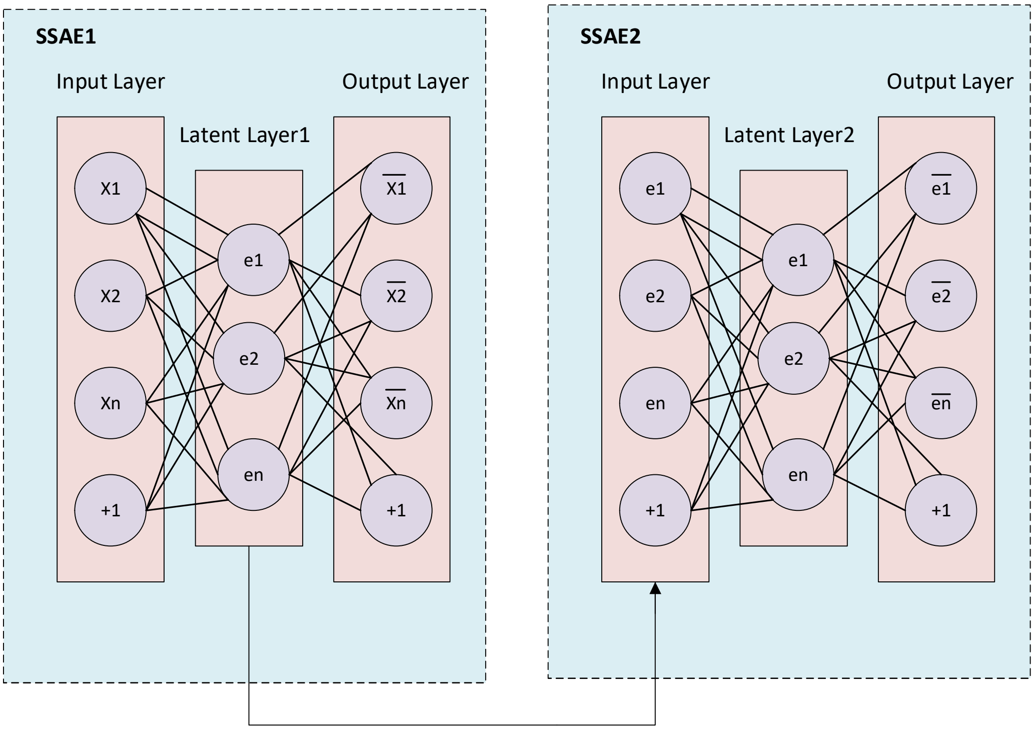


Figure 3. The structure of hidden SSAE layers.

The model performs the greedy layer-wise pre-training of data, considering a stacked autoencoder composed of n layers. The suggested model can be greedily pre-trained to initialize the parameters of the deep network, to train the first layer using the input to obtain the parameters for the first autoencoder in the stack, whereas all other parameters in the remainder of the network remain fixed. By initializing the parameters, the input can be transformed into a vector consisting of the activations (learned features) of the hidden units. The autoencoder can map the input directly to the hidden layer using a parameter called an encoder [54]. The encoding step can transform the high-dimensional input data into lower dimensions. The decoding step involves mapping these learned features from the hidden space back to the reconstruction of that input. We have demonstrated the structure of the SSAE1 hidden layers above in Figure 3, where the SSAE2 layers are stacked together. The output of the first layer of SSAE becomes the input of the next layer of SSAE. Data is compressed at latent layers, which become the input of further layers for better performance. The stacked layered network is connected to the Softmax layer, which performs the prediction of the features attained by the SSAE2. As our proposed method uses a novel technique for facial expression recognition, the autoencoder works in an unsupervised manner in pre-training. It comprises encoder and decoder: The encoder maps the input data and represents it in a new form, whereas the new form of data is then decoded at the output to regenerate the input x' as given in Eqs (1) and (2), where x is the input, and z is the new representation of the input.

$$Z = H(W_x + b) \tag{1}$$

$$X' = H(W'_x + b') \tag{2}$$

In the above Eq (1), h represents the activation function of neurons of the hidden layer. In Eq (2), g represents the neurons of the output layer, W and W' represent the weight matrices, and b and $b^{'}$ represent the bias vectors for encoder and decoder, respectively. SSAE layers have some weights W and biases b, which help to produce better results. Further parameters of the model are utilized to improve the performance of the network. Fine-tuning reduces the error rate observed from the previous epoch which is performed in a supervised manner. After the backpropagation is utilized to fine-tune the whole network, this process minimizes the error rate and refines the model enough to deal with the new samples of datasets.

3.3. Classification model

After the completion of fine-tuning, we trained the FER-SSAE model and applied the classifier Softmax on the last layer. The utilization of the Softmax function is best for multiclass data classification because it maps and predicts the values to probabilities against each expression present in the data. There are seven nodes in the last or output layer of the model that facilitate the network to choose the most desirable features for the representation of each image. Therefore, with the Softmax classifier used for the classification of expressions, this function returns the probability of each class. In our case, it gives the best recognition result against seven expressions. The equation for the Softmax activation function is given below [55].

$$Softmax(z_i) = \frac{\exp(z_i)}{\sum_i \exp(z_i)} \tag{3}$$

In Eq (3), the z is the neuron value, which it presents, in the output, and exp is the non-linear function. Later, the sum of exponential finds is used to divide the values of neurons to perform normalization and subsequently convert them into different probabilities when Softmax activation is applied on the final or last layer of each neuron to recognize the expression successfully. Figure 4 shows how two hidden layers work during the pre-training and fine-tuning stages for the classification of expressions.

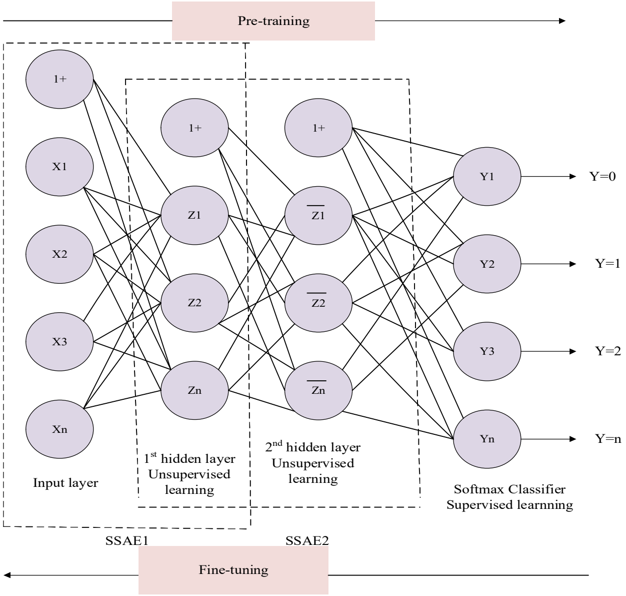


Figure 4. Structure of proposed methodology.

3.4. Testing scheme

After the completion of training the same data, input is provided for testing for the trained model. In this step, we also tested the balanced class for better test results. Some of each class sample are equally validated on our SSAE-FER model for classifications of expressions. The structure of the proposed model is given below in Figure 5, which illustrates the complete overview.

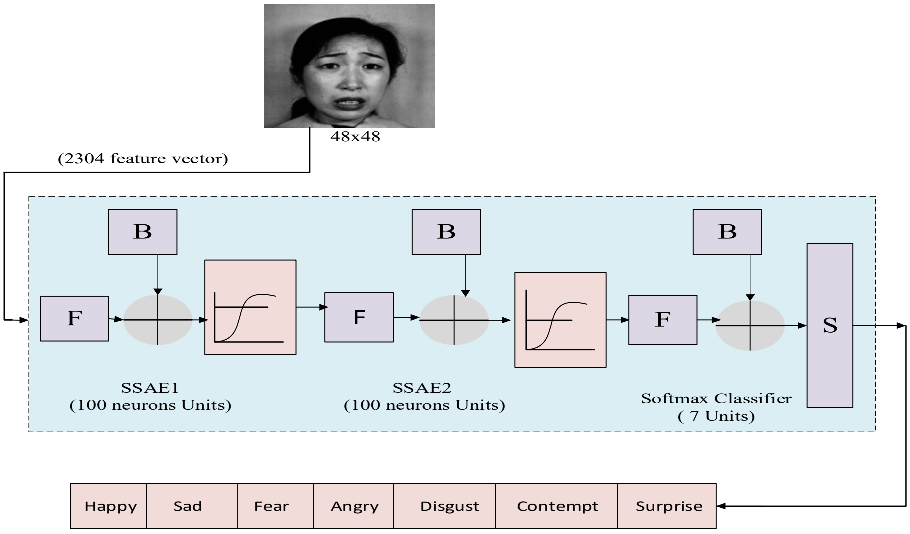


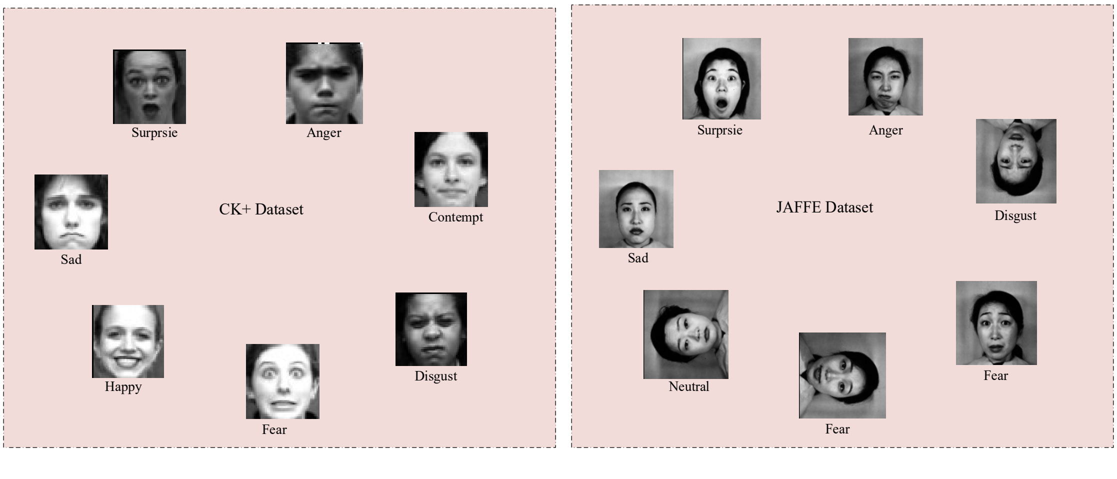
Figure 5. Structure of stacked sparse autoencoder.

4. Results and discussion

In this section, we present the datasets used in this experiment, the parametric settings for the SSAE-FER model and the details of the results.

4.1. Dataset details

JAFFE [56,57] and CK+ [58] were used for the experiment on the SSAE-FER model. The JAFFE dataset contains 10 Japanese female expressions that have seven poses: happy, sad, fear, anger, surprise, neutral and disgust. Several images of each expression are available in the dataset having 256 × 256 pixels resolution. We used a total of 213 2D grayscale images from JAFFE dataset with different classes: anger containing 30 images, disgust in 29 images, fear in 33 images, happiness in 31 images, neutral in 30 images, sadness in 31 images and 29 images containing surprise expressions. Similarly, the CK+ dataset contains 8 expressions, including seven primary expressions plus contempt expressions. The dataset comprises a total of 981 images of different classes were used in our experiment: The happy class contains 207 images, sad 84 images, anger 135 images, fear 75 images, surprise 249 images, disgust 177 images, and the contempt class contains 54 images in our proposed work. The resolution of the CK+ dataset images is 48 × 48 pixels. In [59], the JAFEE dataset is used in facial expression recognition which is available at the following link: https://zenodo.org/record/3451524#.YSSx1I4zaM8. CK+ dataset is also available publically at the following link: https://www.kaggle.com/shawon10/ckplus. A sample of CK+ & JAFFE dataset images is given below in Figure 6. This work was performed on MATLAB 2021a with a Core i7 processor (3.6 GHz CPU) and 32 GB of RAM. In this study, our FER-SSAE model was based on the MATLAB library "Deep learning Toolbox" [60].



(a) (b)

Figure 6. (a) Samples of CK+; (b) Samples of JAFFE dataset.

4.2. Parametric settings

In this section, we provide our results on the basis of our SSAE-FER model. 70% of the data was utilized to train the model, while the rest of the data was used for testing and validation. There were two hidden layers, with 100 neurons at the SSAE1 and similarly 100 neurons on the second layer, SSAE2. The final layer contains 7 neurons to find the most similar features, which help to recognize the expression in each testing image. For pre-training, we set up 200 epochs, and for fine-tuning 6000 epochs with a minimum batch size of 32 were used. The learning rate for pre-training and fine-tuning was 0.0001 with a sparsity of 0.05 and momentum of 0.9. Table 1 shows the parametric settings for our experiment. Our model took 3 hours on CK+ and 13 hours on the JAFFE dataset for fine-tuning. Table 1 shows the parameters which were used to train the SSAE-FER model. Mean Square Error (MSE) was noted at 0.06 on CK+, and the error rate was 0.02 on the JAFFE dataset during the training. The training and validation graphs are given in Figure 7 for both datasets.

Table 1. Parameter setting scheme of CK+ and JAFFE datasets.

Parametric name	Values
Hidden layers	2
Number of neurons at each layer	Layer1 & Layer2 = 100
Number of epochs	200 for pre-training and 6000 for fine-tuning
Learning rate	0.0001
Momentum	0.9
Mini batch size	32
Sparsity	0.5

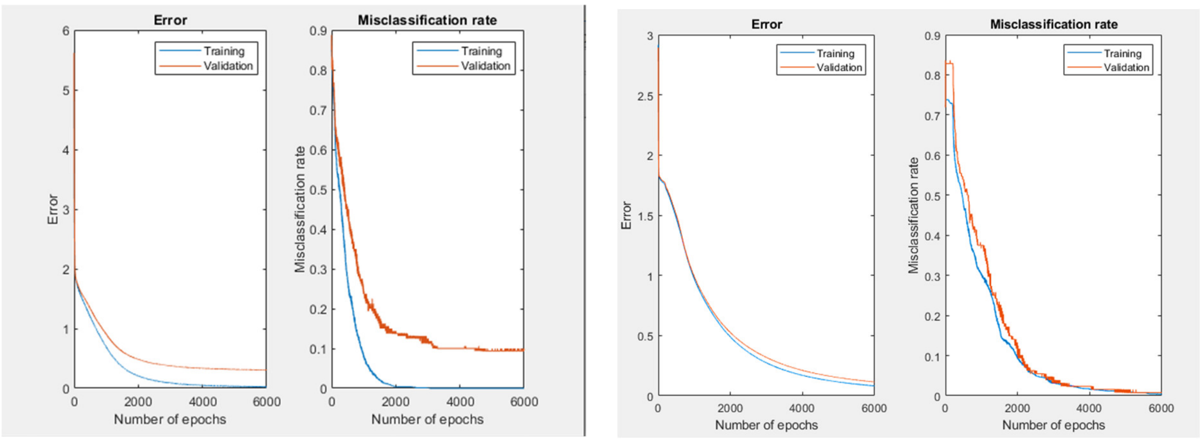


Figure 7. Learning curves during the fine-tuning: (a) Training and validation on CK+; (b) Training and validation on JAFFE dataset.

4.3. Performance evaluation

The following boundaries are given to assess the exhibition of our planned model.

Accuracy: the proportion of the total number of right expectations. It consists of the prediction of seven human expression samples. We can calculate it by using the following equation:

$$\frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{4}$$

Error rate: the total number of predicted cases that were not correct. It consists of both positive and negative samples of seven expressions. We can calculate it by using the following equation:

$$\frac{\mathrm{FP} + \mathrm{FN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{5}$$

Sensitivity or Recall: the proportion of genuine positive occasions effectively recognized is called the true positive rate. We can calculate it by using the following equation:

$$\frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6}$$

Precision: Precision is the ratio of true positives to the total of false positives and true positives. We can calculate it by using the following equation:

$$\frac{\text{TP}}{\text{TP} + \text{FP}} \tag{7}$$

Specificity: This is the proportion of negative occasions accurately distinguished and is also known as the negative rate. We can calculate it by using the following equation:

$$\frac{\text{TN}}{\text{TN} + \text{FN}} \tag{8}$$

True positives (TP) represent those expressions that are correctly identified. False positives (FP) represent the expressions that do not belong to their respective class but the model identifies them as a part of it. True negatives (TN) represent the images that do not belong to another class and are correctly identified as belonging to other classes. False negatives (FN) represent the expressions that belong to a class itself but are identified as another class expression. The results of our model are presented in Tables 2 and 3 for CK+ and JAFFE datasets, respectively.

Table 2. Performance evaluation of the CK+ dataset.

No.	Expression	Precision %	Sensitivity %	Specificity %	Accuracy %
1	Anger	100.00	82.00	100.00	99.00
2	Contempt	100.00	100.00	100.00	100.00
3	Disgust	100.00	100.00	100.00	100.00
4	Fear	100.00	100.00	100.00	100.00
5	Happy	100.00	100.00	100.00	100.00
6	Sadness	95.00	95.10	100.00	96.00
7	Surprise	100.00	100.00	100.00	100.00
	Mean	99.28	96.72	100.00	99.30

Table 3. Performance evaluation of the JAFFE dataset.

No.	Expression	Precision %	Sensitivity %	Specificity %	Accuracy %
1	Anger	96.30	96.30	99.30	98.00
2	Disgust	85.70	92.30	97.50	96.70
3	Fear	91.30	100.00	97.50	99.40
4	Happy	100.00	92.30	100.00	98.90
5	Neutral	86.50	83.90	97.50	95.20
6	Sadness	89.70	89.70	98.10	96.80
7	Surprise	100.00	96.20	100.00	99.40
	Mean	92.90	92.96	98.60	92.50

The CK+ dataset gives us much better results using SSAE-FER on 7 standard expressions. Precision, sensitivity, specificity and accuracy were noted at 99.28, 96.72, 100 and 99.30%, respectively. On the other hand, the JAFEE dataset gives us lower results due to the complexity of the dataset, where precision, sensitivity, specificity and accuracy were noted at 92.90, 92.96, 98.60 and 92.50%, respectively.

4.4. Comparison and discussion

In this section, we compare our results with other techniques which are given in Table 4.

Table 4. Comparison of the proposed model with other methods.

Methodology	Precision %	Sensitivity %	Specificity %	Error rate %	Accuracy %
[61]	-	-	-	10.55	89.45
[62]	-	-	-	17.90	82.10
[49]	88.00	86.00	-	13.64	86.36
[63]	-	-	-	26.20	73.80
Our model on the CK+ dataset	99.28	96.72	100.00	0.70	99.30
Our model on the JAFFE dataset	92.90	92.96	98.60	7.50	92.50

The error rate and accuracy presented in [61] were 10.55 and 89.45% with a novel technique of FER that has a modified classification and regression tree (M-CRT) to deal with the problem in the classification of expressions. The supervised descent and local binary method involve forgetting the global and local features. In [62] a projective complex matrix factorization (proCMF) is introduced, high-dimensional images are used for input, and these are converted into lower dimension subspace. It deals with the complex domain through the optimization problem where the error rate and accuracy were 17.9 and 82.10%. Another novel technique to recognize the expressions is through multimodal automatic emotion recognition (AER) network, which is highly capable in recognizing the expressions with reasonable accuracy. The model achieved 86.36% accuracy with 88% precision [49]. In [63], the author proposed a technique for the FER system to reduce the parameters. A deep learning neural network with a fully connected layer and the global average pooling (GAP) method is applied to achieve 73.80% accuracy. Our proposed method shows a comparatively high recognition rate of 99.30% accuracy on CK+ and 92.50% on JAFFE dataset. Our model on the JAFFE dataset did not perform well due to the complex nature of the dataset, but on CK+, it performed well.

5. Conclusions, limitations and future work

As is quite evident after plenty of research and deliberation, gaining insight into what a person may be feeling is very valuable for many reasons by identifying human feelings from their facial expressions. We have adopted a unique approach, the SSAE-FER model, for the classification of facial expressions. Our model learns the features automatically when input images are given to the model. The pre-training of datasets is achieved in an unsupervised manner and then fine-tuned in a supervised manner. After that, the probability estimation matrix showed the most effective results in the classification of seven basic facial expressions.

Our work was limited to training on CPU-based machines, which is why it took a longer time for training. In the future, we will use a framework that could support GPU, which will improve the training time. Several possible research directions of our proposed model can be utilized for the binary classification of images, such as tumor classification and segmentation. The performance of the proposed model could be enhanced by providing a larger dataset; moreover, it can be used for color-based datasets and real-time scenarios.

Acknowledgments

This research work was supported by Zayed University in Abu Dhabi with research fund #R20102.

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. A. T. Lopes, E. D. Aguiar, A. F. D. Souza, T. Oliveira-Santos, Facial expression recognition with convolutional neural networks: coping with few data and the training sample order, *Pattern Recognit.*, **61** (2017), 610–628. https://doi.org/10.1016/j.patcog.2016.07.026
2. S. S. Hammed, A. Sabanayagam, E. Ramakalaivani, A review on facial expression recognition systems, *J. Crit. Rev.*, **7** (2020), 903–905. Available from: https://www.jcreview.com/admin/Uploads/Files/61aa04ff88cda6.89247605.pdf.
3. S. Rajan, P. Chenniappan, S. Devaraj, N. Madian, Facial expression recognition techniques: a comprehensive survey, *IET Image Proc.*, **13** (2019), 1031–1040. https://doi.org/10.1049/iet-ipr.2018.6647
4. S. H. Ma, S. M. Lai, Y. Sun, Z. C. Pan, Research status and prospect of face expression recognition, in *2019 Chinese Control And Decision Conference (CCDC)*, (2019), 640–646. https://doi.org/10.1109/CCDC.2019.8833483
5. T. A. Rashid, Convolutional neural networks based method for improving facial expression recognition, in *Intelligent Systems Technologies and Applications 2016*, (2016), 73–84. https://doi.org/10.1007/978-3-319-47952-1_6
6. M. A. Jaffar, Facial expression recognition using hybrid texture features based ensemble classifier, *Int. J. Adv. Comput. Sci. Appl.*, **8** (2017), 449–453. https://doi.org/10.14569/IJACSA.2017.080660
7. R. Gupta, Positive emotions have a unique capacity to capture attention, *Prog. Brain Res.*, **247** (2019), 23–46. https://doi.org/10.1016/bs.pbr.2019.02.001
8. A. B. S. Salamh, H. I. Akyüz, A new deep learning model for face recognition and registration in distance learning, *Int. J. Emerging Technol. Learn.*, **17** (2022), 29. https://doi.org/10.3991/ijet.v17i12.30377
9. M. Ahmad, D. Ai, G. Xie, S. F. Qadri, H. Song, Y. Huang, et al., Deep belief network modeling for automatic liver segmentation, *IEEE Access*, **7** (2019), 20585–20595. https://doi.org/10.1109/ACCESS.2019.2896961

10. S. F. Qadri, D. Ai, G. Hu, M. Ahmad, Y. Huang, Y. Wang, et al., Automatic deep feature learning via patch-based deep belief network for vertebrae segmentation in CT images, *Appl. Sci.*, **9** (2018), 69. https://doi.org/10.3390/app9010069
11. I. Hirra, M. Ahmad, A. Hussain, M. U. Ashraf, I. A. Saeed, S. F. Qadri, et al., Breast cancer classification from histopathological images using patch-based deep learning modeling, *IEEE Access*, **9** (2021), 24273–24287. https://doi.org/10.1109/ACCESS.2021.3056516
12. M. Ahmad, J. Yang, D. Ai, S. F. Qadri, Y. Wang, Deep-stacked auto encoder for liver segmentation, in *Advances in Image and Graphics Technologies*, (2017), 243–251. https://doi.org/10.1007/978-981-10-7389-2_24
13. S. F. Qadri, L. Shen, M. Ahmad, S. Qadri, S. S. Zareen, M. A. Akbar, SVseg: stacked sparse autoencoder-based patch classification modeling for vertebrae segmentation, *Mathematics*, **10** (2022), 796. https://doi.org/10.3390/math10050796
14. M. Ahmad, S. F. Qadri, S. Qadri, I. A. Saeed, S. S. Zareen, Z. Iqbal, et al., A lightweight convolutional neural network model for liver segmentation in medical diagnosis, *Comput. Intell. Neurosci.*, **2022** (2022), 7954333. https://doi.org/10.1155/2022/7954333
15. M. Ahmad, S. F. Qadri, M. U. Ashraf, K. Subhi, S. Khan, S. S. Zareen, et al., Efficient liver segmentation from computed tomography images using deep learning, *Comput. Intell. Neurosci.*, **2022** (2022), 2665283. https://doi.org/10.1155/2022/2665283
16. S. F. Qadri, M. Ahmad, D. Ai, J. Yang, Y. Wang, Deep belief network based vertebra segmentation for CT images, in *Image and Graphics Technologies and Applications*, (2018), 536–545. https://doi.org/10.1007/978-981-13-1702-6_53
17. M. Ahmad, Y. Ding, S. F. Qadri, J. Yang, Convolutional-neural-network-based feature extraction for liver segmentation from CT images, in *Eleventh International Conference on Digital Image Processing (ICDIP 2019)*, **11179** (2019), 829–835. https://doi.org/10.1117/12.2540175
18. I. Banerjee, Y. Ling, M. C. Chen, S. A. Hasan, C. P. Langlotz, M. Moradzadeh, et al., Comparative effectiveness of convolutional neural network (CNN) and recurrent neural network (RNN) architectures for radiology text report classification, *Artif. Intell. Med.*, **97** (2019), 79–88. https://doi.org/10.1016/j.artmed.2018.11.004
19. M. Murugappan, A. M. Mutawa, S. Sruthi, A. Hassouneh, A. Abdulsalam, S. Jerritta, et al., Facial expression classification using KNN and decision tree classifiers, in *2020 4th International Conference on Computer, Communication and Signal Processing (ICCCSP)*, (2020), 1–6. https://doi.org/10.1109/ICCCSP49186.2020.9315234
20. M. Qasim, M. Khan, W. Mehmood, F. Sobieczky, M. Pichler, B. Moser, A comparative analysis of anomaly detection methods for predictive maintenance in SME, in *Database and Expert Systems Applications - DEXA 2022 Workshops*, (2022), 22–31. https://doi.org/10.1007/978-3-031-14343-4_3
21. M. Khan, A. Ahmad, F. Sobieczky, M. Pichler, B. A. Moser, I. Bukovský, A systematic mapping study of predictive maintenance in SMEs, *IEEE Access*, **10** (2022), 88738–88749. https://doi.org/10.1109/ACCESS.2022.3200694
22. W. Rafique, M. Khan, N. Sarwar, M. Sohail, A. Irshad, A graph theory based method to extract social structure in the society, in *Intelligent Technologies and Applications*, (2018), 437–448. https://doi.org/10.1007/978-981-13-6052-7_38

23. M. Khan, M. Liu, W. Dou, S. Yu, vGraph: graph virtualization towards big data, in *2015 Third International Conference on Advanced Cloud and Big Data*, (2015) 153–158. https://doi.org/10.1109/CBD.2015.33
24. W. Rafique, M. Khan, X. Zhao, N. Sarwar, W. Dou, A blockchain-based framework for information security in intelligent transportation systems, in *Intelligent Technologies and Applications*, (2019), 53–66. https://doi.org/10.1007/978-981-15-5232-8_6
25. P. Haindl, G. Buchgeher, M. Khan, B. Moser, Towards a reference software architecture for human-AI teaming in smart manufacturing, in *Proceedings of the ACM/IEEE 44th International Conference on Software Engineering: New Ideas and Emerging Results*, (2022), 96–100. https://doi.org/10.1145/3510455.3512788
26. W. Rafique, M. Khan, W. Dou, Maintainable software solution development using collaboration between architecture and requirements in heterogeneous IoT paradigm (Short Paper), in *Collaborative Computing: Networking, Applications and Worksharing*, (2019), 489–508. https://doi.org/10.1007/978-3-030-30146-0_34
27. W. Rafique, M. Khan, N. Sarwar, W. Dou, SocioRank*: A community and role detection method in social networks, *Comput. Electr. Eng.*, **76** (2019), 122–132. https://doi.org/10.1016/j.compeleceng.2019.03.010
28. Z. Hu, J. Tang, P. Zhang, J. Jiang, Deep learning for the identification of bruised apples by fusing 3D deep features for apple grading systems, *Mech. Syst. Signal Process.*, **145** (2020), 106922. https://doi.org/10.1016/j.ymssp.2020.106922
29. M. Iqtait, F. Mohamad, M. Mamat, Feature extraction for face recognition via active shape model (ASM) and active appearance model (AAM), *IOP Conf. Ser.: Mater. Sci. Eng.*, **332** (2018), 012032. https://doi.org/10.1088/1757-899X/332/1/012032
30. H. Jung, S. Lee, J. Yim, S. Park, J. Kim, Joint fine-tuning in deep neural networks for facial expression recognition, in *2015 IEEE International Conference on Computer Vision (ICCV)*, (2015), 2983–2991. https://doi.org/10.1109/ICCV.2015.341
31. M. J. Cossetin, J. C. Nievola, A. L. Koerich, Facial expression recognition using a pairwise feature selection and classification approach, in *2016 International Joint Conference on Neural Networks (IJCNN)*, (2016), 5149–5155. https://doi.org/10.1109/IJCNN.2016.7727879
32. X. Zhao, X. Liang, L. Liu, T. Li, Y. Han, N. Vasconcelos, et al., Peak-piloted deep network for facial expression recognition, in *Computer Vision – ECCV 2016*, (2016), 425–442. https://doi.org/10.1007/978-3-319-46475-6_27
33. R. N. Abiram, P. Vincent, Identity preserving multi-pose facial expression recognition using fine tuned VGG on the latent space vector of generative adversarial network, *Math. Biosci. Eng.*, **18** (2021), 3699–3717. https://doi.org/10.3934/mbe.2021186
34. H. Yang, L. Yin, CNN based 3D facial expression recognition using masking and landmark features, in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, (2017), 556–560. https://doi.org/10.1109/ACII.2017.8273654
35. W. Wei, Q. Jia, G. Chen, Real-time facial expression recognition for affective computing based on Kinect, in *2016 IEEE 11th Conference on Industrial Electronics and Applications (ICIEA)*, (2016), 161–165. https://doi.org/10.1109/ICIEA.2016.7603570

36. B. Huang, Z. Ying, Sparse autoencoder for facial expression recognition, in *2015 IEEE 12th Intl Conf on Ubiquitous Intelligence and Computing and 2015 IEEE 12th Intl Conf on Autonomic and Trusted Computing and 2015 IEEE 15th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom)*, (2015), 1529–1532. https://doi.org/10.1109/UIC-ATC-ScalCom-CBDCom-IoP.2015.274
37. T. Ahmad, H. Mao, L. Lin, G. Tang, Action recognition using attention-joints graph convolutional neural networks, *IEEE Access*, **8** (2019), 305–313. https://doi.org/10.1109/ACCESS.2019.2961770
38. M. Wang, T. C. Yeh, Human action recognition using CNN and BoW methods, 2016. Available from: http://cs229.stanford.edu/proj2016spr/report/053.pdf.
39. C. Shen, K. Zhang, J. Tang, A covid-19 detection algorithm using deep features and discrete social learning particle swarm optimization for edge computing devices, *ACM Trans. Internet Technol.*, **22** (2021), 1–17. https://doi.org/10.1145/3453170
40. R. K. Meleppat, C. R. Fortenbach, Y. Jian, E. S. Martinez, K. Wagner, B. S. Modjtahedi, et al., In Vivo imaging of retinal and choroidal morphology and vascular plexuses of vertebrates using swept-source optical coherence tomography, *Transl. Vision Sci. Technol.*, **11** (2022), 11. https://doi.org/10.1167/tvst.11.8.11
41. K. Ratheesh, L. Seah, V. Murukeshan, Spectral phase-based automatic calibration scheme for swept source-based optical coherence tomography systems, *Phys. Med. Biol.*, **61** (2016), 7652. https://doi.org/10.1088/0031-9155/61/21/7652
42. R. Meleppat, M. Matham, L. Seah, An efficient phase analysis-based wavenumber linearization scheme for swept source optical coherence tomography systems, *Laser Phys. Lett.*, **12** (2015), 055601. https://doi.org/10.1088/1612-2011/12/5/055601
43. R. K. Meleppat, P. Prabhathan, S. L. Keey, M. V. Matham, Plasmon resonant silica-coated silver nanoplates as contrast agents for optical coherence tomography, *J. Biomed. Nanotechnol.*, **12** (2016), 1929–1937. https://doi.org/10.1166/jbn.2016.2297
44. D. Girish, V. Singh, A. Ralescu, Understanding action recognition in still images, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, (2020), 1523–1529. https://doi.org/10.1109/CVPRW50498.2020.00193
45. H. Yang, U. Ciftci, L. Yin, Facial expression recognition by de-expression residue learning, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2018), 2168–2177. https://doi.org/10.1109/CVPR.2018.00231
46. Y. Zhou, B. E. Shi, Photorealistic facial expression synthesis by the conditional difference adversarial autoencoder, in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, (2017), 370–376. https://doi.org/10.1109/ACII.2017.8273626
47. B. Yan, G. Han, Effective feature extraction via stacked sparse autoencoder to improve intrusion detection system, *IEEE Access*, **6** (2018), 41238–41248. https://doi.org/10.1109/ACCESS.2018.2858277
48. S. R. Livingstone, F. A. Russo, The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): a dynamic, multimodal set of facial and vocal expressions in North American English, *PloS One*, 13 (2018), e0196391. https://doi.org/10.1371/journal.pone.0196391
49. M. F. H. Siddiqui, A. Y. Javaid, A multimodal facial emotion recognition framework through the fusion of speech with visible and infrared images, *Multimodal Technol. Interact.*, **4** (2020), 46. https://doi.org/10.3390/mti4030046

50. J. Jang, D. H. Kim, H. I. Kim, Y. M. Ro, Color channel-wise recurrent learning for facial expression recognition, in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (2017), 1233–1237. https://doi.org/10.1109/ICASSP.2017.7952353
51. S. Happy, A. Routray, Robust facial expression classification using shape and appearance features, in *2015 Eighth International Conference on Advances in Pattern Recognition (ICAPR)*, (2015), 1–5. https://doi.org/10.1109/ICAPR.2015.7050661
52. K. M. Koo, E. Y. Cha, Image recognition performance enhancements using image normalization, *Hum.-centric Comput. Inf. Sci.*, **7** (2017), 33. https://doi.org/10.1186/s13673-017-0114-5
53. Y. Liu, Y. Li, X. Ma, R. Song, Facial expression recognition with fusion features extracted from salient facial areas, *Sensors*, **17** (2017), 712. https://doi.org/10.3390/s17040712
54. A. Ng, J. Ngiam, C. Y. Foo, Y. Mai, C. Suen, A. Coates, et al., Unsupervised feature learning and deep learning, 2013. Available from: https://redirect.cs.umbc.edu/courses/pub/www/courses/graduate/678/spring15/visionaudio.pdf.
55. L. Chen, M. Zhou, W. Su, M. Wu, J. She, K. Hirota, Softmax regression based deep sparse autoencoder network for facial emotion recognition in human-robot interaction, *Inf. Sci.*, **428** (2018), 49–61. https://doi.org/10.1016/j.ins.2017.10.044
56. M. J. Lyons, "Excavating AI" Re-excavated: Debunking a fallacious account of the JAFFE dataset, preprint, arXiv:2107.13998.
57. M. J. Lyons, M. Kamachi, J. Gyoba, Coding facial expressions with Gabor wavelets (IVC special issue), preprint, arXiv:2009.05938.
58. T. Kanade, J. F. Cohn, Y. Tian, Comprehensive database for facial expression analysis, in Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580), (2000), 46–53. https://doi.org/10.1109/AFGR.2000.840611
59. S. Eng, H. Ali, A. Cheah, Y. Chong, Facial expression recognition in JAFFE and KDEF datasets using histogram of oriented gradients and support vector machine, in *IOP Conf. Ser.: Mater. Sci. Eng.*, **705** (2019), 012031. https://doi.org/10.1088/1757-899X/705/1/012031
60. R. B. Palm, Prediction as a candidate for learning deep hierarchical models of data, 2012. Available from: https://www2.imm.dtu.dk/pubdb/edoc/imm6284.pdf.
61. L. Du, H. Hu, Modified classification and regression tree for facial expression recognition with using difference expression images, *Electron. Lett.*, **53** (2017), 590–592. https://doi.org/10.1049/el.2017.0731
62. V. H. Duong, Y. S. Lee, J. J. Ding, B. T. Pham, M. Q. Bui, J. C. Wang, Projective complex matrix factorization for facial expression recognition, *EURASIP J. Adv. Signal Process.*, **2018** (2018), 10. https://doi.org/10.1186/s13634-017-0521-9
63. T. Zhang, Face expression recognition based on deep learning, *J. Phys.: Conf. Ser.*, **1486** (2020), 042048. https://doi.org/10.1088/1742-6596/1486/4/042048