Convolutional Sparse Autoencoder for Emotion Recognition

M. Mohana(✉) and P. Subashini

Department of Computer Science, Centre for Machine Learning and Intelligence (CMLI),
Avinashilingam Institute, Coimbatore, India
{mohana_cs,subashini_cs}@avinuty.ac.in

Abstract. Emotion recognition is a hot research area in deep learning and computer vision that analyses expressions from both static and dynamic sequences of facial expressions to reveal human emotional states. In recent decades, deep learning approaches have been exhibiting a superior performance on image representation datasets. However, the convolutional neural network (CNN) requires a larger number of labeled datasets for training and accurate classification results. It is always inevitable, whereas unsupervised representation learning models like autoencoder do not require labeled information for training. Meanwhile, it is difficult to infer the feature map when the size of the CNN layer is increased. To address these challenges, this paper introduced a self-supervised deep learning technique called convolutional sparse autoencoder (CSA) which can learn robust features from small data with unlabeled facial expression datasets. Moreover, sparsity is added in the max pooling layer for the feature map which makes the backpropagation optimizer Adam work efficiently for the CSA training; thus, no complicated optimizer is not involved. Finally, the trained convolutional sparse encoder part is combined with the softmax layer for emotion classification. The performance results demonstrate that the proposed approach achieved 98% of accuracy on the CK+ dataset and outperforms various state-of-the-art methods.

Keywords: Convolutional sparse autoencoder (CSA) · Convolutional neural network (CNN) · Deep learning · Emotion recognition (ER) · Representation learning (RL)

1 Introduction

Emotion recognition has gained popularity over the past decades of its applications [1] in a variety of fields, including the development of intelligent robots, improved emotional understanding, monitoring candidates' expressions during interviews, and so on. Furthermore, it is an essential component of human-robot interaction technology. Humans find it easy to recognize human emotions, but it is difficult for machines. Numerous research has been conducted in the literature, and a few emotions are mentioned as universal expressions that are happy, sad, fear, angry, surprise, and disgust while some studies added neutral and disgust. Furthermore, many challenges such as illumination,

A. E. Hassanien et al. (Eds.): AICV 2023, LNDECT 164, pp. 3–15, 2023.
https://doi.org/10.1007/978-3-031-27762-7_1

pose variation, low resolution, and complex background always have issues with facial expression recognition performance.

There are different styles of approaches and techniques applied to facial emotion recognition like other recognition research. The primary function of FER is to map an expression to the appropriate emotional state. Face detection, pre-processing, feature extraction, and emotion recognition are the four steps in the conventional FER. Except for facial images, the face detection method is used to identify the face in images and remove the background parts. The next pre-processing steps include resizing, cropping, and normalization. Existing works use notable approaches such as histogram-oriented gradients (HOG), scale-invariant feature transform (SIFT), facial action units, and Gabor wavelet transform to extract features for the conventional FER system. Moreover, dimension reduction techniques, such as local binary pattern (LBP), principal component analysis (PCA), and optical flow techniques, also help to extract relevant descriptors from facial images. Prior to the introduction of deep learning techniques in the FER system, those approaches aided in removing irrelevant information from facial images and improving recognizing ability. Conventional feature extraction approaches, however, are insufficient for extracting micro-expressions from facial images.

Representation learning is an effective technique for learning high-level feature representations from images [4, 6]. Convolutional neural networks (CNN) in deep learning are typically applied in numerous image recognition and detection processes [7]. The CNN structure and pooling are important factors in representation learning, and the convolutional layer is used to learn spatial information from static images. The pooling layer is used to decrease the dimension of the features and overfitting issues [13]. However, CNN requires a large number of labeled data when training the model, resulting in high computational costs. Furthermore, the layer size is increasing while the network is becoming deeper for extracting high-level features, which frequently results in a vanishing gradient problem.

To address these challenges, this paper introduced the convolutional sparse autoencoder (CSA) for emotion classification. The presented method is first trained with an unsupervised small amount of facial expression dataset, as illustrated in Fig. 3. Following that, the trained encoder has been combined with the softmax layer and fine-tuned for emotion classification. Here, sparsity [11] has been added to each max pooling layer to enhance and speed up the feature map for efficient feature learning [12]. The outcome of the experiments demonstrates that the presented method is effective for image feature extraction using a data-driven approach and achieves robust labeling for facial expression recognition. This paper's main contribution is as follows:

1. CSA is proposed for the automatic extraction of features from small amounts of data while avoiding the uncertainty of the conventional feature selection process. This technique is effective for dealing with the problem of insufficient data training caused by labeled facial expression images,
2. With the help of the robust power of convolutional kernels, facial traits are directly extracted from raw images. Moreover, this method reduced the size of parameters and overfitting issues compared to the conventional approach, and
3. Any type of image dataset and associated application domains can be extracted using the unsupervised data-driven approach.

The structure of the paper is as follows: Sect. 2 describes the already existing works based on convolutional neural networks and autoencoder on the FER system. Section 3 explains the proposed approach step by step. Section 4 discusses the results and analyses of the proposed network on CK+ datasets. Finally, Sect. 5 summarizes the overall CSA work.

2 Background and Related Works

During recent decades, deep learning techniques, particularly in FER, have excelled at automatic image recognition and detection tasks [4, 17]. The conventional method extracts feature from facial images before classifying emotions based on feature values. Through the evaluation of the FER system, various machine learning methods (K-nearest neighbor, Support vector machine, neural network, principal component analysis, local binary pattern, and linear discriminant analysis) are initially employed. Such techniques have the limitation of extracting features from front views of facial images for FER evaluation. Deep learning techniques based on FER tasks, on the other hand, combine feature selection and emotion classification processes. Numerous studies have reviewed and compared [4–7], and the recent unsupervised learning-based FER is also included [14, 16]. The techniques used in eminent FER methods are briefly described in the studies that follow.

Zhao et al. [4] have proposed the convolutional neural network-based FER system and fine-tuned it with the VGG network. This network has been trained with different sizes of convolutional filters and dropout values to generalize the network. The authors have achieved the performance of this network at 99.33% on CK+, 87.65% on MUG, and 93.33% on RaFD datasets. Mollahosseini et al. [5] introduced deep learning-based architecture to address the issues in FER on different datasets. Two convolutional networks and a pooling layer were used to create this network, which was then followed by inception layers that increased the depth and width of the network during keeping the computing cost stable. It is a single-network architecture that accepts facial images as input and categorizes them based on one of six primary emotions. These experiments were conducted on CK+, DISFA, FERA, SFEW, MultiPIE, MMI, and FER 2013. Minaee et al. [7] proposed a deep network for improving the accuracy of multiple datasets, including FERG, JAFFE, FER-2013, and CK+, based on an attentional convolutional neural network that focused on important features of facial images. This network has achieved some extent performance compared to conventional FER techniques. Jaiswal et al. [6] demonstrated emotion recognition using a convolutional neural network that combines two network features to classify emotions. It has been tested on two different datasets with 70.14% accuracy on FER-2013, and 98.65% on JAFFE. Akhand et al. [8] have introduced a deep convolutional network (DCN) through transfer learning (TL) techniques and fine-tuned it with facial expression data. The limitation of the conventional CNN network on the facial image is features extracted from the frontal view of high-resolution images. For this experiment, the authors include eight different pre-trained networks (ResNet-18, ResNet-34, ResNet-50, ResNet-152, Inception-v3, VGG-16, VGG-19, and DenseNet-161) for improving the accuracy of the FER system. This method experimented on JAFFE and KDFE datasets and achieved 96.51%, and 99.52% respectively.

Unsupervised artificial neural network-based autoencoders is widely used in several real-world applications like anomaly detection, data compression, image denoising, segmentation, and classification in recent decades [16]. Zeng et al. [14] used a deep sparse autoencoder to learn robust features from unlabeled facial image data to recognize facial expressions with high accuracy. Facial features were extracted using both geometric and appearance features instead of extracting features from raw facial images. This technique helped to identify the relevant features. These experimental results show 95.79% accuracy was achieved on CK+ datasets. Liu et al. [15] proposed a stacked sparse autoencoder that uses an optical flow method combined with a deep neural network to reduce the influence of the same expression on different facial expressions. Finally, a softmax layer had been added on top of the layer for emotion classification. This work has experimented on CK+ datasets and achieved 92.3% accuracy. Usman et al.[16] have introduced emotion recognition using deep sparse autoencoder for feature selection and dimensional reduction for facial expression recognition on multiple hidden layers. The authors showed that the stacked autoencoder extracted more relevant features from facial images compared to the conventional approach and achieved 99.60% accuracy on CK+ datasets. The dense network is used in the works mentioned above to recognize facial emotions. Zhang [21] has compared the performance of simple autoencoder and convolutional autoencoder for image pre-processing. The results show that the convolutional autoencoder gives better performance on the image dataset.

3 Convolutional Sparse Autoencoder for Emotion Recognition

The proposed emotion recognition framework uses a convolutional sparse autoencoder (CSA) to extract the latent representation for each class from facial images and a softmax layer to perform classification. This methodology is divided into the following steps: pre-processing, convolutional sparse autoencoder, and emotions classification. Figure 1 shows the proposed network architecture.

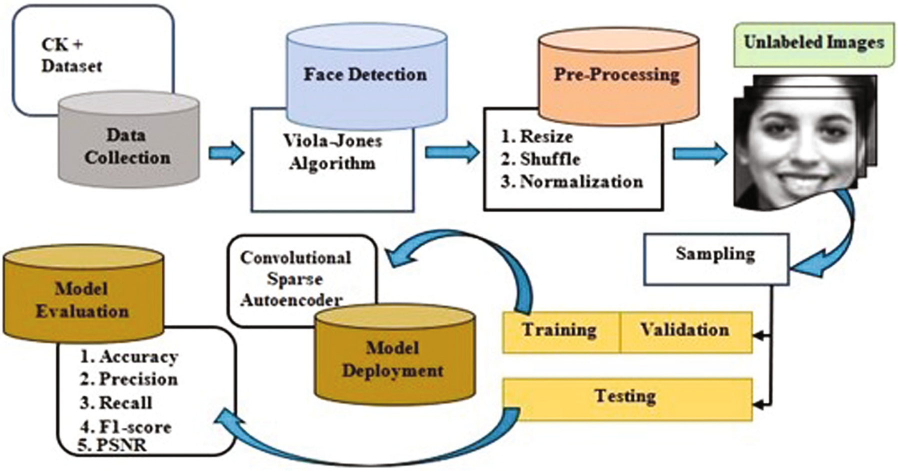

Fig. 1. Overview of the proposed architecture

3.1 Pre-processing

Pre-processing is a critical stage in image processing that is used to improve the relevant features for subsequent steps. The first process of a conventional FER system is face detection which is used to identify and locate the face in images. For facial emotion recognition, the facial part is more sufficient than other parts of features in facial images. For this purpose, a real-time viola-jones face [9] detection algorithm has been employed in this research study to detect the face from raw images (See Fig. 2). This algorithm consists of four parts, namely, haar features, integral images, Adaboost algorithm, and cascade classifiers. Firstly, Haar features consist of black and white regions which produce a single value by the sum of the light regions subtracted by the sum of the black regions. It is used to extract useful information such as edges, and diagonal and straight lines for identifying the human face. Next integral images are used to seed up the haar rectangle feature calculation process. Third, the AdaBoost algorithm is used to build a strong classifier among all available features. Finally, the cascade classifier removed the unnecessary part from facial images except for the facial region. In addition, the detected face has cropped into 48 x 48 pixels whose intensity values have been normalized between 0–1 for reducing the computation time complexity in a neural network.

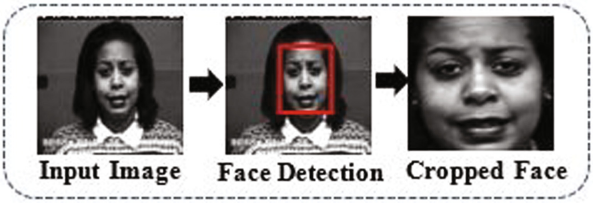

Fig. 2. Face detection process

3.2 Convolutional Sparse Autoencoder

A supervised method is a data-driven feature learning method that updates connection weights through forward and backward training processes. Unsupervised learning directly receives unlabeled input data and learns more relevant features compared to the supervised method. This method significantly reduced the workload for labeling data. Figure 3 shows the overall structure of the proposed method.

In this paper, the unsupervised autoencoder [3], which is made up of three parts: encoder, latent space (code), and decoder, effectively recognizes facial expressions. The autoencoder is primarily used for dimensionality reduction, image denoising, and feature extraction. It gives an output that is the same as the input data and compares its original data. After many iterations, the cost function reaches optimality, which means that the reconstructed output is as close to the input data as possible. The encoder converts the input data into code of the hidden layer by $code(c) = f(w.x + b)$, where f is an activation function, w is a weight x is an input value and b is a bias. The decoder reconstructs the output from code of the hidden layer by $x' = f'(w'.c + b')$, and calculate the mean squared values between input and reconstructed output using the cost function $cost = min \sum_{i=1}^{n} \left|x' - x\right|^2$.

A convolutional autoencoder [2, 18] is a variational of a convolutional neural network that is used for retaining the connected information between pixels of images. The layers in the CNN network help to extract the relevant features and find patterns without human intervention. The process of converting the feature maps input to output is known as a convolutional encoder, and the output is reconstructed using the inverse convolutional operation, which is known as a convolutional decoder. Moreover, the reconstruction error of the convolutional encoder and decoder can be calculated in the same way as the standard autoencoder.

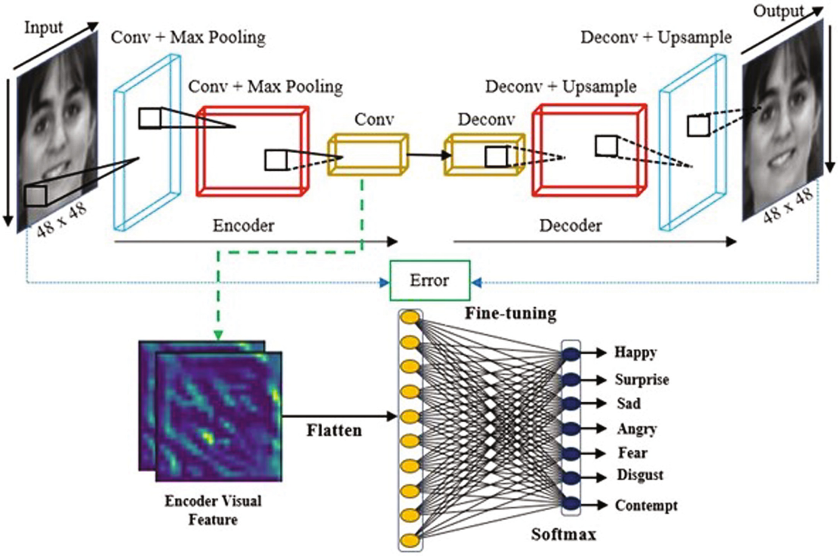

Fig. 3. Proposed convolutional sparse autoencoder (CSA)

The number of convolutional layers, pooling layer, ReLU activation function, neuron size, sparsity, and filter size are the main components of a convolutional sparse autoencoder. The proposed model consists of four convolution blocks of the encoder and three deconvolution blocks of the decoder, each with a convolutional layer and filter size (3 × 3), followed by batch normalization, which is used to generalize the network learning process. After each block, the max pooling (a kernel size 2 × 2) layer with sparsity [12] is used to downsample the feature map of the convolutional encoder. Here, the sparsity helps to highlight the more relevant features for learning. In the convolutional decoder, the convolutional layer, activation function, and batch normalization are also included. After the two convolutional blocks, upsampling layers (kernel size 2 × 2) are used.

First, the input facial image is encoded each time with pixels patch x_i, $i = 1, 2 \ldots x_n$ and multiplied with neuron weightsw_j, where j is used for convolutional calculation. Finally, the output layer o_{ij} is calculated as$o_{ij} = f(w_j.x_i + b)$. Then, output from the convolutional decoder is defined as$x'_i = f'(w'_j.o_{ij} + b')$. Finally, reconstructed error is calculated as$CSA = \frac{1}{P}\sum_{i=1}^{P}\|x_i - x'_i\|$, where p is reconstruction operation of convolutional kernel size with d x d, where $d \leq pixels$.

3.3 Emotion Classification

In the classification part, the last convolutional layer in the encoder part has flattened into a feature vector and fed into a dense layer. A flattened layer is followed by a dense layer containing 128 neurons and a ReLU activation function. For more than one emotion classification, the softmax layer has added the top of the convolutional encoder. In the training process, the encoder has been fine-tuned by a softmax layer with a labeled facial expression dataset after the training of convolution sparse autoencoder. It has been trained with 100 epoch 32 batch size and Adam optimizer with 0.001 learning rate. In addition, categorical cross entropy is used as a loss function for calculating the training and validation accuracy of the presented model performance.

4 Experimental Results and Discussion

4.1 Datasets

For this research study, the CK+ [25] database is used for CSA performance evaluation. This dataset, which was released in 2010, is an expanded version of the Cohn-Kanade (CK), one of the most widely used benchmark datasets for evaluating the FER algorithm. The CK+ dataset contains 593 video sequences from 123 subjects at 30 frames per second and 640 x 490 pixels in resolution, which includes eight basic facial expressions such as 276 happy, 180 anger, 112 sad, 332 surprises, 72 contempt, 236 disgust, 100 fear and 327 neutral used for this evaluation. The length of each video sequence ranges from 10 to 60 frames.

In addition, the proposed approach has been designed using Keras and TensorFlow backend. The hold-out cross-validation method is used in this study, with 80% of the facial images used for training and the remaining 20% used for testing.

4.2 Evaluation Metrics

The following metrics are used in this experiment to assess the performance of the presented network emotion classification. FN stands for False Negative, TP stands for True Positive, FP stands for False Positive, and TN stands for True. To summarize the performance of the classification algorithm is called a confusion matrix from which values of performance are calculated individually.

$$Accuracy = \frac{TP + FN}{TP + TN + FP + FN} \tag{1}$$

Accuracy is the proportion of correctly classified instances to the total number of instances. It can judge positive as positive and negative as negative.

$$Recall/Sensitivity = \frac{TP}{TP + FN} \tag{2}$$

Recall refers to the predicted positive cases from the total positive cases.

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

Precision refers to the number of positive predictions in the positive example classified by the detection system.

$$F1 - Score = \frac{2TP}{2TP + FP + FN} \tag{4}$$

F1-Score is a harmonic mean of combining the values of precision and recall.

$$PSNR = 10\log_{10}\frac{{Max_I}^2}{MSE} \tag{5}$$

The MSE is used to calculate the autoencoder's reconstruction loss, whereas the peak signal-to-noise ratio (PSNR) is used to compare the quality of the original and compressed reconstructed images.

4.3 Experimental Results

This section presents the proposed network performance on test data and compares it with existing works. This network consists of four convolutional blocks of the encoder and three deconvolutional blocks of the decoder. The sparsity parameter is set to 1e−5. Initially, convolutional sparse autoencoder trained with 100 epochs, 32 batch size and Adam optimizer with learning rate 0.001. Figure 4. Shows the loss and reconstruction performances of CSA. The training loss for the proposed network is 0.0228, and the validation loss is 0.0428. Furthermore, MSE and PSNR are two more commonly used metrics. MSE's limitation is that it is highly dependent on image intensity scaling. PSNR prevents these issues by scaling the MSE according to the image range, and the reconstructed facial expression images achieved 70.06 dB PSNR in the presented network.

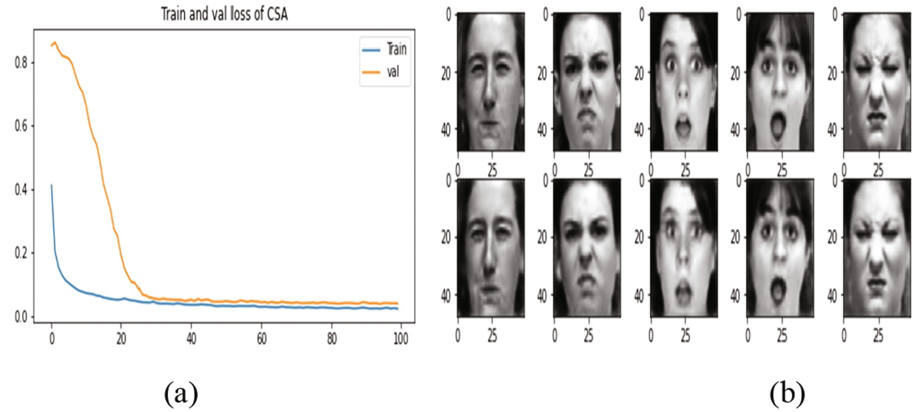

Fig. 4. (a) Loss evaluation of CSA (b) Sample test image, the first row shows the original images, and the second row shows the reconstructed images

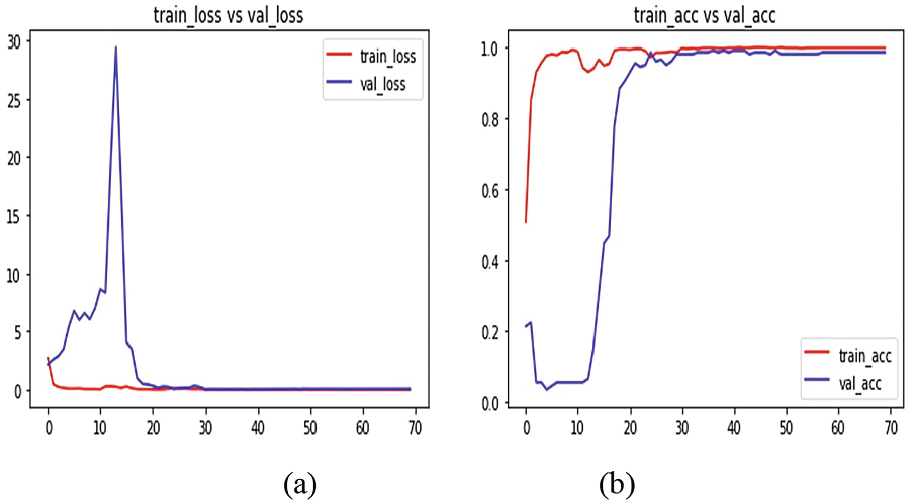

Fig. 5. (a) Training and validation loss of CSA (b) Training validation accuracy of CSA

After training the entire network, the final convolutional encoder layer is flattened into a feature vector and fed into the dense layer. Then, on top of the convolution encoder, a softmax layer with four convolution blocks is added. This network is trained using 70 epochs, 32 batch sizes, and the Adam optimizer with a learning rate of 0.001. Figure 5. Depicts CSA's training and testing performance. The training loss and accuracy, and validation loss and accuracy have been recorded in every epoch. Initially, the loss value is gradually increased due to the random initialization of weights. After the 30th epoch, validation loss decreased step by step and reached a minimum loss of about 0.0726. In the process of backpropagation, the encoding weights of each layer can be improved further for the ability of feature extraction. In Fig. 5. (b), initially, the network starts with higher fluctuations due to imbalanced facial expression images, after the optimizer Adam helped to generalize the network training. Finally, on the CK+ dataset, the presented model achieved 0.98% accuracy, 0.96% precision, 0.98% recall, and 0.97% f1-score. Furthermore, the accuracy of the seven classes is depicted in Fig. 6. The distinct features surrounding the eyes, lips, and nose of the region helped contempt, disgust, happy, and sad reach a greater performance accuracy of 100% among the seven expressions from this depiction. Here anger, fear, and surprise are slightly misclassified as contempt with 0.04%, 0.1%, 0.02% respectively. In general, facial expressions are readily mistaken due to their similar shape and look as well as their distinct variations within the same expression. In addition, table demonstrates how the suggested approach outperforms the other five FER algorithms, as can be observed. The reason for choosing these methods tested on the same CK+ datasets with outstating different approach and performances. From this comparison, convolutional sparse autoencoder discriminate different features and reduce the high dimension features, which also acts as robust facial feature classifier.

Visualizing the information captured behind the CNN network is highly important for evaluating the performance of the model. For this Grad-CAM technique is used

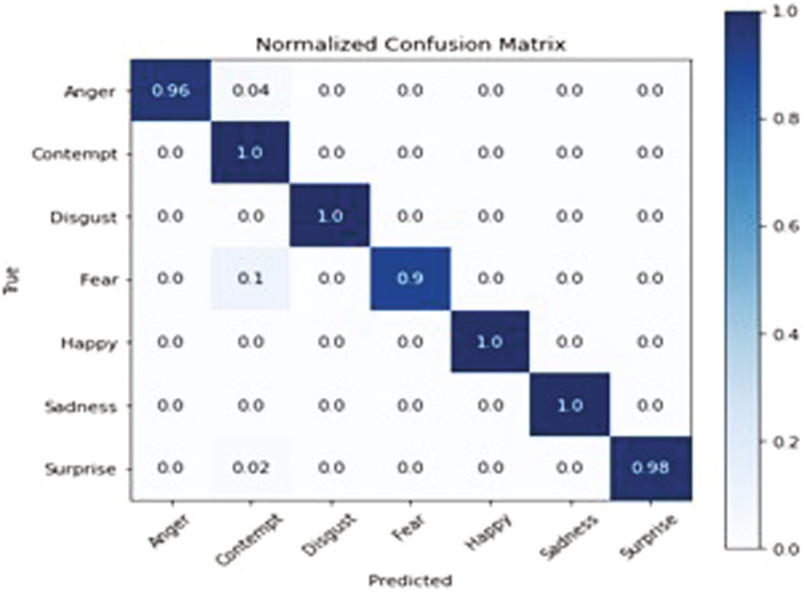

Fig. 6. Confusion matrix of 7 class facial expression recognition on CK+ dataset

here to differentiate between and capture facial emotions. Figure 7 shows the Grad-CAM visualization. The presented model focuses on important aspects of the image, i.e., lips, eyes, and eyebrows which help to distinguish the different emotions. Furthermore, Table 1 shows the proposed method recognition accuracy compared with state-art-of-the techniques results on each expression. The proposed method results are relatively high and reasonable on each facial expression.

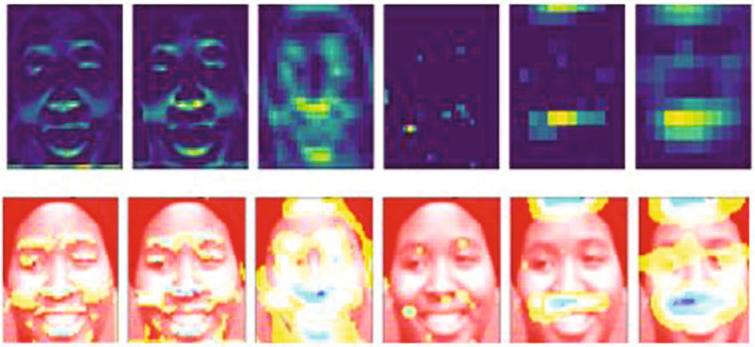

Fig. 7. Grad-CAM visualization [24] of sample emotion

Table 1. Performance evaluation of CSA versus other FER techniques

%	[19]	[20]	[21]	[22]	[23]	Proposed (CSA)
Anger	100	82.50	87.1	87.8	87.00	**96.00**
Contempt	83.34	–	–	–	–	**100**
Disgust	91.52	97.50	90.20	93.33	83.00	**100**
Fear	88.00	95.00	92.00	94.33	89.00	**90.00**
Happy	100	100	98.07	94.20	90.00	**100**
Sadness	85.51	92.50	91.47	96.42	84.00	**100**
Surprise	95.18	92.50	100	98.46	90.00	**98.00**
Average	91.94	93.33	93.14	94.09	87.16	**98.00**

5 Conclusion

A convolutional sparse autoencoder has been proposed in this paper to recognize emotions from facial expressions. This model is fully automated without the need for manual feature extraction. This autoencoder's primary goal is to reduce the issues of overfitting and the large amount of labeled data needed in the conventional FER method. Initially, it has been trained with unsupervised facial expressions. After that convolutional encoder part is trained and fine-tuned by the labeled facial expression dataset. The softmax layer is used to classify each emotion according to features learned by the encoder. Different learning rates, epochs, and batch sizes have been tried on this CSA for getting better configuration hyperparameters. Finally, the proposed model achieved 98% accuracy with a 0.0726 validation loss. Furthermore, this approach has been validated with precision, recall, F1-score, and PSNR metrics. Additionally, this experimental finding has been analyzed with existing state-of-the-art techniques. In the future, for emotion recognition, the physiological signal will be combined with facial expressions to handle the challenges of the real-world environment.

Acknowledgment. The authors wish to express their sincere thanks to the Centre for Machine Learning and Intelligence (CMLI) for providing resources to conduct this research study. This centre is sponsored and supported by the Department of Science and Technology (DST)-CURIE, India.

References

1. Kołakowska, A., Landowska, A., Szwoch, M., Szwoch, W., Wróbel, M.R.: Emotion recognition and its applications. In: Hippe, Z., Kulikowski, J., Mroczek, T., Wtorek, J. (eds.) Human-Computer Systems Interaction: Backgrounds and Applications 3. AISC, vol. 300, pp. 51–62. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-08491-6_5

2. Zhang, Y.: A better autoencoder for image: convolutional autoencoder. In: ICONIP17-DCEC (2018). http://users.cecs.anu.edu.au/Tom.Gedeon/conf/ABCs2018/paper/ABCs2018_paper_58.pdf
3. Bank, D., Koenigstein, N., Giryes, R.: Autoencoders. arXiv preprint arXiv:2003.05991 (2020)
4. Zhao, X., Shi, X., Zhang, S.: Facial expression recognition via deep learning. IETE Tech. Rev. **32**(5), 347–355 (2015)
5. Mollahosseini, A., Chan, D., Mahoor, M.H.: Going deeper in facial expression recognition using deep neural networks. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1–10. IEEE (2016)
6. Jaiswal, A., Raju, A.K., Deb, S.: Facial emotion detection using deep learning. In: 2020 International Conference for Emerging Technology (INCET), pp. 1–5. IEEE (2020)
7. Minaee, S., Minaei, M., Abdolrashidi, A.: Deep emotion: facial expression recognition using the attentional convolutional network. Sensors **21**(9), 3046 (2021)
8. Akhand, M.A.H., Roy, S., Siddique, N., Kamal, M.A.S., Shimamura, T.: Facial emotion recognition using transfer learning in the deep CNN. Electronics **10**(9), 1036(2021)
9. Viola, P., Jones, M.J.: Robust real-time face detection. Int. J. Comput. Vis. **57**(2), 137–154 (2004)
10. Kavukcuoglu, K., Sermanet, P., Boureau, Y.L., Gregor, K., Mathieu, M., Cun, Y.: Learning convolutional feature hierarchies for visual recognition. In: Advances in Neural Information Processing Systems, vol. 23 (2010)
11. Bristow, H., Eriksson, A., Lucey, S.: Fast convolutional sparse coding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 391–398 (2013)
12. Rigamonti, R., et al.: On the relevance of sparsity for image classification. Comput. Vis. Image Underst. **125**, 115–127 (2014)
13. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8689, pp. 818–833. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10590-1_53
14. Zeng, N., Zhang, H., Song, B., Liu, W., Li, Y., Dobaie, A.M.: Facial expression recognition via learning deep sparse autoencoders. Neurocomputing **273**, 643–649 (2018)
15. Liu, Y., Hou, X., Chen, J., Yang, C., Su, G., Dou, W.: Facial expression recognition and generation using sparse autoencoder. In: 2014 International Conference on Smart Computing, pp. 125–130 (2014). IEEE
16. Usman, M., Latif, S., Qadir, J.: Using deep autoencoders for facial expression recognition. In: 2017 13th International Conference on Emerging Technologies (ICET), pp. 1–6 (2017). IEEE
17. Lv, Y., Feng, Z., Xu, C.: Facial expression recognition via deep learning. In: 2014 International Conference on Smart Computing, pp. 303–308 (2014). IEEE
18. Masci, J., Meier, U., Cireşan, D., Schmidhuber, J.: Stacked convolutional auto-encoders for hierarchical feature extraction. In: Honkela, T., Duch, W., Girolami, M., Kaski, S. (eds.) ICANN 2011. ICANN 2011. LNCS, vol. 6791, pp. 52–59. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-21735-7_7
19. Boughida, A., Kouahla, M.N., Lafifi, Y.: A novel approach for facial expression recognition based on Gabor filters and genetic algorithm. Evol. Syst. **13**(2), 331–345 (2021)
20. Uddin, M.Z., Lee, J.J., Kim, T.S.: An enhanced independent component-based human facial expression recognition from video. IEEE Trans. Consum. Electron. **55**(4), 2216–2224 (2009)
21. Zhang, L., Tjondronegoro, D.: Facial expression recognition using facial movement features. IEEE Trans. Affect. Comput. **2**(4), 219–229 (2011)
22. Happy, S.L., Routray, A.: Automatic facial expression recognition using features of salient facial patches. IEEE Trans. Affect. Comput. **6**(1), 1–12 (2014)

23. Mishra, S., Joshi, B., Paudyal, R., Chaulagain, D., Shakya, S.: Deep residual learning for facial emotion recognition. In: Shakya, S., Bestak, R., Palanisamy, R., Kamel, K.A. (eds.) Mobile Computing and Sustainable Informatics. LNDECT, vol. 68, pp. 301–313. Springer, Singapore (2022). https://doi.org/10.1007/978-981-16-1866-6_22
24. Yang, S., Kim, Y., Kim, Y., Kim, C.: Combinational class activation maps for weakly supervised object localization. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2941–2949 (2020)
25. The Extended Cohn-Kanada Database. https://www.ri.cmu.edu/. Accessed 15 Nov 2022