Pattern Recognition Letters 115 (2018) 101–106

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Hybrid deep neural networks for face emotion recognition

Neha Jain[a,*], Shishir Kumar[a], Amit Kumar[a], Pourya Shamsolmoali[b,c], Masoumeh Zareapoor[b]

[a] Department of Computer Science & Engineering, Jaypee University of Engineering and Technology, Guna, India
[b] Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai 200240, China
[c] Advanced Scientific Computing Division, Euro-Mediterranean Centre on Climate Change (CMCC Foundation), Lecce, Italy

ARTICLE INFO

Article history:
Available online 9 April 2018

Keywords:
Emotion recognition
Deep learning
Recurrent neural networks
Convolutional Neural Networks
Hybrid CNN-RNN

ABSTRACT

Deep Neural Networks (DNNs) outperform traditional models in numerous optical recognition missions containing Facial Expression Recognition (FER) which is an imperative process in next-generation Human-Machine Interaction (HMI) for clinical practice and behavioral description. Existing FER methods do not have high accuracy and are not sufficient practical in real-time applications. This work proposes a Hybrid Convolution-Recurrent Neural Network method for FER in Images. The proposed network architecture consists of Convolution layers followed by Recurrent Neural Network (RNN) which the combined model extracts the relations within facial images and by using the recurrent network the temporal dependencies which exist in the images can be considered during the classification. The proposed hybrid model is evaluated based on two public datasets and Promising experimental results have been obtained as compared to the state-of-the-art methods.

1. Introduction

Facial and emotional expressions are the most significant non-verbal ways for expressing internal emotions and intentions.“Facial Action Coding system (FACS) is a useful structure that classifies the human facial actions by their advent on the face using Action Units (AU). An AU is one of 46 minor elements of visible facial motion or its relatedform changes.Facial expressions have worldwide meaning, and these emotions have been accepted for tens and even hundreds of years and it was the main reason for us to select facial expressions for the research”. These days, interest in emotion recognition (ER) has skyrocketed, while it stayed as single main difficulties in the area of human-computer interaction. The cornerstone of the most relevant research is to build a reliable conversation and communication among human and computer (machine). The importance of ER methods can be achieved by either “make humans to understand computer/machine accurately and conversely”. Facial Expression Recognition (FER) is a challenging task in machine learning with a wide-ranging of applications in healthcare, human-computer interaction, and gaming. Emotion recognition is challenging due to several input modalities, have a significant role in understanding it. “The mission of recognizing of the emotions is mostly difficult dueto two main reasons: 1) There is not largely available database of training images and 2) classifying emotion could not be simplebased on whether the input image is static or aevolution frame into a facial expression. The finaldifficulty is mostly for the real-time detection while facial expressions differenthusiastically”. Ekman et al. [1] counted six expressions (surprise, fear, happiness, anger, disgust, and sadness) as main emotional expressions that are common among human beings. Mostly the big overlap between the emotion classes makes the classification task very difficult. This paper proposed a deep learning technique in the context of emotional recognition, in order to classify emotion labels from the images. Too many methods and research has been developed in this regards, however, most current works are appeared focusing on hand-engineered features [2,3]. Now a day’s due to quantity and variety of datasets, deep learning is becoming as mainstream techniques in all computer visions tasks [4,5]. Conventional convolutional neural systems have a noteworthy constraint that they simply handle spatial image. The essential commitment of this work is to display the spatio-worldly development of outward appearances of a man in the Images utilizing a “Recurrent Neural Network (RNN) which embedded with a Convolutional Neural Network (CNN) in a form of CNN–RNN design”. We additionally introduce a neural system based element level combination procedure to join diverse modalities for the last emotion forecast. The pioneering works in emotion recognition based deep learning [6,7] has achieved the state-of-the-art. The cornerstone of these proposed models [6,8] is an average-based aggregation for visual features. A little distinguish from current works, we pro-

* Corresponding author.
E-mail address: neha.juet@gmail.com (N. Jain).

https://doi.org/10.1016/j.patrec.2018.04.010
0167-8655/

posed an RNN to classify the facial emotion. The proposed model explores feature level fusion strategy and proves the moderate improvement by this model. The other parts of the paper are organized as: next section delivers the related work in what we follow. Section 3 presents the proposed network. The results and experiments are included in Section 4. At the end, we have concluded our observation in Section 5.

2. Related work

"Generally, research works in this area have been focused on identifying human emotion in the base of video footage or based on audiovisual records (mixing speech recognition and video techniques). Several papers pursue to identify and match faces [20], nevertheless most works did not use deep learning to extract emotions from images". Customarily, calculations for mechanized outward appearance acknowledgment comprises of three primary modules, viz. enlistment, highlight extraction, and arrangement. Point by point study of various approaches in every one of these means can be found in [9]. "Customary calculations for full of feeling registering from faces utilize designed highlights for example, Histogram of Oriented Gradients [11], Local Binary Patterns [10] and facial historic points [12]". Since the greater parts of these highlights are hand-created for their particular use of acknowledgment, so the generalization in the particular situation is necessary, such as, high variability in lighting, subjects ethnicity, visual determination, and so on. Interestingly, the powerful methodologies for accomplishing a great acknowledgment for series of marking errand are alluded to separate the transient relations of edges in an arrangement. Separating these transient relations have been examined utilizing customary techniques before. Cases of these endeavors are Concealed Markov Models [13,14,47,48] "(which join the data and then apply division on recordings), Spatio Temporal Shrouded Markov Models by combing S-HMM and T-HMM [15], Dynamic Bayesian Networks" [16] is related to multi-tactile data combination paradigm, Bayesian transient models to catch the dynamic outward appearance progress, and Conditional Irregular Fields (CRFs) [17,18] and their augmentations. Recently, "Convolutional Neural Networks" (CNN) has turned into the most mainstream approach in the deep learning techniques. AlexNet [19] depends on the conventional layered engineering which comprises of a few convolution layers, max-pooling layers and Rectified Linear Units (ReLUs). Szegedy et al. [20] presented GoogLeNet which is made out of numerous "Beginning" layers. Commencement applies a few convolutions on the include outline distinctive scales. Mollahosseini et al. [21] have utilized the Inception layer for the undertaking of outward appearance acknowledgment and accomplished best in class comes about. Following the accomplishment of Inception layers, a few varieties of them have been proposed [22]. "RNNs recently have greatly succeeded in handling sequential data such as speech recognition [23], natural language processing [24,25], action recognition [26], and so on. Then RNN is additionally has been improved to treat the images [27] by scanning the parts of images into sequences in certain directions. Due to the capability of recollecting information about the past inputs, RNN has the ability to learn relative dependencies with images, which is advantageous in comparison with CNN. The reason is CNN may fail to learn the overall dependencies because of the locality of convolution and pooling layers. Therefore, RNN is generally combined with CNN in order to achieve better achievement in image processing tasks such as image recognition [28] and segmentation [29]". Conventional Recurrent Neural Networks (RNNs) can learn fleeting progression by mapping input successions to a grouping of concealed states, and furthermore mapping the covered up states to yields. Zhang et al. [30] "proposed a novel deep learning framework called as a spatial-temporal recurrent neural network (STRNN) to unify the learning of two different signal sources into a spatial-temporal dependency model". Khorrami et al. in [31,45,46], developed a method which used the CNN and RNN in order to perform emotion recognition on video data. Chernykh et al. [32] and Fan et al. [33] proposed CNN + RNN models for the video and speech recognition. In spite of the fact that RNNs have demonstrated promising execution on different assignments, it is difficult for them to learn long haul successions. This is mostly the result of vanishing/detonating slopes issue [34] that can be understood by having a memory for recalling and overlooking the past states. "Xie and Hu [42] presented a new CNNmodel that used convolutional modules. tominimize redundancy of same features learned, considers communal information among filters of the same layer, and offers the top set of features for the next layer.A distinguishedapplication of a CNN to real-time detection of emotions from facial expressions is by Oullet [43]. Theymade a game, while a CNN was applied to a video stream to grab the subject's facial expressions, performing as a controller for the game. This work established the possibility of executing a CNN in real-time by means of a running-average of the perceived emotions from the input stream, decreasing the special effects of variation and noise. A latest development by Levi et al. [44] illustrated important upgrading in facial emotion recognition using a CNN. They listed two main drawbacks: 1) the small amount of available data for training deep CNNs and 2) appearance dissimilarity generally affected by dissimilarities in illumination".

Distinct from other work including video and RNN strategies, [35], in this paper we don't utilize LSTMs. However, we utilize IRNNs [36] that is made out of amended straight units (ReLUs) what's more; utilize a unique introduction system in view of scaled varieties of the character grid. These components of IRNNs are gone for giving a substantially less difficult system to managing with the vanishing and detonating inclination issue thought about to the more perplexing LSTM system. Late work has contrasted IRNNs and LSTMs and found that IRNNs can yield equivalent outcomes in a few errands, including issues which include long haul conditions" [36]. We give point by point details of the CNN and the RNN structure the in next Section. Moreover, we concatenated the CNN highlights to a permanent distance feature vectors and furthermore, trained on SVM.

3. Proposed model

The opposition dataset has only a single emotion label for each picture and do not have relation to each casing. This presents a great deal of commotion if the picture labels are utilized as focuses on preparing a CNN on the singular image. Our visual highlights are in this way given by a CNN prepared on a mix of two extra emotion datasets of static pictures. In addition, utilizing extra information covers a bigger assortment of age and character rather than the test information where a similar performing artist/on-screen character might show up in numerous clips. For the CNN training we used two large emotion datasets, MMI Facial Expression Database (TFD) [37] it consists more than 2900 images of 75 subjects and "Japanese Female Facial Expression (JAFFE) Database [38] containing 213 pictures, which have seven basic expressions: angry, sad, surprise, happy, disgust, fear, and neutral".

For the preprocessing, we represent fluctuating lighting conditions (specifically, crosswise over datasets) we connected histogram evening out. We utilized the adjusted appearances gave by the coordinators to remove highlights from the CNN. "The arrangement includes a joined facial key point's location and following methodology clarified in [39]. Extraordinary confront location, as well as arrangement procedures, have been utilized for MMI Facial Expression and the JAFFE Datasets". Keeping in mind the end goal to be ready to use the extra datasets, we re-adjusted all datasets to JAFFE utilizing the accompanying method:

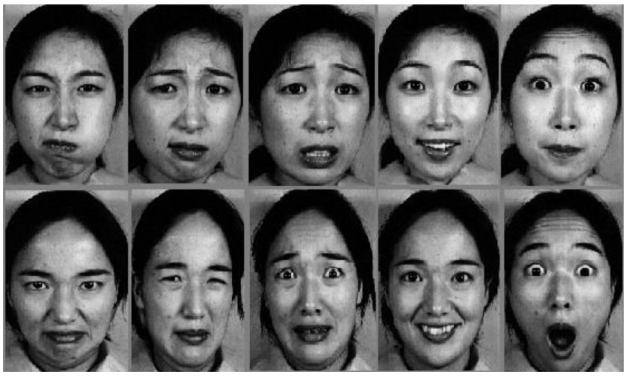

Fig. 1. Sample of JAFFE dataset for five type of emotion (Ang, Sad, Fea, Hap, Sur).

1. We distinguished five facial key focuses for all pictures in the JAFFE and MMI preparing set utilizing the convolutional neural system course strategy in [40].
2. for every dataset, the mean shape have been processed by averaging the directions of main focuses.
3. The datasets have been mapped by utilizing a closeness change among the mean shapes. By processing one change for each dataset the nose, eyes, and mouth is generally in a similar area holding a slight measure of variety. We included an uproarious fringe for MMI and JAFFE-faces as appearances were edited all the more firmly contrasted with JAFFE.
4. JAFFE-faces approval test sets were mapped to utilize the change construed on the preparation set.

Additionally, dataset standardization has been performed by using the standard deviation and mean picture from the consolidated JAFFE and MMI (JAFFE + MMI). Fig. 1 represents the samples face emotion data. For the implementation and the evaluation of the proposed model the 70% of each dataset used for training and the rest 30% for testing.

3.1. Convolution neural network architecture

Emotion Recognition data comes in various sizes and resolutions, so we try to propose a model which can handle any type of input. In our approach, "we considered a class of networks with 6 convolutional layers plus 2 fully connected layers", each with a ReLu activation function, and dropout for training.

Plus 2 fully connected layers", each with a *ReLu* activation function, and dropout for training. Furthermore, we performed regularization for each weight matrix W that limits the size of the weights at individual layer by adding a term to the loss equal to some fixed hyperparameter. We explain these in Eq. (1), where x be the output of a particular neuron in the network and p the dropout possibility.

$$\begin{aligned} &\mathrm{ReLu}(x) = \max(0, x) \\ &\mathrm{Dropout}(x, p) = \begin{cases} x, \text{with prob.} p \\ x, \text{with prob.} 1 - p \end{cases} \\ &Reg(w) = \lambda ||w||_2^2 \end{aligned} \tag{1}$$

Combinations of two deep learning initializer algorithms have been used to perform parameter updates based on the gradient of the loss function called as Momentum and Adam [41,42]. Eq. (2) describes this update, where X_t is parameter matrix at iteration t. v_t is the velocity vector at iteration t, and α is the rate of learning.

$$\begin{aligned} v_t &= \gamma v_{t-1} - a\nabla X_{t-1} \\ X_t &= X_{t-1} + v_t \end{aligned} \tag{2}$$

Eq. (3) illustrates the Adam update and its combination with the momentum update.

$$\begin{aligned} m_t &= \beta_1 m_{t-1} + (1 - \beta_1)\nabla X_{t-1} \\ v_t &= \beta_2 v_{t-1} + (1 - \beta_2)\nabla (X_{t-1})^2 \\ X_t &= X_{t-1} - \frac{a m_t}{\sqrt{v_t + \varepsilon}} \end{aligned} \tag{3}$$

$\beta 1, \beta 2 \in [0,1]$ and ε are hyperparameters, m_t is the momentum vector with t iteration, v_t is the velocity vector, and the learning rate of α. Adam is the actual update algorithm due to information usage for the primary and the secondary moments of the gradient.

The CNN is used primarily for feature extraction and we have just utilized the extra dataset for the training. Accordingly, we hunt down a model that have better communalize to different datasets. Profound models are known to learn portrayals to have better communalize to different datasets. By the way, it has been found out that the deep structure rapidly over-fitted, and communalize severely to the test dataset. This could be because of the generally little measure of marked information accessible for the emotion detection tasks. Consequently, "we build different connections between 6 layers which seem to have decent tended to the over-fitting issues. At the end, we expanded the filter size from 3 to 5 and the numbers of channels are 8-16-32-64-128-256. For the experimentations data augmentation has been used" ((horizontal, vertical and rotation flipping with 0.25 probability), and dropout is used (with the rate of 0.5).

RNNs are a form of neural network that converts the order of inputs into a series of outputs. In separately time step t, an unknown parameter h_t is calculated according to the unknown parameter at time $t-1$ and the input x_t at time t

$$h_t = \sigma(W_{in}x_t + W_{rec}h_{t-1}) \tag{4}$$

While "W_{in} is the weight of input matrix, W_{rec} is the matrix of recurrent and σ is the hidden activation function. Respectively time step similarly calculates the outputs, relying on the existing hidden state":

$$y_t = f(W_{out}h_t) \tag{5}$$

While W_{out} is the result weighted parameters and f is the activation function of the output. An instance of an RNN in which merely the last phase creates the output which illustrated in Fig. 3.

"An RNN model has been used, that previously discussed by using rectified linear units (ReLUs) and recurrent matrix, which is adjusted with scaled deviations of the distinctiveness matrix" [42]. The distinctiveness initialization model confirms good gradient movement at the commencement of training and it consents to train it on moderately extensive orders. The RNN ha been trained to categorize the images by inserting the extracted features of each image from the CNN serially network and finally using the Softmax for the prediction. In the implementation the Gradient clipping rated to 1.0 and a batch size set to 32. We tested the model by using several layers of the CNN as input features and picked the output of every third convolutional layer right after max pooling, as this achieved the highest result on validation data.

3.2. Regression CNN

Firstly we used a single CNN model to train the datasets. At each time trained a single image, the corresponding image passed through the CNN model, the details of the model shown in Fig. 2.

Two fully-connected layers with 200 hidden units for the approximation of the valence label have been used. For the cost function the mean squared error has been used. For the network training stochastic gradient descent while the batch size sets to 32 and the weight decay sets to 1E-4. Moreover, the learning rate at the beginning sets to 5e-3 which decrees by 0.01 every 20 epochs.

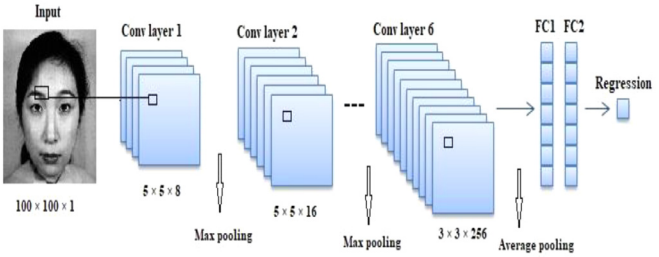

Fig. 2. CNN Architecture, the network contains six convolutional layers containing 8, 16, 32, 64, 128, and 256 filters; each of size 5×5 and 3×3 followed by ReLU activation functions. 3×3 max-pooling layers added just after every first five convolutional layers and average pooling at the last convolution layer. Every convolutional layer has two fully-connected layers and 200 hidden units.

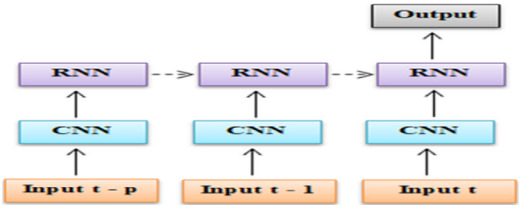

Fig. 3. Hybrid CNN-RNN Network Architecture.

3.3. Combining with recurrent neural networks (RNNs)

In the proposed model like the model which presented by [31], we propose to combine the sequential information by using RNN to spread information. The CNN model used for feature extraction to fix all of its parameters and to eliminate the regression layer. For the processing, when the image passed to the network, 200-dimensional vectors will be extracted from the fully-connected layers. For the assumed time t, we take P frames from the past (i.e. $[t-P, t]$). Then passes every frame from time $t-P$ to t to the CNN and extract P vectors fully for each image. Each and every vector goes through a node of the RNN model. Then every node of the RNN returns some results of valence label. The overall proposed method illustrated in Fig. 3. The mean squared error has been used for the cost function while optimizing.

4. Experiment and evaluation

For the data preprocessing, we initially identify the face in every outline utilizing face and point of interest finder. Then map the distinguished landmark points to characterized pixel areas in a request to guarantee correspondence concerning outlines. After the normalization the nose, mouth and nose organizes, while processing each face image through the CCN mean subtraction and contrast normalization applied. We tested the proposed models on a normal PC with Intel(R) Core(TM) i7-8700 K and 24 GB of RAM.

4.1. Compare the CNN with hybrid CNN-RNN

Fig. 4 shows the loss and the prediction accuracy of the Hybrid CNN-RNN model for training vas validation for one set of the Images. These charts clearly illustrate the smooth performance of the proposed model.

Table 1 presents the prediction accuracy of the proposed single frame regression CNN and Hybrid CNN-RNN technique implemented for predicting valence scores of subjects to developing a set of the dataset. Finally, when combining the information and using the Hybrid CNN-RNN model with the ReLU, a significant performance could be achived.

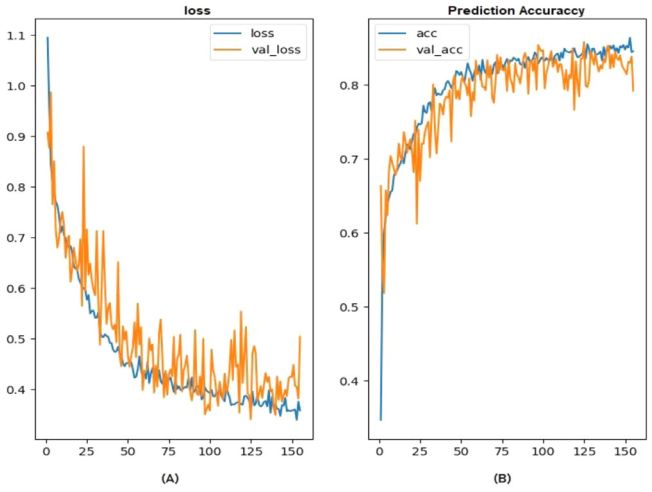

Fig. 4. Loss and the Prediction Accuracy for Hybrid CNN-RNN model.

Table 1
Overall accuracy and mean accuracy for the different models.

Method	Overall accuracy	Mean class accuracy
CNN	76.51%	74.33%
CNN - RNN	91.20%	89.13%
CNN - RNN + ReLU	**94.46%**	**93.67%**

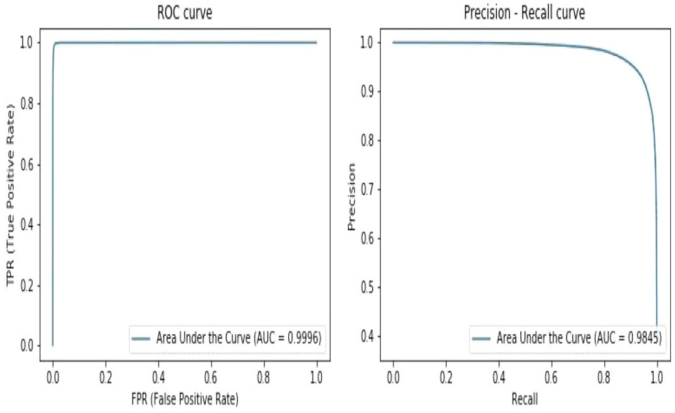

Fig. 5. Roc and the Precision-Recall Curve.

Table 2
Result of altering the number of hidden units.

Method	Prediction accuracy	Loss
Hybrid CNN-RNN, hidden units = 50	92.32%	4.73%
Hybrid CNN-RNN, hidden units = 100	93.57%	4.72%
Hybrid CNN-RNN, hidden units = 150	**94.21%**	**4.43%**
Hybrid CNN-RNN, hidden units = 200	92.53%	4.68%

Fig. 5 displays the Roc curve and the Precision-Recall curve of the proposed hybrid model. As it is visible the proposed model has the ability to with the least number of errors used for the face emotion detection.

We evaluated the special effects of two hyperparameters in the results of Hybrid proposed model, namely the number of hidden units and the number of hidden layers. Table 2 concluded that, the best result can achieve with 150 hidden units and in the other cases rather than improvement in the performance resulted in decreases. Table 3 “shows that increasing the number of hidden layers resulted to improve the overall performance of the proposed

Table 3
Result of altering the number of hidden layers.

Method	Prediction accuracy	Loss
HybridCNN-RNN, hidden layers = 4	94.57%	4.28%
HybridCNN-RNN, hidden layers = 5	94.73%	4.22%
Hybrid CNN-RNN, hidden layers = 6	**94.91%**	**3.98%**

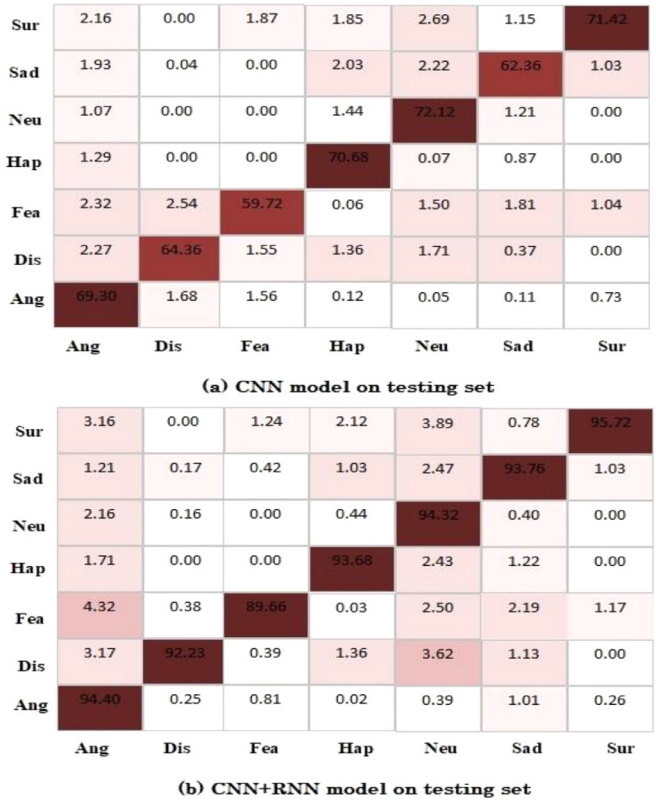

Fig. 6. Confusion matrices on JAFFE Datasets.

Table 4
Proposed model versus other models performance comparison on JAFFE and MMI dataset.

Method	Accuracy of JAFFE	Accuracy of MMI
Zhang et al. [30]	**94.89%**	91.83%
Khorrami et al. [31]	82.43%	81.48%
Chernykh et al. [32]	73%	70.12%
Fan et al. [33]	79.16%	77.83%
Proposed model	**94.91%**	**92.07%**

model. Hence, based on the experiments, the best results obtained by the 6 hidden layers".

The confusion matrices of CNN and Hybrid CNN-RNN models on the testing sets are presented in Fig. 6. Hybrid CNN-RNN model could achieve an accuracy of 94.72%, while a single CNN can reach only to 71.42%. The combined model not only increases the overall accuracy of the proposed CNN model but also it reduces the false detection of the model. As it is clearly visible in the Fig. 6 the best detection are for the Ang, Neu, and Sur emotions.

Table 4 indications the performance of proposed Hybrid CNN-RNN model in comparison with other approaches evaluated on the JAFFE and MMI datasets. The proposed CNN-RNN model achieved equal or greater performance as compared to the four other state-of-the-art methods [30–33].

4.2. Comparison of the proposed model with other approaches

Our model has slightly better performance than the model which proposed by Zhang et al. [30], While, the other models have the lower performance in comparison with the proposed model.

5. Conclusion

In this paper, a model has been proposed for face emotion recognition. We proposed a hybrid deep CNN and RNN model. In addition, the proposed model evaluated under different circumstances and hyper parameters to properly tuning the proposed model. Particularly, it has been found that the combination of the two types of neural networks (CNN-RNN) cloud significantly improve the overall result of detection, which verified the efficiency of the proposed model.

References

[1] P. Ekman, W.V. Friesen, Constants across cultures in the face and emotion, J. Pers. Soc. Psychol. 17 (2) (1971) 124.
[2] S.E. Kahou, P. Froumenty, C. Pal, Facial expression analysis based on high dimensional binary features, ECCV Workshop on Computer Vision with Local Binary Patterns Variants, 2014.
[3] C. Shan, S. Gong, P.W. McOwan, Facial expression recognition based on local binary patterns: a comprehensive study, Image Vis. Comput. 27 (6) (May 2009) 803–816.
[4] N. Kalchbrenner, E. Grefenstette, and P. Blunsom. A convolutional neural network for modelling sentences. arXiv:1404.2188, 2014.
[5] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Adv. Neural Inf. Process. Syst. (2012) 1097–1105.
[6] S.E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, C. Gulcehre, et al., Combining modality specific deep neural networks for emotion recognition in video, International Conference on Multimodal Interaction, ICMI '13, 2013.
[7] M. Liu, R. Wang, S. Li, S. Shan, Z. Huang, X. Chen, Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild, in: International Conference on Multimodal Interaction, ICMI '14, 2014, pp. 494–501.
[8] S.E. Kahou, X. Bouthillier, P. Lamblin, C. Gulcehre, V. Michalski, et al., Emonets: multimodal deep learning approaches for emotion recognition in video, J. Multimodal User Interfaces (2015) 1–13.
[9] E. Sariyanidi, H. Gunes, A. Cavallaro, Automatic analysis of facial affect: a survey of registration, representation, and recognition, IEEE Trans. Pattern Anal. Mach. Intell. 37 (6) (2015) 1113–1133.
[10] C. Shan, S. Gong, P.W. McOwan, Facial expression recognition based on local binary patterns: a comprehensive study, Image Vis. Comput. 27 (6) (2009) 803–816.
[11] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, 1, IEEE, 2005, pp. 886–893.
[12] T.F. Cootes, G.J. Edwards, C.J. Taylor, et al., Active appearance models, IEEE Trans. Pattern Anal. Mach. Intell. 23 (6) (2001) 681–685.
[13] M. Yeasin, B. Bullot, R. Sharma, Recognition of facial expressions and measurement of levels of interest from video, Multimedia, IEEE Trans. 8 (3) (2006) 500–508.
[14] Y. Zhu, L.C. De Silva, C.C. Ko, Using moment invariants and hmm in facial expression recognition, Pattern Recognit. Lett. 23 (1) (2002) 83–91.
[15] Y. Sun, X. Chen, M. Rosato, L. Yin, Tracking vertex flow and model adaptation for three-dimensional spatiotemporal face analysis, Syst. Man Cybern. Part A 40 (3) (2010) 461–474.
[16] N. Sebe, M.S. Lew, Y. Sun, I. Cohen, T. Gevers, T.S. Huang, Authentic facial expression analysis, Image Vis. Comput. 25 (12) (2007) 1856–1863.
[17] B. Hasani, M.M. Arzani, M. Fathy, K. Raahemifar, Facial expression recognition with discriminatory graphical models, in: 2016 2nd International Conference of Signal Processing and Intelligent Systems (ICSPIS), Dec 2016, pp. 1–7.
[18] B. Hasani and M.H. Mahoor. Spatio-temporal facial expression recognition using convolutional neural networks and conditional random fields. arXiv:1703.06995, 2017.
[19] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Adv. Neural Inf. Process. Syst. (2012) 1097–1105.
[20] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.
[21] A. Mollahosseini, B. Hasani, M.J. Salvador, H. Abdollahi, D. Chan, M.H. Mahoor, Facial expression recognition from world wild web, In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, June 2016.
[22] S. Ioffe and C. Szegedy. Batch normalization: accelerating deep network training by reducing internal covariate shift. arXiv:1502.03167, 2015.
[23] A. Graves, A. r. Mohamed, G. Hinton, Speech recognition with deep recurrent neural networks, in: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, 2013, pp. 6645–6649.

[24] A. Graves, N. Jaitly, Towards end-to-end speech recognition with recurrent neural networks, in: Proc. International Conference on Machine Learning, 2014, pp. 1764–1772.
[25] T. Mikolov, M. Karafiát, L. Burget, J. Cernock 'y, S. Khudanpur, Recurrent neural network based language model, in: Proc. INTERSPEECH, 2, 2010, pp. 1045–1048.
[26] A. Sanin, C. Sanderson, M.T. Harandi, B.C. Lovell, Spatiotemporal covariance descriptors for action and gesture recognition, IEEE Workshop on Applications of Computer Vision, 2013.
[27] S. Jain, C. Hu, J.K. Aggarwal, Facial expression recognition with temporal modeling of shapes, in: Proc. IEEE International Conference on Computer Vision Workshops, 2011, pp. 1642–1649.
[28] F. Visin, K. Kastner, K. Cho, M. Matteucci, et al., Renet: A recurrent neural network based alternative to convolutional networks. arXiv:1505.00393, 2015
[29] F. Visin, K. Kastner, A. Courville, Y. Bengio, et al., ReSeg: a recurrent neural network for object segmentation. arXiv:1511.07053, 2015
[30] T. Zhang, W. Zheng, Z. Cui, Y. Zong, Y. Li, Spatial-temporal recurrent neural network for emotion recognition, IEEE Trans. Cybern. (99) (2018) 1–9 arXiv:1705.04515.
[31] P. Khorrami, T.L. Paine, K. Brady, C. Dagli, T.S. Huang, How Deep Neural Networks can Improve Emotion Recognition on Video Data, IEEE Conf. Image Process (ICIP) (2016).
[32] V. Chernykh, G. Sterling, P. Prihodko, Emotion recognition from speech with recurrent neural networks, arXiv:1701.08071v1 [cs.CL], 2017
[33] Y. Fan, X. Lu, D. Li, Y. Liu, Video-based emotion recognition using CNN-RNN and C3D hybrid networks, in: ACM International Conference on Multimodal Interaction (ICMI 2016), 2016, pp. 445–450.
[34] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[35] J. Donahue, L.A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, T. Darrell, Long-Term Recurrent Convolutional Networks for Visual Recognition and Description, IEEE Trans. Pattern Anal. Mach. Intell. 39 (4) (2017) 677–691, doi:10.1109/TPAMI.2016.2599174.
[36] Q.V. Le, N. Jaitly, and G.E. Hinton. A simple way to initialize recurrent networks of rectified linear units. arXiv:1504.00941, 2015.
[37] J. Susskind, A. Anderson, and G. Hinton. The toronto face database. Technical report, UTML TR 2010-001, University of Toronto, 2010.
[38] M.J. Lyons, J. Budynek, S. Akamatsu, Automatic classification of single facial images, IEEE Trans. Pattern Anal. Mach. Intell. 21 (12) (1999) 1357–1362, doi:10.1109/34.817413.
[39] A. Dhall, R. Goecke, J. Joshi, K. Sikka, T. Gedeon, Emotion recognition in the wild challenge 2014: baseline, data and protocol, in: International Conference on Multimodal Interaction, ICMI '14, 2014, pp. 461–466.
[40] Y. Sun, X. Wang, X. Tang, Deep convolutional network cascade for facial point detection, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR '13, 2013, pp. 3476–3483.
[41] I. Sutskever, J. Martens, G. Dahl, G. Hinton, On the importance of initialization and momentum in deep learning, in: Proceedings of the 30th International Conference on Machine Learning, 2013, pp. 1139–1147. D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. CoRR, abs/1412.6980, 2014.
[42] Xie S., Hu H., Facial expression recognition with FRR – CNN, Electron. Lett. 53 (4) (2017) 235–237.
[43] S. Ouellet, Real-time emotion recognition for gaming using deep convolutional network features, CoRR, vol. abs/1408.3750, 2014.
[44] G. Levi, T. Hassner, Emotion recognition in the wild via convolutional neural networks and mapped binary patterns, in: Proc. ACM International Conference on Multimodal Interaction (ICMI), November, 2015.
[45] D.K. Jain, R. Kumar, N. Jain, Decision-based spectral embedding approach for identifying facial behaviour on RGB-D images, Int. Conf. Commun. Netw. 508 (2017) 677–687.
[46] D.K. Jain, Z. Zhang, K. Huang, Hybrid patch based diagonal pattern geometric appearance model for facial expression recognition, in: Conference on Intelligent Visual Surveillance, 2016, pp. 107–113.
[47] D.K. Jain, Z. Zhang, K. Huang, Multi angle optimal pattern-based deep learning for automatic facial expression recognition, Pattern Recognit. Lett. (2017), doi:10.1016/j.patrec.2017.06.025.
[48] D.K. Jain, Z. Zhang, K. Huang, Random walk-based feature learning for micro-expression recognition, 2018. https://doi.org/10.1016/j.patrec.2018.02.004.