Pattern Recognition Letters 120 (2019) 69–74

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Extended deep neural network for facial emotion recognition

Deepak Kumar Jain[a], Pourya Shamsolmoali[b,*], Paramjit Sehdev[c]

[a] Key Laboratory of Intelligent Air-Ground Cooperative Control for Universities in Chongqing, College of Automation, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
[b] Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China
[c] Department of Mathematics and Computer Science, Coppin State University, USA

ARTICLE INFO

Article history:
Available online 11 January 2019

Keywords:
Facial emotion recognition
Deep neural network
Fully convolution network

ABSTRACT

Humans use facial expressions to show their emotional states. However, facial expression recognition has remained a challenging and interesting problem in computer vision. In this paper we present our approach which is the extension of our previous work for facial emotion recognition [1]. The aim of this work is to classify each image into one of six facial emotion classes. The proposed model is based on single Deep Convolutional Neural Networks (DNNs), which contain convolution layers and deep residual blocks. In the proposed model, firstly the image label to all faces has been set for the training. Secondly, the images go through proposed DNN model. This model trained on two datasets Extended Cohn–Kanade (CK+) and Japanese Female Facial Expression (JAFFE) Dataset. The overall results show that, the proposed DNN model can outperform the recent state-of-the-art approaches for emotion recognition. Even the proposed model has accuracy improvement in comparison with our previous model.

1. Introduction

Facial expression is one of the most important features of human emotion recognition. As persons we can assume the impression of someone's emotions by observing their face. Facial emotion recognition has several applications in computer vision, non-verbal human behavior and human-computer interaction. It is a challenging task due to several issues for example similarity of actions, large head poses and etc. Recently, Facial emotion recognition in images has attracted growing attention [5], which is for instances more complicated due to the backgrounds and low-resolution faces. Emotion refers to the feeling, energy, and dispositional effects. The goal of this work is to classify the facial emotion as sad, angry, happy, surprise, fear, disgust, and neutral. Facial expression is one of the most powerful, natural and common signals for human beings to convey their emotional states and intentions.

It is computationally complex and challenging to achieve high recognition rates using conventional feature extraction and classification schemes. This paper proposed a new deep learning model for the emotional recognition, to classify facial emotion from the images. Several techniques have been established in this regards, but, most current works are appeared focusing on hand-engineered features [2,3,10]. Currently, due to the size and variety of datasets, deep neural network is the most appropriate techniques in all the image processing and computer visions tasks [4,5,8]. Deep conventional networks have the ability to simply handle spatial image [6,16,17]. The main commitment of this work is to propose a deep neural network model which contains several convolution layers and deep residual blocks for facial emotion recognition [14,15]. The proposed model is able to learn the subtle features that discriminate the six different facial expressions [9,11,13]. The rest of the paper is organized as follow: next section delivers the related works. Section 3 presents the proposed network. The results and experiments are presented in Section 4. At the end, we have concluded our observation in Section 5.

2. Related work

The researches in the area of facial emotion detection focused on recognizing human emotion based on image or video records. Some recent work pursue to recognize faces in images or video records [20], however, these approaches did not use neural network strategy to extract features from the images.

Based on the features extracted for the recognition, we can differentiate two fundamental methodologies: geometric based approaches and appearance approaches. For the first scenario, the model focus on limiting and tracking specific facial standards, in order to train the model to classify based on relevant positions of

* Corresponding author.
E-mail addresses: pshams55@gmail.com, pshams@sjtu.edu.cn (P. Shamsolmoali).

https://doi.org/10.1016/j.patrec.2019.01.008
0167-8655/

Fig. 1. Sample of dataset for four type of emotion (Sur, Ang, Hap, Neu).

these standards. In [26] the authors proposed a model to track a set of points to classify emotions from sequences. In the similar way [24] the authors applied a transformation on a reduced subset of 117 landmarks for emotion recognition. In form based emotion recognition, a set of features is extracted from pixel images to train the classification model. Ko [27] presented a review article which focuses on recent hybrid deep-learning models. Krestinskaya and James [28], proposed an emotion recognition model based on min-max similarity which reduces the problem of interclass pixel mismatching while classification. This paper analyzes the strengths and the limitations of systems based only on facial expressions or acoustic information. It also discusses two approaches used to fuse these two modalities: decision level and feature level integration.

Busso et al. [29] evaluate the strengths and the boundaries of models based on facial expressions or acoustic material. In addition, the authors discussed two models that used to fuse decision level and feature level integration. Lopes et al. [31] proposed a simple way for facial expression recognition by combining Convolutional Neural Network and some image pre-processing methods.

Regardless of the current progresses, still emotion recognition remains an open problem for the computer vision community. EmotionNet is the largest Challenge in terms of data available, with more than 1 million images (2000 labeled emotions and 950 K unlabeled samples). The first version of the challenge determined that non-frontal faces still pose main problem for the classification algorithms. So, the recognition rates decrease as a function of pitch and yaw rotations [5].

For the last couple of years, Convolutional Neural Networks (CNNs) is one of the most popular approaches in the computer vison. For example, C. Szegedy et al. [20] proposed a model called as GoogLeNet which has numerous convolution layers and this model had novelty due to usage of inception blocks. In every inception block there are several convolution layers which are individually connected to each other and at the last stage the results of convolution layers concatenated and pass to the next convolution layer out of inception block.

3. Datasets and pre-processing

3.1. Dataset details

To train a neural network, a huge amount of labeled data is required to handle the curse of dimensionality [5]. There are several facial expressions datasets are publicly available such as Cohn–Kanade (CK+) [7] and Japanese Female Facial Expression (JAFFE) datasets and were used in this paper. A big part of datasets with emotion labels have facial pose expressions that were recorded in a precise environment, with head pose and constant lighting. The models that trained on this kind of dataset mostly have low performance on the data which have different conditions, particularly in the live and outdoor environment. We performed training on non-similar datasets to have better generalization on the performance.

The CK dataset involves of 486 video sequences from 97 posers, in CK+ the number of sequences has been extended by 22%. The data has been scaled to 640×480 or 640×490 pixels and all

Table 1
Dataset description.

Base	Number of images	
Training	CK+	8000
	JAFFE	200
Testing	CK+	150
	JAFFE	13

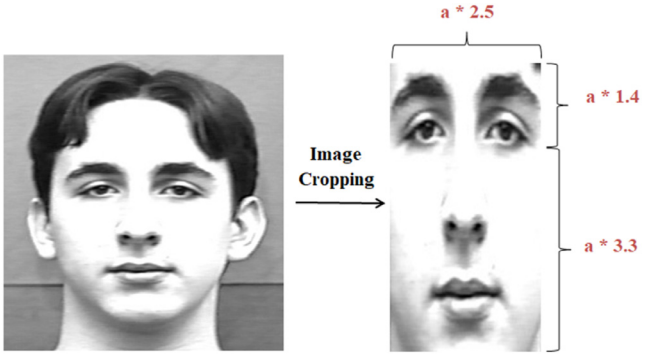

Fig. 2. Image cropping example. This action aims to remove all non-expression features, such as background and hair.

recorded in grayscale. Entire images were captured with the equal lighting and poses. The dataset is quite appropriate to use for all the models of feature extraction. The JAFFE dataset holds 213 images in total of seven facial expressions (six basic facial expressions and one neutral) posed by ten Japanese female models. The images are in size of 256×256 pixels and the expresser have 2–4 samples for every expression. The samples presented in Fig. 1.

The experimental dataset contains in total 8363 images. 8200 images are used for training, and the rest of images for testing and validation. The dataset details are in Table 1.

3.2. Data pre-processing

Image normalization is important pre-processing technique to decrease the inner-class feature mismatch which could be observed as intensity offsets. As the intensity offsets are constant in the local region, Gaussian normalization and standard deviation has been used. The input image is denoted as x(p, q), and y(p, q) is the normalized output image, while p and q are the row and column number of the processed image. The normalized resulted image is computed by Eq. (1) [19], where μ is a local mean and σ is a local standard deviation calculated over a window of $M \times M$ size [12].

$$\psi(\pi,\theta) = \frac{\xi(\pi,\theta) - \mu(\pi,\theta)}{6\sigma(\pi,\theta)} \tag{1}$$

The proposed model has the ability to handle different image sizes without human intermediation. The cropping section is delimited by a vertical factor of 4.7 (considering 1.4 for the above the eyes area and 3.3 for the area below) useful to the gap between the eyes middle point and the center of right eye. The horizontal cropping section is delimited by a factor of 2.5 functional to the similar distance. These factor values were determined empirically. A sample of this process is presented in Fig. 2.

The image intensity and contrast can differ even in the same person images and in the same expression, hence, increasing the variation in the feature vector which could increase the complexity of the problem that the classifier has to solve for each expression. To reduce these problems intensity normalization been used. This technique adapted from [30], which named contrastive equalization. In fact, the normalization process is a two-step scheme: firstly

Fig. 3. The intensity normalization. The original intensity image (left) and its intensity normalized form (right).

subtract local contrast; next, uses divisive local contrast normalization. Initially, the value of all pixels is subtracted from a Gaussian-weighted average of its neighbors. Later, each pixel is divided by the standard deviation of its own neighbor pixels. An example of this procedure is presented in Fig. 3.

Eqs. (2) and (3) are used to calculate the factors of μ and σ while a = (M − 1)/2.

$$\mu(\pi,\theta) = \frac{1}{M^2}\sum_{\kappa=-\alpha}^{\alpha}\sum_{\eta=-\alpha}^{\alpha}\xi(\kappa+\pi,\eta+\theta) \tag{2}$$

$$\sigma(\pi,\theta) = \sqrt{\frac{1}{N^2}\sum_{\kappa=-\alpha}^{\alpha}\sum_{\eta=-\alpha}^{\alpha}[\xi(\kappa+\pi,\eta+\theta)-\mu(\pi,\theta)]^2} \tag{3}$$

3.3. Feature detection

The feature parts that valuable for the facial emotion recognition are forehead, eyebrows, eyes, cheeks and mouth areas. Here, we show the feature detection by computing the local deviation of already normalized image by a window of N × N size. Eq. (4) used for the feature detection with b = (N − 1)/2.

$$\omega(\pi,\theta) = \sqrt{\frac{1}{N^2}\sum_{K=-\beta}^{\beta}\sum_{n=-\beta}^{\beta}[\psi(K+\pi,n+\theta)-\mu^3(\pi+\theta)]^2} \tag{4}$$

Parameter μ' In Eq. (4) denotes the mean of the normalized image y(p, q) and can be computed by Eq. (5).

$$\mu(\pi,\theta) = \frac{1}{N^2}\sum_{\kappa=-\beta}^{\beta}\sum_{\eta=-\beta}^{\beta}[\psi(\kappa+\pi,\eta+\theta)-\mu(\pi+\theta)]^2 \tag{5}$$

In the proposed model the convolution layers extracts features hierarchically, and the fully-connected layer and the softmax layer used for indicating 6 expression classes. The input of the network is 128 × 96 × k for all patches, where k = 3 for color patches and k = 1 for gray patches. The final output is one of the 6 expression classes.

4. Proposed model

In the following subsections the details of proposed model presented.

4.1. Convolution neural network architecture

Firstly, we introduce the use of face detector.

For the face detection we proposed a deep convolution neural network as shown in Fig. 4 which contains six convolution layers (the green blocks) and two blocks of deep residual learning (purple blocks) the details presented in the next section. Right after each convolution layer there is a max pooling layer (orange blocks). The deep residual blocks implemented after 2nd and 4th convolution layer. There are also 2 Fully Connected layers (FC), each with a ReLU activation function, and dropout for training [18,22,23].

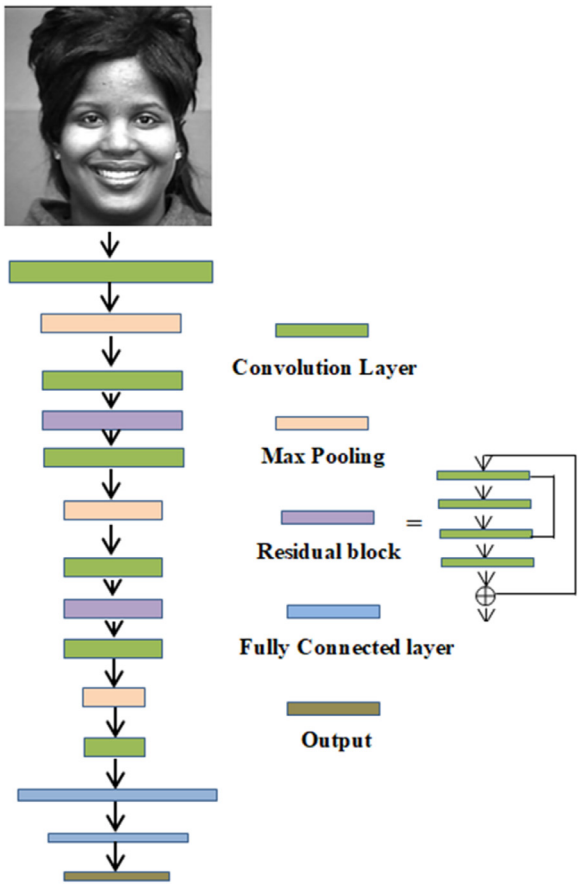

Fig. 4. Architecture of the proposed DNN model.

Table 2
Details of proposed network.

Type	Filter size / stride	Output size
Conv1	5 × 5 / 2	64 × 48 × 32
Maxpool1	2 × 2 / 2	32 × 24 × 32
Conv2	3 × 3 / 1	32 × 24 × 64
Res1	4 Conv	–
Conv3	3 × 3 / 1	32 × 24 × 128
Maxpool2	2 × 2 / 2	16 × 12 × 128
Conv4	3 × 3 / 1	16 × 12 × 128
Res2	4 Conv	–
Conv5	3 × 3 / 1	16 × 12 × 256
Maxpool3	2 × 2 / 2	8 × 6 × 256
Conv6	3 × 3 / 1	8 × 6 × 512
FC1	1024	1024
FC2	512	512

Finally we used Softmax for the classification [21]. The details of proposed network presented in Table 2.

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 a fixed hyper parameter. We explain these in Eq. (1), where *x* be the output of a particular neuron in the network and *p* the dropout possibility.

The DNN is used for feature extraction and we just used the extra dataset for the training.

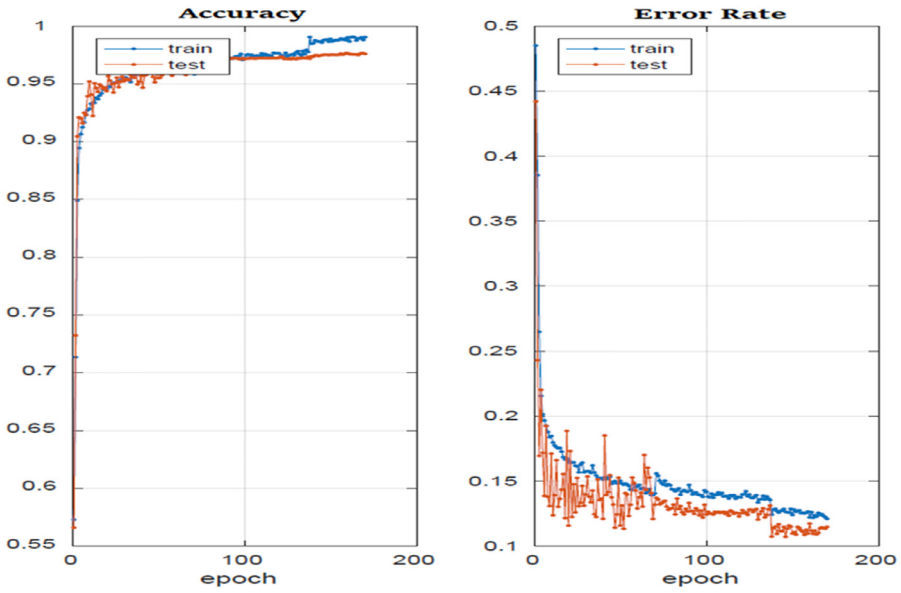

Fig. 5. Loss and the prediction accuracy for the proposed model.

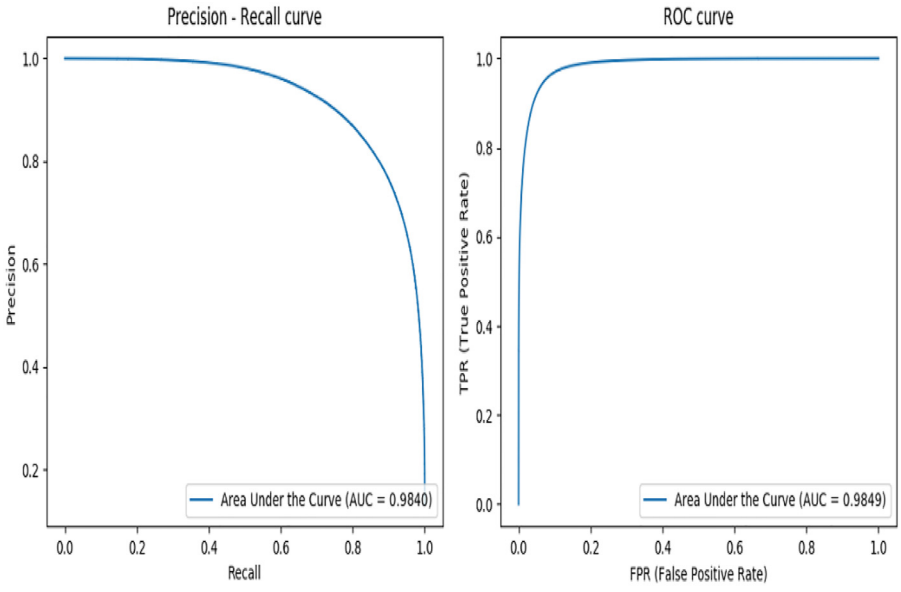

Fig. 6. Roc and the precision-recall curve.

4.2. Residual network

In the proposed model we implemented 2 residual blocks. Each residual block has four convolution layers two short connections and one skip connection. The first convolution layer in residual block has size of $1 \times 1 \times 64$. The second one has size of $3 \times 3 \times 64$. The third one has size of $3 \times 3 \times 128$ and the last one has size of $1 \times 1 \times 256$.

4.3. Regression DNN

Firstly we used a single DNN model to train the datasets. At each time trained a single image, the corresponding image passed through the DNN model, the details of the model shown in Fig. 4.

Two fully-connected layers with 250 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 64 and the weight decay sets to 1E-5. Moreover, the learning rate at the beginning sets to 4e-2 which decrees by 0.005 every 10 epochs.

5. Experiment and evaluation

For all the pre-training settings described in this paper, global average pooling at the last layer has been applied to decrease spatial dimensionality of data before passing to the fully connected layers. In addition, batch normalization is used to improve generalization and optimization.

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 forehead, eyes, nose, mouth while processing each face image through the subtraction and contrast normalization. For tested the proposed DNN model we used a workstation with Intel(R) Core(TM) i7-8700 K and 32GB memory.

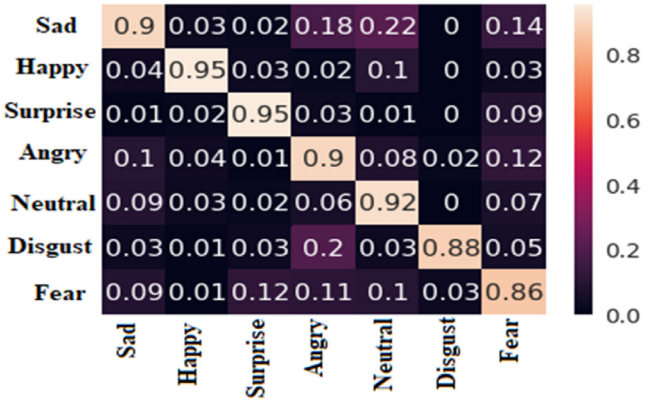

Fig. 7. Confusion matrices on JAFFE Datasets.

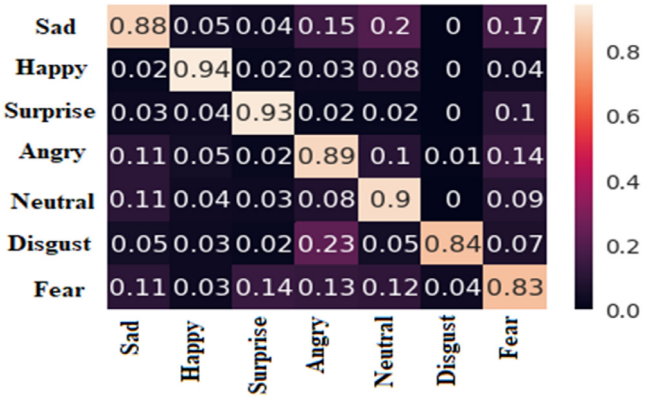

Fig. 8. Confusion matrices on CK+ Datasets.

Table 3
Proposed model versus other models Performance Comparison on JAFFE and MMI dataset.

Method	Accuracy JAFFE	Accuracy CK+
Zhang et al.[24]	94.89%	92.35%
Khorrami et al. [25]	82.43%	81.48%
Chernykh et al. [26]	73%	70.12%
Jain et al. [1]	94.91%	92.71%
Krestinskaya and James [28]	94.89%	92.74%
Lopes et al. [31]	94.86%	92.73%
Proposed model	**95.23%**	**93.24%**

Fig. 5 shows the prediction accuracy of the proposed DNN model for training vas validation on the CK+ dataset. These charts clearly show the smooth performance of the proposed model.

Fig. 6 presents the Roc curve and the Precision-Recall curve of the proposed model. As it is visible the proposed model has the ability to with the least number of errors and high performance for the face emotion recognition.

The confusion matrices of Proposed DNN model on the JAFFE dataset presented in Fig. 7 and its performance on CK+ dataset presented in Fig. 8. The proposed model emotion recognition can reach to 95%. As it is visible in the Figs. 7 and 8 the best recognition are for the Happy, Angry, Neutral, and Surprise emotions.

Table 3 illustrated the overall performance of proposed DNN model as compared to other deep learning recent models on the CK+ and JAFFE datasets. The proposed DNN model gain higher results in comparison with five other models [[1,24,25], and [26]].

Our recent approach has better performance in emotion recognition in comparison with our previous model; in Table 3 we present the performance of recent approaches on the whole JAFFE and CK+ dataset.

As the results shows the proposed model on the JAFFE dataset has 0.32%, 0.34% better performance as compared to Jain et al. [1] and Zhang et al. [24] respectively. On the CK+ dataset the proposed model has 0.53% better performance in comparison with Jain et al. [1] and 0.89% better performance while comparing with Zhang et al. [24] model.

6. Conclusion

In this paper, we present a fully deep neural network model for facial emotion recognition and the model has been tested on two public datasets to assess the performance of the proposed model. Particularly, it has been found that the combination of FCN and residual block cloud considerably improve the overall result, which verified the efficiency of the proposed model.

References

[1] N. Jain, S. Kumar, A. Kumar, P. Shamsolmoali, M. Zareapoor, Hybrid deep neural networks for face emotion recognition, Pattern Recognit. Lett. (2018). https://doi.org/10.1016/j.patrec.2018.04.010.
[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 Vision Comput. 27 (May (6)) (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, in: Advances in neural information processing systems, 2012, pp. 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, 13, ICMI, 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, 14, ICMI, 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 Interf. (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] 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.
[11] 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.
[12] Y. Sun, X. Chen, M. Rosato, L. Yin, Tracking vertex flow and model adaptation for three-dimensional spatiotemporal face analysis, IEEE Trans. Syst. Man Cybern. Part A: Syst. Humans 40 (3) (2010) 461–474.
[13] N. Sebe, M.S. Lew, Y. Sun, I. Cohen, T. Gevers, T.S. Huang, Authentic facial expression analysis, Image Vision Comput. 25 (12) (2007) 1856–1863.
[14] M. Zareapoor, P. Shamsolmoali, J. Yang, Learning depth super-resolution by using multi-scale convolutional neural network, Journal of Intelligent & Fuzzy Systems (2018) 1–11, doi:10.3233/JIFS-18136.
[15] P. Shamsolmoali, M. Zareapoor, J. Yang, Convolutional neural network in network (CNNiN): hyperspectral image classification and dimensionality reduction, IET Image Processing (2018), doi:10.1049/iet-ipr.2017.1375.
[16] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.
[17] 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.
[18] M. Zareapoor, D.K. Jain, J. Yang, Local spatial information for image super-resolution, Cogn. Syst. Res. 52 (2018) 49–57.
[19] 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.
[20] 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.
[21] 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.

[22] P. Shamsolmoali, M. Zareapoor, D.K. Jain, V.K. Jain, J. Yang, Deep convolution network for surveillance records super-resolution, Multimed Tools Appl. (2018) 1–16, doi:10.1007/s11042-018-5915-7.
[23] F. Visin, K. Kastner, K. Cho, M. Matteucci, et al., Renet: a recurrent neural network based alternative to convolutional networks. arXiv preprint arXiv:1505.00393, 2015.
[24] 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.
[25] P. Khorrami, T.L. Paine, K. Brady, C. Dagli, T.S. Huang, How deep neural networks can improve emotion recognition on video data, IEEE Conference on Image Processing (ICIP), 2016.
[26] V. Chernykh, G. Sterling, P. Prihodko, Emotion recognition from speech with recurrent neural networks, arXiv:1701.08071v1 [cs.CL], 2017.
[27] B.C. Ko, A brief review of facial emotion recognition based on visual information, Sensors 18 (2018) 401.
[28] O. Krestinskaya, A.P. James, Facial emotion recognition using min-max similarity classifier, in: Conference on Advances in Computing, Communications and Informatics (ICACCI), 2017, pp. 752–758.
[29] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C.M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, S. Narayanan, Analysis of emotion recognition using facial expressions, in: Speech and Multimodal Information, International Conference on Multimodal Interfaces, 2004, pp. 205–211.
[30] B.A. Wandell, Foundations of Vision, First ed., Sinauer Associates Inc, Sunderland, Mass, 1995.
[31] A.T. Lopes, E. d. Aguiar, A.F. De Souza, T. O.Santos, Facial expression recognition with convolutional neural networks: coping with few data and the training sample order, Pattern Recognit. 61 (2017) 610–628.