Multimedia Tools and Applications (2019) 78:22861–22882
https://doi.org/10.1007/s11042-019-7530-7

Multi-stream CNN for facial expression recognition in limited training data

Javad Abbasi Aghamaleki[1] · Vahid Ashkani Chenarlogh[2]

Received: 10 July 2018 / Revised: 3 February 2019 / Accepted: 20 March 2019 /
Published online: 25 April 2019

Abstract

Limited data is a challenging problem to train Convolutional Neural Networks. On the other hand, acquiring a database in a demanded scale is not a straightforward task. In this paper, handcrafted features along with a multi-stream structure are proposed as a solution to improve performance of limited data via CNN. Three handcrafted features using local binary pattern code extractor and Sobel edge detection operator in horizontal and vertical directions of images have been extracted to apply to the multi-stream CNN model. Our model is based on two distinct structures including three-stream and single-stream structures. The three-stream structure can be employed to improve the recognition rate in facial expression classifiers when the training data is limited. In three-stream structure, each of information channels will be added to distinct streams separately. Furthermore, the transfer learning technique employed and behaviour of VGG16 architecture trained with limited data have been studied to be compared with the proposed method. In addition, input data is expanded by means of rotation, cropping, and flipping. Next, three-stream and single-stream structures are examined while using limited and also expanded training data. We have evaluated the mentioned system in order to compare it with state of the arts for CK+ and MUG databases in both limited-data and expanded-data. The results indicate that by using limited-data, recognition accuracy will be improved through the mentioned strategy. (92.19 to 88.95 in CK+ database and 85.4 to 82.5 in MUG database). Additionally, the performance was improved in comparison with benchmark methods.

Keywords Facial expression recognition · Convolutional neural network · Limited data · Multi-stream structure

✉ Javad Abbasi Aghamaleki
J.a.aghamaleki@du.ac.ir

[1] Faculty of engineering department, Damgham University, Damghan, Iran

[2] ECE Department, Islamic Azad University Science and Research Branch, Tehran, Iran

1 Introduction

Facial Expression Recognition (FER) aims to recognize human internal emotion from one or more observations. In recognition of humans and human expressions by face, global and local approaches have a main role in biometrics. Researchers have tried to use several feature extraction approaches in 2D and 3D dimensions [10]. Ekman et al. identified six facial expressions (anger, disgust, fear, happiness, sadness and surprise) as basic emotion expressions [8]. Facial expression recognition has been employed in several applications such as image understanding, psychological studies and smarter human-computer interfaces. In recent years, several techniques have been introduced using handcrafted features proceeded by a classifier. In a study, fusions of multiple Gabor features were used to improve performance of facial expression recognition system [22]. Facial expression recognition methods are divided into image-based method [43] and additionally video-frames-based method providing results based on the analysis of spatio-temporal sequences of images [6, 41]. Traditional machine learning methods such as Support Vector Machine (SVM) or Bayesian classifiers have been successful aiming at facial expression recognition. These solutions do not have appropriate performance in an unsupervised environment to recognize emotion expressions in captured images. In addition, classifiers tend to be tuned towards single database. Generally, this problem accentuates poor results when applied to unseen images [30].

Although, traditional feature extraction and recognition approaches were successful in different applications such as [24, 25] which have been proposed by Liu et al. to recognize activities from sensor data, but recently computational power has increased and deep learning algorithms have been replaced by traditional artificial intelligent approaches. Unlike traditional machine learning approaches, using handcrafted features and neural networks enable us to extract undefined features from the training database. One of the most notable deep learning approaches is Convolutional Neural Networks (CNNs), which have been widely used for facial expression recognition [26, 31]. CNNs consist of three main neural layers including convolutional layer, pooling layer and fully connected layer. Generally, convolutional layers are alternated with pooling layers followed by fully connected layers form a CNN. A CNN employs various kernels to convolve to the whole images in the convolutional layers to generate various feature maps. Pooling layer follows the convolutional layer to reduce dimensions of feature maps and network parameters. To this end, max pooling is commonly used as pooling layer. Fully connected layers follow the last pooling layer in the network to convert 2D feature maps into 1D feature vectors [11]. Accurate and efficient CNN models have been used in many applications such as face detection [39], video classification [16], moving object recognition [14], action detection [45], or facial expression recognition [17, 27].

2 Related works

Automated FERs usually consist of three main steps including registration, feature extraction, and classification. Extracted features in feature extraction step can be Gabor filter [44], Local Binary Patterns (LBP) [1], or Histogram of Oriented Gradients (HOG) [29]. However, developing an accurate facial expression recognition system is a challenging problem due to limitations such as pose variations, noise, and occlusion. Mohammad Sajjad et al. proposed a hybrid approach by fusing the HOG descriptor with the uniform local ternary pattern to facial sentiment analysis [35]. Ali Mollahosseini et al. [32] suggested deep neural network

architecture to automated FER including two convolutional layers followed by max pooling and four inception layers. This strategy leads to generating less redundant features among feature maps of the same layer that yields a more compact representation of an image [40]. Seyed Mehdi Lajevardi and Magaret Lech attempted to form feature arrays concatenated to six data tensors using logarithmic Gabor filters. Then the tensor was used to train one of the six parallel neural networks to FER [20]. Gil Levi and Tal Hassner overcame less labeled data problem for training emotion classification systems by removing confounding factors from the input images with an emphasis on image illumination variations. To this end, images are being converted to LBP codes to apply to the CNNs [21]. Yang et al. employed multi-modal facial images, such as facial gray scale, depth, and LBP to recognize facial expressions [42].

In [17], authors tried to show advantages of using CNN architectures to baseline methods. Temporal information can provide valuable features for FER. Heechul Jung et al. [15] used deep networks based on two different models. Firstly, deep network extracts temporal features from temporal facial land mark points. Then, image sequences are applied to the second-deep network. These two models have been combined in order to boost the performance of the FER. Dennis et al. [12] proposed an architecture based on the multi-channel CNN for FER. The features which have been extracted using two channels, replaced by a single channel. Information from both channels merged into a fully connected layer and are used for classification.

Shehab khan et al. proposed a deep model with a four-stream hybrid network for group-level emotion recognition [36]. They put forward a four-stream network to hybrid face-location aware global stream, the multi-scale face stream, a global blurred stream, and a global stream to infer the group-level emotion from individual faces employing a multi-scale face network. Ashkani et al. [4] recommended a multi-stream 3D-CNN for human action recognition which is trained by limited data. They further suggested several CNN structure containing single-stream, two-streams, and four-streams. Next, they applied optical flow and gradients features in horizontal and vertical direction to the multi-stream 3D-CNN structures. They showed that in limited data conditions for 3D CNN training, the use of handcrafted features in a four-stream architecture improves the recognition performance.

To the best of our knowledge, training a deep neural network in limited data condition and providing a special CNN structure in this condition have not been explored by the community. Hence, they could be considered as the novelties of this work. In this paper, our contributions can be summarized as follows:

- Providing input images as network input with LBP code extractor and Sobel edge detector.
- Providing three-streams and single-stream CNN architectures to process the input.
- Studying the behaviour and the performance of VGG16 structure to be compared with proposed method in limited data condition.
- Improving the performance in comparison to the state of the art methods with experiments on CK+ and MUG databases.

The rest of the paper is organized as follows: We propose motivation, handcrafted features and CNN architectures in section 3. Experiments on facial expression databases, making use of transfer learning strategy in limited data case and error analysis are proposed in section 4. Finally, we conclude the paper in section 5.

3 Proposed algorithm

3.1 Motivation

Large number of samples are required for convolutional neural networks to learn well, while in most cases, there is no enough data available. On the other hand, some methods such as transfer learning which is proposed to improve the accuracy in limited data conditions can result in disadvantages in limited training data. We have collected a new database from Iranian community to wild facial expression recognition. Indeed, the number of samples is limited in practice for this application. In better words, the procedure of collecting the wild database annotation is expensive and time consuming. On the other hand, CNNs, to adapt itself with a training samples, require a large number of samples to abstract the efficient features. To overcome this challenge, we propose to aid CNN by applying handcrafted features along with multi-stream CNN structure. In our multi-stream CNN, employing several streams leads the training parameters in each CNN to reduce and this is the main philosophy which supports the proposed strategy.

However, the idea of using pre-extracted features as CNNs input data cannot always improve accuracy. As a result, we have shown that when is a good time to use handcrafted features as CNNs input? And also, when the usage of raw data is preferred as CNNs input data? Our observations express that due to lack of access to adequate data for CNNs training, using handcrafted features along with multi-stream CNNs will be more preferable and vice-versa, in case of having enough data for CNNs training, using raw data will lead to better results compared to handcrafted features.

3.2 Handcrafted features

Many algorithms in this field have been introduced. In digital image, the so-called edge is a collection of pixels where local area brightness of the image changes significantly. Therefore, edges are vital parts of images which have remarkable information. Recently, many studies have tried to propose an improved edge detection algorithms which applied Sobel edge detector to extract edges on the image [9]. Additionally, LBP is used in limited data cases to apply to CNNs [21]. The original LBP operator has been proposed by Ojala et al. [34]. This scheme captures local image micro-textures. They are produced by applying threshold on the intensity values of the pixels in small region using central pixel's intensity as the threshold and considering the results as a binary number. In this case, lower values than the threshold become 0 and higher than the threshold become 1. Another LBP which has been used in proposed method, will be extended to the original operator so called uniform LBP [1]. Following parameters have been used for uniform LBP operator: $LBP^{u2}_{P=8,R=1}$ which $u2$ stands for using only uniform patterns, P is the number of circularly symmetric neighbor set point on a circle with radius R.

By Considering the limited training data conditions, dividing raw data into multiple level of information channels leads accuracy to improve. In this case, we considered Sobel edge detection operator in horizontal and vertical directions as highlighted parts of each data and LBP code extractor for removing confounding factors from the input images as a solution to apply to CNN model for facial expression recognition. Figure 1 shows handcrafted features as CNNs inputs.

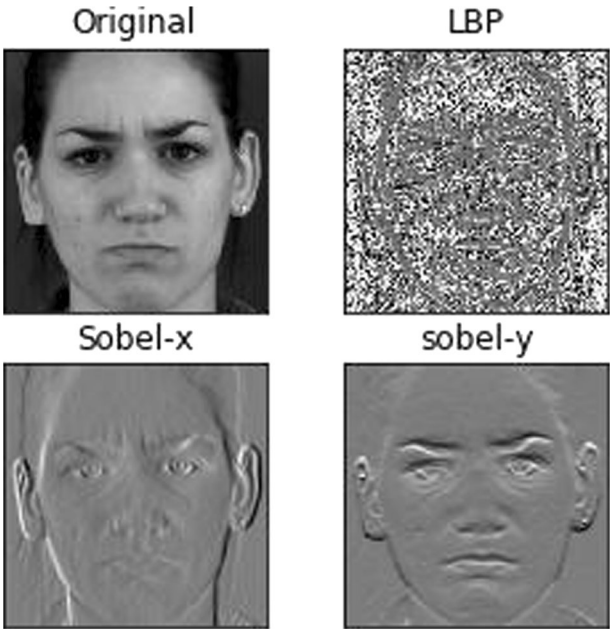


Fig. 1 Three handcrafted features consist of LBP, Sobel edge detection operator in horizontal and vertical directions extracted from an image, to apply to the CNNs

3.3 Pre-processing

We have expanded the data to compare different data states. Figure 2 shows five images produced by transformations. At first, Haar cascade face detection was used to face detection. Then cropped all images along with equalizing all images to 210×210 pixels and then converted to gray scale. To make the model more robust in training step, the data augmentation strategy has been employed. Rotation with two angles (15, 30), horizontal flip and cropping were used to extend each image to five images. Finally, normalization by standard deviation was performed. Normalizing formula is determined as below:

$$\overline{x_i} = \frac{x_i - mean(x)}{s} \tag{1}$$

Where x_i and $\overline{x_i}$ are the element and normalized element, respectively and s is the standard deviation of x.

3.4 Three-stream CNN architecture and learning strategy

CNN architecture is constructed by providing multiple layers of convolution and pooling in an alternating fashion. In this article, we have taken advantages of three-stream CNNs for all the experiments. Illustrated handcrafted features in Fig. 3 are used as input. Proposed structure includes three-streams. LBP, Sobel-x, and Sobel-y added to each stream, separately. Sobel-x and Sobel-y are calculated by computing Sobel edge detection along the horizontal and vertical

Fig. 2 Extended each image to five images using transformations (cropping, rotation and flipping)

directions. It is worth noting, following optimal hyper parameters are selected during trial and error in many experiments. Generally, we have divided our network structure into two main parts including before and after concatenation. CNN structure is composed of three convolutional layers with 16, 32 and 64 filters for all the streams before concatenation. In each of these layers, kernel size is 3×3 with strides 1×1 along with zero-padding and biases used and followed by Rectified Linear Units (ReLU) activation functions which are faster to train than standard sigmoid units [7]. Max pooling layers with 2×2 kernel size with strides 2×2 and biases are replaced after each ReLU activation function in all three layers, which leads to the same number of feature maps with a reduced spatial resolution. After three layers including convolutional, ReLU and max-pooling in each of three streams, we concatenated all three streams. After concatenation step, two convolution layers with 128 and 256 filters along with 3×3 kernel sizes with strides 1×1, zero padding and biases are used. Each of these

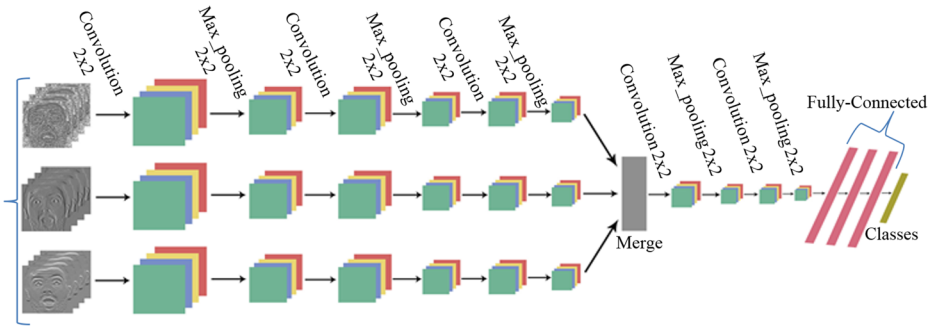


Fig. 3 Proposed three stream CNN model

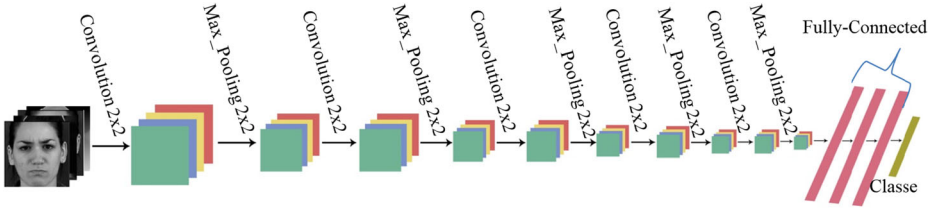


Fig. 4 Single stream CNNs proposed for facial expression recognition

convolutional layers are followed by ReLU activation functions and max pooling layers with parameters like those before concatenation. Last max pooling layer is followed by three fully-connected layers with 1024, 1024 and 512 hidden units, respectively. Each fully-connected layer is tracked with Drop-out algorithm with probability of 0.5 to reduce overfitting [18]. Finally, a softmax layer has been used for classification. The softmax layer contains 6 outputs corresponding to the number of training classes. Ultimately, in order to train our model, Stochastic Gradient Descent (SGD) has been used with constant learning rate in 0.01 to optimize our model. It should be mentioned that we considered batch size of 14 in our experiments.

In the mentioned three-stream CNN structure, the features are extracted from different layers and a feed-forward phase of the network is performed. Given a kind of input, 2D convolution is executed at convolutional layers to extract feature. 2D convolution is formulated as follows:

$$v_{ij}^{xy} = \tanh\left(b_{ij}\sum_{m}\sum_{p=0}^{P_i-1}\sum_{q=0}^{Q_i-1} w_{ijm}^{pq} v_{(i-1)m}^{(x+p)(y+q)}\right) \tag{2}$$

Considering the *j*th feature map in the *i*th layer, the value of a unit at position (x, y), which results v_{ij}^{xy}. in this formula *tanh(.)* is the hyperbolic tangent, b_{ij} represents the bias, m indexes over the set of feature maps in the *(i-1)*th layer connected to the current feature map. Meanwhile w_{ijk}^{pq} indicates kernel value at the position (p,q) connected to the kth feature map, P_i and Q_i are used as height and width of the indicated kernel, respectively.

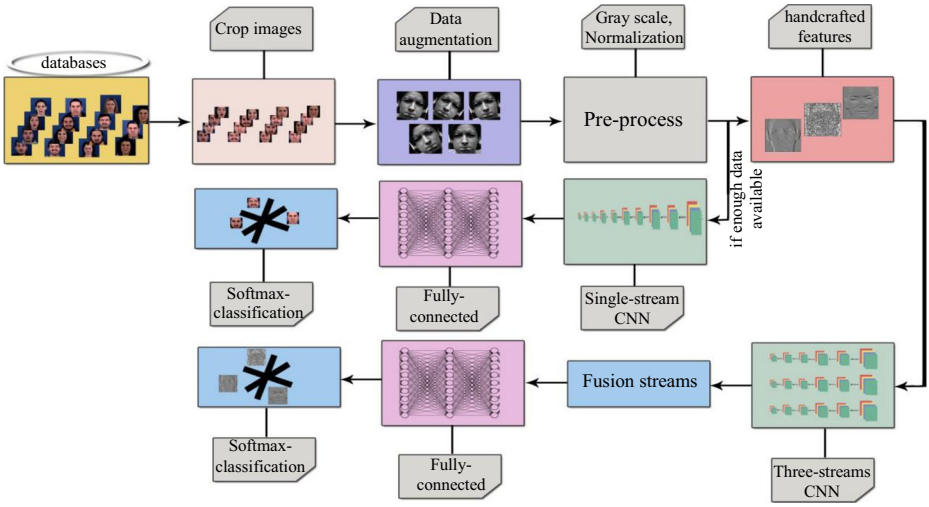


Fig. 5 The overview of proposed strategy for facial expression recognition

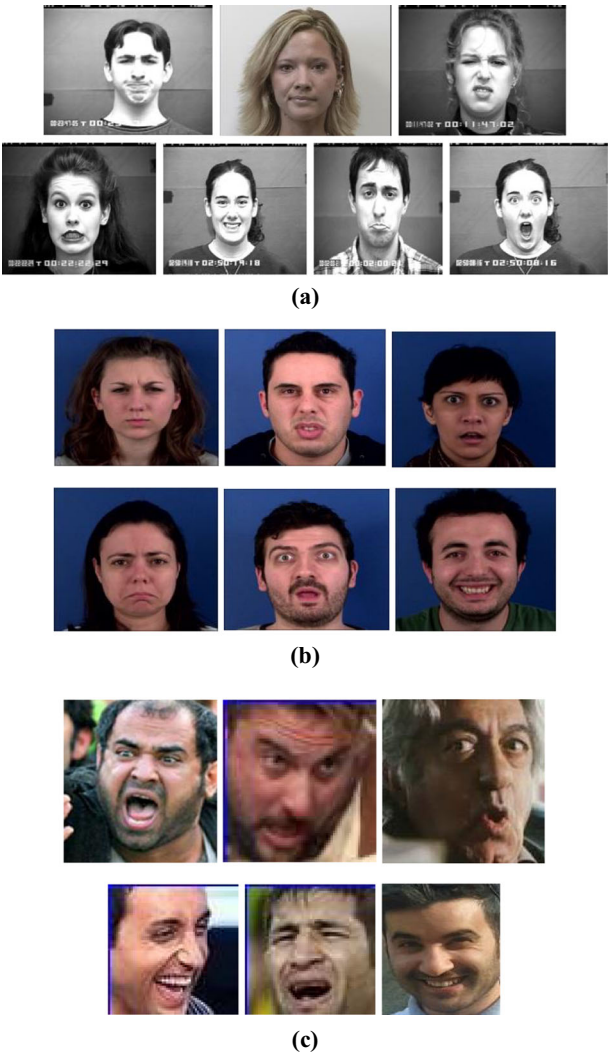


Fig. 6 samples of databases. **a** CK+, **b** MUG and **c** IWFER

3.5 Single-stream CNN architecture

Single-stream CNN is proposed to compare raw images with handcrafted features as CNN inputs. Single-stream CNN involves five convolutional layers with 16, 32, 64, 128, and 256

Table 1 Train and test samples count available in CK+, MUG and IWFER databases

	CK+		MUG		IWFER	
	Train	Test	Train	Test	Train	Test
Main data	277	31	252	28	233	26
Augmented data	1386	154	1260	140	1165	130

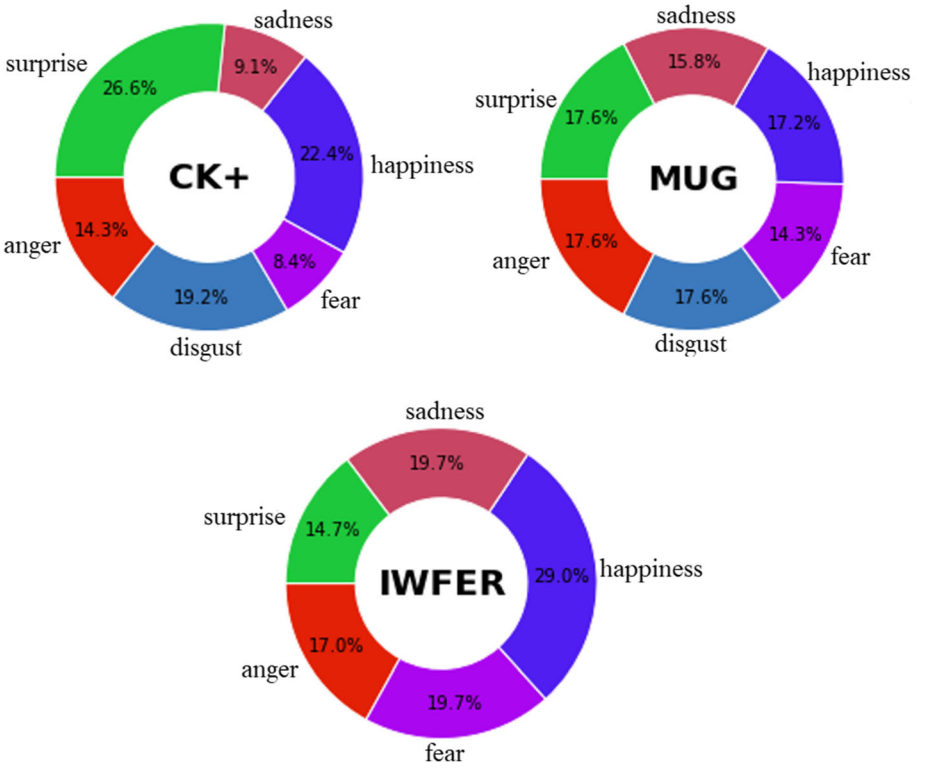


Fig. 7 Classes distribution in CK+, MUG and IWFER databases

filters, respectively. Kernel size, strides, biases, zero padding and other parameters are similar to the mentioned three-stream CNN. After each convolutional layer, ReLU activation function and max-pooling layer are used with 2×2 kernel size and 2×2 strides along with the bias value. Like three-stream structure, last max-pooling layer is followed by three fully-connected

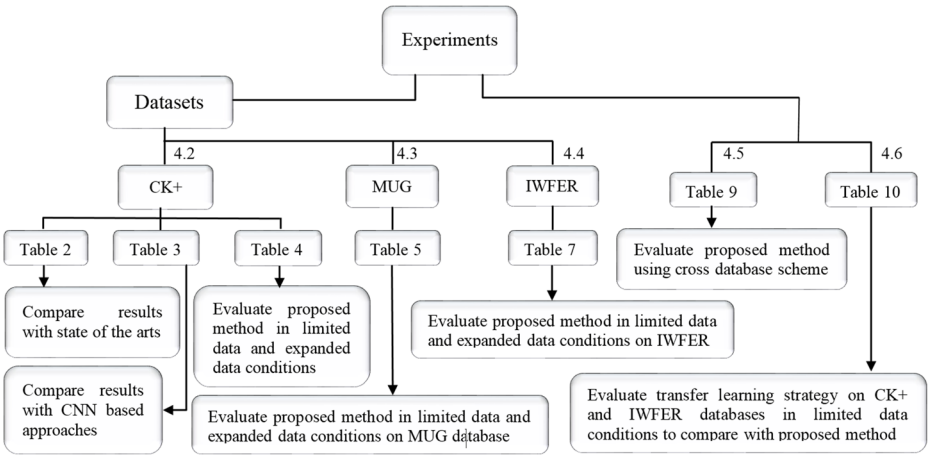


Fig. 8 A block diagram to identify the various experiments and purposed experiments

Table 2 Implementation results of 10-fold cross validation experiment

Methods	CK+ database
Happy et al. [13]	94.14
Alam et al. [3]	99.11
Burkert et al. [5]	99.6
Sun et al. [38]	91.35
Liu et al. [23]	97.70
Kumar et al. [19]	97.91
Proposed (single stream)	**99.61**
Proposed (three stream)	98.18

The bold vaules are used to emphasize the proposed algorithm's results

layers with 1024, 1024, and 512 hidden units, respectively. Fully-connected layers are followed by drop-out algorithms with probability of 0.5 and finally, a softmax layer is used for classification. Figure 4 shows the single stream CNN structure. Figure 5 shows the scheme of proposed algorithm for facial expression recognition.

4 Experiments

This section contains the details of gathered databases, the experiments on proposed method, experiments on VGG16 model, and error analysis of the proposed approach. Due to the fact that these stated approaches are experimental, the results of several tests suggest that using multi-stream structure results in reducing parameters of each CNN. Therefore, it can cause that CNN to have better understanding of functional features. The experiments are discussed further in this section.

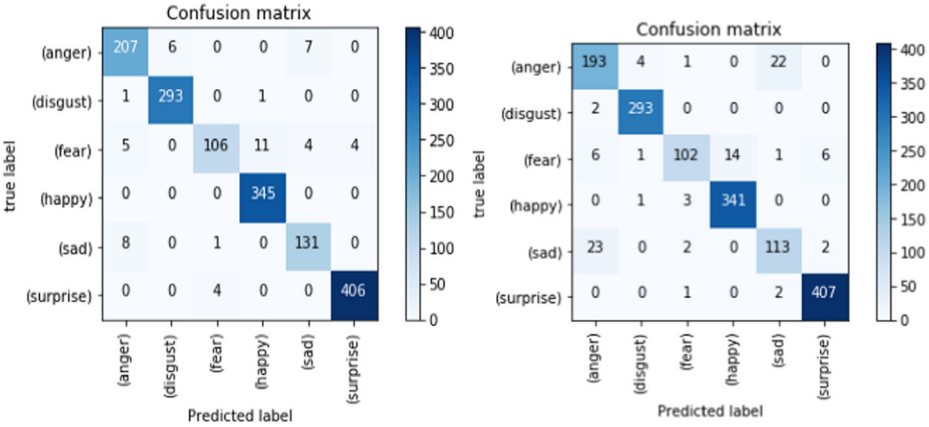
4.1 Databases

Three facial expression databases are used in our experiments. The extended Cohn-Kanade database (CK+) [28] contains 327 image sequences. Each of which is assigned to one of 7 expressions: anger, contempt, disgust, fear, happiness, sadness and surprise. We used six classes without contempt expression. Last frame of each sequence is considered as expressive frame to form our database. Totally, we selected 308 images for 6 classes. In addition, as mentioned above, we extended each image into five images using data augmentation strategy with 210×210 pixels to apply to the CNNs. Next, we split the images into 10 subject independent subsets to evaluate the effectiveness of our proposed methods and then reported the result using 10-fold cross validation strategy.

Table 3 Implementation results of CNN based approaches

Methods	CK+ database
Khorrami et al. [17]	98.3
Lopes et al. [26]	97.81
Lopes et al. [27]	96.76
Meng et al. [31]	95.37
Proposed (single stream)	**99.61**
Proposed (three stream)	**98.18**

The bold vaules are used to emphasize the proposed algorithm's results



(a) (b)

Fig. 9 The confusion matrix of the average 10-fold cross validation for six classes on CK+ database. **a** single-stream, **b** three-streams structure

Multimedia Understanding Group (MUG) database [2] consists of image sequences of 86 subjects performing facial expressions. Each image has resolution of 896 × 896 pixels. In the database, 35 women and 51 men with 20 and 35 years of age participated. In the database the subjects performed the six basic expressions including anger, disgust, fear, happiness, sadness, and surprise. First frame of each sequence was considered as a neutral frame and the best frame of each sequence as expressive frame to form our MUG database. Generally, we added data augmentation strategy and stabilized it with 210 × 210 pixels to apply to the CNN.

The third used database in this work is gathered from web sources. This database is collected from Iranian society. Due to not possessing enough disgust expression samples, we formed our database in five classes (anger, fear, happiness, sadness and surprise) and called it Iranian Wild Facial Expression Recognition (WIFER). Due to noise reduction in collecting process, samples in each class are chosen by hand. Totally, there are 259 images for five classes. The resolution of the samples is 210 × 210 pixels, and they have been compressed using JPEG format. To crop faces in this database, Haar cascade is used for face detection. Images are collected from-side and front viewpoints. The samples of this database are realistic state of the expressions in Iranian community. Different light conditions and different face viewpoints are taken into account as challenges of this database. Figure 6 shows some samples of extended CK+, MUG and IWFER databases. In addition, number of training and test

(a) (b) (c) (d)

Fig. 10 Samples of similar images of fear and happiness expressions in CK+ database. **a** and **b** fear, **c** and **d** happiness moods

Table 4 Recognition accuracy of expanded data and main data results of CK+ database in both single stream and three streams structures

Model	Expanded data	Main data
Single-stream	99.61	88.95
Three-stream	98.18	92.19

samples and distribution of each class in CK+, MUG and IWFER databases, are given in Table 1 and Fig. 7, respectively. Figure 8 is provided to identify the experiments and the purposed experiments.

4.2 Experiment on CK+ database

Two kinds of input data are used in our experiments. Firstly, raw images as single-stream CNN inputs and then three handcrafted features of LBP, Sobel-x and Sobel-y have been extracted from each image to be applied to three-stream CNN. In order to express capability of the proposed approach, experiments are divided into two main parts. At first, 1540 images made-up by data augmentation strategy are used and in the second step just 308 main images of CK+ database are utilized as input in the network. Then, 1540 raw images are applied to single-stream CNN, and three proposed information channels including LBP, Sobel-x and Sobel-y are applied as input to three-stream CNN. Results illustrate that utilizing raw frames gives better performance when enough training data is available. This point indicates that the network is capable of learning the details when enough data is available. In other words, adequate number of correct samples enable the network to learn the desired visual concepts. On the other hand, 1540 samples are not a big data for deep learning community. CK+ database is a very clean database; therefore, it is a special case and has significant recognition accuracy. The results of these two experiments are compared with some state of the art methods in Table 2. Additionally, comparison results with CNN based papers are shown in Table 3.

A combined confusion matrix for single-stream and three stream structure is shown in Fig. 9. Matrix cells are computed from 100 epochs in each fold and then the results of all folds are added together. The outstanding results indicate that happiness, disgust, and surprise moods hold the highest amount of recognition accuracy. However, anger, fear, and sadness moods introduce the lowest amount of accuracy in both single-stream and three-stream structures. As can be seen, some samples of fear have been wrongly predicted as happiness, because some samples of these two classes are similar in CK+ database. Figure 10 shows some similar samples in CK+ database.

In second experiment, only 308 main images of CK+ database have been used without any augmentation. This experiment is also done along with single-stream and three-stream structures. Table 4 shows the results of the experiments. The results express important information including in the case of lack of enough training data. If there is no enough training data available, handcrafted features along with multi-stream CNN structure are preferable to improve the desirable performance. Regarding the vast usage of CNN, the system convergence

Table 5 Recognition accuracy of expanded data and main data of MUG database in both single stream and three streams structures

Model	Expanded data	Main data
Single-stream	91.6	82. 5
Three-stream	89.7	85.4

Table 6 Obtained average recognition accuracy for each classes after 10-fold on expanded MUG dataset

Classes	Single-stream			Three-streams		
	precision	recall	f1-score	precision	recall	f1-score
Anger	77.1	88.2	82.3	84.5	90.3	87.3
disgust	94.5	96.1	94.8	96.8	95.4	95.9
Fear	79.6	69.8	73.4	78.2	66.8	71.2
Happy	95	98.3	96.7	97.6	96.9	97.1
Sad	83.9	71.5	76.8	89	84.1	86.1
Surprise	82.9	85.1	83.4	77.5	87.1	81.3
Average %	85.5	84.83	84.56	87.26	86.76	86.48

is deeply linked to CNN convergence. Therefore, since CNN learning has been tested in several experiments and the output results have been valid, we can finally conclude that convergence does exist in this system as well.

4.3 Experiments on MUG database

For further experiments, MUG database is used. Generally, we first do the experiment with 1400 images made-up with data augmentation strategy during single-stream and three-stream structures. In addition, experiments are performed on MUG database without augmented data to show the effects of our proposed method in limited data conditions. In this case, 280 images are used to apply to both proposed CNN structures. Table 5 shows the experiment results on MUG database. The results represent that providing adequate data for CNN training, and employing raw images are much better than handcrafted features, but in limited data conditions using handcrafted features along with multi-stream structure will be much better than raw images along with a single-stream.

Being real-time and increasing the recognition accuracy while accessing limited-training data can be mentioned as some of the advantages of this system. As CNN demands a large amount of training data, the usage of the mentioned strategy will lead recognition accuracy to increase in these functions, whereas in many CNN functions, just a limited amount of training data might exist. The disadvantage of this strategy is that, after concatenating streams, the general parameters of CNN network will rise. This issue can both reduce the speed of learning in network and might increase the time of learning, too. It has been carried out in a way that by the means of GPU-1080, the amount of time spent on expanded MUG database after 10-folds and 100 epochs in each fold was equal to 110 min and 80 min for three-stream and single-stream structure, respectively.

The average precision, recall, and f1-score accuracy of expression classes after 160 epochs in 10 folds with single-stream and three-stream structure are tabulated in Table 6. Happiness and disgust have the highest performance in both structures, while fear has the worst performance.

Table 7 Recognition accuracy of expanded data and main data of IWFER database in both single stream and three streams structures

Model	Expanded data	Main data
Single-stream	86.56	57.4
Three-stream	82.7	59.2

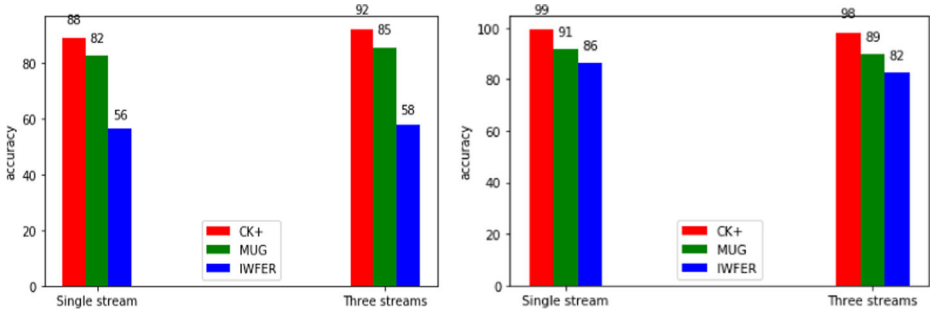


Fig. 11 Comparison results of CK+, MUG and IWFER databases in single stream and three stream structures. **a** main data, **b** expanded data

4.4 Experiments on IWFER database

IWFER database is utilized to show the performance of the proposed approach on real world images. Experiments using three-stream and single-stream structures with same hyper parameters as previous experiments, have been designed. Additionally, expanded data and main data are used as databases. In expanded data, 1295 images are made up with data augmentation strategy in five classes by scaling each image to 210 × 210 pixels. However, in main data, 259 images were applied in five classes. Table 7 shows results of these databases in both three-stream and single stream structures. Experiments show results similar to previous databases in limited data conditions, using handcrafted features along with multi-stream structure, will be much better than raw images in a single-stream. Recognition accuracy is low in comparison with CK+ and MUG databases due to challenges in the database like different light conditions and anomaly in face viewpoints. Fig. 11 shows comparison results of databases. The average precision, recall, and f1-score accuracy of expression classes after 160 epochs in 10 folds with single-stream and three-stream structure are tabulated in Table 8.

4.5 Experiments on cross databases

This section provides an evaluation of proposed approach across CK+ and IWFER databases. In order to homogeniz the number of classes, disgust expression from CK+ database is ruled out. We used one to serve as training and one as test data employing single-stream and three-stream

Table 8 Obtained average recognition accuracy for each class after 10-fold on expanded IWFER dataset

Classes	Single-stream			Three-streams		
	precision	recall	f1-score	precision	recall	f1-score
Anger	84.2	77.6	79.2	78.5	83	80
Fear	72	70.9	71.4	75.2	66	69.5
Happy	88.1	92.9	90.4	89.5	92.6	91
Sad	80.1	79.8	79.9	81.2	84.5	82.2
Surprise	72.6	73.7	72.9	69.3	65.5	67
Average %	79.4	78.98	78.76	78.74	78.32	77.94

Table 9 Evaluation cross databases with different schemes

Model	Train Database	Test Database	Accuracy
Single sream	Main CK+	Main IWFER	36.68
Single stream	Main IWFER	Main CK+	47.39
Single stream	Expanded CK+	Main IWFER	38.22
Single stream	Expanded IWER	Main CK+	50.2
Three stream	Expanded CK+	Main IWFER	35.14
Three stream	Expanded IWFER	Main CK+	53.01
Three stream	Main CK+	Main IWFER	32.82
Three stream	Main IWFER	Main CK+	46.18

structures. Main and expanded data are used for both training and testing. The idea and core assumption behind this test is that self-classification evaluation represents the highest accuracy value. A second motivation is to find appropriate data for training and test steps. Third motivation is to represent the single-stream and three-stream structure performances in cross databases situation. Table 9 presents the results of cross validatation. As can be seen, accuracy is strickingly worse as compared to self-classification evalutions. Another remarkable result is to use IWFER database as training which leads to a better accuracy on CK+ database. This suggests that applying clean data in training step cannot be generalized to unclean samples in test step.

4.6 Experiments on limited data using transfer learning strategy

Transfer learning is a method assumed to increase the accuracy of recognition rate in limited data conditions. Ng et al. have drawn on this approach for emotion recognition on small databases [33]. Proposed transfer learning models were pre-trained on generic ImageNet database. They performed supervised fine-tuning on the network in a two-stage process. Firstly, on the database relevant to facial expressions followed by the contest's database. They illustrated that direct fine-tuning in the transfer learning models for facial expression recognition have led to worse accuracy. This hints at the difficulty of fine tuning deep neural networks with small databases along with irrelevant initialized weights, and also the importance of using auxiliary data in these cases. To compare the transfer learning strategy with the proposed method for training CNN in limited data conditions, the VGG16 [37] model is used. Different

Table 10 Transfer learning using VGG16 on CK+ and IWFER databases

Batch size	Epoch	Optimizer	Units in full connected layers	Dropout	Accuracy	database
32	70	–	–	–	52.6	IWFER
32	130	–	–	–	67.17	CK+
32	130	Adadelta	68, 68	–	54.68	CK+
32	140	Adadelta	128, 128	–	59.68	CK+
32	140	SGD	128, 128	–	Overfit	CK+
32	140	RMSprob	128, 128	–	Overfit	CK+
32	140	SGD	128, 128	0.5	58	CK+
32	150	Adadelta	256, 256	–	59.68	CK+
32	150	Adadelta	2048, 2048	0.5	69.35	CK+
32	150	Adadelta	2048, 2048	0.5, 0.5	77.42	CK+

experiment results on CK+ and IWFER databases using this model are shown in Table 10. Following results can be concluded from experiments:

- Given that the transfer learning models trained on ImageNet database, this strategy cannot provide high accuracy when training data is irrelevant to ImageNet database. In this case, networks need to retrain the model with a large database similar to data source or require multiple times fine-tuning depending on amount of data.
- Training data should have two options. First of all, the distribution of the training data, used by pre-trained model, should be similar to the data that you are going to face during test time. Second, the number of training data for transfer learning should be in a way that does not overfit the model. If there is a limited number of training data, and typical pre-trained model like VGG16 holding millions of parameters, we will be aware of overfitting.
- By taking the facts of Table 10 into account, experiments on IWFER database just had acceptable performance by releasing classification layer. Otherwise, releasing more layers leads to overfitting. It shows that fine tuning is a challenging problem with unclear data.

Many experiments were carried out on the main CK+ and IWFER database using the VGG16 model and Table 10 presents some of them. In this case, experiments have been done in multiple phases. First, we just released classification layer. In this case, experiment on main CK+ database shows accuracy at 67.17. In the second phase, two fully connected layers released and fine tunned the model with many experiments which yield more accuracy at 77.42. Additionally, by releasing the convolutional layers plus more layers, the network will be over-fit due to the data deficit. For robust analysis, similar experiments using VGG16 model have been performed on IWFER database. In this case, proposed approach gives higher accuracy than transfer learning strategy. Figur 12 shows comparison results between VGG16 model and proposed approach on main CK+ and IWFER databases.

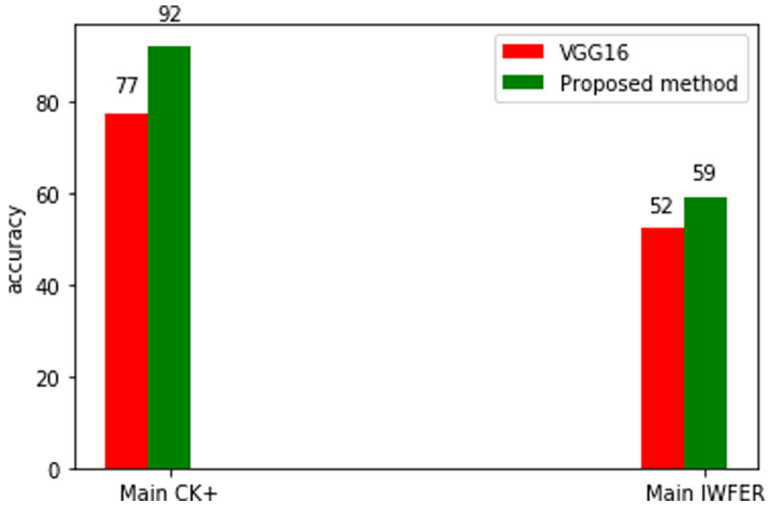


Fig. 12 Comparison results of VGG16 and proposed approach applied on main CK+ and IWFER database

Table 11 Choosing appropriate parameters using various experiments

Model	Epoch	Convolutional stride	Dense	Convolutional filter	Convolutional kernel	Batch	Test accuracy
Three-stream	100	1 × 1	[1024,1024,512]	[16,32,64,128,256]	5 × 5	32	95/79
Three-stream	100	1 × 1	[1024,1024,512]	[8,16,32,64,128]	3 × 3	32	96.26
Three-stream	200	1 × 1	[512,512,256]	[16,32,64,128,256]	3 × 3	32	96.16
Three-stream	100	2 × 2	[1024,1024,512]	[16,32,64,128,256]	3 × 3	32	86.81
Three-stream	100	1 × 1	[2048,2048,1024]	[16,32,64,128,256]	3 × 3	32	95.2
Three-stream	100	1 × 1	[1024,1024,512]	[16,32,64,128,256]	3 × 3	14	98.18
Single-stream	200	1 × 1	[1024,1024,512]	[16,32,64,128,256]	3 × 3	10	99.4
Single-stream	200	1 × 1	[1024,1024,512]	[8,16,32,64,128]	3 × 3	14	98.5
Single-stream	350	1 × 1	[512,512,256]	[16,32,64,128,256]	3 × 3	14	98.9
Single-stream	200	1 × 1	[1024,1024,512]	[16,32,64,128,256]	7 × 7	14	99.48
Single-stream	210	1 × 1	[1024,1024,512]	[16,32,64,128,256]	3 × 3	14	99.61

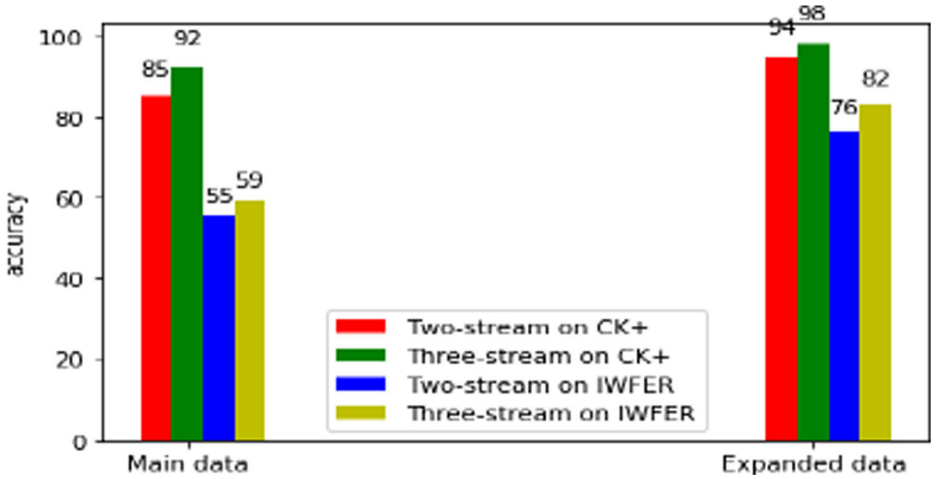


Fig. 13 Comparison results between two-stream and three-stream structures tested on CK+ and IWFER databases

4.7 Error analysis

4.7.1 Best hyper parameters selection

To delve more into the details of selecting the appropriate hyper parameters, various experiments have been performed. Some good performances are listed in the Table 11. Results of experiments show that:

- The most suitable convolutional kernel size is 3×3 with strides1x1.
- Optimal required number of neurons in fully-connected layer is obtained by the test.

4.7.2 Two-stream versus three-stream structure

In this section, two-stream CNN is constructed. Therefore, the Sobel edge detector in horizontal and vertical directions concatenated to be applied to one stream along with applying LBP to another stream. It is noteworthy to mention that the hyper parameters which have been used are similar to three-stream CNN structure. The results of CK+ and IWFER databases show that splitting raw data to more different information channels and then applying these to separate streams lead to higher accuracy. This can be described by the fact that in limited data conditions, improving the performance requires both features as input data and then it should be applied to separate streams. Figure 13 shows comparison results between two-stream and three-stream structures experimented on CK+ and IWER databases.

5 Conclusion and future work

In order to work out the challenge of using limited data in training CNNs, we have expressed applying handcrafted features in a three-stream CNN structure. At first, collection of

handcrafted features has been extracted by adding up Local Binary Pattern (LBP) code extractor and Sobel edge detection operator in horizontal and vertical directions from each image and then they will be added to a three-stream CNN, separately. This proposed strategy will lead to improvement in recognition accuracy during utilizing limited training data. We have concluded that during using limited data for CNN training, the network cannot learn the most important features from limited data. As a result, using pre-extracted features as the most important parts of each image and then executing that on streams separately can possibly help the network to take and learn the most striking parts from each image which finally results in increasing accuracy. Furthermore, by expanded data usage, we declared that providing enough data for CNN training, employing raw images is much better than handcrafted features. In addition, comparing the proposed approach with transfer learning strategy showed that in irrelevant data usage to ImageNet database, proposed approach is much more preferable than transfer learning strategy. We carried out our experiments on CK+, MUG and our own providing data from Iranian community databases and then observed that results coming from databases have actually confirmed our claims. Results show that our structure outperforms the state of the art facial expression recognition systems, tested on CK+ database.

The approach of using handcrafted features with multi-stream structure during the usage of limited data for CNN training caused the better recognition. As a result, in the surveys that their input data is video sequences with limited input data, there is no way to extend video data. Therefore, using handcrafted features with multi-stream structure could be a better approach to improve performance.

References

1. Ahonen T, Hadid A, Pietikainen M (2004) Face recognition with Local Binary Patterns. Proceeding of european conferenceon computer vision, pp 469–481
2. Aifanti N, Papachriston C, Delopoulos A (2010) The MUG facial expression database. proceedings o 11th International workshop on image analysis for multimedia interactive service (WIAMIS), Desenzano, Italy, pp 1–4
3. Alam M, Vidyaratne LS, Iftekharuddin KM (2018) Sparse simultaneous recurrent deep learning for robust facial expression recognition. IEEE Transaction on neural networks and learning systems pp(99): 1–12
4. Ashkani Chenarlogh V, Razzazi F (2018) A multi-stream 3D CNN structure for human action recognition trained by limited data. IET Comput Vis. https://doi.org/10.1049/iet-cvi.2018.5088
5. Burkert P, Trier F, Afzal M.Z et al (2016) DeXpression: Deep convolutional neural network for expression recognition. arXiv: 1509.05371v2, pp 1–8
6. Byeon YH, Kwak KC (2014) Facial expression recognition using 3D convolutional neural network. Int J Adv Comput Sci Appl 5(12):107–112
7. Dahl GE, Sainath TN, Hinton GE (2013) Improving deep neural networks for LVCSR using rectified linear units and dropout. Proceedings of 2013 IEEE International conference on acoustics, speech and signal processing, Vancouver, Canada, pp 8609–8613
8. Ekman P, Friesen WV (1971) Constants across cultures in the face and emotion. J Pers Soc Psychol 17(2): 124–129
9. Gao W, Yang L, Zhang X et al (2010) An improved Sobel edge detection. proceedings of 2010 3rd IEEE International conference on computer science and information technology, Chengdu, China, pp 67–71
10. Ghasemzadeh A, Demirel H (2018) 3D discrete wavelet transform-based feature extraction for hyper spectral face recognition. IET Biometrics 7(1):49–55
11. Guo Y, Liu Y, Oerlemans A et al (2016) Deep learning for visual understanding: a review. Neurocomputing 187:27–48

12. Hamester D, Barros P, Wermter S (2015) Face expression recognition with a 2-channel convolutional neural network. Proceedings of 2015 International joint conference on neural networks, Killarney, Ireland, pp 1787–1794
13. Happy SL, Routray A (2015) Automatic facial expression recognition using features of salient facial patches. IEEE Trans Affect Comput 6(1):1–12
14. He T, Mao H, Yi Z (2017) Moving object recognition using multi-view three dimensional convolutional neural networks. Neural Comput & Applic 28(12):3827–3835
15. Jung H, Lee S, Park S et al (2015) Deep temporal appearance-geometry network for facial expression recognition http://arxiv.org/abs/1503.01532.2, 3, pp 1–9
16. Karpathy A, Toderici G, Shetty S et al (2014) Large scale video classification with convolutional neural networks. Procceding of International Vision and Pattern recognition (CVPR), Columbus, Oh, USA, pp 1725–1732
17. Khorrami P, Paine TL, Huang TS (2015) Do deep neural networks learn facial action units when doing expression recognition?. Proceedings of the IEEE International conference on computer vision workshops, Santiago, Chile, pp 19–27
18. Krizhevsky A, Sunskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. Adv Neural Inf Proces Syst 25:1106–1114
19. Kumar S, Bhuyan MK, Chakraborty BK (2016) Extraction of informative regions of a fact for facial expression recognition. IET Comput Vis 10(6):567–576
20. Lajevardi SM, Lech M (2008) Facial expression recognition using neural networks and log-gabor filters. proceedings of computing: Techniques and applications, DICTA '08. Digital image, Canberra, Australia, pp 77–83
21. Levi G, Hassner T (2015) Emotion recognition in the wild via convolutional neural networks and mapped binary patterns. Proceedings of the 2015 ACM on international conference on multimodal interaction, Washington, USA, pp 503–510
22. Liu WF, Wang Z (2006) Facial expression recognition based on fusion of multiple gabor features. Proceedings of 18th international conference on pattern recognition (ICPR), Hong kong, china, pp 2–5
23. Liu P, Zhou JT, Tsang IWH et al (2014) Feature disentangling machine a novel approach of feature selection and disentangling in facial expression analysis. European conference on computer vision, ECCV, pp 151–166
24. Liu Y, Nie L, Han L et al (2016) Action2Activity: Recognizing complex activities from sensor data. arXiv: 1611.01872v1, pp 1–7
25. Liu Y, Nie L, Liu L et al (2016) From action to activity: sensor-based activity recognition. Neurocomputing 181:108–115
26. Lopes AT, Aguiar ED, Oliveira-santos T (2015) A facial expression recognition system using convolutional networks. Proceedings of 28th SIBGRAPI conference on graphics, patterns and images, Salvador, Brazil, pp 273–280
27. Lopes AT, Aguiar E, De Souza AF et al (2017) Facial expression recognition with convolutional neural networks: coping with few data and the training sample order. Pattern recognition (61): pp 610–628
28. Lucey P, Cohn J.F, Kanade T et al (2010) The extended cohn-kanade database (CK+): A complete database for action unit and emotion-specified expression. Proceedings of 2010 IEEE Computer society conference on computer vision and pattern recognition workshops, San Francisco, USA, pp 94–101
29. Mavadati SM, Mahoor MH, Bartlett K et al (2013) DISFA: a spontaneous facial action intensity database. IEEE transaction on affective computing 4(2):151–160
30. Mayer C, Eggers M, Radig B (2014) Cross-database evaluation for facial expression recognition. Pattern Recognition and Image Analysis 24(1):124–132
31. Meng Z, Liu P, Cai J et al (2017) Identity-aware convolutional neural network for facial expression recognition. Proceedings of 2017 IEEE 12th international conference on automatic face & gesture recognition, Washington-DC, USA, pp 558–565
32. Mollahosseini A, Chan D, Mahoor MH (2016) Going deeper in facial expression recognition using deep neural networks. Proceedings of 2016 IEEE winter conference on applications of computer vision, Lake Placid. USA, pp 1–10
33. Ng H, Nguyen VD, Vonikakis V et al (2015) Deep learning for emotion recognition on small databases using transfer learning. Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, Washington, USA, pp 443–449
34. Ojala T, Pietikainen M, Harwood D (1996) A comparative study of texture measures with classification based on feature distributions. Pattern Recogn 29(1):51–59

35. Sajjad M, Shah A, Jan Z et al (2017) Facial appearance and texture feature-based robust facial expression recognition framework for sentiment knowledge discovery. Cluster Computing, 1–19
36. Shehab Khan A, Li Z, Cai J et al (2018) Group-level emotion recognition using deep models with a four-stream hybrid network. Proceedings of the 20th ACM international conference on multimodal interaction, Boulder, CO, USA, pp 623–629
37. Simonyan K., Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556v6, Apr. 2015, pp 1–14
38. Sun Z, Hu Zp, Chiong R et al (2018) An adaptive weighted fusion model with two subspaces for facial expression recognition. SIViP 12(5):835–843
39. Wang S, Yang B, Lei Z et al (2015) A convolution neural network combined with aggregate channel feature for face detection. IET 6TH International Conference on Wireless, Mobile and Multi-Media (ICWMMN 2012), Beijing, China, pp. 49–55
40. Xie S, Hu H (2017) Facial expression recognition with FRR-CNN. IET Electronics letter 53(4): 235–237
41. Yacoob Y, Davis LS (1990) Recognition human facial expression from long image sequence using optical flow. IEEE transaction on pattern analysis and machine intelligence 18(6):636–642
42. Yang B, Cao JM, Jiang DP et al (2017) Facial expression recognition based on dual-feature fusion and improved random forest classifier. Multimed Tools Appl: 1–23
43. Yu Z, Zhang C (2015) Image based static facial expression recognition with multiple deep network learning. Proceedings of the 2015 ACM on international conference on multimodal interaction, Washington, USA, pp 435–442
44. Zhan Y, Ye J, Niu D et al (2006) Facial expression recognition based on gabor wavelet transformation and elastic templates matching. International journal of image and graphic 6(1): 125–138
45. Zhang M, Gao C, Li Q et al (2018) Action detection based on tracklets with the two-stream CNN. Multimed Tools Appl 77(3):3303–3316

Javad Abbasi Aghamaleki received his BSc degree in electrical engineering from K. N. Toosi University of Technology, Tehran, Iran, in 2007. In 2010, he received his MSc degree in electrical engineering from Shahrood University of Technology, Shahrood, Iran. He received his PhD in electrical engineering from Shahed University, Tehran, Iran, in 2016. Currently, he is an assistant professor at Engineering Faculty, Damghan University, Damghan, Iran. His research fields are image and video forensics and machine vision.

Vahid Ashkani Chenarlogh received the BSc. Degree in electrical and electronic engineering from Qazvin Islamic Azad University, Qazvin, Iran, in 2015. In 2018, He received his MSc. Degree in electrical and telecommunication engineering from Islamic Azad University Science and Research Branch, Tehran, Iran. His current research interests include deep learning, computer vision, image and video processing, and machine learning for robotics.