Check for updates

Facial expression recognition for monitoring neurological disorders based on convolutional neural network

Gozde Yolcu [1,2] · Ismail Oztel [1,2] · Serap Kazan [1] · Cemil Oz [1] ·
Kannappan Palaniappan [2] · Teresa E. Lever [3] · Filiz Bunyak [2]

Abstract

Facial expressions are a significant part of non-verbal communication. Recognizing facial expressions of people with neurological disorders is essential because these people may have lost a significant amount of their verbal communication ability. Such an assessment requires time consuming examination involving medical personnel, which can be quite challenging and expensive. Automated facial expression recognition systems that are low-cost and non-invasive can help experts detect neurological disorders. In this study, an automated facial expression recognition system is developed using a novel deep learning approach. The architecture consists of four-stage networks. The first, second and third networks segment the facial components which are essential for facial expression recognition. Owing to the three networks, an iconize facial image is obtained. The fourth network classifies facial expressions using raw facial images and iconize facial images. This four-stage method combines holistic facial information with local part-based features to achieve more robust facial expression recognition. Preliminary experimental results achieved 94.44% accuracy for facial expression recognition on RaFD database. The proposed system produced 5% improvement than the facial expression recognition system by using raw images. This study presents a quantitative, objective and non-invasive facial expression recognition system to help in the monitoring and diagnosis of neurological disorders influencing facial expressions.

✉ Filiz Bunyak
  bunyak@missouri.edu

[1] Department of Computer Engineering, Sakarya University, 54050 Serdivan, Sakarya, Turkey

[2] Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO 65211, USA

[3] Department of Otolaryngology, University of Missouri, Columbia, MO 65211, USA

🍀 Springer

1 Introduction

Facial expressions are important for human social communication. According to Mehrabian [52], facial expressions are more effective than words in face-to-face communication. Mehrabian revealed that words contribute 7%, voice tone 38%, and body language 55% to effectively communicate a message. Also, impaired facial expressions are a common symptom of many medical conditions. Predominant examples range from childhood neurodevelopmental disorders, such as autism spectrum disorder [14], cerebral palsy [23] and Angelman syndrome [1], and to adult-onset neurological diseases, such as Parkinson's [46, 62], stroke [44], Alzheimer's disease [30], and Bell's Palsy [10]. A comprehensive list of neurological and psychiatric disorders affecting facial expressions can be found in Table 3 of [72]. Effects of neurological and psychiatric disorders on facial expressions are well-known by clinicians and scientists and facial expressions are used for assessment and severity of these disorders. However, clinical evaluation involves subjective and qualitative assessment. Clinical diagnosis and disease monitoring can be challenging for most of the neurological conditions, often requiring invasive and/or expensive medical testing. Thus, non-invasive, low-cost alternatives must be developed. In this study, an automated facial expression recognition system is proposed. If such a system can readily differentiate between a variety of facial expressions, identification of clinically relevant features that distinguish between disease conditions may be possible. These distinctive features can ultimately serve as disease-specific biomarkers to help clinically diagnosis and evaluate the therapeutic response of patients with neurological disorders.

Automated facial expression recognition studies e.g. [33, 36, 61] are usually based on six universal expressions that were defined in the early work of Ekman and Friesen [24]: happy, angry, disgust, sad, surprise and fear. Some studies e.g. [6, 22, 57, 79] have worked on five or less classes of facial expressions. Facial expression recognition studies can be mainly classified as appearance and geometric based methods [59]. While appearance-based methods extract features from texture information of the face [11, 38, 77, 80], geometric-based methods are based on features obtained from distance or shape information of the face components during expressions [29, 59]. Recent years, deep learning-based methods have achieved remarkable success rates in facial expression recognition studies [47, 51, 58]. Deep learning enables automatic learning of complex features required for computer vision [42]. While some facial expression recognition studies are based on whole face information [7, 16, 43, 54, 70], some studies use partial-based information [64].

In this study, a new deep learning approach is presented for automated facial expression recognition. It is the first step towards a non-invasive automated system for clinical diagnosis and neurological disorder monitoring. It focuses on six universal facial expressions defined by Ekman and Friesen for effective and acceptable comparison with the literature.

The proposed method includes four convolutional neural networks (CNN). Three CNNs are structured for segmentation of facial components and one CNN is structured for recognition of facial expressions. In the first CNN, the eyebrow regions are segmented, the second CNN segments eye regions and the third CNN segments mouth regions on a facial image. Thus; component-segmented (eyebrow-segmented, eye-segmented, mouth-segmented) images are obtained. After post-processing, for each face, a final iconize image is formed by combining the corresponding component-segmented images. Finally, the fourth CNN classifies facial expressions combining the final iconize images with corresponding raw facial images. The four-stage CNN system has the following advantages compared to the CNN that only uses raw facial images as an input:

- Guided image analysis: Eyebrow, eye, and mouth regions are essential to recognize facial expressions [84]. The proposed CNN architecture forces the earlier layers of the CNN system to learn to detect and localize these facial regions, thus providing decoupled and guided training.
- Part-based and holistic information fusion: Owing to the proposed system, part-based (first three CNNs) and holistic (fourth CNN) information are fused. Combining part-based and holistic information is improved the accuracy of the recognition.
- Privacy and patient de-identification: If facial expressions are used for medical diagnostic purposes, disease progression or treatment outcome monitoring, de-identification of protected health information is critical (HIPAA regulations). Owing to the final iconize images which are the combined output of the first-three CNN structures; using, archiving, and communication of essential facial features are facilitated while patient privacy is still protected.
- Higher success rate: In order to better show the benefits of proposed 4-cascade architecture, in the experiments the fourth CNN architecture has been trained and tested with raw facial images and higher success rate has been obtained (Table 4).

2 Related works

Numerous papers in the literature report the relationship between psychological and neurological disorders and facial expressions. According to [39], Alzheimer's patients may have a deficiency in facial expression behaviors. According to [82], negative expressions such as fear, disgust, and sadness can be seen in the majority of neurodegenerative disorders. In [27], facial expression abilities of Alzheimer patients, frontotemporal dementia patients, and healthy individuals were examined. The researchers observed that frontotemporal dementia patients had better abilities for positive facial expression than Alzheimer patients. In [12], a tool that uses facial expression to support neurological disorders was presented. In the tool, while a patient watches a video, the system detects the facial expression of the patient and makes disease state predictions based upon the absence or intensity of facial expression. In [32], facial expression abilities of children with and without Autism spectrum disorders were examined. Six universal emotions were compared based on statistical analysis and time-series modeling. A more pronounced group of differences was noted for negative emotions. Dantcheva et al. [20] presented a method for automated facial activities and expression recognition for patients suffering from severe dementia. The system classifies four expression states and activities, namely neutral, smiling, talking, and singing. Dapogny et al. [21] introduced a game that teaches children with autism spectrum disorder how to produce facial expressions.

This study presents a facial expression recognition tool for monitoring neurological disorders. It is the first step towards a non-invasive automated system for clinical diagnosis and neurological disorder monitoring. In order to recognize facial expressions, eyebrow, eye and mouth regions include significant information [84]. Therefore, we have been focused on firstly training our network to segment these facial regions. Then, using these segmented images, we trained our network to recognize facial expressions. In our previous study [81]; we proposed a facial component segmentation method based on training a single network for three kinds of facial regions (mouth, eyes, and eyebrows). In this study, we extended our segmentation method to include three networks, one for each kind of facial component. Because, in human face images; eyebrows, eyes, and mouth have different appearance characteristics. So, each

individual feature can be detected with a specific CNN successfully. Use of three separate networks and filtering process described in section 3.2 increased recognition accuracy. This paper also greatly extends our previous work with new experiments and results, expanded analysis and discussion, new validation dataset, new comparison methods etc.

While classical computer vision techniques can be used to detect facial landmarks, we chose to use deep learning because: (1) Deep learning methods have demonstrated increasing success and popularity in recent years. It has been observed that some deep learning methods outperform simple heuristics or descriptors such as SIFT or Haar-like features [15, 49, 71]. (2) Deep learning based approaches are more adaptable to new modalities and scenarios compared to detection with hand-crafted features and descriptors that are often fine-tuned for specific modalities [69, 71]. (3) Size of Haar-like features can be relatively high, for instance in [75] 160,000 features are located for a 24 × 24 detection window. (4) Using CNNs for both facial landmark detection and facial expression recognition steps ensure a more unified pipeline compared to using hand-crafted features for landmark detection and CNN for facial expression recognition.

3 Methodology

CNN is a popular subfield of deep learning for image analysis. Early CNN [42] was introduced as a neural network architecture, including convolutional and sub-sampling (pooling) layers. In 2012, Krizhevsky et al. [40] achieved significant performance on the Imagenet classification challenge. Since then, deep learning has been widely used in complex computer vision problems such as face detection [48, 60], facial expression recognition [17, 73, 78], biomedical image analysis [18, 45, 50, 55, 56, 66, 68], head pose detection [74], gender classification [37], age classification [8, 35], object detection [83], etc.

CNN structures mostly include convolutional, pooling, activation, batch normalization, fully connected and drop out layers. The inputs are convolved with learned filters and feature maps are generated in convolutional layers (Eq. 1).

$$S(i,j) = (I*K)(i,j) = \sum_m \sum_n I(i+m, j+n) x K(m,n) \Big)$$

(1)

where I is input image, K is kernel and S is the convolution result [31].

In a CNN structure, a non-linear activation layer is often used after the convolutional layers. In this study, for all of CNN structures, Rectified Linear Unit (ReLU) has been used for the activation function. ReLU is defined in Eq. 2.

$$ReLU(x) = \max(0, x)$$

(2)

where x is the input of the activation function [53].

The spatial size of the inputs is reduced in pooling layers, thus the number of parameters and computation are decreased, and the overfitting is controlled. According to [34], dropout layer is also very effective for avoiding overfitting. Dropout layer keeps the network from being too dependent on any neuron. In a fully connected layer, all activations are connected to the previous layer, and a classification or regression task is performed.

In a feedforward network, inputs are passed through the network and the output obtained from the network is compared with the actual output [31]. Using the backpropagation

algorithm, this error propagates backwards to improve training [31]. Backpropagation algorithm performs based on chain rule. For a complete neural network framework with a loss L, the backpropagation computes the gradient of the parameter matrix W and the input x as Eq. 3 and Eq. 4 [76]:

$$\frac{\partial L}{\partial W} = \frac{\partial L}{\partial y} \frac{\partial y}{\partial W} \tag{3}$$

$$\frac{\partial L}{\partial x} = \frac{\partial y}{\partial x} \frac{\partial L}{\partial y} \tag{4}$$

In this study, a cascade CNN architecture has been developed for segmentation-guided facial expression recognition. First three cascade of the CNN architecture segments facial components and forms the component-segmented images. CNN-1, CNN-2, and CNN-3 form eyebrow-segmented, eye-segmented and mouth-segmented images, respectively. After the post-processing step, these component-segmented images are combined to form a final iconize image for each face. Finally, using the final iconize images and the corresponding raw facial images, the fourth CNN classifies facial expressions. The system flow is illustrated in Fig. 1, and, detailed CNN structure is given in Fig. 2.

3.1 CNN for facial component segmentation

In the cascade architecture, the first three CNNs are trained to segment eyebrow, eye and mouth regions from facial images because of the importance of these regions to recognize the expressions. Using the first three CNN outputs, the proposed system forms a face-iconize image that is used by the fourth CNN as input.

In the proposed method, segmentation is handled as a binary classification problem. Every 16×16 block is classified as eyebrow, eye or mouth versus background. Before training of the segmentation networks, training masks have been generated. In order to obtain training masks, Face++ toolkit [63] has been used. The toolkit can detect and localize facial landmarks on a
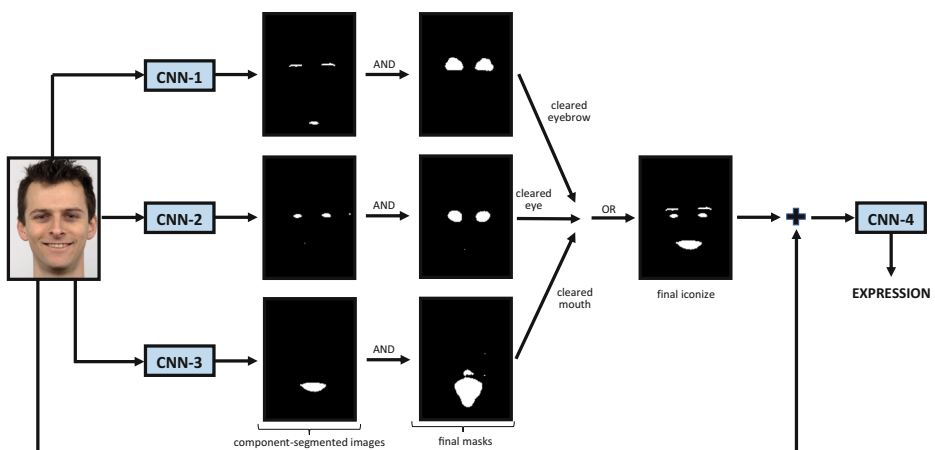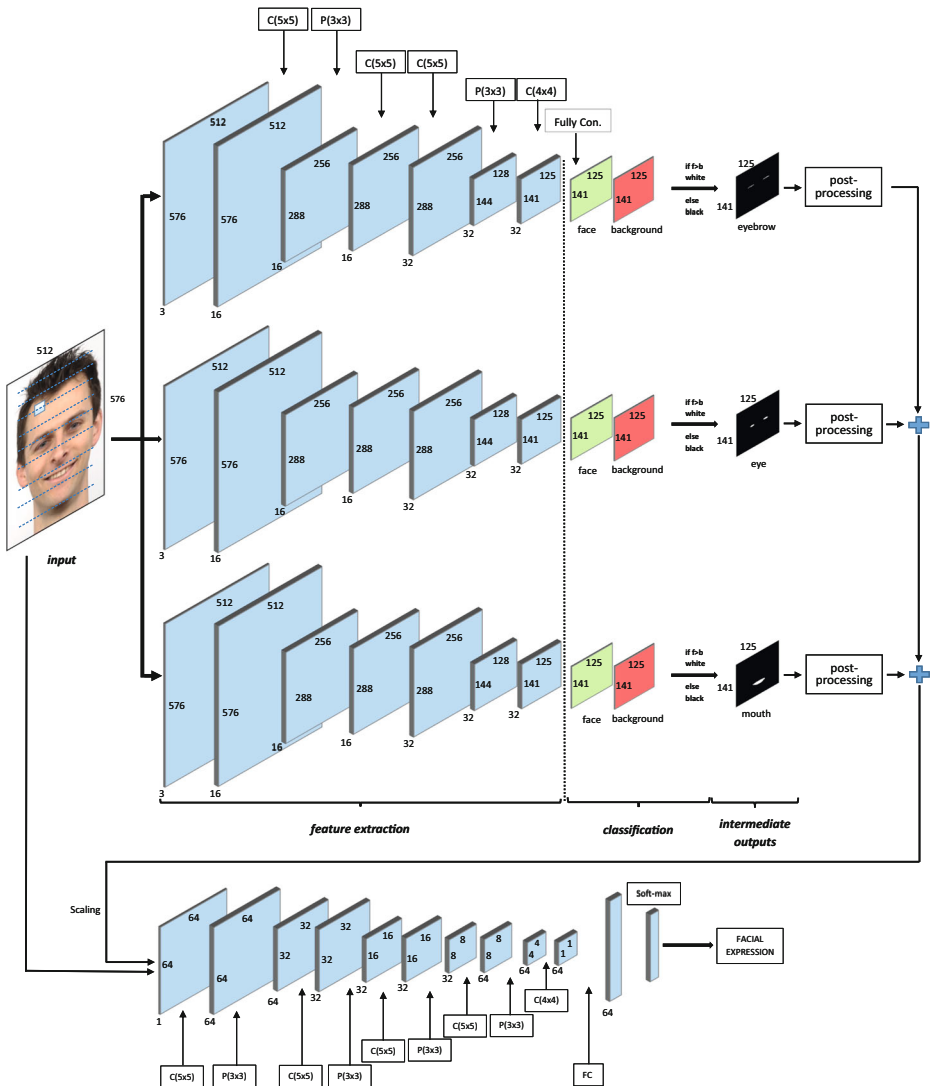

Fig. 1 Proposed system pipeline

Fig. 2 Proposed four-stage CNN structure (first CNN for eyebrow segmentation, second CNN for eye segmentation, third CNN for mouth segmentation and fourth CNN for recognition of facial expression)

face image. After detection of the landmarks, the points of the landmarks have been linked to get polygons for eyebrows, eyes and mouth regions. Finally, the polygons have been filled to obtain final mask images. Figure 3 shows training masks generation steps. In the figure, green pixels show facial components that are eyebrows, eyes, and mouth; red pixels show the rest of the facial components as the background.

The training masks are used for determining majority and mixed classes in the facial component segmentation step. Before the training step of the facial landmark segmentation, original raw images and corresponding training masks have been divided into 16 × 16 non-overlapping blocks as shown in Fig. 4.
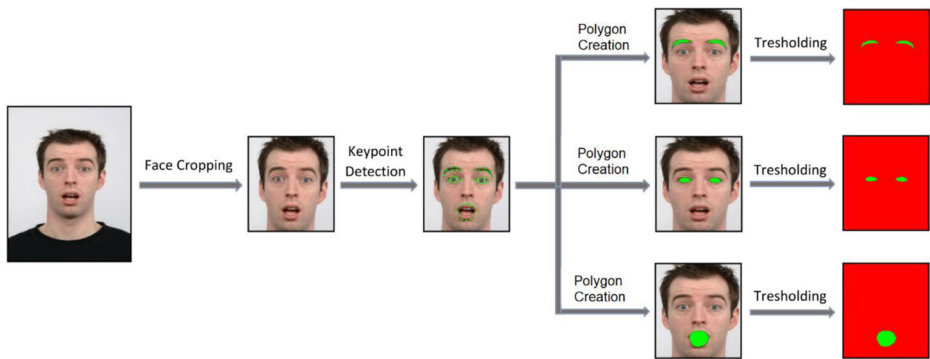
Fig. 3 Training mask generation steps

After image partitioning, the obtained blocks are assigned a label corresponding to one of the following classes facial component, background, or mixed, according to the distribution of pixel labels in the block To determine a block as a background or facial component (eyebrow, eye or mouth), in the corresponding training mask, each green and red pixel numbers are summed separately and when 80% or more of a block is covered by one of the two classes (facial component/background), the block is assigned the label of the majority class. When the percentage of the majority class is less than 80%, the block is marked as mixed and not used during training. These processes are shown in Eqs. 5 and 6. The threshold value 80% is empirically selected.

$$If \ \frac{number\_of\_green\_pixels}{number\_of\_all\_pixels\_of\_the\_block} \geq 80\% \rightarrow block\_is\_facial\_component \qquad (5)$$

$$If \ \frac{number\_of\_red\_pixels}{number\_of\_all\_pixels\_of\_the\_block} \geq 80\% \rightarrow block\_is\_background \qquad (6)$$
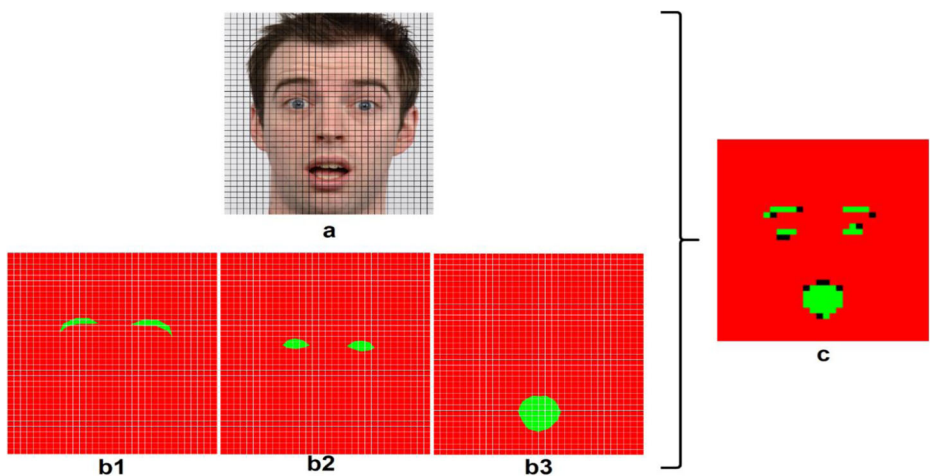

Fig. 4 Non-overlapping blocks on a raw image with corresponding training masks and determined final classes of the blocks (a: raw image, b1,b2,b3: training mask of eyebrow, eye and mouth respectively, c: training blocks, green pixels: facial component, red pixels: background, black pixels: mixed blocks)
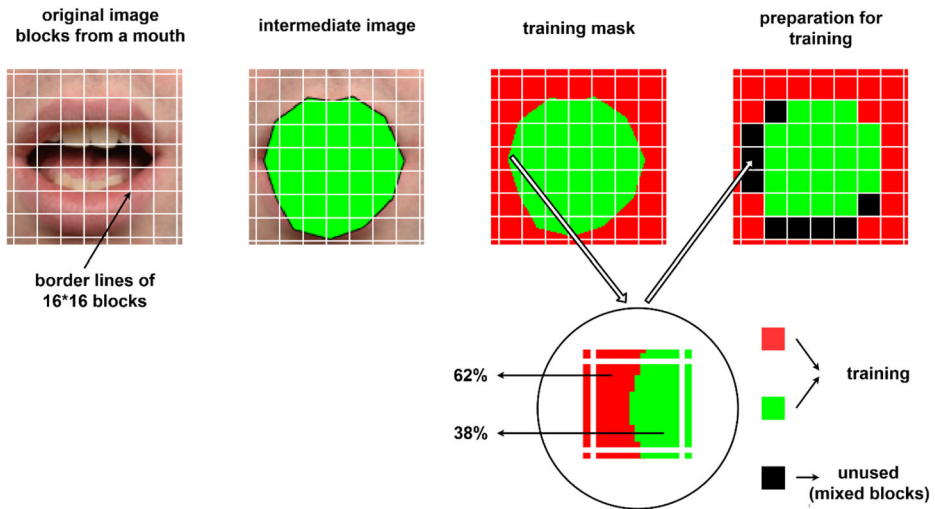
Fig. 5 Determining block status

Mixed class blocks which are shown in Fig. 4c with black pixels include both facial components and background pixels. Thus, using these blocks may be complicated for network training. But majority class blocks include robust information about background or facial components and using only these blocks provides stronger training. In Fig. 5, a block is shown as a sample. The block in training mask has 62% background information and 38% mouth information. The block hasn't 80% either facial component or background pixels and it is ignored during training. Because this block does not include robust information about a class.

After the network training, testing is applied using the whole image (as opposed to 16×16 blocks from the images) as described in [65]. Sliding window processing is efficiently simulated by reducing computation redundancy on overlapping regions. The fully connected layers of the first three CNN networks have two channel scores: for the first CNN; eyebrow versus background, for the second CNN; eye versus background and for the third CNN; mouth versus background scores. Finally, three type component-segmented (eyebrow-segmented, eye-segmented and mouth-segmented) images are obtained from the first three CNNs according to the higher component scores in the fully connected layers. This first three CNN architectures include the following layers: 1) four convolutional layers (layer 1: $16 \ 5 \times 5 \times 3$

Table 1 Proposed CNN architecture layer information. CNN-1,2,3 have 3-channel inputs. CNN-4 has 1,3, and 4-channel inputs. To simplify the table, dropout, pooling, ReLU layers are not listed

Architecture	Layer	Kernel	Filter	Output
CNN-1,2,3 (Facial Component	conv1	5×5	16	16x16x16
Segmentation Input: 16×16 blocks)	conv2	5×5	16	8x8x16
	conv3	5×5	32	4x4x32
	conv4	4×4	32	1x1x32
CNN-4 (Facial Expr.	conv1	5×5	64	64x64x64
Recognition Input: 64×64 full image)	conv2	5×5	32	32x32x32
	conv3	5×5	32	16x16x32
	conv4	5×5	64	8x8x64
	conv5	4×4	64	1x1x64

filters, layer 2: 16 5 × 5 filters, layer 3: 32 5 × 5 filters, and layer 4: 32 4 × 4 filters), 2) two pooling layers and 3) one fully connected layer (see Table 1).

3.2 Segmentation refinement

After obtaining component-segmented (eyebrow segmented) images from the CNN-1 for all images, an empty matrix is created with the component-segmented images size. All the eyebrow-segmented images are added to the empty matrix, respectively and eyebrow intermediate mask is generated. With this process, areas, where the segmented eyebrows are concentrated, are found out. These steps are repeated for eye and mouth intermediate masks. Thus, noisy areas occur due to the lack of density and these areas are cleared with a threshold value. The threshold value is selected 128 from our experiments. Figure 6 illustrates the final masks generation steps.

After generating the final masks, the logical-and operator is applied to each component-segmented image and corresponding final mask for noise reduction (Fig. 1). Thus; cleared eyebrow, eye and mouth iconize images are obtained and combined to form the final-iconize image using logical-or operation. Figure 7 illustrates the combining process.
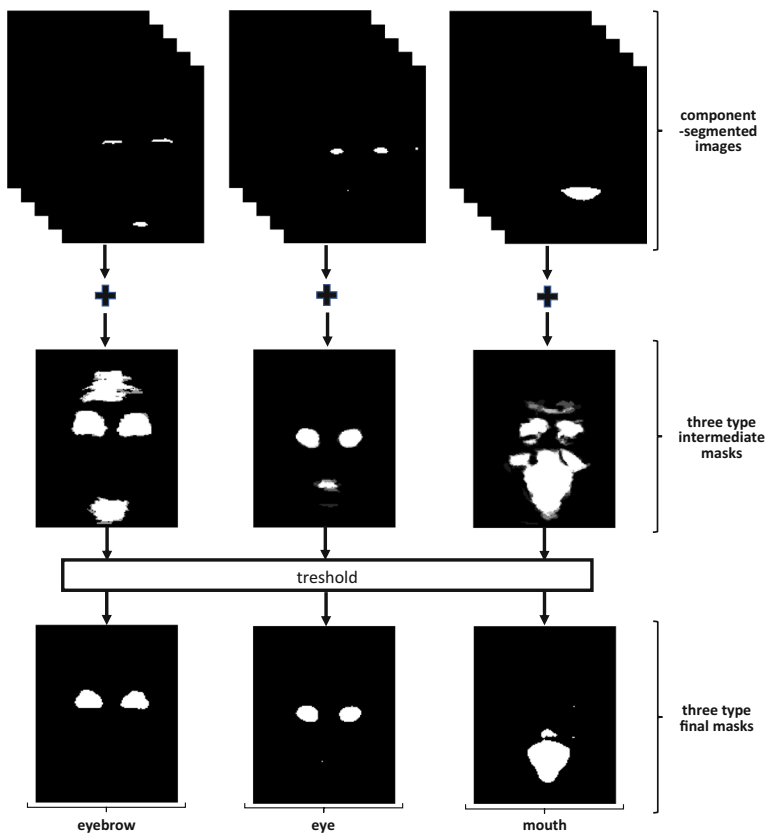

Fig. 6 Final mask generation step. Firstly, each type of components is summed separately and intermediate masks are obtained. In these masks, lose gray pixels are removed after applying a threshold and final masks are formed
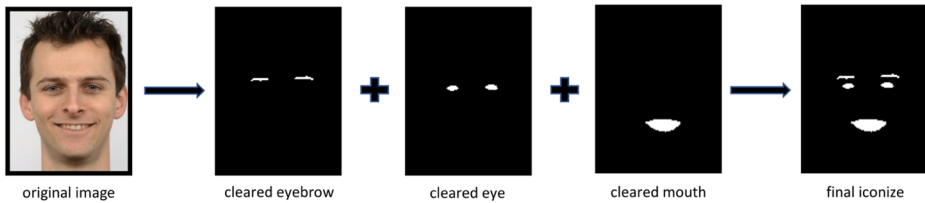
Fig. 7 Final iconize generation

3.3 CNN for facial expression recognition

The fourth CNN structure in the proposed architecture uses the final iconize image (1-channel) combined with the corresponding raw facial images (3-channel). While the first three CNN structures operate on 16×16 blocks from the facial images, the fourth CNN uses a resized whole face image as the input. The fourth CNN architecture contains the following layers: 1) five convolutional layers (layer 1: 64 $5 \times 5 \times 3$ filters, layer 2: 32 5×5 filters, layer 3: 32 5×5 filters, layer 4: 64 5×5 filters, and layer 5: 64 4×4 filters), 2) four pooling layers and 3) one fully connected layer. Table 1 illustrates the layer information of the proposed CNN architecture.

4 Experimental results

In this study, the Radboud Face Database (RaFD) [41] has been used for training and testing steps. RaFD is a public face database that includes facial images of 67 people, consisting of 19 female and 38 male adults; 6 female and 4 male children. Each person has different images with different angles and expressions. The database was generated according to the Facial Action Coding System [25].

4.1 Experiments on facial component segmentation

1608 face images regardless of facial expressions were used for facial component segmentation. These 1608 images were divided into two set for training and testing. First 804 images were used for training, the remaining 804 images were used for testing. If just entire images are used in training process, just 804 images/samples can be used and using less input data will decrease the segmentation results. But our segmentation system divides the images 16×16 non-overlapping blocks and uses these blocks for training data. The system has 926,208 blocks for the input. These blocks were labeled as one of the following classes facial component,

Table 2 The number of the blocks used and unused in facial component segmentation step (Facial component numbers are given as two-fold because of the flipping process)

Class type	For eyebrow segmentation CNN	For eye segmentation CNN	For mouth segmentation CNN
#Facial component class	5864	7066	20262
#Background class	919265	920282	913315
#mixed class	4011	2393	2762
	#Total block: 926208		
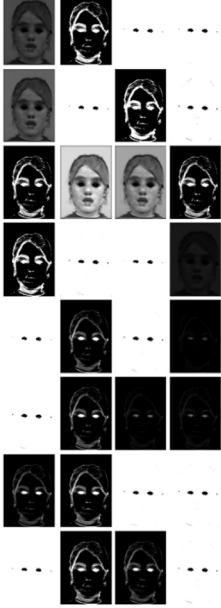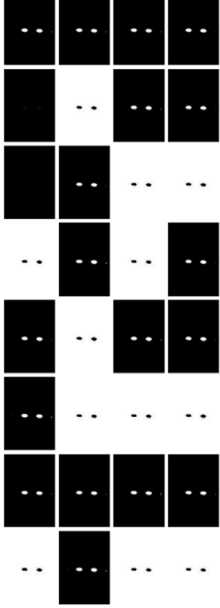
Fig. 8 Layer outputs of the CNN for eye segmentation

background, or mixed, as described in Section 3.1. Blocks which are assigned facial component class were flipped for augmentation, so the number of facial classes were doubled. The number of blocks for different classes used in the training process can be seen in Table 2.

Fig. 9 Transfer learning visual results with the proposed segmentation system results
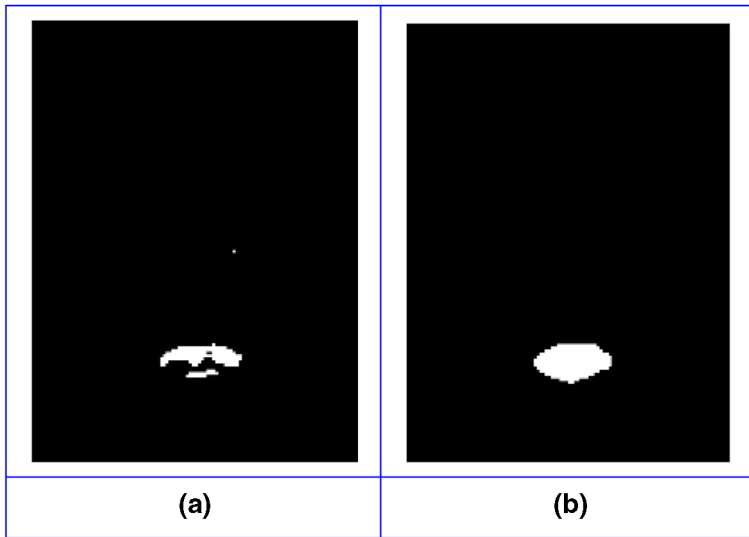
Fig. 10 Effect of different threshold values on block label selections

In the segmentation step, three CNNs work on the image blocks and every CNN has three convolution layers. Intermediate layer outputs for the CNN trained for eye detection and segmentation are shown in Fig. 8 as an example. Obtaining the features through the network can be seen step by step in this figure. In the first convolutional layer, some filters start to learn eye features like the first picture of the "CONV1" column. In the second layer, almost half of the filters learn the feature of the eyes approximately: eye pixels are marked as white pixels by the filters in the pictures. In the last convolutional layer, the network learned the eye features almost completely.

Because the proposed system is designed as a pipeline of facial landmark detection/ segmentation and facial expression recognition networks, other deep segmentation networks can be used to replace the proposed facial landmark segmentation networks. SegNet [9] is a
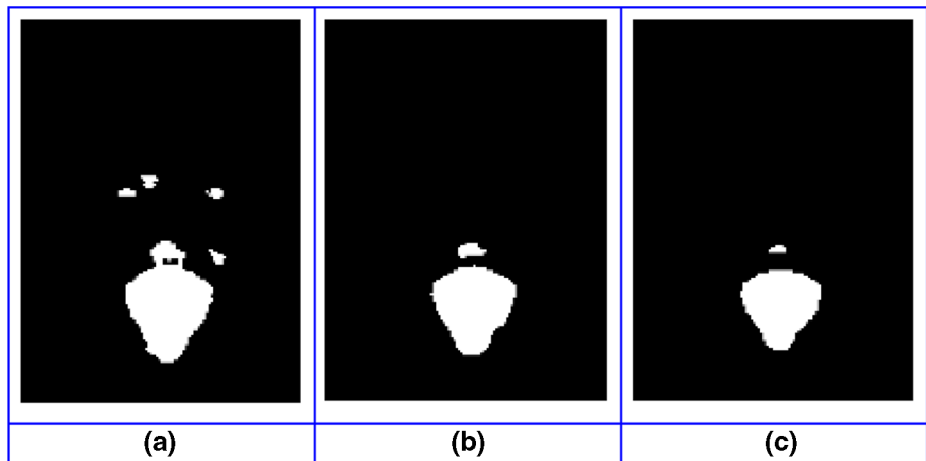
Fig. 11 Effect of different threshold values on generating intermediate masks

Table 3 The effect of different distributions on the result

Training Data (%)	Testing Data (%)	Accuracy (%)
30	70	88.65
50	50	94.11
70	30	94.44
80	20	94.44

popular deep encoder-decoder network used for segmentation tasks [67]. Using RaFD dataset, we have trained three SegNet networks for segmentation of eyebrow, eye and mouth regions. Facial landmark segmentation results obtained from SegNet and the proposed CNNs are shown in Fig. 9. The shown results do not include any post-processing operations. As can be seen in the figure, the proposed networks produce less noisy and more complete detections compared to the SegNet outputs. SegNet uses entire images to segment facial components. The proposed method is more successful when comparing whole based and partial based approach.

To obtain a better segmentation result, a good threshold value should be chosen while determining block labels. If the threshold value is selected too small, the blocks include mixed features for facial components and background. In this situation, a successful training would not be obtained. In order to better show the differences between smaller value and the value of 80%, the threshold value has been selected 50% and the network has been trained again. Figure 10 shows visual testing results for training with the threshold value of 50% and 80%. Figure 10. **a** represents a mouth segmentation result for 50% threshold value and **b** is for 80%.

If the threshold value is selected too large, the blocks include robust features for facial components or background. But, in this case, number of blocks will be too small, especially for facial components. In machine learning problems, sample space size is very important, and it determines the quality of training.

Also, in segmentation refinement stage, if the threshold value is selected too large, intermediate masks will have too much noise. In other case, the pixels of the facial components can be lost. In order to better show the differences between selecting smaller or larger threshold value and the value of 128, the threshold value has been selected 64 and 192. Visual differences have been illustrated in Fig. 11. Figure 11a is a final mouth mask for 64 threshold value, (b) is the result for 128 and (c) is for 192. According to the visual results, smaller threshold value affects increasing the noise, but after a level, to use bigger threshold, almost, doesn't affect the visual result of the final mask. Also, some facial component pixels can be lost because of the face position in the images after segmentation refinement process with a big threshold value.

4.2 Experiments on facial expression recognition

Afraid, angry, happy, sad, surprised and disgusted expressions (1206 frontal face images) have been used in this study. Every image has been cropped to contain only the face regions because of computational time reduction (Fig. 3).

Table 4 Facial expression recognition success rates with the different input channel number

Input	#Channels	Accuracy
Raw Image (Single CNN)	3-channel RGB	89.44%
Final-iconize Image (CNN cascade)	1-channel binary	90.55%
Raw + Final-iconize Images (CNN Cascade)	4-channel	94.44%

Table 5 Proposed cascade CNN structure confusion matrix for the RaFD

Predicted (%) Actual (%)	Anger	Disgust	Fear	Happy	Sad	Surprise
Anger	95.24	3.17	0	0	1.59	0
Disgust	1.59	98.41	0	0	0	0
Fear	0	0	85.71	0	12.70	1.59
Happy	0	1.59	0	98.41	0	0
Sad	6.35	0	3.17	0	90.48	0
Surprise	0	0	1.59	0	0	98.41

Average: 94.44%

In order to examine the effects of the size of the training set, the dataset was divided into different sized training and testing sets. Increasing the size of the training set had a positive effect on the result up to a certain level (Table 3). According to the experiments, 70% - 30% distribution of training and testing sets, achieves facial expression accuracy of 94.44%. To ensure more reliable performance evaluation, for all of the distributions, the testing and training datasets do not include images of the same person even for different expressions.

In this study, CNN-4 has been trained and tested for different inputs with different channel numbers. The proposed system uses 4-channel input, but in order to better show the benefits of the proposed architecture, 1-channel binary final-iconize image and 3-channel raw facial image also have been used by CNN-4. Facial expression recognition results for different inputs with a different number of channels are illustrated in Table 4 using RaFD images with %70 training and %30 testing sets.

For RaFD experiments, facial expression recognition using 1-channel binary final-iconize image outperforms recognition using 3-channel raw facial image by 1.11% (90.55 versus 89,44%). Using final iconize image combined with raw facial image outperforms using 3-channel raw facial image by 5% (94.44% versus 89.44%). Table 5 shows the proposed cascaded CNN architecture confusion matrix using 4-channel input with %70 training and %30 testing sets. The accuracy of the proposed system is 95.24% for anger, 98.41% for disgust, 85.71% for fear, 98.41% for happy, 90.48% for sadness and 98.41% for surprise.

There is not a unified benchmark test set for facial expression recognition, different groups to use different test sets to report their results. Also, most databases are not suitable for the proposed pipeline for various reasons such as resolution, number of facial expressions, labeling, etc. Some databases include images captured in the wild under varying conditions, including partial occlusions are not suitable for our study. Because our goal is to use the

Table 6 Proposed cascade CNN structure confusion matrix for the MUG

| Predicted (%) Actual (%)	Anger	Disgust	Fear	Happy	Sad	Surprise
Anger	91.11	0	0	0	8.89	0
Disgust	2.22	97.78	0	0	0	0
Fear	6.67	0	86.66	0	0	6.67
Happy	0	2.22	0	97.78	0	0
Sad	4.44	0	0	0	95.56	0
Surprise	0	0	8.89	0	0	91.11

Average: 93.33%

Table 7 Performance comparison of the proposed cascaded CNN structure with different studies

Methods	Database	#Expression	Accuracy (%)
HoG + NNE [6]	RaFD, TFEID, JAFFE	5	93.75
Facial Components Detection + KNN [36]	RaFD	6	75.61
Viola & Jones + AAM + ANN [13]	RaFD	7	89.55
Surf Boosting [61]	RaFD	6	90.64
Facial Components Detection + Fuzzy [36]	RaFD	6	93.96
CNN [26]	RaFD	6	94.16
Cascade CNN [81]	RaFD	6	93.43
LBP + SVM [2]	MUG	7	77.14
LBP + Geometric Features + SVM [28]	MUG	6	83.12
CNN [26]	MUG	6	87.68
Gabor + NN [19]	MUG	6	89.29
PCA + SRC [5]	MUG	7	91.27
Landmark points + SVM [3]	MUG	6	92.76
Proposed method	RaFD(3-channel raw image)	6	89.44
	RaFD (1-channel final-iconize)	6	90.55
	MUG (4-channel combine)	6	93.33
	RaFD (4-channel combine)	6	94.44

proposed system for clinical and scientific purposes, specific imaging and pose constraints (i.e. frontal view, no occlusion of facial landmarks, minimum image resolution) have been enforced to ensure the highest accuracy and comparable results across patients and studies. Another problem with some databases is low resolution images. The proposed pipeline is designed for approximately 576 × 512 resolution images where facial landmark shapes are distinct.

In order to evaluate the proposed method, the MUG facial expression database [4] has been also used. The MUG is a video database; thus, the images must be selected from the video sequences. The studies that worked with this database had chosen images with different strategies [2, 3, 5, 19, 28]. Three expression images were selected from each video of each model. These images were grouped into two sets for training (70%) and testing (30%). The training and testing sets were picked so that they do not include images of the same person even for different expressions, to ensure more reliable performance evaluation. The confusion matrix is given in Table 6.

Also, the proposed method has been compared with other studies that use the same databases (Table 7). Our preliminary results achieved a high success rate. Image analysis with high accuracy is very important for medical applications.
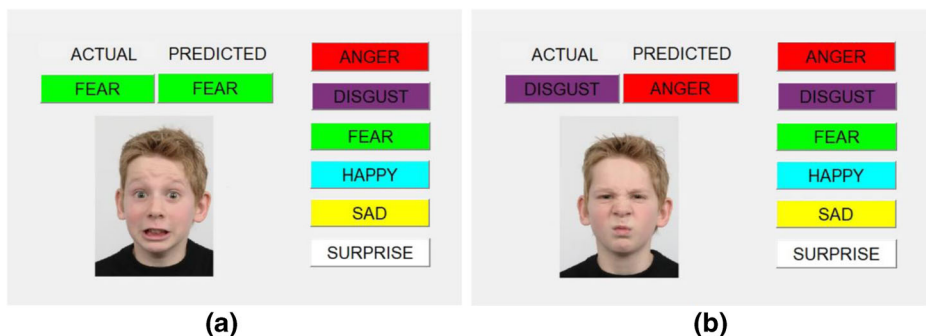
Fig. 12 Proposed system interface (a: correct prediction, b: incorrect prediction)

In the proposed system, facial expression recognition expert knowledge and learned complex features from the deep learning system are used featly. According to our experiments, guided classification improves expression recognition results.

Figure 12 shows two display of the proposed system. Figure 12a, the actual facial expression of the person is fear and proposed system detects it correctly. But, in Fig. 12b, while the actual label is disgust according to RAFD labels, the system prediction is anger.

5 Conclusion

Impaired facial expression can be related to medical disorders. Therefore, a facial expression recognition system can be extremely useful for medical purposes. This paper presents a quantitative, objective and noninvasive system for diagnosis of neurological disorders. Owing to the automated facial expression recognition system, clinically relevant facial expression features can be revealed. Also, it can make a distinction between conditions of disease and can serve as disease specific biomarkers to help in clinical diagnosis and monitoring therapeutic reactions of patients with neurological conditions. This study presents a novel deep learning approach for facial expression recognition which includes four CNN structures. Three types of facial components are segmented in the first three CNNs and an iconize output is formed. The iconize output is combined with raw facial image and used as the input for the last CNN structure. Facial expression classification is performed in the fourth CNN. Owing to the proposed cascade system, integration of part-based and holistic information and guided image classification is ensured. Preliminary results achieved 94.44% accuracy of facial expression recognition for the RaFD database. The proposed system produced 5% success rate than the face recognition system using raw images alone.

References

1. Adams D, Horsler K, Mount R, Oliver C (2015) Brief Report: A Longitudinal Study of Excessive Smiling and Laughing in Children with Angelman Syndrome. J Autism Dev Disord 45(8):2624–2627
2. Agarwal S, Santra B, Mukherjee DP (2018) Anubhav: recognizing emotions through facial expression. Vis Comput 34(2):177–191
3. Aifanti N, Delopoulos A (2014) Linear subspaces for facial expression recognition. Signal Process Image Commun 29(1):177–188
4. Aifanti N, Papachristou C, Delopoulos A (2010) The MUG facial expression database. In: 11th International Workshop on Image and Audio Analysis for Multimedia Interactive services, WIAMIS 2010, pp. 1–4
5. Aina S, Zhou M, Chambers JA, Phan RC (2014) A new spontaneous expression database and a study of classification-based expression analysis methods. In: 2014 22nd European Signal Processing Conference (EUSIPCO), pp. 2505–2509.
6. Ali G, Iqbal MA, Choi T-S (2016) Boosted NNE collections for multicultural facial expression recognition. Pattern Recogn 55:14–27
7. Alphonse AS, Dharma D (2018) Novel directional patterns and a Generalized Supervised Dimension Reduction System (GSDRS) for facial emotion recognition. Multimed Tools Appl 77(8):9455–9488

8. Aydogdu MF, Celik V, Demirci MF (2017) Comparison of Three Different CNN Architectures for Age Classification. In: 2017 IEEE 11th International Conference on Semantic Computing (ICSC), pp. 372–377

9. Badrinarayanan V, Kendall A, Cipolla R (2017) SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. IEEE Trans Pattern Anal Mach Intell 39(12):2481–2495

10. Baugh RF, Basura GJ, Ishii LE, Schwartz SR, Drumheller CM, Burkholder R, Deckard NA, Dawson C, Driscoll C, Gillespie MB, Gurgel RK, Halperin J, Khalid AN, Kumar KA, Micco A, Munsell D, Rosenbaum S, Vaughan W (2013) Clinical Practice Guideline. Otolaryngol Head Neck Surg 149(5):656–663

11. Ben Abdallah T, Guermazi R, Hammami M (2018) Facial-expression recognition based on a low-dimensional temporal feature space. Multimed Tools Appl 77(15):19455–19479

12. Bevilacqua V, D'Ambruoso D, Mandolino G, Suma M (2011) A new tool to support diagnosis of neurological disorders by means of facial expressions. In: 2011 IEEE International Symposium on Medical Measurements and Applications, pp. 544–549

13. Bijlstra G, Dotsch R (2011) FaceReader 4 emotion classification performance on images from the Radboud Faces Database

14. Brewer R, Biotti F, Catmur C, Press C, Happé F, Cook R, Bird G (2016) Can Neurotypical Individuals Read Autistic Facial Expressions? Atypical Production of Emotional Facial Expressions in Autism Spectrum Disorders. Autism Res 9(2):262–271

15. Cha KH, Hadjiiski L, Samala RK, Chan H-P, Caoili EM, Cohan RH (2016) Urinary bladder segmentation in CT urography using deep-learning convolutional neural network and level sets. Med Phys 43(4):1882–1896

16. Chang J, Ryoo S (2018) Implementation of an improved facial emotion retrieval method in multimedia system. Multimed Tools Appl 77(4):5059–5065

17. Chen J, Xu R, Liu L (2018) Deep peak-neutral difference feature for facial expression recognition. Multimed Tools Appl

18. Cheng H-C, Cardone A, Krokos E, Stoica B, Faden A, Varshney A (2017) Deep-learning-assisted visualization for live-cell images. In: 2017 IEEE International Conference on Image Processing (ICIP), pp. 1377–1381

19. da Silva FAM, Pedrini H (2015) Effects of cultural characteristics on building an emotion classifier through facial expression analysis. Journal of Electronic Imaging 24(2):23015

20. Dantcheva A, Bilinski P, Nguyen HT, Broutart J-C, Bremond F (2017) Expression recognition for severely demented patients in music reminiscence-therapy. In 2017 25th European Signal Processing Conference (EUSIPCO), pp. 783–787

21. Dapogny A, Grossard C, Hun S, Serret S, Bourgeois J, Jean-Marie H, Foulon P, Ding H, Chen L, Dubuisson S, Grynszpan O, Cohen D, Bailly K (2018) JEMImE: A Serious Game to Teach Children with ASD How to Adequately Produce Facial Expressions. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pp. 723–730

22. Dornaika F, Moujahid A, Raducanu B (2013) Facial expression recognition using tracked facial actions: Classifier performance analysis. Eng Appl Artif Intell 26(1):467–477

23. Edvinsson SE, Lundqvist L-O (2016) Prevalence of orofacial dysfunction in cerebral palsy and its association with gross motor function and manual ability. Dev Med Child Neurol 58(4):385–394

24. Ekman P, Friesen WV (1971) Constants across cultures in the face and emotion. J Pers Soc Psychol 17(2):124–129

25. Ekman P, Friesen WV (2002) Investigator's Guide to the Facial Action Coding System (FACS)

26. Fathallah A, Abdi L, Douik A (2017) Facial Expression Recognition via Deep Learning. In: 2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA), pp. 745–750.

27. Fernandez-Duque D, Black SE (2005) Impaired recognition of negative facial emotions in patients with frontotemporal dementia. Neuropsychologia 43(11):1673–1687

28. Ghimire D, Jeong S, Yoon S, Choi J, Lee J (2015) Facial expression recognition based on region specific appearance and geometric features. In: 2015 Tenth International Conference on Digital Information Management (ICDIM), pp. 142–147

29. Ghimire D, Lee J (2013) Geometric Feature-Based Facial Expression Recognition in Image Sequences Using Multi-Class AdaBoost and Support Vector Machines. Sensors 13(6):7714–7734

30. Gola KA, Shany-Ur T, Pressman P, Sulman I, Galeana E, Paulsen H, Nguyen L, Wu T, Adhimoolam B, Poorzand P, Miller BL, Rankin KP (2017) A neural network underlying intentional emotional facial expression in neurodegenerative disease. NeuroImage: Clinical 14:672–678

31. Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT Press, Cambridge

32. Guha T, Yang Z, Ramakrishna A, Grossman RB, Hedley D, Lee S, Narayanan SS (2015) On quantifying facial expression-related atypicality of children with Autism Spectrum Disorder. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 803–807

33. Guo M, Hou X, Ma Y (2017) Facial expression recognition using ELBP based on covariance matrix transform in KLT. Multimed Tools Appl:2995–3010
34. Hinton GE, Srivastava N, Krizhevsky A, Sutskever I, Salakhutdinov RR (2012) Improving neural networks by preventing co-adaptation of feature detectors. arXiv
35. Hosseini S, Lee SH, Kwon HJ, Il Koo H, Cho NI (2018) Age and gender classification using wide convolutional neural network and Gabor filter. in 2018 International Workshop on Advanced Image Technology (IWAIT), pp. 1–3.
36. Ilbeygi M, Shah-Hosseini H (2012) A novel fuzzy facial expression recognition system based on facial feature extraction from color face images. Eng Appl Artif Intell 25(1):130–146
37. Jia S, Lansdall-Welfare T, Cristianini N (2016) Gender Classification by Deep Learning on Millions of Weakly Labelled Images. In: 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), pp. 462–467
38. Khan SA, Hussain A, Usman M (2018) Reliable facial expression recognition for multi-scale images using weber local binary image based cosine transform features. Multimed Tools Appl 77(1):1133–1165
39. Kohler CG (2005) Emotion-Discrimination Deficits in Mild Alzheimer Disease. Am J Geriatr Psychiatr 13(11):926–933
40. Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet Classification with Deep Convolutional Neural Networks. Adv Neural Inf Proces Syst:1–9
41. Langner O, Dotsch R, Bijlstra G, Wigboldus DHJ, Hawk ST, van Knippenberg A (2010) Presentation and validation of the Radboud Faces Database. Cognit Emot 24(8):1377–1388
42. Lecun Y (1989) Generalization and network design strategies. In: Pfeifer R, Schreter Z, Fogelman F, Steels L (eds) Connectionism in perspective. Elsevier, Zurich
43. Li Z, Zhang Q, Duan X, Wang C, Shi Y (2018) New semantic descriptor construction for facial expression recognition based on axiomatic fuzzy set. Multimed Tools Appl 77(10):11775–11805
44. Lin J, Chen Y, Wen H, Yang Z, Zeng J (2017) Weakness of Eye Closure with Central Facial Paralysis after Unilateral Hemispheric Stroke Predicts a Worse Outcome. J Stroke Cerebrovasc Dis 26(4):834–841
45. Liu S, Liu S, Cai W, Pujol S, Kikinis R, Feng D (2014) Early diagnosis of Alzheimer's disease with deep learning. In: 2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI), pp. 1015–1018
46. Livingstone SR, Vezer E, McGarry LM, Lang AE, Russo FA (2016) Deficits in the Mimicry of Facial Expressions in Parkinson's Disease. Front Psychol 7
47. Lopes AT, de Aguiar E, De Souza AF, Oliveira-Santos T (2017) Facial expression recognition with Convolutional Neural Networks: Coping with few data and the training sample order. Pattern Recogn 61: 610–628
48. Lou Y, Fu G, Jiang Z, Men A, Zhou Y (2017) PT-NET: Improve object and face detection via a pre-trained CNN model. In: 2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP), pp. 1280–1284.
49. Luus FPS, Salmon BP, van den Bergh F, Maharaj BTJ (2015) Multiview Deep Learning for Land-Use Classification. IEEE Geosci Remote Sens Lett 12(12):2448–2452
50. Mandache D, Dalimier E, Durkin JR, Boceara C, Olivo-Marin J-C, Meas-Yedid V (2018) Basal cell carcinoma detection in full field OCT images using convolutional neural networks. in 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 784–787.
51. Matsugu M, Mori K, Mitari Y, Kaneda Y (2003) Subject independent facial expression recognition with robust face detection using a convolutional neural network. Neural Netw 16(5–6):555–559
52. Mehrabian A (1968) Some referents and measures of nonverbal behavior. Behav Res Methods Instrum 1(6): 203–207
53. Nair V, Hinton GE (2010) Rectified Linear Units Improve Restricted Boltzmann Machines. In: Proceedings of the 27th International Conference on International Conference on Machine Learning, pp. 807–814
54. Nigam S, Singh R, Misra AK (2018) Efficient facial expression recognition using histogram of oriented gradients in wavelet domain. Multimed Tools Appl
55. Oztel I, Yolcu G, Ersoy I, White T, Bunyak F (2017) Mitochondria segmentation in electron microscopy volumes using deep convolutional neural network. In 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 1195–1200.
56. Oztel I, Yolcu G, Ersoy I, White TA, Bunyak F (2018) Deep learning approaches in electron microscopy imaging for mitochondria segmentation. International Journal of Data Mining and Bioinformatics 21(2):91
57. Oztel I, Yolcu G, Oz C, Kazan S, Bunyak F (2018) iFER: facial expression recognition using automatically selected geometric eye and eyebrow features. Journal of Electronic Imaging 27(2):1
58. Pitaloka DA, Wulandari A, Basaruddin T, Liliana DY (2017) Enhancing CNN with Preprocessing Stage in Automatic Emotion Recognition. Procedia Computer Science 116:523–529
59. Pons G, Masip D (2017) Supervised Committee of Convolutional Neural Networks in Automated Facial Expression Analysis. IEEE Trans Affect Comput:1–1

60. Qin X, Zhou Y, He Z, Wang Y, Tang Z (2017) A Faster R-CNN Based Method for Comic Characters Face Detection. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), pp. 1074–1080

61. Rao Q, Qu X, Mao Q, Zhan Y (2015) Multi-pose facial expression recognition based on SURF boosting. In 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 630–635

62. Ricciardi L, Visco-Comandini F, Erro R, Morgante F, Bologna M, Fasano A, Ricciardi D, Edwards MJ, Kilner J (2017) Facial Emotion Recognition and Expression in Parkinson's Disease: An Emotional Mirror Mechanism? PLoS One 12(1):e0169110

63. S. C. Face++ (2017) Face++ Cognitive Services. Available: https://www.faceplusplus.com/. Accessed: 12 Nov 2017

64. Saha P, Bhattacharjee D, De BK, Nasipuri M (2018) Facial component-based blended facial expressions generation from static neutral face images. Multimed Tools Appl 77(15):20177–20206

65. Shelhamer E, Long J, Darrell T (2016) Fully Convolutional Networks for Semantic Segmentation. Cognit Emot 24(8):1377–1388

66. Shpilman A, Boikiy D, Polyakova M, Kudenko D, Burakov A, Nadezhdina E (2017) Deep Learning of Cell Classification Using Microscope Images of Intracellular Microtubule Networks. In: 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 1–6.

67. Simonyan K, Zisserman A (2014) Very Deep Convolutional Networks for Large-Scale Image Recognition

68. Singh S, Srivastava A, Mi L, Chen K, Wang Y, Caselli RJ, Goradia D, Reiman EM (2017) Deep-learning-based classification of FDG-PET data for Alzheimer's disease categories. In: 13th International Conference on Medical Information Processing and Analysis, p. 84.

69. Socher R, Huval B, Bhat B, Manning CD, Ng AY (2012) Convolutional-recursive Deep Learning for 3D Object Classification. In: Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1, pp. 656–664

70. Sultan Zia M, Hussain M, Arfan Jaffar M (2018) A novel spontaneous facial expression recognition using dynamically weighted majority voting based ensemble classifier. Multimed Tools Appl 77(19):25537–25567

71. Sun X, Wu P, Hoi SCH (2018) Face detection using deep learning: An improved faster RCNN approach. Neurocomputing 299:42–50

72. Thevenot J, Lopez MB, Hadid A (2018) A Survey on Computer Vision for Assistive Medical Diagnosis From Faces. IEEE Journal of Biomedical and Health Informatics 22(5):1497–1511

73. Uddin MZ, Khaksar W, Torresen J (2017) Facial Expression Recognition Using Salient Features and Convolutional Neural Network. IEEE Access 5:26146–26161

74. Venturelli M, Borghi G, Vezzani R, Cucchiara R (2018) Deep Head Pose Estimation from Depth Data for In-Car Automotive Applications. In: Understanding Human Activities Through 3D Sensors, pp. 74–85.

75. Viola P, Jones MJ (2004) Robust Real-Time Face Detection. Int J Comput Vis 57(2):137–154

76. Wei B, Sun X, Ren X, Xu J (2017) Minimal Effort Back Propagation for Convolutional Neural Networks. Computing Research Repository

77. Wu C, Huang C, Chen H (2018) Expression recognition using semantic information and local texture features. Multimed Tools Appl 77(9):11575–11588

78. Wu B-F, Lin C-H (2018) Adaptive Feature Mapping for Customizing Deep Learning Based Facial Expression Recognition Model. IEEE Access 6:12451–12461

79. Xie X, Lam K-M (2009) Facial expression recognition based on shape and texture. Pattern Recogn 42(5):1003–1011

80. Xie W, Shen L, Yang M, Jiang J (2018) Facial expression synthesis with direction field preservation based mesh deformation and lighting fitting based wrinkle mapping. Multimed Tools Appl 77(6):7565–7593

81. Yolcu G, Oztel I, Kazan S, Oz C, Palaniappan K, Lever TE, Bunyak F (2017) Deep learning-based facial expression recognition for monitoring neurological disorders. In: 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 1652–1657

82. Yuvaraj R, Murugappan M, Sundaraj K (2012) Methods and approaches on emotions recognition in neurodegenerative disorders: A review. In 2012 IEEE Symposium on Industrial Electronics and Applications, pp. 287–292

83. Zhang H, Wang K, Tian Y, Gou C, Wang F-Y (2018) MFR-CNN: Incorporating Multi-Scale Features and Global Information for Traffic Object Detection. IEEE Trans Veh Technol:1–1

84. Zhong L, Liu Q, Yang P, Liu B, Huang J, Metaxas DN (2012) Learning active facial patches for expression analysis. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2562–2569

Gozde Yolcu received her BS, MS and Ph.D. degrees in Computer Engineering at Sakarya University, in 2011, 2014 and 2019, respectively. She is a Research Assistant at the Department of Computer Engineering, Sakarya University since 2012. She was a visiting scholar at the Department of Electrical Engineering and Computer Science, University of Missouri-Columbia, USA in 2017–2018. Her research interests include image processing, computer vision, machine learning, biomedical image analysis and deep learning.

Ismail Oztel received his BS,MS and Ph.D. degrees in computer engineering from the Sakarya University in 2011, 2014 and 2018, respectively. He is a research assistant at the Department of Computer Engineering, Sakarya University, Turkey. He was a visiting scholar at the Department of Electrical Engineering and Computer Science, University of Missouri-Columbia, USA in 2017–2018. His current research interests include machine learning, computer vision, biomedical image analysis, deep learning and virtual reality.

Serap Kazan received her BS degree in Electrical and Electronics Engineering, MS degree in Computer Engineering and Ph.D. degree in Electrical and Electronics Engineering at Sakarya University, in 2000, 2003 and 2009, respectively. She is an assistant professor in the Department of Computer Engineering, Sakarya University. Her research interests include computer vision and machine learning.

Cemil Oz received his BS degree in Electronics and Communication Engineering in 1989 from Yildiz Technical University and his MS degree in Electronics and Computer Education in 1993 from Marmara University, Istanbul. During the M.S. studies, he worked as a lecturer in Istanbul Technical University. He completed his Ph.D. in 1998. He worked as a research fellow in University of Missouri-Rolla, MO, USA. He has been working as a professor in Computer and Information Sciences Faculty, Department of Computer Engineering in Sakarya University. His research interests include robotics, vision, artificial intelligence, virtual reality and pattern recognition.

Kannappan Palaniappan is a professor in the Electrical Engineering and Computer Science Department. He has received several notable awards, including the National Academies Jefferson Science Fellowship (first in Missouri), the NASA Public Service Medal for pioneering contributions to (Big Data) scientific visualization of petabyte-sized archives, the Air Force Summer Faculty Fellowship, the Boeing Welliver Summer Faculty Fellowship, and MU's William T. Kemper Fellowship for Teaching Excellence. At NASA's Goddard Space Flight Center, he co-founded the Visualization and Analysis Lab that has produced a number of spectacular Digital Earth visualizations used by search engines (BlueMarble), museums, magazines and broadcast television. He is co-inventor of the Interactive Image SpreadSheet for handling large multispectral imagery, and he developed the first massively parallel semi-fluid cloud motion analysis algorithm using geostationary satellite imagery. In 2014, his team won first place at the IEEE Computer Vision and Pattern Recognition (CVPR) Change Detection Workshop video analytics challenge. In 2015, the team was a finalist in the CVPR Video Object Tracking Challenge, and in 2016, the team won the best paper award at the CVPR Automatic Traffic Surveillance Workshop and also was selected as a finalist for a best student paper award at the IEEE Engineering in Medicine and Biology Society conference (EMBC 2016). He has several U.S. patents, including one for moving object detection using the flux tensor split Gaussian model and the other for fast bundle adjustment to accurately estimate the pose of airborne camera sensor systems. Research projects have been funded by National Institutes of Health, the Air Force Research Laboratory, Army Research Laboratory, NASA, the National Science Foundation and others. His current, multidisciplinary interests in computer vision, high performance computing, data science and biomedical image analysis range across orders of scale from sub-cellular microscopy at the molecular level to aerial and satellite remote sensing imaging at the macro level.

Teresa Lever graduated from East Caroline University with a doctorate in communication science and disorders. Prior to that, she worked as a speech-language pathologist in California and Washington for eleven years. Her research focuses on swallowing disorders and the complications that arise from certain diseases. She is a member of the Society of Neuroscience and recently joined the University of Missouri faculty.

Filiz Bunyak received her BS and MS degrees from the Istanbul Technical University, Turkey and Ph.D. degree from University of Missouri-Rolla, USA. She is an assistant research professor of Electrical Engineering and Computer Science at University of Missouri-Columbia, USA. Her research interests include image processing, computer vision, and pattern recognition with emphasis on biomedical image analysis, aerial and wide-area surveillance, visual tracking, data fusion, segmentation, level set and deep learning methods.