FACIAL EXPRESSION RECOGNITION USING COMBINED PRE-TRAINED CONVNETS

Raid Saabni[1, 2] and Alon Schclar[1]

[1]School of Computer Science,
The Academic College of Tel-Aviv Yaffo, Tel-Aviv, Israel
[2]Traiangle R&D Center, Kafr Qarea, Israel

ABSTRACT

Automatic Facial Expression Recognition (AFER), has been an active research area in the past three decades. Research and development in this area have become continually active due to its wide range of potential applications in many fields. Recent research in the field presents impressive results when using Convolution Neural Network (CNN's, ConvNets). In general, ConvNets proved to be a very common and promising choice for many computer vision tasks including AFER. Motivated by this fact, we parallelly combine modified versions of three ConvNets to generate an Automated Facial Expression Recognition system. This research aims to present a robust architecture and better learning process for a deep ConvNet. Adding four additional layers to the combination of the basic models assembles the net to one large ConvNet and enables the sophisticated boosting of the basic models. The main contribution of this work comes out of this special architecture and the use of a two-phase training process that enables better learning. The new system we present is trained to detect universal facial expressions of seven\eight basic emotions when targeting the FER2013 and FER2013+ benchmarks, respectively. The presented approach improves the results of the used architectures by 4% using the FER2013 and 2% using FER2013+ data sets. The second round of training the presented system increases the accuracy of some of the basic models by close to 3% while improving the accuracy of the whole net.

KEYWORDS

Automatic Facial Expression Recognition, Convolutional Neural Networks, Machine Learning, Boosting, Deep Learning.

1. INTRODUCTION

Psychologists found that verbal and vocal parts of a message contribute only 41% of its meaning while facial movements and expressions contribute 55% of the effect of that message. This fact means that the facial part does the major contribution to human communication and interaction [17]. Therefore, developing Automatic FER applications would be widely applicable for many real-world tasks, which can get the significant benefit of reliable systems that automatically recognize facial expressions and emotions. Some of such fields, are Human-Computer Interface, Human Emotion Analysis, Image Retrieval, User Profiling, Medical Care and Cure, Video Games, Neuro Marketing, and many more. People can vary significantly in the way they show their expressions for even the same person and expression, which makes AFER a more challenging problem. Images also can vary in brightness, background, and pose, and these variations are emphasized when considering different persons with variations in shape, ethnicity, and other factors.

David C. Wyld et al. (Eds): ICAITA, CDKP, SAI, NCO, CMC, SOFT, MLT, AdNLP - 2020
pp. 95-106, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.100908

Facial expression recognition is a task naturally done by humans daily, but it is a complex task for computer programs. This problem is challenging for computers because it is very hard to extract and classify expression's features when images may vary a lot not only in the way that the subjects show their expression but also due to different conditions of lighting, brightness, position, and background. Other difficulties may include face position and direction, face partial occluding by objects in the scene, or due to bad light conditions causing high variations of illumination, which may easily lead to losing main features of facial expressions.

The work done in the 1970s by the psychologist Paul Ekman [20] and his colleagues, is an important milestone in the study of facial expressions and human emotions. This important work has significant importance and a large influence on the development of modern-day automatic facial expression recognizers. This work leads to adapting and developing the comprehensive Facial Action Coding System (FACS), which has since then become the standard for facial expression recognition research. Facial expressions are extremely important in any human interaction, and additional to emotions, it also reflects on other mental activities, social interaction, and physiological signals. Ekman et. el. identified six facial expressions that are universal across all cultures: Anger, Disgust, Fear, Happiness, Sadness, and Surprise. These six emotions in addition to the natural one, are still used for most of the modern automatic facial expression research. State of the art research on competitions and challenges in the field such as Emotion Recognition in the Wild (*EmotiW*) and Kaggle's Facial Expression Recognition Challenge *FER2013* use these seven emotions in their competitions.

In traditional approaches, facial expression recognition usually consists of three main steps. In the first step, the system detects the face region from an image or sequence of images. This is mostly followed by a pre-processing step to emphasize the relevant features and neglect the irrelevant data. In the next step, features are extracted from the region of interest. Selecting a compact and effective facial representation and features from the face image is a vital step for successful facial expression recognition. The last step uses the extracted features to train and obtain a classifier. Many of the recent systems targeting *AFER* are based on Convolutional Neural Networks (*ConvNet*), using existing and new variations of ConvNet architectures. These approaches present many of the state-of-the-art results in tasks of object classification including facial expression recognition. Unlike traditional approaches, in many cases, no human crafted, and designed features are needed, and the system operates as a start-to-end technique.

In the proposed system, we have designed a start-to-end system based on *ConvNet*s. Motivated by our previous work [24], we have used a pre-processing step to extract the region of interest (*ROI*), which has already proved to improve results. We have also used normalization and data augmentation as an additional practice generally used to improve generalization ability [27,29,25]. In section 3, we review some of the previous works in the related filed. In section 4, we present an overview of our approach. Detailed experimental results of the proposed system are presented in section 5 and concluding and future work directions are presented in section 6.

2. PREVIOUS WORK

Though much progress has been made, automatic recognizing of facial expressions with high accuracy remains difficult due to the complexity and variability of facial expressions. Generally, in classic approaches, the system includes two stages: feature learning and selection, and classifier construction. In the first stage, features are extracted from either static images or video sequences of images, to characterize facial appearance/geometry changes caused by activation of target expression. There are two common approaches to extract facial features: geometric and appearance feature-based methods. The geometric features measure the displacements of certain parts of the face such as eyebrows, eyes, mouth lines, and corners. This is based on the

assumption that expressions affect the relative position and size of various features, and that by measuring the movement and relative position of certain facial points, we can determine the underlying facial expression. In such a case, finding and tracking a crucial point in the face region is an important task of geometric feature measurement and face region analysis.

The idea of appearance-based methods assumes that emotions cause changes to face textures, such as wrinkles, bulges, forefront, regions surrounding the mouth, and eyes when performing a particular action. In appearance-based methods, image filters are applied to regions of interest, which can be any specific region in a face image, to extract feature vectors. These methods include Principal Component Analysis (PCA), Locality Preserving Projections (LPP), Linear Discriminate Analysis (LDA), Gabor wavelets, Local Binary Pattern (LBP), and others. As suggested by the psychological studies, the information specific areas such as around nose, eyes, and mouth are more critical for facial expression analysis. Therefore, a subset of features, which are the most effective to distinguish one expression from the others, are often selected to improve the recognition performance. In the second step, a classifier is obtained by training on a data set [10,14,18,19]. Recently, unsupervised feature learning approaches especially those based on Sparse-Coding [5,28,30] and Deep Learning Networks, have been employed to extract underlying features from facial images and have shown promising results in facial expression recognition and analysis.

(A)

(B)

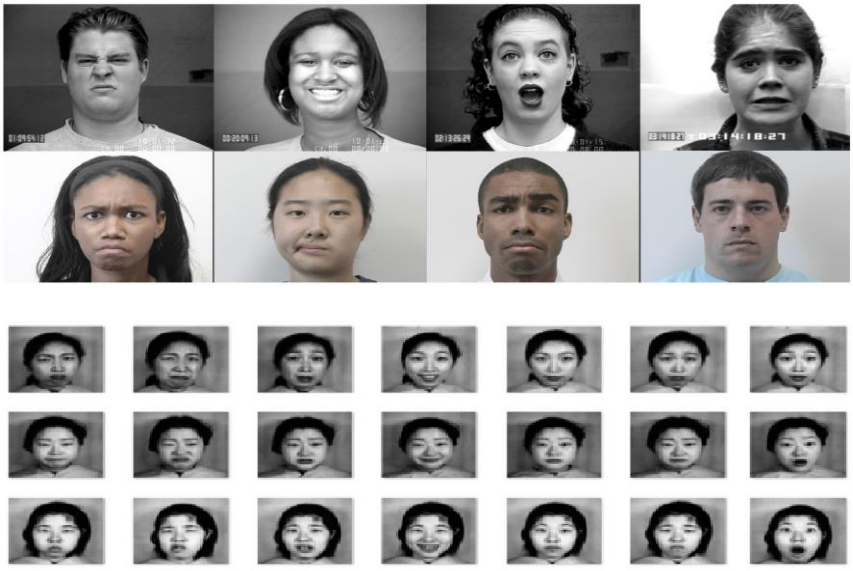


Figure 1: (A) Samples of images from the (CK+) benchmark.
(B) Some samples of the different emotions from the JAFFE data set.

Convolutional Neural Networks are a category of Neural Networks that have proven very effective in image recognition and classification areas. *ConvNet*s have been successful in identifying faces, objects, traffic signs, and many other computer vision tasks. *ConvNet*s work better for image recognition and classification because they can automatically capture spatial features of the inputs due to their large number of filters. These features and filters are not hand-designed but are learned as a part of the training process. This fact makes Neural Networks in general and *ConvNet*s specifically a better choice for start-to-end solutions for computer vision tasks. Yu and Zhang [29] achieved state-of-the-art results in 2015 on the EmotiW2015 data set, by using an ensemble of *ConvNet*s having five convolutional layers each and using stochastic pooling rather than max pooling. They randomly perturb the input images to get an extra boost of 2-3% inaccuracy. They applied transformations to the input images at train time. At test time,

their model-generated predictions for multiple perturbations of each test example and voted on the class label to produce a final answer.

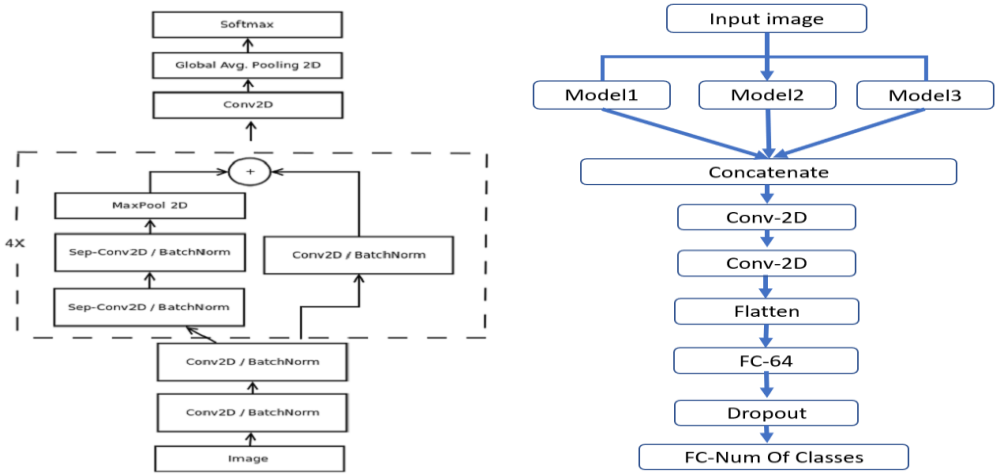
Kim et. al. [12] achieved a height test accuracy on EmotiW2015 by using an ensemble-based method with varying network architectures and parameters. They used a hierarchical decision tree and an exponential rule to combine decisions of different networks rather than simply using a simple weighted average to improve the results. They initialized weights by training networks on other FER data sets and using these weights for fine-tuning. Mollahosseini et. al. [1] have achieved the state-of-the-art results on FER2013 using a *ConvNet* which is consisted of two convolutional layers, max-pooling, and 4 Inception layers as introduced by *GoogLeNet*. The proposed architecture received a low-test accuracy of 47% when tested on the EmotiW2015 data set. Pramerdorfer and Kampel[22], review the state of the art in image-based facial expression recognition using *ConvNet*s, and highlight algorithmic differences and their performance impact and by that identify existing bottlenecks and consequently directions for advancing this research field. Furthermore, they demonstrate that overcoming one of these bottlenecks leads to a performance increase. They used an ensemble of modern deep *ConvNet*s to obtain a test accuracy of 75.2% on *FER2013*.

Saravanan et. al. [25], experimented with several different models, including decision trees and neural networks, and find that *ConvNet*s work better for image recognition tasks since they can capture special features of the inputs due to their large number of filters. They propose a model consists of six convolutional layers, two max-pooling layers, and two fully connected layers. Upon turning off the various hyperparameters, this model achieved a final accuracy of 60%. In [9], the author reviewed the development of FER using VGGNet, ResNet, GoogleNet, and AlexNet tested on *FER2013*. After making some improvements based on the original methods of FER and training on the *FER2013* data set with different revised ways, the best result of accuracy they got is 64.24%.

Burkert et. al. [3], propose a convolutional neural network architecture for facial expression recognition. The proposed architecture is independent of any hand-crafted feature extraction and performs better than the earlier proposed convolutional neural network-based approaches. They tested their system on the standard datasets Extended Cohn-Kanade (CK+) and MMI to achieves an accuracy of 99.6% on (CK+) and 98:63% for MMI. For a comprehensive survey refer to [4,13,22,26]

3. OUR APPROACH

We have used three well-known *ConvNet* models already used for automatic facial expression recognition. Results were close to what has been reported in the literature and a minor improvement has been recorded when we slightly modified their architectures.



(A) (B)

Figure 2: Model (A) is based on the architecture presented in [27] and Model (B) is our final ConvNet combining all the three pre-trained models and the four additional layers.

The novel contribution of the presented research came out when we used a combination of pre-trained versions of these three models to one large *ConvNet*. After combining these models in parallel, we added four additional layers and re-trained the net using two schemes. In the first, we have retrained the net while freezing the weights of the pre-trained layers, and at the second scheme, we retrained the whole net including the weights of the pre-trained layers.

The first scheme has been motivated by the idea of boosting or ensemble of classifiers using the four additional layers to learn the boosting parameters and lead to a 1%-2% improvement of accuracy. The second scheme enabled re-training the pre-trained models in addition to the whole net, and it improved the results by more than 4%. An interesting result of the re-training process came out when the updated re-trained weights of the basic models improved their results while training the whole net. All reported results were conducted on the *JAFFE* [16], (*CK+)* [15], *FER2013* [6], and *FER2013+* [7] data sets and will be presented in details at section 4.

3.1. Face Localization and Detection

Encouraged by results of previous works [9,24], we start facial expression analysis, by detecting and localizing the face in the given image. Locating the face within an image is termed as face detection and localization. As been reported in many papers the first and one of the best options to consider is the one developed in 2004 by Viola and Jones. The method is very fast and could rapidly detect frontal view faces by applying the AdaBoost learning algorithm on a simple class of features. The Authors, achieve excellent performance by using novel methods that could compute the features very quickly and then rapidly separate the background from the face [21].

The Viola-Jones algorithm uses five patterns to extract Haar-like features which are assumed to hold all the information needed to characterize a face. The number of the resulted Haar-like features is huge; therefore, the use of the integral image technique allows us to calculate them at a very low computational cost. To make sense of these features which can be seen mostly as week classifiers, the Ada-boost [8] algorithm is used to generate a strong and accurate classifier based on a small set of the week classifiers. Additional use of the Ada-boost method enables generating cascade classifiers, which produces a fast rejection mechanism of non-face areas efficiently. Many ***24 x 24*** images of faces are used to train and obtain a face detection algorithm

in real-world images quickly and efficiently. In the proposed system, we have used the Viola-Jones algorithm to detect the exact window of the main part of the face and then locate the eyes and mouth within that face. In the next step we have normalized the face image to include the face from the limits of the eyes horizontally, and eyes to mouth positions vertically. We crop the target window based on previous calculations from the image and re-size back to the original size.

3.2. Convolutional Neural Network Architectures and Pre-Trained Models

In this section, we present the three models we have used and combine to generate the proposed system. The input to these models is a ***W x H*** grey level images, where the width ***W*** and the height ***H*** are derived from the size of images used in the specific benchmark. In all models we have used Data Augmentation (*DA*), to generate more samples for the training set by applying transformations such as rotation, crop, shifts, shear, zoom, flip, reflection, and normalization. This is required usually to increase the size and the variability of the training set when it is not enough to learn data representations. We also have used *L2* regularization of the weights to apply penalties on layer parameters during optimization which are incorporated in the loss function that the networks optimize. Batch normalization has been used after each layer to normalize the activation of the previous layer at each batch, to maintain the mean activation close to zero and the activation standard deviation close to one. This practice acts as a regularizer to handle the problem of internal shift co-variation.

The first model we have used is a deep ConvNet with very few free parameters when compared to other deep models in the area. The model presented in [27] is motivated by the idea of reducing the number of free parameters usually exist as weights in the final fully connected layers. This architecture combines the deletion of the final fully-connected layers and the inclusion of the combined depth-wise separable convolutions and residual modules, and by that could speed up the process of training. It uses Average Pooling having the same number of feature maps as the number of classes in the last convolution layer and the soft-max activation function, which enables completely removing the fully connected layers, see *Model(A)* in Figure 2. This architecture is a standard fully convolutional neural network composed of 9 convolution layers. Using Global Average Pooling reduces each feature map into a scalar value by taking the average over all elements in the feature map to force the network to extract global features from the input image. This architecture exchanges the 2-d convolutional layers with depth-wise separable convolutions which are composed of two different layers depth-wise convolutions and point-wise convolutions. This architecture succeeds to reduce the number of the weights within the ***ConvNet*** to approximately 600,000 parameters comparing to more than 140 million in the VGG-16 ConvNet and achieved an accuracy of 66% on the *FER-2013* data set [27].

VGG-16 is one of the state-of-the-art architectures for convolutional neural networks. This architecture consists of 16 weight layers that include 13 convolution layers followed by three fully connected layers. All layers use a ***3x3*** filter size of one pixel for stride and padding. The convolutional layers are divided into 5 groups and each group is followed by a max-pooling layer. In the VGG-16, each group includes a Convolutional layer with several filters and ends with a Max-pooling layer carried out over a 2***x2*** window with stride 2. The number of filters of each convolutional layer is multiplied by 2 when moving from one group to the next and starts from 64 in the first group and 512 in the last group, where all these features are of the size 3***x3***. An additional group of three fully connected layers, which have most of the weights of the net comes after the first 5 groups. The first two layers in this group have 4096 nodes each, and the third contains seven/eight channels (one for each class), The first two layers use *ReLU* activation function, and the last one uses the *SoftMax* activation function for the final classification, See Model (C) in Figure 3 for the full architecture.

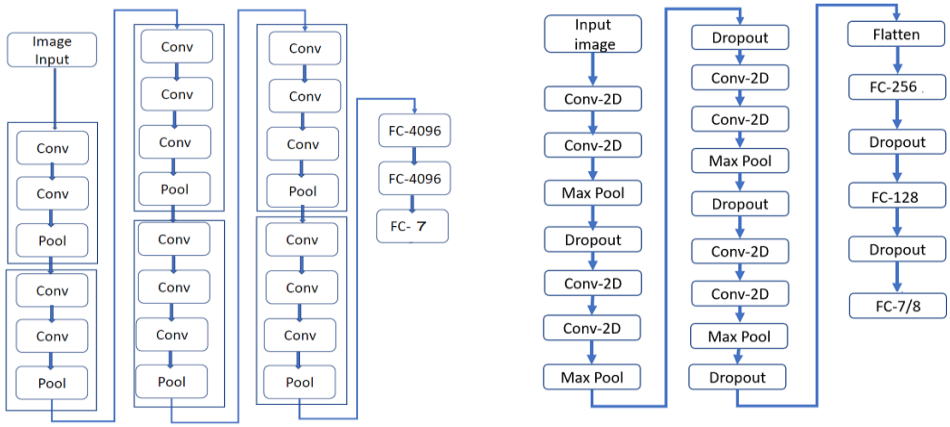


Figure 3: Model (C) is based on the VGG16 ConvNet Architecture and Model (D) is our modified version of VGG16 with less free parameters.

The third system was inspired by the VGG architecture and consists of 11 layers with close to 1,460,000 free trainable parameters. It has 4 Groups of two Conv-2D layers followed by the max-pooling layer. The number of filters of the convolutional layers in each group is multiplied by 2 when moving from one group to the next and starts from 32 in the first one, leading to that the last group reaches 256 filters of 3***x3*** size. After the fourth group comes a group of three fully connected layers which had most of the weights of the net where the first two have 256 and 128 nodes, and the third contains seven/eight channels (one for each class) using *ReLU* and *Softmax* activation functions for the final classification, See Model (D) in Figure 3 for the full architecture.

The full system uses the three models as basic building components and combines them to one large convolutional neural network. The output of each model is a flatten layer with ***N*** nodes derived from the number of classes (***N*** is the number of classes which is 7 and 8 for FER and FER+ data sets receptively). The output layers of the three models having the same size are concatenated to a three channels layer. Two Convolutional layers follow the concatenation layer and the final two layers are fully connected layers with 64 and ***N*** nodes respectively, See Model (B) in Figure 2 for the full architecture.

4. DATA SETS AND EXPERIMENTAL RESULTS

To evaluate our system, we have used four standard benchmarks, the (JAFEE), the (CK+), and two versions of the FER2013 benchmark. Because of the inaccurate labeling problem of the (FER2013) images reported in [29] and other papers, we have used the re-tagged version of (FER2013) named (FER2013+) [7,29], see Figure 4 for examples of such labeling errors. The JAFFE [16] data set contains 213 images of Japanese females, collected by Komachi and Yoba at Kyushu University, Japan. Ten subjects were asked to pose several different facial expressions, where pictures were taken, through remote control while looking towards the camera. Original images have been rescaled and cropped such that the eyes are roughly at the same position with a resolution of ***256x256*** pixels. The number of images corresponding to each of the 7 categories of expression is roughly the same, few of them are shown in Figure 1 (B). The second benchmark is the (CK+) database [15] which contains labeled image sequences for 123 subjects, where each sequence has one of 6 expressions (contempt replaced disgust, natural is the first image in each sequence), see Figure 1 (A).

For each image sequence, only the last frame (the peak frame) is provided with an expression label. Some of the subjects did not have all the 6 expressions, so the final number of subjects we have used for this case was only the 10 subjects from whom we have all the 6 expression categories. Three images from the last five frames were extracted for training/testing purposes from each image sequence in order to expand the data set. Two images of the first two frames from each sequence were extracted and labeled as neutral expressions. To evaluate the presented approach, we have generated data sets for JAFFE and CK+ benchmarks taking out all images of all subjects. In the next step, we have divided them randomly into three sets which we have used for training, validation, and testing.

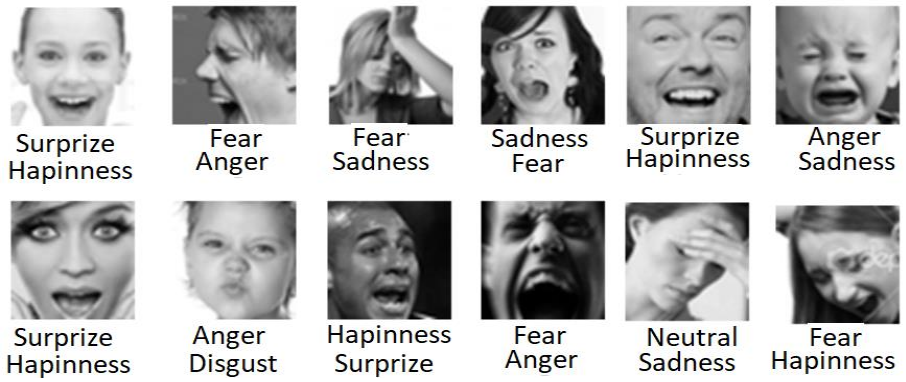


Figure 4: Samples of images from the FER2013 benchmark with re-tagged labels. The upper labelling is the original FER2013 labelling and the lower is the FER2013+ labelling.

The *FER2013* data set [6], is provided by the Kaggle community website and consists of about 37000 greyscale images of faces with the size ***48x48***. The images are pre-processed and registered so that the face is centred and occupies about the same amount of space in each image. Each image is categorized into one of the seven classes that express different facial emotions. The data set is divided into three different sets with the sizes, 29000, 4000, and 4000 images for training, validation, and testing, respectively. The images in the original FER data set was filtered and labeled by human labelers with emotion-related keywords, but the label accuracy is not very high [11], see a few such examples in Figure 4. The authors in [7,29], re-tag the *FER2013* data set using crowdsourcing. For each input image, they asked crowd taggers to label the image into one of 8 classes while adding the 'contempt' emotion as an additional one to the existing 7 classes. The taggers are required to choose one single emotion for each image and the gold standard method has been adopted to ensure the tagging quality.

Table 1: The confusion matrix with accuracy rates for each emotion category using the CK+.

Results of the CK+ data set							
Class	Neutral	Surprise	Anger	Disgust	Fear	Happiness	Sadness
Neutral	98.8 %	0.0 %	0.2 %	0.6 %	0.2 %	0.2 %	0.0 %
Surprise	0.0 %	98.5 %	0.7 %	0.0 %	0.5 %	0.3 %	0.0 %
Anger	0.9 %	0.0 %	95.5 %	0.6 %	0.9 %	0.6 %	1.5 %
Disgust	0.0 %	0.0 %	1.7 %	97.8 %	0.0 %	0.3 %	0.2 %
Fear	0.3 %	0.6 %	1.6 %	2.2 %	94.3 %	0.4 %	0.6 %
Happiness	0.1 %	1.1 %	0.0 %	0.0 %	0.7 %	98.1 %	0.0 %
Sadness	0.1 %	0.1 %	2.2 %	0.0 %	3.5 %	0.0 %	94.1 %
Average				96.81 %			

Ten taggers were asked to label each image, thus obtaining a distribution of emotions for each face image. They randomly chose 10000 images from the data set and assume that the majority of the 10 labels are a good approximation to the ground truth labels. When they have fewer taggers, they compute how many of the majority agree with the ground truth emotion and show that when there are 3 taggers, the agreement is merely 46%. With 5 taggers, the accuracy improves to about 67% and, with 7 taggers, the agreement improves to above 80%. they concluded that the number of taggers has a high impact on the final label quality [29]. With 10 annotators for each face image, they generate a probability distribution of emotion capture by the facial expression, which enables experiment to be held with multiple schemes during training (Categorical and Probability).

Table 2: The confusion matrix with accuracy rates for each emotion category using JAFFE.

Results of the JAFFE data set							
	Neutral	Surprise	Anger	Disgust	Fear	Happiness	Sadness
Neutral	98.5 %	0.0 %	0.2 %	0.8 %	0.3 %	0.2 %	0.0
Surprise	0.2 %	98.2 %	0.4 %	0.0 %	0.6 %	0.5 %	0.1 %
Anger	0.5 %	0.0 %	97.8 %	0.7 %	0.2 %	0.3 %	0.5 %
Disgust	0.0 %	0.0 %	1.0 %	98.4 %	0.0 %	0.3 %	0.3 %
Fear	0.2 %	0.6 %	0.4 %	0.4 %	96.7 %	0.2 %	0.5 %
Happiness	0.5 %	0.7 %	0.0 %	0.0 %	0.6 %	98.2 %	0.0 %
Sadness	0.6 %	0.3 %	0.2 %	0.0 %	0.5 %	0.1 %	98.3 %
Average				96.28 %			

Table 3: The confusion matrix with accuracy rates for each emotion category using the FER.

Results of the Fer2013 data set							
	Neutral	Surprise	Anger	Disgust	Fear	Happiness	Sadness
Neutral	74.1 %	2.2 %	1.2 %	1.3 %	8.6 %	3.5 %	8.3 %
Surprise	2.1 %	84.1 %	1.5 %	1.6 %	1.7 %	6.9 %	2.1
Anger	7.1 %	0.0 %	72.2 %	0.0 %	8.9 %	1.7 %	10.1 %
Disgust	4.0 %	2.1 %	9.1 %	73.2 %	5.4 %	3.9 %	2.3
Fear	2.4 %	12.2 %	1.0 %	3.2 %	71.1 %	0.0 %	10.1 %
Happiness	3.1 %	2.2 %	3.1 %	0.1 %	3.1 %	84.3 %	3.1
Sadness	9.1 %	1.7 %	8.2 %	0.0 %	8.7 %	1.5 %	70.8 %
Average				74.4 %			

Table 4: The confusion matrix with accuracy rates for each emotion category using the FER+.

Results of the FER+ data set								
	Neutral	Surprise	Anger	Disgust	Fear	Happy	Sad	contempt
Neutral	88.7%	2.1 %	0.0 %	0.1 %	3.5%	4.8 %	.08 %	0.0%
Surprise	6.9 %	87.1 %	0.0 %	1.8 %	0.7%	2.5 %	1.0 %	0.0%
Anger	6.9 %	0.0%	88.1%	0.0 %	0.0%	3.6 %	0.4 %	0.0%
Disgust	17.8%	3.2 %	21.8 %	55.9%	0.4%	1.0 %	0.0 %	0.0%
Fear	5.8 %	25.1 %	4.0 %	2.2 %	56.2%	0.0 %	6.7 %	0.0%
Happy	3.6 %	3.2 %	0.4 %	0.0 %	0.0%	92.8%	0.0 %	0.0%
Sad	15.7%	1.2 %	3.0 v	0.0 %	2.6%	0.8 %	73.7%	0.0%
	23.1%	1.8%	15.0 %	0.0 %	4.2 %	0.9 %	15.2%	39.8%
Average				85.1 %				

The second scheme enables training all weights of the large net and increases the accuracy rates by more than 4%. When training only the weights of the additional four layers results were 2% lower. It is important to notice, that when we used the second scheme for training, the accuracy rates of the model (D) ascended from 70.3% to 72.5% which did not happen with many rounds of direct training of this model. The same results with modest improvement (only 2.2%) were achieved on the FER+ data set with a final accuracy rate of 85.1%. As we also can see in Table 4, the proposed method performs well for the emotions Neutral, Happiness, Surprise, Sadness, and Anger, and worst for the remaining emotions. On the other hand, the data set have very few examples of these emotions and mostly with less quality. The total accuracy result are higher than the results achieved on the original *FER2013* data set due to the right re-tagging and the small size of the misclassified categories.

5. CONCLUSIONS

In this research, we present a novel architecture that parallelly aggregates three different ConvNets followed by additional four layers. This approach is a novel way of boosting existing architectures and by using the two-phases process of training, it enables better and faster conversion of the learning step. The two-phase training process increases the accuracy of the complete model by 3.5% on average and the accuracy of the basic models by 2% on average. In the scope of future work, we plan to target the same approach using a combination of a different number of the same basic models with different sizes and layer numbers. Using the same training process, we anticipate an improvement in the learning process in terms of speed and accuracy.

REFERENCES

[1] D. Chan A. Mollahosseini and M. H. Mahoor. Going deeper in facial expression recognition using deep neural networks. The 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY., pages 1–10, 2016.

[2] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd sourced label distribution. In Proceedings of the 18th ACM International Conference on Multimodal Interaction, ICMI 2016, pages 279–283, New York, NY, USA, 2016. Association for Computing Machinery.

[3] Peter Burkert, Felix Trier, Muhammad Zeeshan Afzal, Andreas Dengel, and Marcus Liwicki. Dexpression: Deep convolutional neural network for expression recognition, 2015.

[4] Claude Chibelushi and Fabrice Bourel. Facial expression recognition: A brief tutorial overview 2002.

[5] Rania Salah El-Sayed, Ahmed El Kholy, and Mohamed Youssri ElNahas. Robust facial expression recognition via sparse representation and multiple gabor filters. International Journal of Advanced Computer Science and Applications, 4(3), 2013.

[6] FER2013. Challenges in representation learning: Facial expression recognition challenge facial-expression-recognition-challenge. http://www.kaggle.com/c/challenges-inrepresentationlearning, 2013.

[7] FERPlus. Fer plus emotion label. https://github.com/Microsoft/FERPlus, 2016.

[8] Yoav Freund and Robert E. Schapire. A decisiontheoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 55(1): pages 119– 139, 1997.

[9] Yijun Gan. Facial expression recognition using convolutional neural network. pages 1–5, 08 2018.

[10] G.U.Kharat and S.V. Dudul. Emotion recognition from facial expression using neural networks. Human computer systems interaction advances in intelligent and soft computing, 2009.

[11] Goodfellow I.J., Lee M., Hirose A., Hou ZG., and Kil R.M. (eds). Challenges in representation learning: A report on three machine learning contests. Neural Information Processing, Lecture Notes in Computer Science, 8228, 2013.

[12] Bo-Kyeong Kim, Jihyeon Roh, Suh-Yeon Dong, and Soo-Young Lee. Hierarchical committee of deep convolutional neural networks for robust facial expression recognition. Journal on Multimodal User Interfaces., 10: pages 173–189, 2016.

[13] S. Li and W. Deng. Deep facial expression recognition: A survey. IEEE Transactions on Affective Computing., pages 1–1, 2018.
[14] Lin. Facial expression recognition based on geometric features and geodesic distance. International Journal of Signal Processing, Image Processing and Pattern Recognition, 7(1): pages 323–330, 2014.
[15] J. F. Cohn Lucey, J. Saragih T. Kanade, Z. Ambadar, and I. Matthews. The extended cohn-kanade dataset (ck+): A complete expression dataset for action unit and emotionspecified expression. In CVPR Workshops, pages 94–101, 2010.
[16] M. Lyons, J. Budynek, and S. Akamatsu. Ieee trans. pattern analysis and automatic classification of single facial images. Machine Intelligence, 13: pages 252–263, 1991.
[17] Alka Gupta M. and L. Garg. A human emotion recognition system using supervised self-organizing maps. In IEEE International Conference on Computing for Sustainable Globle Development (INDIACOM), pages 654– 659, 2014.
[18] Jharna Majumdar and Ramya Avabhrith. Human face expression recognition. International Journal of Emerging Technology and Advanced Engineering Website, 4(7): pages 559–565, 2014.
[19] S. Mariooryad and C. Busso. Exploring cross-modality affective reactions for audiovisual emotion recognition. Affective Computing, IEEE Transactions on, 4(2): pages 183–196, April 2013.
[20] Ekman P. and Friesen W. Unmasking the face: A guide to recognizing emotions from facial expressions. Consulting Psychologists press, palto Alta, CA, 1975.
[21] Viola Paul and Jones Michael. Robust real-time face detection. International Journal of Computer Vision, 57(2): pages 137–154, 2004.
[22] Christopher Pramerdorfer and Martin Kampel. Facial expression recognition using convolutional neural networks: State of the art. ArXiv, 1612.02903, 2016.
[23] R. Rosenthal. Conducting judgment studies: Some methodological issues. The New Handbook of Methods in Nonverbal Behavior Research, 2008.
[24] R. Saabni. Facial expression recognition using multi radial bases function networks and 2-d gabor filters. In 2015 Fifth International Conference on Digital Information Processing and Communications (ICDIPC)., pages 225–230, 2015.
[25] Akash Saravanan, Gurudutt Perichetla, and K. S. Gayathri. Facial emotion recognition using convolutional neural networks. CoRR, abs/1910.05602, 2019.
[26] Yingli Tian, Takeo Kanade, and Jeffrey Cohn. Facial Expression Recognition, chapter 19, pages 487–519. 2011.
[27] Matias Valdenegro-Toro, Octavio Arriaga, and Paul Pl¨oger. Real-time convolutional neural networks for emotion and gender classification. In 27th European Symposium on Artificial Neural Networks, ESANN 2019, Bruges, Belgium, 2019.
[28] Z.-L. Ying, Z.-W. Wang, and M.-W. Huang. Facial expression recognition based on fusion of sparse representation. Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence, pages 457–464, 2010.
[29] Z. Yu and C. Zhang. Image based static facial expression recognition with multiple deep network learning. In ICMI 2015: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction., pages 435– 442, 2015.
[30] S. Zafeiriou and M. Petrou. Nonlinear non-negative component analysis algorithms. IEEE T-IP, 19(4): pages 1050–1066, 2010.

AUTHORS

Raid Saabni is a senior lecturer and researcher at the school of computer science in the Academic College of Tel-Aviv Yafo and the TRD Center. He received his BSc in Mathematics and Computer Science in 1989 and his MSc and Ph.D. in computer science from the Ben-Gurion University of the Negev in 2006 and 2010, respectively. His research interests are: Machine learning, Computer vision, Historical Document Image Analysis, Handwriting Recognition, Image Retrieval, and Image and Signal Processing.

Alon Schcalar is a Senior Lecturer at the School of Computer Science at the Academic College of Tel-Aviv Yafo. He received his BSc, MSc (Summa Cum Laude) and PhD in Computer Science from Tel Aviv University. He co-authored over 30 papers in leading scientific journals and conferences. His areas of interest include machine learning (both unsupervised and supervised learning), dimensionality reduction, data mining, ensemble methods, image and signal processing and computer vision.

© 2020 By AIRCC Publishing Corporation. This article is published under the Creative Commons Attribution (CC BY) license.