FACE EXPRESSION RECOGNITION USING CONVOLUTION NEURAL NETWORK (CNN) MODELS

Nahla Nour [1,] Mohammed Elhebir [2] and Serestina Viriri [3]

[1]Computer Science, Sudan University of Science and Technology, Khartoum, Sudan
[2]Department of Computer Science, University of Gezira, Gezira, Sudan
[3] Computer Science, University of KwaZulu-Natal, Durban, South Africa

ABSTRACT

This paper proposes the design of a Facial Expression Recognition (FER) system based on deep convolutional neural network by using three model. In this work, a simple solution for facial expression recognition that uses a combination of algorithms for face detection, feature extraction and classification is discussed. The proposed method uses CNN models with SVM classifier and evaluates them, these models are Alex-net model, VGG-16 model and Res-Net model. Experiments are carried out on the Extended Cohn-Kanada (CK+) datasets to determine the recognition accuracy for the proposed FER system. In this study the accuracy of AlexNet model compared with Vgg16 model and ResNet model. The result show that AlexNet model achieved the best accuracy (88.2%) compared to other models.

KEYWORDS

Deep learning, Alex-Net, Vgg-Net & Res-Net

1. INTRODUCTION

Face is known as the organ which shows emotion, it is the strongest “channel” of non-verbal communication. The face possesses an ability of showing non obvious emotion as signals of facial expressions like a smile face shows happiness, a frown face shows disapproval or sadness, a face with a wide-open eyes shows surprise and a face with a curled lip shows disgust. Recognizing these signals through machines can make human-machine interaction stronger and comported. Facial expression recognition (FER) is an important non-verbal channels through which human’ internal intent and emotions can be recognized by Human Machine Interaction (HMI) systems.

There are several areas where FER is been applied such as human machine interaction as used in security surveillance, computer games and social robot. FER is also applied in behavioural science to get social facts (origin, gender, and age) and also applied in medical science to monitor pain, depression, anxiety and for mental retardation treatment. Despite that humans easily perceive most facial expressions, but there is still difficulty in getting reliable expression recognition by machine. The main difficulties in FER are obtaining optimum pre-processing, feature extraction and classification, most especially under variable conditions of input data, the head pose, environmental disorder and illumination, various causes of variation in face. Problems remain even when deep learning is applied to FER despite its feature learning ability. Firstly, many training data are required by deep neural networks to be free from over fitting. Nevertheless, the subsisting facial expression databases are insufficient for training the common

DOI:10.5121/ijgca.2020.11401

recognized neural network with deep framework which derived the highest hopeful accurate rate in tasks for recognition of objects. Moreover, various personal details such as background, gender, age, ethnic and level of expression allow high inter-subject variation to exist [2]. In addition to subject identity bias, variations in pose variability, occlusions and lighting effect which are known in unrestricted scenarios of face expression. They are all not connected with facial expressions longitudinally and consequently provide strength to the deep network necessity to overcome variation in the large intra-class and to train active expression precise representing symbols. In this study, we institute and present developments in study on an introduced method and technique for feature extraction, face detection and classification which utilizes VGG-19, Alex-Net and Reset-Net architecture to improve FER performance and solve the mentioned difficulties. Deep learning methods gave a considerable results when used for feature extraction and classification, specifically Convolutional Neural Networks (CNN) frameworks that are biologically motivated multistage one which automatically trained hierarchies of constant features [3]. The ConvNets include a multistage input image processing for hierarchical extraction and high-rank representing feature. This inspired us to come up in this study with an efficient approach method that depends on ConvNets for recognition of facial expression. We propose a latest framework in which the input of the system is seen as the image; after that, Convolutional Neural Network models are used for predicting the facial expression label that has to be among the following [4]: neural, happiness, anger, surprise, sadness, and disgust. And these models are evaluated. We organized the other sections of this study as follows: Section 2 presents a general background on researchers preceding tasks applying Alex-Net and Reset-Net for recognition of face expression, Section 3 gives detailed description of the proposed method and architecture, Section 4 provides the results of the experiment, evaluates and discusses the results in detail. In conclusion, Section 5 gives the concluding part of the paper.

2. FACE EXPRESSION BACKGROUND

Deep learning is a kind of machine learning in which a model learns for carrying out classification function direct from text, images, or sound. Deep learning is applied using a neural network architecture. The word deep indicates the number of layers in the network, more layers increase the depth of the network. Deep learning algorithm performs a repeated task, each time tweaking it a little to increase the result, so that train’s computer system to perform what comes to humans naturally: learning from examples [5]. Deep learning is the main technology for several applications like driverless cars, it enables them to distinguish pedestrians and to recognize a stop sign [6]. Recently, for a good reason deep learning has been getting lots of attention and giving results that were impossible before [4]. Deep learning models could get an accurate state-of-the-art that occasionally surpasses human ability. They have been trained by using a large number of identified data as well as neural network frameworks with number of layers [4]. Deep learning models have been trained by using large of identified data and neural network frameworks which automatically learn features from the data then manual extraction of features is not needed [7].

2.1. Convolution Neural Network

Convolutional neural networks are presently among the best flamboyant algorithms for deep learning with image data. They have been successfully applied in computer vision tasks, and the robustness in object recognition localization in variant images is proven by the results. Thorough research on automatic analysis of expression was publicized recently [8], [9], [10]. These publications have established a set of standard algorithmic pipelines for FER. Though, their concentration is on classic methods, and deep learning has seldom been reviewed. In the recent time, FER has been examined according to deep learning in [11], but this is a very short review without introductions on FER datasets and technical details on deep FER. Hence, we carry out in this study, a systematized study on deep learning for FER duty that depend on videos (image sequences) and non-moving image. Aiming to provide new researchers on this field a general description of the systematized architecture and high skills of deep FER. It consists of input and output layer. Convolutional layers, fully-connected layers and max-pooling layers are examples of in-between layers. CNN architectures vary in the number and type of layers applied for its particular application. The most common CNN architectures are GoogLeNet [12], VGGNet [13], AlexNet [14], ResNet [15].

2.1.1. Convolutional Layer

Convolutional layer has always been the first layer of CNN, it extracts features, and it consists of several feature maps. Each neuron in a feature map is connected to a small region, called the local receptive field, through a set of shared weights and a single shared bias. The main two advantages and reasons why Convolutional layer is preferable over fully connected layer are; firstly, parameter sharing; where all neurons share equal weights and bias in a feature map, which causes fast training as a result of a high reduction in the number of parameters, and will eventually help in building deep networks [16].

2.1.2. Pooling Layer

After convolutional layer, the pooling layer is usually implemented with the aim of reducing the spatial resolution of the feature maps, speeding the computation and extracting prominent features. Max-pooling is the most used technique for pooling layer, where each pooling unit is equal to the largest element in a receptive field.

2.1.3. Fully Connected (FC) Layer

The fully connected layer input must be a vector, so we have to first flatten the output features from the two layers (convolutional and pooling). Then, we may pass them to the output layer, where a Soft-Max classifier or sigmoid are used for predicting the input class label. Besides these networks, there are also many common derived architectures. In [17], [18], region-based CNN (RCNN) [19] was utilized for learning features for FER. In [20], Faster R-CNN [105] was employed to signify facial expressions through generation of proposals of a high class region. In addition, 3D CNN was introduced by Ji et al. [21] to acquire motion information encoded in various close structures for recognizing actions through 3D convolutions. The well-designed C3D was proposed by Tran et al, [22] which utilizes 3D convolutions on large-scale supervised training datasets to learn patio-temporal features. Several related researches ([23] [24]) have used this network for FER.

2.2. Related Work

Recently, a significant improvement has been made by researchers in developing expression classifiers [25], [26], [27]. Many deep learning techniques which function on data representing symbols, such as Convolutional Neural Networks (CNNs) have been developed in order to acquire better facial expression representation. There is similarity between the CNN used in [28] and the used one for building the FACS model in terms of concept. The input is broken down into features and the deeper the layer the harder the representation becomes, which builds on the former layer. Then the last feature representations are utilized for classification. CNNs are not similar to the FACS model [29] as a matter of fact that, if enough samples are shown, the network may possibly learn an overall smile symbolizing rather than using a specific and predefined structure symbolizing. In the year 2015, Zhang and Yu employed CNN for FER in EmotiW. They used a complete unit of CNNs [30] and discomposed the input images randomly to get a 2-3% increase in accuracies. Kahou et al. [31], used CNNs to evaluate two experiments, where they pre-trained a standard CNN in the first experiment employing the Acted Facial Expression in the Wild (AFEW) dataset [32] while the second one was held on Toronto Facial Dataset. Both experiments were used to recognize a neural expression and six universal facial expressions. This paper has proven an achievement that increased network accuracy. Dennis H. et. al [33], applied a facial image to CNN in his task, he combined the information from the network to achieve 94.4% accurate recognition. The method achieved a higher recognition rate compared to others. We propose in this study a common CNN framework models to predict common facial image expressions. This approach combines standard methods such as Viola Jones method for face and facial region detection and extraction [34], [35] and convolutional neural network (Alex-Net and Reser-Net) for feature extraction and image sequences are involved. For classification, we used Support Vector Machine to compare each of the two models with the state- of the- art.

3. METHODOLOGY

This section presents the suggested method used in this study. Which is organized in three main phases: Data-pre-processing, feature extraction, classification steps for recognizing face expression. Figure 1 illustrate our methodology

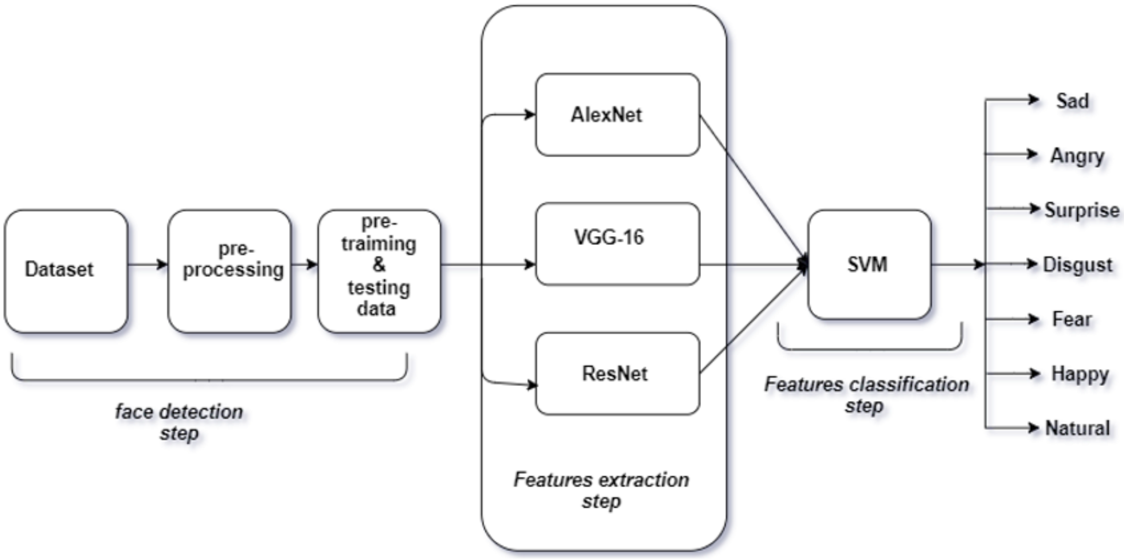


Figure 1: Research Methodology

3.1. Face Expression Dataset

The CK+ dataset contains 593 image sequences from 123 subjects within the age of 18 to 50, including male and females from different ethnic groups comprising of 69% female and 19% from African, American and Asians [36]. Ck+ Dataset image sample is illustrated in Figure 1

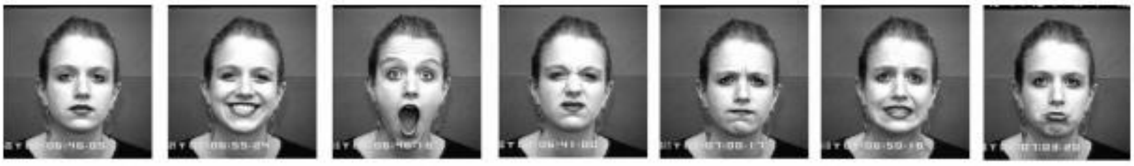


Figure 2 A sample of images from the CK+ dataset

The proposed method begins with pre-processing and uses Viola-Jones face detection framework for face detection. The face was first detected, cropped and face components were extracted and normalized 224 x 224 x 3 pixel size for Res-Net50 and Res-Net101 CNN Models, also 128 x128 pixels for Alex-Net CNN model. When image scale is reduced it helps in reducing a must learned information by the network and also ensures faster training and reduces memory cost [1]. The detected face components were applied as the input into the first layer of CNN to make them compatible with the input size of CNN Pre-Trained Models and convert any grayscale images to RGB images using an augmented Image Data store to resize image and convert image to RG

3.2. Prepare Training and Test Image Sets

Separate the sets into two, the separation should be randomized to avoid being bias. Pick 80 % of the selected images from each set for the training data and the remaining 20 %, for the test data.

3.3. CNN Pre-Trained Models

In this paper, we applied four models CNN pre-trained convolution neural networks as the proposed facial expression recognition methods, which are Alex-Net, VGG, ResNet-101 and ResNet-50 and compared the four methods with state of the art methods. A pre-trained model is a kind of model which has been trained on a large standard dataset to address a problem similar to the one we aim to address. The process diagram of the proposed facial expression recognition system is shown in Fig.1 it is arranged in three stages; Image Pre-Processing where we used Viola Jones algorithm for detecting Face and facial parts, facial Feature extraction and feature classification using CNN.

3.3.1. Alex-Net:

It was the winning architecture for ImageNet competition. It uses large filter and stride size at the first layer. It uses 8-layer network, with an image input size of 48 by 48.

3.3.2. VGG-19

Our VGG-19 has an input of 48x48 RGB image. The image is sent via a bunch of convolution layers, where 3x3 filters are utilized. It contains 19 weight layers which involves 16 layers of convolution with 3x3 filter size and 3 layers of fully connected, followed by stack of convolutional layers. Each first two of 3 fully connected layers have 4096 channels, the third performs 7-way ILSVRC classification and therefore consists of 7 channels (every class has a channel). SVM layer is the last layer. Fig. shows Vgg-19 networks.

3.3.3. ResNet-101

Research VGG ResNet-101 is a convolutional neural network which has undergone training on over one million images within ImageNet database. It contains 347 layers and has an ability to categorize images into 1000 object classes. It possesses 224 by 224 as size of an image input.

3.3.4. ResNet-50

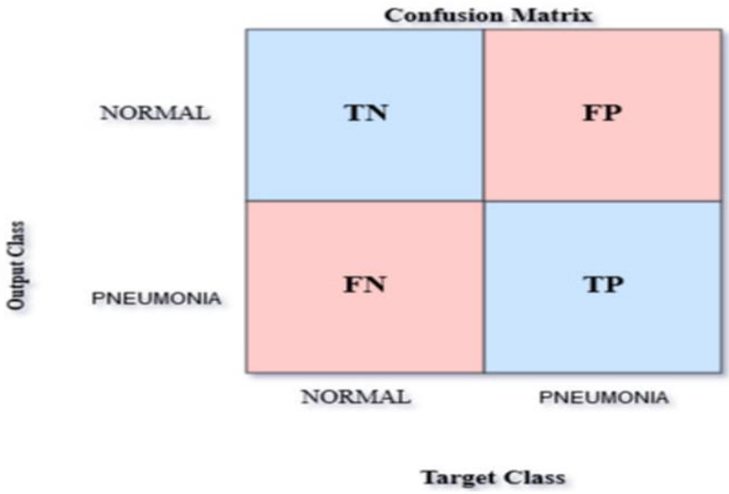
ResNet-101 is a convolutional neural network which has undergone training on over one million images within ImageNet database. It contains 177 layers and can categorize images into 1000 object classes. It also has 224 by 224 as size of an image input.

3.4. Train a Multiclass SVM Classifier using Pre-trained CNN

There was an extraction of the learned image features from a pre-trained CNN for feature extraction, the precedent layer to the classification layer called fc1000 was used for feature extraction through activation method. These features were latterly utilized in training and testing SVM classifier.

3.5. Evaluation Protocol

A method performance in facial expression image analysis is typically measured by specificity, sensitivity, and F1 score. Sensitivity measures the proportion of real positive samples which are correctly recognized, in which the percentage of Pneumonia image that is correctly classified as Pneumonia. Therefore, it is computed according to the following definition:



Confusion Matrix

Output Class		
NORMAL	TN	FP
PNEUMONIA	FN	TP
	NORMAL	PNEUMONIA

Target Class

Figure 3 depicted confusion matrix.

True Positives (TP): The number of cases where the model predicted yes and the person have the predicted expression, True Negatives (TN): the number of cases where the model predicted no and the person does not have that expression. False positives (FP): Model predicted yes, but actually they do not have the predicted expression. False Negatives (FN): Model predicted no, but actually they have the predicted expression Figure 3 depicted confusion matrix.

Sensitivity measures the proportion of actual positives samples that are correctly identified, in which the percentage of expression that is correctly classified as correct one. Therefore, it is computed according to the following definition:

$$\textbf{Sensitivity} = \frac{TP}{TP+FN}\ \textbf{TP}\backslash\ \textbf{(TP+TN)}\frac{TP}{TP+FN} \qquad \textbf{Eq. 1}$$

Where TP (true positive) is the number of images which have been successfully detected, and FN (false negative) is the number of images which have been detected by the method. In contrast, specificity measures the proportion of identified negatives samples, in which the percentage normal images correctly classified as normal. In this manner, specificity is computed as:

$$\text{Specificity} = TP \backslash (TP+FN)\ \frac{TN}{TP+FN} \qquad \textbf{Eq. 2}$$

Where TN (true negative) is the number of normal patients which have been successfully classified, and FP (false positive) is the number of normal patients which have been wrongly classified as Pneumonia. And, F1-score measures the average F1 score through different class labels which is computed as:

$$\mathbf{F1= (PPV \times TPR) \backslash\ (PPV+TPR)} \qquad \textbf{Eq. 3}$$

Where PPV

$$\mathbf{PPV = TP \backslash\ (TP+TN)}$$

And TPR

$$\mathbf{TPR =}\ \frac{TP}{TP+FN}\ \mathbf{TP \backslash\ (TP+TN)}$$

The accuracy was the fraction of the predicted labels that were correctly computed as:

$$\mathbf{Accuracy\ = TP\text{-}TN \backslash\ (TP+TN+FP+FN)} \qquad \textbf{Eq. 4}$$

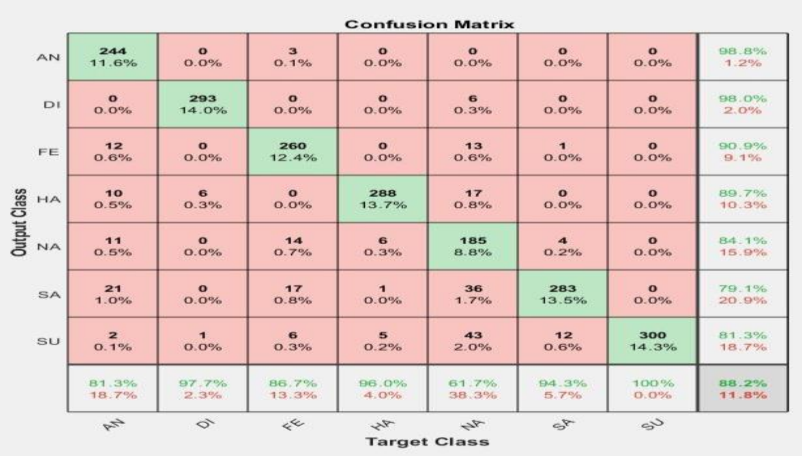
Table 1. Comparison of results with approaches in metric measurements

	Sensitivity	Specificity	F1-score	Accuracy
Alex-Net	88.2 %	88.1%	89.3%	88.2%
VGG-16	83.7%	82.9%	85.1%	88.2%
ResNet-101	81.4%	81.2%	82.6%	81.6%
ResNet-50	81.1%	80.9%	80.4%	81.9%

4. EXPEREMENT RESULT AND DISCUSSION

The proposed system operates on Intel Core i5- CPU @ 2.7 GHz with 8GB. MATLAB 2019 a device was used for the method evaluation and performed the task of classification and feature selection. The training subset of 70% was used to train the network for classification and 30% for testing subset was used to test how probable a facial image is correspond to a specific class of facial expression.

CNN training is accomplished in a supervised method applying the standard across-validation algorithm. The average recognition accuracy is used for evaluation of network performance. An average of 0.1s was taken by image pre-processing while an average of 0.2s was given to image classification. Figure 2, 3, 4 & 5 show the confusion matrix of the recognition accuracies for seven facial expressions with the uses of the training weights achieved with maximum accurate rate. The recognition rate obtained when using Alex-net, VGG-Net, Res-Net101, Res-Net50 methods is 99.3%, 98.4%, 99.1%, 97.9% respectively.



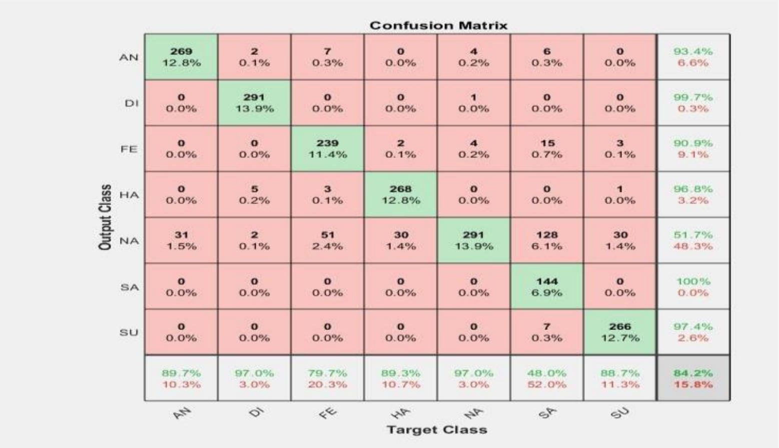
Confusion Matrix

Output Class

	AN	DI	FE	HA	NA	SA	SU	
AN	244 11.6%	0 0.0%	3 0.1%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	98.8% 1.2%
DI	0 0.0%	293 14.0%	0 0.0%	0 0.0%	6 0.3%	0 0.0%	0 0.0%	98.0% 2.0%
FE	12 0.6%	0 0.0%	260 12.4%	0 0.0%	13 0.6%	1 0.0%	0 0.0%	90.9% 9.1%
HA	10 0.5%	6 0.3%	0 0.0%	288 13.7%	17 0.8%	0 0.0%	0 0.0%	89.7% 10.3%
NA	11 0.5%	0 0.0%	14 0.7%	6 0.3%	185 8.8%	4 0.2%	0 0.0%	84.1% 15.9%
SA	21 1.0%	0 0.0%	17 0.8%	1 0.0%	36 1.7%	283 13.5%	0 0.0%	79.1% 20.9%
SU	2 0.1%	1 0.0%	6 0.3%	5 0.2%	43 2.0%	12 0.6%	300 14.3%	81.3% 18.7%
	81.3% 18.7%	97.7% 2.3%	86.7% 13.3%	96.0% 4.0%	61.7% 38.3%	94.3% 5.7%	100% 0.0%	88.2% 11.8%

Target Class

Figure 4. Confusion matrix (%) of cross-validation testing method for 7-classes emotional in CK dataset using Alex-Net



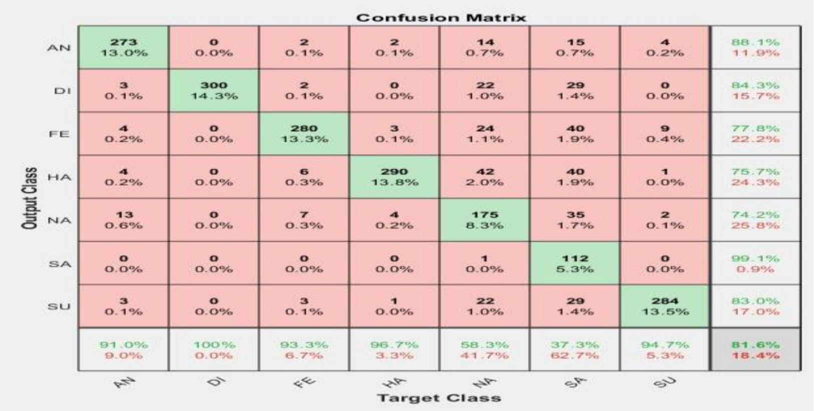
Confusion Matrix

Output Class

	AN	DI	FE	HA	NA	SA	SU	
AN	269 12.8%	2 0.1%	7 0.3%	0 0.0%	4 0.2%	6 0.3%	0 0.0%	93.4% 6.6%
DI	0 0.0%	291 13.9%	0 0.0%	0 0.0%	1 0.0%	0 0.0%	0 0.0%	99.7% 0.3%
FE	0 0.0%	0 0.0%	239 11.4%	2 0.1%	4 0.2%	15 0.7%	3 0.1%	90.9% 9.1%
HA	0 0.0%	5 0.2%	3 0.1%	268 12.8%	0 0.0%	0 0.0%	1 0.0%	96.8% 3.2%
NA	31 1.5%	2 0.1%	51 2.4%	30 1.4%	291 13.9%	128 6.1%	30 1.4%	51.7% 48.3%
SA	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	144 6.9%	0 0.0%	100% 0.0%
SU	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	7 0.3%	266 12.7%	97.4% 2.6%
	89.7% 10.3%	97.0% 3.0%	79.7% 20.3%	89.3% 10.7%	97.0% 3.0%	48.0% 52.0%	88.7% 11.3%	84.2% 15.8%

Target Class

Figure 5. Confusion matrix (%) of cross-validation testing method for 7-classes emotional in CK dataset using VGG-16



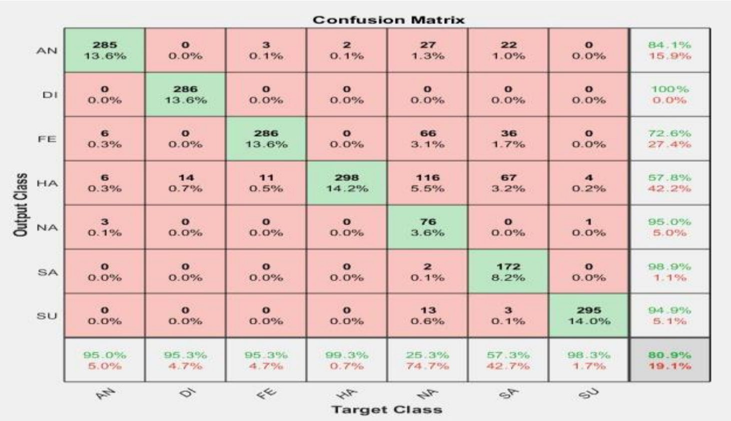
Confusion Matrix

Output Class

	AN	DI	FE	HA	NA	SA	SU	
AN	273 13.0%	0 0.0%	2 0.1%	2 0.1%	14 0.7%	15 0.7%	4 0.2%	88.1% 11.9%
DI	3 0.1%	300 14.3%	2 0.1%	0 0.0%	22 1.0%	29 1.4%	0 0.0%	84.3% 15.7%
FE	4 0.2%	0 0.0%	280 13.3%	3 0.1%	24 1.1%	40 1.9%	9 0.4%	77.8% 22.2%
HA	4 0.2%	0 0.0%	6 0.3%	290 13.8%	42 2.0%	40 1.9%	1 0.0%	75.7% 24.3%
NA	13 0.6%	0 0.0%	7 0.3%	4 0.2%	175 8.3%	35 1.7%	2 0.1%	74.2% 25.8%
SA	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 0.0%	112 5.3%	0 0.0%	99.1% 0.9%
SU	3 0.1%	0 0.0%	3 0.1%	1 0.0%	22 1.0%	29 1.4%	284 13.5%	83.0% 17.0%
	91.0% 9.0%	100% 0.0%	93.3% 6.7%	96.7% 3.3%	58.3% 41.7%	37.3% 62.7%	94.7% 5.3%	81.6% 18.4%

Target Class

Figure 6. Confusion matrix (%) of cross-validation testing method for 7-classes emotional in CK dataset using ResNet-101



	AN	DI	FE	HA	NA	SA	SU	
AN	285 13.6%	0 0.0%	3 0.1%	2 0.1%	27 1.3%	22 1.0%	0 0.0%	84.1% 15.9%
DI	0 0.0%	286 13.6%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
FE	6 0.3%	0 0.0%	286 13.6%	0 0.0%	66 3.1%	36 1.7%	0 0.0%	72.6% 27.4%
HA	6 0.3%	14 0.7%	11 0.5%	298 14.2%	116 5.5%	67 3.2%	4 0.2%	57.8% 42.2%
NA	3 0.1%	0 0.0%	0 0.0%	0 0.0%	76 3.6%	0 0.0%	1 0.0%	95.0% 5.0%
SA	0 0.0%	0 0.0%	0 0.0%	0 0.0%	2 0.1%	172 8.2%	0 0.0%	98.9% 1.1%
SU	0 0.0%	0 0.0%	0 0.0%	0 0.0%	13 0.6%	3 0.1%	295 14.0%	94.9% 5.1%
	95.0% 5.0%	95.3% 4.7%	95.3% 4.7%	99.3% 0.7%	25.3% 74.7%	57.3% 42.7%	98.3% 1.7%	80.9% 19.1%

Figure 7, Confusion matrix (%) of cross-validation testing method for 7-classes emotional in CK dataset using ResNet-50

We compared the average recognition accuracy of our proposed methods with other facial expression recognition methods. The comparison of the performance accuracy achieved with the proposed method and the other methods on the CK+ database are shown in Table 2. It was proven that the proposed methods achieved a higher recognition accuracy comparing to other existing methods in literature.

Table 2: Comparison with methods in literature

Methods	
AlexNet[37]	61.7%
VGG-19 [38]	76.73%
ResNET [37]	63.1%
Ours AlexNet	88.2%
Ours VGG-19	84.2%
Ours ResNet	81.6%

Generally, our method explicitly inherits the advantage of information gathered from frontal image using different models of CNN and working with hidden layers, and hence it naturally improves the final predication accuracy. The disadvantage of our approach is that facial expression recognition based on face images can achieve promising results, facial expression is only one modality in realistic human behaviours.

5. CONCLUSION

We used deep learning to examine the VGG-19, AlexNet and ResNet architectures for facial emotion recognition. We succeeded by achieving satisfactory results in CK+ dataset as the results demonstrated. We further improved these models, for AlexNet model and SVM classifier have highest accuracy (88.2%), followed by VGG16 model with (84.1%) accuracy and followed by AlexNet model with (81.6%) accuracy. Finally the result of this study can help in facial expression recognition.

Work can be improve through the uses of other datasets or if we use another deep learning architecture so with a different approach, the result can be improved. In addition to the use Data augmentation techniques to increase the efficiency of training and use larger datasets to increase accuracy and finally design an online application to recognize the expression from real image.

REFERENCES

[1] C. Ruoxuan, L. Minyi, and L. Manhua Facial Expression Recognition Based on Ensemble of Multiple CNNs, CCBR 2016, LNCS 9967, pp. 511-578, Springer International Publishing AG 2016.M. F. Valstar, M. Mehu, B. Jiang, M. Pantic, and K. Scherer,

[2] Metaanalysis of the first facial expression recognition challenge,IEEE Transactions on Systems, Man, and Cybernetics, Part B(Cybernetics), vol. 42, no. 4, pp. 966979, 2012.

[3] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE, vol. 86, no. 11, pp. 2278 2324, 1998

[4] P. Ekman and D. Keltner, Universal facial expressions of emotion, California Mental Health Research Digest, vol. 8, no. 4, pp.151158, 1970.

[5] Eddy S anchez-DelaCruz1 and Pilar Pozos-Parra2: Machine learning-based classification for diagnosis of neurodegenerative diseases. Instituto Tecnol ogico y de Estudios Superiores de Occidente, Tlaquepaque, Jalisco, November 15, (2018).

[6] J. Schmidhuber, ”Deep learning in neural networks: An Overview.” Neural Networks, Vol. 61, pp. 85-117, (2015).

[7] Ji, S., Xu, W., Yang, M., Yu, K.: 3D convolutional neural networks for human action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 35(1), 221231, (2013)

[8] Zeng, Z., Pantic, M., Roisman, G.I. and Huang, T.S., 2008. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE transactions on pattern analysis and machine intelligence, 31(1), pp.39-58.

[9] Bettadapura, V., 2012. Face expression recognition and analysis: the state of the art. arXiv preprint arXiv:1203.6722.

[10] Rao, J. and Su, X., 2004, July. A survey of automated web service composition methods. In International Workshop on Semantic Web Services and Web Process Composition (pp. 43-54). Springer, Berlin, Heidelberg.

[11] Giannopoulos, P., Perikos, I. and Hatzilygeroudis, I., 2018. Deep learning approaches for facial emotion recognition: A case study on FER-2013. In Advances in hybridization of intelligent methods (pp. 1-16). Springer, Cham.

[12] Szegedy, C., et al. Going deeper with convolutions. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[13] Simonyan, K. and A. Zisserman, Very deep convolutional networks for large-scale image recognition. ArXiv preprint arXiv:1409.1556, 2014.

[14] Krizhevsky, A., I. Sutskever, and G.E. Hinton. Imagenet classification with deep convolutional neural networks. in Advances in neural information processing systems. 2012.

[15] He, K., et al. Deep residual learning for image recognition. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[16] Michael, A. Nielsen: Neural Networks and Deep Learning. Chapter 6. Determination Press (2015).

[17] B. Sun, L. Li, G. Zhou, X. Wu, J. He, L. Yu, D. Li, and Q. Wei, Combining multimodal features within a fusion network for emotion recognition in the wild, in Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. ACM, 2015, pp. 497 502.

[18] R. Girshick, J. Donahue, T. Darrell, and J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 580587.

[19] J. Li, D. Zhang, J. Zhang, J. Zhang, T. Li, Y. Xia, Q. Yan, and L. Xun, Facial expression recognition with faster r-cnn, Procedia Computer Science, vol. 107, pp. 135140, 2017.

[20] S. Ren, K. He, R. Girshick, and J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, in Advances in neural information processing systems, 2015, pp. 9199.

[21] S. Ji, W. Xu, M. Yang, and K. Yu, 3d convolutional neural networks for human action recognition, IEEE transactions on pattern analysis and machine intelligence, vol. 35, no. 1, pp. 221231, 2013.

[22] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, Learning spatiotemporal features with 3d convolutional networks, in Computer Vision (ICCV), 2015 IEEE International Conference on. IEEE, 2015, pp. 44894497.

[23] Y. Fan, X. Lu, D. Li, and Y. Liu, Video-based emotion recognition using cnn-rnn and c3d hybrid networks, in Proceedings of the 18th ACM International Conference on Multimodal Interaction. ACM, 2016, pp. 445450.

[24] D. Nguyen, K. Nguyen, S. Sridharan, A. Ghasemi, D. Dean, and C. Fookes, Deep spatio-temporal features for multimodal emotion recognition, in Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on. IEEE, 2017, pp. 12151223.

[25] Y. Muttu, H. G. Virani, ”Effective face detection feature extraction neural network based approaches for facial expression recognition”, IEEE International Conference on Information Processing (ICIP), pp. 102-107, Dec 2015.

[26] N. Mousavi, H. Siqueira, P. Barros, B. Fernandes and S. Wermter, ”Understanding how deep neural networks learn face expressions,” in IEEE International Joint Conference on Neural Networks (IJCNN), 2016.

[27] S.A.M Al-Sumaidaee, Facial Expression Recognition Using Local Gabor Gradient Code-Horizontal Diagonal Dedscriptor School of Electrical and Electronic Engineering, Newcastle University, England, UK, 2015.

[28] H.C.Santiago, T.Ren,and G. D.C. Cavalcanti Facial expression Recognition based on Motion Estimation, Neural Networks (IJCNN), 2016 International Joint Conference , Electronic ISSN:2161-4407. 03 November 2016.

[29] H.C.Santiago, T.Ren,and G. D.C. Cavalcanti Facial expression Recognition based on Motion Estimation, Neural Networks (IJCNN), 2016 International Joint Conference , Electronic ISSN:2161-4407. 03 November 2016.

[30] J. Li, and E.Y . Lam Facial expression recognition using deep neural networks,Imaging Systems and Techniques (IST), 2015 IEEE International Conference on, 1-6

[31] Z. Yu and C. Zhang. Image based static facial expression recognition with multiple deep network learning. ICMI Proceedings.

[32] Kahou, Samira Ebrahimi, et al. ”Combining modality specific deep neural networks for emotion recognition in video.” Proceedings of the 15th ACM on International conference on multimodal interaction. ACM, 2013.

[33] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, Collecting Large,Richly Annotated Facial-Expression Databases from Movies, IEEE Multimedia, vol. 19, no. 3, pp. 3441, 2012.

[34] Dennis Hamester et al., Face Expression Recognition with a 2-Channel Convolutional Neural Network, International Joint Conference on Neural Networks (IJCNN), 2015.

[35] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In Proc. of CVPR, 2001

[36] S. Ouellet, Real-time emotion recognition for gaming using deep convolutional network features, arXiv preprintarXiv: 1408.3750, 2014.

[37] Ioffe, S. and Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167.

[38] Fan, Y., Lam, J.C. and Li, V.O., 2018, October. Multi-region ensemble convolutional neural network for facial expression recognition. In International Conference on Artificial Neural Networks (pp. 84-94). Springer, Cham.