JKAU: Comp. IT. Sci., **Vol**. **12** No. 1, pp: 51 – 61 (1444 A.H. / 2023 A.D.)
DOI: 10.4197/Comp.12-1.5

Facial Expression Recognition Based on Well-Known ConvNet Architectures

Taima Alrimy[1], Ahad Alloqmani[1], Abrar Alotaibi[1,2], Nada Aljohani[1] and Salma Kammoun[1]

[1]Computer Science Department, Faculty of Computing and Information Technology
King Abdulaziz University, Jeddah, Saudi Arabia
[2]Computer Science Department, College of Computer Science and Information Technology
Imam Abdulrahman bin Faisal University, Dammam, Saudi Arabia
[1]{ talrimy, asalaamahalloqmani, Naljohani0084}@stu.kau.edu.sa , amotaibi@iau.edu.sa and smohamad1@kau.edu.sa

***Abstract*—The Convolution Neural Network (CNN) is the most widely used deep learning architecture as it has broken most world records for recognition tasks. Facial Expression Recognition (FER) systems that use classical feature-based techniques, especially CNN's, is best for classifying images. This paper used three CNN-based methods, which are VGG-16, Inception-v3, and Resnet50-V2 network architectures, to classify facial expressions into seven classes of emotions: happy, angry, neutral, sad, disgust, fear, and surprise. The face expression dataset from Kaggle and JAFFE dataset were used to compare the accuracy between the three architectures to find the pretrained network that best classifies models. The results showed that VGG-16 network architecture produced a higher accuracy (93% in JAFFE and 54% in Kaggle) than the other architectures.**

Keywords— Facial Expression Recognition (FER), Convolution Neural Network (CNN), VGG-16, Inception-v3, Resnet50-V2

I. INTRODUCTION

One of the commonly used biometric traits is the face. A face recognition system is a biometric artificial intelligence that identifies and verifies a person by detecting and analyzing facial regions. Recently, the human-computer interaction system studies facial expression to recognize human emotion and use it in many applications, like education so teachers could learn if students understood the course from their expressions [1], and in the medical field to recognize patients' expressions as a supporting tool for patient care [2]. Facial expression recognition (FER) plays the primary role in many intelligence systems fields such as artificial intelligence, robotics, and security, as a facial expression can sometimes guarantee the driver's life safety [3]. In computer vision, facial expression recognition is a common and complex research topic because many factors affect facial expressions, such as age, pose, occlusion, and lighting.

A facial expression recognition (FER) system has mainly three phases: preprocessing, feature extraction, and classification [3]. The first phase of FER is preprocessing, which could be used to improve the performance of FER. The primary preprocessing methods are normalization, localization, face alignment, and region of interest (ROI), which use the median filter, viola jones algorithm, scale-invariant feature transform (SIFT) and regulating the face dimensions respectively. The second phase of the FER system is feature extraction. The five types of feature extraction methods are edge method, geometric feature method, texture feature method, patch method, and global and local feature method [3]. The edge-based method uses line edge map (LEM), the dynamic two strip algorithm (Dyn2S), and graphics- processing unit-based active shape model (GASM) descriptors. As for the geometric feature-based method, it uses local curvelet transform (LCT) descriptors, while the texture feature-based method uses local binary pattern (LBP) and Gabor filter descriptors. The patch-based method uses two processes, extract and match the patches. Both the global and local feature-based methods use principal component analysis (PCA) descriptors. The last phase is classification, in which the classifier classifies the expression into classes like happiness, anger, sadness, fear, surprise, neutral, and disgust [4]. There are many classification techniques and methods such as minimum distance classifier (MDC), the KNN (k – Nearest Neighbors) algorithm, support vector machine (SVM), hidden Markov model

(HMM), ID3 decision tree (DT) and convolution neural network (CNN) [5].

Traditional facial expression recognition systems perform feature selection manually and then develop a suitable classifier based on the selected features [1]. With science and technology developments, many researchers proposed other techniques, but most have limitations like the loss of the image's original feature information. A deep learning algorithm reduces the difficulty of extracting features manually as it only takes the original image as input to find the feature from the original data by applying some filtering techniques and learning processes. Then it automatically extracts features through some hidden layers for classification. Recently, deep learning algorithms have broken most of the world records of the recognition tasks; hence, it is considered a vast development in recognition tasks [6]. CNN is the most widely used deep learning architecture which includes two parts: feature extraction (hidden layers: convolution and pooling) and a classifier [1]. The convolution is responsible for generating a feature map by using a filter. After the convolution layer, the pooling layer is performed. The standard pooling method is a max-pooling that takes the most activated value of a feature. In the classifier part, the essential layers of CNN are the fully connected layer and softmax layer [4]. The underlying CNN architecture is illustrated in Fig. 1.

Developing CNN from scratch is a complex task due to the enormous computation power needed because convolution itself requires an expensive operation. Therefore, the concept of transfer learning is used, which allows transferring the learning of other pretrained models to the data. So, popular pretrained models can be used instead of creating a custom neural network to extract features from the selected dataset.

In this paper, three CNN-based methods are applied to classify facial expressions into one of

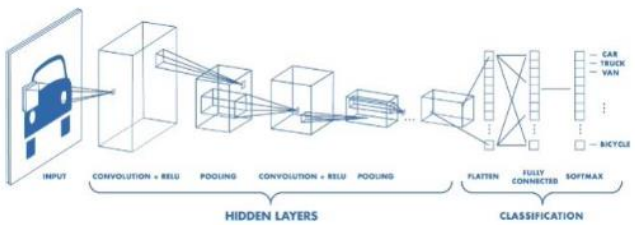

Fig. 1. The basic CNN Architecture [7]

the following seven emotions: happy, angry, neutral, sad, disgust, fear, and surprise using VGG-16, Inception-v3, and Resnet50-V2 network architectures. The facial expression dataset from Kaggle [8] and the JAFFE dataset [9] are used to compare the accuracy between the three architectures to find the pretrained network that best classifies models.

The remaining of the paper is structured as follows. Section 2 represents some of the related works done in the same area. Section 3 introduces the three CNN architectures we used in our work. Section 4 describes the datasets and the experimental results for each architecture used in Section 3. In Section 5, we discuss the results, and Section 6 concludes the paper. Future work is presented in Section 7.

II. Related Works

Many researchers nowadays are working on the topic of facial expression detection using various methods and algorithms. It is proven that CNN is the best technique according to [1] who evidenced in their survey on CNN- based facial expression recognition that using classical feature-based techniques to transact with large quantities of data could waste much time. So, the researchers found that applying deep learning techniques are the best, especially CNNs for classification of images. Also, they presented the overall main issues and their solutions, such as data augmentation and pre- processing.

Starting from 2015, [10] proposed a CNN-based approach and compared it with two baseline approaches by using two feature extraction methods (LBP and SIFT). They tested the two feature-based approaches (LBP + SVM, SIFT + SVM) on the extended Cohn-Kanade dataset (CK+) [11] and compared them with other CNN-based approaches. They observed that their proposed approach is very effective with 83% accuracy. They also offered a deep CNN architecture with a facial detection model and two improved frameworks, which are likelihood loss and hinge loss. They achieved an 80% accuracy in six classes of emotions (sad, happy, angry, surprise, fear, disgust). At the same time, [12] presented their main contribution, which is the

segmentation of a face into two regions, which achieved high performance at a low cost. The accuracy achieved by their proposed approach was 99%.

In 2016, [13] proposed several compact CNNs that solve facial expression recognition problems. They assembled several subnets (3 different CNN) separately. They scored 65.03% accuracy on the FER2013 dataset [14] in seven classes of emotions (sad, happy, angry, surprise, fear, disgust, neutral). Moreover, [15] proposed a novel approach for facial expression recognition in which each pairwise classifier uses a particular subset. Their proposed approach outperformed existing methods and achieved a high 98.91% accuracy requiring less complexity.

Furthermore, [16] described a part-based hierarchical bidirectional recurrent neural network (PHRNN) and a multi-signal convolutional neural network (MSCNN) for extracting facial features from extract temporal and spatial features, respectively. Also, [17] offered a novel peak piloted deep network (PPDN) for facial expression recognition with a peak gradient suppression (PGS)method for network optimization and achieved a 99.3% accuracy. In addition, [6] proposed a CNN architecture for recognizing facial expressions that used automatic facial expression recognition. It has a wide range of applications like human-computer interaction and safety systems, because nonverbal cues are necessary forms of communication and play an important role in interpersonal communication. The proposed architecture claim is independent of any hand-crafted feature extraction. The researchers used the CKP dataset and the MMI Database to evaluate whether the architecture is competitive. The offered architecture proved to be very effective on the dataset with an average accuracy higher than state-of-the-art approaches that preceded this approach.

Later in 2017, [18] used CNN (with 9 layers) with the viola jones algorithm for facial detection; their method of real- time emotion recognition reached an accuracy around 90% on FER2013 dataset in seven classes of emotions. Besides, [19] proposed a novel I2CNN approach for facial expression recognition based on multi-scale global images and local facial patches. Their proposed I2CNN approach achieved significant performance on the CK+ dataset by 98.3% accuracy. Moreover, [20] proposed a new local pattern called local directional ternary pattern (LDTP) for feature extraction using support vector machines (SVM) to classification facial expressions. Their approach outperforms other current state of the arts based on their experimental results.

In 2018, researchers continued to improve the recognition process, as shown in [21], which proposed a human-centric convolutional neural network (HCCNN) architecture which learns specific facial regions independently. They combined the output with support vector machine (SVM) and achieved 93.3% accuracy on the CK+ dataset. [2] also applied extract facial features from deep convolutional neural networks (CNN) in deep learning and classified the facial expression using softmax classifier. The proposed algorithm for facial expression recognition is based on depth volume and network. It takes the facial expression image as the input to the CNN and trains the CNN network. Then uses the trained network to recognize facial expressions. The researchers used the Japanese Female Facial Expression Database (JAFFE) expression dataset and the CK+ dataset to evaluate the effectiveness of the algorithm. The results of the experiment showed that this algorithm performed facial expression recognition better than the traditional methods. The researchers encountered some limitations as there were not enough samples for training, as a result, CNN was unable to quickly find links between samples as artificially defined features. Overall, the experiment showed promising results in classifying various expressions, but some still need further training and improvement. Furthermore, the purpose of [22] was to use a method based on CNN to classify each facial image as one of the seven facial expressions: angry, disgust, neutral, sad, fear, surprise, and happy. They designed a new convolutional neural network structure based on facial expression recognition characteristics. A convolution kernel was used to extract implicit features, and max pooling was used to reduce the

dimensions of the extracted implicit features. They built the structure of the neural network model to detect the main facial expression automatically. The architecture that was used in this research is AlexNet network. The datasets FER and CK+ were used to perform the experiments. This research presented a comparison between the accuracy of facial expression classification using SVM and DCNN. It was evidenced that DCNN provides better classification accuracy (accuracy of SVM is 76%, and of DCNN is 88%).

Recently in 2019, [23] proposed a CNN average weighting method to recognize real-time facial expressions that improves the robustness of the recognition process. The researchers took into consideration the problem of high-speed camera capturing and image characteristics changes that may occur in real- time systems. Therefore, they offered a method that refers to the previous image for averaging instead of the immediate output to facilitate recognition and reduce the interference from the image's characteristics. The experiment was performed in a fixed environment using a computer and a webcam. The results indicated that the accuracy and robustness of facial expression recognition were significantly improved compared to the results of applying only the convolution neural network (CNN). Also, the objective of [24] was to discover ways to improve the current facial expression recognition system. So, they proposed a probable solution and that developed a facial expression recognition system based on a convolutional neural network with data augmentation. The methodology of this research was using CNN with data augmentation, and the dataset that was used was gathered from different datasets. As a result, the proposed model was not biased to any specific dataset. The proposed model in this study outperformed any other existing model in terms of accuracy (96.24 %). This model was successful in keeping a nearly equal and higher recognition rate for each class.

Table 1 summarizes all the related work described above.

TABLE I. RECENT STUDIES OF FACIAL EXPRESSION RECOGNITION

Ref.	Year	Method	Dataset	Accuracy
[10]	2015	Feature-based CNN (LBP and SIFT)	CK+	83
[12]	2015	CNN (segmentation of face into two regions)	KDEF	99
[13]	2016	3 different CNN subnets	FER2013	65.03
[15]	2016	LBP, WLD, Pairwise classifier	JAFFE, CK	98.91
[16]	2016	Hierarchical bidirectional recurrent neural network (PHRNN)	MMI	81.18
[17]	2016	A novel peak piloted deep network (PPDN)	CK+	99.3
[6]	2016	CNN architecture	CKP and MMI	99.6
[18]	2017	CNN (with 9 layers) with viola jones algorithm	FER2013	90
[19]	2017	Novel deep learning-based framework (I2CNN method)	CK+	98.3
[20]	2017	Local directional ternary pattern (LDTP)	CK+	94.19
[21]	2018	A human-centric Convolutional Neural Network (HCCNN)	CK+	93.3
[2]	2018	CNN using Softmax classifier	JAFFE and CK+	87.2
[22]	2018	CNN using AlexNet network	FER and CK+	88
[23]	2019	CNN average weighting method	Real-Time	-
[24]	2019	CNN with data augmentation	Combined datasets	96.24

III. TECHNICAL DESCRIPTION

This paper uses three CNN-based network architectures which are VGG-16, Inception-v3m and Resnet50-V2. In all network architectures, the hybrid parameters, which are learning rate, batch size and epoch, are changed and adjusted to improve the accuracy of the results corresponding to the selected dataset.

A. VGG-16 Network Architecture

VGG-16 pretrained neural network is a network well trained on massive datasets. It has a 224 x 224 RGB image input to the cov1 layer. The image is passed across convolutional stack (conv) layers, where the filters were used with a minimal receptive field: 3x3 (which is the smallest size to capture the notion of right/left, down/up, center). Also, one of the configurations utilizes a 1x1 convolution filter, the convolution stride is set to 1 pixel, and the resolution of spatial padding of conv layer input is preserved after convolution. Five max-pooling layers carry out spatial pooling, and max- pooling is performed,

with stride two over a 2x2 pixel window. Three layers of fully- connected (FC) follow a convolutional stack of layers; each of the first two layers have 4096 channels while the third performs 1000-way ILSVRC classification and contains 1000 channels. The last layer is the soft-max layer. The FC-layer configuration is similar in all networks. All the hidden layers are equipped with the rectification (ReLU) non-linearity. Local response normalization (LRN) is not present in any network in which normalization increases memory consumption and computation time [25]. Regrettably, there are two significant drawbacks with VGGNet:

1) *It is excruciatingly slow to train.*
2) *The network architecture's weights are extremely large (in terms of disk/bandwidth).*

VGG16 is larger than 533MB due to its depth and number of FC nodes, making it difficult to deploy. VGG16 is used in various deep learning classification problems. It is an excellent learning building block because it is easy to implement; nevertheless, smaller network architectures are often preferred [25]. Fortunately, a VGG- 16 trained network is available in Keras. The weights of the image-net network are used [26]. Fig. 2 illustrates the VGG-16 network. The VGG-16 network contains several convolutional layers supported by some FC dense layers and a softmax output layer for the classification. Combining features from the convolutional layers is the dense layers' responsibility, and this helps in the final classification. So, the main idea is re-using the pretrained network, so when the VGG-16 network is used on the dataset, all dense layers must be replaced, and another dense layer and a dropout layer are added to avoid overfitting [27].

B. *Inception-v3 Network Architecture*

The Inception network is different from the other networks that came before it by improving the computing resources utilization inside the network. This was accomplished by designing it in such a way that the network's depth and width could be increased while the computational budget remained constant. Its improvement caused the creation of several versions of the network, such as Inception v1, Inceptionv2, Inception v3, Inception v4, and Inception-ResNet [29]. The goals of the network are to avoid representational bottlenecks, mainly at early stages in the network; increasing the activations per tile in a convolutional network allows for clearer features. Also, spatial aggregation is done on lower- dimensional embeddings with little loss in representational power and balance width and depth of the network. Every iteration of the network improves upon factorizing convolutions, auxiliary classifier, and efficient grid size reduction [29]. Inception Net v3 introduced the use of the following:

1) *RMSProp optimizer.*
2) *Factorized 7x7 convolutions.*
3) *In the auxillary classifiers, use BatchNorm.*
4) *Label smoothing.*

As shown in Fig. 3, the Inception v3 network stacks 11 inception modules, each of which is made up of pooling layers and convolutional filters with rectified linear units (ReLU) as activation function. The model's input is two-dimensional images of 16 horizontal brain sections placed on 4x3x4 grids generated during the preprocessing step. To the final concatenation layer, two fully connected layers of sizes 265 and 128 are added. As a means of regularization, a dropout with a rate of 0.6 is used before the FC layers. The model is fine-tuned after being pre-trained on the ImageNet dataset using the Adam optimizer with a batch size of 30, epoch size of 75, and learning rate of 0.001 [29].

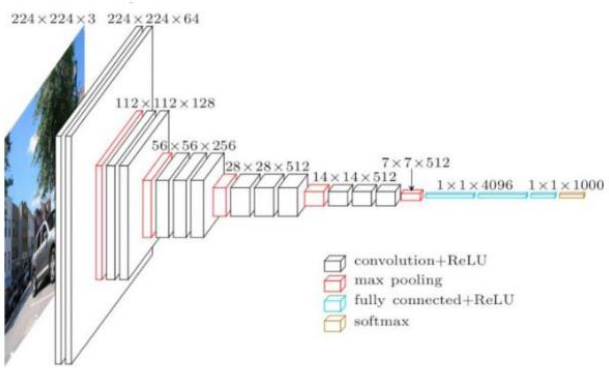

Fig. 2. VGG-16 Network Architecture [28]

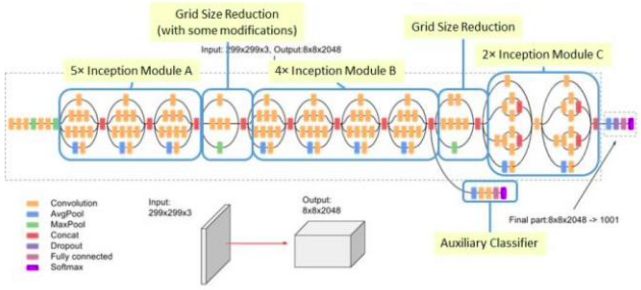

Fig. 3. Inception-v3 Network Architecture [29]

C. *Resnet50-V2 Network Architecture*

ResNet-50 is a CNN-based network architecture trained on over a million images from the ImageNet database. It has an image input size of 224-by-224 and can learn deep feature representations for a broad range of images. The network is able to classify images into 1000 categories and is 50 layers deep [30]. It first introduces the concept of skip connection. This works with the network stacking convolution layers one after the other; it creates a shortcut the goes through the convolution block, skipping a few layers of the stack and adding the original input to the convolution block's output [30].

The goal of the network is to flow information from earlier layers in the model to later ones, and the architecture is used to pass information from the down-sampling layers to the up-sampling layers without saturating the accuracy.

ResNet50 introduced the use of skip connection, V2 employs batch normalization prior to each weight layer, which is the main difference between ResNetV2 and the original (V1).

ResNet50V2 network uses five stages of modules, each of which is made up of pooling layers and convolutional filters with rectified linear units (ReLU) as activation function. To the final concatenation layer, three FC layers of sizes 512, 265, and 128 are added. As a means of regularization, a dropout with a rate of 0.6 is used before the FC layers. The model is fine-tuned after being pre-trained on the ImageNet dataset using the Adam optimizer with a batch size of 30, epoch size of 75, and learning rate of 0.001. Fig. 4 shows the architecture of the Resnet50-V2 network.

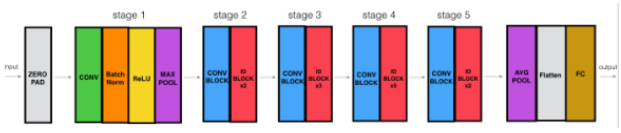

Fig. 4. Resnet50-V2 Network Architecture [30]

IV. Experimenal Results

In this section, we present the datasets used in this paper. The first one is the facial expression dataset from Kaggle [8], which is made up of grayscale images of faces 48x48 pixels in size. Each image represents a different face expression (0=angry, 1=disgust, 2=fear, 3=happy, 4=sad, 5=surprise, 6=neutral).

The dataset contains approximately 36K images [31]. The other dataset is JAFFE [9] which has seven expressions for ten Japanese females. It includes 220 images with 256x256-pixel resolution.

The two datasets were split into train, validation, and test subsets. The JAFFE dataset was spliced using cross-validation function from SKlearn, which split the data into 60% for training, 20% for validation, and 20% for testing While the Kaggle dataset was split by Kaggle online. The hyperparameters (epoch, batch size, and learning rate) were manipulated and changed until the best results were achieved.

The models were evaluated by computing the accuracy and loss-function of the training, validation, and testing subsets of the dataset. Multiple values for the hyperparameters were tested. For epoch size: 5, 10, 20, 60, 75 and for batch size: 7, 8, 16, 20, 30, 32, 60. As for the learning rate, the adaptive Adam optimizer was used, and the learning rates of 0.001 and 0.0001 were tested.

In the JAFFE dataset experiment, the best results were achieved when the epoch was set to 30, batch size to 7, and learning-rate to 1e-5. From the result of the JAFFE dataset, it is shown that the VGG-16 network had the best performance, achieving an accuracy of 93%, and a loss of 0.3, followed by InceptionV3 with accuracy of 85%.

The last network is ResNet50-V2, with an accuracy of 72%. Fig. 5 and Fig. 6 demonstrate the train and validation accuracy and loss for the VGG-16 network in the JAFFE dataset. Fig. 7 demonstrates the classification report and Fig. 8 illustrates the confusion matrix of VGG-16 in the JAFFE dataset, which show the score of accuracy for each of the seven classes. Fig. 9 shows the output of VGG-16 in the JAFFE dataset.

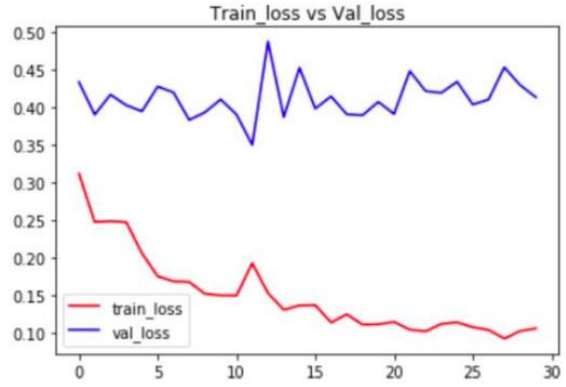

Fig. 5. Loss of the VGG-16 network in JAFFE dataset

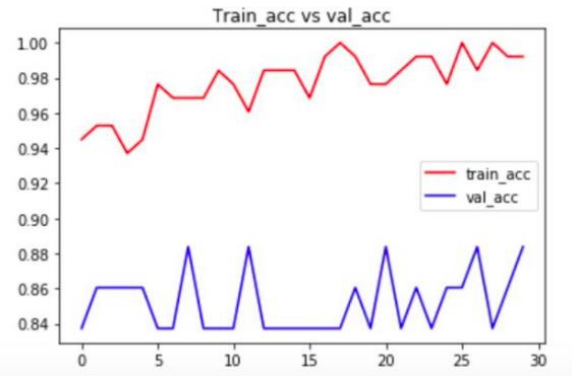

Fig. 6. Accuracy of the VGG-16 network in JAFFE dataset

```
[INFO] evaluating network...
              precision    recall  f1-score   support

       ANGRY       1.00      1.00      1.00         4
    DISGUST        1.00      0.77      0.87        13
        FEAR       1.00      1.00      1.00         7
       HAPPY       1.00      1.00      1.00         7
     NEUTRAL       0.67      1.00      0.80         2
         SAD       0.71      1.00      0.83         5
    SURPRISE       1.00      1.00      1.00         5

    accuracy                           0.93        43
   macro avg       0.91      0.97      0.93        43
weighted avg       0.95      0.93      0.93        43
```

Fig. 7. The classification report of VGG-16 in JAFFE dataset

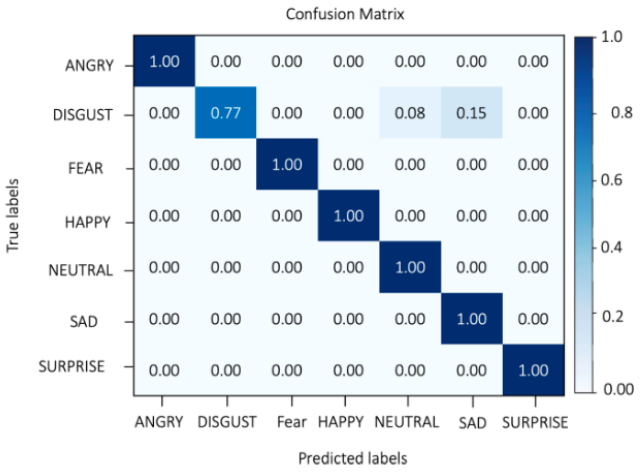

Fig. 8. Confusion matrix of VGG-16 in JAFFE dataset

In the Kaggle dataset experiment, the best results were achieved when the epoch was set to 75, batch size to 60 for training 30 for validation, and learning-rate to 1e-5. From the result of the Kaggle dataset, it showed that the VGG- 16 network was the best performance, which achieved an accuracy of 54%, followed by InceptionV3 with accuracy 50%. The last network is ResNet50V2, with an accuracy of 49%. Fig. 10 demonstrates the train and validation accuracy and loss for the VGG-16 network in the Kaggle dataset. Fig. 11 demonstrates the classification report and Fig. 12 illustrates the confusion matrix of VGG-16 in the Kaggle dataset, which show the score of accuracy for each of the seven classes. Fig. 13 shows the output of VGG-16 in the Kaggle dataset. Finally, the results proved that the VGG-16 network architecture is best compared to the other network architectures (InceptionV3, ResNet50V2) on the Kaggle and JAFFE datasets. The accuracy comparisons between the applied networks on the two datasets are illustrated in Fig. 14 and Fig. 15.

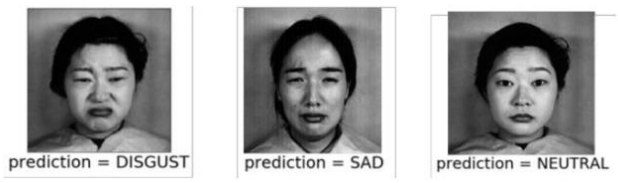

Fig. 9. Some output of VGG-16 in JAFFE dataset

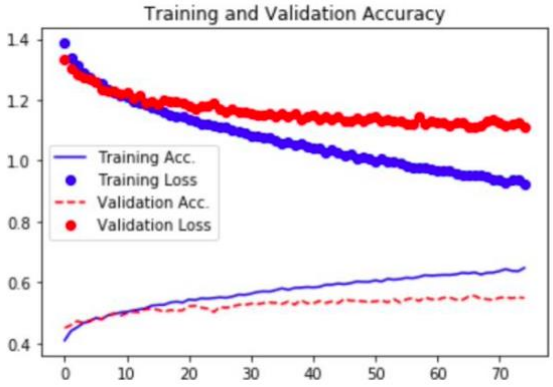

Fig. 10. Train and Validation accuracy value of VGG-16 in Kaggle dataset

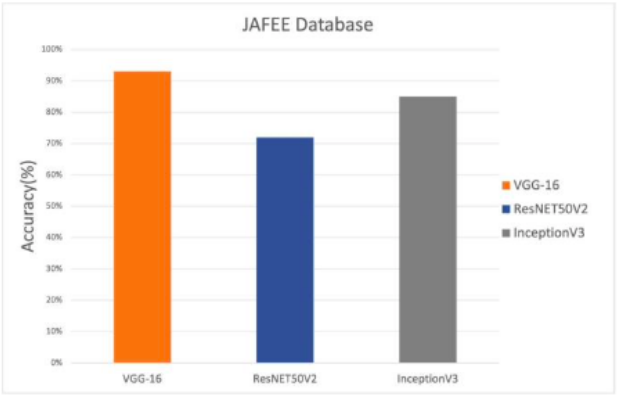

Fig. 14. JAFFE dataset accuracies

```
[INFO] evaluating network...
Classification Report
              precision    recall  f1-score   support

       angry       0.58      0.48      0.53        93
        fear       0.50      0.32      0.39       103
     neutral       0.58      0.61      0.60        82
         sad       0.39      0.67      0.50        85
    surprise       0.81      0.70      0.75        79

    accuracy                           0.54       442
   macro avg       0.57      0.56      0.55       442
weighted avg       0.57      0.54      0.54       442
```

Fig. 11. The classification report of VGG-16 in Kaggle dataset

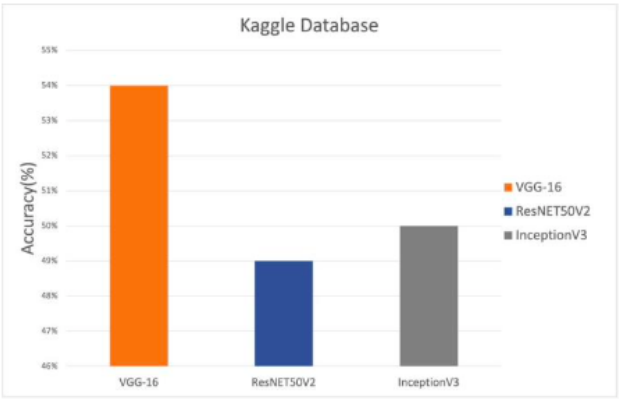

Fig. 15. Kaggle dataset accuracies

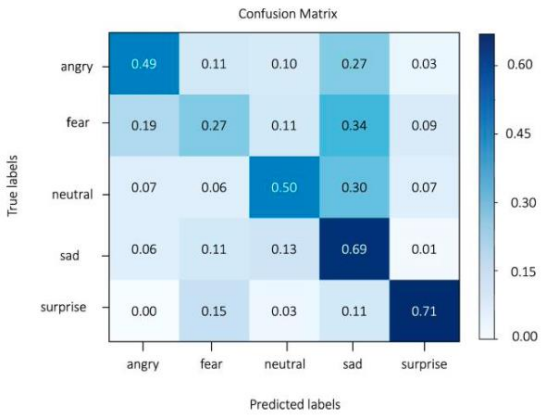

Fig. 12. Confusion matrix of VGG-16 in Kaggle dataset

V. Discussion

In the experiments, all seven classes in the JAFFE dataset were used as for the Kaggle dataset; only five classes were used, which are angry, fear, neutral, sad, and surprise. The two remaining classes were discarded because they caused an irredeemable imbalance between classes. The class size ranged between 3800-4200 images per class, but the 'disgust' class contained around 400 images, while the 'happy' class contained 7800 images. Different methods to solve the imbalance issue were performed. Still,

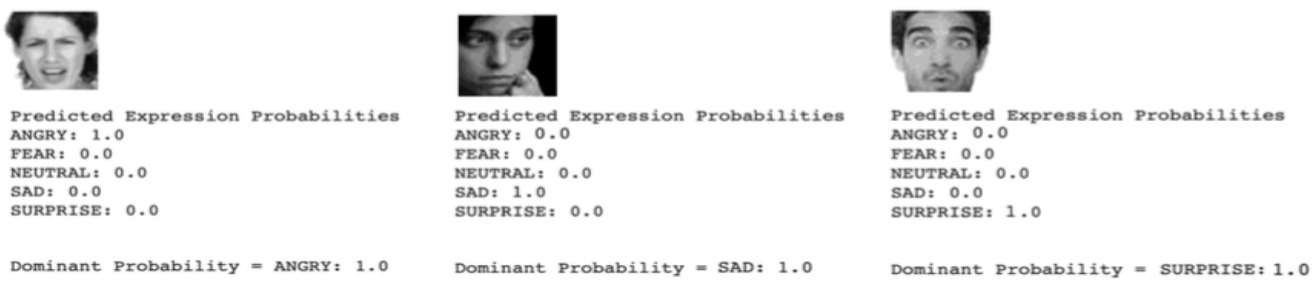

Fig. 13. Some output of VGG-16 in Kaggle dataset

none proved to be a good solution, such as giving a different weight to each class that affects the amount of information learned from each class to give more attention to information gained from classes with a lower number of images. The solution did not provide the required performance, and since solving the imbalance by collecting more data was outside the scope of the course project, the two classes with major imbalances were discarded. The effect of the imbalance on the dataset was clear when the winner of the Kaggle competition for this dataset only scored 0.71 accuracy, and all the competitors scored 0.69 and lower. Another possible reason for the low accuracy is that the size of the images is too small (48x48), which results in bad image quality, and the input size of VGG-16 is 224x224. For InceptionV3, it is 96x96, and for ResNet50V2, it is 96x96. So, it is possible that when the size was increased, the image may have lost some features. On the other hand, the JAFFE dataset is well-arranged, and the size of images is 256x256, which is close to the input size of the applied networks. Hence, the accuracy was far better in the JAFFE dataset compared to the Kaggle dataset. The biggest encountered challenge was the lack of computational power since the data is huge and required massive computations. Even the Google colab has limited GPU access, which caused slow progress.

VI. Conclusion

In conclusion, facial expression recognition (FER) represents the primary goal of many intelligence systems, and it has mainly three phases which are preprocessing, feature extraction, and classification. This paper applied three CNN techniques, which are VGG-16, Inception-v3, and Resnet50- V2 network architectures. The facial expression dataset from Kaggle and the JAFFE dataset were used to evaluate and compare the accuracy between the three architectures and to find the pretrained network that best classifies the images. Many challenges were encountered while implementing the models, such as lack of computational power and poor quality of images in the dataset. The results showed that VGG-16 network architecture has a higher efficiency compared to the other architectures.

VII. Future Work

Due to a lack of time, many different tests and experiments have been postponed, like using data augmentation to improve the accuracy and trying another network architecture. Also, future work concerns looking deeper into rearranging and modifying the Kaggle dataset to get the best accuracy. Moreover, the research can be extended by recognizing micro expressions from the eyes and mouth.

References

[1] S. Vyas, H. B. Prajapati and V. K. Dabhi, "Survey on face expres- sion recognition using cnn," in 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), pp. 102– 106, IEEE, 2019.

[2] M. Wang, Z. Wang, S. Zhang, J. Luan and Z. Jiao, "Face expres- sion recognition based on deep convolution network," in *2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pp. 1–9, IEEE, 2018.

[3] I. M. Revina and W. S. Emmanuel, "A survey on human face expression recognition techniques," *Journal of King Saud University-Computer and Information Sciences*, 2018.

[4] N. A. Sheth and M. M. Goyani, "A comprehensive study of geometric and appearance based facial expression recognition methods," *Inter- national Journal of Scientific Research in Science, Engineering and Technology*, vol. 4, no. 2, p. 163, 2018.

[5] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Transactions on Affective Computing*, 2020.

[6] P. Burkert, F. Trier, M. Z. Afzal, A. Dengel and M. Liwicki, "Dexpres- sion: Deep convolutional neural network for expression recognition," *arXiv preprint arXiv:1509.05371*, 2015.

[7] S. Patel and J. Pingel, "Introduction to deep learning: What are convolutional neural networks," *Mathworks UK*, 2019.

[8] J. Oheix, "Face expression recognition dataset," Jan 2019.

[9] M. Lyons, M. kamachi and J. Gyoba, "The japanese female facial expression (jaffe) dataset," 1998.

[10] Z. Yu and C. Zhang, "Image based static facial expression recognition with multiple deep network learning," in *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pp. 435–442, 2015.

[11] A. Cohn-Kanade, "Extended coded expression database," *Unter: http://www. pitt. edu/˜ emotion/ck-spread. htm*, 2018.

[12] A. Hernandez-Matamoros, A. Bonarini, E. Escamilla-Hernandez, M. Nakano-Miyatake and H. Perez-Meana, "A facial expression recog- nition with automatic segmentation of face regions," in *International Conference on Intelligent*

Software Methodologies, Tools, and Tech- niques, pp. 529–540, Springer, 2015.

[13] K. Liu, M. Zhang and Z. Pan, "Facial expression recognition with cnn ensemble," in *2016 international conference on cyberworlds (CW)*, pp. 163–166, IEEE, 2016.

[14] "Fer-2013: Wolfram data repository." https://datarepository.wolframcloud.com/resources/FER-2013, 2013.

[15] M. J. Cossetin, J. C. Nievola and A. L. Koerich, "Facial expression recognition using a pairwise feature selection and classification ap- proach," in *2016 International Joint Conference on Neural Networks (IJCNN)*, pp. 5149–5155, IEEE, 2016.

[16] K. Zhang, Y. Huang, Y. Du and L. Wang, "Facial expression recognition based on deep evolutional spatial-temporal networks," *IEEE Transac- tions on Image Processing*, vol. 26, no. 9, pp. 4193–4203, 2017.

[17] X. Zhao, X. Liang, L. Liu, T. Li, Y. Han, N. Vasconcelos and S. Yan, "Peak-piloted deep network for facial expression recognition," in *European conference on computer vision*, pp. 425–442, Springer, 2016.

[18] G. R. Kumar, R. K. Kumar and G. Sanyal, "Facial emotion analysis using deep convolution neural network," in *2017 International Confer- ence on Signal Processing and Communication (ICSPC)*, pp. 369–374, IEEE, 2017.

[19] C. Zhang, P. Wang, K. Chen and J.-K. Ka¨ma¨ra¨inen, "Identity-aware convolutional neural networks for facial expression recognition," *Jour- nal of Systems engineering and Electronics*, vol. 28, no. 4, pp. 784–792, 2017.

[20] B. Ryu, A. R. Rivera, J. Kim and O. Chae, "Local directional ternary pattern for facial expression recognition," *IEEE Transactions on Image Processing*, vol. 26, no. 12, pp. 6006–6018, 2017.

[21] D. Learning and C. N. Network, "Human centric facial expression recognition,"

[22] E. Ahmed, T. A. Abir, J. A. Siraji and B. Khulna, "Automatic facial expression recognition using convolutional neural network (cnn)," 2018.

[23] K.-C. Liu, C.-C. Hsu, W.-Y. Wang and H.-H. Chiang, "Real-time facial expression recognition based on cnn," in *2019 International Conference on System Science and Engineering (ICSSE)*, pp. 120–123, IEEE, 2019.

[24] T. U. Ahmed, S. Hossain, M. S. Hossain, R. ul Islam and K. Andersson, "Facial expression recognition using convolutional neural network with data augmentation," in *2019 Joint 8th International Conference on Informatics, Electronics & Vision (ICIEV) and 2019 3rd Interna- tional Conference on Imaging, Vision & Pattern Recognition (icIVPR)*, pp. 336–341, IEEE, 2019.

[25] M. ul Hassan, "Vgg16 convolutional network for classification and detection," *Neurohive. Dostopno na: https://neurohive. io/en/popular- networks/vgg16/[10. 4. 2019]*, 2018.

[26] G. Sharma, "Real time facial expression recognition." https://medium.datadriveninvestor.com/real-time-facial-expression- recognition-f860dacfeb6a, Nov 2018.

[27] Hvass-Labs, "Hvass-labs/tensorflow-tutorials." https://github.com/Hvass-Labs/TensorFlow-Tutorials. Accessed on 2019-11-28.

[28] S. Das, "Cnn architectures: Lenet, alexnet, vgg, googlenet, resnet." https://medium.com/analytics-vidhya/cnns-architectures-lenet-alexnet- vgg-googlenet-resnet-and-more-666091488df5, 2019. Accessed on 2019-11-28.

[29] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, "Re- thinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

[30] S. Patel and J. Pingel, "Introduction to deep learning: What are convolutional neural networks," *Mathworks UK*, 2019.

[31] J. Oheix, "Face expression recognition with deep learning." https://www.kaggle.com/jonathanoheix/face-expression-recognition- with-deep-learning, Jan 2019.

التعرف على تعبيرات الوجه استنادًا إلى بنى الشبكات العصبية (CNN)

تيماء الريمي[١] ، عهد اللقماني [١] ، ابرار العتيبي [1,2] ، نداء الجهني[1]، سلمى كمون[1]

[١] قسم علوم الحاسبات، كلية الحاسبات وتقنية المعلومات، جامعة الملك عبد العزيز، جدة، المملكة العربية السعودية
[2] قسم علوم الحاسب، كلية علوم الحاسب وتقنية المعلومات، جامعة الامام عبدالرحمن بن فيصل الدمام، المملكة العربية السعودية

المستخلص. تعد شبكة Convolution Neural Network (CNN) هي بنية التعلم العميق الأكثر استخدامًا لأنها حطمت معظم الأرقام القياسية العالمية لمهام التعرف. تعتبر أنظمة التعرف على تعبيرات الوجه (FER) التي تستخدم التقنيات القائمة على الميزات الكلاسيكية، وخاصة تقنيات CNN، هي الأفضل لتصنيف الصور. استخدمت هذه الورقة ثلاث طرق قائمة على CNN، وهي VGG-16 و Inception-v3 و Resnet50-V2، لتصنيف تعابير الوجه إلى سبع فئات من المشاعر: سعيد، غاضب، محايد، حزين، اشمئزاز، خوف، ومفاجأة. تم استخدام مجموعة بيانات تعبير الوجه من مجموعة بيانات Kaggle وJAFFE لمقارنة الدقة بين البنى الثلاثة للعثور على الشبكة المحددة مسبقًا التي تصنف النماذج على أفضل وجه. أظهرت النتائج أن بنية شبكة VGG-16 أنتجت دقة أعلى (٩٣٪ في JAFFE و٥٤٪ في Kaggle) من البنى الأخرى.

الكلمات المفتاحيةـ التعرف على تعبيرات الوجه، الشبكة العصبية الالتفافية (CNN)، VGG-16, Inception-v3, Resnet50-V2