Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.049645

ARTICLE

CapsNet-FR: Capsule Networks for Improved Recognition of Facial Features

Mahmood Ul Haq[1], Muhammad Athar Javed Sethi[1], Najib Ben Aoun[2,3], Ala Saleh Alluhaidan[4,*], Sadique Ahmad[5,6] and Zahid farid[7]

[1]Department of Computer System Engineering, University of Engineering & Technology, Peshawar, 25000, Pakistan

[2]College of Computer Science and Information Technology, Al-Baha University, Alaqiq, 65779-7738, Saudi Arabia

[3]REGIM-Lab: Research Groups in Intelligent Machines, National School of Engineers of Sfax (ENIS), University of Sfax, Sfax, 3038, Tunisia

[4]Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh, 84428, Saudi Arabia

[5]EIAS: Data Science and Blockchain Laboratory, College of Computer and Information Sciences, Prince Sultan University, Riyadh, 11586, Saudi Arabia

[6]Department of Computer Sciences, Bahria University Karachi Campus, Karachi, 541004, Pakistan

[7]Department of Electrical Engineering, Abasyn University, Peshawar, 25000, Pakistan

*Corresponding Author: Ala Saleh Alluhaidan. Email: ASALluhaidan@pnu.edu.sa

Received: 13 January 2024 Accepted: 08 March 2024 Published: 15 May 2024

ABSTRACT

Face recognition (FR) technology has numerous applications in artificial intelligence including biometrics, security, authentication, law enforcement, and surveillance. Deep learning (DL) models, notably convolutional neural networks (CNNs), have shown promising results in the field of FR. However CNNs are easily fooled since they do not encode position and orientation correlations between features. Hinton et al. envisioned Capsule Networks as a more robust design capable of retaining pose information and spatial correlations to recognize objects more like the brain does. Lower-level capsules hold 8-dimensional vectors of attributes like position, hue, texture, and so on, which are routed to higher-level capsules via a new routing by agreement algorithm. This provides capsule networks with viewpoint invariance, which has previously evaded CNNs. This research presents a FR model based on capsule networks that was tested using the LFW dataset, COMSATS face dataset, and own acquired photos using cameras measuring 128 × 128 pixels, 40 × 40 pixels, and 30 × 30 pixels. The trained model outperforms state-of-the-art algorithms, achieving 95.82% test accuracy and performing well on unseen faces that have been blurred or rotated. Additionally, the suggested model outperformed the recently released approaches on the COMSATS face dataset, achieving a high accuracy of 92.47%. Based on the results of this research as well as previous results, capsule networks perform better than deeper CNNs on unobserved altered data because of their special equivariance properties.

KEYWORDS

CapsNet; face recognition; artificial intelligence

1 Introduction

Face recognition (FR) is one of the computer vision fields that is being researched the most because of its wide range of possible applications [1]. Recent years have seen the proposal of numerous novel deep learning-based face identification and verification algorithms [2], with generally very good recognition accuracy for distinct human faces under carefully regulated settings [3]. However, there are a lot of obstacles in the way of developing pose-robust face recognition.

A recent study [4] found that most FR algorithms perform worse when moving from frontal-frontal to frontal-profile face verification. This implies that the most challenging obstacle for face identification in the actual world is still stance variation [5]. The pose is described as a mix of face configuration and viewpoint. This work aims to provide a useful model for the recognition of unconstrained faces with large pose variations [6].

Convolutional Neural Networks (CNNs) have made significant progress in image recognition and visual processing across a wide range of applications, setting the standard for handling challenging computer vision tasks [7]. However, they fall short in three unique and significant ways that highlight how far distant they are from the brain processes vision [8], as demonstrated by the challenge of facial identification. First, there is no encoding of an object's orientation and position, which was a significant challenge for early CNNs [9]. While the use of augmented images helped to alleviate this problem, CNNs still fail to recognize a modified item if its position and orientation are not included in its training data [10]. To generalize to all points of view, an unlimited amount of data would be required.

Second, the method used by CNNs to produce predictions is easily fooled [11]. A CNN examines a picture and extracts meaningful elements that are subsequently used to classify it. If all components are present but in different positions, the CNN will still categories it as an object despite the jumbled feature distribution [12]. This is usually caused by the pooling layer doing feature subsampling that ignores feature position and location. A face with ears instead of eyes, for example, will still be classed as a face by a CNN, although a human obviously recognizing that it is not a face [13].

Finally, CNNs route information in a fashion that differs significantly from how the brain works. While CNNs route all information from low to high levels through all neurons, the brain routes specific information to specialized areas that are better at interpreting specific types of information [14]. Although computer vision and face recognition have enormous promise to revolutionize a variety of industries, CNNs continue to fall short of the flexibility and accuracy of the human brain [15].

A convolutional neural network (CNN) can recognize an image as a face even if it appears differently to a human observer if the ears and eyes are switched [16]. Because CNN are not as good at understanding the spatial relationships between images as the human brain is [17].

Hinton et al. have proposed capsule networks to overcome the difficulties with CNNs and to develop networks that more closely resemble the visual brain. In this research, CapsNet (Capsule Networks) are used to solve a face recognition challenge to investigate the workings of this unique network and compare it to the brain.

The main contribution of this paper is:

- Capsule Networks have shown excellent results in character identification, and this research extends prior work by using Capsule Networks to a face recognition problem on three distinct datasets.

- The suggested model has been tested for four different scenarios:
  - Face image resolution analysis.
  - Face pose analysis.
  - Face occlusion and non-uniform illumination analysis
  - Face rotation with non-uniform illumination and blurred face image analysis.
- Based on its distinct equivariance properties, the proposed model outperforms other baseline FR methods when compared to them.

2 Literature Review

2.1 Capsule Network

The hierarchical structure of the entities in images is beyond the capacity of the standard convolutional neural network [18,19]. "Capsules" could be used to learn part-whole relationships while retaining spatial information, according to Hinton et al. [20]. Because of their hierarchical feature representation, which improves generalization to pose variations and lessens reliance on data augmentation, Capsule Networks (CapsNets) are used [21]. Their value lies in their dynamic routing system, part-whole handling capabilities, and possible interpretability in applications such as medical image analysis and facial recognition. Researchers are interested in CapsNet since its authors in [22] first presented it as a more successful picture recognition system.

In [23], the relationship between the observer (position) and the entity is learned by the matrix capsule. Additionally, several strategies for implementing and enhancing the capsule design have been put forth [24], and [25]. On more complex datasets, DeepCaps, a deep capsule network architecture that was first shown in [26], has improved performance.

CapsNet has recently been integrated into several applications. Human action recognition [27], medical imaging [28], agricultural image classification [29], face forgery detection [30], character recognition [31], COVID-19 classification [32], face recognition [33] and other real-world problems [34] have been solved using these networks. Regular convolutional neural networks are replaced with capsule networks as discriminators in [35]. In order to achieve lung cancer screening, Mobiny et al. [36] use a dynamic routing method that is consistent. CNN-CapsNet is recommended by researchers in [37] for the successful classification of remote sensing photo scenes. Our work is an improved version of capsules, which creates arbitrary frontal faces in the deep feature space by using capsules.

2.2 Face Recognition

Research on CNN has significantly sped up the development of facial recognition techniques. A unified face verification, recognition, and clustering system was proposed by Schroff et al. [38]. Several approaches [39,40] has been examined to handle face pose changes. Researchers have developed methods for face frontalization [41], pose-specific identification training with deep models [42], and deformation of 3D facial markers [43].

A face frontalization sub-net (FFN) and a discriminative learning sub-net (DLN) are used to create a pose invariant model (PIM) by Zhao et al. [44]. Because PIM provides high-fidelity frontalized face photos, it is essentially a pixel-level alignment method. In contrast, the method in [45] directly addresses feature-level alignments and extracts deep features for face recognition using deformable convolutions with a spatial displacement field. Our method is lightweight and easy to deploy, unlike previous research that requires well-designed data augmentation or multi-task training. Haq et al. [46]

proposed a pose, illumination and occlusion invariant PAL-based face recognition model. Their proposed model achieved high accuracy as compared to the published algorithms. However, their algorithm was not tested on rotated and blurred images.

The authors of reference [47] introduced a revolutionary FR method dubbed MagFace, which is aimed to learn integrated features. They were able to attain an accuracy rate of 95.97% with their proposed approach, outperforming other existing algorithms.

The researchers in [48] created a Laplacian face approach (LFA) that generates Laplacian faces by using optimal linear approximation of eigenfaces. According to their published data, the LFA-based face recognition system has substantially reduced error rates. Furthermore, Simonyan et al. [49] identified faces using densely collected Scale Invariant feature Transform (SIFT) features using Fisher vectors. The proposed method achieved an overall accuracy of 87.7% for the standard Labelled Faces in the Wild (LFW) dataset.

Taigman et al. [50] unveiled DeepFace, a deep learning-based face recognition system that achieved high accuracy in tests. DeepFace used a deep convolutional neural network to learn features from raw face pictures and achieved a verification accuracy of 97.35% on the Labelled Faces in the Wild dataset.

The authors of [51] offer a method for recognizing goat faces using CNN, a deep learning network. The authors compiled a dataset of goat face photos and used it to train a CNN. They were able to recognize individual goats with excellent accuracy even when the photographs were taken in different lighting and environmental circumstances.

Although CNN has reached very high accuracy in face recognition, with Google's FaceNet obtaining 99.63% accuracy on the Labelled Faces in the Wild dataset, these networks are incapable of giving accurate spatial relationships between high-level elements due to several factors. For example, if the positions of the nose and mouth on a face are switched, convolutional neural networks will still classify the image as a face. To categorize the image as a face in a capsule network, the two active capsules representing the mouth and nose must have the correct spatial connection to activate the following capsule, which in this case will be the face. If the predictions for the mouth and nose agree, then the mouth and nose must be in the correct spatial relationship to form a face. Figs. 1 and 2 present recent FR algorithms tested on LFW dataset and COMSATS face dataset.

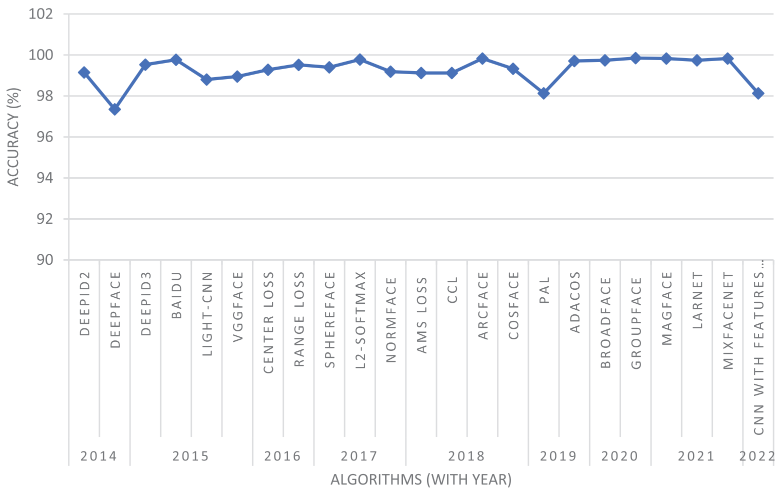

Figure 1: Recent face recognition algorithms tested on LFW face dataset

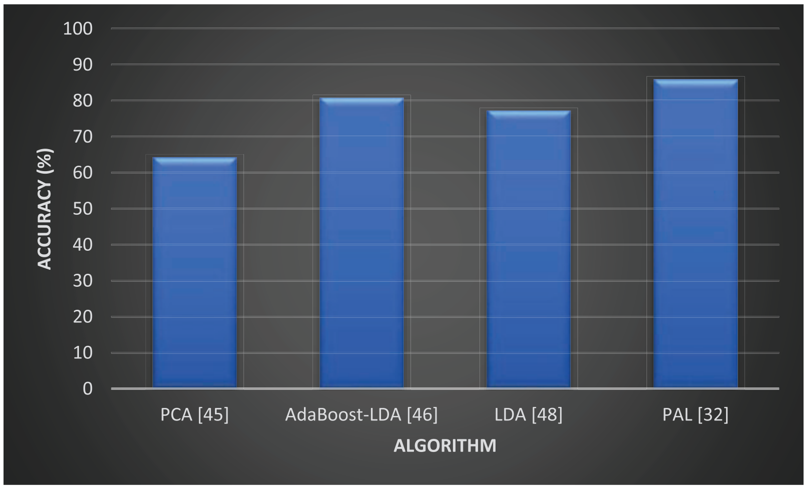

Figure 2: Recent face recognition algorithms tested on COMSATS face dataset

3 Proposed Methodology

3.1 Preprocessing Step

Face recognition preprocessing typically consists of a sequence of procedures designed to prepare input images for an accurate and efficient face identification. Image resolution (in pixels) varies between datasets. For this experiment, these images were reduced to 128 × 128 pixels, 40 × 40 pixels, and 30 × 30 pixels.

The initial stage of our proposed approach is to find and align the face so that the eyes, nose, and mouth are in the same place in each image. Several algorithms [52,53] have been employed in the literature for this goal, each with its own set of strengths and weaknesses. Face Landmark Localization (FLL) [54] was chosen for the proposed approach because it aligns the face by detecting 68 face landmarks in an image. As a result, precise points of these traits are recognized to find the human face. The face has distinct features such as the top of the chin, mouth, eyes, nose, and eyebrows. Therefore, to find the human face, specific points of these traits are identified. Pose variation is handled by the FLL with considerable efficiency. The process of face alignment involves building a tringle by joining the eye locations of facial landmarks to the image's normal, followed by calculating the angles between the eye points and the normal. Subsequently, the computed angles are used to rotate the face images. After face alignment, the face region is cropped from the input image using Eq. (1)'s minimum horizontal and vertical FLL points. The FLL and image cropping pseudo code are shown in Algorithm 1.

$$I_{face} = \sum_{i_{min-k}}^{i_{max+k}} \sum_{j_{min-l}}^{j_{max+l}} I_{mage} \tag{1}$$

where the lowest and maximum horizontal FLL points are represented by the variables i_{min} and i_{max}. The minimum and highest vertical FLL points are represented by the variables j_{min} and j_{max}, respectively. However, k and l are predefined variables and employed to crop the desired facial area. Fig. 3 illustrates the FLL and face cropping procedures.

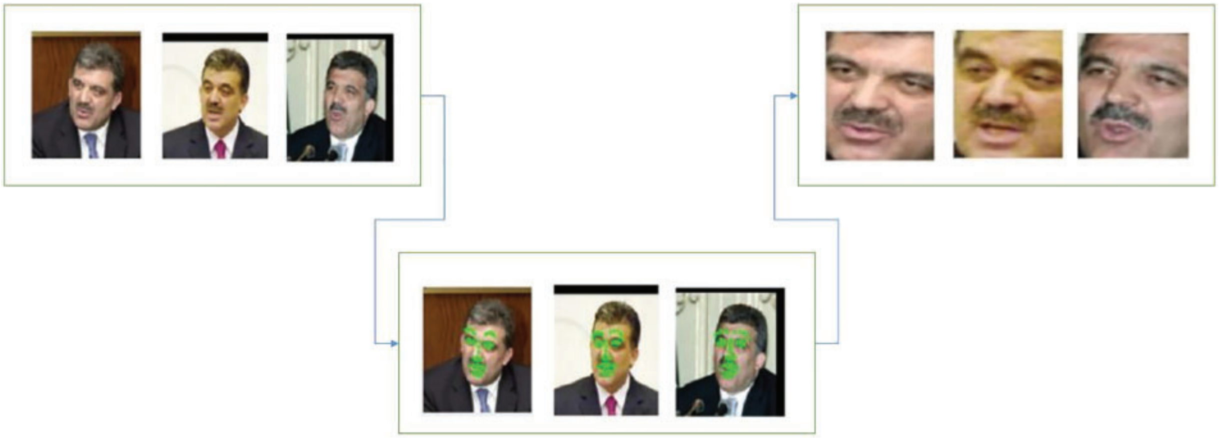

Figure 3: The FLL and face cropping

Algorithm 1: Pseudo code of FLL and image cropping

```
1. Take image.
2. Use the Haar-cascade classifier to detect faces.
3. Loop through detected face to detect landmarks
4. Find the face landmarks on the face.
5. Make a triangle with the tip of your eyes.
6. Determine how far the eyes are from the image normal.
7. Rotate the picture using the computed angles and normal line.
8. Crop the face according to the horizontal and vertical point’s threshold.
9. Move to step 1 for the next image.
```

3.2 Machine Learning Approach

For this paper, a three-layer capsule network is employed to identify faces as presented in Fig. 4. The first convolutional layer translates the pixel intensities of the input images to local feature detector activity, which are then used as inputs to the second convolutional layer. The major capsules are located in the second layer known as convolutional capsule layer. Convolutional layer consists of 32 convolutional channels, each comprising 8-dimensional capsules. These dimensions can express characteristics such as hue, position, size, orientation, deformation, texture, and so forth. Using routing by agreement, each 8-dimensional vector output is sent as input to all 16-dimensional capsules in the layer above. The last layer contains 16 dimensional capsules per class.

A capsule network is made up of two distinct elements. The problem of representing multidimensional entities is tackled in the first segment. This method groups together features with comparable attributes into units called “Capsules.” The purpose of the second component is to activate higher-level features, and it does so via using routing by agreement. It entails the agreement of lower-level features to activate higher-level characteristics.

A capsule network first splits the input image into subsets according to the regions that make up each subset. It functions on the presumption that there is only one feature per region, with one instance of that feature at most. The suggested model’s capsule is what it is known as. It records the movement, color, and position of objects. In contrast to CNNs, which represent solitary and unrelated scalar values, capsule networks link neurons together based on their multi-dimensional properties. This grouping makes routing by agreement possible, which makes it easier to activate features at a higher level.

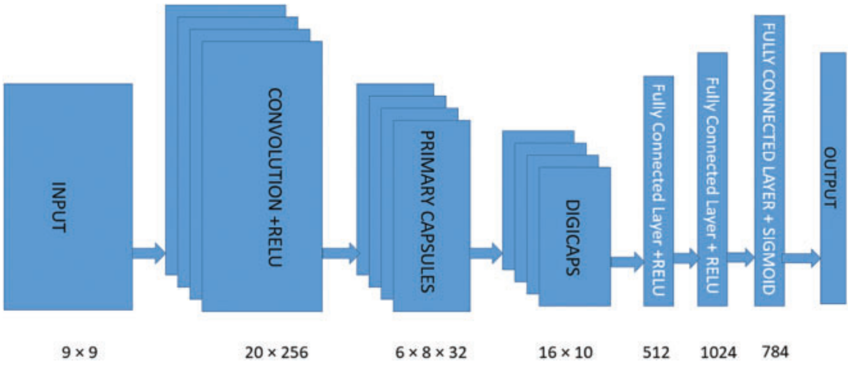

Figure 4: Capsule network architecture

3.3 Training

Dynamic Routing is employed as part of the training procedure to learn the coupling coefficients, c_{ij}, between capsules in the second and final layer. All log prior probabilities, b_{ij}, are set to zero at the start of each routing. A log prior probability is the likelihood of capsule i in layer l being connected with capsule j in layer l + 1. For r iterations to calculate, the prior probabilities are updated.

Using the routing softmax given in Eq. (2), the coupling coefficients (c_{ij}) between each capsule in layers l and l + 1 are first calculated. This guarantees that all of the capsules in layer l + 1 and a capsule in layer l's coupling coefficients sum up to one.

$$c_{ij} = \frac{\exp\left(b_{ij}\right)}{\sum_k \exp\left(b_{ij}\right)} \tag{2}$$

The total input (s_j) to a capsule j in layer l + 1 is then calculated using Eq. (3), which is a weighted sum of all prediction vectors ($\widehat{u_{j|i}}$). Eq. (4) defines a prediction vector as the output of capsule i in layer l multiplied by a weight matrix between capsules i and j.

$$S_j = \sum_i c_{ij}\hat{u}_{j|i} \tag{3}$$

$$\hat{u}_{j|i} = W_{ij}u_i \tag{4}$$

The probability that the entity a capsule represents is present in the current input is indicated by the output vector (v_{ij}) of the capsule. Eq. (5), a nonlinearity that guarantees small vectors are shrunk to almost zero and long vectors are squished to slightly less than one, is used to obtain this result. The agreement between each capsule j's current output and the prediction vector that capsule i produced is used to update the log prior probability.

$$V_j = \frac{||S_j||^2}{1 + ||S_j||^2}\frac{S_j}{||S_j||} \tag{5}$$

The batch size and learning rate have been set to 128 and 0.001, accordingly, based on the parameters used in the capsule network described in Dynamic Routing between Capsules [4]. The hyperparameters of min faces per sample were set to 25 in order to strike a compromise between the size of the data set and the number of faces per label.

4 Experimental Analysis

Experiments are carried out on a Dell Precision Tower 7810 equipped with a GTX 1080 GPU and 32 GB of RAM (Random Access Memory). The aforementioned is a better workstation for handling real-time sophisticated algorithms. Simulations are performed using MATLAB and Python. As evaluation measures in the experiment, accuracy and loss are used.

4.1 Datasets Used

The proposed model has been tested for four different scenarios: (i) face image resolution analysis. (ii) face pose analysis. (iii) face occlusion and non-uniform illumination analysis. (iv) face rotation with non-uniform illumination and blurred face image analysis.

4.1.1 Face Image Resolution Analysis

Several images were taken using the camera in a variety of locations, including rooms, parking lots, forests, and public markets, as shown in Fig. 5. We obtained a total of 1243 varied images using a mobile camera, including 76 subjects. After image acquisition face detection has been done using AdaBoost face detection algorithm to acquire only face images and remove the background. Faces were recognized by the AdaBoost face detection algorithm in 1217 face images of 76 participants chosen to test the proposed approach. Experiments are carried out by adjusting the size of facial images, for example, 128 × 128 pixels, 40 × 40 pixels, and 30 × 30 pixels.

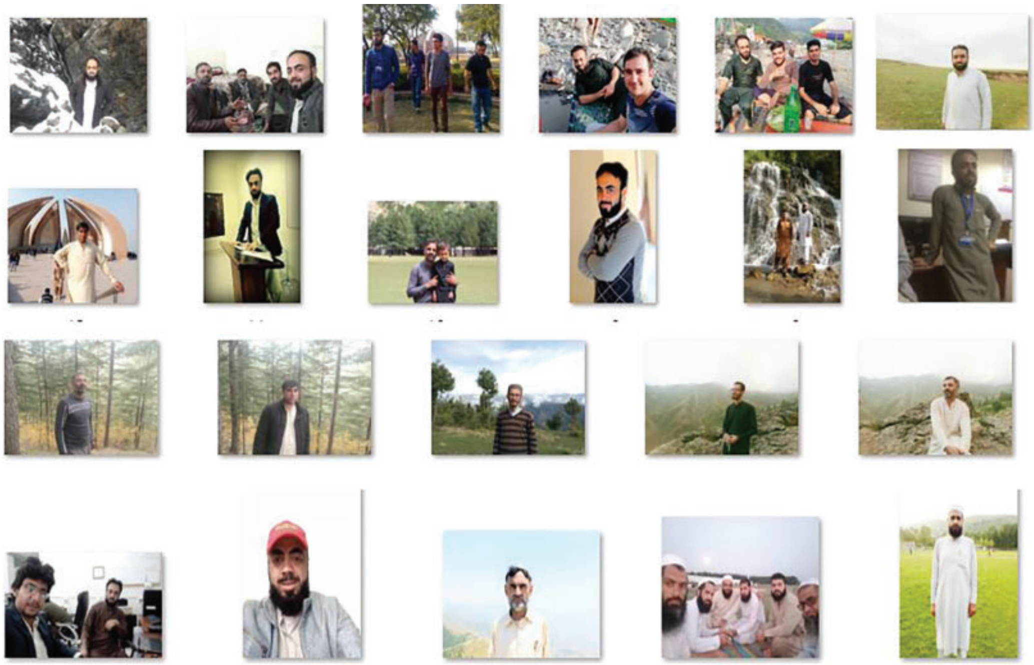

Figure 5: Images taken with mobile camera

4.1.2 Face Pose Analysis

The suggested model was tested on the COMSATS face dataset to see how effective it was in face pose analysis. The COMSATS face dataset contains 850 photos of people in various face stances (0°, ±5°, ±10°, ±15°, ±20°, ±25°, ±30°, ±35°, ±55°) of 50 different subjects. These photos were captured at COMSATS University over the course of five months under real-world conditions. Figs. 6 and 7 show a few samples with poses from the COMSATS face dataset.

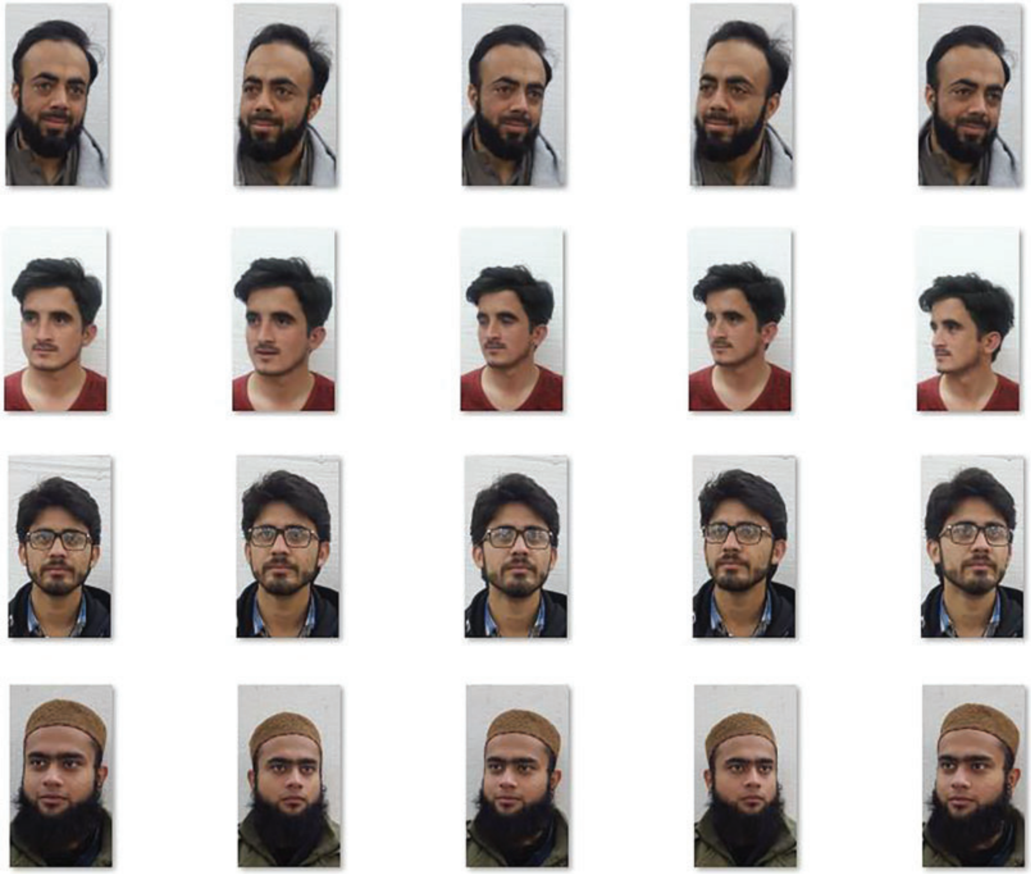

Figure 6: COMSATS face dataset individuals

4.1.3 Face Occlusion and Non-Uniform Illumination Analysis

To test the effectiveness of the proposed model against occlusion with pose variation and illumination variation, the LFW dataset was used. The LFW dataset is a well-known computer vision benchmark dataset. The LFW dataset images were collected from several sources, including the internet, and hence differ in quality, lighting, and position. As a result, the LFW dataset provides a challenging test for facial recognition algorithms. Face detection, recognition, verification, and identification investigations have all made extensive use of the LFW dataset. The LFW dataset has been used to develop and test several algorithms, and it is commonly used as a standard benchmark to compare the performance of various methods. Fig. 8 shows several examples of LFW datasets. Only a subset of the LFW dataset was employed, with the initial condition that each subject includes at least ten faces images. This resulted in a total of 191 distinct subjects within a collection of 5431 photos. This dataset was separated into two parts: 4328 photos for training and 1103 images for testing.

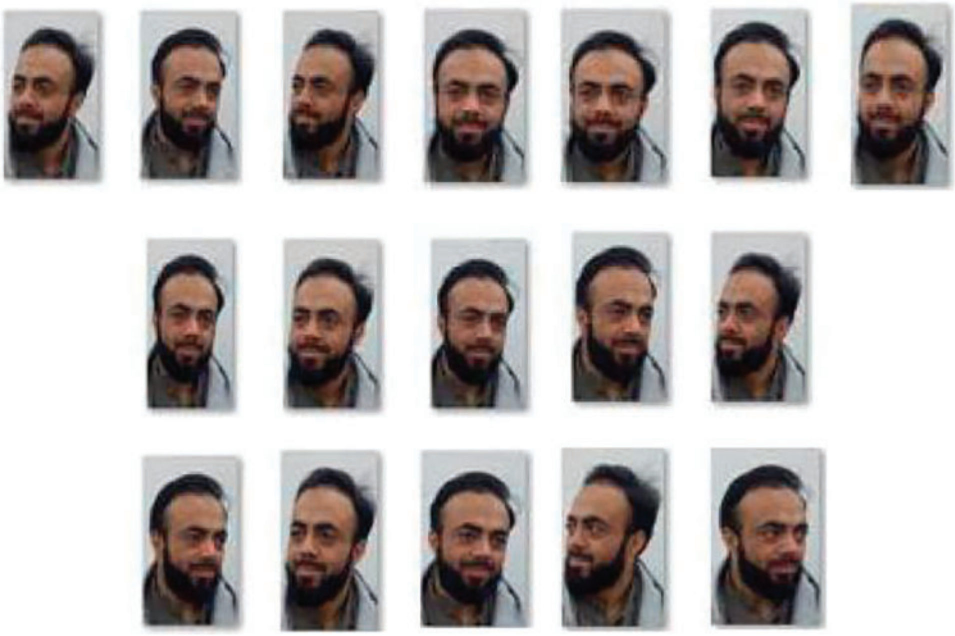

Figure 7: Individual of COMSATS face dataset with seventeen various face postures

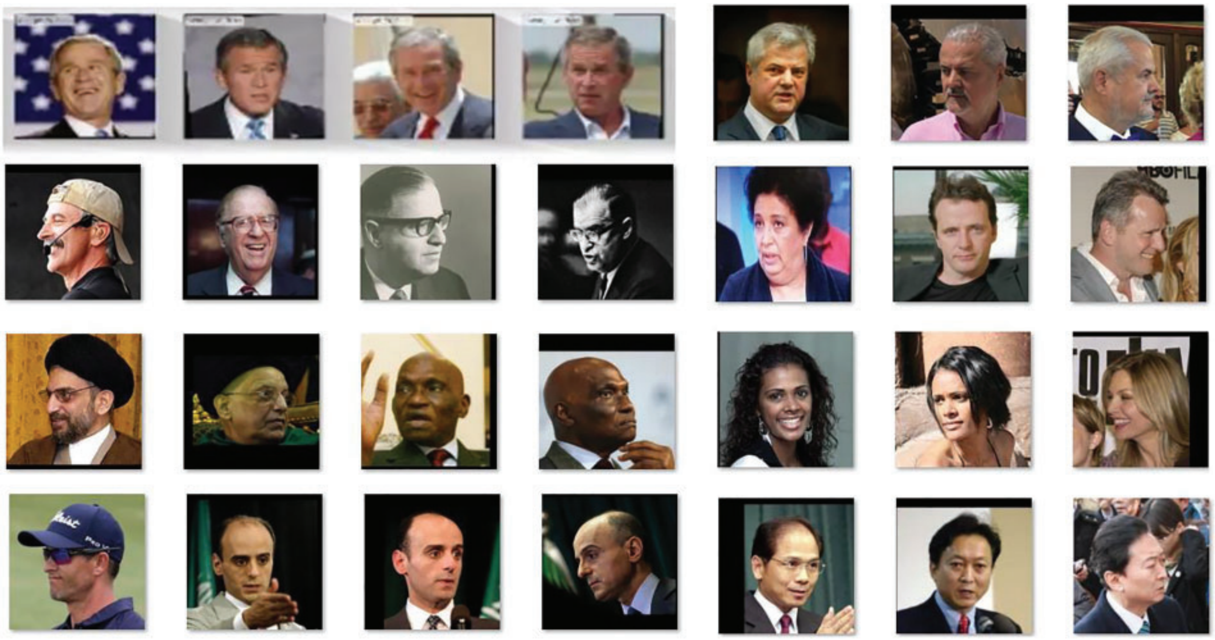

Figure 8: LFW dataset images

4.1.4 Face Rotation with Non-Uniform Illumination and Blurred Face Image Analysis

About 1221 face images were rotated 90 degrees, 1157 images 180 degrees, and 438 images 270 degrees to test the suggested model's resilience to rotated faces. In addition, 107 face shots were made blurry to evaluate the performance of the proposed model on these blurry images. Fig. 9 displays the LFW dataset's twisted and blurred facial images.

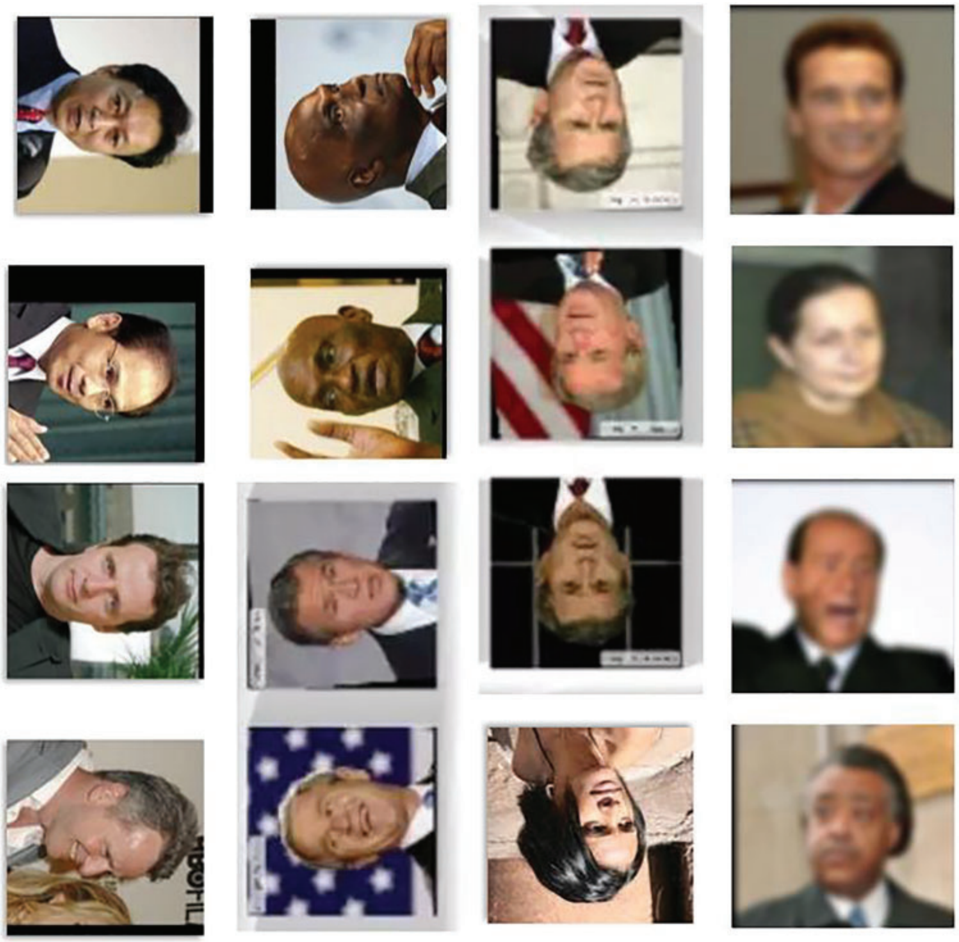

Figure 9: LFW dataset with rotated and blurred images

5 Results

The suggested model performed effectively while recognizing faces images having several difficulties. Fig. 10 displays the validation accuracy of the gathered camera photos for 128 × 128 pixels, 40 × 40 pixels, and 30 × 30 pixels. The recognition accuracy after 100 epochs is 95.29% for 128 × 128 pixels, 94.81% for 40 × 40 pixels, and 91.26% for 30 × 30 pixels.

Fig. 11 depicts the suggested model's recognition accuracy on the COMSATS face dataset for seventeen different positions. The total accuracy of the suggested model for the COMSATS face dataset is 92.47%. On the LFW dataset, the proposed model's facial recognition performance was approximately 97.3%; however, after rotating and blurring images, as illustrated in Fig. 12, it fell to 95.82%.

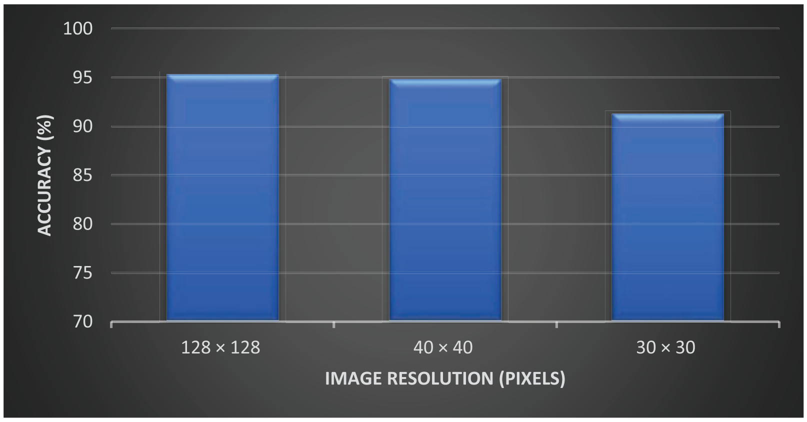

Figure 10: Validation accuracy of camera images for 128 × 128 pixels, 40 × 40 pixels, and 30 × 30 pixels

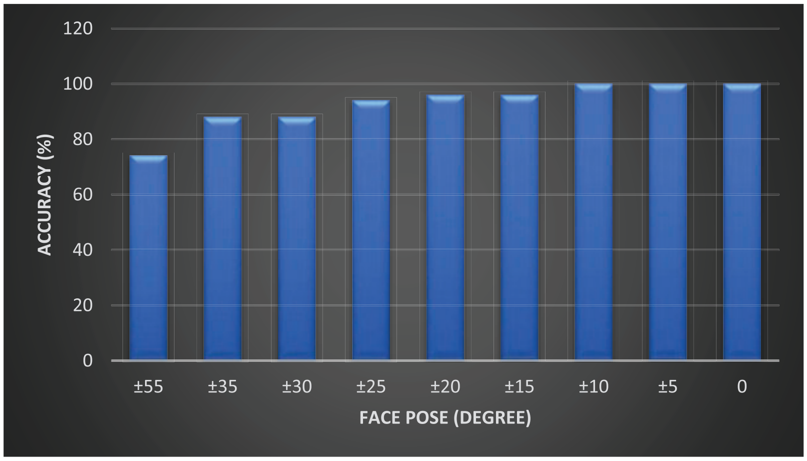

Figure 11: Recognition accuracy of the suggested model on the COMSATS face dataset for seventeen different positions

The suggested model was contrasted with five of the most sophisticated algorithms currently in use: (i) PAL [43], (ii) CapsNet algorithm [45], (iii) Eigen Faces [55], (iv) AdaBoost-LDA [56] and (v) Multi Modal Deep Face Recognition (MM-DFR) approaches [57]. Figs. 13 and 14 display the comparisons between the proposed model and the latest publicly available algorithms.

Figs. 13 and 14 provide a summary of the observations.

- The proposed FR method outperformed all prior FR strategies, as illustrated in Fig. 13, with 97.3% recognition accuracy on the LFW dataset.
- With an accuracy of 92.47%, the proposed FR algorithms outperform the conventional FR algorithms on the COMSATS face dataset (Fig. 14).

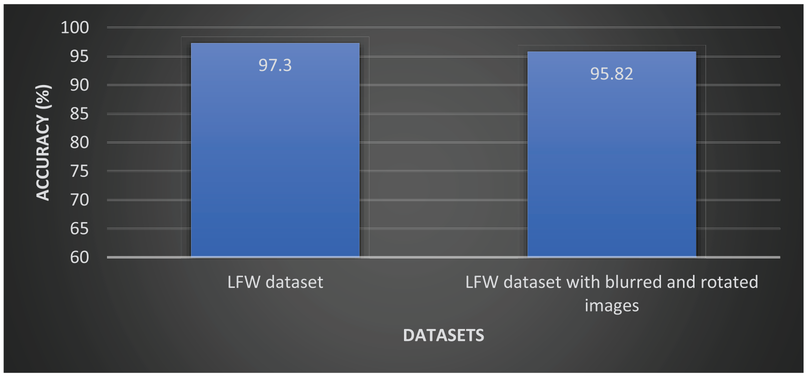

Figure 12: Accuracy of proposed model recognition on LFW dataset

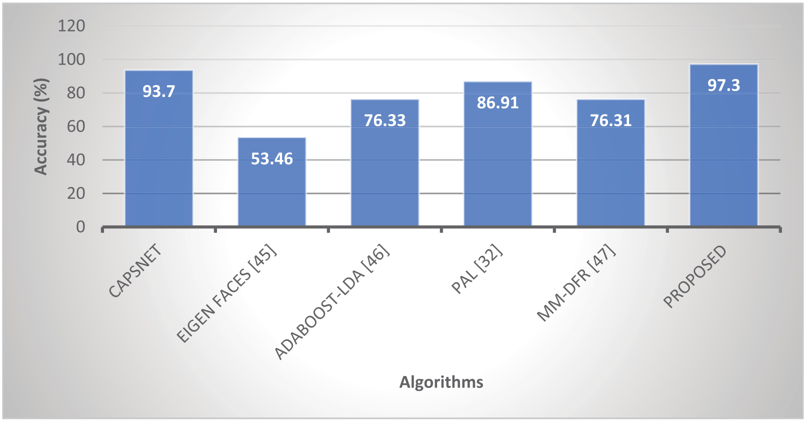

Figure 13: Recognition accuracy on the LFW dataset

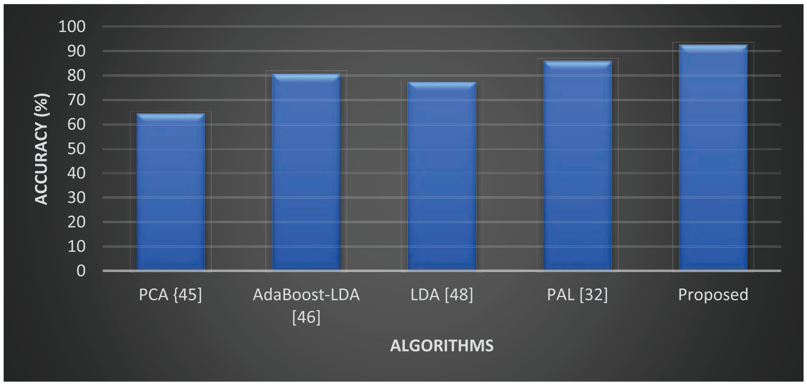

Figure 14: Recognition accuracy on the COMSATS face dataset

The performance of the published FR models and the proposed model on each facial pose of the COMSATS face dataset is shown in Fig. 15.

Fig. 15 highlights the following points:

- All algorithms achieved 100% recognition accuracy for frontal facial positions (±0^0, ±5^0, and ±10^0).
- The suggested model outperformed rival FR algorithms and attained satisfactory precision for non-frontal face images.
- In terms of face pose recognition, the proposed FR model performed more effectively than the most recent best FR approach (PAL).

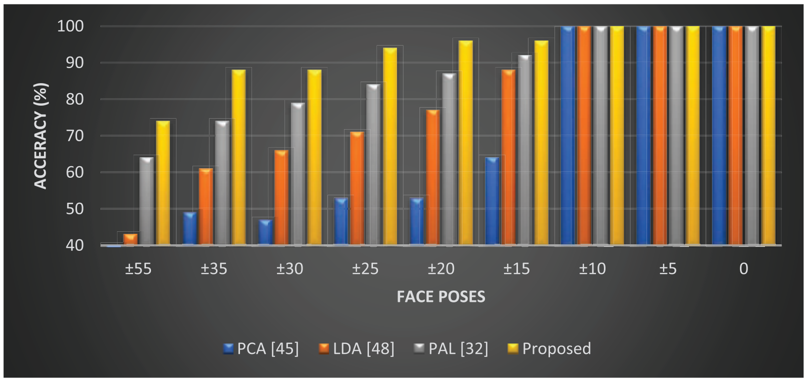

Figure 15: Recognition accuracy on COMSATS face dataset for seventeen different poses

6 Limitations and Future Directions:

The limitations of the proposed approach may be:

- The quantity and variety of the training data may have an impact on the model's performance. The model's capacity to generalize to a wider range of settings can be improved by augmenting the dataset with more varied facial images and variants.
- Although proposed model performed efficiently but still unable to recognize images having high occlusion.

Moreover, the proposed face recognition algorithm could be used to identify criminal suspects more accurately and efficiently. Additionally, facial recognition algorithms may be employed in border security to help identify people who might be on watch lists or who might be trying to enter the nation illegally.

In the future, we will concentrate on making the proposed model more efficient for real-time uses including augmented reality, robotics, and video processing. In order to guarantee low latency performance, this involves model compression, hardware acceleration, and other methods.

7 Conclusion

This research presents a Capsule Networks based FR model that was tested on images captured by a camera with an image size of 128 × 128 pixels, 40 × 40 pixels, and 30 × 30 pixels, as well as the COMSATS face dataset and the LFW dataset. The findings of this work are:

- The trained model demonstrated promising recognition accuracy of 95.29% for 128 × 128 pixels, 94.81% for 40 × 40 pixels, and 91.26% for 30 × 30 pixels on own collected images.

- The accuracy of the trained model has decreased to 95.82% on unseen LFW dataset faces subjected to occlusion or rotation, as compared to the proposed model's 97.3% accuracy for the LFW dataset.
- With a high accuracy of 92.47%, the proposed model outperformed the recently published method utilized for COMSATS face dataset evaluation.
- Due to their special equivariance qualities, capsule networks perform better than deeper CNNs on unobserved altered data, according to prior research and the findings of this study.

Acknowledgement: The authors would like to thank Prince Sultan University for their support.

Funding Statement: This research was funded by Princess Nourah bint Abdulrahman University Riyadh, Saudi Arabia with Researchers Supporting Project Number: PNURSP2024R234.

Author Contributions: Conceptualization, M.U.H. and M.A.J.S.; methodology, M.U.H., M.A.J.S., A.S.A., Z.F., and S.A.; validation, M.U.H., A.S.A., S.A. Z.F., and N.B.A.; data curation, M.U.H., M.A.J.S., S.A., and N.B.A.; writing—original draft preparation, M.U.H. and Z.F.; writing—review and editing, M.A.J.S., S.A., and N.B.A.; visualization, M.U.H. and S.A.; supervision, M.A.J.S., A.S.A., S.A., and N.B.A.

Availability of Data and Materials: The datasets used in this study's generation and/or analysis may be found at: https://www.kaggle.com/datasets/mahmoodulhaq/comsats-face-dataset and https://vis-www.cs.umass.edu/lfw/#download. Accessed: Oct. 20, 2023.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] H. Ullah, M. U. Haq, S. Khattak, G. Z. Khan, and Z. Mahmood, "A robust face recognition method for occluded and low-resolution images," in *2019 Int. Conf. Appl. Eng. Math. (ICAEM)*, Taxila, Pakistan, 2019, pp. 86–91, doi: 10.1109/ICAEM.2019.8853753.

[2] F. Munawar *et al.*, "An empirical study of image resolution and pose on automatic face recognition," in *2019 16th Int. Bhurban Conf. Appl. Sci. Technol. (IBCAST)*, Islamabad, Pakistan, 2019, pp. 558–563, doi: 10.1109/ibcast.2019.8667233.

[3] R. Xu *et al.*, "Depth map denoising network and lightweight fusion network for enhanced 3D face recognition," *Pattern Recognit.*, vol. 145, no. 11, pp. 109936, Jan. 2024. doi: 10.1016/j.patcog.2023.109936.

[4] M. U. Haq, M. A. J. Sethi, R. Ullah, A. Shazhad, L. Hasan and G. M. Karami, "COMSATS face: A dataset of face images with pose variations, its design, and aspects," *Math. Probl. Eng.*, vol. 2022, no. 1, pp. 1–11, May 2022. doi: 10.1155/2022/4589057.

[5] D. Khan, P. Kumam, and W. Watthayu, "A novel comparative case study of entropy generation for natural convection flow of proportional-Caputo hybrid and Atangana baleanu fractional derivative," *Sci. Rep.*, vol. 11, no. 1, pp. 35, Nov. 2021. doi: 10.1038/s41598-021-01946-4.

[6] W. Zheng, M. Yue, S. Zhao, and S. Liu, "Attention-based spatial-temporal multi-scale network for face anti-spoofing," *IEEE Trans. Biom. Behav. Identity Sci.*, vol. 3, no. 3, pp. 296–307, Jul. 2021. doi: 10.1109/TBIOM.2021.3066983.

[7] S. Cong and Y. Zhou, "A review of convolutional neural network architectures and their optimizations," *Artif. Intell. Rev.*, vol. 56, no. 3, pp. 1905–1969, Jun. 2022. doi: 10.1007/s10462-022-10213-5.

[8] B. H. Zou, C. Cao, L. Wang, S. Fu, T. Qiao and J. Sun, “FACILE: A capsule network with fewer capsules and richer hierarchical information for malware image classification,” *Comput Secur*, vol. 137, pp. 103606, Feb. 2024. doi: 10.1016/j.cose.2023.103606.

[9] R. Budiarsa, R. Wardoyo, and A. Musdholifah, “Face recognition for occluded face with mask region convolutional neural network and fully convolutional network: A literature review,” *Int. J. Electr. Comput. Eng.*, vol. 13, no. 5, pp. 5662, Oct. 2023. doi: 10.11591/ijece.v13i5.pp5662-5673.

[10] X. Zheng, Y. Fan, B. Wu, Y. Zhang, J. Wang and S. Pan, “Robust physical-world attacks on face recognition,” *Pattern Recognit.*, vol. 133, no. 3, pp. 109009, Jan. 2023. doi: 10.1016/j.patcog.2022.109009.

[11] P. Hedman, V. Skepetzis, K. Hernandez-Diaz, J. Bigün, and F. Alonso-Fernandez, “On the effect of selfie beautification filters on face detection and recognition,” *Pattern Recognit. Lett.*, vol. 163, no. 1, pp. 104–111, Nov. 2022. doi: 10.1016/j.patrec.2022.09.018.

[12] M. U. Haq, M. A. J. Sethi, and A. U. Rehman, “Capsule network with its limitation, modification, and applications—A survey,” *Mach. Learn. Knowl. Extraction*, vol. 5, no. 3, pp. 891–921, Aug. 2023. doi: 10.3390/make5030047.

[13] Y. M. Wu, L. Cen, S. Kan, and Y. Xie, “Multi-layer capsule network with joint dynamic routing for fire recognition,” *Image Vis. Comput.*, vol. 139, pp. 104825, Nov. 2023. doi: 10.1016/j.imavis.2023.104825.

[14] F. Boutros, V. Štruc, J. Fiérrez, and N. Damer, “Synthetic data for face recognition: Current state and future prospects,” *Image Vis. Comput.*, vol. 135, no. 10, pp. 104688, Jul. 2023. doi: 10.1016/j.imavis.2023.104688.

[15] Z. Huang, S. Yu, and J. Liang, “Multi-level feature fusion capsule network with self-attention for facial expression recognition,” *J. Electron. Imag.*, vol. 32, no. 2, pp. 193, Apr. 2023. doi: 10.1117/1.JEI.32.2.023038.

[16] J. M. Sahan, E. I. Abbas, and Z. M. Abood, “A facial recognition using a combination of a novel one dimension deep CNN and LDA,” *Mater. Today: Proc.*, vol. 80, no. 8, pp. 3594–3599, Jan. 2023. doi: 10.1016/j.matpr.2021.07.325.

[17] Z. Chen, J. Chen, G. Ding, and H. Huang, “A lightweight CNN-based algorithm and implementation on embedded system for real-time face recognition,” *Multimedia Syst.*, vol. 29, no. 1, pp. 129–138, Aug. 2022. doi: 10.1007/s00530-022-00973-z.

[18] E. Oyallon and S. Mallat, “Deep roto-translation scattering for object classification,” in *2015 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, 2015, pp. 2865–2873, doi: 10.1109/CVPR.2015.7298904.

[19] E. Worrall, S. J. Garbin, D. Turmukhambetov, and G. J. Brostow, “Harmonic networks: Deep translation and rotation equivariance,” in *2017 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, 2017, pp. 7168–7177, doi: 10.1109/CVPR.2017.758.

[20] G. E. Hinton, A. Krizhevsky, and S. D. Wang, “Transforming auto-encoders,” *Lect. Notes Comput. Sci.*, pp. 44–51, 2011. doi: 10.1007/978-3-642-21735-7_6.

[21] R. Kosiorek, S. Sabour, Y. W. Teh, and G. E. Hinton, “Stacked capsule autoencoders,” in *33rd Conf. Neural Inf. Process. Syst. (NeurIPS 2019)*, Vancouver, Canada, Jun. 2019, vol. 32, pp. 15486–15496.

[22] S. Sabour, N. Frosst, and G. E. Hinton, “Dynamic routing between capsules,” Cornell University, vol. 30, pp. 3856–3866, Oct. 2017. doi: 10.48550/arXiv.1710.09829.

[23] G. E. Hinton, S. Sabour, and N. Frosst, “Matrix capsules with EM routing,” in *ICLR 2018*, Vancouver, Canada, 2018, vol. 115.

[24] Y. Zhao, T. Birdal, H. Deng, and F. Tombari, “3D point capsule networks,” in *2019 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, 2019, pp. 1009–1018, doi: 10.1109/CVPR.2019.00110.

[25] J. E. Lenssen, M. Fey, and P. Libuschewski, “Group equivariant capsule networks,” in *32nd Conf. Neural Inf. Process. Syst. (NeurIPS 2018)*, Montréal, Canada, Jun. 2018, vol. 31, pp. 8844–8853.

[26] J. Rajasegaran, V. Jayasundara, S. Jayasekara, H. Jayasekara, S. Seneviratne and R. Rodrigo, “DeepCaps: Going deeper with capsule networks,” in *2019 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, 2019, pp. 10717–10725. doi: 10.1109/CVPR.2019.01098.

[27] V. Jayasundara, D. Roy, and B. Fernando, "FlowCaps: Optical flow estimation with capsule networks for action recognition," in *2021 IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Waikoloa, HI, USA, 2021, pp. 3408–3417. doi: 10.1109/WACV48630.2021.00345.

[28] Y. Wang, D. Ning, and S. Feng, "A novel capsule network based on wide convolution and multi-scale convolution for fault diagnosis," *Appl. Sci.*, vol. 10, no. 10, pp. 3659, May 2020. doi: 10.3390/app10103659.

[29] B. Janakiramaiah, G. Kalyani, L. V. N. Prasad, A. Karuna, and M. G. Krishna, "Intelligent system for leaf disease detection using capsule networks for horticulture," *J. Intell. Fuzzy Syst.*, vol. 41, no. 6, pp. 6697–6713, Dec. 2021. doi: 10.3233/JIFS-210593.

[30] K. Lin *et al.*, "IR-capsule: Two-stream network for face forgery detection," *Cogn. Comput.*, vol. 15, no. 1, pp. 13–22, Jun. 2022. doi: 10.1007/s12559-022-10008-4.

[31] S. Zhuo and J. Zhang, "Attention-based deformable convolutional network for Chinese various dynasties character recognition," *Expert. Syst. Appl.*, vol. 238, no. 9, pp. 121881, Mar. 2024. doi: 10.1016/j.eswa.2023.121881.

[32] H. Malik, T. Anees, A. Naeem, R. A. Naqvi, and W. K. Loh, "Blockchain-federated and deep-learning-based ensembling of capsule network with incremental extreme learning machines for classification of COVID-19 using CT scans," *Bioeng.*, vol. 10, no. 2, pp. 203, Feb. 2023. doi: 10.3390/bioengineering10020203.

[33] W. Zheng, M. Yue, S. Zhao, and S. Liu, "Attention-based spatial-temporal multi-scale network for face anti-spoofing," *IEEE Trans. Biom. Behav. Identity Sci.*, vol. 3, no. 3, pp. 296–307, Jul. 2021. doi: 10.1109/TBIOM.2021.3066983.

[34] A. Marchisio, G. Nanfa, F. Khalid, M. A. Hanif, M. Martina and M. Shafique, "SeVuc: A study on the security vulnerabilities of capsule networks against adversarial attacks," *Microprocess. Microsy.*, vol. 96, no. 9, pp. 104738, Feb. 2023. doi: 10.1016/j.micpro.2022.104738.

[35] W. Jaiswal, Y. AbdAlmageed, Y. Wu, and P. Natarajan, "CapsuleGAN: Generative adversarial capsule network," Cornell Univ., Feb. 2018. doi: 10.48550/arxiv.1802.06167.

[36] Mobiny and H. van Nguyen, "Fast CapsNet for lung cancer screening," *Lect. Notes Comput. Sci.*, pp. 741–749, 2018. doi: 10.1007/978-3-030-00934-2_82.

[37] W. Zhang, P. Tang, and L. Zhao, "Remote sensing image scene classification using CNN-CapsNet," *Remote Sens.*, vol. 11, no. 5, pp. 494, Feb. 2019. doi: 10.3390/rs11050494.

[38] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *2015 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, 2015, pp. 815–823. doi: 10.1109/CVPR.2015.7298682.

[39] S. Sengupta, J. C. Chen, C. Castillo, V. M. Patel, R. Chellappa and D. W. Jacobs, "Frontal to profile face verification in the wild," in *2016 IEEE Winter Conf. Applications of Computer Vision (WACV)*, Lake Placid, NY, USA, 2016, pp. 1–9. doi: 10.1109/WACV.2016.7477558.

[40] S. Wang, Y. Wu, Y. Chang, G. Li, and M. Mao, "Pose-aware facial expression recognition assisted by expression descriptions," in *IEEE Transactions on Affective Computing*, 2023. doi: 10.1109/TAFFC.2023.3267774.2023.

[41] S. Y. Wu, C. T. Chiu, and Y. C. Hsu, "Pose aware RGBD-based face recognition system with hierarchical bilinear pooling," in *21st IEEE Int. NEWCAS Conf. (NEWCAS)*, Edinburgh, UK, 2023, pp. 1–5. doi: 10.1109/NEWCAS57931.2023.10198097.

[42] H. He, J. Liang, Z. Hou, H. Liu, Z. Yang and Y. Xia, "Realistic feature perception for face frontalization with dual-mode face transformation," *Expert. Syst. Appl.*, vol. 236, no. 5, pp. 121344, Feb. 2024. doi: 10.1016/j.eswa.2023.121344.

[43] H. Joo, T. Simon, and Y. Sheikh, "Total capture: A 3D deformation model for tracking faces, hands, and bodies,," in *2018 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, 2018, pp. 8320–8329. doi: 10.1109/CVPR.2018.00868.

[44] J. Zhao *et al.*, "Towards pose invariant face recognition in the wild," in *2018 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, 2018, pp. 2207–2216. doi: 10.1109/CVPR.2018.00235.

[45] M. He, J. Zhang, S. Shan, M. Kan, and X. Chen, "Deformable face net for pose invariant face recognition," *Pattern Recognit.*, vol. 100, no. 10, pp. 107113, Apr. 2020. doi: 10.1016/j.patcog.2019.107113.

[46] M. U. Haq, A. Shahzad, Z. Mahmood, A. Shah, N. Muhammad and T. Akram, "Boosting the face recognition performance of ensemble based LDA for pose, non-uniform illuminations, and low-resolution images," *KSII Trans. Int. Inf. Syst.*, vol. 13, no. 6, pp. 3144–3164, Jun. 2019. doi: 10.3837/tiis.2019.06.021.

[47] Q. Meng, S. Zhao, Z. Huang, and F. Zhou, "MagFace: A universal representation for face recognition and quality assessment," in *2021 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, 2021, pp. 14220–14229. doi: 10.1109/CVPR46437.2021.01400.

[48] X. F. He, S. C. Yan, Y. X. Hu, P. Niyogi, and H. J. Zhang, "Face recognition using Laplacianfaces," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 27, no. 3, pp. 328–340, Mar. 2005. doi: 10.1109/TPAMI.2005.55.

[49] K. Simonyan, O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Fisher vector faces in the wild," *presented at the British Mach. Vision Conf.*, Jan. 2013. doi: 10.5244/C.27.

[50] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *2014 IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, 2014, pp. 1701–1708. doi: 10.1109/CVPR.2014.220.

[51] M. Billah, X. Wang, J. Yu, and Y. Jiang, "Real-time goat face recognition using convolutional neural network," *Comput. Electron. Agr.*, vol. 194, no. 2, pp. 106730, Mar. 2022. doi: 10.1016/j.compag.2022.106730.

[52] K. Sobottka and I. Pitas, "Face localization and facial feature extraction based on shape and color information," in *Proc. 3rd IEEE Int. Conf. Image Process.*, Lausanne, Switzerland, vol. 3, 1996, pp. 483–486. doi: 10.1109/ICIP.1996.560536.

[53] F. Tsalakanidou, S. Malassiotis, and M. G. Strintzis, "Face localization and authentication using color and depth images," *IEEE Trans. Image Process.*, vol. 14, no. 2, pp. 152–168, Feb. 2005. doi: 10.1109/TIP.2004.840714.

[54] E. King, "DLib-ML: A machine learning toolkit," *J. Mach. Learn. Res.*, vol. 10, no. 60, pp. 1755–1758, Dec. 2009. doi: 10.5555/1577069.1755843.

[55] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cognitive Neurosci.*, vol. 3, no. 1, pp. 71–86, Jan. 1991. doi: 10.1162/jocn.1991.3.1.71.

[56] J. Lu, K. N. Plataniotis, A. N. Venetsanopoulos, and S. Z. Li, "Ensemble-based discriminant learning with boosting for face recognition," *IEEE Trans. Neural Netw.*, vol. 17, no. 1, pp. 166–178, Jan. 2006. doi: 10.1109/TNN.2005.860853.

[57] Ding and D. Tao, "Robust face recognition via multimodal deep face representation," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2049–2058, Nov. 2015. doi: 10.1109/TMM.2015.2477042.