2D AND 3D FACE RECOGNITION BASED ON DEEP CONVOLUTIONAL NEURAL NETWORK AND 3D RECONSTRUCTION APPROACH

WISAM KAREEM THAJEAL AL-HRAISHAWI, Dr. Zurina Muda

Fakulti Teknologi dan Sains Maklumat, Universiti Kebangsaan Malaysia
43600 UKM Bangi, Selangor Malaysia.

wesamkarem@gmail.com, zurinam@ukm.edu.my

Copyright@FTSM UKM

ABSTRACT

Face Recognition (FR) has become one of the most crucial applications employed in the present day and is utilised for broad purposes, e.g. law enforcement, commercial applications, and security. Whilst FR has been improved significantly following the adoption of Deep Learning (DL) techniques, which varied Convolutional-Neural-Network (CNN)-based models has been proposed for FR, there remain many challenges in FR which must be addressed by scholars. The FR can take the form of 2D or 3D where 2D FR is the most popular, but the 3D FR provides higher accuracy. There is a challenge in the gathering of a 3D face images dataset which is quite a complicated process. Thus, this study aims to compare the efficiency of 3D FR against the 2D FR based on different deep CNN-based models. A model is proposed to reconstruct a 3D face-image dataset from a 2D face-image dataset via the use of the PRNet model combined with the MediaPipe tool. The study also introduced a framework (comparative study) to identify the CNN-based model which is most suitable for FR. Two CNN-based models are used through fine-tuning, namely VGG-16 and DenseNet-201 with a custom CNN. These three models are implemented in two experiments, first one is 2D FR that is implemented by using a 2D face image dataset namely PinsFace. The second experiment is 3D FR which is performed by using new generated 3D face image dataset. All these models are evaluated based on accuracy and error rates. The accuracy results of 2D FR for custom CNN, VGG-16, and DenseNet-201 models are 81.13%, 83.76 %, and 89.72% respectively. Whilst, the accuracy results of 3D FR for the same models are 82.16%, 88.97%, and 93.33%. In conclusion, the experiment results demonstrated that the accuracy results are increased by 1.03%, 5.21 %, and 3.61 for CNN, VGG-16, and DenseNet-201 models respectively in the 3D FR in comparison with 2D FR. These show that 3D FR has greater efficiency than 2D FR, particularly in handling the specific issues related to images, e.g. age variations, pose variations, illuminations, etc. The findings reveal that the DenseNet-201 is the most suitable model for FR for both images in 2D and 3D.

Keyword: 2D AND 3D FR, DEEP CNN, 3D RECONSTRUCTION APPROACH

I. INTRODUCTION

Deep Learning (DL) has emerged after certain significant advancement steps of Machine Learning (ML). DL is known as sub-techniques of a wider group of ML. The applications of ML have been expanded from data classification to speech recognition, computer vision, bioinformatics, classification of extremely big data, etc., by using DL. In terms of its use, DL, as a technique, models high-level data abstractions via, leveraging deep-networks architectures which are constructed using numerous linear and non-linear transformations. This results in a system is extremely intelligent, mimicking the human brain's functions, i.e. expressing complex facts derived from events in the real-world, and thus made it affordable and easy for users for formulating judgments more accurate.

Amongst the most promising algorithms in the ML domain, DL is based on comprehending data representation and is rooted in hierarchical learning. DL itself has significantly influenced computer vision performance in numerous areas – e.g. picture categorisation and object identification – that were unreachable previously. In terms of its applications, DL is utilised in research pertaining to signal processing, pattern recognition, and graphical modelling (Purwins et al.2019), computer vision (Bao et al. 2019), speech recognition (Xue et al. 2019), language recognition (Imran & Raman 2020), audio recognition (Al-Emadi et al. 2019), and Face Recognition (FR) (Prasad et al. 2020).

Security demands are extremely relevant and the technology has a great deal of potential for law enforcement and commercial use; in light of this, recent times have seen scholars single out FR technology as a red-hot research area (Prasad et al. 2020). FR system can be defined as a technology that identifies and matches a specific human-face from a given video or image against; this done through measuring facial features of the image with those in its own face's database. There exist two modes of functioning within FR. The first is the verification mode, i.e. one:one matching which is utilised to choose a face from a database containing numerous faces, so as to establish whether the face information belongs to the certain individual. The another is the identification-mode, i.e. one:many matching; this includes taking a person's biometrics and comparing those biometrics to a database containing possible faces (Shepley 2019).

DL algorithms are now being paid increasing levels of attention in the area of FR technology, with numerous researchers showing an interest in studying 3D FR (Zhang et al. 2019). On the one hand, the information extraction of 3D face-image is deemed the main function within 3D FR: if the steps of alignment and detection of the face are effective, this may enhance 3D FR's overall performance – something that is crucial in security systems as well as in commercial 3D FR application. In opposite, scholars have developed certain approaches based on utilising DL for 3D FR; those researchers proved that DL systems exhibit much better performance than ML with presence of a sizeable number of 3D face-images.

Copyright@FTSM UKM

Whilst 3D FR can perform FR with a higher degree of efficiency than can 2D face recognition. However, the lack of, or very small number of, publicly-available 3D face datasets has forced scholars and researchers to utilise the 2D face dataset, as the latter is affordable and widely available within FR (Fime et al. 2021; Yu et al. 2020). This is due to the fact that 3D data cannot be collected from websites as 2D faces can. Moreover, creating a dataset via using 3D scanning or infrared laser beams is a difficult and time-expensive task. As such, a few researchers have put forward the prospect of using different methods for the reconstruction of 3D face-image datasets from available 2D face-image datasets (Li et al. 2020). Indeed, 2D images are linked to issues that are considered the main challenges in applying recognition for sizeable datasets, since some of the extracted features might well be impacted by numerous different kinds of challenges, e.g. lighting variations, pose, occlusion, and differences in facial expressions (Nonis et al. 2019; Zhang et al. 2019; Zhu et al. 2017).

Whilst the utilisation of DL in FR has seen promising results, it is nevertheless still impacted by numerous problems. There are two obstacles in using DL approaches in the realistic FR scenarios. First, accuracy becomes unstable with the addition of more face image, since different DL networks have a variety of generalisation abilities when it comes to extracting the characteristics of images. Second, the deeper a DL model's layers, the greater the number of complexities which arise during the processing of large quantities of image data (Fime et al. 2021; Li et al. 2020, there is the potential for the recognition rate to be impacted by the model's complexity.

Based on the problems and challenges discussed, this research performs a comparison between 2D and 3D FR, by using three types of CNN-based widespread models: DenseNet-201, the CNN, and the VGG-16. These three models are employed to compare the error and accuracy rates of 3D and 2D FR. Moreover, in order to tackle the issues linked to 2D images and generate a new 3D face dataset, this study also reconstruct model from 2D dataset to produce a 3D face images dataset. A new 3D dataset can be generated for this domain, without relying on the small number of available 3D datasets which are excessively employed by past scholars and researchers.

II. LITERATURE REVIEW

A. FACE RECOGNITION

the FR system's initial task is to capture an image via a database, video, or camera, following which that image is passed on to the subsequent stage of the FR system – something that is discussed in the current section:

Copyright@FTSM UKM

Face Detection: The main aim of this stage is to identify the face of a individual in a database face-images or a captured face-images. The face detection method assesses whether or not the given image contains within it a face image; as soon as the face has been detected, it is time for the output to be passed on to the next stage (pre-processing).

Pre-processing: The pre-processing step for FR involves the removal of undesired noise, changing lighting conditions, blur, and shadowing effects; indeed, this removal is made possible by pre-processing techniques. The feature extraction procedure is carried out once a fine smooth facial image has been obtained.

Feature Extraction: Via the use of a feature extraction algorithm, it is possible for features of the face to be extracted at this stage. Indeed, dimension reduction, information packing, salience extraction, and noise cleansing are all conducted through extractions. Face features are commonly transformed into either a constant-dimension of vectors or a group of dots with their corresponding positions following present stage. The two main types of Feature Extraction methodology are the global and local characteristics-based approaches. Thus, also is possible to classify a human face according to local, as well as global, characteristics. Global features are not as discriminative as localised characteristics, and thus they are easier to capture.

Face Recognition: Once feature extraction is over, the final phase examines the representation of each face – a step which is employed to recognise the faces' identities, and this makes it possible to achieve FR; in order to attain the above goal, there is the need for a face database to be built. Numerous photographs are captured for each person, and the features of those photographs are retrieved before then being saved within the database. Upon being presented for recognition, a face image first undergoes pre-processing, face detection, and feature extraction, following which the feature is compared to each face class that has been recorded in the database.

B. DEEP LEARNING

DL methods are basically based on Artificial-NN (ANN) and representation learning. The methods of DL Learning are divided into supervised, semi-supervised or unsupervised (LeCun et al. 2015). Deep-learning architectures, e.g. Deep-NN (DNN), Deep-Belief-Networks (DBN), Deep-Reinforcement-Learning (DRL), Recurrent-NN (RNN) and Convolutional-NN (CNN), are implemented within various areas namely speech-recognition, computer-vision, natural-language processing (NLP), machine-translation, bioinformatics, medical image analysis, material inspection, climate science, and boardgame programs. Within these fields, the above-mentioned architectures obtained promised results which are competitive, also in certain conditions surpass, human-expert performance (Li et al. 2022). This research uses the CNN and the CNN-based models to construct a

Copyright@FTSM UKM

comparison framework of 2D and 3D FR, as there are only a few DL and CNN-based methods which have been implemented in the 3D FR field (Jribi et al. 2021).

C. Convolutional Neural Network

Convolutional-NN (CNN) constitutes the most popular DL architecture, and is based on an artificial neural network. A basic CNN comprises the following: one or more convolution layers, pooling, single or various numbers of fully-connected layers, and an output-layer, this algorithm is known as a distinctive form of multi-layer NN which offers benefits over feedforward neural networks, especially in terms of fewer connections and parameters; indeed, this facilitates the training process.

D. CNN-Based Models for FR

Whilst the CNN does deliver exceptional performance on FR tasks, in 2012 a large deep-CNN, named AlexNet (Krizhevsky et al. 2012), was created by Krizhevsky et al. and demonstrated outstanding performance on the ImageNet-Large Scale-Visual Recognition-Challenge (ILSVRC) (Russakovsky et al. 2015). The ILSVRC is utilised as an evaluation process for algorithms in object detection and image classification; it provides the researchers with the ability to compare progress in detection. Over the years, AlexNet's success has become the inspiration for different CNN models, including ZFNet (Zeiler & Fergus 2014), VGGNet (Simonyan & Zisserman 2015), GoogleNet (Szegedy et al. 2015), ResNet (He et al. 2016), DenseNet (Huang et al. 2017), CapsNet (Sabour et al. 2017), and SENet (Hu et al. 2018) etc.

After reviewing the CNN-based models which currently exist, it was found that, amongst these models, the VGG and DenseNet are more robust and promising when it comes to enhancing the FR's performance. As a consequence, this study focuses on utilising the VGG, and DenseNet, in addition to employing a CNN algorithm to construct the framework comparison, whilst, finally, it highlights the most suitable model, i.e. that which achieves the highest accuracy on the 2D and 3D FR datasets. The main benefit of employing CNN-based models in this research is the ability to fine-tune the existing pre-trained models, as opposed to training these models from scratch.

E. 3D FEATURE EXTRACTION

Facial-feature extraction is a task which involves extracting facial signs as a features such as eyes, mouth, nose etc. from individual face-images. The said process is vital for the initialisation of processing techniques such as FR, face tracking, or facial expression recognition (Benedict & Kumar

Copyright@FTSM UKM

2016). Face-landmark detection constitutes a computer vision function, employed for detecting and tracking a human face's key points; it can be applied to many problems. As an example in this regard, key points which stem from landmarking can be employed to detect the pose position and rotation of a human's head. In this way, it can establish whether or not a driver is paying attention. Moreover, it can be employed to more easily apply an augmented reality. With this said, in numerous applications it is not essential to have a comprehensive understanding of the concepts of face landmark detection. There exist many libraries which are used, e.g. MediaPipe.

MediaPipe is a framework utilised for the construction of ML pipelines, which are in turn used for the processing of time-series data such as videos and audio, etc. This cross-platform framework functions for servers/desktops, iOS, Android, and embedded devices, e.g. Jetson Nano and Raspberry Pi. As a tool, MediaPipe is utilised to implement ML-based computer vision models. Such tool produced by Google, it contains numerous different kinds of computer vision solutions, e.g. object-detection, face-detection, and pose estimation. After reviewing the existing feature extraction approaches, we will use the landmarks technique for features extraction in this study. MediaPipe tool is used as mesh landmarks which is a lightweight ML structure. This tool enables us to focus only on the three main sub-regions including the eyes, nose, mouth (eyes/iris, mesh, lips) and eventually the facial feature location points will be extracted to boost the quality of 3D FR by improving the accuracy of the recognition.

F. RECONSTRUCTION OF 3D FROM 2D FACE IMAGE

Although the 3D facial images provide a more accurate characterisation of the face, it is more difficult to obtain such images than it is with 2D photos. This is due to the fact that there is a higher price attached to the more complicated imaging process involved in 3D facial analysis systems used to acquire the 3D facial-information, such as stereo-vision systems (Beeler et al. 2010, 2011), 3D laser scanners (Lee et al. 1995), and RGB-D camera (such as Kinect). The stereo-vision systems and 3D laser scanners obtain scans of high-quality facial, but need controlled circumstances and costly instruments. Whilst, RGB-D camera is easier to use and not as expensive, but those scans produced are constrained in terms of quality (Yang et al. 2015). Consequently, certain researchers have attempted to develop alternative approaches which are able to generate 3D face-images by using a solely one 2D face-images in order to tackle this lack in the easiest manner.

DL has been shown to be an extremely efficient technique in numerous fields, e.g. 3D reconstruction in the computer vision area. The aforementioned has prompted researchers to begin employing different DL techniques, which have emerged over the last two decades, in the reconstruction of 3D face-images via using 2D face-images. This also led to the generation of, and

Copyright@FTSM UKM

learning from, actual representation of data training, thus avoiding the challenge of having enough ground-truth data.

To further aid in the training of input and output neural networks, deep learning models may construct their own representation, which preserves 3D face data. It is found that, amongst these reconstruction methods, the CNN-based reconstruction models have the most promise and success in 3D face reconstruction (Morales et al. 2021). Thus, this study utilised the PRNet model for the reconstruction of 3D face-images from 2D face-images. Feng et al. (2018) proposed an unconstrained end-to-end method namely as a Position-Map Regression-Network (PRNet).

G. RELATED WORKS

In the FR field, DL methods have received considerable attention, and numerous academics have realised the value of studying 3D FR. On the one hand, the obtaining of 3D face-information is a vital period in 3D FR, since effective face-detection and alignment has the ability to boost overall 3D FR performance, which is crucial when it comes to security as well as commercial 3D FR systems. In the current section, this work is analysed from the perspective of how the CNN and CNN-based models were employed in FR (2D and 3D FR), the reconstruction method, feature extraction, type of dataset, and constraints.

In the work of Huang and Chen (2022), the FR model is proposed that contains several components includes feature-restoration, feature-extraction, and embedding matching components. The feature-restoration component based on a two branches architectures of the CNN for producing the featured of face-image, then the illumination step is applied to enhance the image's lights. Following this, feature-extraction is utilised for encoding the feature-image and transforming to an embedded form, to be utilised in last step for identification and verification in matching module.

In Jeevan et al.'s (2022) study, an experimental comparison was conducted by utilising various existing CNN-based models to recognise masked (occluded faces) face images. The authors attemcpted to alter hyper-parameters such as loss functions, network architectures, and training methods of the models which were used. Subsequently, these models were evaluated through the use of different datasets of normal face images and masked face datasets.

In the work of Wang and Zhang (2021), an FR model-based DCNN is put forth, and is referred to as a multidimensional feature network (MFNet). This model comprises multidimensional feature extraction as well as a DL module. The first element is designed for extracting the features from spaces, dimensions, and channels, so as to address the occluded face images. Conversely, the

Copyright@FTSM UKM

DCNN is employed with the goal of ameliorating the recognition accuracy rate. This model recorded an accuracy of 90.35% when it was tested with the Yale face dataset.

In another work of (You et al. 2020), a multi-channel deep-network for 3D FR was proposed. In this study, the geometric information is computed for every 3D face image through the utilisation of their own piecewise-linear triangular mesh structure. This method is leveraged via adopting the pre-trained VGG-Face model (Parkhi et al. 2015), following which the model is fine-tuned with the newly-generated multi-channel features from the network's improved input layer. The FR accuracy of this model achieved 98.6% when tested across two 3D datasets, namely Bosphorus and TexasFRD.

In the work of Peng et al. (2020), a new method was suggested for FR based on the inception of the ResNet V1 and V2 network models to ameliorate performance accuracy. The authors attempted to improve the learning model of ResNet V1 and V2 through changing the training parameters, e.g. residual scaling factor, whilst they also used other various activation functions, including Leaky ReLU and PReLU, instead of ReLU activation; this enhancement gave rise to improvement of the stability of the training. The improved method experimented with different datasets, such as VGGFace2, MS1MV2, IJBB and LFW datasets. However, the training parameters increased, thus leading to increases in the model's complexity.

Two CNN-based models, namely Lightened CNN and VGG-Face, were employed in the work of Prasad et al. (2020). Both models showed that the DL model is robust againstc various FR challenges, such as varying misalignments, variant pose, variant occlusions, and illuminations.

Face spoofing detection methods based on learned features using the CNN series were introducced in a study carried out by Yu et al. (2021) using DenseNet-121 to reflect the characteristics of various frequency bands of an image. The simulation results showed that DenseNet exhibited good face spoofing detection performance.

Low-quality data was used in Mu et al. (2019). They proposed a MultiScale Feature-Fusion (MSFF) with the Spatial-Attention-Vectorization (SAV) modules for constructing a discriminative and compact-CNN. They subsequently proposed a data processing methods involving point-cloud retrieved, refinement of the surface, and data-augmentation, so as to tune the algorithms used. The simulation results outperformed many deep-CNNs, including ResNet 34 and VGG 16.

The study of Feng et al. (2018) conducted a comparison of the use of several reconstruction mechanisms for reconstructing ca 3D face-image from solely one 2D face-image by the help of 3D face ground truth (real 3D facial images) scan of the same subjects; the aim was to evaluate the performance of 3D reconstruction methods. Their results illustrated that the texture-based

Copyright@FTSM UKM

reconstruction method performed better than the landmark-based reconstruction method, but was very time consuming, since the use of in-depth information can ameliorate the resilience of a 3D face system.

III. METHODOLOGY

This section puts forth an overview of the methodology and delivers a description of the work's process; the methodology comprises three key stages. The first stage revolves around reconstructing a 3D face-image from the 2D face-image and then extracting the 3D facial features by using facial landmarks (MediaPipe) to prepare the image for 3D FR. The second stage is the implementation of 2D FR through utilising three approaches: CNN, VGG-16, and DenseNet-201. The last stage is the implementation of the 3D FR, which involves employing the same three models applied in 2D FR. Finally, evaluation metrics for 2D FR, as well as for 3D FR, will be introduced in relation to accuracy and model error rate. Python is used as a programming language to put into action the methodology of this work.

An overview of the proposed research methodology for this study is elaborated in Figure 1.

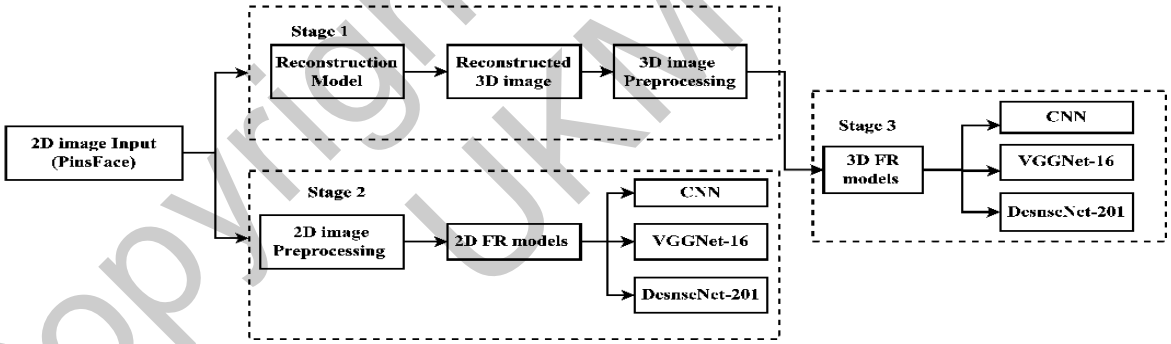

Figure 1. Proposed research methodology overview.

The methodology flowchart begins by reading the input 2D face image dataset to stage 1 and stage 2, following which the output of stage 1 will be inputted to stage 3 as the 3D face image dataset. For stage 2 and stage 3, there exists a similar flowchart procedure which is adhered to in this work, as illustrated in 1. The process begins with gathering the required dataset, which comes from the Facial Recognition Dataset available on Kaggle for the 2D FR. Subsequently, in stage 1, reconstruction and 3D pre-processing approaches are proposed to produce the 3D dataset from the 2D face image which is available. The dataset is separated into 20% for testing and 80% for training. The CNN algorithm, as well as another two models (VGG-16 and DenseNet-201), is used. Following the running of the Python code, for this study, we utilise two of the most common metrics, employed in almost all of the

Copyright@FTSM UKM

literature, for evaluation of the efficiency of FR; the said metrics are the accuracy and the loss of the FR model.

1. DATASET

This study utilised a new 2D face image dataset known as Pins Face Recognition (PinsFace), which is available on Kaggle (Burak 2020). Such a dataset is used in this work for two reasons: the first is to avoid employing the same datasets that were used in the literature, and which have been exhausted by previous researchers. The second reason is that there is the need for a dataset which contains all 2D challenges, e.g. alignment, age variation, pose variation, occlusion, lighting, and so on. The PinsFace dataset is gathered from the images posted on Pinterest by using Scrapper-Bot, designed using Python and Selenium. The total number of image faces is 17,534 faces from 105 subjects; the images are distributed unequally amongst each identity.

2. TRANSFER-LEARNING

The designing of a Deep-NN is complicated, as it is data hungry and very time consuming. Thus, this study adopts the transfer learning concept, whereby the transfer learning mechanism focuses on taking advantage of past knowledge regarding certain models that has been gained from solving related or different tasks. The aforementioned mechanism contributes significantly in addressing the key challenge of DL, namely data scarcity. More specifically, the first layers (convolution layers) of a deep neural network are not limited to specific datasets or tasks, and so those layers are able to be re-implemented with general datasets; this is achieved by adjusting the layers to a specific type of dataset.

In this work, transfer-learning is applied to the two employed models, namely VGG-16 and DenseNet-201, via transferring their pre-trained weights, which were learned by examining the ImageNet, CIFAR, and SVHN datasets. Subsequently, the remaining hyper-parameters of their architecture network layers are fine-tuned using our new hyper-parameter configuration with the used dataset (PinsFace).

3. PERFORMANCE METRICS

In this study, we utilise two of the most common metrics, employed in almost all of the literature, for evaluation of the efficiency of FR; the said metrics are the accuracy and the loss of the FR model. Accuracy is a rate utilised to indicate how the output data match the original input data. In terms of its definition, accuracy is the percentage of the number of successfully predicted classes compared to the total-number of inputted images. It can be represented as shown in

Copyright@FTSM UKM

equation 1… Whilst Model loss is the most common metric used in DL, and shows how much the model fits with the input dataset. This can be calculated by finding the summation of the errors of the input dataset and dividing that by the total-number of the input data, as illustrated in equation 2…

$$Acc = \frac{Nr}{Nt} \quad \ldots (1)$$

Where Acc denotes the accuracy, Nr is the successfully predicated classes, and Nt is the total of the original data.

$$Model\_loss = \frac{Total\ No.of\ errors}{Total\ No.of\ input\ data} \quad \ldots(2)$$

IV. Experiment

The experimental environment specifications occupy a vital role in the implementation of DL techniques. Thus, before examining the results of this work, it is best to explicate the environment specifications which were utilised to implement the stages of the FR models. The Python language is employed as a programming language, due to its numerous benefits, e.g. wide community, simplicity, platform independence, and frameworks which fit very well with DL, e.g. Scikit-learn, Pandas, Keras, and TensorFlow. Environment platform: Google Colab notebook, Processor: GPU Tesla T4 from NVIDIA version 460.32.03, RAM Memory 12.6 GB. Storage and Google Drive with 200 GB.

Finally, two experiments are carried out: the first involves applying these three approaches with 2D FR, whilst the second pertains to 3D FR. Prior to this, the reconstruction process was implemented based on using the PRNet model to reconstruct the 3D facial-image from a solely one 2D face-image. The following sections illustrate and demonstrate the results. The 2D FR and 3D FR results are visualised by employing the matplotlib, which is a comprehensive library for visualising data interactively as figures in Python.

V. Result and Discussion

In order to reconstruct a 3D facial-image from a 2D facial-image, the PRNet model is employed in this stage. The 2D face image dataset is fed into the PRNet; the total number of images stands at 17,534 faces of 105 subjects, and those images are distributed unequally amongst each identity. Table 1. presents the results of the PRNet.

Copyright@FTSM UKM

Table 1. PRNet model results

Total no. of input images	Total no. of output images	Loss rate
17,534 2D face images	17,497 3D face images	0.0021

As seen in Table 1, the PRNet has been implemented in this work with high efficiency, and a low loss rate (equal to 0.0021). More concretely, this loss rate represents only 37 2D face images which the PRNet was unable to reconstruct into 3D facial images. After investigating this result, we discovered that numerous different challenges may be behind such a loss rate, e.g.: these 37 2D face images have high occlusion (hair covers most of the face area), low lighting (some images are almost black because of the dark), and pose variations. For these reasons, the PRNet has failed to detect the 2D face.

Figure 2. presents the comparison between the 2D and reconstructed 3D facial-images datasets in terms of the accuracy of these three models used. The accuracy results of 2D FR for custom CNN, VGG-16, and DenseNet-201 models are 81.13%, 83.76 %, and 89.72% respectively. Whilst, the accuracy results of 3D FR for the same models are 82.16%, 88.97%, and 93.33%. The simulation results illustrate that there exists an improvement in accuracy by 1.03%, 5.21 %, and 3.61 % for the CNN, VGG-16, and DenseNet-201 models respectively, from using 2D to 3D. This is due, as previously stated, to the increase in the facial features obtained when the 3D reconstruction dataset is employed, as well as to the background removing, since that reduces the noise in the model during training.

With regards to the obtained results, it is possible for use to infer that the transfer learning mechanism and the increasing number of layers (network depth) of the pre-trained model (DenseNet-201 and VGG-16), improved accuracy in 2D as well as in 3D FR. Above aforementioned factors, the accuracy of 3D FR was increased due to the high discriminative facial features that were extracted by the MediaPipe landmarks. The said 3D facial features proved their ability to handle the 2D image's challenges, e.g. occlusions, age-variation, pose-variation, as well as lighting etc. Such challenges constitute the main reason why the recognition achieved the lowest accuracy rate in the 2D FR experiment. It can thus be concluded that, all of these factors, together contribute to increasing the efficiency of the 3D FR more so than they do the 2D FR.

Additionally, the results of the DenseNet-201 model reveal higher and stronger generalisation capability when compared with other models. The simulation results also demonstrate that the DenseNet-201 model is the best for the 2D and 3D FR.

Copyright@FTSM UKM

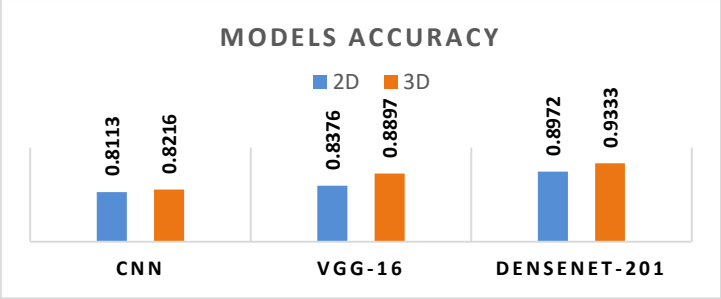

Figure 2. Models accuracy comparison

VI. Conclusion

FR is a popular and vital research domain in the areas of computer-vision, image processing, and pattern recognition. This is due to its noteworthy role in practical applications, e.g. in security, law enforcement, and commercial use, etc. Although FR has been ameliorated via the use of the DL techniques, it still faces challenges, such as 2D and 3D image issues, as well as the challenges of the DL techniques themselves. As such, this study proposed, and indeed adopted, number of approaches with a view to achieving the targeted research objectives and thus answering the specified key research questions.

A 3D reconstruction model (Stage 1) which consists of the PRNet model integrated with the MediaPipe tool to reconstruct 3D facial images from 2D facial images, to pre-process, and extract discriminative 3D facial features. Using this model, the research puts forth an accurate evaluation of 2D and 3D FR by employing the same dataset in both these processes. Regarding the obtained results, it is possible to conclude that the PRNet has been implemented in this work and achieved high efficiency as well as a low loss rate equal to 0.0021.

comparison framework between the 2D FR (Stage 2) and 3D FR (Stage 3), along with an extensive evaluation process. The proposed comparison was presented through applying the custom CNN, VGG-16, and DenseNet-201 models, once with a 2D dataset and secondly with a 3D dataset. This was achieved by conducting two experiments; each experiment involved using the same models in both 2D and 3D FR. The results highlight that the VGG-16 recorded results closer to the DenseNet-201 in 3D FR, but that the DenseNet-201 model outperformed the other models in both 2D and 3D FR, with the highest accuracy rates standing at 89.72% and 93.33% respectively. Based on these results, it can be stated that the DenseNet-201model is most suitable for FR in both 2D and 3D.

In terms of future work related to this study, there are several steps which could be applied in order to extend the work of this study, e.g. employing another large-scale dataset such as CASIA-WebFace, applying an additional Deep CNN-based model that may effectively increase the

Copyright@FTSM UKM

comparison bands such as ReseNet18, ResNet50, and ResNet164, utilising another 3D reconstruction approaches such as FR3DNet and employing 3D facial feature extraction approaches e.g. Subclass Discriminant Analysis (SDA) and PCA, to find the most suitable 3D reconstruction approach with optimal results. All of the above may lead to the discovery of the most accurate and suitable deep CNN models for FR and extending the comparison scale of this study, thus lowering the training cost, generating greater accuracy, and reducing loss of FR, whilst also generalising 3D reconstruction models for any large-scale dataset.

REFERENCES

Burak. 2020. Pins face recognition.

Cai, Y., Lei, Y., Yang, M., You, Z. & Shan, S. 2019. A fast and robust 3D face recognition approach based on deeply learned face representation. *Neurocomputing 363*: 375–397.

Feng, Y., Wu, F., Shao, X., Wang, Y. & Zhou, X. 2018. Joint 3D Face Reconstruction and Dense Alignment with Position Map Regression Network. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 11218 LNCS pp. 557–574.

Feng, Z.H., Huber, P., Kittler, J., Hancock, P., Wu, X.J., Zhao, Q., Koppen, P. & Raetsch, M. 2018. Evaluation of dense 3D reconstruction from 2D face images in the wild. *Proceedings - 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018*.

Fime, A.A., Sikder, D., Rabbi, J., Al-rakhami, M.S., Sen, O. & Fuad, M. 2021. Recent Advances in Deep Learning Techniques for Face Recognition (July).

Grishchenko, I., Ablavatski, A., Kartynnik, Y., Raveendran, K. & Grundmann, M. 2020. Attention Mesh: High-fidelity Face Mesh Prediction in Real-time.

He, K., Zhang, X., Ren, S. & Sun, J. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 2016-Decem pp. 770–778. IEEE.

Hu, H., Shah, S.A.A., Bennamoun, M. & Molton, M. 2017. 2D and 3D face recognition using convolutional neural network. *IEEE Region 10 Annual International Conference, Proceedings/TENCON 2017-Decem*: 133–138.

Hu, J., Shen, L. & Sun, G. 2018. Squeeze-and-Excitation Networks. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K.Q. 2017. Densely Connected Convolutional Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 154 pp. 2261–2269. IEEE.

Huang, Y.H. & Chen, H.H. 2022. Deep face recognition for dim images. *Pattern Recognition 126*: 108580.

Jeevan, G., Zacharias, G.C., Nair, M.S. & Rajan, J. 2022. An empirical study of the impact of masks on face recognition. *Pattern Recognition 122*: 108308.

Krizhevsky, A., Sutskever, I. & Hinton, G.E. 2012. ImageNet Classification with Deep Convolutional Neural Networks Alex. In *Advances in Neural Information Processing Systems*

Copyright@FTSM UKM

25. pp. 1097–1105.

Li, L., Mu, X., Li, S. & Peng, H. 2020. A Review of Face Recognition Technology. *IEEE Access 8*: 139110–139120.

Mu, G., Huang, D., Hu, G., Sun, J. & Wang, Y. 2019. Led3D: A lightweight and efficient deep approach to recognizing low-quality 3D faces. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

Nonis, F., Dagnes, N., Marcolin, F. & Vezzetti, E. 2019. 3D Approaches and Challenges in Facial Expression Recognition Algorithms—A Literature Review. *Applied Sciences 9*(18): 3904.

Parkhi, O.M., Vedaldi, A. & Zisserman, A. 2015. Deep Face Recognition (Section 3): 41.1-41.12.

Peng, S., Huang, H., Chen, W., Zhang, L. & Fang, W. 2020. More trainable inception-ResNet for face recognition. *Neurocomputing 411*: 9–19.

Prasad, P.S., Pathak, R., Gunjan, V.K. & Ramana Rao, H. V. 2020. Deep Learning Based Representation for Face Recognition. In Kumar, A. & Mozar, S. (eds.). Vol. 570 pp. 419–424. Singapore: Springer Singapore.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C. & Fei-Fei, L. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision 115*(3): 211–252.

Sabour, S., Frosst, N. & Hinton, G.E. 2017. Dynamic routing between capsules. In *Advances in Neural Information Processing Systems*. Vol. 2017-Decem pp. 3857–3867. Neural information processing systems foundation.

Shepley, A.J. 2019. Deep Learning For Face Recognition: A Critical Analysis.

Simonyan, K. & Zisserman, A. 2015. Very deep convolutional networks for large-scale image recognition. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings* 1–14.

Szegedy, C., Wei Liu, Yangqing Jia, Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. & Rabinovich, A. 2015. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 1–9. IEEE.

Wang, M. & Deng, W. 2021. Deep face recognition: A survey. *Neurocomputing 429*: 215–244.

Wang, X. & Zhang, W. 2021. Anti-occlusion face recognition algorithm based on a deep convolutional neural network. *Computers and Electrical Engineering 96*(PA): 107461.

Xu, R., Liu, X., Wan, H., Pan, X. & Li, J. 2021. A feature extraction and classification method to forecast the pm2.5 variation trend using candlestick and visual geometry group model. *Atmosphere 12*(5).

You, Z., Yang, T. & Jin, M. 2020. Multi-channel Deep 3D Face Recognition 1–9.

Yu, C., Zhang, Z. & Li, H. 2020. Reconstructing A Large Scale 3D Face Dataset for Deep 3D Face Identification.

Yu, S.G., Kim, S.E., Suh, K.H. & Lee, E.C. 2021. Face Spoofing Detection Using DenseNet. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 12616 LNCS pp. 229–238.

Zeiler, M.D. & Fergus, R. 2014. Visualizing and Understanding Convolutional Networks. In *Neurocomputing*. Vol. 187 pp. 818–833.

Copyright@FTSM UKM