Journal of Integrated and Advanced Engineering (JIAE)
Vol. 3, No. 1, March 2023: 37-46
http://asasijournal.id/index.php/jiae
http://doi.org/10.51662/jiae.v3i1.84

Learning a Multimodal 3D Face Embedding for Robust RGBD Face Recognition

Ahmed Rimaz Faizabadi[1,2], Hasan Firdaus Mohd Zaki[1,2]*, Zulkifli Zainal Abidin[1,2], Muhammad Afif Husman[1,2], Nik Nur Wahidah Nik Hashim[1,2]
[1]Department of Mechatronics, Kulliyyah of Engineering, International Islamic University, Malaysia
[2]Centre for Unmanned Technologies (CUTe), International Islamic University, Malaysia

Abstract
Machine vision will play a significant role in the next generation of IR 4.0 systems. Recognition and analysis of faces are essential in many vision-based applications. Deep Learning provides the thrust for the advancement in visual recognition. An important tool for visual recognition tasks is Convolution Neural networks (CNN). However, the 2D methods for machine vision suffer from Pose, Illumination, and Expression (PIE) challenges and occlusions. The 3D Race Recognition (3DFR) is very promising for dealing with PIE and a certain degree of occlusions and is suitable for unconstrained environments. However, the 3D data is highly irregular, affecting the performance of deep networks. Most of the 3D Face recognition models are implemented from a research aspect and rarely find a complete 3DFR application. This work attempts to implement a complete end-to-end robust 3DFR pipeline. For this purpose, we implemented a CuteFace3D. This face recognition model is trained on the most challenging dataset, where the state-of-the-art model had below 95% accuracy. An accuracy of 98.89% is achieved on the intellifusion test dataset. Further, for open world and unseen domain adaptation, embeddings learning is achieved using KNN. Then a complete FR pipeline for RGBD face recognition is implemented using a RealSense D435 depth camera. With the KNN classifier and k-fold validation, we achieved 99.997% for the open set RGBD pipeline on registered users. The proposed method with early fusion four-channel input is found to be more robust and has achieved higher accuracy in the benchmark dataset.

Keywords:
3D Face Recognition;
Expression Invariant;
Facial Depth Maps;
Intel RealSense;

Article History:
Received: August 22, 2022
Revised: September 24, 2022
Accepted: October 24, 2022
Published: March 10, 2023

Corresponding Author:
Hasan Firdaus Mohd Zaki,
Department of Mechatronics,
Kulliyyah of Engineering,
International Islamic University,
Centre for Unmanned Technologies (CUTe), International Islamic University, Malaysia
Email: hasanzaki@iium.edu.my

INTRODUCTION

The current wave of Industrial Revolution 4.0 (IR 4.0) will mainly rely upon machine vision to drive the need for industrial automation. Human and robot coworkers, known as collaborative robots or cobots, have collaborated to complete tasks in various environments. Cobots play a vital role in IR 4.0 revolution. However, many of the robots developed are blind. Hence high-precision machine vision will be a critical part of IR 4.0, making intelligent machines capable of interacting in collaborative environments and making decisions [1]. Many such applications require Face Analysis and Recognition [2][3].

Face Recognition (FR) is a method of identifying or validating a person's identity. The human face has highly non-rigid characteristics that have very discriminative features. Humans can identify each other with ease. Identifying faces from computers started as early as the 1960s and became popular with Eigenfaces [4] in the 1990s. It became famous as a non-invasive biometric with the advancement in technologies. Significant advancements in face recognition techniques can be grouped into four phases. The millstones can be called i) Holistic learning, ii) Locally handcrafted techniques, iii) Shallow learning, and iv) Deep learning. Phase-I uses holistic approaches. It dominated in the 1990s and spanned till early 2000 [5][6]. The locally handcrafted feature extraction became popular in early 2000 [7]. In the next decade, shallow learning with the local feature reached an accuracy of 95% on the LFW dataset [7]. The breakthrough in deep learning technology driven by improved computer hardware and algorithms and the availability of large datasets made a new revolutionary phase with the advent of AlexNet [8] in 2012. DeepFace [9] and DeepID [10] achieved state-of-the-art performance in 2014, and research has shifted to deep-learning-based approaches. It took three decades to increase shallow recognition from 60% to 90%. In comparison, deep learning took its performance to 99.8% using a deeper pipeline on the LFW dataset in just three years.

2D machine vision has many limitations [7], such as parallax. A parallax is an apparent displacement of an object due to a change in perspective and depth of focus. Other issues faced by FR are changing ambient light and variations in contrast. The 2D methods are also prone to spoofing or other attacks. The 3D data is rich in information. 3D cameras are becoming affordable and prevalent with the advancement in camera technology. This work proposes a complete, real-time, implementable 3D face recognition pipeline for practical use in this work.

The organization of the paper is Section 1 introduces face recognition, Section 2 provides a literature review and discusses advancements in 3DFR. Section 3 describes the methodology and components of the proposed system in detail. Section 4 the results of the proposed multimodal 3D deep face recognition model as a feature extractor for RGBD face recognition applications in an open world. Section 5 is the conclusion.

MATERIAL AND METHODS

Literature Review

In 2012, researchers began utilizing Deep Learning for visual tasks on ImageNet [8]. Deep CNN has a significant advantage over traditional processing methods of images and videos. In contrast, Recurrent Neural Network (RNN) processes continuous data such as voice and text [11][12]. Zhou et al. [13] proposed a real-time 3DFR system that employs a trained two-level cascade classifier and preprocesses RGB and depth data. Goswami et al. [14] suggest the unification of 2D and 3D information to accomplish a hybrid face recognition, applying techniques of entropy and saliency to construct a descriptor and utilizing geometrical analysis of 3D fiducial points.

Large-scale face datasets used for the train deep learning model improvise recognition accuracy. Deep learning models can learn facial features and depict rich internal data information with the assistance of large datasets. 2D face datasets on a massive scale can be done by data scraping from the internet. Due to the lack of large-scale 3D face datasets available, it is challenging to train discriminative in-depth features for 3D facial models compared to the 2D face dataset. To solve this problem, Kim et al. [15] proposed a frontal 3D scan, producing a 2.5D depth map and extracting the depth map features using the VGG16 network to represent the 3D face. VGG Face gave an excellent result on Bosphorus (99.240%). Except for the Bosphorus dataset, their results do not outperform the state-of-the-art conventional methods. Zhang et al. [16] proposed an expression and pose, invariant 3D face recognition. It directly takes 3D point clouds as input. However, its performance lacks considerably over FR3DNet with or without finetuning. It also needs an effective mechanism to handle the distribution gap between synthesized data and real 3D faces. A specialized Deep CNN model trained over a large dataset for 3D face recognition is proposed by Gilani et al. [17]. The FR3DNet uses the three-channel images generated from 3D point cloud data. However, Zhang et. al. [16] and Gilani et. al. [17] both use a synthetic 3D face dataset for training. FR3DNet uses two more maps than Kim et.al. [15]. Using more channels helps minimize the loss of 3D information but incurs additional memory and computation costs.

It should be remarked that 3D facial recognition is still an open field for improvement, either because it demands high computational power or lacks a large dataset to train algorithms or validate results. A complete survey of 3D facial recognition is presented in [18][19]. The proposed method should be a 3DFR application capable of running on embedded platforms like [20]. Modern 2D face recognition applications use large training datasets of millions of images and challenging testing datasets benchmarks. However, face recognition applications are deployed with different scenarios in the unconstrained real world and deal with unseen data. Generalized face recognition is more challenging and less studied. The generalized face recognition system should deal with unseen domains without updating or finetuning deep learning models. In 3D, such a face recognition application is rarely attempted in literature. Without making any assumptions about the target domain, this work aims to investigate and improve upon how 3D modalities can play a part in the construction of a generalised face recognition system.

Methodology

The proposed work methodology is accomplished in the following stages. First, the CNN architecture for 3D face is designed, trained, and finetuned on a challenging dataset. Later, the robust 3D face recognition model developed in the previous stage is used for developing a classifier for an open-set application gallery. Then the final 3D deep FR pipeline for the open world is implemented for effective recognition. Each stage is discussed in the following subsections.

CNN Architecture

The backbone of CuteFace3D is like ResNet-50 architecture. Since we use the residual units in our CNN architecture illustrated in Figure 1, it takes four-channel input of 224*224 instead of 3 channel input images, as shown in Figure 1. The last layer uses the most widely used SoftMax layer to classify 1200 identities from the intellifusion dataset. The SoftMax loss is presented as follows:

$$L_1 = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}} \tag{1}$$

Where $x_i \in \mathbb{R}^d$ is a deep extracted feature of the *i-th* sample of the class y_i, the embedding from the previous layer (avgpool) can be extracted for open-set recognition challenges with a vector size of 2048. These embeddings need to be more discriminative. On the other hand, the softmax loss does not optimize the embedding quality to achieve a higher similarity to interclass variations or diversities. From the literature, the gap is evident when SoftMax is used in deep CNN for face recognition. The intraclass variations [21, 22, 23, 24] are not handled effectively using the SoftMax function. This situation suits studying the extra channel used in the 3D multichannel face recognition. It will be evident if embedding quality is highly discriminative despite being trained on SoftMax loss due to the fusion of depth channel with RGB. Without using any specialized loss functions, such ArcFace [25]. Alternatively, without the special learning metrics used in SphereFace [26] for learning large margin features to estimate the discriminant power of the depth image fusion to RGB. This is one of the primary motives of this work to identify the significance of additional channel input in multichannel 3D face recognition.

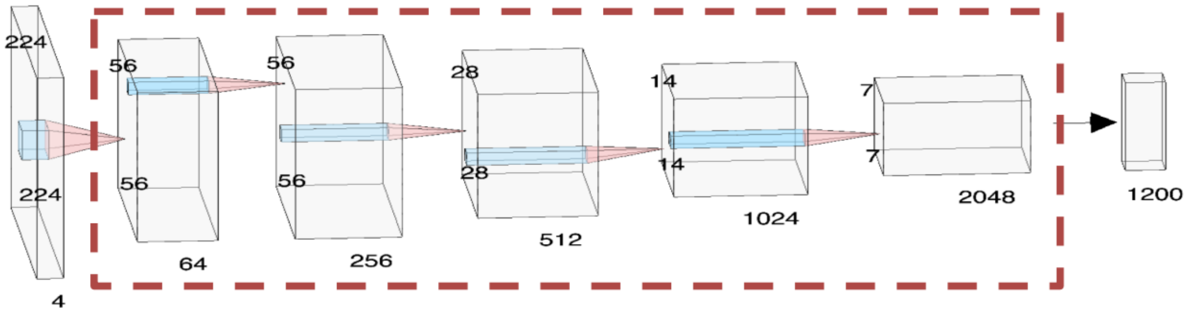

Figure 1. CNN architecture with four channel input for an early fusion of depth and RGB images [27]

Figure 1 shows that the parameters are not substantially increased compared to ResNet-50 after modifications. The overall size of the model has been increased only by nearly 2MB. The ReLU activation function is used to assess the impact of fusion alone. Further can experiment with CReLU or PReLU as used in previous studies. Most 3D face recognition applications use training datasets of a few thousand scans and small testing datasets benchmarks. In this attempt of 3DFR, a training dataset of nearly four hundred thousand RGB-D scans and over forty thousand test images from the intellifusion dataset are used. The Intellifusion dataset is described in next.

Training and Finetuning

The CNN-based FR model with multimodal images is developed using the Intellifusion dataset. The dataset used is a high-quality dataset that uses different domains: age, ethnicity, expression, and occlusion. The training phase of CuteFace3D is illustrated in Figure 2. Training faces preprocessed using MTCNN are fed to CNN with a SoftMax layer. The Adam optimizer is utilized with a learning rate of 0.001. After every seven epochs, the learning rate decreases by 0.1. The model is trained for 50 epochs, and test accuracy is calculated using SoftMax for final prediction. Named this 3DFR model called CuteFace3D as a reference to the Center for Unmanned Technologies (CUTe).

Further, as shown in Figure 3, the trained CuteFace3D model is used by dropping fully connected layers (FC) to extract RGBD face embeddings. The extracted embeddings of RGBD face scan with deep features are a vector of 2048 size. The embedding for each scan in the gallery is fed as input to the classifier. Then the similarity or dissimilarity metric or classifier can be employed, as shown in Figure 3. This work will compare the results using classifiers such as KNN for different values of k. If this is robust and discriminative, it is worthy of use in unseen domains and open world scenarios.

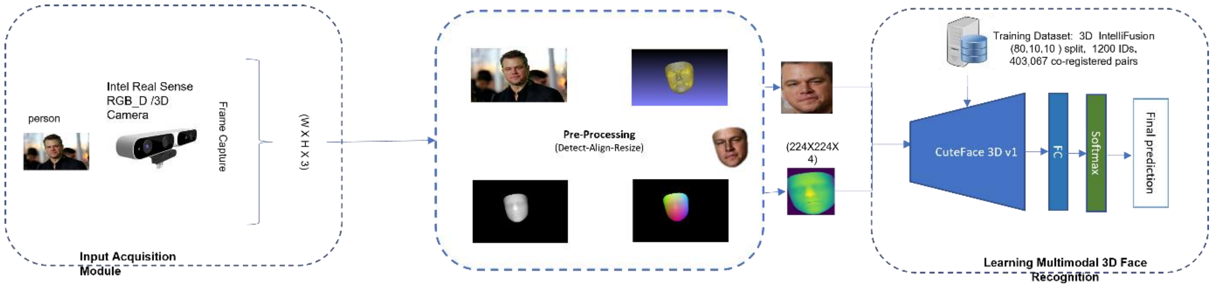
person
Intel Real Sense RGB_D /3D Camera
Frame Capture
(W X H X 3)
Input Acquisition Module
Pre-Processing
(Detect-Align-Resize)
(224X224X 4)
Training Dataset: 3D IntelliFusion (80,10,10) split, 1200 IDs, 403,067 co-registered pairs
CuteFace 3D v1
FC
Softmax
Final prediction
Learning Multimodal 3D Face Recognition

Figure 2. Training of CuteFace3D

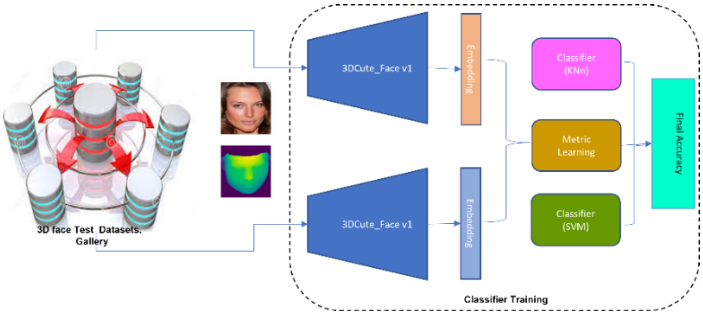

Figure 3. Learning from 3D face embeddings for classifier training

3D Deep Face Recognition Pipeline

The proposed methodology of a robust 3D face recognition system comprises four modules, namely a) Image acquisition module, b) Feature extraction module, c) Classifier module, and d) Inference module, as depicted in the system diagram illustrated in Figure 4.

The acquisition module captures a video stream from the RealSense camera using *pyrealsense2* python wrapper. The depth and RGB streams are aligned and co-registered for further consideration. When aligned pairs of frames are available, they are converted to *NumPy* arrays. A depth map is converted to an 8-bit color map. Then *dlib* deep face detector is applied to capture the face in the frame. If the face is found, it is cropped, preprocessed, and stored as registered with other user details. Fifty frames are extracted in every user registration. The inference module can also use the acquisition module before applying face recognition.

The feature extraction module comprises a novel 3DCuteFace deep learning model using multimodal learning as described. The registered user scans are used to extract a feature vector size of 2048. The feature vectors from the gallery are used for training a classifier, and a best-trained classifier is deployed for the face recognition task. The inference module acquires the image for inference as mentioned in the acquisition module and invokes the inference engine. The captured depth map and face scan will be fed to 3DCuteFace, and embedding is extracted. The embeddings are fed to the classifier model for final prediction. Thus, 3D face recognition in the real world can be achieved in an open-world environment.

Training Dataset

An Intellifusion RGB-D dataset contains 403,067 pairs of face images of 1,205 people. Each pair of face images is registered and includes RGB and depth images. It was issued during the international 3D face recognition algorithm challenge 2019. It incorporates huge variations. A few challenges of PIE are shown in Figure 5. Depth images are not shown here as they will be more distinct only for expression and extreme pose with no impact of illumination or background clutter.

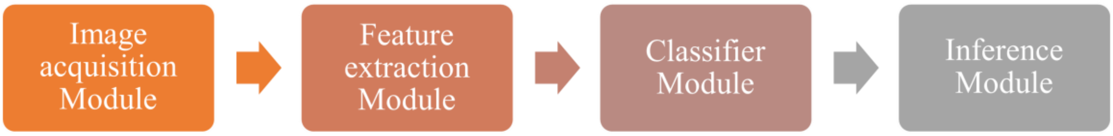

Figure 4. Face recognition pipeline using 3D embeddings from CuteFace3D for classifier and inference on real-time video streams.

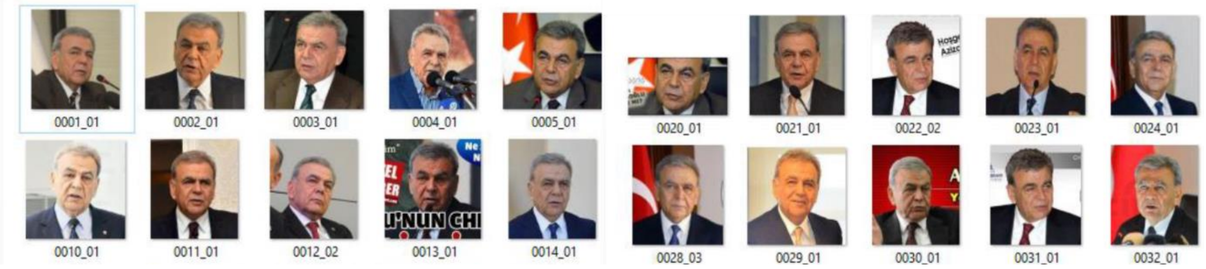

Figure 5. RGB Images with different challenges related to pose, expression, illumination, and occlusion for the same ID.

The exact train test split mechanism by X. Xiong et al. [26] is adopted. That is 90-10 split performed for training and testing, respectively. The IDs with less than ten samples were excluded, and after cleaning the dataset, a total of 361,799 face scans from 1200 identities were used for training. For the test dataset, 40,809 registered pairs were separated using the closed set approach. It means identities in the test set will always be present in the training set. The split list for the test set is 10%, validation and training set is 90% to keep results comparable.

Application Data

The data from the RealSense camera is collected for making a 3D face recognition system. The RGB and depth stream are recorded in an uncontrolled environment. For each subject, 50 frames are captured in the gallery. After the registration of users is completed. An FR pipeline is completed by training the SVM and KNN classifier for registered users. Figure 6(a) and Figure 6(b) show the raw input of depth and RGB images. The preprocessed depth map and RGB face of a subject are illustrated in Figure 6(c) and Figure 6 (d).

RESULTS

CuteFace3D converges well with training and finetuning of the model with hyperparameters described in Section 3.2. The loss and accuracy of training can be seen in Figure 7. The orange line is for training, and the blue line indicates validation loss and accuracy, respectively. Figure 7(a) indicates the loss function of CuteFace3D on the training and validation set, and Figure 7(b) shows the accuracy of CuteFace3D on the training and validation set for {Citation}up to 14 epochs. The training is carried out for 50 epochs. It took over seven days on a single GPU TitanXp system. After 50 epochs, the training accuracy was 99.67%, with a loss of 0.0379, evaluation accuracy of 99.77%, and a loss of 0.221. In the future, the CNN can be trained for different tasks and challenges and will be ensembled for robust face analysis.

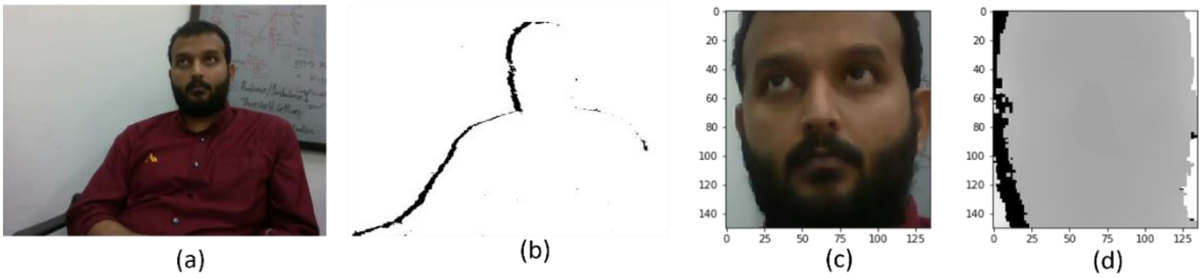

Figure 6. a) Raw RGB image b) Depth map captured from RealSense technology c) Preprocessed RGB image of a subject in face recognition system gallery d) Preprocessed face depth map without any reconstruction.

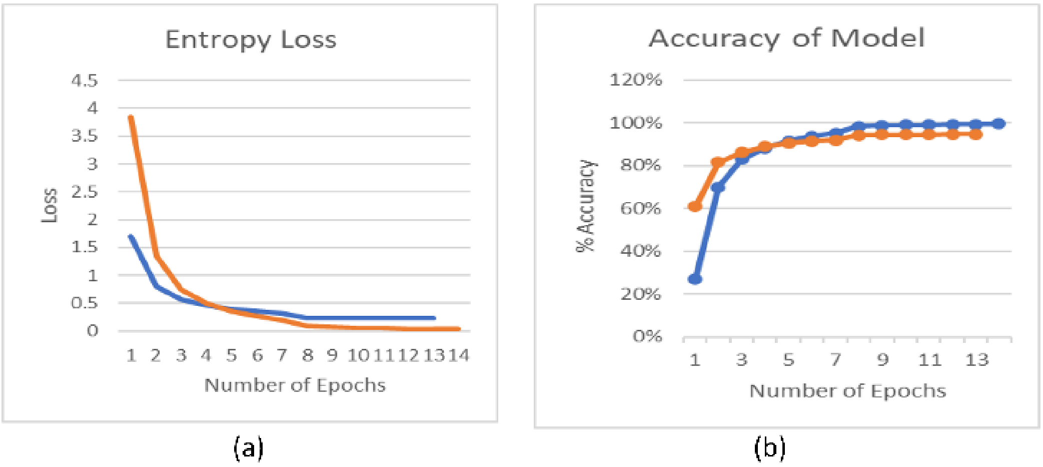

Figure 7. a) Training and validation loss b) Training and validation accuracy of CuteFace3D

The results obtained for an Intellifusion RGB-D dataset surpassed the performance of the most advanced method found in the research literature, as seen in Table 1. The CuteFace3D model outperforms model [28] by approximately 16% and Model (A) [27] by over 10%. Model (B) uses pre-trained weights and has an accuracy of 94.64 percent. The CuteFace3D has a 4.25 percent lower error rate than Model (B).

The application gallery was collected for about 85 subjects, as described in section 3.5, and RGBD embeddings were extracted for 3DFR application development using the CuteFace3D model, as described in Figure 3. Then k-fold validation KNN is applied for classifier training. The KNN accuracy trained on the gallery population with error rate is shown for 10-fold validation, as illustrated in Figure 8. 3D face embedding of registered user gallery using KNN with 10-fold validation achieved an accuracy of 99.997%.

Table 1. Accuracy of 3D face recognition model using Intellifusion RGB-D dataset

Sl No.	*Reference*	*Accuracy*
1	Tongyan Gong [27]	82%
2	X. Xiong et al. [26] (A)	88.36%
3	X. Xiong et al. [26] (B)	94.64%
4	CuteFace3D V1 (Ours)	98.89%

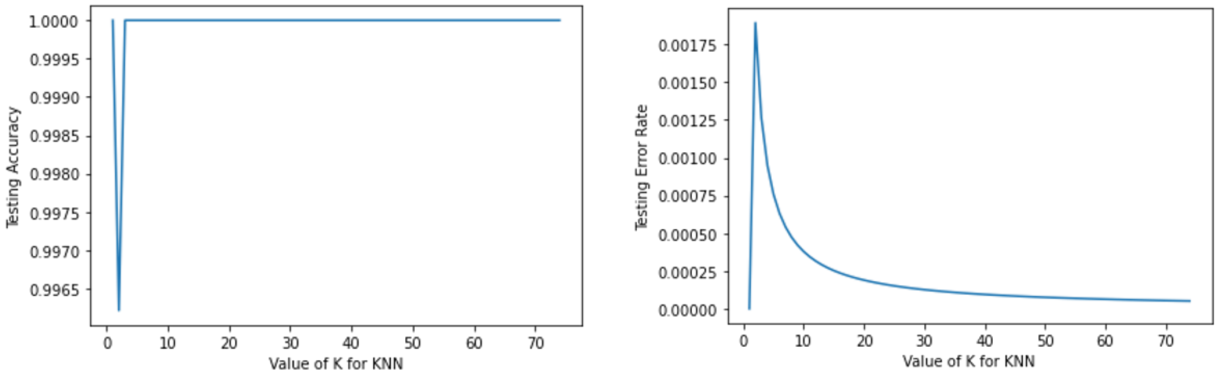

Figure 8. KNN Classifier result for FR population

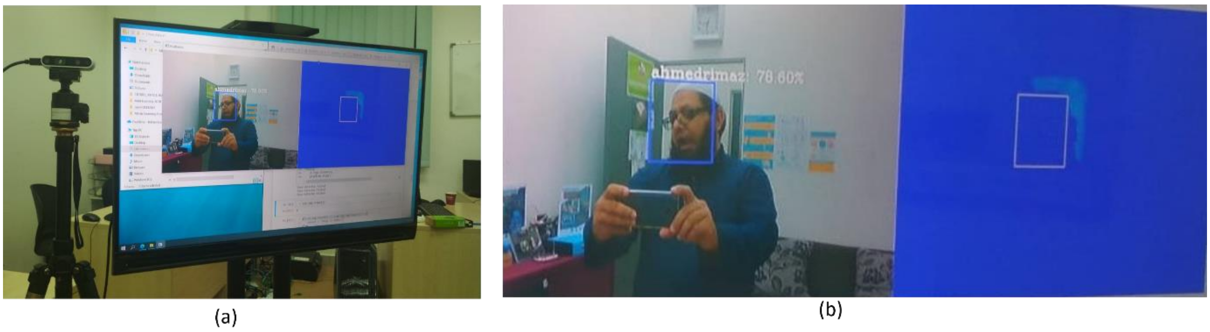

Figure 9. a) Development setup of 3D face Recognition, the raw RGB image, and depth map are captured from RealSense camera technology D435 b) Output with face recognition in the presence of extreme pose and expression along with RGB image and depth map.

The application development setup with the output of face recognition using a proposed system is shown in Figure 9(a) and its recognition with depth map in Figure 9(b). The person in action is very similar to the user represented in Figure 6 in terms of facial hair and outlook. In addition, the spectacles, extreme expression, pose, and skull cap is also introduced. In the gallery, there was no image of a person with such extreme expression or open mouth. Despite such extreme variations, a subject has been recognized with very fair accuracy. Similar results were observed in all the registered users of the 3DFR application.

CONCLUSION

A complete, robust 3DFR pipeline is successfully implemented and tested. The proposed work has achieved an accuracy of 98.89%, with an improvement of over 4% from the state-of-the-art. The improvements are achieved by tweaking the CNN architecture for early RGB and depth fusion. This method is found to be more discriminative. The model is improvised with finetuning the hyperparameters such as Adam optimizer and using PReLU. The proposed 3DFR can effectively work with an extreme pose, expression, self-occlusion, facial hairs, and spectacles. The effectiveness is achieved through detailed experiments on embedding vector size and found the size 2048 optimal in the case of 3DFR applications in an open set. A domain adaptation pipeline uses an embedding of 2048 size. 3D face embedding of registered user gallery using KNN with 10-fold validation surpassed the accuracy of 99.997%. So, it has well demonstrated that the proposed 3DFR model can be used more effectively in a practical 3DFR real-time application pipeline. The RGBD camera with low resolution and low-quality depth map using RealSense D435 can work effectively without incurring the additional computational cost for reconstruction or quality enhancement.

ACKNOWLEDGMENT

This research is funded by the Ministry of Higher Education of Malaysia (FRGS19-159-0768). We also thank NVIDIA for donating the GPU used for this research.

REFERENCES

[1] A. Hentout, M. Aouache, A. Maoudj, and I. Akli, "Human–robot interaction in industrial collaborative robotics: a literature review of the decade 2008–2017," *Advanced Robotics*, vol. 33, no. 15–16, pp. 764–799, 2019, doi: 10.1080/01691864.2019.1636714

[2] O. Bongomin, G. G. Ocen, E. O. Nganyi, A. Musinguzi, and T. Omara, "Exponential Disruptive Technologies and the Required Skills of Industry 4.0: A Review," *Journal of Engineering*, vol. 2020, ID: 4280156, 2020, doi: 10.1155/2020/4280156

[3] C. Filippini *et al.*, “Facilitating the Child–Robot Interaction by Endowing the Robot with the Capability of Understanding the Child Engagement: The Case of Mio Amico Robot,” *International Journal of Social Robotics*, vol. 13, no. 2, pp. 1–13, 2020, doi: 10.1007/s12369-020-00661-w

[4] M. A. Turk and A. P. Pentland, "Face recognition using eigenfaces," *Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Maui, HI, USA, 1991, pp. 586-591, doi: 10.1109/CVPR.1991.139758.

[5] P. N. Belhumeur, J. P. Hespanha and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: recognition using class specific linear projection," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711-720, July 1997, doi: 10.1109/34.598228.

[6] B. Moghaddam, W. Wahid and A. Pentland, "Beyond eigenfaces: probabilistic matching for face recognition," *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, Nara, Japan, 1998, pp. 30-35, doi: 10.1109/AFGR.1998.670921.

[7] M. Wang and W. Deng, “Deep Face Recognition: A Survey,” *Neurocomputing*, vol. 429, pp. 215–244, Mar. 2021, doi: 10.1016/j.neucom.2020.10.081.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.

[9] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, “DeepFace: Closing the Gap to Human-Level Performance in Face Verification,” in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2014, pp. 1701–1708. doi: 10.1109/CVPR.2014.220.

[10] Y. Sun, Y. Chen, X. Wang, and X. Tang, “Deep Learning Face Representation by Joint Identification-Verification,” in *Advances in Neural Information Processing Systems*, 2014, vol. 27.

[11] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.

[12] A. Ashraf, A. Sophian, A. A. Shafie, T. S. Gunawan, N. N. Ismail, A. A. Bawono, " Detection of Road Cracks Using Convolutional Neural Networks and Threshold Segmentation," *Journal of Integrated and Advanced Engineering (JIAE)*, vol. 2, no. 2, pp. 123-134, 2022, doi: 10.51662/jiae.v2i2.82

[13] W. Zhou, J. Chen, and L. Wang, “A RGB-D face recognition approach without confronting the camera,” in *2015 IEEE International Conference on Computer and Communications (ICCC)*, Oct. 2015, pp. 109–114. doi: 10.1109/CompComm.2015.7387550.

[14] G. Goswami, M. Vatsa, and R. Singh, “RGB-D Face Recognition with Texture and Attribute Features,” *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 10, pp. 1629–1640, Oct. 2014, doi: 10.1109/TIFS.2014.2343913.

[15] D. Kim, M. Hernandez, J. Choi, and G. Medioni, “Deep 3D face identification,” in *2017 IEEE International Joint Conference on Biometrics (IJCB)*, Denver, CO, Oct. 2017, pp. 133–142. doi: 10.1109/BTAS.2017.8272691.

[16] Z. Zhang, F. Da, and Y. Yu, “Data-Free Point Cloud Network for 3D Face Recognition,” *ArXiv191104731 Cs*, Nov. 2019, Accessed: May 04, 2020. [Online]. Available: http://arxiv.org/abs/1911.04731

[17] S. Zulqarnain Gilani and A. Mian, “Learning from Millions of 3D Scans for Large-Scale 3D Face Recognition,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, Jun. 2018, pp. 1896–1905. doi: 10.1109/CVPR.2018.00203.

[18] S. Zhou and S. Xiao, “3D face recognition: a survey,” *Human-centric Computing and Information Sciences*, vol. 8, no. 1, p. 35, Dec. 2018, doi: 10.1186/s13673-018-0157-2.

[19] J. R. Barr, K. W. Bowyer, P. J. Flynn and S. Biswas, “Face Recognition from Video: An overview”, *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 26, no. 5, 2012, doi: 10.1142/S0218001412660024

[20] M. A. Zulkhairi, Y. M. Mustafah, Z. Z. Abidin, H. F. M. Zaki, and H. A. Rahman, “Car Detection Using Cascade Classifier on Embedded Platform,” in *2019 7th International Conference on Mechatronics Engineering (ICOM)*, Oct. 2019, pp. 1–3. doi: 10.1109/ICOM47790.2019.8952064.

[21] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou, “AgeDB: The First Manually Collected, In-the-Wild Age Database,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jul. 2017, pp. 1997–2005. doi: 10.1109/CVPRW.2017.250.

[22] 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 Conference on Applications of Computer Vision (WACV)*, Mar. 2016, pp. 1–9. doi: 10.1109/WACV.2016.7477558.

[23] T. Zheng and W. Deng, “Cross-Pose LFW: A Database for Studying Cross-Pose Face Recognition in Unconstrained Environments,” *Technical Report 18–01*, Beijing University of Posts and Telecommunications, pp. 1-6, 2018

[24] T. Zheng, W. Deng, and J. Hu, "Cross-Age LFW: A Database for Studying Cross-Age Face Recognition in Unconstrained Environments." *arXiv*, Aug. 28, 2017. Accessed: Aug. 11, 2022. [Online]. Available: http://arxiv.org/abs/1708.08197

[25] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, pp. 4685–4694. doi: 10.1109/CVPR.2019.00482.

[26] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "SphereFace: Deep Hypersphere Embedding for Face Recognition," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, Jul. 2017, pp. 6738–6746. doi: 10.1109/CVPR.2017.713.

[27] X. Xiong, X. Wen, and C. Huang, "Improving RGB-D face recognition via transfer learning from a pretrained 2D network," in *International Symposium on Benchmarking, Measuring and Optimization*, 2019, pp. 141–148, doi: 10.1007/978-3-030-49556-5_14

[28] T. Gong and H. Niu, "An Implementation of ResNet on the Classification of RGB-D Images," in *Benchmarking, Measuring, and Optimizing*, Cham, 2020, pp. 149–155. doi: 10.1007/978-3-030-49556-5_15.