Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 07, No. 04, December 2021.

UNCONSTRAINED EAR RECOGNITION USING TRANSFORMERS

Marwin B. Alejo

(Received: 3-Aug.-2021, Revised: 5-Sep.-2021, Accepted: 12-Sep.-2021)

ABSTRACT

The advantages of the ears as a means of identification over other biometric modalities provided an avenue for researchers to conduct biometric recognition studies on state-of-the-art computing methods. This paper presents a deep learning pipeline for unconstrained ear recognition using a transformer neural network: Vision Transformer (ViT) and Data-efficient image Transformers (DeiTs). The ViT-Ear and DeiT-Ear models of this study achieved a recognition accuracy comparable or more significant than the results of state-of-the-art CNN-based methods and other deep learning algorithms. This study also determined that the performance of Vision Transformer and Data-efficient image Transformer models works better than that of ResNets without using exhaustive data augmentation processes. Moreover, this study observed that the performance of ViT-Ear is nearly like that of other ViT-based biometric studies.

KEYWORDS

Deep learning, Neural networks, Transformers, Vision transformer, Data-efficient image transformers, Ear recognition.

1. INTRODUCTION

Biometric recognition is an information system technology that allows identifying any person by his/her unique personal characteristics. Several studies use the common unique traits of an individual, such as fingerprint [1]–[3], face [4], [5], iris [6]–[8], iris and voice [9], [10], gait [11], [12] and ECG and EEG [13]–[15] for biometric recognition. However, recent studies suggested using ears for biometric recognition due to its advantages over using each of these common biometric traits [16]–[19]. Ear-based biometric recognition is both a science and technology that identify and authenticate individuals by their ear images in a constrained or unconstrained environment [20]. This method gained a momentum of interest in computational method research and application due to many advantages over other forms of biometric recognition. Although ear recognition offers numerous advantages over fingerprint, iris and face, it still faces significant levels of difficulty and challenges in unconstrained environments [21]-[22].

Modern studies utilize image processing algorithms, machine learning techniques or the fusion of both for the computational method of ear-based biometric recognition. One of these papers that utilizes these algorithms is Kavipriya et al. [23]. Similar to the enhanced method of Cheribet and Mazouzi [24], their method uses the canny edge detection algorithm and contour tracking method for ear biometric and personal identification. The paper of Mangayarkarasi et al. [25] proposed the same ear recognition method, but using only the contour method. The study of Jiddah and Yurtkan [26] presented an ear recognition method utilizing the used ear image dataset's fused geometric and texture features. The works of Zarachoff et al. [27] presented the 2D Wavelet-based Multi-Band PCA (2DWMBPCA) method, inspired by PCA (Principal Component Analysis) –a machine learning technique– for an ear-based biometric recognition. The paper of Sajadi et al. [28] utilized the genetic algorithm to extract the local and global features of ear images for ear recognition. While these ear biometric methods achieved exemplary results, most recent studies suggested using deep learning algorithms –a machine learning technique– in developing an ear-based biometric recognition method.

Deep learning algorithms are the most prevalent technology applied in the computational studies of ear recognition methods. Most of these deep learning studies utilize an improved architecture to learn from a single image [29]. The paper of Khaldi et al. [30] proposed the use of deep unsupervised active learning for ear recognition using the AMI (Mathematical Image Analysis), USTB2 (University of

M. Alejo is with Electrical and Electronics Engineering Institute, University of the Philippines – Diliman, Quezon City, Philippines and also is with Computer Engineering Department, National University, Manila, Philippines. Emails: marwin.alejo@eee.upd.edu.ph and mbalejo@national-u.edu.ph, ORCID:0000-0002-0926-1261.

Science and Technology Beijing), AWE (Annotated Web Ears) datasets and GAN (Generative Adversarial Network) for image coloring. Their method achieved recognition rates of 100.00%, 98.33% and 51.25% on the used datasets. The works of Lei et al. [31] used the SSD_MobileNet_v1 model on USTB datasets and achieved a recognition accuracy of 99%. Ying et al. [32] designed a DCNN (Deep Convolutional Neural Network) architecture called ear-recognition-Net for the ear recognition task. Their approach achieved a recognition rate of 95% to 98%. Chowdhury et al. [33] proposed using a handcrafted neural network algorithm for robust ear recognition and achieved a recognition accuracy of 98.2%. The study of Alshazly et al. [34] proposed using pre-trained AlexNet, VGGNet, Inception, ResNet and ResNeXt models for unconstrained ear recognition EarVN1.0 dataset by transfer learning and fine-tuning. Their method determined that ResNeXt is the best model for the task with a recognition accuracy of 95.85%. On a similar note, the papers of Alejo and Hate [35] and Almisreb et al. [36] utilized transfer learning for unconstrained ear recognition tasks on pre-trained deep convolutional neural network models. The works of Alejo and Hate achieved the recognition accuracy of 97.3%, 93.3%, 96.7%, 94.7%, 100.00%, 96.7%, 87.3%, 86.7% and 81.3% on AlexNet, GoogLeNet, Inception-v3, Inception-ResNet, ResNet-18, ResNet-50, SqueezeNet, ShuffleNet and MobileNet, while 100% was obtained on AlexNet in the works of Almisreb et al., one of the newest developed algorithms of deep learning. Although unrelated to ear recognition, the papers of Zhong and Deng [37] and George and Marcel [38] are among the early studies that adapted TNN or Transformer Neural Network as part of the method of their face recognition pipeline. Moreover, no published or presented study proposed the use of Transformer Neural Network for ear recognition; hence, an open opportunity.

Inspired by the facts and studies above, this paper aimed to investigate the effectiveness of the Transformer Neural Network on unconstrained ear recognition in terms of recognition accuracy performance. Furthermore, this paper (1) provided a deep learning pipeline for unconstrained ear biometric recognition by ViT (Vision Transformer) and DeiT (Data-efficient image Transformer) models and (2) compared the recognition accuracy performance of ViT and DeiT with the recognition accuracy performance of other methods based on deep learning, particularly the CNN.

The organization of the rest of this paper is as follows: Section 2 of this paper discusses Transformers; Section 3 discusses the Transformer-driven deep learning pipeline of this paper; Section 4 compares and discusses the results of this paper with the results of other relevant studies and Section 5 discusses the conclusion of this study.

2. TRANSFORMERS AND THEIR VISION-CENTRIC MODELS

Transformer Neural Network (or Transformers) is a novel deep learning technique developed by Vaswani et al. [39] with a self-attention mechanism as its core. It is a simple and scalable solution that exceeds the state-of-the-art results of the architectures based on RNN (Recurrent Neural Network) and CNN (Convolutional Neural Network) on NLP tasks (Natural Language Processing). The ongoing effort of several studies extended Transformers onto Computer Vision tasks [40], allowing the introduction of deep learning models, like DETR (Detection Transformer) [41] and Deformable DETR [42] for object detection, Axial-DeepLab [43] and Cross-Model Self-Attention [44] for image segmentation, Image Transformer [45], Image GPT [46], Transformer-induced Biases [47], TransGAN [48] and SceneFormer [49] for image generation and CLIP (Contrastive Language-Image Pre-training) [50], ViT (Vision Transformer) [51] and DeiT (Data-efficient image Transformers) [52] for object recognition. Furthermore, this paper focuses only on implementing Transformers through Vision Transformer and Data-efficient image Transformer models due to constraints with the used computational resources. The following subsections briefly discuss these two models of object recognition.

2.1 Vision Transformer (ViT)

The absence of an end-to-end object recognition architecture through Transformers provided an avenue for the development of Vision Transformer or ViT. Vision Transformer (ViT) is a brainchild algorithm of Dosovitskiy et al. [51] that employs a modified transformer network architecture to operate on images instead of text directly. Vision Transformer aims to provide an object recognition architecture without relying on CNN. Moreover, while Vision Transformer consumes extensive

computational resources during implementation over CNN [53], the works of Naseer et al. [54] emphasize that (1) ViT demonstrates strong robustness over occlusions, spatial patch-level permutations, adversarial perturbations and common natural signal corruptions for object recognition, (2) ViT performance for shape recognition is comparable to that of humans and (3) ViT exceptionally generalizes pre-trained ImageNet models or transfer learning for new domains of object recognition.

The Vision Transformer of this paper divides the input image into grid square patches flattened into a single vector by joining all the channels of pixels in a patch and linearly injecting each by the desired dimension. Moreover, this model employs a learnable position embedding into each patch and allows the Transformer network to learn the image's positional patch. Figure 1 shows the architecture of the Vision Transformer for unconstrained ear recognition (Vit-Ear) as adapted from the original paper [51]. Like the concept of transfer learning in CNN, the ViT architecture of this study replaced and fine-tuned the final layer to satisfy the recognition requirements accordingly.

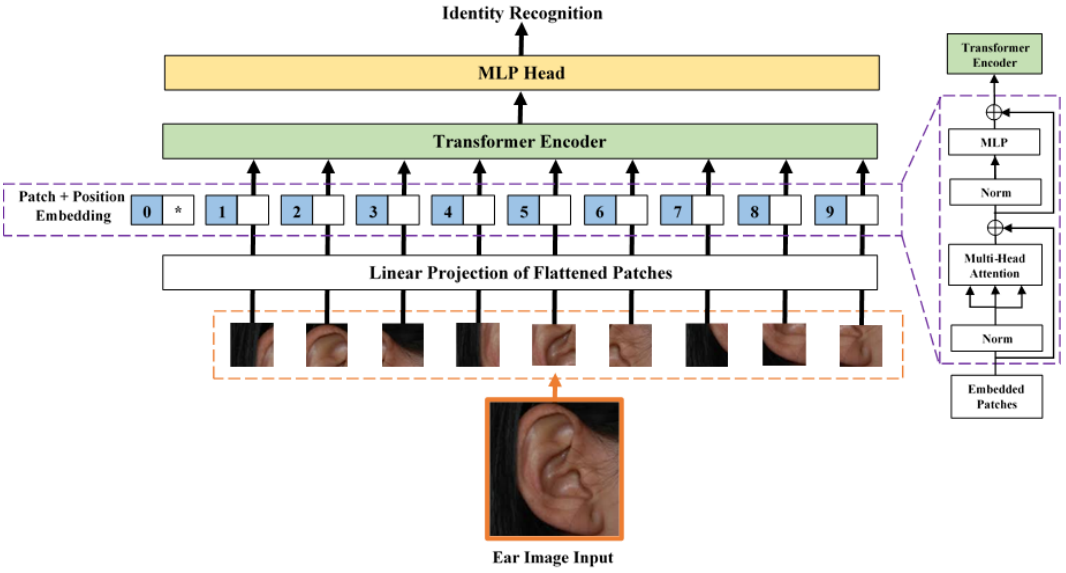

Figure 1. Vision transformer architecture of this study as adapted from the original paper [51].

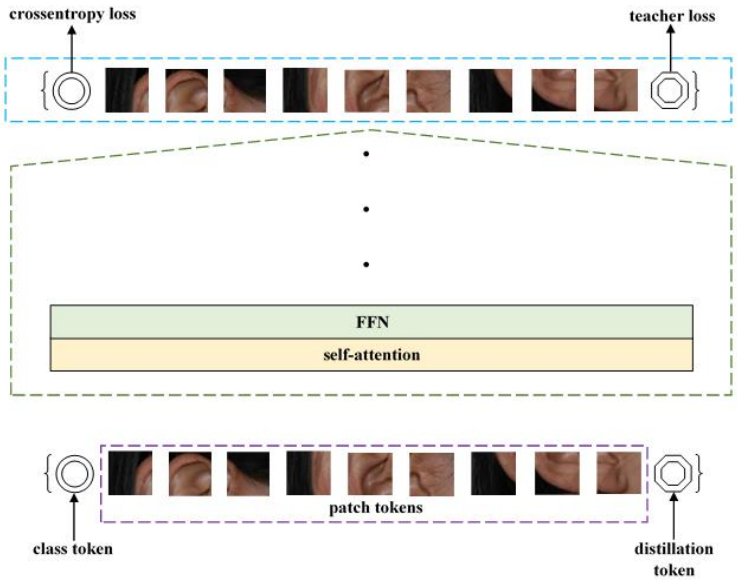

Figure 2. Data-efficient image transformer architecture of this study as adapted from the original paper [52].

2.2 Data-efficient Image Transformer (DeiT)

Data-efficient image Transformers (DeiTs) is another Transformer-based object recognition algorithm developed by Touvron et al. [52] with ViT in its core. DeiT aimed to overcome the excessive usage of

computational resources while exceeding the performance accuracy of CNN-based methods for object recognition. DeiT can maintain a ~60% accuracy on object recognition while zero accuracies were obtained for CNN on ImageNet task [54]. This is due to DeiT's use of a teacher-student strategy on its Transformer neural network to train directly on the used datasets. This teacher-student strategy of DeiT relies on distillation tokens to ensure that the student (model) learns through attention from the teacher. Figure 2 shows the DeiT architecture of this study (DeiT-Ear) as adapted from the original paper [52].

3. METHODOLOGY

The methodology of this study consisted of four subsequent phases: (1) Dataset and Input Data, (2) Data Preprocessing, (3) Training and Modeling and (4) Classification. Figure 3 visually shows these phases.

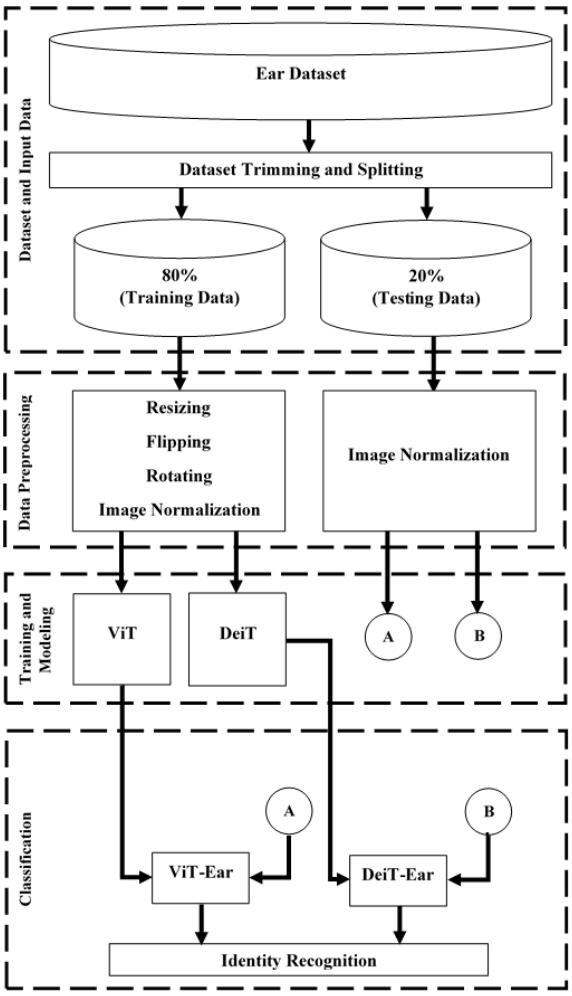

Figure 3. Deep learning pipeline of this study.

3.1 Dataset and Input Data

This study uses two different ear databases: (1) EarnVN1.0 dataset [55] and (2) UERC (Unconstrained Ear Recognition Challenge) dataset [56]. The EarVN1.0 dataset is the largest ear images dataset mainly collected for the task of ear recognition. It consisted of unprocessed ear images in the wild (unconstrained) of 164 individuals, with each having ~180 images for a totality of 28,412 ear images. The UERC dataset consisted of 3,300 ear images of 330 distinct identities. Due to the computational resources' constraint, this study considered only the first 20 classes of the EarVN1.0 dataset for a total of ~4000 ear images and the first ten classes of the UERC dataset for a total of 100 ear images. Furthermore, this study partitioned the trimmed dataset by 80% training and 20% testing dataset. The output of this phase is a set of training and testing datasets of the two databases. Figures 4 and 5 show samples of ear images of the used EarVN1.0 and UERC datasets.

Figure 4. Sample ear images of EarVN1.0 dataset.

Figure 5. Sample ear images of UERC dataset.

3.2 Data Pre-processing

This study pre-processed the partitioned training dataset by (a) resizing each ear image into 224 square pixels, (b) horizontal and (c) vertical flipping, (d) rotating by 30 degrees and (e) normalization using the standard ImageNet normalization values. Moreover, this study pre-processed the testing/validation dataset by resizing 224 square pixels and normalizing each resized ear image using the standard ImageNet normalization values. The output of this phase is a set of pre-processed training and testing/validation ear images. Figure 6 shows the sample output of these processes.

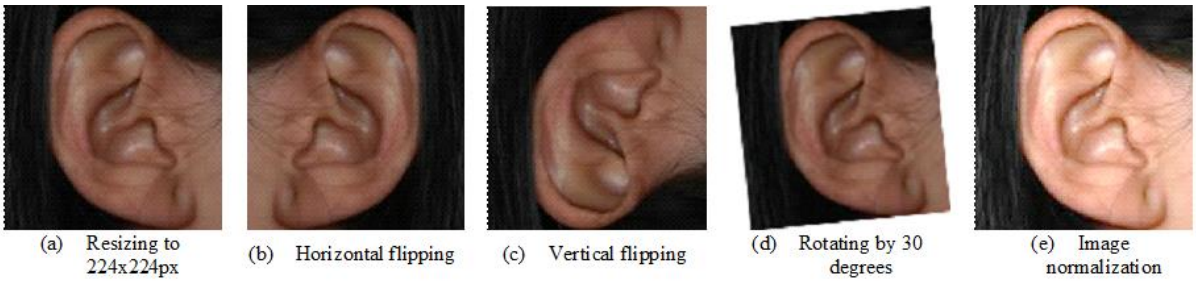

Figure 6. Sample output of each data processing process.

3.3 Training and Modeling – ViT and DeiT Implementation

Following both the description of ViT and DeiT in their respective papers [51], [52] and the context of transfer learning [50], [57], [58] due to the statistics of the used datasets, this study implemented these architectures with their pre-trained ImageNet-21k model and fine-tuned each of the architecture's final layers to only recognize 20 people from the used EarVN1.0 dataset and ten people from the used UERC dataset. This study implemented ViT using PyTorch-XLA on Google Colab TPU with eight as batch size, a learning rate of 0.00002, a gamma rate of 0.7 and 20 epochs. On the contrary, this study implemented the DeiT using PyTorch on Google Colab GPU with 32 as batch size, a learning rate of 0.001 and epoch values of 20, 30, 40 and 50. The used optimizer of this study in both ViT and DeiT is Adam. Table 1 summarizes the used training and modeling configuration of this study in implementing ViT and DeiT.

Table 1. Training and modeling configuration of this study.

Configuration	ViT	DeiT
Batch size	8	32
Learning rate	0.00002	0.001
Epoch	20	20, 30, 40, 50
Optimizer	Adam	Adam
Gamma	0.7	Not applicable

3.4 Classification

This phase utilizes the pre-processed testing datasets of this study. This phase aimed to determine the recognition accuracy of the trained unconstrained ear recognition models of this study on ViT and DeiT. The output of this phase is a comparative analysis of the recognition performance of ViT and DeiT in the context of unconstrained ear recognition over the recognition accuracy of other existing deep learning methods, like CNN. Since there are no published studies that proposed the use of Transformers for ear recognition, this paper considered comparing the results of this study with those of the existing CNN-based unconstrained ear recognition studies and considered transfer learning as the common ground.

4. Results and Discussion

4.1 ViT and DeiT on EarVN1.0

This paper's trained unconstrained ear recognition model on Vision Transformer (ViT) achieved a training accuracy of 100.00% and a recognition accuracy of 95.31% with a loss of 26.36%. The implementation of this model took 20 minutes and 40 seconds to train the ViT model on the preprocessed EarVN1.0 training dataset. This study also observed that overfitting occurs when training the ViT beyond 20 epochs. Overfitting is a phenomenon when the observed recognition accuracy is higher than the observed training accuracy [59].

On the contrary, the implementation of DeiT on the preprocessed EarVN1.0 dataset took ~15 minutes. The trained DeiT model of this study achieved a training accuracy of 100.00% in all the specified epoch configurations and a recognition accuracy of 88.33% with a loss of 1.4% on 20 epochs, 93.33% recognition accuracy with 1.02% loss on 30 epochs, 96.11% recognition accuracy with 0.88% loss on 40 epochs and 96.11% recognition accuracy with 0.69% loss on 50 epochs. This paper observed that recognition accuracy remains at 96.11% when training the DeiT model beyond 50 epochs.

4.2 ViT and DeiT on UERC

Given that the used UERC dataset of this study consisted of only ten subjects with each having ten images, the ViT model of this paper on the UERC dataset achieved a training accuracy of 100.00% and a recognition accuracy of 96.48% with a loss of 20.08%. The implementation of this model took ~16 minutes to train on the preprocessed UERC training dataset. Like the observations on the implementation of ViT on the EarVN1.0 dataset, overfitting occurs in this ViT implementation when training beyond 20 epochs.

The implementation of DeiT on the preprocessed UERC dataset took ~10 minutes. It achieved a training accuracy of 100% in all the epoch configurations and a recognition accuracy of 94.45% with a loss of 0.98% on 20 epochs, 97.81% recognition accuracy with a loss of 0.93% on 30 epochs and 100.00% recognition accuracy on 40 to 50 epochs with a loss of 0.43% to 0.51%. Overfitting also occurred in this implementation when training on epochs beyond 50.

4.3 Comparative Results

These recognition results of ViT and DeiT are closely comparable to the results of the relevant studies on state-of-the-art CNN-based methods through transfer learning. The results of this paper achieved a comparable result to the works of Lei et al. [31], Alshazly et al. [34], Alejo and Hate [35] and Almisreb et al. [36]. The performance of the trained DeiT model of this paper on EarVN1.0 on 20 epochs is similar to the recognition accuracy performance of the SqueezeNet and ShuffleNet models of Alejo and Hate, while the trained ViT and DeiT models on EarVN1.0 on 30 to 50 epochs and

UERC on 20 epochs are comparable to the performance of inception-based, ResNet-50 and MobileNet models of Alejo and Hate and the ResNext model of Alshazly et al. The trained DeiT models on UERC with 30 to 50 epochs achieved a comparable result over the SSD_MobileNet model of Lei et al. and the AlexNet and ResNet-18 models of Alejo and Hate and Almisreb et al. Furthermore, Alejo and Hate took 2.5 hours to develop their ResNet models for the unconstrained ear recognition task considering extensive data augmentations to achieve 100% recognition accuracy, while the transformer network of this paper took ~10 minutes to achieve the same recognition accuracy without relying heavily on data augmentation. Hence, this paper also proved the claim of Chen et al. [60] that Vision Transformers can achieve similar or more excellent results than ResNets even without extensive data augmentation. Table 2 shows the comparative results of the method of this study with those of the other related studies.

Table 2. Comparative results of this study over the results of methods of other CNN-based studies.

Study	**Dataset**	**Common Method**	**Method**	**Results in % (Recognition Accuracy)**
This study	EarVN1.0	Transfer Learning	Vision Transformer	95.31
			Data-efficient image Transformers	88.33 @ 20 epochs
				93.33 @30 epochs
				96.11 @ 40-50 epochs
This study	UERC	Transfer Learning	Vision Transformer	96.48
			Data-efficient image Transformers	94.45 @ 20 epochs
				97.81 @30 epochs
				100.00 @ 40-50 epochs
Lei et al. [27]	USTB2 (University of Science and Technology Beijing)	Transfer Learning	SSD_MobileNet_v1	99.00
Alshazly et al. [30]	EarVN1.0 [49]	Transfer Learning	ResNeXt	95.85
Alejo and Hate [31]	Handcrafted (own)	Transfer Learning	AlexNet	97.30
			GoogLeNet	93.30
			Inception-v3	96.70
			Inception-ResNet	94.70
			ResNet-18	100.00
			ResNet-50	96.70
			SqueezeNet	87.30
			ShuffleNet	86.70
			MobileNet	91.30
Almisreb et al. [32]	Handcrafted (own)	Transfer Learning	AlexNet	100.00

5. CONCLUSION

This paper investigated the use of Transformers in unconstrained ear recognition, particularly the Vision Transformer (ViT) and Data-efficient image Transformers (DeiT). This paper also provided a deep learning pipeline that employed these models. Like the concept of transfer learning on pre-trained state-of-the-art CNN architectures, this study replaced the final layer of ViT and DeiT to enable the Transformer network to learn the features from the extracted training ear images of the EarVN1.0 and UERC datasets. The ViT-Ear or Vision Transformer on the unconstrained ear recognition model of this study achieved a recognition accuracy of 95.31% on EarVN1.0 dataset and 96.48% on UERC dataset. The DeiT-Ear or Data-efficient image Transformers on this paper's unconstrained ear recognition model achieved a recognition accuracy of 88.33%, 93.33% and 96.11% on EarVN1.0 dataset and 94.45%, 97.81% and 100.00% on UERC dataset.

This paper determined that Transformers through ViT and DeiT achieved comparable or excellent results compared to state-of-the-art CNN-based methods for unconstrained ear recognition. Both the

ViT and DeiT achieved a similar recognition score of Inception-v3 and ResNet-50, but with faster modeling time, thus proving that Transformer networks work similarly or better than ResNets regardless of the particularity of the computer vision task. Additionally, this paper observed that the performance of ViT-Ear is like the recognition rate of a recently published face recognition method based on ViT, hence inferring that ViT might achieve an approximate 95% in biometric recognition studies regardless of the used modalities.

Future studies suggest exploring and investigating the performance of DeepVit (Deep Vision Transformer), CaiT (Class-Attention in Image Transformers), T2TViT (Tokens-to-Tokens Vision Transformer), CrossViT (Cross-Attention Multi-Scale Vision Transformer), PiT (Pooling-based Vision Transformer), LeViT (Vision Transformer in ConvNet's Clothing for Faster Inference) and CvT (Convolutions to Vision Transformers) on ear recognition and other biometric recognition modalities. Since this paper is an initial study of the ear recognition on Transformers and considering the limitations of conducting the experiments of this study, the proponent also suggested providing a comparative performance of these Transformer models over the results of this study.

ACKNOWLEDGMENTS

The author would like to acknowledge Professor Rowel Atienza of EEE Institute of the University of the Philippines – Diliman for his guidance and knowledge contribution towards the development of this study. Also, the author would like to acknowledge Dr. Rhandley Cajote for allowing the author to partake in the DSP Laboratory of the same institute. Moreover, this research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors and did not have conflict of interests from anybody.

REFERENCES

[1] S. Shi, J. Cui, X. L. Zhang, Y. Liu, J. L. Gao and Y. J. Wang, "Fingerprint Recognition Strategies Based on a Fuzzy Commitment for Cloud-Assisted IoT: A Minutiae-based Sector Coding Approach," IEEE Access, vol. 7, pp. 44803–44812, DOI: 10.1109/ACCESS.2019.2906265, 2019.

[2] I. Elzein and M. Kurdi, "Analysis of Embedded Fingerprint Biometric Recognition System Algorithm," Proc. of the 12th IEEE International Symposium on Advanced Topics in Electrical Engineering (ATEE 2021), DOI: 10.1109/ATEE52255.2021.9425124, Bucharest, Romania, Mar. 2021.

[3] M. H. Hersyah, D. Yolanda and H. Sitohang, "Multiple Laboratory Authentication System Design Using Fingerprints Sensor and Keypad Based on Microcontroller," Proc. of the IEEE International Conference on Information Technology Systems and Innovation (ICITSI 2020), pp. 14–19, DOI: 10.1109/ICITSI50517.2020.9264969, Bandung, Indonesia, Oct. 2020.

[4] M. Sahu and R. Dash, "Study on Face Recognition Techniques," Proc. of the 2020 IEEE Int. Conf. on Communication and Signal Processing (ICCSP 2020), pp. 613–616, Chennai, India, Jul. 2020.

[5] A. A. Sukmandhani and I. Sutedja, "Face Recognition Method for Online Exams," Proc. of the IEEE International Conference on Information Management and Technology (ICIMTech 2019), pp. 175–179, DOI: 10.1109/ICIMTECH.2019.8843831, Jakarta/Bali, Indonesia, Aug. 2019.

[6] C. S. Hsiao, C. P. Fan and Y. T. Hwang, "Design and Analysis of Deep-learning Based Iris Recognition Technologies by Combination of U-Net and EfficientNet," Proc. of the 9th IEEE Int. Conf. on Information and Education Technology (ICIET 2021), pp. 433–437, Okayama, Japan, Mar. 2021.

[7] H. D. Rafik and M. Boubaker, "A Multi Biometric System Based on the Right Iris and the Left Iris Using the Combination of Convolutional Neural Networks," Proc. of the 4th IEEE Int. Conf. on Intelligent Computing in Data Sciences (ICDS 2020), DOI: 10.1109/ICDS50568.2020.9268737, Fez, Morocco, Oct. 2020.

[8] S. D. Shirke and C. Rajabhushnam, "Biometric Personal Iris Recognition from an Image at Long Distance," Proceedings of the International Conference on Trends in Electronics and Informatics (ICOEI 2019), vol. 2019-April, pp. 560–565, DOI: 10.1109/ICOEI.2019.8862640, Apr. 2019.

[9] R. Giorgi, N. Bettin, S. Ermini, F. Montefoschi and A. Rizzo, "An Iris+Voice Recognition System for a Smart Doorbell," Proc. of the 8th IEEE Mediterranean Conference on Embedded Computing (MECO 2019), DOI: 10.1109/MECO.2019.8760187, Budva, Montenegro, Jun. 2019.

[10] O. Tymchenko, B. Havrysh, O. O. Tymchenko, O. Khamula, B. Kovalskyi and K. Havrysh, "Person

Voice Recognition Methods," Proc. of the IEEE 3rd Int. Conf. on Data Stream Mining and Processing (DSMP 2020), pp. 287–290, Aug. 2020.

[11] E. M. Owaidah, K. S. Aloufi and J. H. Alkhatib, "Gait Recognition for Saudi Costume Using Kinect Skeletal Tracking," Proc. of the 2nd Int. Conf. on Computer Applications and Information Security (ICCAIS 2019), DOI: 10.1109/CAIS.2019.8769552, Riyadh, Saudi Arabia, May 2019.

[12] H. M. L. Aung and C. Pluempitiwiriyawej, "Gait Biometric-based Human Recognition System Using Deep Convolutional Neural Network in Surveillance System," Peoc. Of IEEE Asia Conference on Computers and Communications (ACCC 2020), pp. 47–51, DOI: 10.1109/ACCC51160.2020.9347899, Singapore, Sep. 2020.

[13] R. Srivastva, A. Singh and Y. N. Singh, "PlexNet: A Fast and Robust ECG Biometric System for Human Recognition," Information Sciences, vol. 558, pp. 208–228, DOI: 10.1016/J.INS.2021.01.001, May 2021.

[14] M. Wang, K. Kasmarik, A. Bezerianos, K. C. Tan and H. Abbass, "On the Channel Density of EEG Signals for Reliable Biometric Recognition," Pattern Recognition Letters, vol. 147, pp. 134–141, DOI: 10.1016/J.PATREC.2021.04.003, Jul. 2021.

[15] W. Cui, Z. Wang and Y. Li, "ECG-based Biometric Recognition under Exercise and Rest Situations," Biomedical Engineering Advances, p. 100008, DOI: 10.1016/J.BEA.2021.100008, Jul. 2021.

[16] Z. Wang, J. Yang and Y. Zhu, "Review of Ear Biometrics," Archives of Computational Methods in Engineering, vol. 28, no. 1, pp. 149–180, DOI: 10.1007/S11831-019-09376-2, Nov. 2019.

[17] A. Abaza, A. Ross, C. Hebert et al., "A Survey on Ear Biometrics," ACM Computing Surveys (CSUR), vol. 45, no. 2, pp. 1-35, DOI: 10.1145/2431211.2431221, Mar. 2013.

[18] Ž. Emeršič, V. Štruc and P. Peer, "Ear Recognition: More than a Survey," Neurocomputing, vol. 255, pp. 26–39, DOI: 10.1016/J.NEUCOM.2016.08.139, Sep. 2017.

[19] L. P. Etter, E. J. Ragan, R. Campion, D. Martinez and C. J. Gill, "Ear Biometrics for Patient Identification in Global Health: A Field Study to Test the Effectiveness of an Image Stabilization Device in Improving Identification Accuracy," BMC Medical Informatics and Decision Making, vol. 19, no. 1, pp. 1–9, DOI: 10.1186/S12911-019-0833-9, Jun. 2019.

[20] B. Bhanu, "Ear Shape for Biometric Identification," Encyclopedia of Cryptography and Security, pp. 372–378, DOI: 10.1007/978-1-4419-5906-5_738, 2011.

[21] A. Kamboj, R. Rani and A. Nigam, "A Comprehensive Survey and Deep Learning-based Approach for Human Recognition Using Ear Biometric," The Visual Computer, vol. 2021, pp. 1–34, DOI: 10.1007/S00371-021-02119-0, 2021.

[22] S. Ntshangase and D. Mathekga, "Ear Recognition for Young Children," Proc. of the IEEE International Multidisciplinary Information Technology and Engineering Conference (IMITEC 2019), DOI: 10.1109/IMITEC45504.2019.9015852, Vanderbijlpark, South Africa, Nov. 2019.

[23] P. Kavipriya, M. R. Ebenezar Jebarani, T. Vino and G. Jegan, "Ear Biometric for Personal Identification Using Canny Edge Detection Algorithm and Contour Tracking Method," Materials Today: Proceedings, DOI: 10.1016/J.MATPR.2021.03.351, Apr. 2021.

[24] M. Cheribet and S. Mazouzi, "A New Adapted Canny Filter for Edge Detection in Range Images," Jordanian Journal of Computers and Information Technology (JJCIT), vol. 7, no. 3, pp. 278-291, DOI: 10.5455/JJCIT.71-1620428305, Sep. 2021.

[25] N. Mangayarkarasi, G. Raghuraman and A. Nasreen, "Contour Detection Based Ear Recognition for Biometric Applications," Procedia Computer Science, vol. 165, pp. 751–758, DOI: 10.1016/J.PROCS.2020.01.016, Jan. 2019.

[26] S. M. Jiddah and K. Yurtkan, "Fusion of Geometric and Texture Features for Ear Recognition," Proc. of the 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT 2018), DOI: 10.1109/ISMSIT.2018.8567044, Ankara, Turkey, Dec. 2018.

[27] M. Zarachoff, A. Sheikh-Akbari and D. Monekosso, "Single Image Ear Recognition Using Wavelet-based Multi-band PCA," Proc. of the 27th IEEE European Signal Processing Conference (EUSIPCO 2019), vol. 2019-September, DOI: 10.23919/EUSIPCO.2019.8903090, A Coruna, Spain, Sep. 2019.

[28] S. Sajadi and A. Fathi, "Genetic Algorithm Based Local and Global Spectral Features Extraction for Ear Recognition," Expert Systems with Applications, vol. 159, p. 113639, DOI: 10.1016/J.ESWA.2020.113639, Nov. 2020.

[29] S. F. A. Abuowaida and H. Y. Chan, "Improved Deep Learning Architecture for Depth Estimation from Single Image," Jordanian Journal of Computers and Information Technology (JJCIT), vol. 6, no. 4, pp. 434–445, DOI: 10.5455/JJCIT.71-1593368945, Dec. 2020.

[30] Y. Khaldi, A. Benzaoui, A. Ouahabi, S. Jacques and A. Taleb-Ahmed, "Ear Recognition Based on Deep Unsupervised Active Learning," IEEE Sensors Journal, Early Access, vol. 2021, DOI: 10.1109/JSEN.2021.3100151, 2021.

[31] Y. Lei, B. Du, J. Qian and Z. Feng, "Research on Ear Recognition Based on SSD-MobileNet-v1 Network," Proceedings of the Chinese Automation Congress, (CAC 2020), pp. 4371–4376, DOI: 10.1109/CAC51589.2020.9326541, Nov. 2020.

[32] T. Ying, W. Shining and L. Wanxiang, "Human Ear Recognition Based on Deep Convolutional Neural Network," Proc. of the 30th Chinese Control and Decision Conference (CCDC 2018), pp. 1830–1835, DOI: 10.1109/CCDC.2018.8407424, Jul. 2018.

[33] M. Chowdhury, R. Islam and J. Gao, "Robust Ear Biometric Recognition Using Neural Network," Proc. of the 12th IEEE Conference on Industrial Electronics and Applications (ICIEA 2017), vol. 2018-February, pp. 1855–1859, DOI: 10.1109/ICIEA.2017.8283140, Feb. 2018.

[34] H. Alshazly, C. Linse, E. Barth and T. Martinetz, "Deep Convolutional Neural Networks for Unconstrained Ear Recognition," IEEE Access, vol. 8, pp. 170295–170310, DOI: 10.1109/ACCESS.2020.3024116, 2020.

[35] M. Alejo and C. P. G. Hate, "Unconstrained Ear Recognition through Domain Adaptive Deep Learning Models of Convolutional Neural Network," International Journal of Recent Technology and Engineering, vol. 8, no. 2, DOI: 10.35940/ijrte.B2865.078219, 2019.

[36] A. A. Almisreb, N. Jamil and N. M. Din, "Utilizing AlexNet Deep Transfer Learning for Ear Recognition," Proc. of the 4th International Conference on Information Retrieval and Knowledge Management: Diving into Data Sciences (CAMP 2018), pp. 8–12, DOI: 10.1109/INFRKM.2018.8464769, 2018.

[37] Y. Zhong and W. Deng, "Face Transformer for Recognition," arXiv, arXiv:2103.14803v2, [Online], Available: https://arxiv.org/abs/2103.14803v2, Mar. 2021.

[38] A. George and S. Marcel, "On the Effectiveness of Vision Transformers for Zero-shot Face Anti-Spoofing," Proc. of IEEE International Joint Conference on Biometrics (IJCB), pp. 1–8, DOI: 10.1109/IJCB52358.2021.9484333, Shenzhen, China, 2021.

[39] A. Vaswani et al., "Attention Is All You Need," Advances in Neural Information Processing Systems, vol. 2017-December, pp. 5999–6009, [Online], Available: https://arxiv.org/abs/1706.03762v5, 2017.

[40] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan and M. Shah, "Transformers in Vision: A Survey," arXiv, [Online], Available: http://arxiv.org/abs/2101.01169, 2021.

[41] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov and S. Zagoruyko, "End-to-End Object Detection with Transformers," Proc. of the European Conference on Computer Vision, Part of the Lecture Notes in Computer Science Book Series, vol. 12346, pp. 213–229, [Online], Available: https://arxiv.org/abs/2005.12872v3, 2021.

[42] X. Zhu, W. Su, L. Lu, B. Li, X. Wang and J. Dai, "Deformable DETR: Deformable Transformers for End-to-End Object Detection," arXiv, [Online]. Available: https://arxiv.org/abs/2010.04159v4, Oct. 2020, Accessed: Aug. 02, 2021.

[43] H. Wang, Y. Zhu, B. Green, H. Adam, A. Yuille and L.-C. Chen, "Axial-DeepLab: Stand-alone Axial-attention for Panoptic Segmentation," Proc. of the European Conference on Computer Vision, Part of the Lecture Notes in Computer Science Book Series, vol. 12349, pp. 108–126, [Online], Available: https://arxiv.org/abs/2003.07853v2, 2021.

[44] L. Ye, M. Rochan, Z. Liu and Y. Wang, "Cross-modal Self-attention Network for Referring Image Segmentation," Proc. of the IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR), vol. 2019, pp. 10494–10503, DOI**:** 10.1109/CVPR.2019.01075, CA, USA, 2019.

[45] N. Parmar et al., "Image Transformer," Proc. of the 35th Int. Conf. on Machine Learning (ICML 2018), vol. 9, pp. 6453–6462, [Online], Available: https://arxiv.org/abs/1802.05751v3, Feb. 2018.

[46] M. Chen et al., "Generative Pretraining from Pixels," Proc. of the 37th International Conference on Machine Learning, pp. 1691–1703. [Online]. Available: http://proceedings.mlr.press/v119/chen20s.html, Nov. 2020.

[47] P. Esser, R. Rombach and B. Ommer, "Taming Transformers for High-resolution Image Synthesis," arXiv, [Online], Available: https://arxiv.org/abs/2012.09841v3, Dec. 2020.

[48] Y. Jiang, S. Chang and Z. Wang, "TransGAN: Two Pure Transformers Can Make One Strong GAN and That Can Scale Up," arXiv, [Online], Available: http://arxiv.org/abs/2102.07074, Feb. 2021.

[49] X. Wang, C. Yeshwanth and M. Nießner, "SceneFormer: Indoor Scene Generation with Transformers," arXiv, [Online], Available: https://arxiv.org/abs/2012.09793, Dec. 2020.

[50] A. Radford et al., "Learning Transferable Visual Models from Natural Language Supervision," arXiv, [Online], Available: http://arxiv.org/abs/2103.00020, Feb. 2021.

[51] A. Dosovitskiy et al., "An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale," arXiv, [Online], Available: http://arxiv.org/abs/2010.11929, Oct. 2020.

[52] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles and H. Jégou, "Training Data-efficient Image Transformers & Amp; Distillation through Attention," arXiv, [Online], Available: https://arxiv.org/abs/2012.12877, Dec. 2020.

[53] A. Bakhtiarnia, Q. Zhang and A. Iosifidis, "Single-layer Vision Transformers for More Accurate Early Exits with Less Overhead," arXiv, [Online], Available: https://arxiv.org/abs/2105.09121v1, May 2021.

[54] M. Naseer, K. Ranasinghe, S. Khan, M. Hayat, F. S. Khan and M.-H. Yang, "Intriguing Properties of Vision Transformers," arXiv, [Online], Available: https://arxiv.org/abs/2105.10497v2, May 2021.

[55] V. T. Hoang, "EarVN1.0: A New Large-scale Ear Images Dataset in the Wild," Data in Brief, vol. 27, p. 104630, DOI: 10.1016/J.DIB.2019.104630, Dec. 2019.

[56] Ž. Emeršič et al., "The Unconstrained Ear Recognition Challenge 2019 - ArXiv Version with Appendix," arXiv, [Online], Available: https://arxiv.org/abs/1903.04143v3, Mar. 2019.

[57] K. Weiss, T. M. Khoshgoftaar and D. D. Wang, "A Survey of Transfer Learning, " Journal of Big Data, vol. 3, no. 1, pp. 1–40, DOI: 10.1186/s40537-016-0043-6., Dec. 2016.

[58] B. Zoph, V. Vasudevan, J. Shlens and Q. V. Le, "Learning Transferable Architectures for Scalable Image Recognition, "Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 8697–8710, DOI: 10.1109/CVPR.2018.00907, 2018.

[59] X. Ying, "An Overview of Overfitting and Its Solutions," Journal of Physics: Conference Series, vol. 1168, no. 2, DOI: 10.1088/1742-6596/1168/2/022022, Mar. 2019.

[60] X. Chen, C.-J. Hsieh and B. Gong, "When Vision Transformers Outperform ResNets without Pretraining or Strong Data Augmentations," arXiv, arXiv: 2106.01548, [Online], Available: https://arxiv.org/abs/2106.01548, Jun. 2021.

ملخص البحث:

لقد وفّرت أفضلية الأُذُن في تمييز الأشخاص مقارنةً بغيرها من وسائل التّمييز مجالاً خصْباً للباحثين لإجراء الدّراسات على طُرُق الحوسبة. تقدّم هذه الورقة طريقةً قائمةً على التّعلُّم العميق للتّمييز غير المقيَّد للأشخاص عن طريق الأُذُن باستخدام شبكة المحوّلات العصبيّة: محوّل الصُّور (ViT) ومحوّلات الصُّور فعّالة البيانات (DeiTs).

النّماذج المستخدمة في هذه الدّراسة لتمييز الأشخاص عن طريق صُوَر الأُذُن حقّقت دقّة تمييزٍ مماثلةً أو أفضل من الطّرق المستخدَمة في هذا المجال والتي تستند على الشّبكات العصبيّة الالتفافيّة (CNNs) وغيرها من خوارزميات التّعلُّم العميق. كذلك بينت النّماذج المستخدمة في هذه الدّراسة أنها تعمل بصورة أفضل من حيث الأداء مقارنة بشبكات (ResNets) دون استخدام عمليات زيادة استنزافيّة للبيانات. علاوة على ذلك، لاحظت هذه الدّراسة أنّ أداء تقنية (ViT) المستخدمة في هذه الدّراسة كان مُقارباً لأداء مثيلاتها من التّقنيات المستخدمة في الدّراسات البيومتريّة الأخرى.