ISSN 2278-3091
Volume 9, No.1.1, 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse8091.12020.pdf
https://doi.org/10.30534/ijatcse/2020/8091.12020

Deep transfer learning for ear recognition: A comparative study

Ali Abd Almisreb[2], Nursuriati Jamil[1]*, Syamimi M. Norzeli[2], Norashidah Md Din[2]
[1]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Malaysia
[2]Institute of Energy Infrastructure, Universiti Tenaga Nasional, Jalan IKRAM-UNITEN, Kajang, Selangor, Malaysia
*lizajamil@computer.org

ABSTRACT

Transfer Learning is an efficient platform to solve problems with little amount of data. In this paper, the performance of three well-known Convolution Neural Network (CNN)-based learning model that are AlexNet, VGG16 and VGG19 for human identification based on ear images are compared. The respective convolution neural networks (CNNs) are fine-tuned to customize it to the ear images dataset. The last fully connected later is replaced with another fully connected layer to recognize 10 classes instead of 1000 classes. A total of 3,000 ear images are captured and augmented from 10 male subjects aged between 18 to 27 years old. To train the fine-tuned CNN –based networks, 2,500 images are used and the remaining 500 image are allocated for validation. The proposed fine-tuned CNN-based networks performed well in ear recognition as validation accuracy achieved 100% for all 10 male subjects.

Key words: Transfer learning ear recognition, AlexNet VGG16, VGG19

1. INTRODUCTION

Deep learning has become a leading research platform that has found its way to be applied in several research areas such as image classification [1] [3]-[5] and speech recognition. It has been increasingly applied to biometrics applications as well [2]. In particular, the fields of machine learning and pattern recognition have gained substantial recognition lately along with the use of Deep Convolutional Networks. As well known, the Deep Convolutional Networks are consists ofcompound layers of networks that are basically learned through each layer. However, the data extraction grows with each level of layers. For example, in face recognition, layers of the higher level maintain the identities of a person, which is invariant to alterations such as illumination and pose [2]. Therefore, how to employ the Deep Learning concept for identifying humans using limited ear dataset is the main challenge here [6] [7]. Generally, transfer learning provides the ability to utilize the pre-trained networks by fine-tuning it with domain-specific data. The primary role of transfer learning is to recycle knowledge attained in a previous training process, in order to boost the learning procedure in a new and more complex mission. In other words, transfer learning provides an appropriate key for speeding up the learning procedure in image classification [7], [8], image recognition [9], eye tracking [10] and gaming [11]. This concept is especially helpful for the challenging task of learning classifiers that needs to perform well when only few training examples are provided [12]. On the other hand, [13] mentioned that the learning of each new task begins from scratch. However, this requires huge amount of training data. Even though many machine learning methods also require learning, the training and testing datasets are sourced from the same feature domain. This might be challenging or in some cases are inapplicable as datasets are scarce. Transfer learning proposes a solution to reuse the previously learned knowledge to some problems where little information is offered and furthermore improving the learning task.

The rest of this paper is organized as follows: Section II discusses related literature on application of transfer learning using convolution neural network (CNN) in different domains. The latest literature related to ear recognition using deep learning is also presented. In Section III, the architecture of our CNN AlexNet is described by fine-tuning training parameter values such as weights of pre-trained network and layers. Results and discussion of the ear recognition is presented in Section IV followed by recommendation for future work in Section V.

2. RELATED WORK

The use of ear as a candidate for biometrics has slowly gained attention in the past decade. Experiments by Iannarelli [14] proved the uniqueness of human ears. Human ears are relatively static in size and structure throughout an individual's life. Ear images can be obtained remotely without any direct contact with the sensor compared to thumbprint and iris pattern. Compared to fingerprint, ears are organ that are passively used in daily life, therefore is less prone to injury. In addition, it is also unaffected by facial expression which is one issue in facial recognition. The uniqueness of ears and its advantages make the ear as an alternative candidate in biometrics [15]. While deep learning has been widely used in biometrics such as face and fingerprint recognition, ear recognition using deep learning is still scarce due to lack of large-scale datasets [6] [7]. At the time of writing, only four published work [6] [7] [16] [17] of employing deep learning for ear recognition are found, with two papers from the same team. Galdámez et al. [17] first adopted deep learning using CNN for ear recognition of video images that are taken in controlled environment. Since video streaming requires recognition to be done rapidly, a highly optimized CNN architecture is necessary. Even though the results are encouraging, the network training is done in a non-collaborative environment of controlled images. In [6], a new ear image dataset comprising images collected from the web with high degree of variability is introduced. The same authors [7] then developed a CNN-based model for ear recognition using limited training data that was gathered in challenging environment. They also compared the performance of AlexNet [18] architec-

ture and the 16-layer VGG model architecture [19]. Ear detection is also improved when [16] implemented Multiscale Fast Region-Based Convolution Neural Network (R-CNN) by including information of the ear location other than the morphological characteristics of the ear.

In this paper, three CNN-based deep transfer learning models are evaluated for the purpose of ear recognition. We propose transfer learning using AlexNet architecture as this model is recommended as a starting point for applying deep learning for all tasks. VGGNet has a similar architecture to AlexNet with more convolutional layers. Only VGG-16 and VGG-19 are experimented as they have the deepest layers in VGGNet configurations.

3. METHODOLOGY

In this section, the ear image data collection is described followed by the architecture of the pre-trained CNNs. Transfer learning in this paper is done by fine-tuning each of the pre-trained CNN and the added layers are presented.

3.1. Ear image dataset

Successful training of CNN requires huge amount of datasets as CNN architecture is consists of millions of parameters. Therefore, data augmentation is necessary to facilitate training in any CNN model. In this paper, ten undergraduate students from a university aged between 18 to 27 years old are used as the subjects for the data collections. Three hundred (300) images are captured using and Android smartphone from the left and right ear of each person, totalling to 3000 ear images. The images are captured under normal room lighting and outdoor using natural lighting. Variations of the ear images are produced by performing geometric transformation on them. The person may be sitting or standing and the distance of the person and camera varies. For some images, parts of the ear images are occluded by hairs and hairlines. For transfer learning, tens or hundreds of images are adequate to train the CNN model [20]. Twenty-five (250) ear images from each person is used for training, while five (50) images each are used as testing dataset. Therefore, a total of 2500 ear images (83.4%) is allocated for training and 500 images (16.4%) for validation and testing. Samples ear images in the dataset can be seen in Fig. 1. No other pre-processing operations are done to the image.

Figure 1: Ear images are captured and augmented from 10 male subjects

3.2. AlexNet

In this paper, we use AlexNet which is a Convolutional Neural Network that was developed by [18]. AlexNet CNN is chosen as it is the most studied CNN [21] and because of its suitable tradeoffs between speed and accuracy [21]. It contains 8 learned layers, 5 convolutional layers and 3 fully-connected layers. The last fully-connected layer connects to 1000 classes and the rest of the network are considered as a feature extractor. AlexNet can yield 4096-dimensional feature vector for each image, which contains the activations of the hidden layer immediately prior the output layer. This network accepts input image of size 227x227x3 as a requirement of the input image layer.

3.3. VGGNet

VGG model improved its architecture over AlexNet by using multiple 3X3 kernel-sized filters instead of the 11 and 5 kernel-sized filters employed by AlexNet in the first and second convolutional layers. With smaller filters, the depth of VGG model is increased enabling it to learn more complicated features. The width (the number of layers) of convolutional layers in VGG models is comparably small, beginning from 64 in the first layer and increased by factor of 2 after each max-pooling, reaching 512 at the final

3.3.1. VGG16

The network has 41 layers. There are 16 layers with learnable weights: 13 convolutional layers, and 3 fully connected layers [22]. VGG16 accepts images with a size of 224x224x3

3.3.2. VGG19

The network has 47 layers. There are 19 layers with learnable weights: 16 convolutional layers, and 3 fully connected layers [22]. This network accept input image with size of 224x224x3.

A summary of AlexNet, VGG16 and VGG19 configurations are presented in Table 1 with the fine-tuned layers specified in bold and discussed further in section 3.4.

3.4. Fine-tuning the CNN models

The earlier layers of the pre-trained CNNs are retained as fixed feature extractor for our ear dataset. The first step in transfer learning is to replace the last three layers in the pre-trained CNN with a set of layers that can classify 10 classes (i.e. to identify 10 subjects). Discussion of the fine-tuning are further elaborated in section 3.4.1 and 3.4.2.

Table 1: AlexNet [18], VGG16 [19] and VGG19 [19] in a nutshell

ConvNet Configuration		
AlexNet	VGG16	VGG19
input (227x227 RGB image)	**input (224x224 RGB image)**	**input (224x224 RGB image)**
conv11-96	conv3-64	conv3-64
maxpool	conv3-64	conv3-64
conv5-256	maxpool	maxpool
maxpool	conv3-128	conv3-128
conv3-384	conv3-128	conv3-128
conv3-384	maxpool	maxpool
conv3-256	conv3-256	conv3-256
maxpool	conv3-256	conv3-256
FC-4096	conv3-256	conv3-256
FC-4096	maxpool	conv3-256
FC-64	conv3-512	maxpool
FC-10	conv3-512	conv3-512
softmax	conv3-512	conv3-512
	maxpool	conv3-512
	conv3-512	conv3-512
	conv3-512	maxpool
	conv3-512	conv3-512
	maxpool	conv3-512
	FC-4096	conv3-512
	FC-4096	conv3-512
	FC-64	maxpool
	FC-10	FC-4096
	softmax	FC-4096
		FC-64
		FC-10
		softmax

3.4.1. Fine-tuning AlexNet

For Alexnet, we added a fully connected layer with filter size 64x64, in order to adopt with the new output (10 subjects). This is notated by FC64 in Table 1. A Rectified Linear Unit (ReLU) layer (i.e. softmax in Table I) is also added to improve the non-linear problem-solving ability as suggested by [23]. Furthermore, ReLU layer not only has the ability to get the network trained several times faster, it also does not produce any gradient vanishing effect due to activation non-linearities of sigmoid or tanh units [24]. One more fully connected layer is added with 10 output neurons (i.e. FC-10 in Table 1) to enable the classification of 10 subjects. In order to learn faster in the new layers than in the transferred layer, we initialized the weights in the last fully connected layer with 10. We also initialized the neuron biases in this layer with the constant 20. This initialization accelerates the early stages of learning by providing the ReLUs with positive inputs.

3.4.2. Fine-tuning VGG16 and VGG19

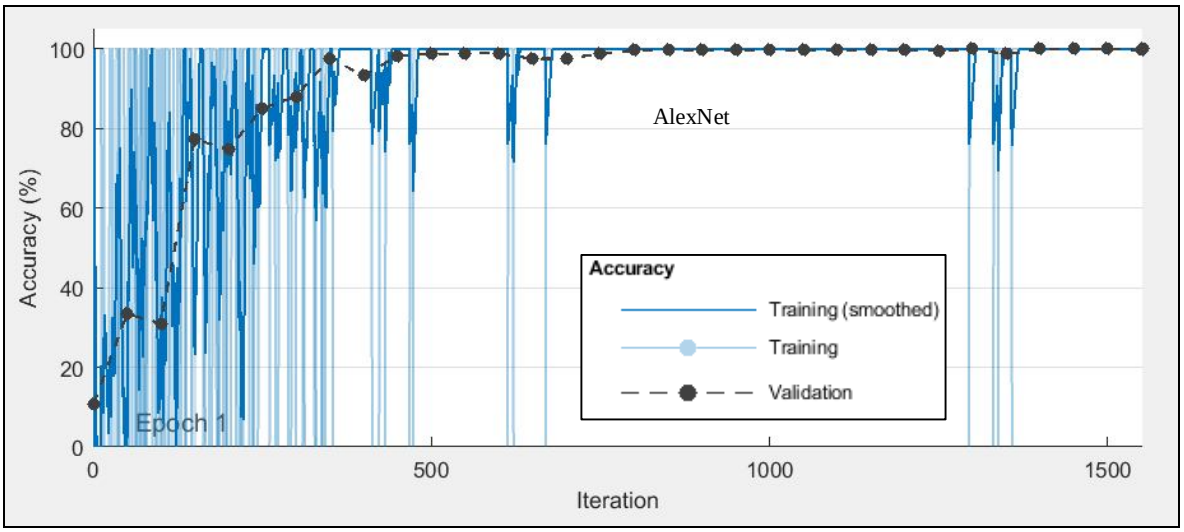
For VGG16 and VGG19, all the images are resized into 224x224x3 to correlate with input layer. Then, we added a fully connected layer with filter size 64x64, Softmax and classification output layer.

4. Results and discussions

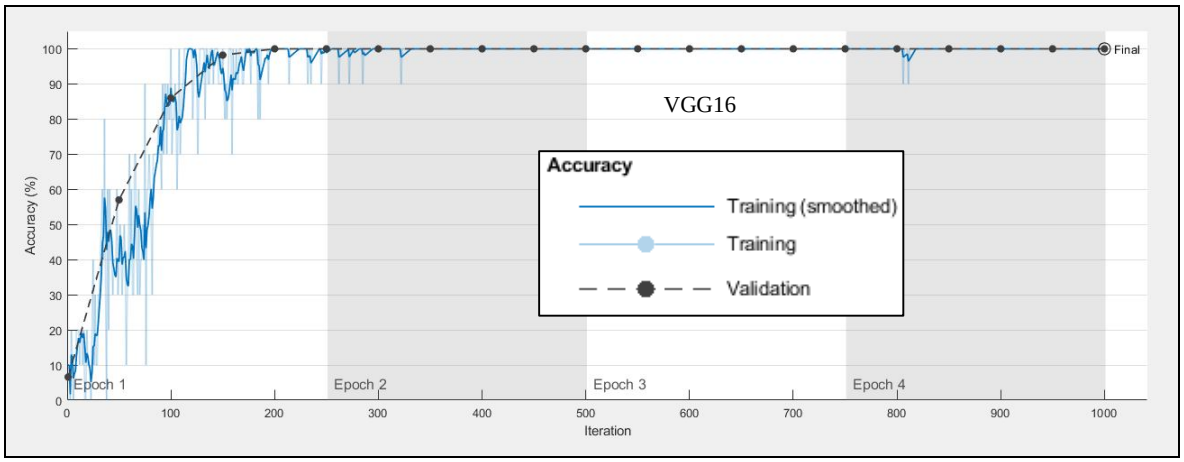
The training parameters for AlexNet, VGG16 and VGG 19 are set the same way for comparisons of their performances. The fine-tuned CNN models are trained using 250 ear images of each class. However, the training is conducted using Multiple CPUs. The minimum batch size is set to 1 to increase the training process and avoid the low memory issues. We setup the maximum training times to 20 and the number of iterations at 50000. However, the training finished after 1551 iterations as the validation accuracy stabilized. The minimum learning rate in this experiment is 0.0001 which is a small value to slow down learning in the transferred layers. The training function used in all models is Stochastic Gradient Descent with Momentum (SGDM), momentum of 0.9.

Fig 2(a) shows training and validation outcomes for AlexNet, while Fig. 2(b) and 2(c) show the outcomes for VGG16 and VGG19, respectively. In general, training momentums of all models begin slowly and achieved a low accuracy rate of less than 10%. Since ImageNet classes are largely on animals and objects, the database is not trained on ear images. Due to small minimum batch size, the training is very erratic in the beginning particularly for AlexNet. The training for AlexNet took near to 500 iterations to stabilize, while VGG16 stabilizes at 300 iterations and VGG19 at 200 iterations. This is not surprising as VGGNets have more layers compared to AlexNet. On the other hand, the validation accuracy increased dramatically over a short period of time due to the use of 50 images at each training time. All three CNN models achieved validation accuracy of 100%. AlexNet model achieved 100% accuracy at close to 700 iterations, while VGG16 achieved the feat at 200 iterations and VGG19 championed at 150 iterations. This confirmed what has been mentioned in the literature that transfer learning does not require too many images to produce successful classification.

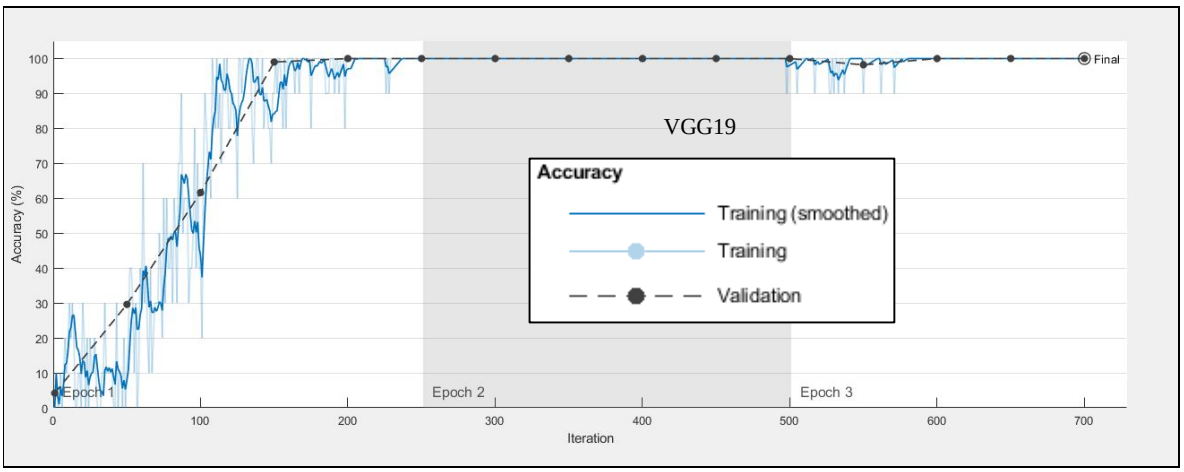
The loss rate is illustrated in Fig. 3(a) for AlexNet, 3(b) for VGG16 and 3(c) for VGG19. Notice that the loss ratio for all models is small (the maximum loss value is nearly equal to 21) even though the data set is small and this is the advantage of using transfer learning. Validation using 50 images is conducted and the confusion matrix is demonstrated in Fig. 4. As can be seen, all 10 subjects are correctly identified achieving validation accuracy of 100%.



(a)

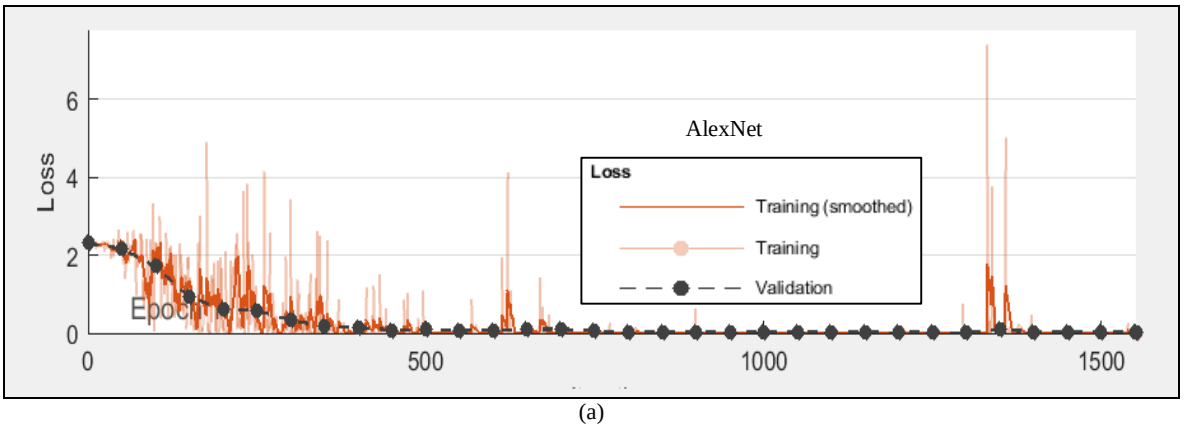


(b)

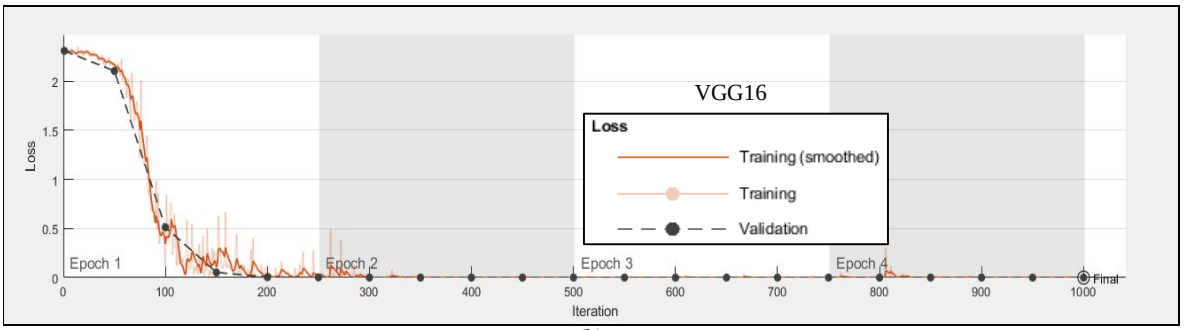


(c)

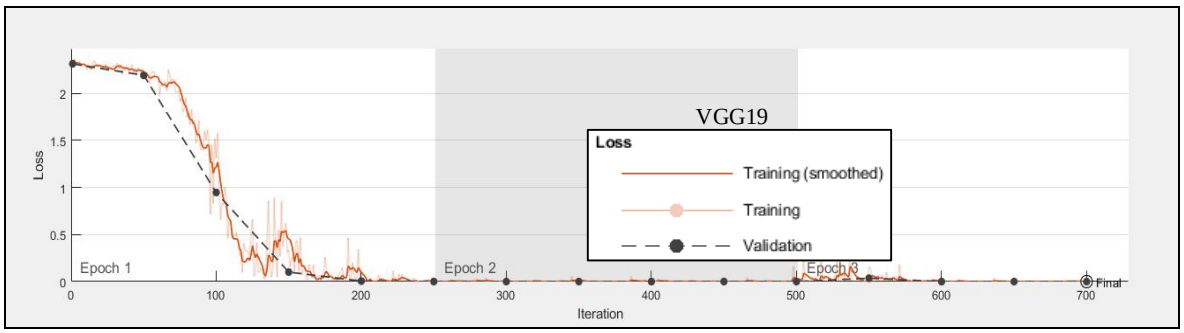
Figure 2: Training and validation results of (a) AlexNet (b) VGG16 and (c) VGG19.



(a)



(b)



(c)

Figure 3: Loss rate results of (a) AlexNet (b) VGG16 and (c) VGG19.

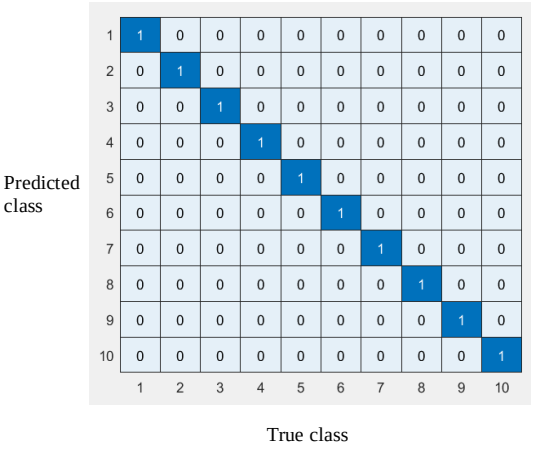


Figure 4: Confusion matrix of all transfer learning models.

5. CONCLUSION

This paper aims to investigate the performance of three CNN-based transfer learning models that are AlexNet, VGG16 and VGG19 in the domain of human recognition based on the ear image. Inspired by the fact that the transferred CNNs need less amount of data compared to the CNNs trained from scratch, we chose fine-tuned the transfer learning models to recognize 10 classes only using 2,500 ear images for training and 500 images for validation. We use Stochastic Gradient Descent with Momentum (SGDM) function to train the network. The proposed fine-tuned networks achieved 100% validation accuracy. However, the architecture of the tested models can be improved as real-time testing with occlusion and pose variations may cause recognition problems. Future work should be done by comparing the efficiency of the models to appreciate the worthiness of choosing the appropriate transfer model. It is

interesting to note that even though AlexNet has a much simpler configuration compared to VGGNet, it has achieved equal accuracy rate. Further tests need to be conducted to measure the robustness of each transfer learning models.

ACKNOWLEDGEMENT

Due acknowledgement is accorded to the Ministry of Education and Research Institute of Research, Management and Innovation, Universiti Teknologi MARA for the funding received through the Niche Research Grant Scheme, 600-RMI/NRGS 5/3 (10/2013).

REFERENCES

[1] Ghazi MM, Yanikoglu B & Aptoula E. (2017). Plant identification using deep neural networks via optimization of transfer learning parameters. *Neurocomputing, 235*, 228-235.

[2] Tay NNW, Botzheim J & Kubota N. (2016). Joint probabilistic approach for real-time face recognition with transfer learning. *Robotics and Autonomous Systems, 75*, 409-421. https://doi.org/10.1016/j.robot.2015.09.002

[3] Boufenar C, Kerboua A & Batouche M. (2018). Investigation on deep learning for off-line handwritten Arabic character recognition. *Cognitive Systems Research, 50*, 180-195.

[4] Lee SH, Chan CS, Mayo SJ & Remagnino P. (2017). How deep learning extracts and learns leaf features for plant classification. *Pattern Recognition, 71*, 1-13.

[5] Bloom V, Argyriou V, & Makris D. (2016). Hierarchical transfer learning for online recognition of compound actions. *Computer Vision and Image Understanding, 144*, 62-72.

[6] Emeršič Ž, Štruc V & Peer P. (2017). Ear recognition: More than a survey. *Neurocomputing, 255*, 26-39.

[7] Emeršič Ž, Štepec D, Štruc V & Peer P. (2017). Training convolutional neural networks with limited training data for ear recognition in the wild. Available online https://arxiv.org/abs/1711.09952. https://doi.org/10.1109/FG.2017.123

[8] Quattoni A, Collins M & Darrell T. Transfer learning for image classification with sparse prototype representations. *Proceedings of the IEEE Conference on IEEE Computer Vision and Pattern Recognition CVPR 2008,* (June 2008), pp:1-8.

[9] Hu, D., & Yang, Q. Transfer learning for activity recognition via sensor mapping. *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain* (2011), pp:1962-1967.

[10] Shell J, Vickers S, Coupland S & Istance H. Towards dynamic accessibility through soft gaze gesture recognition. *Proceeding of 12th IEEE UK Workshop on Computational Intelligence* (September 2012), pp:1-8.

[11] Sharma M, Holmes MP, Santamaría JC, Irani A, Isbell Jr CL & Ram A. Transfer learning in real-time strategy games using hybrid CBR/RL. *Proceeding of IJCAI,* (January 2007), Vol. 7, pp: 1041-1046.

[12] Al-Halah Z, Rybok L & Stiefelhagen R. (2016). Transfer metric learning for action similarity using high-level semantics. *Pattern Recognition Letters, 72*, 82-90.

[13] Mihalkova L, Huynh T & Mooney RJ, Mapping and revising Markov logic networks for transfer learning, *Proceedings of AAAI* (July 2007), Vol. 7, pp: 608-614.

[14] Iannarelli AV, *Ear identification*, Paramont Publishing Company, A.V. Iannarelli, Ear Identification, (1964)

[15] Ariffin SMZ, Jamil N & Rahman PNMA, (2017). Thermal and visible image fusion for ear recognition, *Journal of Telecommunication, Electronic and Computer Engineering (JTEC), 9*(2-3), 49-53.

[16] Zhang Y & Mu Z, (2017). Ear detection under uncontrolled conditions with multiple scale faster region-based convolutional neural networks, *Symmetry, 9*(4), 53.

[17] Galdámez PL, Raveane W & Arrieta AG. (2017). A brief review of the ear recognition process using deep neural networks, *Journal of Applied Logic*, *24*, 62-70. https://doi.org/10.1016/j.jal.2016.11.014

[18] Krizhevsky A, Sutskever I & Hinton GE, (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 1097-1105.

[19] Simonyan K & Zisserman A. Very deep convolutional networks for large-scale image recognition, (2014) pp. 1-14, available pnline:*arXiv preprint arXiv:1409.1556*.

[20] Shell J & Coupland S. (2015). Fuzzy transfer learning: methodology and application. *Information Sciences, 293*, 59-79.

[21] Hoo-Chang S, Roth HR, Gao M, Lu L, Xu Z, Nogues I., ... & Summers RM. (2016). Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging*, *35*(5), 1285.

[22] Chatfield K, Simonyan K, Vedaldi A & Zisserman, A. Return of the Devil in the Details: Delving Deep into Convolutional Nets. (2014) pp. 1-11, available online: *arXiv preprint arXiv:1405.3531*. https://doi.org/10.5244/C.28.6

[23] Azarov E, Vashkevich M, Likhachov D & Petrovsky AA. Real-time voice conversion using artificial neural networks with rectified linear units. *Proceedings of INTERSPEECH* (August 2013*)*, pp: 1032-1036.

[24] Glorot X, Bordes A & Bengio Y. Deep sparse rectifier neural networks. *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics* (June 2011), pp: 315-323.