

Human identification and classification network based on object hierarchy

Loai Alamro¹, Yuhanis Yusof², Nooraini Yusoff^{3,4}

¹Department of Cybersecurity Engineering, College of Engineering, Al-Naji University, Baghdad, Iraq

²School of Computing, College of Arts and Sciences, Universiti Utara Malaysia, Changlun, Malaysia

³Faculty of Data Science and Computing, Universiti Malaysia Kelantan, UMK City Campus, Kota Bharu, Malaysia

⁴Institute for Artificial Intelligence and Big Data, Universiti Malaysia Kelantan, UMK City Campus, Kota Bharu, Malaysia

Article Info

Article history:

Received Oct 24, 2023

Revised Jul 12, 2026

Accepted Aug 12, 2026

Keywords:

Attribute classification

Deep learning

Object hierarchy

Object identification

Object recognition

ABSTRACT

In the real world, objects are categorized hierarchically. However, deep learning (DL)-based computer vision (CV) approaches typically learn to classify objects in a flat manner, which limits their ability to generalize to new objects and scenarios. To address this limitation, a new network called human identification and classification network (HICNet) is proposed for hierarchical-based object identification and classification. HICNet is a hierarchical object identification and classification network with feature extraction layers, an identification branch, and a classification branch. HICNet was evaluated on the labeled faces in the wild (LFW) and labeled faces in the wild-A (LFWA) datasets, where it achieved an identification accuracy of 99.27% and an average classification accuracy of 98.61%. These results demonstrate that HICNet is able to classify objects beyond the training sample(s) by tracing the hierarchy and correlation of the object. HICNet can be used in a variety of CV applications, such as visually impaired aids and robot guidance as it provides a comprehensive insight into the objects.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Loai Alamro

Department of Cybersecurity Engineering, College of Engineering, Al-Naji University

Baghdad, Iraq

Email: loai.alamro@alnaji-uni.edu.iq

1. INTRODUCTION

Recently, motivated by the success of deep learning (DL), computer vision (CV) approaches have enabled computers to understand images or videos at a high level. Therefore, DL architectures, especially convolutional neural network (CNN), have been widely used for object recognition [1]–[10], object identification [11]–[29] and human head attribute classification [30]–[42]. However, the success rate of these approaches is limited when they are used in real-world applications, such as visually impaired people and robot guidance. The main problem with these approaches is that they are limited to training data and cannot identify or recognize objects outside the scope of this data. This problem results from the limitations of DL in global generalization, hierarchical learning, and correlation learning [43]–[46]. Therefore, these approaches learn to identify or recognize objects in a shallow way (i.e. they learn the categories of the same level) without being able to follow the hierarchy of objects.

In the real world, an object can be assigned to different categories (labels) that represent the object's family tree. These categories are organized hierarchically and known as the object hierarchy, which is used to recognize the type of object or identify the identity of the object [47]. The category at the top of the hierarchy represents the most important generic object type [47], [48]. The lower categories represent a specific

subtype of object. The categories at the bottom of the object hierarchy represent the identity of the object, i.e. the most characteristic categories of an object. Figure 1 shows an example of the human object hierarchy. As can be seen in Figure 1, the human category is at the top of the hierarchy, which is divided into two main types: male and female. Male and female are divided into several sub-categories, namely: man, woman, boy, and girl. In addition, the subcategories at the bottom of the hierarchy represent the identities used to differentiate between objects of the same kind, for example, names such as (Ahmed and Hind) can be used to differentiate between people.

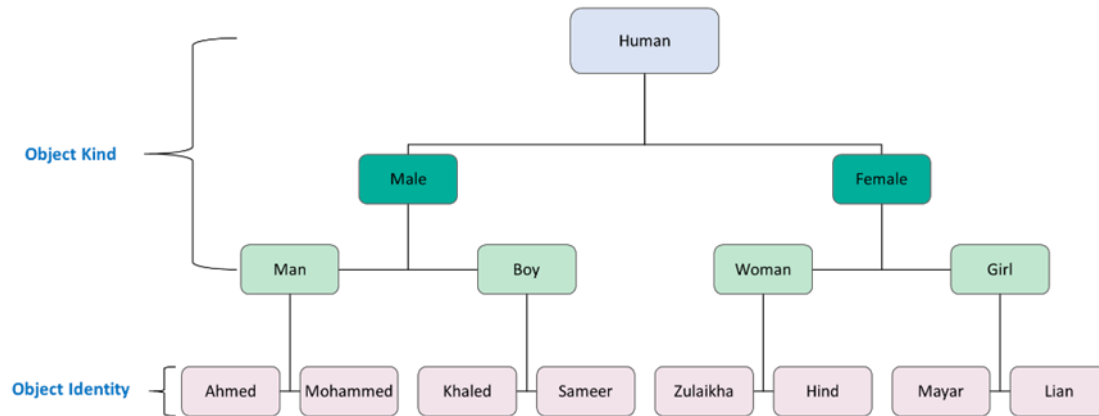


Figure 1. An example of an object hierarchy

Humans are dependent on the object hierarchy to adapt and identify or recognize the object without prior knowledge. In doing so, they first try to determine the identity of the object, and if this fails, they try to recognize its type or subtype, or even do this all at once. In contrast, DL-based CV approaches cannot adapt like humans because they cannot transfer between the categories of the object hierarchy. This problem, as mentioned earlier, is due to the limitations of DL in global generalization, hierarchical learning, and correlation learning. Therefore, these approaches fail in identifying or recognizing objects outside the scope of the training data. Real-world applications must mimic human performance to increase their success rate. Accordingly, there is an urgent need for an approach that can identify an object, recognize its type, or at least classify its parts by tracing the hierarchy of the object. To achieve this, another challenge must first be overcome. This challenge is that the visual features are different for different subsets of objects. Therefore, the identity of an object, its type and its parts should have different features [49]–[51] in order to be distinguished. However, DL-based CV approaches only rely on extracting high-level features that are not semantic enough to recognize parts or subsets of object parts [50], [52], [53]. Therefore, in order to identify an object, recognize its type and classify its parts, appropriate features must be extracted for each task.

Consequently, this study aims to overcome these challenges and proposes to integrate the tasks of object identification and object part classification into a single network for hierarchical-based object identification and classification. The proposed network is called the human identification and classification network (HICNet). The HICNet consists of i) feature extraction layers to extract the appropriate feature level for each task, and ii) identification branch and classification branch to identify human identity and classify its head parts. The rest of the article is organized as follows. Section 2 gives a brief background on visual geometry group-face (VGG-Face) [16] and multi-output convolutional neural network (MOCNN) [42]. The proposed network with the parameters used in building the layers of the network is described in detail in section 3. The training setup of the proposed network in section 4. Section 5 presents the experimental results of the proposed network. Finally, section 6 contains the conclusion.

2. BACKGROUND ON VISUAL GEOMETRY GROUP-FACE AND MOCNN

The main objective of this study is to develop a DL network that integrates the tasks of object identification and human head attribute classification to identify and classify humans based on the object hierarchy. To achieve this goal, this study proposes the integration of the VGG-Face net [16] with the MOCNN. Therefore, two configurations, A and B, are considered for integration in this study, which are further explained in section 3. In addition, the details of the VGG-Face net and the MOCNN are presented in the next subsections.

2.1. Visual geometry group-face

VGG-Face is a deep CNN introduced by Parkhi *et al.* [16] in 2015 for face recognition. VGG-Face is based on the visual geometry group network 16 (VGGNet)-16 architecture [54] proposed by the visual geometry group (VGG) in 2014. VGG-Face consists of 26 layers with an input layer, 16 convolutional layers, five max-pooling layers, three fully linked layers and a linear softmax activation as the last layer. The input layer is fed with colored face images of size 224×224 pixels. All convolutional layers are followed by a rectified linear unit (ReLU) activation, while each convolutional block is followed by a max-pooling layer, except for the last block, which is followed by three fully-linked layers. The fully linked layers are listed as convolution layers because they are a special case of convolution. These layers employ filters of a size is equal to the size of the input data to capture data from the entire image. In addition, a dropout is applied at a rate of 0.5 after the first and second fully-connected layers. Finally, the last fully-connected layer is followed by a flattened layer before softmax activation. The exact components of the VGG-Face are shown in Figure 2 and the structure of the VGG-Face is detailed in Table 1, including the parameters used in building the layers.

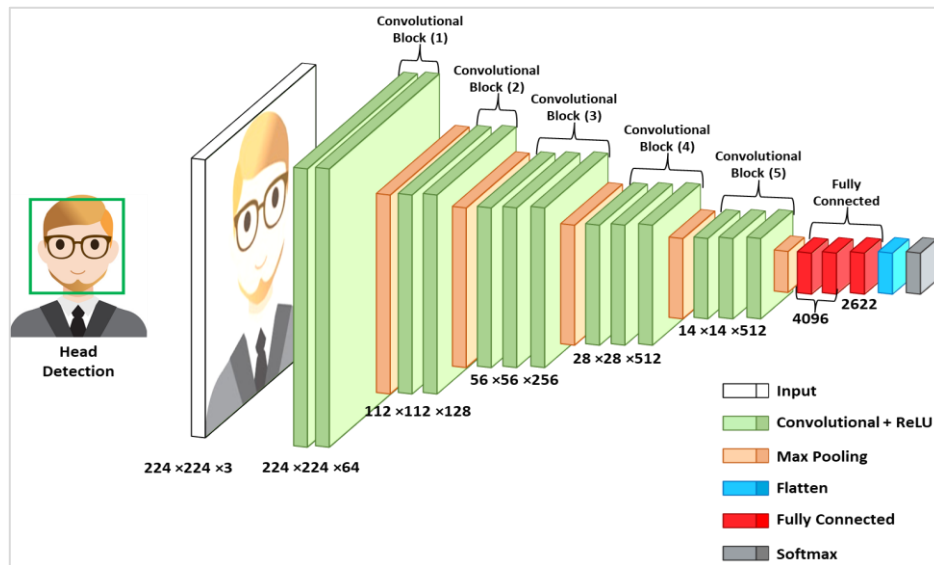


Figure 2. The components of VGG-Face

Table 1. Architecture details of the VGG-Face net

Layer type	Input size	Parameters	Filter size	Pooling size	Output size
Input	$224 \times 224 \times 3$	-	-	-	$224 \times 224 \times 3$
Convolutional (1)	$224 \times 224 \times 3$	1,792	3×3	-	$224 \times 224 \times 64$
Convolutional (2)	$224 \times 224 \times 64$	36,928	3×3	-	$224 \times 224 \times 64$
Max Pooling (1)	$224 \times 224 \times 64$	-	-	2×2	$112 \times 112 \times 64$
Convolutional (3)	$112 \times 112 \times 64$	73,856	3×3	-	$112 \times 112 \times 128$
Convolutional (4)	$112 \times 112 \times 128$	147,584	3×3	-	$112 \times 112 \times 128$
Max Pooling (2)	$112 \times 112 \times 128$	-	-	2×2	$56 \times 56 \times 128$
Convolutional (5)	$56 \times 56 \times 128$	295,168	3×3	-	$56 \times 56 \times 256$
Convolutional (6)	$56 \times 56 \times 256$	590,080	3×3	-	$56 \times 56 \times 256$
Convolutional (7)	$56 \times 56 \times 256$	590,080	3×3	-	$56 \times 56 \times 256$
Max Pooling (3)	$56 \times 56 \times 256$	-	-	2×2	$28 \times 28 \times 256$
Convolutional (8)	$28 \times 28 \times 256$	1,180,160	3×3	-	$28 \times 28 \times 512$
Convolutional (9)	$28 \times 28 \times 512$	2,359,808	3×3	-	$28 \times 28 \times 512$
Convolutional (10)	$28 \times 28 \times 512$	2,359,808	3×3	-	$28 \times 28 \times 512$
Max Pooling (4)	$28 \times 28 \times 512$	-	-	2×2	$14 \times 14 \times 512$
Convolutional (11)	$14 \times 14 \times 512$	2,359,808	3×3	-	$14 \times 14 \times 512$
Convolutional (12)	$14 \times 14 \times 512$	2,359,808	3×3	-	$14 \times 14 \times 512$
Convolutional (13)	$14 \times 14 \times 512$	2,359,808	3×3	-	$14 \times 14 \times 512$
Max Pooling (5)	$14 \times 14 \times 512$	-	-	2×2	$7 \times 7 \times 512$
Convolutional (fully connected 1)	$7 \times 7 \times 512$	102,764,544	-	-	$1 \times 1 \times 4,096$
Dropout	$1 \times 1 \times 4,096$	-	-	-	$1 \times 1 \times 4,096$
Convolutional (fully connected 2)	$1 \times 1 \times 4,096$	16,781,312	-	-	$1 \times 1 \times 4,096$
Dropout	$1 \times 1 \times 4,096$	-	-	-	$1 \times 1 \times 4,096$
Convolutional (fully connected 3)	$1 \times 1 \times 4,096$	10,742,334	-	-	$1 \times 1 \times 2,622$
Flatten	$1 \times 1 \times 2,622$	-	-	-	2,622
Softmax	2,622	15,079,627	-	-	2,622

2.2. Multi-output convolutional neural network

The MOCNN is a multiple-output convolutional network introduced by Alamro *et al.* [42] in 2022 for human head attribute classification. The MOCNN consists of sixteen layers with one input layer, two convolutional layers, two pooling layers, one flattened layer, five fully connected layers and five output layers. Similar to the proposed low-level feature extraction network, all convolutional layers of the proposed MOCNN contain ReLUs units for linear transformation and nonlinear mapping. The input layer of the proposed network was fed with images of size $(178 \times 218 \times 3)$ or $(256 \times 256 \times 3)$ pixels. In addition, a 5×5 convolutional filter was used in the first convolutional layer to implement the convolutional operations and extract the features. Next, a 2×2 top pooling layer was used to reduce the dimensionality of the outputs of the convolutional layer. The outputs of the maxpooling layer are fed into the second convolutional layer to perform its convolutional operations. In the second convolutional layer, a 3×3 convolutional filter was used to perform the convolutional operations and create the feature map. A maxpooling layer of size 2×2 was used after the second convolutional layer to reduce the dimensionality of the generated feature map. Furthermore, a flattening layer was used to reduce the outputs of the maxpooling layer into a 1-dimensional vector. Then, a multi-output classifier is connected to the flatten layer to classify the parts of the human head. The classifier consists of five fully connected layers, each connected to an output layer with a sigmoid classifier to predict the probability of the categories and implement the human head attributes classification. The components of the MOCNN are shown in Figure 3 and the exact structure of the MOCNN is detailed in Table 2, including the parameters used to build the layers.

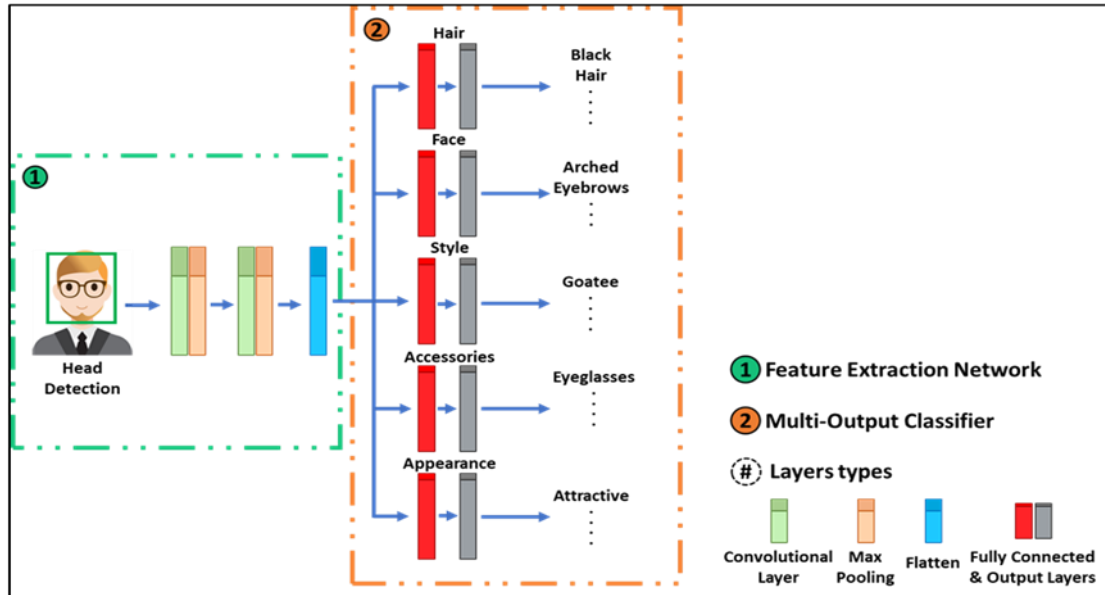


Figure 3. The components of the MOCNN

Table 2. Details of MOCNN structure

Layer type	Input size	Parameters	Filter size	Pooling size	Output size
Input	$218 \times 178 \times 3$	-	-	-	$218 \times 178 \times 3$
Convolutional 1	$218 \times 178 \times 3$	896	3×3	-	$218 \times 178 \times 32$
Max pooling 1	$218 \times 178 \times 32$	-	-	2×2	$109 \times 89 \times 32$
Convolutional 2	$109 \times 89 \times 32$	-	-	-	$109 \times 89 \times 64$
Max pooling 2	$109 \times 89 \times 64$	-	-	-	$54 \times 44 \times 64$
Flatten	$54 \times 44 \times 64$	-	-	-	152,064
Fully connected 1	152,064	38,928,640	-	-	256
Fully connected 2	152,064	38,928,640	-	-	256
Fully connected 3	152,064	38,928,640	-	-	256
Fully connected 4	152,064	19,464,320	-	-	128
Fully connected 5	152,064	19,464,320	-	-	128
Output 1	256	2,313	-	-	9
Output 2	256	2,827	-	-	11
Output 3	256	2,313	-	-	9
Output 4	128	645	-	-	5
Output 5	128	774	-	-	6

3. PROPOSED NETWORK

The structure of the VGG-Face consists of sixteen layers: an input layer, thirteen convolutional layers, three fully connected layers and an output layer, as described in subsection 2.1. The MOCNN network, on the other hand, consists of one input layer, two convolutional layers, five fully connected layers and five output layers, as described in subsection 2.2. In this context, it should be noted that both networks have a number of identical components. In practice, the input layer is identical in both networks because the size of the input data is the same in both networks. Furthermore, the first and second convolutional layers of the VGG-Face are identical to the second convolutional layer of the MOCNN, as filter of the same number and size (i.e. $64 \times 3 \times 3$) are used in these layers. Therefore, this study considers two configurations, A and B, based on two different ways of integrating the VGG-Face with the MOCNN.

The integration in configuration A is based on the sharing of the input layer by both networks. The integration in configuration B, on the other hand, not only shares the input layer, but also the first convolutional layer of the VGG-Face with the second convolutional layer of the MOCNN. Based on this strategy, configuration A consists of two parallel branches, one responsible for identifying humans and the other for classifying head attributes. The structure of the identification branch adopts the structure of the VGG-Face layers, while the structure of the classification branch adopts the structure of the MOCNN layers. The integration process is carried out without any change to the standard layers of both networks, except for the unification of the input layer. The detailed structure of configuration A is shown in Figure 4.

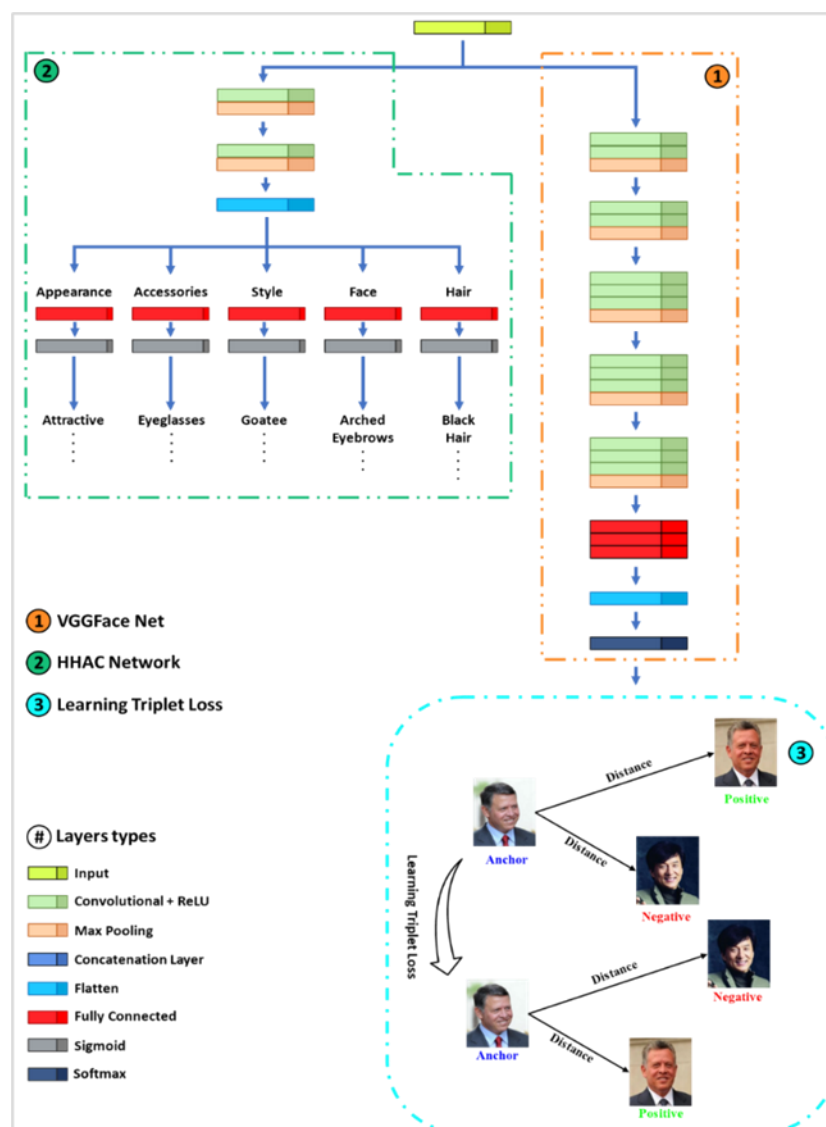


Figure 4. Structure of the network (A) proposed to integrate the object identification network and the multi-output classification network

Configuration B, on the other hand, consists of the same branches as configuration A, but with a different integration process. The integration process of configuration B consisted of two steps: i) unification of the input layer and ii) integration of the first convolutional layer of the VGG-Face with the second convolutional layer of the MOCNN network by adding a concatenation layer. Then, the concatenation layer is connected to the second convolutional layer of the VGG-Face and the flattened layer of the MOCNN to form the identification and classification branches. The integration of the convolutional layers of both networks allows multiple levels of features to be extracted and each branch of the network to be supplied with the appropriate level of features to perform the assigned task. The detailed structure of the configuration B is shown in Figure 5.

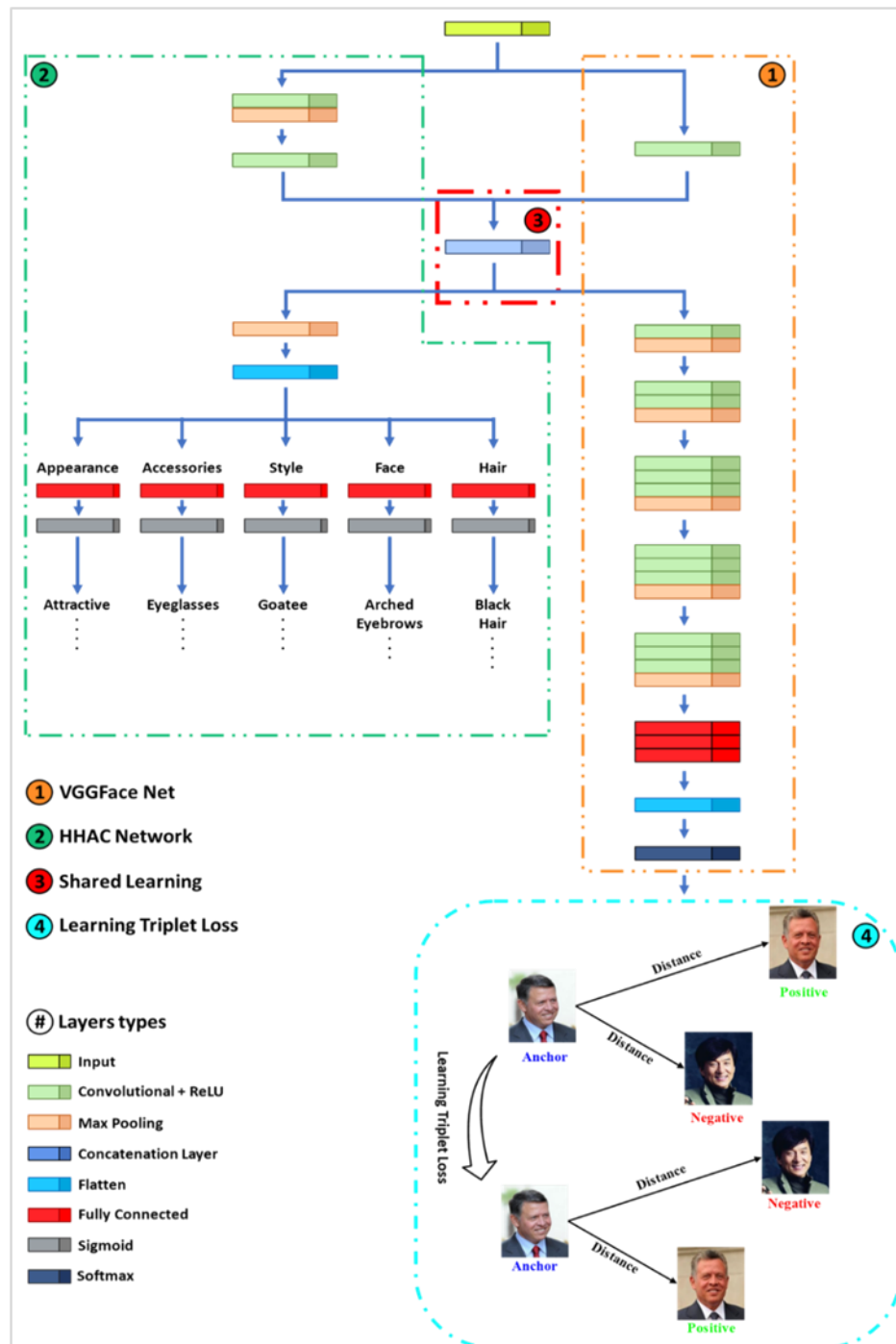


Figure 5. Structure of the network (B) proposed for the integration of the object identification network and the multi-output classification network

4. TRAINING SETUPS FOR INTEGRATION OF VGG-FACE AND MOCNN

This section describes the training procedure for the two proposed networks (A and B). The proposed networks were implemented based on the open-source platform TensorFlow 2.0, using Intel Core i7-7700 CPU, 64 Gigabyte RAM and NVIDIA GEFORCE RTX 3090 GPU to accelerate the training process. The training procedure follows the training details of VGG-Face [18] and MOCNN [45]. In general, networks A and B are trained for 74 epochs with a batch size of 64. The size of the input images is 224×224 pixels, randomly cropped from the images in the dataset. The identification branch is regularized by a weight drop of 5×10^{-4} and a dropout rate of 0.5 is applied to the first and second fully connected layers. In addition, the learning rate of the network was initially set to 10^{-2} and decreased three times by a factor of 10 when the validation accuracy stopped increasing. On the other hand, the learning rate of the classification branch was first set at 10^{-4} and then lowered to 10^{-6} . Furthermore, learning in the classification branch is stopped after ten epochs to avoid overfitting. In both networks (A and B), triple-loss learning is implemented by freezing all layers in the branches of each network, except for the last fully connected layer in the identification branch. This layer is then learned for 10 epochs using Stochastic gradient descent (SGD) with a fixed learning rate of 0.25.

5. EXPERIMENT RESULTS

This section presents the results of the proposed A and B networks and discusses their implications, interpretations and significance. The aim of this study was to overcome the limitations of current DL-based CV approaches [1]–[29] in tracing the object hierarchy, which affect their performance in real-world applications. In this study the focuses on tracing the hierarchy of the human object. Therefore, the experiment was conducted on the labeled faces in the wild (LFW) [55] and labeled faces in the wild-A (LFWA) [55] datasets. LFW is a large-scale dataset introduced in 2007 for face verification. LFW is composed of 13,233 color images of 5,749 persons, where 1,680 identities have two or more images. LFWA is an updated version of LFW dataset, which was presented in 2015 for human head attributes classification. The LFWA inherits the images of the LFW dataset, then it provides each image with 73 annotations of different parts of the human head. The proposed networks are evaluated by comparing the accuracy of the identification branch of both networks and comparing the average classification accuracy of the classification branch of the networks. In addition, the classification accuracy of the proposed networks is compared for each human head attribute.

Through rigorous analyses and investigations, this study found that the proposed network B outperforms the proposed network A. The proposed network A achieved an identification accuracy of 98.95%, which is equivalent to the accuracy of the original VGG-Face net, as shown in Figure 6. Similarly, network A achieved an average classification accuracy of 97.93%, which is also equivalent to the average classification accuracy of the original MOCNN, as shown in Figure 7. In contrast, the proposed network B outperformed the proposed network A in both identification accuracy and average classification accuracy. Network B achieved an identification accuracy of 99.27%, which is an improvement of 0.32 over network A and the original network. In addition, network B achieved an average classification accuracy of 98.61%, an improvement of 0.68% over Network A.

The results presented in this paper emphasize the efficiency of the proposed network B, which showed superior performance compared to network A and the established baseline models. Network B showed remarkable improvements in identification accuracy and average classification accuracy, outperforming both network A and the original benchmark networks. Furthermore, the results show that Network B consistently outperformed network A in all evaluated groups of human head attributes, as shown in Figure 8 (see Appendix). Interestingly, although the two branches of network A use the same input layer, there is no performance improvement over the VGG-Face and MOCNN networks before integration. This is mainly due to the fact that the two network branches are not involved in any activity (e.g. feature extraction) along the network pipeline, but each works on its own. In contrast, network B achieved a significant improvement compared to the performance of the original networks (i.e. VGG-Face and MOCNN) and network A. This improvement is due to the integration of the first convolutional layer of the identification branch with the second convolutional layer of the classification branch, which enables joint feature learning.

The experimental results show the ability of the proposed network B to trace the hierarchy of human objects, allowing it to identify, recognize and classify humans even when they are outside the scope of the training data. For example, even if the proposed network cannot identify the identity of a person, it is able to recognize him as human and determine their gender, i.e. male or female. Moreover, it can characterize him by classifying the parts of his head, such as the hair color, the shape of the face and nose. Accordingly, the

performance of the proposed network outperforms existing DL-based CV approaches such as object recognition [1]–[10] and object identification [11]–[29] in real-world applications.

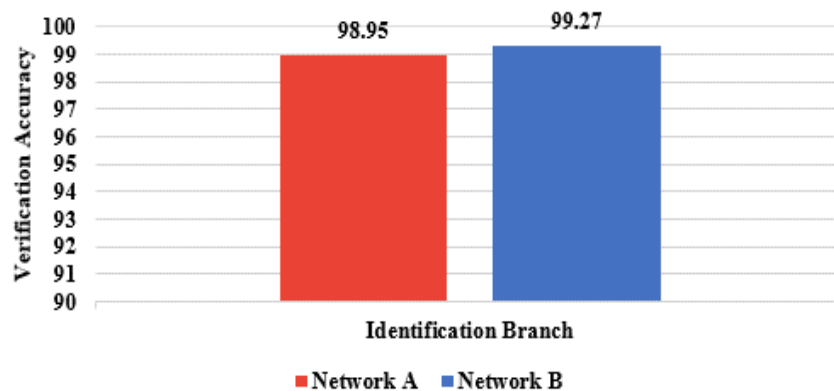


Figure 6. Comparison of the verification accuracy (in %) between the identification branches of network A and network B on LFW dataset

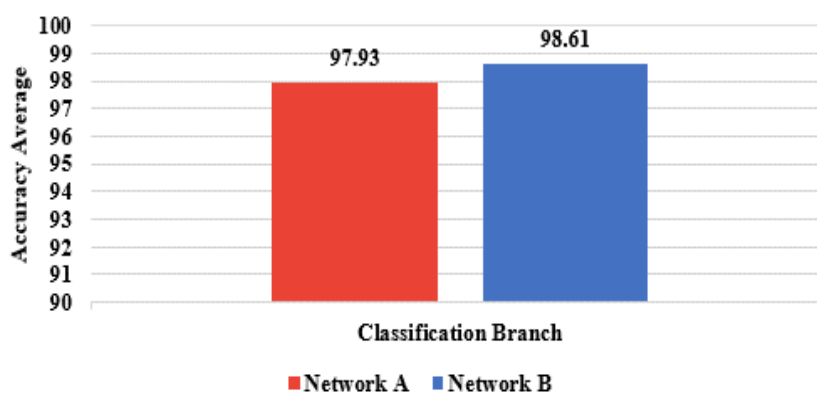


Figure 7. The average accuracy (in %) comparing between the classification branches of network A and network B on the LFWA dataset

6. CONCLUSION

This study presents the integration process of a human identification network with a human head attribute classification network to form a unified network for hierarchical-based object identification and classification. The study uses the VGG-Face and the MOCNN to develop the unified network. In this study, two configurations, A and B, have been proposed based on different ways of integrating VGG-Face and MOCNN. The integration in network A is based on unifying the input layers of both networks into one main network. Integration into network B is achieved by unifying the input layer and integrating the first convolutional layer of the VGG-Face with the second convolutional layer of the MOCNN using a concatenation layer to form the main network. Experiments on the LFW and LFWA datasets for identification and classification tasks have shown that network B is a better network A and even has higher accuracy compared to the original networks (i.e. the standalone network). Furthermore, the proposed network B is able to trace the object hierarchy, which gives it an advantage over existing DL-based CV approaches in real-world applications. The proposed network is limited to identifying and describing only one part of a human object—namely the head. In the future, however, other human attributes must also be considered to obtain a complete and comprehensive picture of a human being.

ACKNOWLEDGEMENTS

This research was supported by the Ministry of Higher Education (MoHE) of Malaysia.

FUNDING INFORMATION

This research was supported by the Ministry of Higher Education (MoHE) of Malaysia through Fundamental Research Grant Scheme (FRGS/1/2022/ICT02/UUM/02/1).

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Loai Alamro	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			
Yuhanis Yusof						✓				✓		✓	✓	✓
Nooraini Yusoff						✓				✓		✓		

C : Conceptualization

M : Methodology

So : Software

Va : Validation

Fo : Formal analysis

I : Investigation

R : Resources

D : Data Curation

O : Writing - Original Draft

E : Writing - Review & Editing

Vi : Visualization

Su : Supervision

P : Project administration

Fu : Funding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The authors confirm that the data supporting the findings of this study are available within the article.

REFERENCES

- [1] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015, doi: 10.1109/TPAMI.2015.2389824.
- [2] R. Girshick, "Fast R-CNN," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec. 2015, pp. 1440–1448, doi: 10.1109/ICCV.2015.169.
- [3] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.
- [4] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: object detection via region-based fully convolutional networks," 2023, *arXiv: 1605.06409*.
- [5] Z. Li, C. Peng, G. Yu, X. Zhang, Y. Deng, and J. Sun, "Light-head R-CNN: in defense of two-stage object detector," 2017, *arXiv: 1711.07264*.
- [6] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 936–944, doi: 10.1109/CVPR.2017.106.
- [7] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 779–788, doi: 10.1109/CVPR.2016.91.
- [8] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 6517–6525, doi: 10.1109/CVPR.2017.690.
- [9] W. Liu et al., "SSD: single shot multibox detector," in *European conference on computer vision (ECCV)*, 2016, pp. 21–37, doi: 10.1007/978-3-319-46448-0_2.
- [10] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 2999–3007, doi: 10.1109/ICCV.2017.324.
- [11] 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*, 2014, pp. 1701–1708, doi: 10.1109/CVPR.2014.220.
- [12] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," *Proceedings of the 28th International Conference on Neural Information Processing Systems – Volume 2*, 2014, pp. 1988–1996.
- [13] Y. Sun, D. Liang, X. Wang, and X. Tang, "DeepID3: face recognition with very deep neural networks," 2015, *arXiv: 1502.00873*.
- [14] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: a unified embedding for face recognition and clustering," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 815–823, doi: 10.1109/CVPR.2015.7298682.
- [15] J. Liu, Y. Deng, T. Bai, Z. Wei, and C. Huang, "Targeting ultimate accuracy: face recognition via deep embedding," 2015, *arXiv: 1506.07310*.
- [16] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proceedings of the British Machine Vision Conference 2015*, 2015, pp. 41.1–41.12, doi: 10.5244/C.29.41.
- [17] W. Liu, Y. Wen, Z. Yu, and M. Yang, "Large-margin softmax loss for convolutional neural networks," *ICML'16: Proceedings of the 33rd International Conference on International Conference on Machine Learning*, vol. 48, 2017, pp. 507–516.

- [18] X. Zhang, Z. Fang, Y. Wen, Z. Li, and Y. Qiao, "Range loss for deep face recognition with long-tailed training data," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 5419–5428, doi: 10.1109/ICCV.2017.578.
- [19] R. Ranjan, C. D. Castillo, and R. Chellappa, "L2-constrained softmax loss for discriminative face verification," 2017, *arXiv: 1703.09507*.
- [20] F. Wang, X. Xiang, J. Cheng, and A. L. Yuille, "NormFace," in *Proceedings of the 25th ACM international conference on Multimedia*, Oct. 2017, pp. 1041–1049, doi: 10.1145/3123266.3123359.
- [21] M. A. Hasnat, J. Bohné, J. Milgram, S. Gentric, and L. Chen, "Von mises-fisher mixture model-based deep learning: application to face verification," 2017, *arXiv: 1706.04264*.
- [22] J. Deng, Y. Zhou, and S. Zafeiriou, "Marginal loss for deep face recognition," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jul. 2017, pp. 2006–2014, doi: 10.1109/CVPRW.2017.251.
- [23] 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)*, Jul. 2017, pp. 6738–6746, doi: 10.1109/CVPR.2017.713.
- [24] X. Qi and L. Zhang, "Face recognition via centralized coordinate learning," 2018, *arXiv: 1801.05678*.
- [25] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, Jul. 2018, doi: 10.1109/LSP.2018.2822810.
- [26] H. Wang *et al.*, "CosFace: large margin cosine loss for deep face recognition," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 5265–5274, doi: 10.1109/CVPR.2018.00552.
- [27] Y. Zheng, D. K. Pal, and M. Savvides, "Ring loss: convex feature normalization for face recognition," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 5089–5097, doi: 10.1109/CVPR.2018.00534.
- [28] X. Wu, R. He, Z. Sun, and T. Tan, "A light CNN for deep face representation with noisy labels," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 11, pp. 2884–2896, Nov. 2018, doi: 10.1109/TIFS.2018.2833032.
- [29] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: additive angular margin loss for deep face recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4685–4694, doi: 10.1109/CVPR.2019.00482.
- [30] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev, "PANDA: pose aligned networks for deep attribute modeling," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2014, pp. 1637–1644, doi: 10.1109/CVPR.2014.212.
- [31] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec. 2015, pp. 3730–3738, doi: 10.1109/ICCV.2015.425.
- [32] E. M. Rudd, M. Günther, and T. E. Boulton, "MOON: a mixed objective optimization network for the recognition of facial attributes," in *European Conference on Computer Vision (ECCV)*, 2016, pp. 19–35, doi: 10.1007/978-3-319-46454-1_2.
- [33] K. He, Z. Wang, Y. Fu, R. Feng, Y.-G. Jiang, and X. Xue, "Adaptively weighted multi-task deep network for person attribute classification," in *Proceedings of the 25th ACM international conference on Multimedia*, Oct. 2017, pp. 1636–1644, doi: 10.1145/3123266.3123424.
- [34] H. Guo, X. Fan, and S. Wang, "Human attribute recognition by refining attention heat map," *Pattern Recognition Letters*, vol. 94, pp. 38–45, Jul. 2017, doi: 10.1016/j.patrec.2017.05.012.
- [35] U. Mahbub, S. Sarkar, and R. Chellappa, "Segment-based methods for facial attribute detection from partial faces," *IEEE Transactions on Affective Computing*, vol. 11, no. 4, pp. 601–613, Oct. 2020, doi: 10.1109/TAFFC.2018.2820048.
- [36] M. Xu *et al.*, "Bio-inspired deep attribute learning towards facial aesthetic prediction," *IEEE Transactions on Affective Computing*, vol. 12, no. 1, pp. 227–238, Jan. 2021, doi: 10.1109/TAFFC.2018.2868651.
- [37] N. Zhuang, Y. Yan, S. Chen, and H. Wang, "Multi-task learning of cascaded CNN for facial attribute classification," in *2018 24th International Conference on Pattern Recognition (ICPR)*, Aug. 2018, pp. 2069–2074, doi: 10.1109/ICPR.2018.8545271.
- [38] S. Huang, X. Li, Z.-Q. Cheng, Z. Zhang, and A. Hauptmann, "GNAS: a greedy neural architecture search method for multi-attribute learning," in *Proceedings of the 26th ACM international conference on Multimedia*, Oct. 2018, pp. 2049–2057, doi: 10.1145/3240508.3240588.
- [39] E. Hand and R. Chellappa, "Attributes for improved attributes: a multi-task network utilizing implicit and explicit relationships for facial attribute classification," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, Feb. 2017, doi: 10.1609/aaai.v31i1.11229.
- [40] J. Cao, Y. Li, and Z. Zhang, "Partially shared multi-task convolutional neural network with local constraint for face attribute learning," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 4290–4299, doi: 10.1109/CVPR.2018.00451.
- [41] L. Mao, Y. Yan, J.-H. Xue, and H. Wang, "Deep multi-task multi-label CNN for effective facial attribute classification," *IEEE Transactions on Affective Computing*, vol. 13, no. 2, pp. 818–828, Apr. 2022, doi: 10.1109/TAFFC.2020.2969189.
- [42] L. Alamro, Y. Yusof, and N. Yusoff, "Multi-output convolutional neural network for automatic human head attributes classification," *International Journal of Intelligent Engineering and Systems*, vol. 15, no. 5, pp. 219–231, Oct. 2022, doi: 10.22266/ijies2022.1031.20.
- [43] G. Marcus, "Deep learning: a critical appraisal," 2018, *arXiv: 1801.00631*.
- [44] X. Zhu and M. Bain, "B-CNN: branch convolutional neural network for hierarchical classification," 2017, *arXiv: 1709.09890*.
- [45] Y. Guo, Y. Liu, E. M. Bakker, Y. Guo, and M. S. Lew, "CNN-RNN: a large-scale hierarchical image classification framework," *Multimedia Tools and Applications*, vol. 77, no. 8, pp. 10251–10271, Apr. 2018, doi: 10.1007/s11042-017-5443-x.
- [46] D. Roy, P. Panda, and K. Roy, "Tree-CNN: a hierarchical deep convolutional neural network for incremental learning," *Neural Networks*, vol. 121, pp. 148–160, Jan. 2020, doi: 10.1016/j.neunet.2019.09.010.
- [47] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, Dec. 2015, doi: 10.1007/s11263-015-0816-y.
- [48] A. Zweig and D. Weinshall, "Exploiting object hierarchy: combining models from different category levels," in *2007 IEEE 11th International Conference on Computer Vision*, 2007, pp. 1–8, doi: 10.1109/ICCV.2007.4409064.
- [49] Z. Wang, X. Wang, and G. Wang, "Learning fine-grained features via a CNN tree for large-scale classification," *Neurocomputing*, vol. 275, pp. 1231–1240, Jan. 2018, doi: 10.1016/j.neucom.2017.09.061.
- [50] B. Banerjee and S. Chaudhuri, "Scene recognition from optical remote sensing images using mid-level deep feature mining," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 7, pp. 1080–1084, Jul. 2018, doi: 10.1109/LGRS.2018.2822779.
- [51] L. Wang, S. Guo, W. Huang, Y. Xiong, and Y. Qiao, "Knowledge guided disambiguation for large-scale scene classification with multi-resolution CNNs," *IEEE Transactions on Image Processing*, vol. 26, no. 4, pp. 2055–2068, Apr. 2017, doi: 10.1109/TIP.2017.2675339.
- [52] S. Guo, W. Huang, L. Wang, and Y. Qiao, "Locally supervised deep hybrid model for scene recognition," *IEEE Transactions on Image Processing*, vol. 26, no. 2, pp. 808–820, Feb. 2017, doi: 10.1109/TIP.2016.2629443.

- [53] H. Nakayama and T. Tsuda, "Efficient two-step middle-level part feature extraction for fine-grained visual categorization," *IEICE Transactions on Information and Systems*, no. 6, pp. 1626–1634, 2016, doi: 10.1587/transinf.2015EDP7358.
- [54] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015, *arXiv: 1409.1556*.
- [55] M. Knoche, S. Hormann, and G. Rigoll, "Cross-quality LFW: a database for analyzing cross-resolution image face recognition in unconstrained environments," in *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, Dec. 2021, pp. 1–5, doi: 10.1109/FG52635.2021.9666960.

APPENDIX

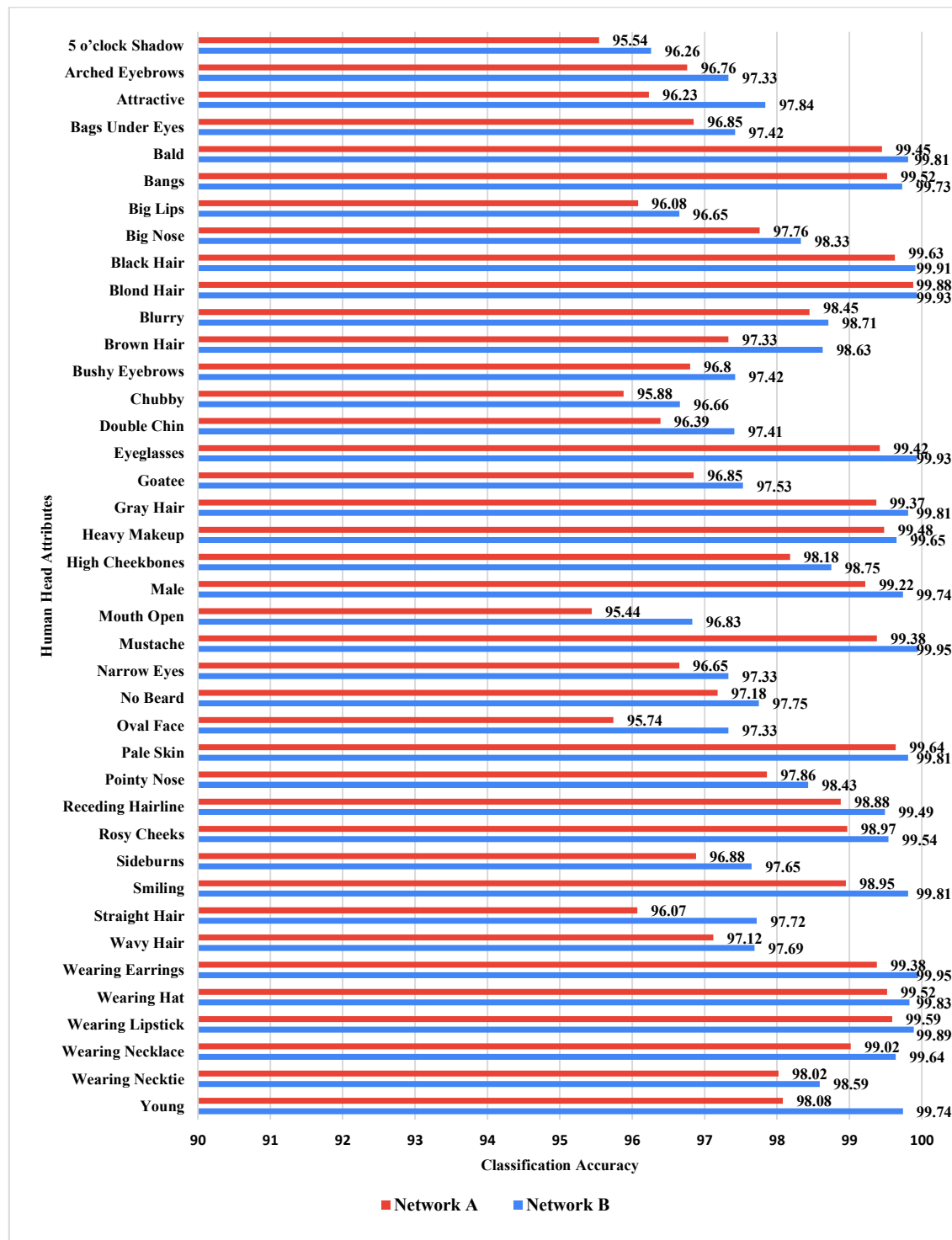
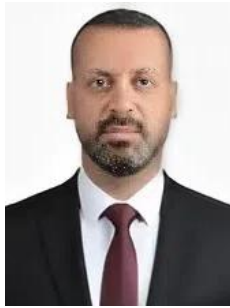


Figure 8. Comparison of classification accuracy (in %) between network A and network B for the 40 categories of the LWFA dataset

BIOGRAPHIES OF AUTHORS



Loai Alamro    is the head of the Department of Cybersecurity Engineering at College of Engineering, Al-Naji University in Baghdad, Iraq. He has also held several important positions at Al-Turath University in Baghdad, Iraq, most notably as a faculty member in the Department of Computer Science, coordinator of the Department of Cybersecurity Sciences, and subsequently head of the Department of Cybersecurity Sciences in the College of Science. His academic journey began with a bachelor's degree in Computer Science from Al-Turath University in 2007. He continued his studies at Universiti Utara Malaysia, earning a Master of Science in Information Technology in 2016, followed by a Ph.D. in artificial intelligence in 2024. His doctoral research focused on developing a DL framework for hierarchical-based object identification and description. His research interests encompass CV, image processing, multimedia forensics, pattern recognition, object recognition, DL, and machine learning. He has contributed to scholarly publications, including a study on camera efficiency and image resolution presented at the Conference of Open Innovation Association, FRUCT, in 2024. His academic and professional endeavors reflect a commitment to advancing CV and cybersecurity education and research. He can be contacted at email: loai.alamro@alnaji-uni.edu.iq.



Yuhanis Yusof    is an associate professor in the School of Computing, College of Arts and Sciences, Universiti Utara Malaysia, Malaysia. She obtained her Ph.D. in Computer Science from Cardiff University, UK. She also holds a M.Sc. degree in Computer Science from Universiti Sains Malaysia and Bachelor of Information Technology from Universiti Utara Malaysia. Her research interest is broadly in data analysis and management for large scale computing. This includes data mining (discovering patterns of interest from data), data warehousing, information retrieval and optimization. Currently, she is involved in several research projects involving data mining, machine learning and swarm computing (in particular the artificial bee colony, firefly algorithm, bat algorithm and grey wolf optimizer). Her work has been applied in various domains including business management, finance, transportation, education, networking and medical. The undertaken work is supported by various schemes including the Fundamental Research Grant, Exploratory Research Grant, Knowledge Transfer Programme, Take-off Innovation Programme and Universiti Grant. She can be contacted at email: yuhanis@uum.edu.my.



Nooraini Yusoff    is currently an associate professor at Faculty of Data Science and Computing, and a research fellow of Institute of Artificial Intelligence and Big Data at Universiti Malaysia Kelantan. She received her B.Sc. in Information Technology specializing in Artificial Intelligence, and M.Sc. in Intelligent System from Universiti Utara Malaysia. In 2012, she obtained her Ph.D. from University of Surrey, United Kingdom, in the field of Computational Neuroscience. She has more than 20 years, experience in artificial intelligence, data analytics, machine learning and system integration. She has involved in various AI and data analytics projects including flood disaster managements, Ministry of Higher Education graduate data analysis project, analysis on effectiveness and relevancy of ICT programs at Malaysia public universities, telecommunication fraud detection and agriculture data analytics projects. She has published a number of national and international journals and conference proceedings. She is also the reviewer of some reputable journals in the related fields. She can be contacted at email: nooraini.y@umk.edu.my.