UTM UNIVERSITI TEKNOLOGI MALAYSIA
INTERNATIONAL JOURNAL OF INNOVATIVE COMPUTING
ISSN 2180-4370
Journal Homepage : https://ijic.utm.my/

Real-Time Hand Gesture Recognition Using YOLO and (Darknet-19) Convolution Neural Networks

Raad Ahmed Mohamed
Computer Science, Iraqi Commission for Computer and informatics
Informatics Institute for Postgraduate Studies
Baghdad, Iraq

Karim Q Hussein*
Mustansiryha University-Faculty of Science
Computer Science Dept.
Baghdad, Iraq
Email: Raadahmed130@yahoo.com

Submitted: 30/11/2022. Revised edition: 31/3/2023. Accepted: 31/3/2023. Published online: 13/9/2023
DOI: https://doi.org/10.11113/ijic.v13n1-2.422

***Abstract*—There are at least three hundred and fifty million people in the world that cannot hear or speak. These are what are called deaf and dumb. Often this segment of society is partially isolated from the rest of society due to the difficulty of dealing, communicating and understanding between this segment and the rest of the healthy society. As a result of this problem, a number of solutions have been proposed that attempt to bridge this gap between this segment and the rest of society. The main reason for this is to simplify the understanding of sign language. The basic idea is building program to recognize the hand movement of the interlocutor and convert it from images to symbols or letters found in the dictionary of the deaf and dumb. This process itself follows mainly the applications of artificial intelligence, where it is important to distinguish, identify, and extract the palm of the hand from the regular images received by the camera device, and then convert this image of the movement of the paws or hands into understandable symbols. In this paper, the method of image processing and artificial intelligence, represented by the use of artificial neural networks after synthesizing the problem under research was used. Scanning the image to determine the areas of the right and left palm. Non-traditional methods that use artificial intelligence like Convolutional Neural Networks are used to fulfill this part. YOLO V-2 specifically was used in the current research with excellent results. Part Two: Building a pictorial dictionary of the letters used in teaching the deaf and dumb, after generating the image database for the dictionary, neural network Dark NET-19 were used to identify (classification) the images of characters extracted from the first part of the program. The results obtained from the research show that the use of neural networks, especially convolution neural networks, is very suitable in terms of accuracy, speed of performance, and generality in processing the previously unused input data. Many of the limitations associated with using such a program without specifying specific shapes (general shape) and templates, hand shape, hand speed, hand color and other physical expressions and without using any other physical aids were overcome through the optimal use of artificial convolution neural networks.**

***Keywords*—Deaf and dumb, sign language, gesture detection. hand gestures, region-based convolution, human hands, DarkNet-19 and YOLO**

I. INTRODUCTION

In computer vision, one of the most demanding tasks is detecting hands in unconstrained surroundings with high accuracy. Hand-related activities such as gesture recognition, hand behavior research, human-machine communication, and the understanding of sign languages are all connected to this technology. As a first step in action recognition, recognizing hands is one of the most challenging tasks in sign language since the shapes and movements of hands can vary widely. It is possible for a person's hand to vary in size, skin tone, and look. In addition, factors such as illumination fluctuations complicate the work even further. There has been a lot of interest in video frame detection of the human body in the recent decade. For face, upper body, and human body detection, the Viola & Jones algorithm was one of the most widely utilized. We had to come up with a new algorithm or approach for detecting human hands because the previous one was unable to do so effectively. The detection and classification of objects can be accomplished using a variety of classical methods, such as SURF, SIFT, PCA, and many more. According to previous findings, the suggested earlier techniques failed to identify, locate, and recognize human hands effectively. A system capable of accurately detecting and classifying human hand motion in a wide range of variations is therefore an essential tool for our research. Using Convolutional

Neural Networks, such as YOLO convolutional neural networks, ALEX Network, VGG-19, Dark Net.

II. Related works

MATLAB was used by Shreya Shi Narayan Sawant in 2014 to construct a real-time Sign Language Recognition system that can identify 26 gestures from the Indian Sign Language (ISL). Webcams are used to capture all of the indications and symptoms. In order to extract characteristics, the HSV color model is employed to preprocess these indicators. The Principal Component Analysis (PCA) approach is used to compare the characteristics generated. After comparing the attributes of a captured sign with those in a testing database, the shortest Euclidean distance for sign identification is determined. In the end, the recognition of gestures is translated into textual and auditory data. People who are deaf or blind can now communicate with one another through the use of this new innovation [1]. For example, skin detection, image filtering and image segmentation were used to address hand gesture identification. The technology is capable of recognizing American Sign Language (ASL) [2]. Statistics-based hand gestures, a subset of American Sign Language, were used by Hasnain A. Hasan and Dr. Jabbar Raheem Rashed in 2017 to avoid a hand gesture identification approach (ASL). An image of a hand motion is subjected to four steps in the proposed methodology: preprocessing, normalization, feature extraction, and classification. Using wavelet neural networks, which can deal with extremely complex interactions, information models are constructed. MATLAB is used to simulate the real-world system [3].

Vision-based hand signal identification frameworks have been created since hands are an important form of communication. Hand tracking, segmentation, extraction of features, and classifications are provided based on earlier work. The purpose of this technology is to enable two people, one of whom is unable to talk, to communicate by converting voice to finger sign and finger sign to speech. In Finger Sign (Gesture), finger signs are used to spell either spoken or written sentences. You can deal with a wide range of issues by using computer vision and pattern recognition [4]. In 1918, B. Sapkota, M. K. Gurung, P. Mali, and Gupta used hardware parts to create People who are deaf or hard of hearing can now communicate more easily thanks to the Smart Glove for Sign Language translation. In order to collect data, an Arduino microcontroller and a Bluetooth module are used, together with a glove embedded with sensors to identify various sign language gestures [5].

According to Norah. A, the discipline of Human-Computer Interaction (HCI) comprises a wide range of interactions, including gestures. There are many ways to recognize gestures, but one of the most common is using a camera to record them. This study uses a different algorithm to investigate the issue. Wavelet transforms and empirical mode compression were employed in this study in order to extract visual features and detect 2D or 3D hand movements. Another method for learning and classifying data besides Convolutional Neural Networks is the Artificial Neural Network classifier [6].

Towards the year 2020, Dinh-Son Tran Real-time fingertip identification and hand motion recognition can be accomplished with an RGB-D camera and a 3D convolution neural network (3DCNN). This system is capable of extracting fingertip locations and real-time motion recognition with excellent accuracy and reliability. They test the interface's accuracy and resilience by evaluating how well it recognizes various hand movements [7-16]. In 2020, Johansson, P., Ester, M., Yonatan, P., and Anastasya, M. will prevent an algorithm known as model E was used in this study to recognize the actual hand gestures. Kaggle.com was used to obtain American Sign Language (ASL) datasets [8].

III. Work Objective

Hand gesture recognition was utilized to construct a communication system for the deaf and mute. DARK NET-19 and YOLO net were trained to identify hands in video frames, while the alphabet dictionary was used to categories the identified hand image.

IV. CNN's methodology

According to the prior results, it was determined that the motioned earlier approaches are not suitable for detecting, localizing, and recognizing the hands of human subjects. This means that authors must employ a strategy that is clever, general, and efficient in dealing with a wide range of hand movements. Convolutional Neural Networks (CNNs) are the most well-known and widely used. The architecture of the DARKNE-50 neural net is described in this work showing in Fig. 1. The primary notion behind neural nets is to identify the network's intended purpose. For example, there are neural networks that are used for classification, while others are used for object detection and recognition and other processing. As a result, this is the first stage in completing the proposal. A key part of our proposal is the ability to distinguish human hand gestures by using nets that can find human hands precisely despite ambient effects. The second stage is to identify the segmented human hand region that has been found.

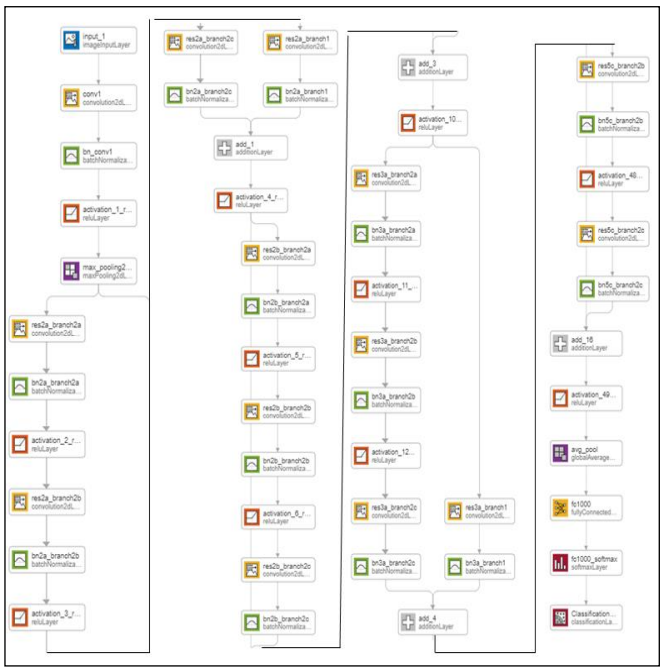

Fig. 1. Darknet-50 design

V. An Object Detection Neural Network

Detecting and locating a specific object is the primary objective of these networks. Object Detection Neural Networks are a type of such network. Numerous different network topologies (number of layers, layer types, and activation functions) can be used to achieve a certain network objective. The YOLO neural network was used in this experiment to detect and recognize objects.

A. You Only Look Once (YOLO) Neural Network

An object detection and localization method known as YOLO convolutional neural networks was used in these projects (human hand). There are two stages in the training process for each sort of neural network: Training and Application see in Fig. 2.

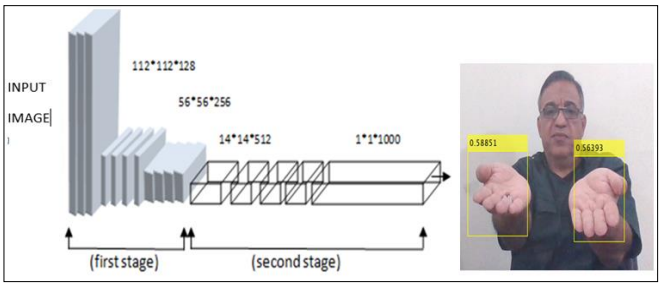

Fig. 2. YOLO network structure design

B. Workout Stage

Preparation for use is as simple as providing input data and requesting an output that corresponds to it. Depending on the data to be processed (i.e., the circumstance under consideration), the input and output data formats may change. When utilizing Yolo network, the main goal is to find the human hand and place it (contained in a box) anywhere in a video or image, regardless of the limitations and constraints specified previously. Fig. 3 shows the input/output format of the training data for this network.

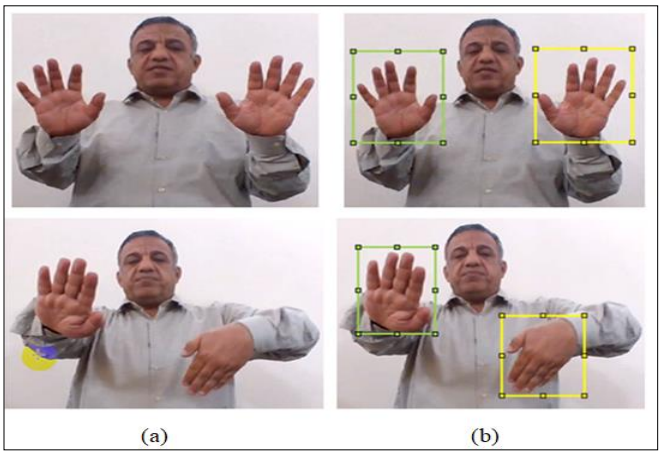

Fig. 3. The YOLO network training set input-output form. Row (a) contains input images, while Row (b) has outputs

The networks must be trained using a large number of training sets, such as the ones listed above. Therefore, Authors need to create these sets that cover at least all of the human hand positions in various imaging settings. Such training sets were created using a video streaming capturing technique in this study.

C. Algorithm.

Algorithm: Workout Stage:
Input: capture the video feed
Output: Sets of images used to detect hands

- For each video frame, there will be a caption.
- The ROI (regions of interest) can be found by employing a video labeler.
- Save ROI as a picture.
- Save both the input image and the ROI image in a single file.
- Create a second set of images by rotating and resizing the original images.
- Create a training database of all sets.
- End

D. Generation of Training Sets

To generate training data. A video labeler was employed. Once the video stream has been converted into pictures, each image is labeled with the region of interest (ROI) by encircling the human hands area with boxes. Once the desired number of training sets has been completed, the process must be repeated. The final number of training sets created was500, with no replicates that span a wide range of human hand configurations.

- The dimensions and conditions of the training stage

The training stage can begin after setting the learning parameter, which must be provided accurately and exactly in order to guarantee a growth and stabilization of the training curve after determining the architecture of the networks utilized see Table I.

TABLE I. SHOW THAT THE LEARNING PARAMETER WAS SPECIFIED PRECISELY

SOLVER	
Solver	sgdm
Initial Learn Rate	0.001
BASIC	
ValidationFrequency	15
MaxEpochs	15
MiniBatchSize	20
ExecuctionEnvironment	Auto
SEQUENCE	
SequenceLenght	Longest
SequencePaddingValue	0
SequencePaddingdirection	Right
ADVANCED	
L2Regularization	0.0001
GradientThresholdMethod	12norm
GrandientThreshold	Inf
ValidationPatience	Inf
Shuffle	Every-epoch
CheckpointPath	-
LearnRateSchedule	none
LearnRateDropFactor	0.1
LearnRateDropPeriod	10

A wide range of parameters, including activation function, learning rate, momentum rate, and Epoch number, must be provided before the network training process can begin. Other factors have an impact on how quickly the learning process evolves. Authors discovered that learning rate, momentum rate, and epoch number are the most interesting variables. These settings were set to the values displayed in the table above in this project. When the training begins, it can be monitored as shown in Fig. 4 as a function of the number of iterations. Table II covers the entire training procedure, as well as the training materials. Fig. 5 depicts the configuration's output outcomes, together with the location of those outputs.

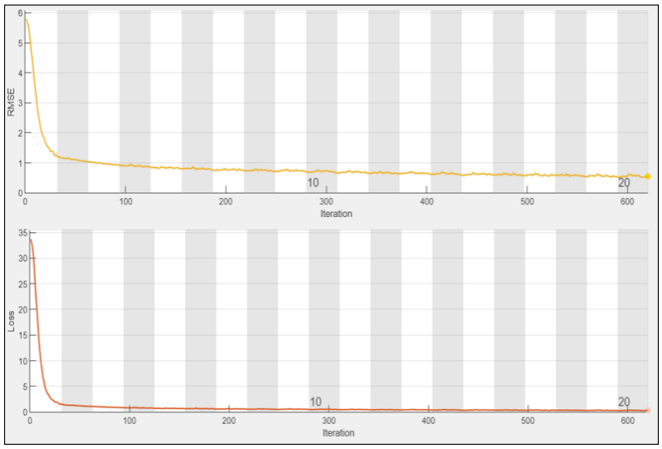

Fig. 4. Show how the learning process evolves over time as a function of the number of repetitions

TABLE II. SUMS UP THE TRAINING PROCEDURE IN A NUTSHELL

Epoch	iteration	time Elapsed (hh:mm:ss)	mini-batch RMSE	mini-batch Loss	Base Learning Rate
1	1	0:00:05	6.31	39.8	1.00E-04
3	50	0:03:42	1.37	1.9	1.00E-04
6	100	0:07:24	0.87	0.8	1.00E-04
8	150	0:11:12	0.69	0.5	1.00E-04
11	200	0:14:55	0.52	0.3	1.00E-04
14	250	0:18:38	0.59	0.3	1.00E-04
16	300	0:22:23	0.46	0.2	1.00E-04
19	350	0:26:12	0.5	0.3	1.00E-04
22	400	0:30:03	0.38	0.1	1.00E-04
24	450	0:33:53	0.01	0.7	1.00E-04
27	500	0:37:43	0.4	0.2	1.00E-04
29	550	0:41:32	0.32	0.1	1.00E-04
32	600	0:45:18	0.44	0.2	1.00E-04
35	650	0:49:06	0.27	7.10E-02	1.00E-04
37	700	0:52:53	0.29	8.70E-02	1.00E-04
40	750	0:56:54	0.23	5.20E-02	1.00E-04
43	800	1:00:59	0.27	7.10E-02	1.00E-04
45	850	1:05:04	0.28	8.00E-02	1.00E-04
48	900	1:09:08	0.21	4.60E-02	1.00E-04
50	950	1:13:12	0.23	5.30E-02	1.00E-04
53	1000	1:17:17	0.1	3.20E-02	1.00E-04
56	1050	1:21:23	0.18	3.40E-02	1.00E-04
58	1100	1:25:27	0.2	4.00E-02	1.00E-04
60	1140	1:28:41	0.19	3.50E-02	1.00E-04

Fig. 5. YOLO network training set result

- Beginning of the application process

Upon completion of the initial training phase, this is done only once; Application can be utilized at any time. The output can be defined at the application stage, which begins with the application of input data sets using the weighting factor (the networks layer coefficients).

VI. Classification Step

For this suggested project, Authors also have a classification step, which is responsible for translating hand gestures into printed characters. Based on the Convolutional Neural Network for Classification, this component is constructed using Dark Net-19 type classification net, with its own unique configuration of layers and training methods. Nets like this one often receive images of hand gestures that have been tracked from a video stream and output characters (sign language and abbreviations) associated with those gestures, which are typically the most commonly used abbreviations. Abbreviations numbers were used to define the sign language dictionary's 26 characters (here Authors proposed to use 8 different symbols). The configuration and response of (DarkNet) convolutional neural networks will be outlined here. As a result, Dark Net-19 will be used for this phase of the project.

A. Dark Net-19

Dark Net-19, which has 64 layers as depicted in Fig. 6. It can be seen in the training curve graph as shown in Fig. 7, the network accuracy (almost 99.53 %), despite a considerable increase in learning times, which is not of much importance here as the training procedure can be completed in one sitting (offline).

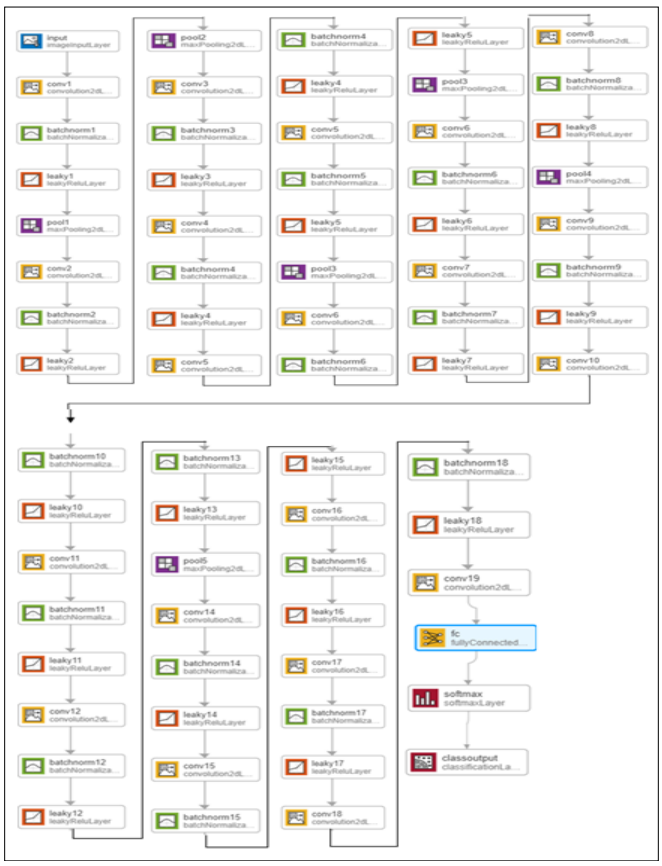

Fig. 6. The 64-layer design of the dark net-19

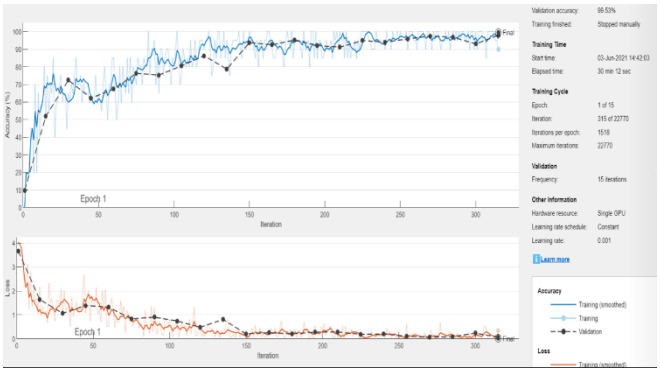

Fig. 7. Dark net -19 (64 layers) training

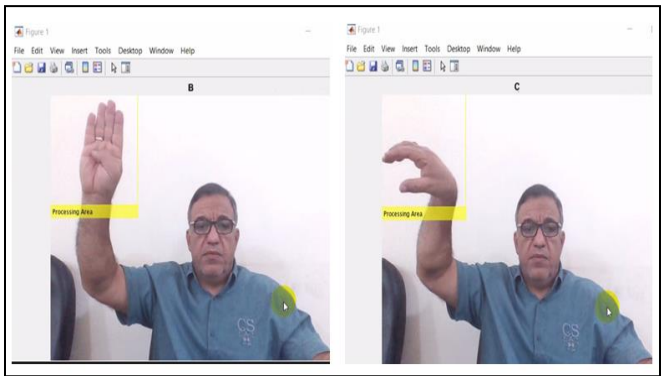

Fig. 8. Output results of Dark net-19 (64 layers)

The overall analysis to used (DarkNet-19) network can be given be the Table III from its accuracy and reliability can be found.

TABLE III. COMPARATIVE ANALYSIS OF THE USED NEURAL NET (DARK-NET19) FOR CLASSIFICATION

Network	Layers Number	Learnable layers	Fully Connected Layers	Time (sec) ~	Training Accuracy %	Validation
DarkNet -19	64	37	1	53	99.53	100

From the table, the accuracy as well as the validation accuracy. The main network architecture reflects memory burden on the computer and that means processing time. According to that the most suitable net to be used here is DarkNet-19. is the most suitable neural net that can adopted in the project.

VII. RESULTS AND DISCUSSION

Hand gesture recognition was examined using a variety of classic methods. For the all-sign dictionary, these methods were shown to be ineffective at separating the hand from the image and classifying the images into sign language characters. It is quite difficult to employ these procedures for general-purpose purposes because of the human variety in scale and orientation. Non-traditional approaches like CNN must be used, regardless of the difficulty and particular computing requirements, because this is the only way to get accurate results. When it came to processor speed, memory storage, and video processing, the computer used in this project had a very high specification. Authors used a Core i7 PC with 16GB of RAM and 8GB of VRAM (single GPU). In the beginning, Authors prepared a training set for the YOLO net object detection network, which was used to identify human hands. As shown in Fig. 3. As a first step, Authors recorded videos of hands in a variety of positions and orientations to ensure the image's universality. In order to locate the human hand in an image with great accuracy, many experiments were conducted. All three epochs and sizes of training sets are examined in these trials. It was discovered that increasing the number of training sets and epochs improved the responsiveness and accuracy of the network. However, the processing time that results from expanding and using this network is a significant disadvantage. To put it another way, YOLO networks eat up computing power and RAM. As a result, raising the epoch number will lengthen the training stage's processing time. As for the training sets, they also affect the training in a batch-like manner (video ram depending). Training sets, epoch numbers, and machine specifications all influence how long training takes, as shown in Table I. You don't have to worry about long-term training sets in this field because you may complete this level offline and then use it later online.

VIII. CONCLUSIONS

The neural networks specially the convolution neural networks are a promising technique for artificial intelligence applications as in human sign language. Neural networks in hand sign language are better than the traditional methods in computer vision.

YOLO Convolutional neural network is very promising in video and vision application like in this present project, where it's responsible to locating and detection of (ROI). For that reason, this kind of network YOLO v2 was adopted.

Hand detection and localization method that was proposed in this project is very promising method using YOLO net compared with Google pose estimation API regarding the facilities of Google Company in processing sites and datasets. The used dataset in this project consists of 499 images while the dataset of Google was nearly 13 million images.

The final quality of the system frame per second will depend mainly on the internet connection strength as well as the used computer specification especially memory size and GPU unit.

ACKNOWLEDGMENTS

The authors would like to thank the Iraqi Commission for Computer and Informatics, as well as the Informatics Institute for Post Grad, for their help and encouragement in performing this work.

REFERENCES

[1] Waheed, S. R., Suaib, N. M., Rahim, M. S. M., Adnan, M. M., & Salim, A. A. (2021, April). Deep Learning Algorithms-based Object Detection and Localization Revisited. *Journal of Physics: Conference Series* (Vol. 1892, No. 1, p. 012001). IOP Publishing.

[2] Sawant, S. N. (2014). Sign language recognition system to aid deaf-dumb people using PCA. *Int. J. Comput. Sci. Eng. Technol. (IJCSET)*, *5*(05).

[3] Farzi, A., & Tarjomannejad, A. (2015). Prediction of phase equilibria in binary systems containing acetone using artificial neural network. *International Journal of Scientific & Engineering Research*, *6*(9), 358-363.

[4] Rashed, J. R., & Hasan, H. A. (2017). New method for hand gesture recognition using wavelet neural network. *Journal of Engineering and Sustainable Development*, *21*(01), 2520-0917.

[5] Salim, A. A., Bakhtiar, H., Bidin, N., & Ghoshal, S. K. (2018). Unique attributes of spherical cinnamon nanoparticles produced via PLAL technique: Synergy between methanol media and ablating laser wavelength. *Optical Materials*, *85*, 100-105.

[6] Waheed, S. R., Rahim, M. S. M., Suaib, N. M., & Salim, A. A. (2023). CNN deep learning-based image to vector depiction. *Multimedia Tools and Applications*, 1-20.

[7] Salim, A. A., Bakhtiar, H., Krishnan, G., & Ghoshal, S. K. (2021). Nanosecond pulse laser-induced fabrication of gold and silver-integrated cinnamon shell structure: Tunable fluorescence dynamics and morphology. *Optics & Laser Technology*, *138*, 106834.

[8] Tran, D. S., Ho, N. H., Yang, H. J., Baek, E. T., Kim, S. H., & Lee, G. (2020). Real-time hand gesture spotting and recognition using RGB-D camera and 3D convolutional neural network. *Applied Sciences*, *10*(2), 722.

[9] Pratama, Y., Marbun, E., Parapat, Y., & Manullang, A. (2020). Deep convolutional neural network for hand sign language

recognition using model E. *Bulletin of Electrical Engineering and Informatics*, *9*(5), 1873-1881.

[10] Mohamed, R. A., & Hussein, K. Q. (2021). New Technique withConvolution Neural Networks (R-CNN's) Model for Hand Detection. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, *12*(14), 1657-1665.

[11] Salim, A. A., Bakhtiar, H., Shamsudin, M. S., Aziz, M. S., Johari, A. R., & Ghoshal, S. K. (2022). Performance evaluation of rose bengal dye-decorated plasmonic gold nanoparticles-coated fiber-optic humidity sensor: A mechanism for improved sensing. *Sensors and Actuators: A. Physical*, *347*, 113943.

[12] Salim, A. A., Bakhtiar, H., & Ghoshal, S. K. (2021). Improved fluorescence quantum yield of nanosecond pulse laser ablation wavelength controlled cinnamon nanostructures grown in ethylene glycol medium. *Optik*, *244*, 167575.

[13] Waheed, S. R., Adnan, M. M., Suaib, N. M., & Rahim, M. S. M. (2020, April). Fuzzy logic controller for classroom air conditioner. *Journal of Physics: Conference Series* (Vol. 1484, No. 1, p. 012018). IOP Publishing.

[14] Kadhim, K. A., Najjar, F. H., Waad, A. A., Al-Kharsan, I. H., Khudhair, Z. N., & Salim, A. A. (2023). Leukemia Classification using a Convolutional Neural Network of AML Images. *Malaysian Journal of Fundamental and Applied Sciences*, *19*(3), 306-312.

[15] Waheed, S. R., Saadi, S. M., Rahim, M. S. M., Suaib, N. M., Najjar, F. H., Adnan, M. M., & Salim, A. A. (2023). Melanoma Skin Cancer Classification based on CNN Deep Learning Algorithms. *Malaysian Journal of Fundamental and Applied Sciences*, *19(*3), 299-305.

[16] Waheed, S. R., Sakran, A. A., Rahim, M. S. M., Suaib, N. M., Najjar, F. H., Kadhim, K. A., Salim, A. A. & Adnan, M. M. (2023). Design a Crime Detection System based Fog Computing and IoT. *Malaysian Journal of Fundamental and Applied Sciences*, *19*(3), 345-354.