New Technique withConvolution Neural Networks (R- CNN's) Model for Hand Detection

Raad Ahmed Mohamed* [1]**Dr.Karim Q. Hussein***[2]

*[1] Computer Science, Iraqi commission for computer and informatics, Informatics Institute for Postgraduate Studies, Baghdad, Iraq. Raadahmed130@yahoo.com
*[2]Assist. Prof. /Computer Science Dept./Faculty of Science/ Mustansiryha University / Baghdad, Iraq .
Karimzzm@yahoo.com ,karim.q.h@uomustansiriyha.edu.iq

Abstract: The ability to listen and say the word are the most important aspects of communication, but many of us are unlucky because we were not born with this skill from God. These people are deaf and dumb. Many studies are currently being conducted to address the difficulties that these members of our society encounter in communicating with ordinary people. It is extremely difficult for mute (deaf and dumb) people to communicate their information to the general public. Because the average person is not adequately prepared to grasp various sign. It gets very difficult to communicate between these two types of people. Our research is only for the purpose of assisting these mute (deaf and dumb) people in leading a better life. Many computer vision tasks involving human hands, such as hand pose estimation, hand identification of gestures, human behavior analysis, and so on, are performed by humans. include hand detection as a critical pre-processing technique. However, due to the diverse appearance diversities of human hands, such Strong diffraction, weak resolution, fluctuating levels of light, a variety of hand gestures, and complicated interactions between hands and things or other hands are all factors to consider. such as (different hand forms, hand tracking, skin colors, scales, illuminations, orientations, gesture analgesia), accurately detecting hands is a difficult activity. in color pictures, as well as Interaction between humans and machines, recognizing of sign languages and so on). To address this problem, a region-based convolution neural networks (R-CNN's) was used, in which hand regions are discovered and hand appearances are recreated simultaneously using attributes derived from a region proposal. The R-CNN was shown to be suitable for hand gesture detection with acceptable error.

Keywords:hand detection, identification of gestures, region-based convolution, human–machine interaction.

1. Introduction

Among the most important but challenging issues in computer vision is robust hand detection in unconstrained environments. It's linked to a variety of activities involving hands, such as recognition of hand gestures, study of hand behavior, Interaction between humans and machines, and the recognition of sign languages. When it comes to action recognition, hand recognition is frequently the initial step, and it is frequently one of the most difficult tasks for sign language recognition since the forms and movements of the hands can vary greatly. A hand may, be differ from person to other, in size, skin color and appearance. Furthermore, the task is more complex when environmental effects play a significant role like lighting variations. . In the last decade, human body detection in video frames has received a lot of attention. Viola &Jones algorithm for face, upper body and human body detection was one of the most popular used algorithmsin human - machine interaction. Human hands cannot be detected by the previous algorithm, which force us to write another algorithm or method for human hand detection. There are many traditional methods that can be used to detect and classify objects like SURF, SIFT, PCA and more other. Many researchers have suggested methods for hand detection like.

2. Related works

In 2014 Shreyashi Narayan Sawant is used MATLAB to create a real-time Sign Language Recognition system that recognizes 26 motions from Indian Sign Language. Using a webcam, the signs
are caught. The HSV color model is used to pre-process these indications in order to extract features. The Principal Component Analysis (PCA) method is used to compare the produced features. The smallest Euclidean distance is obtained for sign recognition after comparing attributes of captured sign with testing database. Finally, the gesture recognition is transformed to textual and audio format. This technology allows deaf-blind persons to converse with one another.[1]

Traditional methods like Skin Detection, , Image Filtering, Image Segmentation and Template Matching to addressed hand gesture recognition .The system has the ability to recognize (ASL) American sign Language.[2]

In 2017 Hasnain A. Hasan, and Dr. Jabbar Raheem Rashed, prevent a method for hand gesture detection using Statistic hand gestures, which is a subset of American Sign Language (ASL). Preprocessing, normalizing, extraction of features, and classification are the four processes that the suggested methodology takes an image of a hand motion through. Because it can handle very complicated interactions, the wavelet neural network is used to develop information models. MATLAB is used to model the real-world system.[3]

As the hands plays an important communication mode, vision-based hand signal recognition frameworks have been developed. Various approaches for hand tracking, segmentation, extraction of features, and classifications are presented, based on previously announced work. The goal of this technique is to allow two persons, one speaking weakly and the other speak normally, to communicate by converting speech to finger sign and finger sign to speech. Finger sign (Gesture) is a subset of Sign Language that uses finger signs to spell spoken or written sentences. It is possible to cope with a wide range of problems by effectively utilizing computer vision and pattern recognition.[4]

B. Sapkota, and M. K. Gurung, and P Mali, and Gupta,in 1918 using hardware parts Smart Glove for Sign Language translation was used to provide a simple means of communication for people who are deaf or hard of hearing. This project comprises of a sensor-equipped glove, which detects various sign language gestures and feeds the information to Arduino and a Bluetooth module to transport data to an android phone. [5]

Human-Computer Interaction (HCI), according to Norah Alnaim, is a vast field that encompasses a variety of interactions, including gestures. Gesture recognition is the process of recording gestures that are created in a specific way and then being detected by a device such as a camera. This study examines the problem using a different algorithm. This study used image processing technologies including Wavelet Transforms and Empirical Mode Compression to extract picture features in order to detect 2D or 3D hand movements. Apart from Convolution Neural Networks, the Artificial Neural Network (ANN) classifier is used to train and categorize data (CNN).[6]

In 2020 Dinh-Son Tran Use an RGB-D camera and a 3D convolution neural network, they offer a unique approach for real-time fingertip detection and hand motion recognition (3DCNN). This system can extract fingertip locations and recognize motions in real time with high accuracy and reliability. They evaluate hand gesture recognition across a variety of gestures to demonstrate the interface's accuracy and robustness.[7]

3. WORK OBJECTIVE

In this work, CNN was used to address hand gesture recognition to build deaf and dumb communication system. YOLO net was trained to detect hands from video frames, while ALEX, DARKNET-19, RESNET were used to classified the detected hand image according to the Indian alphabet dictionary.

4. Region-basedConvolution Neural NetworkCNNs (R-CNNs)

In order to build the suggested system,nontraditional methods were used to overcome the major drawbacks of image processing, human diversity (specially with size, skin color ...etc. of human hands), image capturing console, light and shadows variation. All these variables affect the video images subsequently the detection and classification accuracy.

The main architecture of the proposed system consists of two stages; a-Hand detection stage and b- character classification

A: Hand Detection stage

In this stage YOLO neural network was used. The main architecture of this net is illustrated in Fig (1). It consists of 25 layers from input to output. This kind of net is very promising to detect objects (in our case is human hands) after suitable training. The main part of the network training is how to utilize suitable training sets. Therefore, a dataset of more than 700 extracted images for human hand from video movie montaged for this reason Fig (2). This step was time and burden consuming but essential to achieve high accuracy.

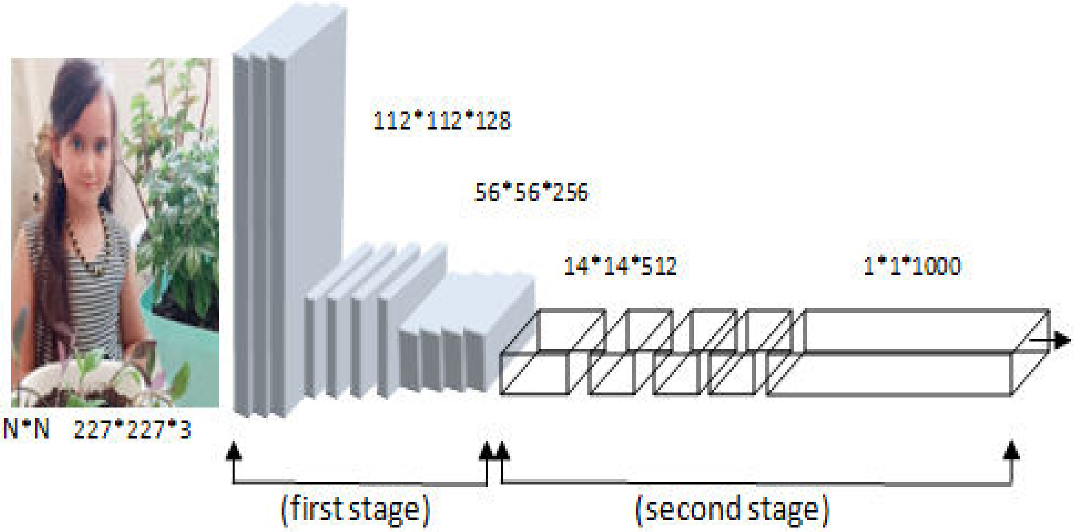

Fig. (1) YOLO network structure design.

B: character classification stage

In this stage ALEX, RESNET, VGG-16 and DARKNET-19 were used to classify the extracted hand image from stage 1 into alphabet characters according to the Indian Sign Dictionary. To use these nets, they have to be train first. In training stages 1200 image was generated for each alphabet character as well as space and end character using web camera and written program with MATLAB. So, to train network like ALEX more than 36000 images was used. The reason behind using multiple nets is to find the optimal net according to time and accuracy.

4. Algorithm

1.Create the first sub-segmentation of the input image.

2.Recursively merge similar bounding boxes into larger ones.

3.Build area proposals for object detection using these larger boxes.

Color similarity, texture similarity, area size, and other factors are considered in Step 2. In this post, we went through the selective search algorithm in great detail.

5. R-CNN Challenges

Selective Search is a really static algorithm which does not allows for any learning. As a consequence, region suggestions for object recognition may be erroneous.

Given the fact that all 2000 candidate suggestions have been submitted. It takes a very long time to train a network. We also need to train many stages individually (CNN architecture, bounding box regressed). As a result, implementation moves at a glacial pace.

Because analyzing an image with a bounding box regression model takes roughly 50 seconds, R-CNN could be used in near real - time.

Because all of the feature maps for the area proposed must be retained. It also increases the number of disc memory accessible for preparations.

6. Results and discussion

Many traditional methods were investigated to fulfil hand gesture recognition like SURF, SIFT, ORB and Skin Color Detection. It was found that these methods fail to isolate hand as well as classify the images into sign language characters, especially for the all-sign dictionary. These methods were scale, orientation dependent, which is very difficult to reach specially for general purpose uses according to the human diversity.

This is why one have to use the nontraditional methods like CNN regardless the complexity and special computer specifications needs. The specification of the used computer in this work was very high in processing speed, memory storage and video processing card.

A computer of core i-10 with 16 GB ram, 8GB VRAM (single GPU) was used. The beginning of the work was to prepare the training set of the object (human hand) detection network, which was YOLO net. Object labeler program was used to determine the objective region of the video frame as in Fig. (2). First, we have shot a video for different hands location, scales and orientation to ensure the generality of hand location in the image. Many experiments were carried out to ensure high accuracy tracking and detection as well as location the human hand from an image. These experiments cover, training set number, network epoch number and anchor size. It was found that increasing both training set and epoch number increases the network accuracy and response. But the major drawback of increasing and using such network is the processing time. As a matter of fact, using Yolo network consume processing time and memory. Therefore, increasing epoch number will increase the processing time (CPU speed depending) of training stage. The same thing can be said for the training sets, which will affect the training batch (video ramdepending). Eventually the training time will be last long depends on training sets, epoch number and computer specification as shown in Table (1). The good thing in this field, that you can fulfill this stage offline and used it later online, so there is no problem of long-time training set.

Table (1): training parameters of YOLO network

epoch	iteration	Time (h: m: s)	Mini batch RMSE	Mini batch LOSE	Base Learn RATE
1	1	00:00:07	6.31	39.8	1.0000e-$_{04}$
3	50	00:03:43	1.06	1.1	1.0000e-$_{04}$
6	100	00:07:24	0.71	0.5	1.0000e-$_{04}$
9	150	00:11:09	0.60	0.4	1.0000e-$_{04}$
12	200	00.14:45	0.47	0.2	1.0000e-$_{04}$
15	250	00:18:50	0.49	0.2	1.0000e-$_{04}$
18	300	00:22:40	0.37	0.1	1.0000e-$_{04}$
21	350	00:26:26	0.35	0.1	1.0000e-$_{04}$
24	400	00:30:12	0.29	8.3$_{e-02}$	1.0000e-$_{04}$
27	450	00:33:58	0.30	8.9$_{e-02}$	1.0000e-$_{04}$
30	500	00:37:44	0.26	6.7$_{e-02}$	1.0000e-$_{04}$

Fig. (2) Training area sample of YOLO network

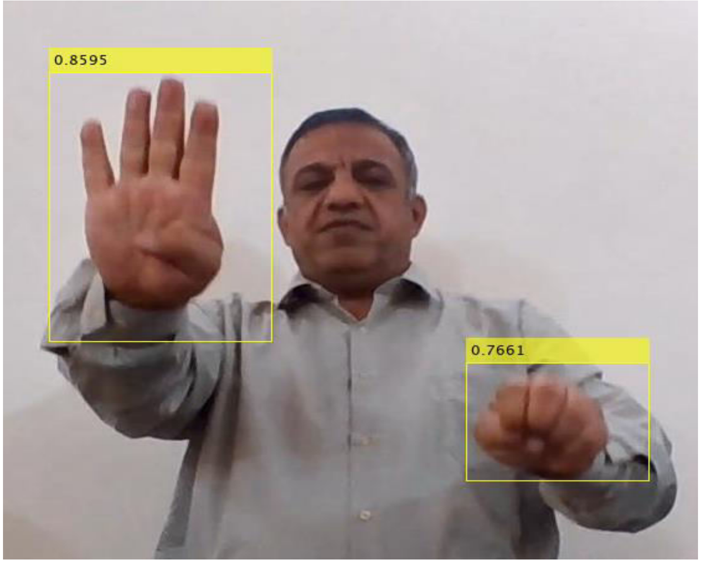

Fig. (3) Output detection of YOLO network.

Fig. (3) shows the detection and labeling the human hands from a video frame in real time with is more than 20 processed frames. One of the most important things that we have seen that increasing the processed frames will generate an ambiguity (error) in character classification in the second stage of the system. One knows that the final shape of any sign character can be distinguished simply, but what about generating the sign character shape more than 20 time per second as a stream to the classifier (this will be explained in the second stage of the system).

The second stage (character classifier) consist of CNN network like ALEX net, VGG-19 and Darknet to classify the all-sign language dictionary (Indian dictionary) with sign of space and end that used to separate the phrase and end the statements. Generally, the structure of the used net is well known and they were modified to satisfy our problem under consideration. Generally, such networks demand time as the in the first stage and the effective parameters for this network as in the first stage (training sets and epoch number). The training set for each sign character consist of

1200 image coversdifferent scale, orientation and reflection and a total of more than 32000 training images. Table (2) summaries the training procedure and Fig. (4) shows samples of the used sign language characters.

Table (2): Alex network training schedule

Epoch	Iteration	Time (h:m:s)	Mini batch Accuracy	Mini batch LOSE
1	1	00:00:01	3.13%	5.5299
2	1000	00:00:42	100.00%	0.0007
3	2000	00:13:10	100.00%	7.9149e-05
4	2950	00:18:32	100.00%	0.0001
5	3950	00:24:11	100.00%	7.0750e-05
6	4950	00:29:52	100.00%	1.6950e-06
7	5900	00:33:50	100:00%	3.3924e-05
8	6900	00:39:25	100:00%	7.2695e-06
9	7850	00:46:55	100.00%	0.0030
10	8850	00:53:12	100.00%	2.2852e-05
11	9850	00:58:39	100.00%	3.5317e-06
12	10800	01:04:24	100.00%	1.5348e-06
13	11800	01:09:57	100.00%	1.1737e-05

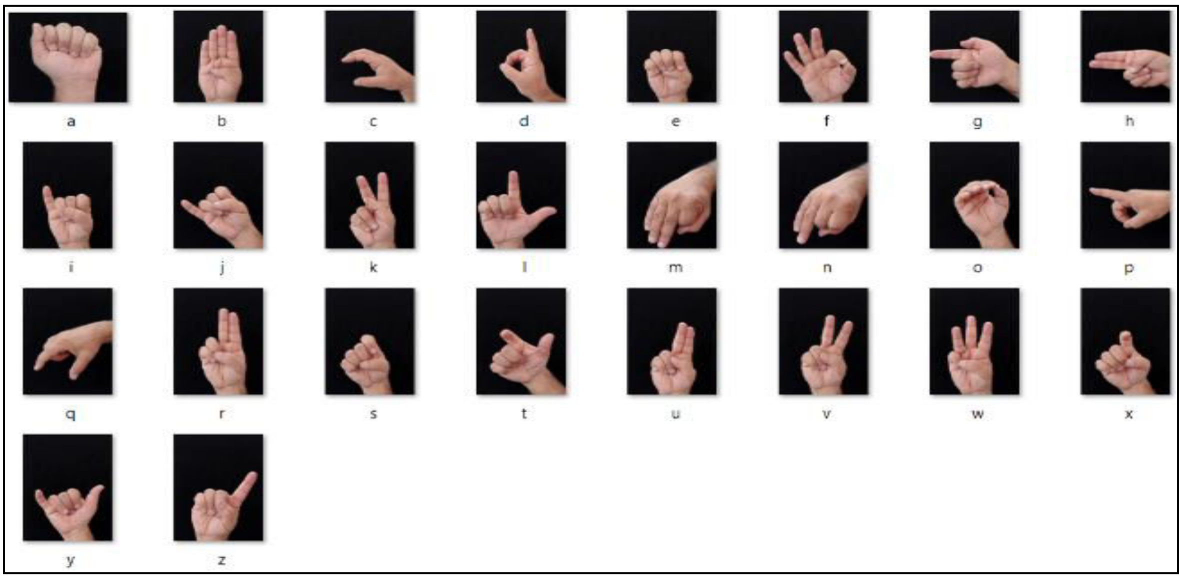

Fig. (4) Sample of sign language characters.

Finally, it can be concluded that the proposed system has the ability to detect classify stream of video frames into printed characters and statements with more than 20 frame per second. We have mentioned before that the high number of processed frames will generate ambiguity for character classification. When a user wants to form a character the transition shape will fed an a deform character shape into the classifier generating an error character classification. This challenge can be overcome by applying a classification accuracy threshold to distinguished the final shape and the corresponding classification.

After completing the proposed system and the optimal training parameters, we have turned to train the system to recognize abbreviation words rather than a character as shown in Fig, (5). This trend will create a burden onto the used network, but it will succeed eventually after modifying the network architecture. YOLO together with Alex, VGG-16 and Darknet -19 network found to be very suitable for this kind of job.

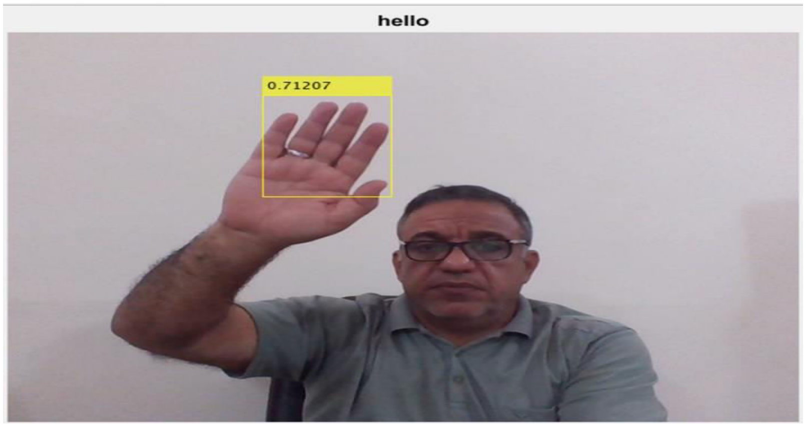

Fig. (5) abbreviation sign for hello word

7. Acknowledgments

The authors wish to thank the Iraqi commission for computer and informatics, as well as the Informatics Institute for Post Grad, for their support and encouragement in conducting this study.

8. References

[1] Shreyas Narayan Sawant, "Sign Language Recognition System to aid Deaf-dumb People Using PCA", International Journal of Computer Science & Engineering Technology (IJCSET), ISSN: 2229-3345, Vol. 5, No. 05, 2014.

[2] A Farzi, and A. Tarjomannejad, "Prediction of phase equilibria in binary systems containing acetone using artificial neural network", International Journal of Scientific & Engineering Research, Volume 6, 338 ISSN 2229-5518, Issue 9, September 2015.

[3] Dr. Jabbar Raheem Rashed, Mr. Hasanain Abbas Hasan, "New Method for Hand Gesture Recognition Using Wavelet Neural Network", Journal of Engineering and Sustainable Development, Vol. 21, No.01, ISSN 2520-0917, January 2017.

[4] P. Bhat, and A. Ansari, and S.D'silva, and A.Baphna, "Speaking System for Mute", IOSR Journal of Engineering (IOSRJEN), ISSN (e): 2250-3021, ISSN (p): 2278-8719, Volume 5, PP 31-36, 2018.

[5] B. Sapkota, and M. K. Gurung, and P Mali, and. Gupta, "Smart Glove for Sign Smart Language Translation Using Arduinos",1st KEC Conference Proceedings, September 27, 2018.

[6] N. Alnaim, "Hand Gesture Recognition using Deep Learning Neural Networks", Ph.D. Thesis, Department of Electronic & Computer Engineering, Brunel University London, 2019.

[7] D. Son, Tran, and N. Huynh Ho, and H. Jeon.Y,and Eu. T,Baek,"Real-Time Hand Gesture Spotting and Recognition Using RGB-D Camera and 3D Convolutional Neural Network",Electronics and Computer Engineering, Chonnam National University,doi:10.3390/app10020722, 2020.