Real-time sign language recognition system

Article *in* International Journal of Health Sciences · August 2022

DOI: 10.53730/ijhs.v6nS4.12206

3 authors, including:

Hawraa H. Abbas

43 PUBLICATIONS **215** CITATIONS

SEE PROFILE

Haider Ismael Shahadi

University of Kerbala

52 PUBLICATIONS **288** CITATIONS

SEE PROFILE

How to Cite:
Mohammedali, A. H., Abbas, H. H., & Shahadi, H. I. (2022). Real-time sign language recognition system. *International Journal of Health Sciences*, *6*(S4), 10384–10407. https://doi.org/10.53730/ijhs.v6nS4.12206

Real-time sign language recognition system

Atyaf Hekmat Mohammedali
Electrical and Electronic Engineering, College of Engineering, University of Karbala, Karbala, Iraq
Corresponding author email: atyaf.h@s.uokerbala.edu.iq

Hawraa H. Abbas
Electrical and Electronic Engineering, College of Engineering, University of Karbala, Karbala, Iraq
Email: hawraa.h@uokerbala.edu.iq

Haider Ismael Shahadi
Electrical and Electronic Engineering, College of Engineering, University of Karbala, Karbala, Iraq
Email: haider_almayaly@uokerbala.edu.iq

***Abstract*---**One of the more natural ways of interaction between the machine and human is offered by Human gestures which are useful for sign language recognition (SLR). Sign language is the speaking tongue of a segment of people who are known as Deaf people. they are the most people who benefit from the (SLR) by the Human-Computer Interaction (HCI). Most natural people cannot understand these kinds of languages, so Deaf people fail to communicate with them without this helper which achieves by the (HCI). This study introduced a proposed system that offered SLR in real-time for some gestures from the American sign language (ASL), by using one of the most suitable deep learning-based architectures that were called Convolutional neural networks (CNN) and choosing the Squeezenet module After comparing it with a traditional machine learning system that relied on extracting [HOG] features. Squeezenet presented the best result in accuracy reached to (100%) in the off-time testing and about (97.5%) in the real-time with a competition time of about 3.3sec for capturing the image, predicting it, and converting it to a text and spoken sentence. The system achieved this result without making any preprocessing for the image, which gave it simplicity and low computing time.

***Keywords*---**sign language recognition, squeezenet, convolution neural network (CNN), (HOG) features, real-time static hand.

Introduction

From the beginning of human civilization, language is essential to human interaction. It is the way how they communicated with others and without it, many thoughts might have killed in the minds of their owners. With the diversity of civilizations and the different lands erected by the language varies from one society to another. In every society, there is a group for whom language means nothing but signs and gestures. They are deaf and dumb people who use one of (138-300) sign languages that found around the world(Hurroo & Elham, 2020). World Health Organization (WHO) said on (1 April 2021) that there are (432 million) adults and (34 million) children who suffer from hearing loss. Usually, deaf people lose the ability to speak especially those who are born deaf, so sign language for the first way that they used to describe their own needs and what they want from the other normal people(D. Kumar et al., 2022).

Sadly, these signs or gestures mean nothing to all those who do not have a deaf one in their family or acquaintances. For this reason, HCI achieved good results to bridge this communication gap. It will be a tool for deaf people to communicate their thoughts, and also an excellent interpretation to understand what the deaf need from non-sign language users(Adeyanju et al., 2021). Gestures may be done by using hands only or with one part of the body like head, lips and ayes. For hand gestures signers use: one hand, or two hand. Static or dynamic mode. Represent (alphabet) of any language or some words in this language(Ekbote & Joshi, 2016).

Sign language gets its special name from the natural language that gives its meaning like American sign language(ASL), Bangla Sign Language, Korean Sign Language, Arabic sign language, Indian Sign Language,... Etc. (Haque et al., 2019) HGR is an essential element of HCI, which represents translating sign language for different forms of gestures. The first interaction that a computer makes it when the gesture take its way to the device which is done from three types of approaches: sensor-based, vision-based, and hybrid-based(Ahmed et al., 2018) Sensor-based approach system extracts the measurements of a hand, i.e., hands position, the orientation of joints, and the velocity of the hand. It can be conducted by using specific sensors, and microcontrollers such as smart gloves, leap motion controller, Microsoft Kinect, and digital camera. Although a sensor-based system offers a high recognition rate it is stile expensive and uncomfortable for the users, especially for wearable ones(Shin et al., 2021)

Vision-based approach system focuses on interpreting the hand's action in different modes of interaction. The system requires only a camera which uses in capturing the gesture image that is sent to the computer to extract its features. Applications developed using a vision-based approach are cheaper than the others but sometimes It is affected by the characteristics of the environment in which the gesture images are taken such as lighting, occlusion (the hand gesture is occluded by other parts of the body), and complexity of the backgrounds(Jayshree R. Pansare & Ingle, 2016) Hybrid-based approach system tries to make a mix of the two above approaches to improve system performance such as using of smart gloves with RGB color model (Red, Green and Blue) cameras(Van Culver, 2004)

All these approaches offer some features to the system about the gestures which are get its classification after passing through some preprocessing and processing steps. These three steps are different from one system to another according to the type of features that extracted. Features can be extracted by one of the feature extraction methods like: (HOG, LBP, and SURF) then traditional classifiers are used for recognition such as (SVM, KNN). In another direction there is an automatically hidden extraction by one of the deep learning techniques like: (CNN and ANN) can be used, they can make the two steps (extract feature and classification) invisibly but they will need more training time for learn the machine(Oudah et al., 2020)

Related work

Working in real-time is the most thing that require for many applications which used HGR models. A system is said to be working in real-time when gives a fast correct response at the most available minimum time after the signer makes a gesture. For this reason many researchers tried to offer these characteristics in their systems.

The procedure of any system must begin from the offline step by making a data collection of the gestures which a translation is wanted, then the system must divide this data into two groups: training and testing data. The training data is used to learn the machine in some steps, while the testing data is used to make validation to that learning. The time that is needed for the offline learning step is not so much important, because the required speed is in real-time. Exactly the contrary sometimes the long training time may be a cause of high accuracy later.
In the real-time, the system will use the information stored about the offline step to make recognition for any new entry gesture to the system(Jaramillo-Yánez et al., 2020). Here there is a literature review for some proposed systems that worked in real-time and depended on (a vision-based approach) from 2011 to 2021.

Author in (Dardas & Georganas, 2011) was proposed a system for (10)-bare hand posture with cluttered background to extract SIFT features via bag-of-features after made a face subtraction then skin detection, and hand posture contour comparison algorithm. K-means clustering and SVM classifier was used to achieve (96.23% as accuracy in 0.017 sec) in real-time.(2011)

Author in(Jayashree R. Pansare et al., 2012) recognized single-hand static gestures for 26-ASL alphabet by the system which contained: preprocessing step for detecting the hand skin from the input image then converted it to the binary form, after that a median filter was used to reduce “salt and pepper” noise. Gaussian filter of size 5 × 5 was used for smoothing, then Morphological operation applied to be ready to extract region of interest (ROI). At last edge detection by “Sobel” method was used to extracted features and classify the result by Euclidian distance to get an accuracy of (90%) in (0.5 sec) recognition time. (2012)
Author in(Ahmad & Akhter, 2013) tried to offer a method that converted the contour of the image to orientation based hash codes, that for projecting it to a (3D address) space bounded by hamming distance to get an accuracy of about (82.1%) and the average time recognizing was (7.36ms) for 10- static hand

gesture. Also, this system was designed to make a Background Subtraction, Color segmentation, and Select Region of Hand Gesture (ROI region of interest) (2013)

Author in(Rahaman et al., 2003) had a system that converted a two static hand gesture of Bengali Sign Language. Haar-like cascaded classifier was used to detect hand area, Skin color-based method for segmentation, then image was converted to a binary form at last, KNN classifier was used to achieve a (96.46% accuracy) and (93.56milisec). (2014) Author in (Jayshree R. Pansare & Ingle, 2016)used Edge Orientation Histogram (EOH) features to achieve an accuracy of (88.26%) by using the Support Vector Machine (SVM) classifier. This was done within recognition time of (0.5sec) for ASL gestures which capturing by web camera in complex background. The preprocessing step was done by converting the image color and removing the noise and morphological operation. Skin detector was applied too. [2016]

Author in (M. M. Islam et al., 2017) proposed a system which began with a preprocessing step for the capture image by making a resizing to (260×260),RGB to binary, median filter, and rotated images. Then extracted fingertip finder, eccentricity, and elongatedness as A features. The accuracy was (94.32%) in real time environment after using (ANN neural network) for classification the study did not explained what the overall time that required for recognition the gesture but the response time to take an image frame of the gesture it was 10sec. (2017)

Author in(Taskiran et al., 2018) developed a system recognized ASL gestures in real time by using CNN deep learning architecture to learn the machine with Massey University dataset. After that, for the real time testing converted the image from RGB to YCbCr space, then noise reduction was done, later a skin color was determined using the convex hull algorithm for a certain frame of hand gesture, then utilized the stored information from the training step to get the recognition for the input frame. The accuracy that was achieved was 98.05%.(2018)

Author in (Bohra et al., 2019) was getting an average prediction time of 0.000805 seconds with an accuracy of 99% when using a system that began with capturing 40 different gestures (2400 images for each one) to get 96000 as a total number, then the preprocessing step started with converting the image to the grayscale. Hand detection was done by defining the range of color of the skin and the segmentation was performed. Reducing the noise by median blur to find the contours in the image better. After that maximum contoured area was representing the hand gesture. At last Convolutional Neural Network(CNN) was used for the classification step. The system was proposed for two way(sign to text) and (text to sign).(2019)

Author in (Kadhim & Khamees, 2020) proposed a real-time ASL recognition system depending on The CNN architecture by using the VGGnet. The dataset was of 61.614 images with around at least 2200image for each of the 28-classes. All the images were resized to (224 x 224) pixels and labeled to be suitable for feeding to VGGnet at (70:30)%rate between the training data and the validation

data. This system took more than 7-days to train the model, at last the accuracy was 98.65%.(2020)

Author in (Mujahid et al., 2021)proposed a model based on YOLO(You Only Look Once)v3 and (DarkNet-53 CNN). YOLO was used to label the training input data which was finger-pointing positions of numbers: (1, 2, 3, 4, 5). YOLO was used for the affine transformation by increasing 2-fold, each image was duplicated for reading and training, both left and right hand, by flipping it horizontally. Then fed the results to the DarkNet-53 which was used to train the dataset, and for the detection, YOLOv3 was used. The accuracy of that system was 97.68%.(2021) All the chosen systems presented in the above review, although they were achieved high accuracy in some of them and less computing time there was an obvious complexity because of the preprocessing or the processing steps, except that in (Kadhim & Khamees, 2020) which used the CNN technique but the long training time (7-days) that needed for learning the machine was not attractive if compared to a system like what our study that proposed.

So systems of (HGR) are becoming more critical when they tested the translation in real-time. This study offers a SLR for some static one hand gestures of (ASL) in real-time without complexity of preprocessing steps such as the segmentation for the images to separate the interested region which represent the hand only. The main approach used is the vision-based by a webcam, so. The main contribution of this study is proposing a HGR system in real-time by testing: (extract HOG features then select some of the most important features)&(CNN by Squeezenet architecture) without segmentation process for a collection hand gestures images dataset that belongs to a female, male and a child, then making a comparison between the results to choose the most suitable system which achieve the aims that offers:

- High accuracy.
- Fast recognition in minimum real-time.
- less memory without any segmentation or complex preprocessing.
- Applicate the system in the smart environment by converting the gestures to spoken sentences to act as orders for the smart device “Alexa”.

Materials and Methods

searching for the optimal system that achieves the most assured results makes the researcher obliged to move in different directions to get the required reliability. In this study, 2-architectures systems is submitted to achieve exacted recognition in the real time. These two choices were based on the results obtained from our previous work that was submitted for publication in the TEM Journal , where the study was presented there on the basis of testing more than one type of features extraction algorithm: (Euclidean dis., geodesic dis., HOG, LBP, SURF and GLCM), and more than one type of feature selection algorithms: (CFS, ILFS, DGUFS, Relief, and LASSO), and more than one type of classification algorithm (SVM, KNN, and Bagged decision tree), and the best results were From HOG as a feature extraction algorithm and from LASO as a feature selection algorithm as well as KNN as the best classifiers.

This same work was done with deep learning algorithms CNNs, where tests were made in the same our study for several architectures: Googelnet, Alexnet, in addition to Squeezenet, which provided the best results as well. Each system must begin in the offline and after saving the classification results, we move on to trying it in real time.

Traditional Machine Learning [TML] system:

This system which is shown in Figure(1) one of the classical techniques that depend on extracting the features of the image manually, selecting the most substantial features, at last sending to one of the classification methods for learning the machine and then storing the result to run the system in real-time later. Different types of features can be extracted from the image by this sequence of steps, depending on the characteristics of this image and the accurate result of any type

Figure 1 the block diagram for (FE-FS) architecture in the off-time

Features should contain the important information to discriminate between classes, be insensitive to unnecessary input variable, and be limited in number to allow efficient computation of different classifiers and reduce the amount of training data needed(G. Kumar & Bhatia, 2014) Feature selection(FS) was employed in our system for removing irrelevant, redundant, and noisy information from the data, often leading to better performance in learning and classification tasks (Roffo, 2016). After testing many types of (FE)&(FS), a histogram of oriented gradient (HOG) (Mahmud et al., 2019) feature was used for extraction, and for feature selection technique The Least Absolute Shrinkage and Selection Operator (LASSO) (Wan et al., 2007)method was used to rank the features and to reduce the unneeded ones. This choice because of the best result that achieved in experiment as compared with other methods.

Dataset collection

The sign gesture used is a Static one-hand gesture independent of other body parts in different environments and backgrounds that differ in complexity. Three volunteers helped in collecting the images by the webcam of the laptop to be as (i) male gestures(ii)female gestures (iii)child gestures. Figure(2) shows some examples.

Figure 2 : hand gesture images for male, female, and child

The total input images in the offline step are (600)image,(100)image for every gesture because we chose only six gesture from the ASL: (a, b, c, i, v, 5).

Preprocessing

To find out how different sizes affect the recognition process, the photos used in the analysis are scaled up and down. The optimum image size must be carefully considered because different image sizes transmit different information. Image resizing's goal is to reduce the data size, which reduces processing time(Barnouti, 2016)

So this step in our system is limited to reducing the size of the image to (128*128) pixels. This choice of image size is the best in the results of the classification step (accuracy & computing time) among the three tests that we were done with the [HOG] feature extraction algorithm: (256*256),(128*128), and (64*64).

Features extraction

In the system design for recognizing gesture, the feature extraction method (FE) helps in the limitation of data dimension. So the most important decision is how to extract the features. Because the feature vector for the image is obtained here. Vectors for all the input images are assembled as a matrix and fed to the next stage, which is the classification step, where the matrix is used to train the classifier. Histogram Oriented Gradient [HOG] is a good feature that can be extracted because it is independent of preprocessing and illumination of images. In the local portion of the image, the (HOG) focuses on the shape of the object and it is based on occurrences of gradient orientation. It is providing the edges and their direction(Mahmud et al., 2019) at the first, the image must be divided into smaller regions called (cells) and then the gradients and orientation are calculated. At last, a Histogram for every one of these regions would generate separately.

The cell size chosen is [8*8] for one image, which would lead to the final matrix of (k*8100) where k is the number of the images. Also, here many tests were made before choosing the cell size from these forms: {[2*2],[4*4], and [8*8]} with every image size that we were testing in the (2.1.2) subsection. That means by these three forms of cells and those three image sizes the experiment was repeated nine times before our suitable choice [8 8]cell and (128*128)size. Figure(3) shows an image and its [HOG] features for one sign.

When looking closely at the image that showed the HOG features, we will notice that there is a clear distinction for the Oriented Gradient points of the hand boundaries by changing the direction of its vectors depending on those boundaries. The stability of this feature with every time the image is changed for the same sign (which was captured with different backgrounds and environments) helped in giving a very excellent classification later.

Figure 3. Original one sign image and with its [HOG]

Feature selection

One of the important processes in the image processing is the feature selection(FS) step because the image had different details and some of them do not effective, if the selected information about the image is strong, the classification step will be more efficient. So the goal of this step is how to get this efficient result by the least number of data for each image. (Figueira-Domínguez et al., 2020)

Three types of methods developed for this purpose:

- Filter methods are a variable elimination, the highly-rated features are picked and applied to a predictor using filter methods, which act as pre-treatment to rank the features.
- Wrapper methods are also a variable elimination, the predictor is wrapped in a search algorithm that a subset will be found to give a highest predictor performance.
- Embedded methods are without dividing the data into training and testing sets. These methods involve variable selection as a part of the training process.
-

For every method there are many algorithms were submitted like: 'cfs' Sorts features according to pairwise correlations [Filter method], 'DGUFS' Dependence Guided Unsupervised Feature Selection [Wrapper method], and 'LASSO' Least Absolute Shrinkage and Selection Operator [Embedded method] (Roffo, 2016) After testing some types, our choice was the LASSO algorithm because of its best results in improving the accuracy and reduction of the memory space by removing some features [F=8100] to be [selected features=8000] the procedure was by ranking the features from the strongest to the weakest ones then removing the last 100 weakest features. After preparation of the features vectors for the images in the dataset, They are grouped as a matrix and carried over to the classification stage.

Classification

Needing for Classifier is in order to recognize hand gestures from each other many traditional Machine Learning methods are used for that. There are two categories of machine learning techniques: supervised and unsupervised. Supervised machine learning is a way of learning a system to recognize classes in input data in order to estimate future data. Supervised machine learning applies a set of known training data to labelled training data(Mapari, 2014) Two classifiers are used in this study to make a comparison between the results: support vector machines (SVM) and k-nearest neighbor (KNN).

- Support vector machine (SVM) is a supervised learning model which is used for regression and classification of objects. Each data sample is represented as a point in a plane in SVM, and the training set categories are separated by gaps that are made as obvious and as broad as feasible for successful classification. Hyperplanes can divide the data into classes by pass through these gaps. SVM use some of functions that are defined as the kernel. It take data as input and transform it into the required form. (Lahoti et al., 2018) In this study kernel function that used was a cubic function, auto scale mode, and one vs one multiclass method the data would be divided into training and testing data training ones for learning the machine, and testing data for prediction to know the accuracy of the application.
- The k-NN technique is a non-parametric method that is used for classification, regression, and distance metric. The magnitude of k can be changed it is known as cross-validation. If the value of k is low, then the curve is overfitting; if the value of k is too high, then the curve is underfitting. Measuring the distance is by either Euclidean distance, City block/ Manhattan distance, or Minkowski distance function(Jain & Salau, 2019) In our work Euclidian distance is used for measuring the distance between points and chose the nearest distance between them. It work by taking a data point and looking at the 'k' closest labeled data points which was then a signed the label of the majority of the 'k' closest points.

Deep learning using a convolution neural network (CNN) system

One type of machine learning is deep learning and one of the most important techniques for image classification in deep learning is a convolution neural network (CNN). Many models are designed as (CNN) nets such as Googlenet, Alexnet, and the one that we are chosen squeezenet(M. R. Islam et al., 2018). Fig.(4) shows the block diagram for this system which is tested in the off-time then several real-time tests are done:

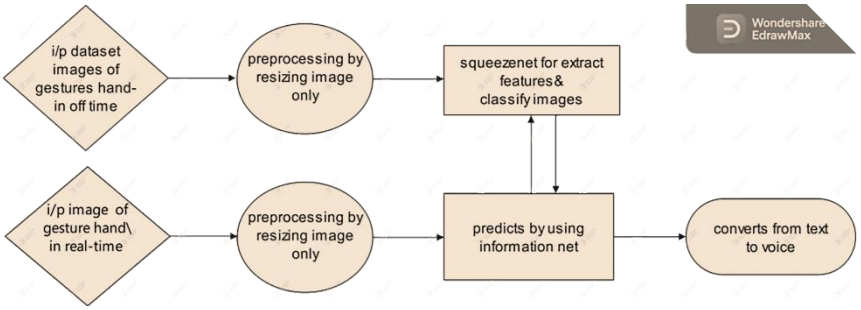

Figure 2 Block diagram for deep learning (CNN) architecture

Squeezenet is a pre-trained algorithm, with a number of layers, and a limited size of the input image. the goal of designing the squeeze net was to achieve creating a smaller neural network with more fewer parameters that can easily fit into computer memory and transmitted over a computer network more easily too. It includes of 10 layers; begins with a convolution layer, then there are 8-fire modules. At the last, another convolution layer is there. as it shown in Fig.(5). Squeezenet is with 50 x fewer parameters than Alexnet (Hameed et al., 2020).

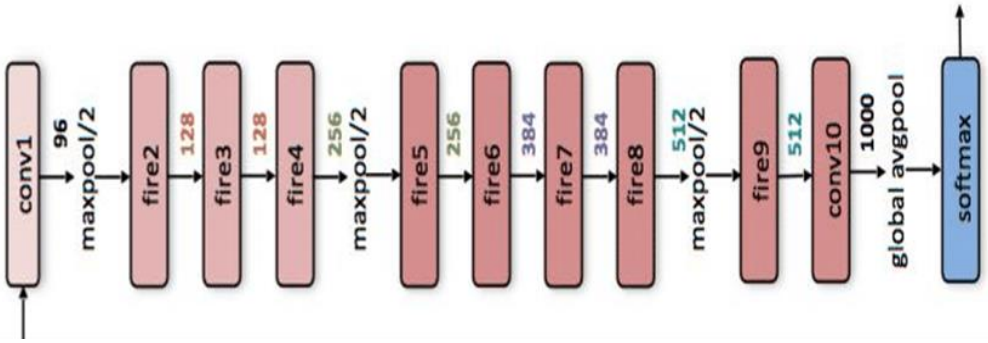

Figure 3 Squeezenet architecture

In our work the images that collected as described in subsection(2.1.1) were resized to 227x227x3 pixels to be suitable for the input layer of the net. The maximum number of epochs was set as 10, the Initial Learn Rate was 10^{-4}, and the data was divided with a 70% randomised ratio that meant the validation data was 30% randomised.

Results and Discussions

Results from the systems in our study were in two phases (i) off-time, (ii) real-time. This is done by storing the results that are getting in the off-time phase classification and then getting the recognition for every capturing sign image in real-time.

Off-time phase

After collecting the images dataset as shown in subsection (2.1.1), the dataset has been entered into the systems for (training and testing) phases and then saved the results of machine learning for the real-time test phase.

Results for [TML] system

For the two classifiers that we used(SVM,KNN) the valedation result is 100% accuracy but they are differ in time that needed for that SVM:98.356 sec, and with KNN:36.216 sec. the cross validation that we chose was:5-fold. this is shown in Fig(6).

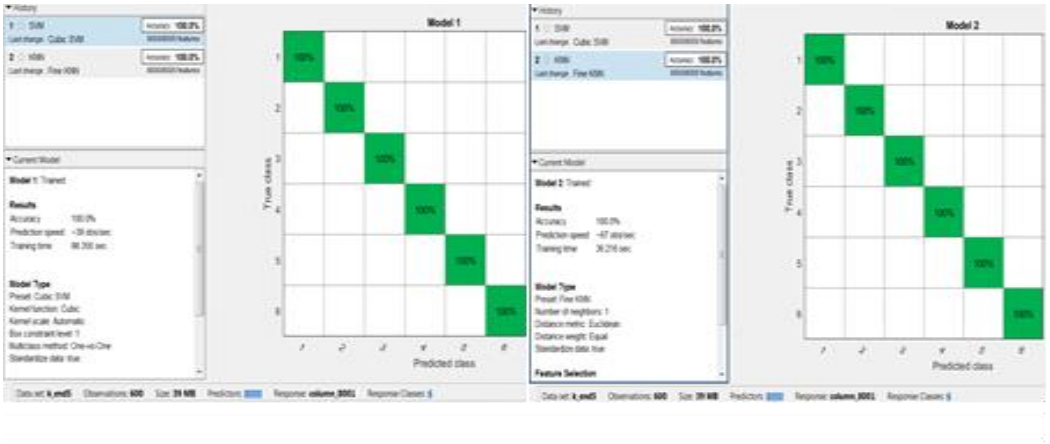

Figure 6: Results for the classification step by used (SVM),(KNN)

Results for Deep learning using a (CNN) system

Part of the analysis result for the squeezenet architecture that used in this test was shown in Fig(7), which shows the discretion of the input image in the first step and then indicated the specifications of the layers in a row. While Fig(8) shows the training progress of the work to give validation accuracy of 100% in training time of (163min37sec) for all the 10-epochs. Moving on to Fig(9) gave a note for the results of prediction for a six randomly selected gestures each time from the validation data. The time required to predict the signs using the stored data from the training phase was 12.927 sec for every six gestures.

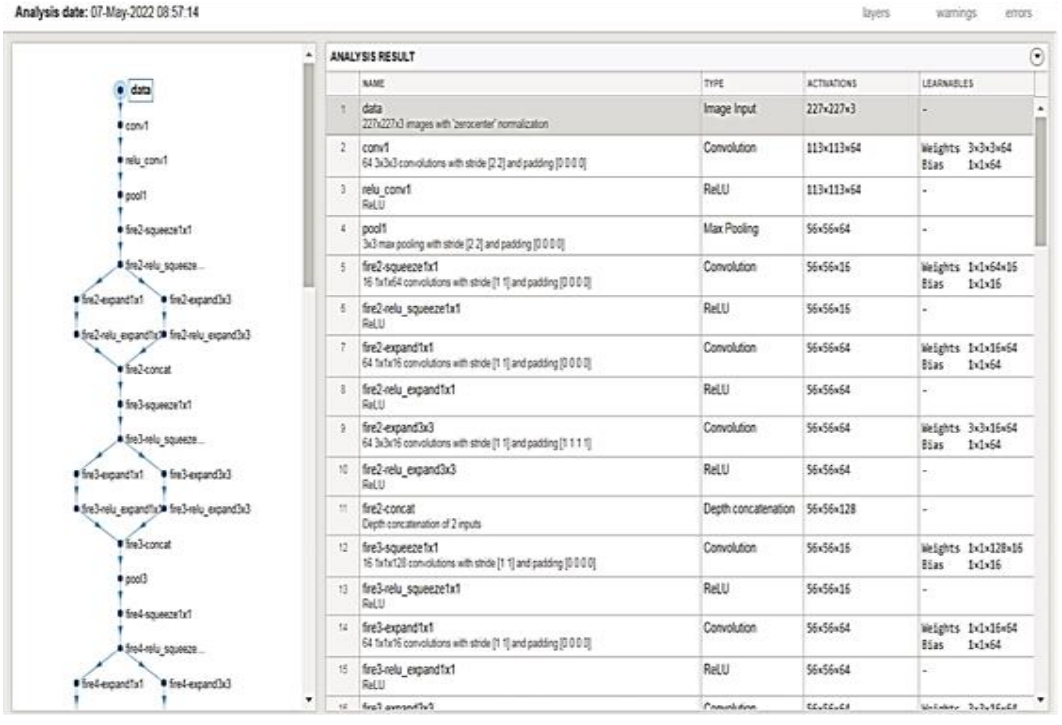

Figure 7: analysis result for the squeesenet architecture

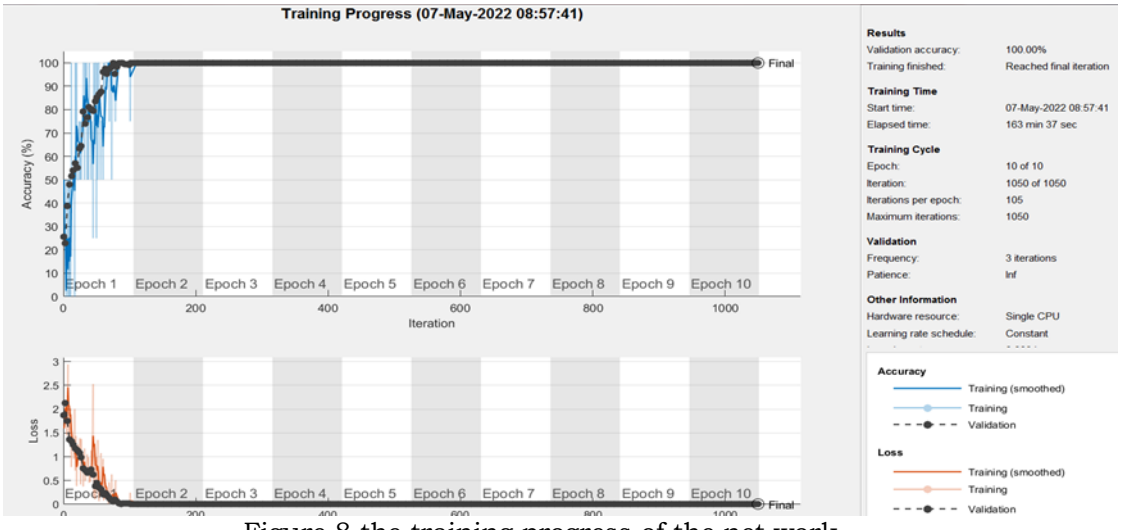

Figure 8 the training progress of the net work

Figure 9: results of prediction for a six randomly selected gestures, the process was repeated

Real-time phase

Real-time system means that this system can achieve its response at the time that the user used it. In the case of recognizing the sign language, the user needs to obtain a translation of his hand gesture at the moment the picture is taken when he is standing in front of the camera of the system used.

Results of the instantaneous response

With the two proposed systems, an algorithm was implemented at first for giving an instantaneous sign prediction and the output shows the label name of the predicted class as a title for the captured image in fractions of a second. With every change that the user makes in his gesture which is captured immediately by the camera of the system, there is a display for the gesture image that is titled with its meaning as a text. The first system sometimes failed to give a correct momentary prediction, but with the second system, the results always were correct. Fig(10) shows the instantaneous response which gives a high accuracy although the colour of the chosen background is very close to the skin colour of the hand.

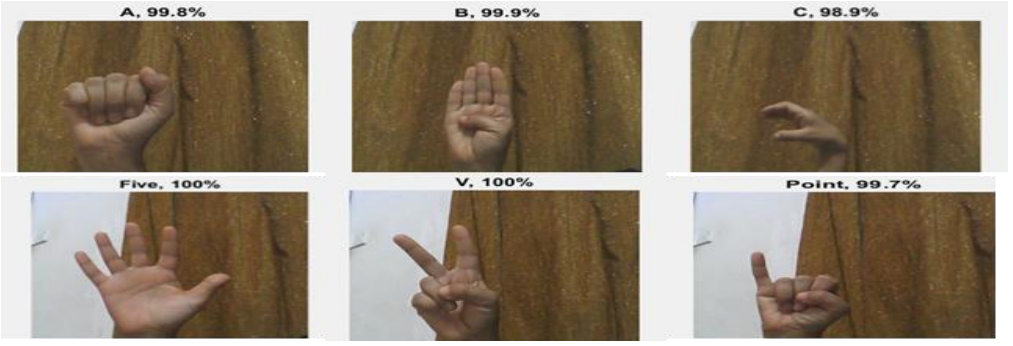

Figure 4 the instantaneous response for the system

To achieve our goal of interacting in smart environments, we not only included the text as a result but also added voice output. So the algorithm is adopted with clear and audible audio output implementation which may need a little longer time. Usually, systems are designed to show the output of the prediction process as a text which represents the labels of the classes that were classified in the classification step, and this will be considered (gesture to written text) translating. In our work, after the prediction step was done the system allocates an (order Phrase) for every label and by the 'switch function' in Matlab the predicted label will be a (text) order then by the synthesis net it converts into an audible sound, however, for explaining the results the output here is with a titled image as well as the sound Many applications may use the hand sign recognition technique, in our real-time test which convert the sign language to sound orders that can be guided to the [Alexa] device which is A virtual smart assistant developed by Amazon to use in smart environments, it is an intelligent personal assistant (IPA), (Lopatovska et al., 2019).

Speech Synthesis system

A speech synthesizer converts written text into speech using a computerized voice. It is an output where a computer simulates a voice and reads the word aloud. So it is generally called text-to-speech, not only should machines speak clearly, but they should also mimic the voices of people of various ages and genders. The text-to-speech engine is becoming more and more common due to the rise in the use of digital services and the increased reliance on voice recognition. Synthesizer is used for many purposes, it is helped users, especially those (blind, deaf-mute, and handicapped) who are needing special care and support. it is helping in modern education techniques also(Suryawanshi et al., 2014).

The intention of a text-to-speech system is to convert a randomly given wording into a speak waveform. The most important form of text-to-speech systems are processing the text and producing Speech. So the speech creation component's goal is to create an auditory waveform from scratch. The pieces of the recorded words have been concatenated in an attempt to create speech(Shirbahadurkar & Bormane, 2009). The Speaker System that used here is Speech Synthesizer system net with:

•Speaker Rate = -4 which represented the speed of the speech so the choice must be made with care so as to ensure that Alexa can distinguish the spoken command.

•Speaker Volume = 100 This choice represents the intensity of the sound, so the distance between the speaker and Alexa affects how much the volume is raised or lowered

•Speaker gender= female.

Table(1) shows the sound orders chosen for each one of the six signs.

Table 1 sound options that chosen for each one of six signs

	Sign in ASL	sound option
1	A	Good morning Alexa
2	B	Open the door
3	C	Good night Alexa
4	Point	Close the door
5	Five	Play alarm
6	V	Play music

Results for TML system

Although the best result that getting in the off-phase, there were many tests done here for choosing the suitable size of image and cells to extract [HOG] features. In fact the results in the off-time phase usually were very good in accuracy, except the choosing of (64*64) which gave (98.1%) in off-time and there were a different computing times for every test. When the choice was (256*256) with [8 8]cell or (128*128) with [4 4]cell the computing time was triple times than that when the choice was (128*128) with [8 8]cell. That means it will also be effected in real-time for any image to be selected for recognition and will cause a slow system. So the choosing form was (128*128) pixels and [8 8]cell represent the best.

Gesture to text results

With the [HOG] system some sign were always get a true recognition they were [a, c, 5] to give the sound orders good morning Alexa, good night Alexa, and Play alarm. But with other signs like [b, v,..] the results sometime be false. Elapsed time is 4.43seconds from capturing the image until to give the sound order sentence, Figure(11) shows the true case for all images in complex backgrounds.

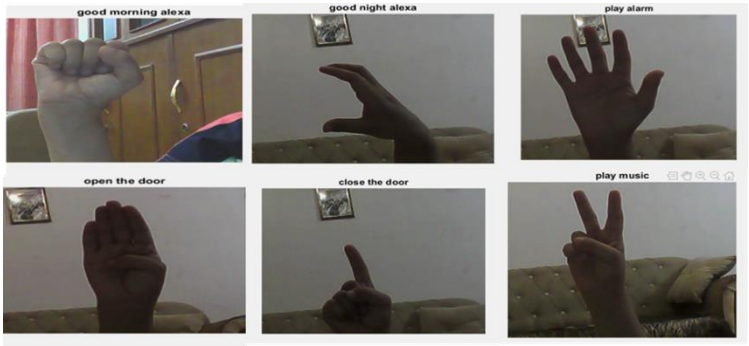

Figure 5 The true cases of recognition for all signs with complex background

With black background there were a true recognition also as shown in Figure(12).

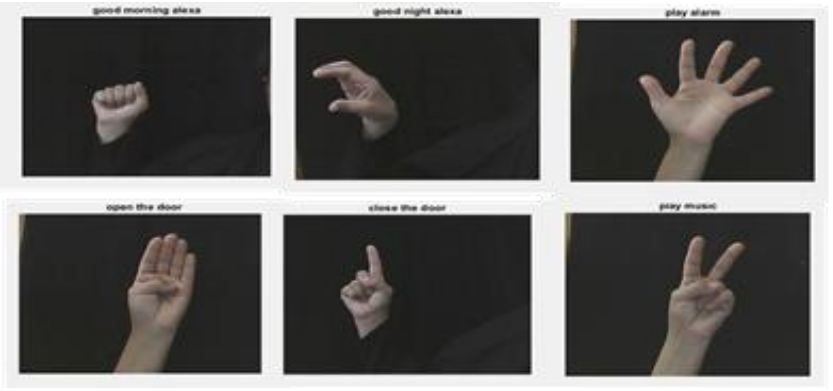

Figure 12: The true cases of recognition for all gestures with black

But these results were not the only ones, as there were some results that were wrongly predicted, especially with (b, point, v). the wrong result appeared when the test repeated for many time and different background or when rotated the hand gesture in different direction such as those in Figure (13). These wrong results do not mean the failure of the system's performance in full, as it would not have appeared until after sporadic repetition attempts, but in this work, there is a trial to obtain infinite accuracy in performance because we are trying to adopt the system in smart environments. Those environments may be educational institutions, health care homes, or hospitals dealing with critical or dangerous cases. So the probability of error cannot be accepted, even if it is 10%.

Instability in results may be treated by increasing the images in the dataset more than that we chose, but we will not consider these to be the ideal solution. Although this system gave a high accuracy in off-time, we are really disappointed in real-time, and this result may help researchers in the future not to rely on the results that they will obtain from the validation data.

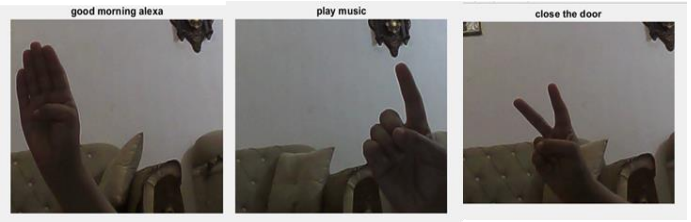

Figure 6 The false case of recognition for some images with complex background

Results for Deep learning using a (CNN) system

In this test the results in the off-time are almost the same as in the real-time phase so to prove the accuracy of the results, there were many tests done with:

- ✓ different background (complex & black).
- ✓ different persons (child & male & female).
- ✓ different distance from the camera (0.75 & 2)meter.

Figure(14) represents multiple attempts from a child to represent some gestures, and as it is clear, despite her inability to represent the signs completely correctly, our system was always able to correctly predict.

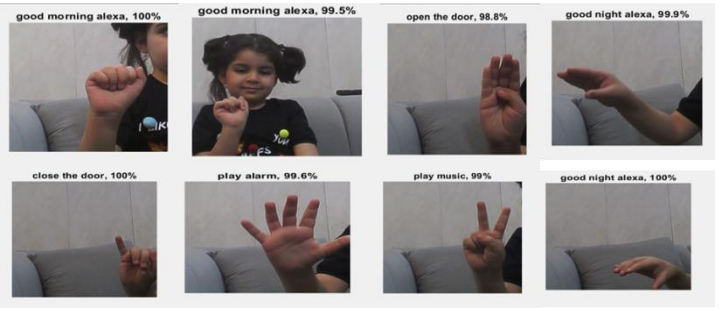

Figure 7 Testing for child gestures

For more complex background and (male & female) hand gestures Figures(15), and (16) a true and high accuracy even when rotating the gesture or when the face is included with the image which gives similar skin colour between the hand and the face, so this makes distortion in the image data for the system and makes the prediction step more difficult. The selected backgrounds have colour details that are close to skin tone also in an experimental attempt by us to ensure the correctness of the results.

Figure 8 Testing for female gestures with complex background

Figure 9 Testing for male gestures with complex background

In the Figures(17),(18), and (19) there are an attempt to experiment with black background gestures that were not included in the dataset images when training the machine. Also here the face is included sometimes and the distortion because of making a movement in the hand at capturing the image, but the result still true always.

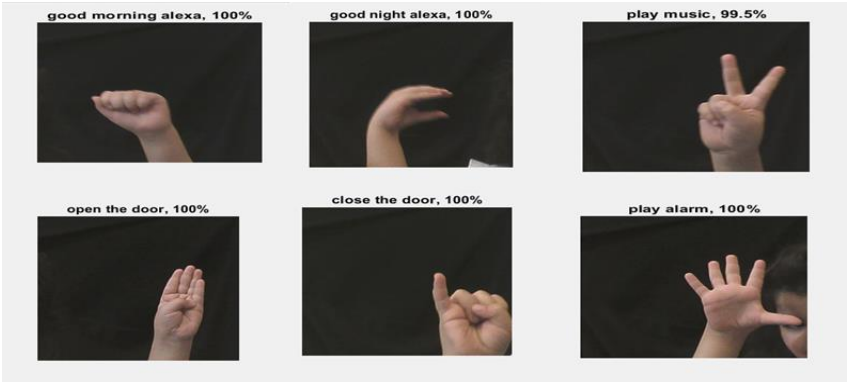

Figure 10 Testing for child gestures with black background

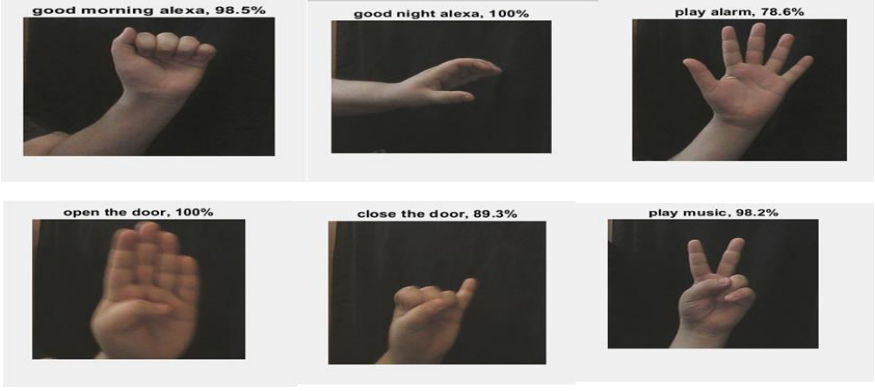

Figure 11 Testing for male gestures with black background

Figure 12 Testing for female gestures with black background

From the observation of Figures (20),(21), and (22) we find that it is better to make a condition for good prediction since we tried to predict by the experiment for gestures that are about two meters away from the camera, and what was obtained were incorrect prediction results in some of them, as shown. Therefore, the condition that we will set for the practical use of our system is that the location be at a distance of 75cm or 1m from the camera, as a maximum, to ensure a correct and guaranteed prediction.

In fact, the incorrectness of the results here does not weaken the confidence in the system, because we basically did not put images with a distance of 2 meters in the dataset, and because this case is not considered the ideal case for dealing with such systems. It is better to make the response of the system only to the signers close to the camera, because it may be run in public environments where it is possible to pass by those who do not need to use the system and any sign from them may cause in turn-on the system without any need.

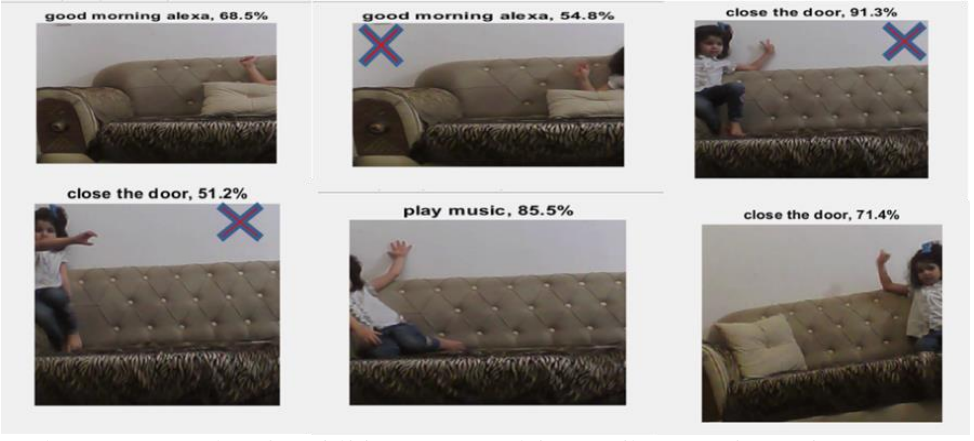

Figure 13 Testing for child gestures with 2m distance from the camera

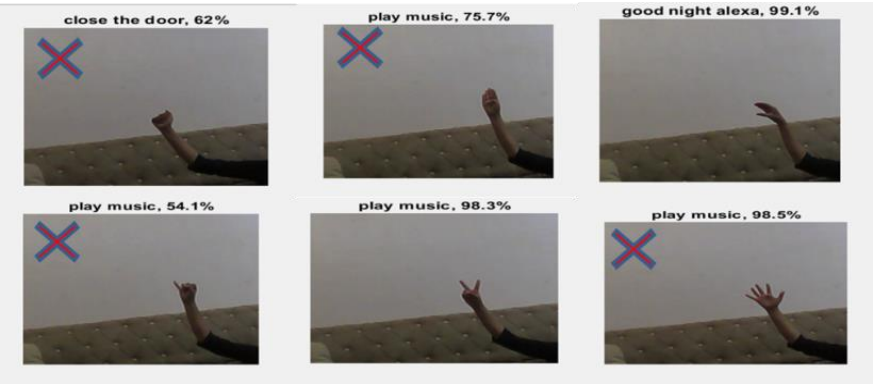

Figure 14 Testing for female gestures with 2m distance from the camera

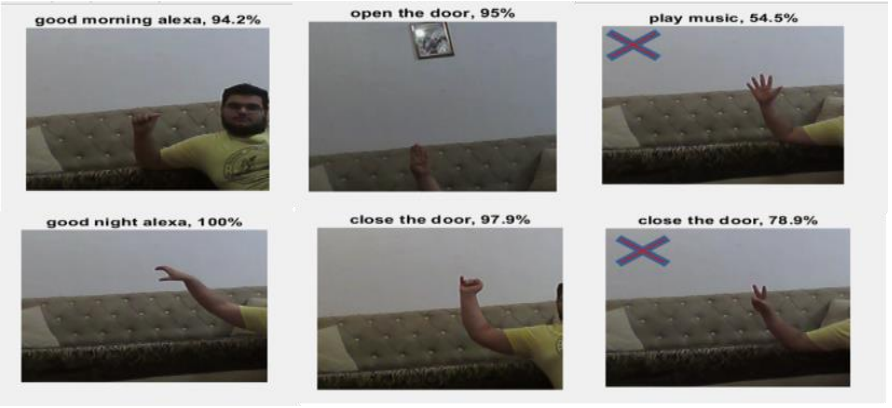

Figure 15: Testing for male gestures with 2m distance from the camera

For making a comparison with previous studies that used Squeezenet, author in (Kasukurthi et al., 2019) used it but in off-time for the surrey finger dataset and validation accuracy is 83.29%. This result as compared with ours which was shown in tuble.2 as a summary for the most frequent results :

Table 2 a summary for the accuracy results by the proposed system

	Child		Male		Female	
	Complex	black	Complex	black	Complex	black
A	100	100	99.9	98.5	9302	100
B	98.8	100	97.4	100	100	93.2
C	100	100	100	100	99.9	96.3
V	99	99.5	99.9	98.2	100	100
Five	99.6	100	87	78.2	100	99.8
Point	100	100	90	89.3	99.9	90.8
Accuracy rate	99.57	99.92	95.7	94.22	98.9	96.72

From the above table the accuracy of the system will be:

- ✓ For complex background: 98.1%
- ✓ For black background: 97%
- ✓ For over all cases: 97.5%

And the competition time about 3.3sec for capturing the image, predicting it, and converting to a spoken sentence for every image

Conclusion

Squeezenet which was one of the convolution neural network (CNN) models, has proven highly effective in classification and prediction for more than one shape in real-time, with different conditions, for different locations, and in multiple light environments for the static hand. The most interesting thing about our system is that it requires no preprocessing for the image before classification or prediction. Squeezenet has a good characteristic in achieving more fewer parameters which will fit into computer memory and transmitted over a computer network more easily too.

The results that were obtained using the HOG in real-time were not the same as they are in the off-time, which calls for uncertainty about the results achieved by the systems in the experiments in the off-time Also, converting the sign gesture into a voice prompted the implementation of this in the smart home system by communicating with the Alexa device, which provides an excellent service for the deaf-mute community and makes artificial intelligence services within the reach of their limited capabilities. Our future work is planning to make a prediction system to detect the dynamic gestures, which represents complete sentences for the deaf mute and convert them into spoken words.

References

Adeyanju, I. A., Bello, O. O., & Adegboye, M. A. (2021). Machine learning methods for sign language recognition: A critical review and analysis. *Intelligent Systems with Applications*, *12*, 200056. https://doi.org/10.1016/j.iswa.2021.200056

Ahmad, S. U. D., & Akhter, S. (2013). Real time rotation invariant static hand gesture recognition using an orientation based hash code. *2013 International Conference on Informatics, Electronics and Vision, ICIEV 2013*. https://doi.org/10.1109/ICIEV.2013.6572620

Ahmed, M. A., Zaidan, B. B., Zaidan, A. A., Salih, M. M., & Lakulu, M. M. Bin. (2018). A review on systems-based sensory gloves for sign language recognition state of the art between 2007 and 2017. *Sensors (Switzerland)*, *18*(7). https://doi.org/10.3390/s18072208

Barnouti, N. H. (2016). *Improve Face Recognition Rate Using Different Image Pre-Processing Techniques Nawaf Hazim Barnouti American Journal of Engineering Research (AJER). 4*, 46–53.

Barra Novoa, R., Limari, K., & Limari, P. (2021). Current overview of assistance bioethics committees in Chile. *International Journal of Health & Medical Sciences*, *4*(1), 95-101. https://doi.org/10.31295/ijhms.v4n1.1492

Bohra, T., Sompura, S., Parekh, K., & Raut, P. (2019). Real-Time Two Way

Communication System for Speech and Hearing Impaired Using Computer Vision and Deep Learning. *Proceedings of the 2nd International Conference on Smart Systems and Inventive Technology, ICSSIT 2019*, *Icssit*, 734–739. https://doi.org/10.1109/ICSSIT46314.2019.8987908

Dardas, N. H., & Georganas, N. D. (2011). *Real-Time Hand Gesture Detection and Recognition.pdf*. *60*(11), 3592–3607.

Ekbote, J., & Joshi, M. (2016). *A Survey on Hand Gesture Recognition for Indian Sign Language*. 1039–1044.

Figueira-Domínguez, J. G., Bolón-Canedo, V., & Remeseiro, B. (2020). *Feature Selection in Big Image Datasets*. 40. https://doi.org/10.3390/proceedings2020054040

Hameed, N., Shabut, A., Hameed, F., Cirstea, S., Harriet, S., & Hossain, A. (2020). Mobile-based skin lesions classification using convolution neural network. *Annals of Emerging Technologies in Computing*, *4*(2), 26–37. https://doi.org/10.33166/AETiC.2020.02.003

Haque, P., Das, B., & Kaspy, N. N. (2019). Two-Handed Bangla Sign Language Recognition Using Principal Component Analysis (PCA) and KNN Algorithm. *2nd International Conference on Electrical, Computer and Communication Engineering, ECCE 2019*, 1–4. https://doi.org/10.1109/ECACE.2019.8679185

Hurroo, M., & Elham, M. (2020). *Sign Language Recognition System using Convolutional Neural Network and Computer Vision*. *9*(12), 59–64.

Islam, M. M., Siddiqua, S., & Afnan, J. (2017). Real time Hand Gesture Recognition using different algorithms based on American Sign Language. *2017 IEEE International Conference on Imaging, Vision and Pattern Recognition, IcIVPR 2017*. https://doi.org/10.1109/ICIVPR.2017.7890854

Islam, M. R., Mitu, U. K., Bhuiyan, R. A., & Shin, J. (2018). Hand gesture feature extraction using deep convolutional neural network for recognizing American sign language. *2018 4th International Conference on Frontiers of Signal Processing, ICFSP 2018*, *September*, 115–119. https://doi.org/10.1109/ICFSP.2018.8552044

Jain, S., & Salau, A. O. (2019). An image feature selection approach for dimensionality reduction based on kNN and SVM for AkT proteins. *Cogent Engineering*, *6*(1). https://doi.org/10.1080/23311916.2019.1599537

Jaramillo-Yánez, A., Benalcázar, M. E., & Mena-Maldonado, E. (2020). Real-time hand gesture recognition using surface electromyography and machine learning: A systematic literature review. *Sensors (Switzerland)*, *20*(9), 1–36. https://doi.org/10.3390/s20092467

Kadhim, R. A., & Khamees, M. (2020). A real-time american sign language recognition system using convolutional neural network for real datasets. *TEM Journal*, *9*(3), 937–943. https://doi.org/10.18421/TEM93-14

Kasukurthi, N., Rokad, B., Bidani, S., & Dennisan, D. A. (2019). *American Sign Language Alphabet Recognition using Deep Learning*. http://arxiv.org/abs/1905.05487

Kumar, D., Jain, A., & Kumar, M. (2022). *Interaction with IoT Comfort technologies for Deaf and Dumb People*. 503–510. https://doi.org/10.1109/iceca52323.2021.9676069

Kumar, G., & Bhatia, P. K. (2014). A detailed review of feature extraction in image processing systems. *International Conference on Advanced Computing and Communication Technologies, ACCT*, 5–12. https://doi.org/10.1109/ACCT.2014.74

Lahoti, S., Kayal, S., Kumbhare, S., Suradkar, I., & Pawar, V. (2018). Android Based American Sign Language Recognition System with Skin Segmentation and SVM. *2018 9th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2018*, 1–6. https://doi.org/10.1109/ICCCNT.2018.8493838

Lopatovska, I., Rink, K., Knight, I., Raines, K., Cosenza, K., Williams, H., Sorsche, P., Hirsch, D., Li, Q., & Martinez, A. (2019). Talk to me: Exploring user interactions with the Amazon Alexa. *Journal of Librarianship and Information Science*, *51*(4), 984–997. https://doi.org/10.1177/0961000618759414

Mahmud, I., Tabassum, T., Uddin, M. P., Ali, E., Nitu, A. M., & Afjal, M. I. (2019). Efficient Noise Reduction and HOG Feature Extraction for Sign Language Recognition. *2018 International Conference on Advancement in Electrical and Electronic Engineering, ICAEEE 2018*, *November*, 1–4. https://doi.org/10.1109/ICAEEE.2018.8642983

Mapari, R. B. (2014). *Real Time Sign Language Translator*. 1–4.

Mujahid, A., Awan, M. J., Yasin, A., Mohammed, M. A., Damaševičius, R., Maskeliūnas, R., & Abdulkareem, K. H. (2021). Real-time hand gesture recognition based on deep learning YOLOv3 model. *Applied Sciences (Switzerland)*, *11*(9). https://doi.org/10.3390/app11094164

Oudah, M., Al-Naji, A., & Chahl, J. (2020). Hand Gesture Recognition Based on Computer Vision: A Review of Techniques. *Journal of Imaging*, *6*(8). https://doi.org/10.3390/JIMAGING6080073

Pansare, Jayashree R., Gawande, S. H., & Ingle, M. (2012). Real-Time Static Hand Gesture Recognition for American Sign Language (ASL) in Complex Background. *Journal of Signal and Information Processing*, *03*(03), 364–367. https://doi.org/10.4236/jsip.2012.33047

Pansare, Jayshree R., & Ingle, M. (2016). Vision-based approach for American Sign Language recognition using Edge Orientation Histogram. *2016 International Conference on Image, Vision and Computing, ICIVC 2016*, 86–90. https://doi.org/10.1109/ICIVC.2016.7571278

Rahaman, M. A., Jasim, M., Ali, M. H., & Hasanuzzaman, M. (2003). Real-time computer vision-based Bengali sign language recognition. *2014 17th International Conference on Computer and Information Technology, ICCIT 2014*, *May 2019*, 192–197. https://doi.org/10.1109/ICCITechn.2014.7073150

Roffo, G. (2016). *Feature Selection Library (MATLAB Toolbox)*. http://arxiv.org/abs/1607.01327

Shin, J., Matsuoka, A., Hasan, M. A. M., & Srizon, A. Y. (2021). American sign language alphabet recognition by extracting feature from hand pose estimation. *Sensors*, *21*(17), 1–19. https://doi.org/10.3390/s21175856

Shirbahadurkar, S. D., & Bormane, D. S. (2009). Marathi language speech synthesizer using concatenative synthesis strategy (spoken in Maharashtra, India). *2009 2nd International Conference on Machine Vision, ICMV 2009*, 181–185. https://doi.org/10.1109/ICMV.2009.52

Suryawanshi, S., Itkarkar, R., & Mane, D. (2014). High quality text to speech synthesizer using phonetic integration. *International Journal Of Advanced Research In Electronics And Communication Engineering*, *3*(2), 77–82.

Taskiran, M., Killioglu, M., & Kahraman, N. (2018). A Real-Time System for Recognition of American Sign Language by using Deep Learning. *2018 41st International Conference on Telecommunications and Signal Processing, TSP 2018*, 1–5. https://doi.org/10.1109/TSP.2018.8441304

Van Culver, R. (2004). A hybrid sign language recognition system. *Proceedings - International Symposium on Wearable Computers, ISWC*, 30–33. https://doi.org/10.1109/iswc.2004.2

Wan, R., Vegas, L., Carlo, M., & Ii, P. (2007). Computational Methods of Feature Selection. *Computational Methods of Feature Selection*. https://doi.org/10.1201/9781584888796

Widana, I.K., Sumetri, N.W., Sutapa, I.K., Suryasa, W. (2021). Anthropometric measures for better cardiovascular and musculoskeletal health. *Computer Applications in Engineering Education*, *29*(3), 550–561. https://doi.org/10.1002/cae.22202