International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 01 | Jan 2022 www.irjet.net p-ISSN: 2395-0072

Sign Language Recognition using Machine Learning

Ms. Ruchika Gaidhani[1], Ms. Payal Pagariya[2], Ms. Aashlesha Patil[3], Ms. Tejaswini Phad[4]

[5]***Mr. Dhiraj Birari***, *Dept. of Information Technology, MVPs KBT College of Engineering, Maharashtra, India*

Abstract - *Our goal is to develop a model that can detect hand movements and signs. We'll train a simple gesture detecting model for sign language conversion, which will allow people to communicate with persons who are deaf and mentally challenged. This project can be performed using a variety of methods, including KNN, Logistic Regression, Nave Bayes Classification, Support Vector Machine, and CNN. The method we have chosen is CNN because it has a higher level of accuracy than other methods. A computer program written in the programming language Python is used for model training based on the CNN system. By comparing the input with a pre-existing dataset created using Indian sign language, the algorithm will be able to understand hand gestures. Users will be able to recognize the signs offered by converting Sign Language into text as an output. by a sign language interpreter This approach is implemented in Jupyter Lab, which is an add-on to the Anaconda documentation platform. To improve even more, we'll convert the inputs to black and white and accept input from the camera after applying the Background subtraction approach. Because the mask is configured to identify human skin, this model doesn't need a simple background to work and can be constructed with just a camera and a computer.*

Key Words: ISL, Data set, Convolutional Neural Network, Accuracy, Haar cascade.

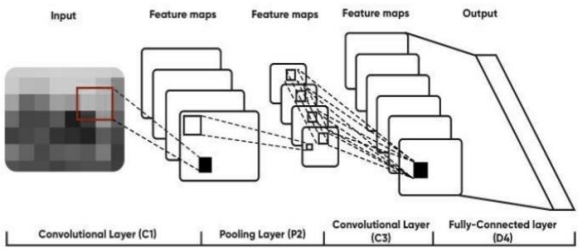
1.INTRODUCTION

In today's world, we have computers that operate at extremely high rates, performing massive amounts of calculations in a fraction of a second. We humans are now attempting to accomplish a goal in which the computer will begin to think and work like a human person. This necessitates the most fundamental quality, 'learning.' This brings us to the topic of artificial intelligence. AI states that a computer may start or begin executing activities on its own, without the need for human interaction. To achieve this, a computer must first learn how to react to inputs and variables. The computer should be taught with a large amount of data, the data required to train the computer is determined by the desired output and the machine's operation. We create a computer model that can identify human hand motions; there are numerous apps that function with hand gestures that we observe in our daily lives. Look at the console in our living room; connect it to a sensor, and we'll be able to play tennis with our hands. We have created a touch detection model that translates sign language into speech. There are a number of devices that rely on touch detection, whether for security or entertainment. Sign language is a vision-based language that uses a combination of material, body language, and gestures, fingers, and orientation, posture, and hand and body movements, as well as eyes, lips, and wholeness. facial expressions and speech. A variety of signature circuits exist, just as there are regional variations in spoken language. Through gestures and gestures, facial expressions, and lip movements, we can spell the letters of each word with our fingers and maintain a certain vocabulary of Indian Sign Language (ISL), American Sign Language (ASL), and Portuguese Signature (PSL). Sign language can be separated or used continuously. People interact with a single sign language by performing a single word movement, while continuous sign language can be a series of steps that form a coherent sentence. All methods of identifying hand movements are often classified as based on vision and based on measurements taken by sensors embedded in gloves. This vision-based process uses human computer interactions to detect touch with bare hands. OpenCV is used to detect sign languages in this project. Uses a webcam to detect user touch; Our model can distinguish handmade touches with bare hands, so we will not need gloves for this project. OpenCV is used to detect sign languages in this project. Uses a webcam to detect the movement of a user's hand, with words displayed on the screen as output. Our project aims to help people who do not know sign language well by identifying and converting man-made symbols into legible characters. Machine learning, especially CNN, can help us achieve this. We want to use an image from the web camera/phone camera to train a model that can predict text (using IP camera software and OpenCV). Because the model was trained using a well-known dataset (ASL), there will be enough data for the training algorithm to produce a precise and accurate model. Because the model's code is written in Python, the project can be completed on a simple computer without the use of high-end processing units or GPUs. The software Jypter Lab, where this model is trained and implemented, is built on the Anaconda Documentation platform. The many concepts involved, as well as the technique for carrying out this project, are addressed in detail in the following sections of the paper.

1.1 Sign Language Recognition

Without this notion, our entire model would be impossible to materialize. Deep neural networks are one of the classes of deep neural networks, and CNN is one of them. CNN is used in a variety of fields, the majority of which include visual images. A CNN is made up of many neurons whose values, or weights, can be changed to achieve the desired result. These biases and weights can be learned. Non-linearity is caused by the dot products made between the neurons. The key distinction between regular neural networks and convolutional neural networks is CNN's

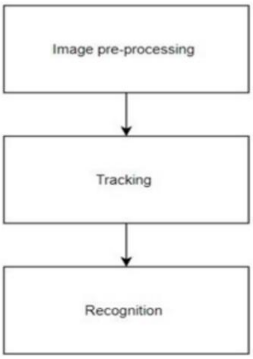
assumption that all of the inputs are explicit pictures. As a result, given our project relies around photographs, this will be the optimal method for training the model. As a result, CNN will be able to benefit from the architecture being bound in a meaningful way with images as input explicitly; a neuron layout is done in three dimensions: width, depth, and height. The volume required for activation is referred to as depth. A convolutional network,



as shown in Figure, consists of a number of layers, and the volume that exists is translated to another form using a function (differentiable). The pooling layer, convolutional layer, and fully-connected layer are the layers that make up CNN architecture. A convolutional network architecture is created by stacking these in the correct order. Now that these layers have extracted features from the image that are unique to a property, the loss function must be minimized. With this formula. It can be done in such a way that the sample's positive classes are marked by M. The CNN score for each positive class is denoted by confusion matrix, and the scaling factor is denoted by 1/M. Backdrop Subtraction- This refers to the process of removing the background from a photograph. The foreground and background parts are separated using this technique. This is accomplished with the use of a mask that is created according to the user's preferences. This method is used to detect through static cameras. Background subtraction is critical for object tracking, and it can be accomplished in a variety of methods.

1.2 How does the image get recognized by computer?

There is a lot of information all around us, and our eyes selectively pick it up, which is different for everyone depending on their preferences. Machines, on the other hand, see everything and absorb every image before converting the data into 1s and 0s that the computer can comprehend. How does this transformation happen? Pixel. The smallest unit of a digital image that may be displayed or projected on a display device is the pixel. The image has multiple intensity levels at various points, which are represented by numbers; for our photographs, we have shown values consisting of only one value (grayscale), which is assigned automatically by the computer based on the strength of the darkness or level. Grayscale and RGB are the two most used methods for identifying images. Imagine a black and white image in grayscale; these images have only two colours. The colour black is considered to be or is used as a measurement for the weakest intensity, with white being the brightest. The computer assigns the values based on the darkness levels. RGG stands for red, green, and blue in RGB. When all of these colours are combined, they form a colour. Only these three colours can be used to define any colour on the planet. Each pixel is checked and a value is extracted by the computer.



1.3 Proposed Methodology

Figure 2 shows the steps we took to create a model that uses CNN to predict the text of a hand gesture/symbol. We also utilise strategies like background subtraction (see Figure 3) to increase the model's performance by removing the background light/ambience.

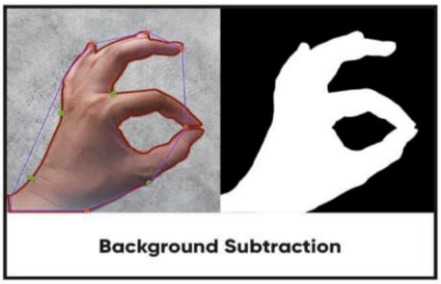


Figure 3

Three components are required to construct a Sign Language Recognition system:

1) The data set depicted in Figure 4

2) Create a model (In this case we will use a CNN)

3) A platform on which to implement our model (We are going to use OpenCV).

Figure 4

Proper hardware is necessary to train a deep neural network (e.g., a powerful GPU). This project does not necessitate the use of a strong GPU. However, using online tools like Google Collab is still the best method to go about it. The dataset will be modified from the National Institute of Standards and Technology. Our collection contains several photos of the American Sign Language alphabet (excluding J and Z), with a total of 784 pixels per image, resulting in a 28x28 image size.

Our data is in CSV (Comma-separated values) format, with the pixel values stored in train X and test Y. The image label is stored in the train Y and test Y variables. Moving on to pre-processing, the trained x and y both include an array of all pixel values. These values will be used to build a picture. We divide the array into 28x28 pixel sections because our image is 28x28 pixels. This dataset is used to train our model. To recognize the alphabets, we utilise CNN (Convolutional Neural Network). Keras is what we utilise. Our model, like every other CNN, starts with a few layers like Conv2D and MaxPooling, which are followed by fully linked layers. The initial Conv2D layer gets the shape of the input image, and the final completely linked layer gives us 26 alphabets as output. To make our training more consistent, we use a Dropout after the 2nd Conv2D layer.

In the last layer, Soft Max is employed as the activation function, which gives us the probability for each alphabet as an output. Table 1 shows the proposed model.

Layer (Type)	Output Shape	Param#
Conv2d_1(Conv2D)	(None,28,8,8)	80
Max_pooling_1(MaxPooling2)	(None,14,14,8)	0
Conv2d_2(Conv2D)	(None,14,14,16)	1168
Dropout_1(Dropout)	(None,14,14,16)	0
Max_pooling2d_2(MaxPooling2)	(None,3,3,16)	0
Dense_1(dense)	(None,3,3,128)	2176
Flatten_1(Flatten)	(None,1152)	0
Dense_2(dense)	(none,26)	29978

2. Emotion Recognition

Facial expressions play an important role in communication. It is a basic communication bridge, overcoming this barrier to any visually based system with sign language and hand gestures alone does not get effective results. In everyday life facial expressions are an important part of meaningless communication. Face-to-face contact is therefore widely accepted. Coding and emotion recording is an important aspect of facial expression. There are six basic emotions which are anger, disgust, fear, joy, sadness, and surprise.

Thus, this study approaches primarily the Haar Cascade approach and the in-depth learning approach using the Convolution Neural Network. Face detection using a Haar-based Cascade separator is an effective way to obtain an object [5]. Deep neural network has the same type as CNN in high network depth and algorithm process. CNN's basic concept is almost identical to that of the Multilayer Perceptron, however, each neuron at CNN will be activated using a two-dimensional structure. CNN can only be used for image data and voice recorder with a two-dimensional structure. CNN has many learning categories that are used to learn input element based on input feature. Layers are in the form of price vectors. This layer extraction feature contains a convolution layer and a composite layer. In the convolution layer, the outflow of neurons will be calculated, each calculating its own weight and until small-sized images are connected to the input volume.

2.1 CNN for facial expression detecting

The CNN architecture consists of 6 convolutional layers, 2 sub-sample layers, 12 convolution layers, and a Neural Network of 2 sub-samples.

2.2. Haar Cascade Classifier

Obtaining an object using a Haar-based cascade classifier is a way to find an object by Paula Viola and Michael Jones. In 2001, they proposed the paper title "Fast Object Recovery using the Boosting Cascade for Simple Features". Haar cascade is a set of Haar-like Features that are integrated to form a separator. Feature is the number of pixels in writing taken from the number of pixels in an empty space. The base of the face detector is 24 x 24. From that basic face detector, where about 160k is possible Haar-Like Feature. However, not all of these features are used

2.3 Experimental Results

We conducted experiments using a training dataset and testing data set in real-time by a video camera. we use varied epochs and get a different result in MSE and Accuracy. From the experiments we can say that there is significant decreasing means square error as the epoch of training data raises.

Epoch	Num of Training Data	Num of Testing Data	MSE	Model Accuracy
30	28,709	70	0,8652	67%
50	28,709	70	0,6754	75%
75	28,709	70	0,5214	81%
100	28,709	70	0,4192	85%
150	28,709	70	0,3356	89%
200	28,709	70	0,2912	92%

The table presents the test result, using a variety of periods in training and testing the data. The results suggest that the higher the epoch value and therefore the MSE value gets lower value. Similarly, the accuracy of the model was very accurate when the period was high. These activities show that CNN's approach is good for image testing and training.

Table 2. Facial Emotion Detection Classification

Epoch	Angry	Happy	Disgust	Sad	Neutral	Surprise	Fear
30	60%	70%	50%	60%	60%	60%	60%
50	70%	70%	70%	70%	70%	70%	70%
75	70%	80%	70%	70%	70%	70%	70%
100	80%	90%	80%	70%	70%	80%	70%
150	80%	100%	80%	80%	80%	80%	80%
200	90%	100%	80%	80%	90%	90%	80%

Table- Facial emotion detection Classification

By measuring the facial emotion in seven classes, the accuracy rates for each epoch varies. The problem when we test using video where the facial expression model is still cannot distinguish between fear and sad expression, happy and surprise expression.

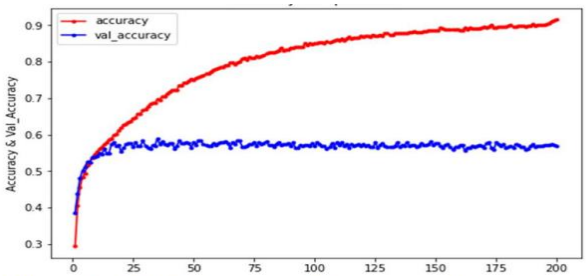


Figure 4. Model Accuracy when running the program

In the graph the accuracy of the model facial sensor detection increased the value increasing from the beginning of the process, i.e. the minimum MSE value, in order to obtain the maximum accuracy of the model.

3. CONCLUSIONS

We have developed the structure of the Convolutional Neural Network (CNN) that a large amount of epoch, appropriately a square measure error has a small value, however, the accuracy of the model will be increased for Face Sign Language Recognition. There are seven categories of facial expressions, we assess real-time data using the Haar - Cascade Classifier. we train the model and create the perception

REFERENCES

[1]. Dalal N, Triggs B. Histograms of Oriented Gradients for Human Detection. IEEE Xplore Digital Library.

[2]. Dalal N. Lecture notes on Histogram of Oriented Gradients (HOG) for Object Detection. Joint work with Triggs B, Schmid C.

[3]. Deniz O, Bueno G, Salido J and Torre F. De la. Face recognition using Histograms of Oriented Gradients, Pattern Recognition. Letters 32 (2011) 15981603

[4]. Jurafsky D, Martin J H. Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition and Computational Linguistics. 2nd edition, Prentice-Hall, 2009.