Sign Language Detection and Recognition

Gokul Kumar K
Department of Computer Science and Engineering
SRM Institute of Science and Technology, Vadapalani
Chennnai, India

I Imran Mohammed
Department of Computer Science and Engineeting
SRM Institute of Science and Technology, Vadapalani
Chennai, India

Soni Jatin Mahendra
Department of Computer Science and Engineering
SRM Institute of Science and Technology, Vadapalani Chennai, India

S Niveditha
Department of Computer Science and Engineering
SRM Institute of Science and Technology, Vadapalani
Chennai, India.

Abstract— Hand gesture/sign is one of the communication methods of non-verbal communication. The collection of these hand gesture forms a sign language. Sign Language is popularly used by deaf-mute people for communicating among themselves and also to other people. It has always been a great challenge for normal people to communicate with the deafmute people as it is tough for them to understand their gestures. Hence the key idea of the paper is to bridge the communication gap between the normal people and the deafmute people. Various sign language systems have been developed, but the systems are neither flexible nor are they cost effective. Hence this paper proposes an effective and user friendly hand gesture recognition system which helps the deafmute people to communicate with normal people easily. In the present developing technology, Gesture Recognition and Pattern Recognition has been the field of research. Hand gesture being an important mode of communication in nonverbal communication it plays a vital role in daily life. The proposed paper provides a user-friendly way of communication by using the CNN algorithm.

Keywords—CNN, Sign Language, Gesture Recognition, OpenCV, ROI, Relu, Silhouette, Pooling, Histogram.

I. INTRODUCTION

Sign Language Recognition is one of the popular topics of research in the emerging trend. Sign Language Recognition projects help to fill the communication gap between the specially-abled (deaf-mute) people and the normal people. Usually, deaf-mute people feel difficult to express their feelings to others as the normal feel tough to understand their gesture language. Hence, the opportunities for the deaf-mute people are low. This can be avoided by creating a platform that can communicate between the deafmute people and normal people by recognising the specially-abled people's gesture and converting it to text/voice and vice versa. Hence, in this paper we aim to initialise this long term project by successfully completing until formation of words through the letters recognised out of the gestures of deaf-mute people. We were also successful in recognising certain words that have their own gestures. This proves to be a one big successful step for creation of the Sign Language Recognition platform. To make it effective economically and qualitatively, we have used simple technologies such as Python, Tensorflow, Keras and OpenCV that are available easily. Hence, this paper aims to provide a user-friendly and efficient platform for Sign Language Recognition.

II. ARCHITECTURE DIAGRAM

The architecture diagram mentioned below is followed for the proposed project. Initially in the input image the hand skin tone is recognised and the Skin colour segmentation is carried out. This is the process of converting the coloured image to the black white format where the hand is alone recognised. This is carried by using the Region of Interest Extraction (ROI) where the other parts of image are ignored and only the skin colour in the image is highlighted as white in colour. So now we get image of the hand gesture after ROI is performed. Then the largest connected component that is a silhouette image of the hand is extracted from the image we get after ROI is performed.

After we get the silhouette image, feature extraction is carried out providing the gesture input that is made by the hand to be recognised and compared to the already available dataset which is sent to the model to train the system and classify the gesture accurately. Then the gesture input is classified and the gesture to which it is matching maximum is recognised and given as output. This process is done using the CNN which sends the input gesture to all its four layers where the input undergoes the shrinking, extracting and classifying functions of the layers respectively.

This is further developed to form words by combining the letters that are recognised. This allows us to move a step forward in establishing a full-fledged Sign Language Recognising project that could make communication easy between the specially abled people who are deaf or dumb and the normal people. This project is done using python platform with the use of Tensorflow, Keras and OpenCV which are one of the most efficient technologies available to work on image processing and classification.

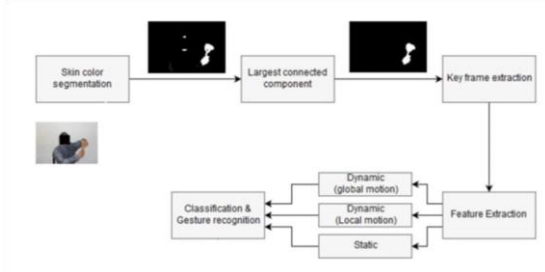

Fig. 1: Architecture Diagram of Proposed System

III. TECHNICAL MODULES

A. Setting Histogram

Setting histogram is one of the important modules of Sign Language Recognition. In context to image processing, the histogram shows how various values of colors occur in the image that can be represented graphically. This is important in order to train the system about your hand and lighting conditions. This module has to be done whenever a new person is going to show gestures or the lighting conditions of the place where the project is executed changes. In the proposed system, setting the histogram is achieved by a window which contains 50 squares in the form of 10 rows and 5 columns. You will have to place your palm in the squares so that the system recognizes your hand.

Fig. 2: Setting Histogram

B. Converting to Black and White

As soon the histogram is set, it is important for us to convert the images to black and white form which facilitates the system to work easily. Hence, Region of Interest Extraction (ROI) algorithm is used to extract the largely connected areas of your hand and highlighting it in white. The background and the unwanted components around the hand are ignored making them black. The output of this module will be a clear black and white silhouette image of the hand which can be used for further processes.

Fig. 3: Black and White Silhouette Image

C. Gesture Recognition

Once the system is set with the histogram and trained with your hand in black and white form, the next step is to start recognising the gestures. For this we have to first provide the system with dataset which can be used by the model to train itself. In this system we have used around 44 gestures where each gesture has 200 images as training data for the model to get trained. Now, when you make a gesture, the particular frame is taken which is converted to black and white silhouette image. This image is compared with the dataset given and then classified by recognising the word/letter to which it matches maximum. This classification is achieved with the help of Convolution Neural Network (CNN) algorithm. Hence, the letters/words are recognised successfully.

D. Letter to word

Certain words in sign language do not have dedicated gesture itself. So, it is conveyed by the deaf-mute people by spelling them out of each letter. Considering, this in mind, it is important to stack up the letters to form a word. The proposed system has this facility as well. In this module, it is important that the person holds the particular letter gesture for continuous 3 frames. Once recognised to stay in 3 continuous frames, it is stacked on behind the other to form a word. This method also allows us to move a step ahead in achieving the proper Sign Language Recognition.

IV. ALGORITHMS

A. Region of interest extraction

It is known as ROI algorithm, which is used for extracting palm region of the input video irrespective of any sleeve type or wristband. This algorithm is applied only if the height of bounding box is larger than a fixed height H. Otherwise it is assumed to be in full sleeve.

Understanding how a computer reads an image: Basically a computer has three channels through which it can read an coloured image which are namely Red, Green, and Blue which are popularly known as RGB channels. So each of these channels have their own respective pixel values. When a computer gets the input video, videos are sequence of frames and frames are actually images that are collection of pixels. In our work above the American Sign Language gestures are recognised accurately only after 15 frames.

As in our case we are dealing with white and black images (greyscale images) so only two channels are being used. The darkness of the pixel is identified as a value ranging from 0 to 1. The darkest part is identified as 1 whereas completely white part is identified as 0. Often simple images have values between 0 and 255. Where 255 is known the highest or brightest value and zero is known as the darkest. We chose the range 0 to 255 because this is how the computer is storing 8-bit numbers. If we want to get the values between 0 and 255 to be ranging between 0 and 1 all we need to do is divide the maximum value to 5 to normalize it to 0 and 1. This is where ROI comes in the picture. After we set histogram for our hand gesture the ROI extracts the palm region of the input video and helps for further gesture recognition.

B. Convolutional Neural Networks (CNN)

Convolutional Neural Networks are deep neural networks or special type of feed forward artificial neural network used to process data that have a grid-like topology, for example images that can be represented as a 2-D array of pixels. A CNN model is divided into four main layers: Convolution Layer, Non-Linearity (Relu Layer), Pooling Layer and Fullyconnected Layer (Classification).

C. Convolution Layer

The main function of convolution layer is to extract features from the input image. It matches the input image features by learning image features using small squares of input data and thus maintaining the spatial relationship between pixels. Usually it is followed by Relu layer.

D. Relu Layer

The function of this layer is to replace all negative pixel values in the feature map by zero due to which it is also known as an element-wise operation. Introducing nonlinearity in a convolution network is the main function of this layer.

E. Pooling Layer

The main function of this layer is to shrink the size of the input image. It is also called down sampling, which reduces the dimensionality of each feature map but does not change the important data.

F. Fully-connected Layer

The main function of this layer is to classify the input image into different classes based on training data using features which we got from previous layers. The combination of all these previous layers is used to create a CNN model. The last layer is a fully connected layer where the actual classification happens.

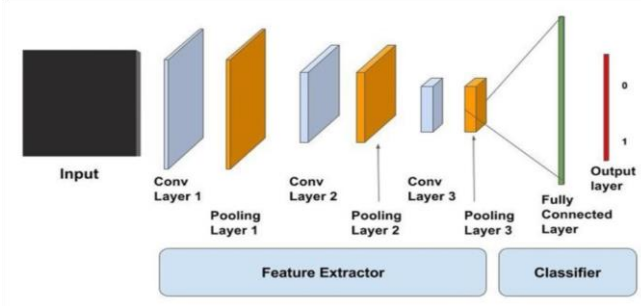

Fig. 4: Convolution Neural Network

After we get the input the CNN plays its role by passing the input through its 4 layers. CNN becomes a lot better at seeing similarity other than whole image matching schemes as it compares the image piece by piece, and the pieces it finds for are known as features and also by finding rough feature matches, in roughly the same position between the two images.

In the first layer the input is compared with filters or features which are actually squares (by default image patch size is taken of 9 pixels) containing pixel values of already existing gestures with values 1 or -1. So multiplication is done between the pixel values of a filter and the corresponding pixel values of a bit out of the input and we will be storing the multiplied values in a matrix. Now the matrix values are added up and divided by the total number of values in the matrix, we will get some value after division and that value is stored at the corresponding bit of the input where our filter is currently. Similarly the filter is moved throughout the image and corresponding values are stored at the respective bits (another matrix formed). Similarly the same procedure is repeated with every feature and corresponding matrices are formed.

The matrices we formed contains negative pixel values in it, this is where second layer works and converts all the negative pixel values to 0 and keeps the positive values same. It is done for all the matrices we get after the first layer. After this the third layer comes into the picture where it shrinks the image stack to smaller size usually does it with picking up a window size (usually 2 or 3), picking up stride (usually 2) and moving the window over the filtered images choosing the maximum value from each window and storing it to the corresponding bits.

After all this procedure stacking up of all these layers takes place where the whole procedure from layer 1-3 repeats where it shrinks and changes the input even more and finally we get some values.

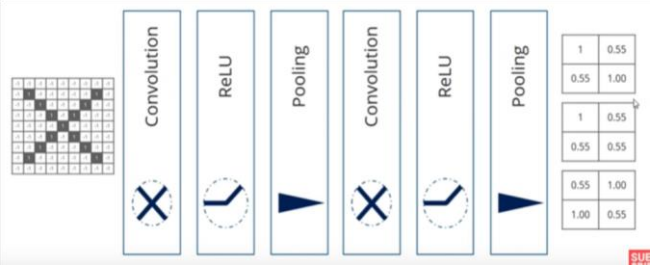

Fig. 5: Matrix formation

The final values we get after the input passes through all the above layers, the fourth layer comes into the picture where the actual classification happens. We take the above images which are shrunk and put them into a single list. So if we notice there are certain values in the list which are high at different positions for different gestures respectively. In case of the above example 1,4,5,10 and 11th value in the list are high. So basically now CNN knows that if the values are higher at 1,4,5,10 and 11th position it would classify it as X. So this is how the four layers of the CNN works together to classify an input image close to the image which we have in our trained model using the existing dataset of gestures.

V. RESULT

The aim of the project was to recognise the American Sign Language hand gestures in real time. The above work shows that we were able to predict the hand gestures with high accuracy with the model we trained using the dataset of 44 gesture samples having 2400 images of each gesture.

After recognising the hand gestures we also were able to form a word using the gestures to add more functionality to the project which made communication easier between specially-abled people who could not hear or speak.

VI. FUTURE SCOPE

The project can further be extended to convert the words to a full fledges sentence. This can provide a platform for creating a full-fledged sign language converter to a normal communication language. The words that are developed into sentences can be spelt out and on further work with emotions. This will become a good bridge for the deaf and mute people to connect with the normal people. The communication will be made even easier.

ACKNOWLEDGMENT

We would like to extend our sincere thanks to SRM Institute of Science and Technology, Vadapalani Campus for providing us with all facilities required in successful completion of the project. We express our hearty gratitude to Dr. S. Prasanna Devi, HOD – CSE for supporting us in all aspects. We are obliged to express our sincere thanks to Ms. S. Niveditha for guiding us throughout, providing complete support and encouraging towards the completion of the project.

REFERENCES

[1] Hand Gesture Recognition using Computer Vision, 2013 by Ashutosh Samantaray, Sanjaya Kumar Nayak, Ashis Kumar Mishra.

[2] Sign Language Translator Application Using OpenCV, 2017 by L Triyono, E H Pratisto, S A T Bawono, F A Purnomo, Y Yudhanto and B Raharjo.

[3] STUDY OF VISION BASED HAND GESTURE RECOGNITION USING INDIAN SIGN LANGUAGE, 2014 by Archana S. Ghotkar and Dr. Gajanan K. Kharate

[4] GESTURE RECOGNITION SYSTEM, 2018 by Prakhyath Rai, Ananya Alva, Gautami K. Mahale, Jagruthi S. Shetty, Manjushree A. N.