ASL Based Hand Image Gesture Recognisation using Convolution Neural Network(CNN)

Harishma Tharakkal Haridas
COMP5112WC Student
Lakehead university

Kevinkumar Patel
COMP5112WC Student
Lakehead university

Nischay Trivedi
COMP5112WC Student
Lakehead university

Pinak Divecha
COMP5112WC Student
Lakehead university

Shiv Hansoti
COMP5112WC Student
Lakehead university

Sabah Mohammed
COMP5112WC Supervisor
Lakehead University

***Abstract*—Of multiple human-computer-interactions (HCI), HCI dependent hand movement may be the most common and intuitive way to connect between humans and computers, because it strongly mimics how a person communicates with each other. HCL plays a key function in bridging the gap in the deployment of IT in modern cities. Hand gestures are commonly accepted as a promising form of HCI, the identification of human hand motions using different methods. It is an intuitive and normal way of discovering massive and complicated files, video games, virtual reality, health care, etc. Although the HCI-based hand gesture market is massive, developing a comprehensive hand gesture recognition network remains a challenging problem for conventional vision-based solutions, which are severely restricted by the quality of optical sensor input. Moreover, in this paper we are proposing a gesture recognition method using convolution neural networks (CNN) for American Sign Language (ASL) which is the most efficient technique for feature extraction and classification. Image-based gesture recognition uses CNN because it provides more accuracy than others and CNNs are very good feature extractors. The procedure involves the application of Back Projection, contour generation, Binary Threshold, and segmentation during preprocessing, in which they contribute to better feature extraction. After recognizing gestures correctly, we used Natural Language Processing (NLP) techniques to form a sentence. To verify the robustness of the proposed system, measured metrics collected during training are analyzed and addressed. The results are collected using a live web camera.**

Keywords—Gesture Recognition, Convolution Neural Network, American Sign Language, Natural Language Processing, Spell Checker.

I. Introduction

Gestures are a non-verbal communication form that will originate from any bodily motion or state that usually comes from face or hand. Users can use simple gestures to interact and manage with devices without touching them. Many approaches are made using cameras and computer vision algorithms to interpret sign language which is the predominant language of those who are deaf or hard to understand and is often used by people with impaired hearing and speech. Gesture recognition is often viewed as a method for computers to begin learning the body language, building a connection between machines and humans than primitive text user interfaces or graphical user interfaces that still restrict input to the keyboard and mouse. Gestures are categorized into static gestures and dynamic gestures. The posture of the body or the gesture of the hand denotes a sign in static gesture whereas the movement of the hand or the body conveys some messages in dynamic gesture.

Typically, Sign recognition is related to image understanding. It mainly consists of two phases: sign detection and sign recognition. Sign detection is an extracting feature of a certain object with respect to certain parameters. Sign recognition is recognizing a certain shape that defines the object from the remaining shapes. There are two methods that are mainly used to perform gesture recognition. One relies on wearable, professional electromagnetic devices whereas the other one based on computer vision. The former one performs well but is costly and unusable in some environments. The latter involves camera-based gesture recognition. The hand region is detected using an input device such as a camera and features like skin color sensitive to the lighting condition and feature points are extracted in order to understand the gesture.

In order to detect the gestures from the images most efficient deep learning model is Convolution Neural Network(CNN). CNN can do much better than other classic neural networks on images because it takes advantage of the inherent properties of images. Simple feedforward neural network doesn't see any order in inputs but CNN takes advantage of the local spatial coherence of the images. CNN is fast and decreases the number of operations needed to process on the image by using convolution on patches of adjacent pixels.

In this research work we are working with American Sign Language. The ASL is shown in image 1. In order to detect the ASL, we would be using the CNN model. We would first take an image of the gesture from the video camera and perform pre-processing. After doing pre-process we would pass the image into our CNN model and classify the gesture.

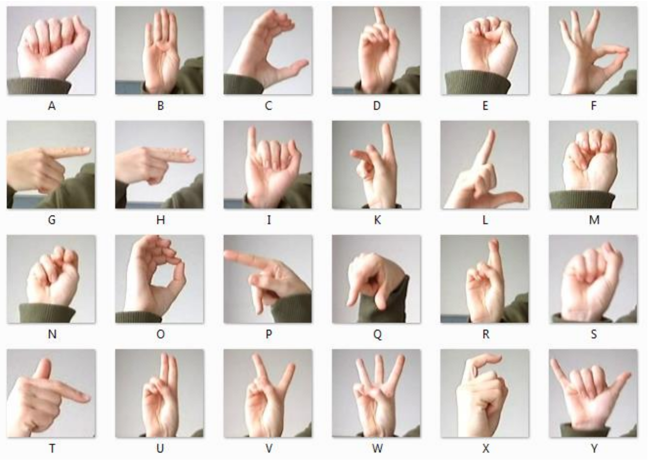


Fig 1: American Sign Language. Source: Kaggle

II. Related Research Work

There are many ways to classify the image. In order to classify the Gesture Images most efficient way is to use a convolution neural network. They [1] acquired 500 images of 9 different gestures using the webcam to evaluate the model. After separating the dataset into a 70:30 ratio is able to achieve an accuracy of 98.74% using CNN. Detecting the hand gesture using Deep Learning is another method. They [2] have used the Gaussian mixture model of skin color for identifying the skin color from the background and in order to separate the hand they used haar features. After identifying the hand, the gesture's picture was used in order to classify the number using CNN. They have created a dataset consisting of numbers using hand gestures and in total they have used 16000 images.

Feature extraction is an important step in gesture recognition. Fingertip finder, eccentricity, elongatedness, pixel segmentation and rotation are different feature extraction methods. They [3] have used this feature extraction method to extract features from American Sign Language (ASL) dataset. In their research they have used a Multilayer feed-forward neural network with a backpropagation training algorithm which is a part of Artificial Neural Network (ANN). They got the highest recognition rate of 85.2 % with the fingertip finder feature extraction method. In [3] this paper they examined in detail the design of the previously implemented Convolutional Long Short-Term Memory Recurrent Neural Network (CNN LSTM) for gesture recognition tasks. They used a Deconvolutional Neural Network to visualize the features of the original input picture which each of the kernels of our CNNLSTM model's convolutional layers learned to extract. They have got 80.10% frame level accuracy with their proposed model.

In [4] this paper they compared various techniques with their proposed model and showed that using skeletonization algorithm they got the better results compare to the other model. They also tested CNN model in which they got the 93.63% accuracy whereas with their proposed model they got the 96.01% accuracy which is even better. So, the method used in this paper reduces the dependence of gesture recognition on devices significantly in complex environment. While in [5] this study, they implemented an end-to-end incorporation of a CNN into an HMM thus analyzing the CNN's outputs in a completely Bayesian manner. They proposed a hybrid CNN-HMM framework integrating CNN's deep discriminative capabilities with HMM's sequence simulation capabilities thus adhering to Bayesian concepts. They achieved results of 38.3% and 38.8% on dev and test respectively.

III. Methodology

We are using ASL and manually created a dataset for gesture recognition. In order to properly detect the gesture, we have used the flow shown in figure 2, 3 and 4. To predict the accurate result from image, we have used Convolution Neural Network (CNN) (figure 5). A Convolutional Neural Network (ConvNet/CNN) could be a Deep Learning algorithm that may absorb an input image, assign importance (learnable weights and biases) to varied aspects/objects within the image and capable of differentiating one from the opposite. The pre-processing required in a real ConvNet is very lower as compared to different classification algorithms. While in primitive methods filters are hand-engineered, with enough training, ConvNets have the flexibility to find out these filters/characteristics.

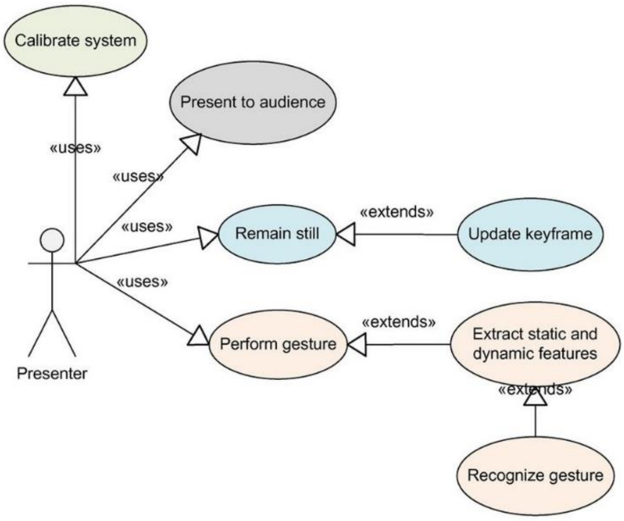


Fig 2. Use case

For the training step, all images are converted to binary images with a 50 x 50 scale. Later these images are trained with a CNN model. For the testing purpose we have used a live webcam. Before predicting the target image, it is essential to recognize the gesture accurately. Then we have to pre-process an image so that it will extract only the important features from the image and remove noise. To do this, we have implemented several image processing steps such as Flip the image, Back projection, Blur techniques, threshold and Find contours. The basic idea behind all this method is to concentrate on hand features, regardless of background and intensity. To mask the skin color, we have used a range of [0, 180, 0, 256]. Finally, we get the binary output of a hand in white color with a black background. Later these images are fed to CNN to check the result.

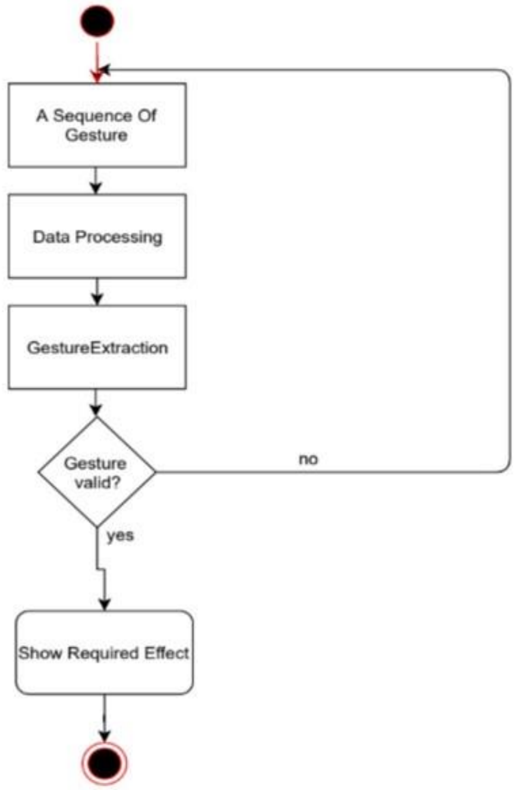


Fig 3. Flow diagram

The architecture of a ConvNet is analogous to it of the connectivity pattern of Neurons within the Human Brain and was inspired by the organization of the cortical region. Individual neurons reply to stimuli only during a restricted region of the sight view called the Receptive Field. a set of such fields overlap to hide the complete cortical area. In the proposed CNN model, we have used 3 convolutional layers followed by Maxpooling layer and at last we fully connected layer that gives the output.

A ConvNet is able to successfully take the Spatial and Temporal dependencies in a picture through the device of appropriate filters. The architecture achieves a much better fitting to the image dataset recognition to the reduction within the number of parameters required and reusability of weights. In other words, the network will be trained to grasp the sophistication of the image better.

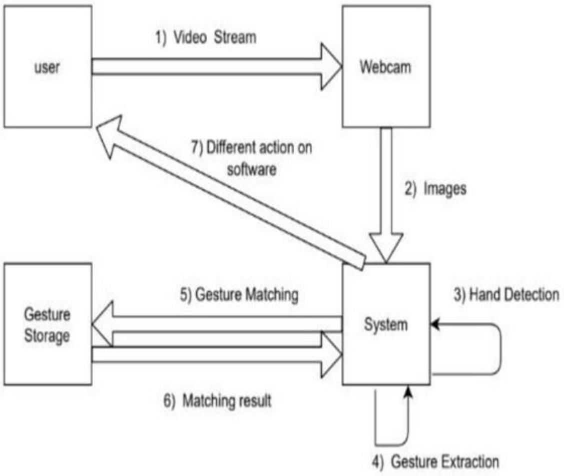


Fig 4. Colab diagram

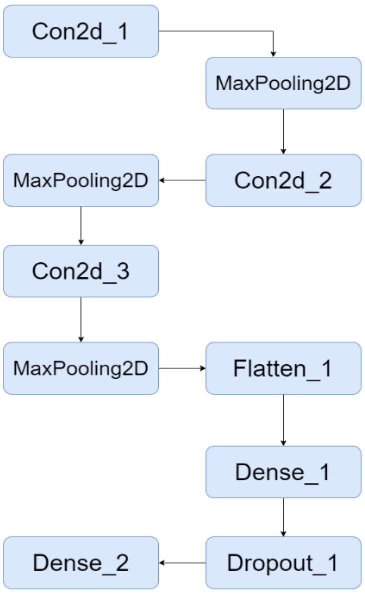
We can assume how computationally intense things would get once the pictures reach dimensions like 8K (7680×4320). The role of the ConvNet is to scale back the pictures into a form that is less complicated to process, without losing features which are critical for getting an honest prediction. this can be important after we are to style an architecture that isn't only good at learning features but is also scalable to massive datasets.

We have used the Tensorflow frame in order to develop the Convolution model. The layer details are shown in figure 2.

IV. Prototyping

In order to implement the ASL based gesture recognition we have used the following steps:

- We have taken the picture of the hand in such a way that it is clear, and we get proper histogram of the image.
- After setting the histogram, part of the loading dataset arrives. In loading images, the manually created dataset and dataset used from the MNIST platform are taken and split for training and testing purposes.
- After loading the dataset, it's time for training a model. Convolution Neural Network was used for the training dataset. Tensorflow was used in building the model.
- In CNN, we augment the 5x5x1 image into a 6x6x1 image so apply the 3x3x1 kernel over it, we discover that the convolved matrix seems to be of dimensions 5x5x1.
- Similar to the Convolutional Layer, the Pooling layer is liable for reducing the spatial size of the Convolved Feature. this can be to decrease the computational power required to process the info through dimensionality reduction. Furthermore, it's useful for extracting dominant features which are rotational and positional invariant, thus maintaining the method of effectively training of the model.
- Adding a Fully Connected layer could be a (usually) cheap way of learning non-linear combinations of the high-level features as represented by the output of the convolutional layer.



-

Fig 5. CNN Model

- Now that we've got converted our input image into an appropriate form for our Multi-Level Perceptron, we shall flatten the image into a column vector. The flattened output is fed to a feed-forward neural network and backpropagation applied to each iteration of coaching.
- After probing the above process, we've got successfully enabled the model to grasp the features. Now, we are flattening the ultimate output and feed it to a daily Neural Network for classification purposes.
- Created a CNN which looks plenty almost like this MNIST classifying model using both Tensorflow and Keras. If you would like to feature more gestures you would possibly have to add your own layers and also tweak some parameters, that you just need to do on your own.
- Then used the model which was trained using Keras on a video stream.

- As of today, we got stored the 44 gestures that are 26 alphabets and 10 numbers of Yankee signing and a few other gestures. And trained the model on these images.

- After training the model it is time to check the model. Testing is finished on the premise of the live gesture performed ahead of the camera. supported those gesture a trained model tries to recognize the hand gesture and display the alphabets of the gesture on the screen.

- Gestures were tested using live or colorful camera and a threshold binary image which helps to optimize the output.

- In ASL there is some gesture which is hard to recognize as they have similar hand gesture which causes the problem to generate a wrong word. In order to overcome the problem, we have used the spell check using the python library which corrects the wrongly spelled spelling.

V. Conclusion

In this paper, we have proposed a real-time vision-based static hand gesture recognition system for American Sign Language which is capable of translating the gesture into text along with the sound. The proposed Convolution Neural Network (CNN) architecture achieved high success rates at a relatively low computational cost. It was above the methodologies which are mentioned in the related works, establishing the robustness of the introduced method. In addition, the architectures proposed reached accuracy very close to the architectures already described in the literature, although they are much simpler and have lower computational costs. This is possible due to the proposed image processing methodology, in which we convert the image into a greyscale. A database consists of 43 gestures and for each gesture, we have 2400 images. As for data pre-processing, skin color algorithm in HSV color space was used to segment the skin area from the background, and a method that can adaptively determine the upper and lower limits of skin color range, instead of fixing them as a specific value is developed. Moreover, in order to resolve the orientation problem, we proposed a rotation method that can accurately determine the rotation angle without losing useful information. For classification, CNN is used with a successful rate of ---- training accuracy.

[1] Jiang, Xinyun, and Wasim Ahmad. “Hand Gesture Detection Based Real-Time American Sign Language Letters Recognition Using Support Vector Machine.” 2019 IEEE

[2] Islam, Md. Mohiminul, Sarah Siddiqua, and Jawata Afnan. “Real Time Hand Gesture Recognition Using Different Algorithms Based on American Sign Language.” 2017 IEEE

[3] Zhan, Felix. “Hand Gesture Recognition with Convolution Neural Networks.” 2019 IEEE

[4] Mcguire, R.m., J. Hernandez-Rebollar, T. Starner, V. Henderson, H. Brashear, and D.s. Ross. “Towards a One-Way American Sign Language Translator.” Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004

[5] Raimundo F. Pinto, Carlos D. B. Borges, Antônio M. A. Almeida and Iális C. Paula. “Towards a One-Way American Sign Language Translator.” 10 Oct 2019