SIGN LANGUAGE RECOGNITION

CONVOLUTION NEURAL NETWORKS

INT247 – Machine Learning Foundation Practical

Maddula Vishnu Vardhan Reddy – 11909034 – KM015 – A23

Department of Computer Science and Engineering, Bachelor of Technology
Lovely Professional University, Punjab, India

ABSTRACT

Sign language is the only tool of communication for the person who is not able to speak and hear anything. Sign language is a boon for the physically challenged people to express their thoughts and emotion. In this work, a novel scheme of sign language recognition has been proposed for identifying the alphabets and gestures in sign language. With the help of computer vision and neural networks we can detect the signs and give the respective text output.

American Sign Language (ASL) is the most popular standard for sign language in North America. However, it can be tough for those who do not know sign language to communicate with those who cannot communicate easily without it. This project serves as an introduction into the world of using ML to recognize sign language.

Keywords: Sign Language Recognition, Convolution Neural Network, Image Processing, Streamlit, American Sign Language

Dataset: https://www.kaggle.com/grassknoted/asl-alphabet
The dataset contains 87,000 200x200 pixel images; 3000 images for each letter of the alphabet, in addition to space, delete, and nothing.

Objective: The goal of this project was to create a convolutional neural network to recognize the ASL alphabet. Additionally, build a web app with Streamlit.

INTRODUCTION

Speech impaired people use hand signs and gestures to communicate. Normal people face difficulty in understanding their language. Hence there is a need of a system which recognizes the different signs, gestures and conveys the information to the normal people. It bridges the gap between physically challenged people and normal people. ASL (American Sign Language) is an important communication way to convey information among deaf people. By visual signing, the brain processes linguistic information; this signing includes shape, movement, and placement of the hands, as well as facial expressions and body movements. ASL is not a universal language, each country has its own language, and in each region of each country, we can find dialects. Due to communication problems, it is very difficult for the deaf community the inclusion in school, job, and personal environments. Plenty of research works in automatic Sign Language Recognition (SLR) has been being published in the last two decades ago.

There are three types of automatic sign language recognition systems: 1) namely sentence; 2) words; 3) fingerspelling. Fingerspelling (alphabetic sign language) is considered an essential part of learning sign language for new users and helps signers to perform signs for names of people, cities, and other words without known signs. There are some published works in which authors propose systems for ASL alphabet recognition.

ASL alphabet recognition is a very difficult task due to high interclass similarities and high intraclass variations. In order to overcome this, in this paper, we propose to use a Convolutional Neural Network (CNN) in order to give the computer the ability of similarity learning and thus, reduce the interclass similarity and the intraclass variation of the non-linear representation of images of each sign of the ASL alphabet.

RELATED WORK

Characterization of sign language is between two parameter one being manual and other non-manual. The manual parameter consists of motion, location, hand shape, and hand orientation. The non-manual parameter includes facial expression, mouth movements, and motion of the head. Sign language does not include the environment which kinesics does. Few terms are use in the sign language like signing space, which refers to signing taking place in 3D space and close to truck and head. Signs are either one-handed or two-handed. When only the dominant hand is in use to perform the signs, they are denoted as one-hand signs else when the non -dominant hand also comes in the phase it is termed as two- handed signs.

Sign language when evolved is different from spoken language so the grammar of the sign language is primarily different from spoken language. Unspoken language, the structure of the sentence is one-dimensional; one word followed by another, while in sign language, a simultaneous structure exists with a parallel temporal and spatial configuration. As based on these characteristics, the syntax of sign language sentence is not as strict as in spoken language. Formation of a sign language sentence includes or refers to time, location, person, base. In spoken languages, a letter represents a sound. For deaf nothing comparable exists. Hence the people, who are deaf by birth or became deaf early in their lives, have very limited vocabulary of spoken language and faces great difficulties in reading and writing.

LITERATURE

The sign language hand gesture recognition is a well contributed research area with a lot of different approaches in implementing it, in this chapter, I review a variety of such approaches for hand sign language gesture recognition. mt entire literature review on hand gesture recognition can be categorized into three groups, image processing/statistical modelling-based recognition, classic machine learning based recognition and deep learning-based recognition. Triesch and Malsburg in 1996 developed an ASL hand gesture recognition system, aiming at performing accurate gesture recognition even for images captured with complex background. In this system, they have eliminated hand segmentation assuming hand images input. For hand representation or feature extraction, the system uses Gabor filters as the Gabor filters are known to resemble the receptive fields of visual cortex. Upon obtaining the Gabor features, they performed a similarity matching technique for gesture recognition called Elastic Bunch- Graph Matching (EBMS) technique.

OpenCV (Open-Source Computer Vision Library) is a library of programming functions mainly aimed at real-time computer vision. Originally developed by Intel, it was later supported by Willow Garage then Itseez (which was later acquired by Intel). The library is cross-platform and free for use under the open-source BSD license.

A neural network is a series of algorithms that endeavours to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates. In this sense, neural networks refer to systems of neurons, either organic or artificial in nature. Neural networks can adapt to changing input; so the network generates the best possible result without needing to redesign the

output criteria. The concept of neural networks, which has its roots in artificial intelligence, is swiftly gaining popularity in the development of trading systems. A neural network works similarly to the human brain's neural network. A "neuron" in a neural network is a mathematical function that collects and classifies information according to a specific architecture. The network bears a strong resemblance to statistical methods such as curve fitting and regression analysis. Convolutional neural networks (CNN) are a special architecture of artificial neural networks, proposed by Yann LeCun in 1988. CNN uses some features of the visual cortex. One of the most popular uses of this architecture is image classification. For example, Facebook uses CNN for automatic tagging algorithms, Amazon — for generating product recommendations and Google — for search through among users' photos.

For experimentation I have considered a subset of 29 gestures from American sign language and the dataset consists of 89000 images with 200X200 from Kaggle. The system has achieved an accuracy of 98%.

METHODOLOGY

1. ACQUISITION OF DATA (CAMERA INTERFACING)

This is a primary and essential step in sign recognition whole process. Camera interfacing is necessary task to capture images with the help of Webcam. Now a days lots of Laptops are coming with inbuilt camera system so that's helps lot for capturing images to process it further. Gestures can be captured by inbuilt camera to detect hand movements and position. Capturing 30fps will be sufficient to process images; more input images may lead to higher computational time and will make system slow and vulnerable.

2. IMAGE PROCESSING

Image pre-processing contains removing unwanted noise, adjusting brightness and contrast of the image, cropping the image as per requirement. In this process contains image enhancement, segmentation and colour filtering process.

3. IMAGE ENCHANCEMENT AND SEGMENTATION

As images captured by webcam is RGB images, but RGB images are very much sensitive for various light conditions therefore RGB information convert into YCbCr. Where Y is luma component which denotes luminance information of image, and Cb , Cr are chromo components which give colour information of images red difference and blue difference. Luminance component may create problems so only chrominance components get process further. After that YCbCr image converted to binary image.

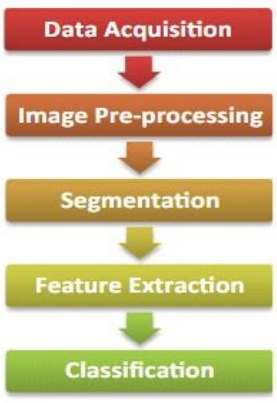

Figure 1: Image processing Flow

Proposed Method

One of the biggest challenging tasks in ASL alphabet recognition, as mentioned above, is the high interclass similarities and the high intraclass variance. In this paper, we propose a CNN architecture which can overcome these two problems performing a similarity learning and thus, reducing the interclass similarities and the intraclass variance among images.

The basic objective of this project is to develop a computer based intelligent system that will enable dumb people significantly to communicate with all other people using their natural hand gestures. The idea consisted of designing and building up an intelligent system using image processing, machine learning and artificial intelligence concepts to take visual inputs of sign language's hand gestures and generate easily recognizable form of outputs. Hence the objective of this project is to develop an intelligent system which can act as a translator between the sign language and the spoken language dynamically and can make the communication between people with hearing impairment and normal people both effective and efficient. The system is we are implementing for Binary sign language, but it can detect any sign language with prior image processing.

For experiments, at first, we used small CNN network architectures, for example, one architecture was composed of 4 convolutional layers and 1 fully connected layer, but this architecture was overfitted, and despite of having used a high Dropout rate, the network did not converge. We conclude from this experiment that the last feature maps were too small, and it was difficult for the network to have good learning.

Thus, I decided to increase the number of convolutional layers to 8 and to conserve the size of the feature maps using paddings, as well as to increase the number of dense layers because they are responsible for encoding; this architecture achieved a validation accuracy of 98%. This value of accuracy was too small, so we decided to add two more convolutional layers as well as to increase the number of neurons of the last dense layer. The proposed scheme was selected because it showed a better performance compared to the rest of the experimental architectures.

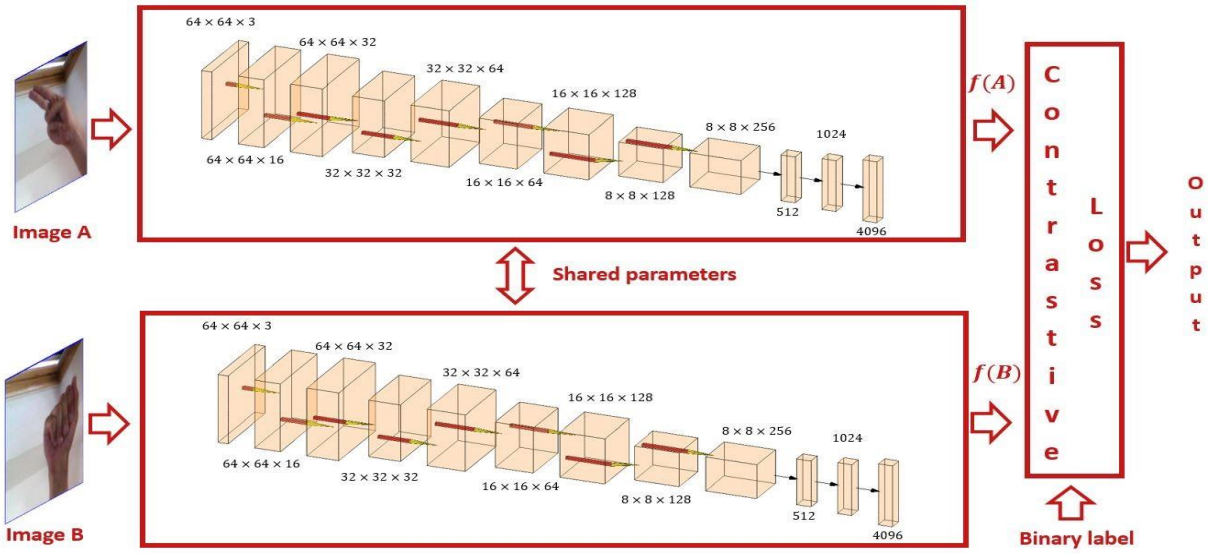

Fig. 2. The proposed architecture consists of two identical CNN which are sharing their parameters. Each network gets a representation of the input image and then they are fed into the contrastive loss for similarity learning. The output of the Siamese architecture is a score that indicates the similarity of the image pair

Dataflow Diagram

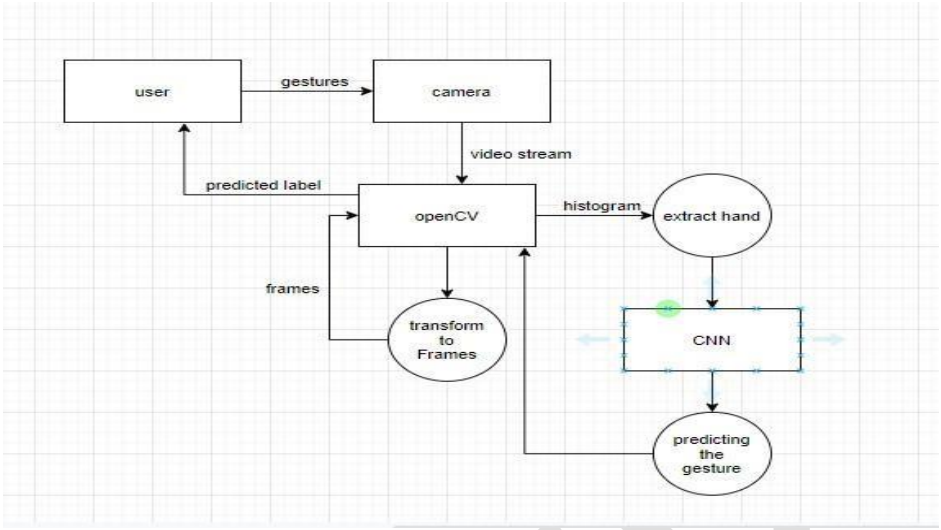

Fig.3: Dataflow Diagram for Sign Language Recognition

Use Case Diagram

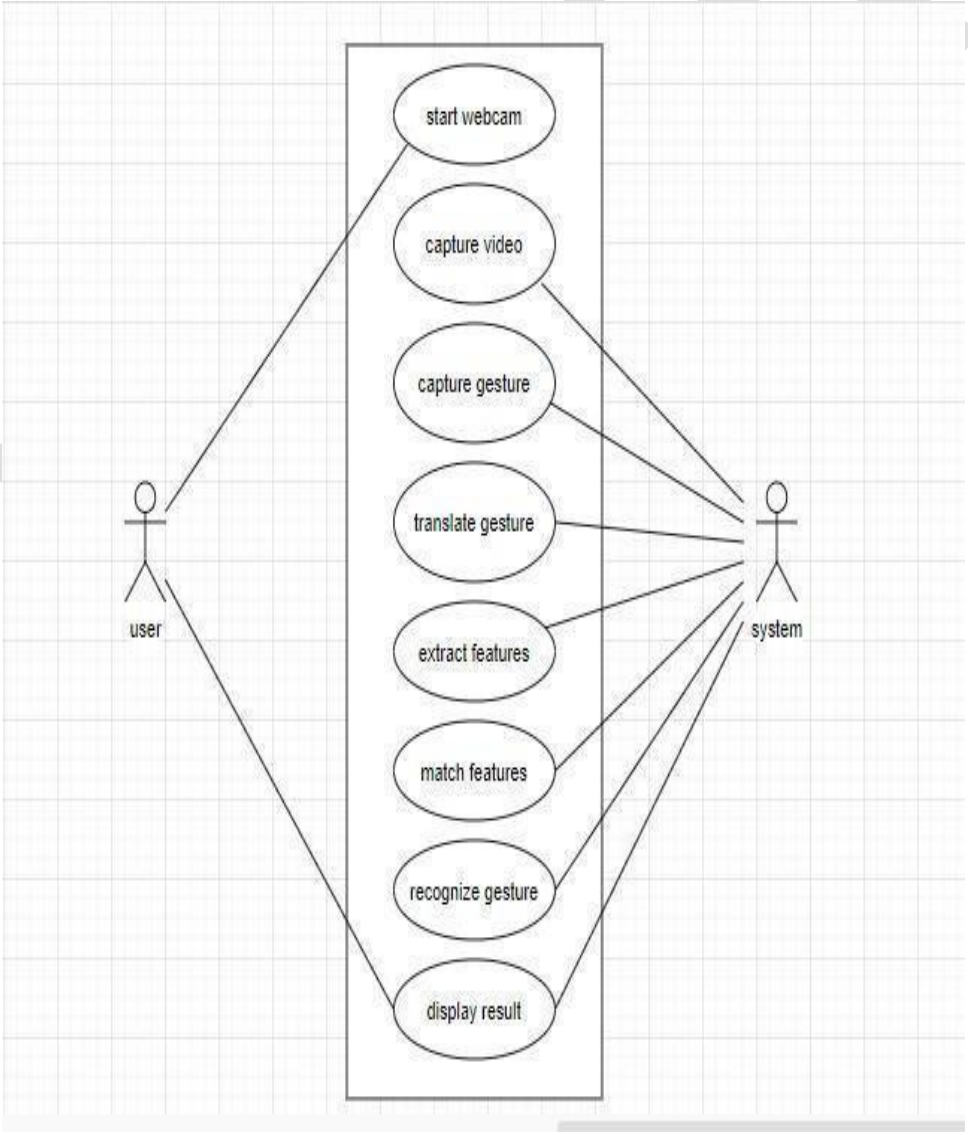

Fig. 4: Use case diagram of sign language recognition System

AMERICAN SIGN LANGUAGE (ASL):

American Sign Language (ASL) is the non-verbal way of communication based on English language. Which can be expressed by movements of the hands and face. It is the primary language of many North Americans who are deaf and find difficulties in hearing. It is not a universal sign language. Different sign languages are used in different countries or regions. For example, British Sign Language (BSL) is a different language compared to ASL so the person who knows ASL may not understand BSL.ASL is forth most used Language in US.

ASL is a language completely segregated and different from English. ASL contains all the significant features of language, with its own rules for pronunciation, word formation, and word order. While every language has ways of indicating different functions, such as asking question instead of making a statement.

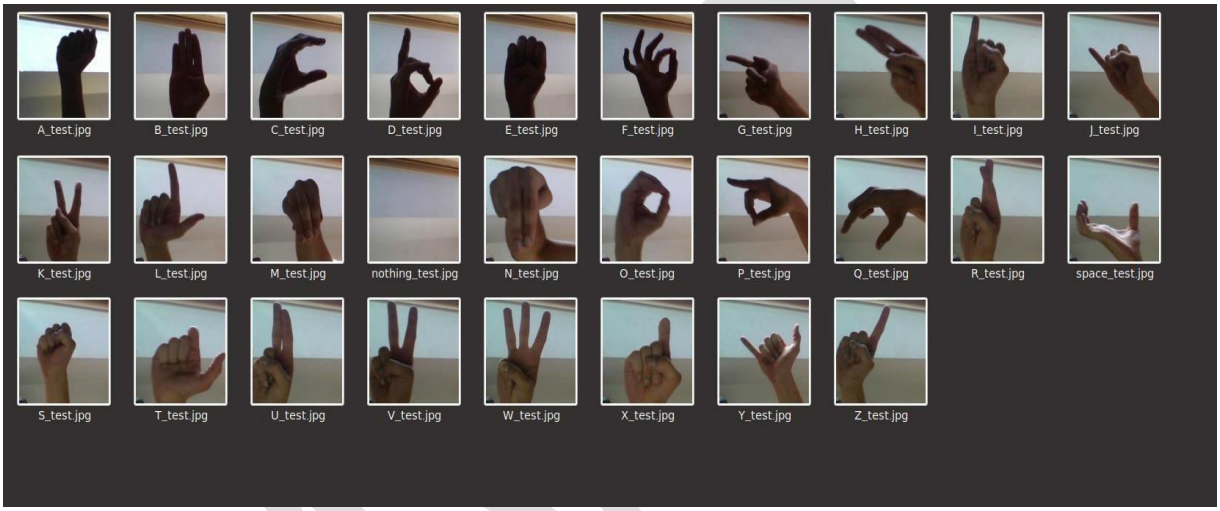

Fig.5: Sample pictures of training data

CLASSIFICATION: CONVOLUTION NEURAL NETWORK

Image classification is the process of taking an input (like a picture) and outputting its class or probability that the input is a particular class. Neural networks are applied in the following steps:

1) One hot encodes the data: A one-hot encoding can be applied to the integer representation. This is where the integer encoded variable is removed and a new binary variable is added for each unique integer value.

2) Define the model: A model said in a very simplified form is nothing but a function that is used to take in certain input, perform certain operation to its best on the given input (learning and then predicting/classifying) and produce the suitable output. 3) Compile the model: The optimizer controls the learning rate. We will be using 'adam' as our optimizer. Adam is generally a good optimizer to use for many cases. The adam optimizer adjusts the learning rate throughout training. The learning rate determines how fast the optimal weights for the model are calculated. A smaller learning rate may lead to more accurate weights (up to a certain point), but the time it takes to compute the weights will be longer. 4) Train the model: Training a model simply means learning (determining) good values for all the weights and the bias from labelled examples. In supervised learning, a machine learning algorithm builds a model by examining many examples and attempting to find a model that minimizes loss; this process is called empirical risk minimization. 5) Test the model

A convolutional neural network convolves learned featured with input data and uses 2D convolution layers.

Convolution Operation:

In purely mathematical terms, convolution is a function derived from two given functions by integration which expresses how the shape of one is modified by the other. Convolution formula:

$$(f * g)(t) \stackrel{\text{def}}{=} \int_{-\infty}^{\infty} f(\tau)\, g(t-\tau)\, d\tau$$

Relu Layer:

Rectified linear unit is used to scale the parameters to non-negative values. We get pixel values as negative values too . In this layer we make them as 0's. The purpose of applying the rectifier function is to increase the non-linearity in our images. The reason we want to do that is that images are naturally non-linear. The rectifier serves to break up the linearity even further in order to make up for the linearity that we might impose an image when we put it through the convolution operation. What the rectifier function does to an image like this is remove all the black elements from it, keeping only those carrying a positive value (the grey and white colors).The essential difference between the non-rectified version of the image and the rectified one is the progression of colors. After we rectify the image, you will find the colors changing more abruptly.

Pooling Layer:

The pooling (POOL) layer reduces the height and width of the input. It helps reduce computation, as well as helps make feature detectors more invariant to its position in the input This process is what provides the convolutional neural network with the "spatial variance" capability. In addition to that, pooling serves to minimize the size of the images as well as the number of parameters which, in turn, prevents an issue of "overfitting" from coming up. Overfitting in a nutshell is when you create an excessively complex model in order to account for the idiosyncracies we just mentioned The result ofusing a pooling layer and creating down sampled or pooled feature maps is a summarized version of the features detected in the input. They are useful as small changes in the location of the feature in the input detected by the convolutional layer will result in a pooled feature map with the feature in the same location. Thiscapability added by pooling is called the model's invariance to local translation.

Fully Connected Layer:

The role of the artificial neural network is to take this data and combine the features into a wider variety of attributes that make the convolutional network more capable of classifying images, which is the whole purpose from creating a convolutional neural network. It has neurons linked to each other, and activates if it identifies patterns and sends signals to output layer .the outputlayer gives output class based on weight values, For now, all you need to know is that the loss function informs us of how accurate our network is, which we then use in optimizing our network in order to increase its

effectiveness. That requires certain things to be altered in our network. These include the weights (the blue lines connecting the neurons, which are basically the synapses), and the feature detector since the network often turns out to be looking for the wrong features and has to be reviewed multiple times for the sake of optimization. This full connection process practically works as follows:

- The neuron in the fully connected layer detects a certain feature; say, a nose.
- It preserves its value.
- It communicates this value to the classes trained images.

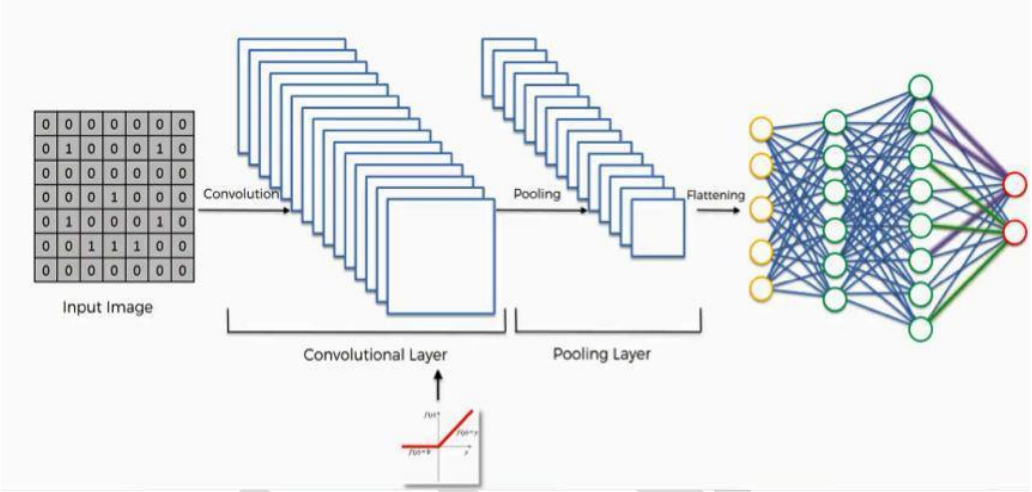

Fig.6: Layers involved in CNN

Experimental Results

The dataset we used for this paper is a sub-set from ASL Alphabet [4] dataset from Kaggle. This dataset consists of 26 ASL alphabet signs (from A to Z) and 3 classes labeled as “SPACE”, “DEL” and “NOTHING”, which according to the authors of the dataset, these are very helpful for real-time applications. Something that is important to mention is that in this dataset, “J” and “Z” are considered static signs

.

The training was done using torch and neural networks as frameworks on the Visual Studios platform with a single 16GB Nvidia Tesla P100 GPU. After 17 epochs, the training loss and training accuracy were 0.0164 and 0.9870, respectively, and achieved a validation loss and a validation accuracy of 0.0245 and 0.9764, respectively.

Confusion Matrix:

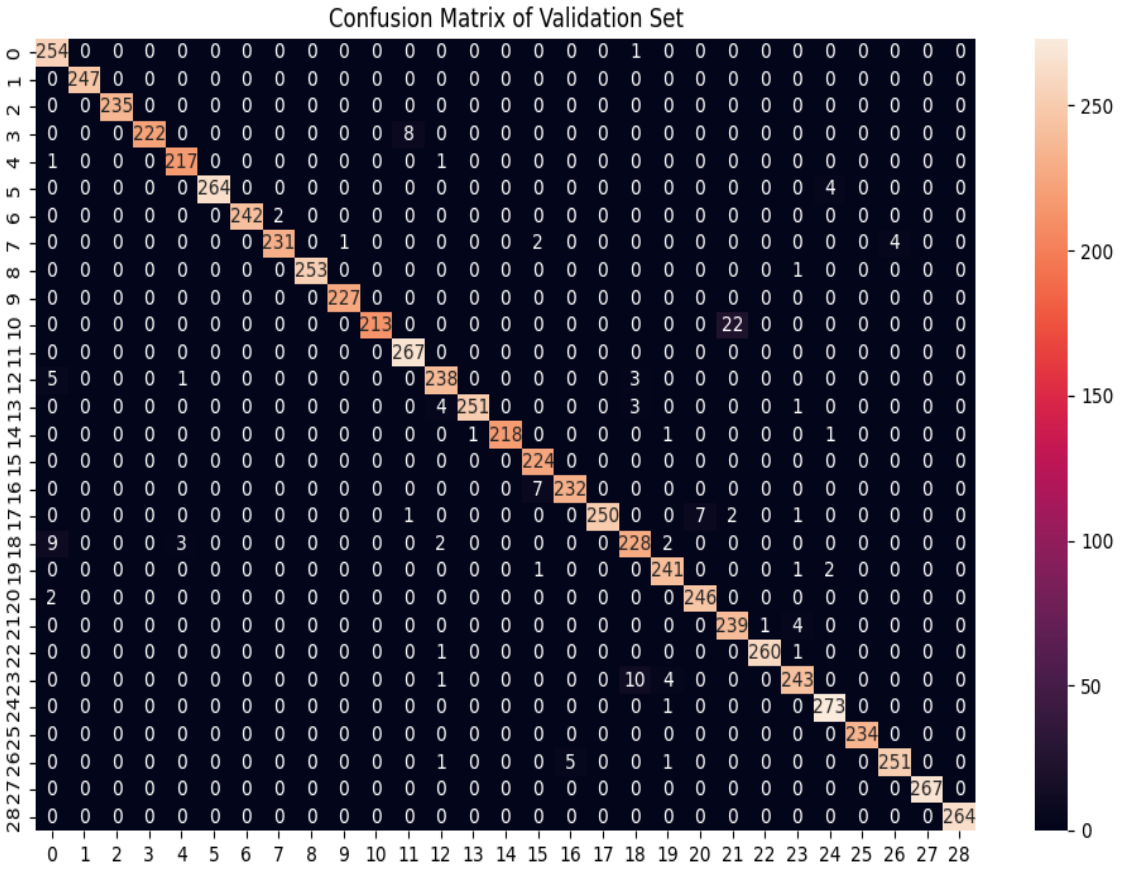

Fig.7: Confusion Matrix

Test Result:

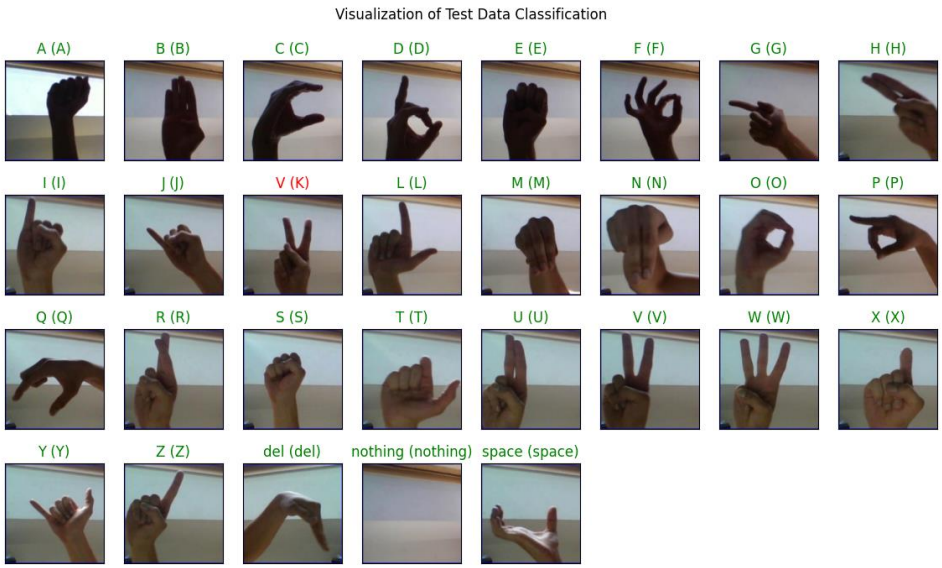

Fig 8: Test Data Resuls

Conclusion

Sign language is not only important for people who are deaf, but also for people who want to communicate with them. Nowadays, the deaf community faces struggle due to the communication gap that exists between hearing people and them. It is very important to develop a system for sign language translation to overthrow this communication wall.

In this paper, I propose a system to carry out the simplest task in ASL recognition, which is ASL alphabet recognition. One of the most challenging tasks in this field is the high interclass similarity and high intraclass variation in ASL alphabet recognition. Then, our hypothesis was to obtain image encoding where those belonging to the same class should be separated by a small distance (low variation) and at the same time by a large distance (low similarity) from those who belong to a different class. Experimental results show that our hypothesis is correct since we achieved to reduce the interclass similarity and intraclass variation, with some poor results in two pairs of classes. However, in general, I considered the proposed scheme performed well at classifying.

The model presents a performance pretty good to identify the static images of the sign alphabetic language. The system shows that the first stage can be useful for deaf persons or with speech disability for communicating with the rest of the people who do not know the language. In this work, the developed hardware architecture is used as image recognizing system but it is not only limited to this application, it means, the design can be employed to process other type of signs.

As future work, it is planned to add to the system a learning process for dynamic signs, as well as to prove the existing system with images taken in different position. Several applications can be mentioned for this method: finding and extracting information about human hands, which can be apply in sign language recognition that it is transcribed to speech or text, robotics, game technology, virtual controllers and remote control in the industry and others.

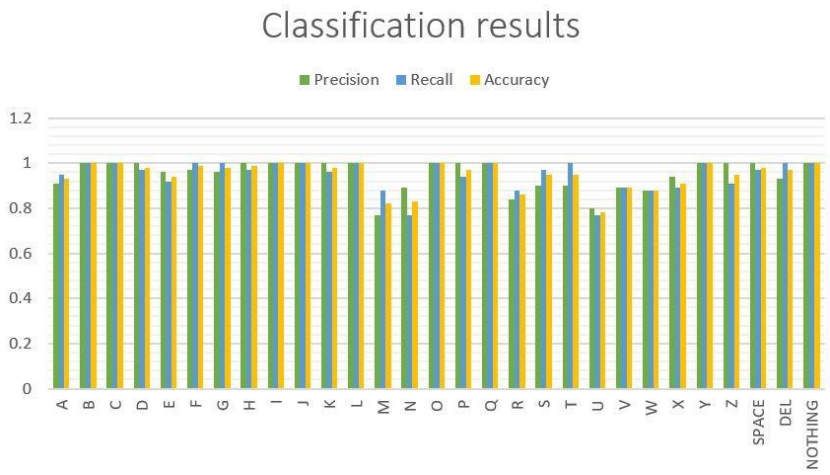

Fig.9: Precision, Recall and Accuracy Classification (Credits: Kaggle)

Acknowledgments

I would like to thank my Professor Dr Sagar Pande for all the given support to accomplish this research.

References

1. **Aly, W., Aly, S., & Almotairi, S. (2019).** User-independent american sign language alphabet recognition based on depth image and pcanet features. *IEEE Access*, Vol. 7, pp. 123138–123150.
2. **Cao Dong, Leu, M. C., & Yin, Z. (2015).** American sign language alphabet recognition using microsoft kinect. *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 44–52.
3. **Fierro, A., Nakano, M., Yanai, K., & Perez, H.(2019).** Siamese and triplet convolutional neural networks for the retrieval of images with similar contents. *Informacion Tecnologica*, Vol. 30, No. 6, pp. 243–254.
4. **Kaggle (2020).** Kaggle homepage. [Online available]: https://www.kaggle.com/ grassknoted/asl-alphabet. [Accessed: 20/06/2020].
5. **Kuznetsova, A., Leal-Taixe, L., & Rosenhahn,´ B. (2013).** Real-time sign language recognition using a consumer depth camera. *2013 IEEE International Conference on Computer Vision Workshops*, pp. 83–90.
6. **Maqueda, A. I., del Blanco, C. R., Jaureguizar, F.,& Garc´ıa, N. (2015).** Human–computer interaction based on visual hand-gesture recognition using volumetric spatiograms of local binary patterns. *Computer Vision and Image Understanding*, Vol. 141, pp. 126–137.
7. **Nai,W. Liu, Y., Rempel, D., & Wang,Y. (2017).** Fast hand posture classification using depth features extracted from random line segments. *Pattern Recognition*, Vol. 65, pp. 1–10.
8. **Pugeault, N. & Bowden, R. (2011).** Spelling it out: Real-time asl fingerspelling recognition. *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pp. 1114–1119.
9. **Sahoo, J. P., Ari, S., & Ghosh, D. K. (2018).** Hand gesture recognition using dwt and f-ratio based feature descriptor. *IET Image Processing*, Vol. 12, No. 10, pp. 1780–1787.
10. **Salem, A. & Vadera, S. (2017).** A convolutional neural network to classify american sign language fingerspelling from depth and colour images. *Expert Systems*, Vol. 34, No. 3, pp. 1–18.
11. **Tao, W., Leu, M. C., & Yin, Z. (2018).** American sign language alphabet recognition using convolutional neural networks with multiview augmentation and inference fusion. *Engineering Applications of Artificial Intelligence*, Vol. 76, pp. 202–213.