Real Time ISL Prediction Using Deep Learning

Mukul Joshi

CHAPTER 1 – INTRODUCTION

1.1 Outline

Sign Language is a language incorporating non-verbal correspondence that takes its origin in visual cues and hand gestures. Within the language there are fully established hand movements with their individual grammatical significance and inbuilt formation systems. Sign is composed of various gestures formed by shapes of the hand, its movements and orientations. Presently, sign is the primary language of those people who have hearing aid or who are hard of hearing and also used by those who can hear but cannot physically speak. It is a complex but complete language which involves and is not limited to the movement of hands, facial expressions, and postures of the body. Sign language is not universal. Every country has its own native sign language. Each sign language has its own rules of grammar and word orders. The problem arises when the differently-abled try to communicate using sign language with the people who are unaware of this language and its grammar. So it becomes necessity to develop an automatic and interactive interpreter to bridge the gap in communication. Sometimes, the people with hearing and speech impediment seek the help of an intermediary sign language interpreter, a person proficient in both sign and regular speech, so as to translate their thoughts to common people and vice versa. However, this way turns out to be very costly and does not work out all the time. Thus, the need to introduce a system which can automatically recognize the sign language gestures becomes a priority. Introducing such a system would significantly bridge the gap between differently-abled and people who use speech as a form of communication in the society. Often, the sign language in use is different depending upon the culture and language. Particularly, Indian sign language (ISL) is used by the people who have difficulty in speaking and hearing in India. ISL is a standard and well-developed way of communication for differently-abled people in India and those speaking the English language. Different symbols are involved for different alphabets in the Indian Sign Language.

An effective solution to this problem is one based on computer vision-oriented gesture recognition, which involves image processing techniques. Consequently, this category faces more complexity. Sign languages are fully functional yet, are not universally used nor understood, and the easiest way to perceive them is by creating a computer recognition model

that through video Sequence captures spatial movements and records them, predicting the translation system for sign language using Convolutional Neural Networks (CNN). This is divided into three main segments: the system design, the data set, and the deep learning model training and evaluation. This model is a CNN network i.e. model which is trained on pre-processed sign image data set without any complex pre-processing wherein we can directly input live video stream with camera by a signer and fed to the network.

The whole idea behind this dissertation is to detect and predict different signs from a set of predefined signs of Indian sign language.

1.2 Visual Recognition

Visual Recognition or machine recognition is defined by a software's innate ability to recognize and classify images by understanding the contents of images and recognizes images for scenes, objects, faces, colours, food, and other subjects which offer insights into visual content. Computers can use machine vision technologies together with a camera and AI software to realize image recognition.

Visual recognition is employed to perform an out sized number of machine-based visual tasks, like labelling the content of images with meta-tags, performing image content search and guiding autonomous robots, self-driving cars and accident avoidance systems.

While human and animal brains recognize objects without challenge, computers have difficulty with the task. Software for image recognition requires deep machine learning. It performs the simplest with CNN i.e. convolutional neural net processors because the specific task otherwise requires massive amounts of power for its compute-intensive nature.

The major steps in image recognition process are to collect and organize data, build a predictive model and use it to acknowledge images. Its algorithms can function by use of comparative 3D models, appearances from different angles using edge detection or by components. Image recognition algorithms are trained on pre-labelled pictures with guided learning and method then are used as data sets for future references.The main goal for the present and future applications of image recognition includes the accessibility for the visually impaired and enhanced researched capabilities. Many companies including, but not limited to Google, Facebook, Microsoft, Apple and Pinterest are heavily investing resources and research into image recognition and related applications.

As an example to the system presently one of Google's most recent venture on image recognition is a tool called –Google Lens‖ which helps scan real time images with the help of –Google Assistant an AI powered assistant.

CHAPTER 2 – PROJECT DESCRIPTION

2.1 Problem Statement

There are various problems faced by sign language speakers especially when it comes to correspondence with non-sign language speaker’s further creating gaps in the counter communication of thoughts and ideas. This brings into light the issues with feasibility and accessibility of a sign language model.

Although previously models have been devised on the same account, the problem arises when it comes to a collective platform where all the data regarding the same is readily available.

Moreover, languages face alterations with time and so do Sign, a model where the system can keep up with change locally and provide significant up gradation does not exist.

A counter method is to create a system where the data can be stored and updated according to need and a solid front end platform is provided for easier user access and usage.

The approach is to build a cost effective model that is readily accessible.

In this age of technology where everyone has significant access to the internet on web browsers on smart phones the idea is to implement a front end system that will readily access images through image recognition and simultaneously convert it to text or speech.

This would solve the collective problem of lack of availability of a platform for Indian Sign Language speakers and also provide a vast scope of research and development in this particular field.

2.2 Project Perspective

The main agenda of a this Project is to perform real time sign language analysis through the deep learning model as the area itself has a scope of vast improvement, it will help mitigate the communication barrier between the differently abled and people who inherently use speech as a form of communication. To accomplish this we will focus on real time image capturing and focus on segmentation, edge detection and grey scale conversion of the image to avoid any possible errors in image detection. Next is the application of CNN to use predictive measures whilst performing data recognition and to look for the most suited image matches based on the existing data set. The Last step is to create a front end platform to display the achieved text/ speech conversion system for easier user readability.

Deep Learning

Deep learning otherwise called deep structured learning is important for a more extensive group of AI techniques dependent on artificial neural networks with representation learning. Learning can be regulated, semi-managed or unsupervised.

Deep learning architectures, for example, deep neural networks, deep belief networks, recurrent neural networks and convolutional neural networks have been applied to fields including PC vision, machine vision, speech recognition, characteristic language handling, sound acknowledgment, informal organization sifting, machine interpretation, bioinformatics, drug plan, clinical picture examination and material review, where they have delivered results that have surpassed human intervention.

In deep learning, a convolutional neural organization (CNN, or ConvNet) is a segment of deep neural networks, most normally applied to breaking down visual imagery. They are otherwise called shift invariant or space invariant artificial neural networks (SIANN), in light of their shared load architecture and translation invariance characteristics. They have applications in picture and video recognition, recommender systems, classification of images, medical image investigation, regular language processing, and cost related time series.

2.3 Implementation Details

Vision System

The vision system is composed of a camera i.e. a front facing camera/webcam.

Hardware Requirements

A multi-core processor (i5 or higher) or an and similar processor, Minimum 8 gb RAM, Space in disk (SSD preferred), Minimum Output display (1280x720), Cuda enabled Nvidia GPU for training Deep Learning model, Internet connectivity (up to 500 kbps of speed)

Software Requirements

OS (Linux/windows) Linux (Ubuntu) preferred, Deep Learning API's, Computer Vision supporting libraries, Python programming language, Jupyter notebook/ VS studio code

CHAPTER 3 – METHODS AND MATERIALS

3.1 Data Set

The Indian sign language dataset obtained included more than 40,000 images of sign language actions and around 1200 for each symbol (A to Z and 0-9) represented specifically. Since the images obtained were from a single source the discrepancy between the images and its background is limited but is also susceptible to noise and background which can be erased through blur for an ideal binary dataset to work upon.

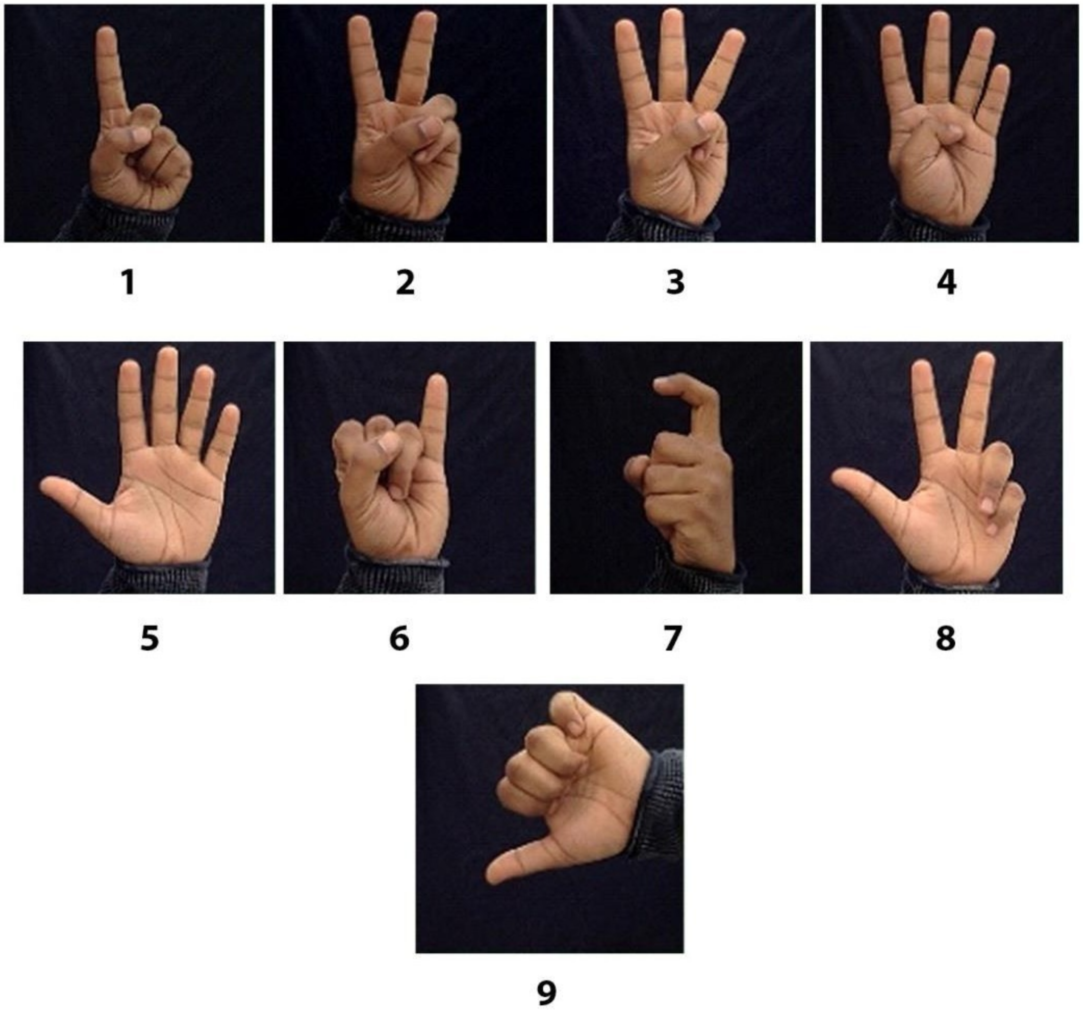


Fig 3.3 Indian Sign Language Dataset

3.2 Technologies Used

Python

Python is an interpretive, high-level and general-purpose programming language. Created by Guido van Rossum and first published in 1991, Python's design Philosophy emphasizes the readability of code with its impressive use of large white space. Python is interpreted, there’s no need to compile the software until you run it. This is equivalent to both PERL and PHP. It can prompt and communicate directly with the interpreter to write your programs. Also, it is Object-oriented it supports Object-Oriented style or programming technique that encapsulates code inside items.

OpenCV

OpenCV (Open Source Computer Vision Library) is a library of programming functions specifically designed for real-time computer vision. Originally developed by Intel, it was later supported by Willow Garage and Itseez (later acquired by Intel). The library is cross-platform and free to use under the open-source Apache 2 License. OpenCV features GPU acceleration for real-time operations beginning in 2011.

Officially launched in 1999, the OpenCV project was originally an Intel Research effort to advance CPU-intensive applications, part of a series of initiatives including real-time ray tracing and 3D display walls.[4] The key contributors to the project included a range of Intel Russia optimization experts and the Intel Performance Library Team.

If the library detects Intel's Integrated Performance Primitives on the device, these proprietary tailored routines will be used to speed up the operation.

CNN

In deep learning, a Convolutional Neural Network (CNN, or ConvNet) is a class of deep neural networks most widely used for visual imaging analysis. They are also known as shift invariant or space invariant artificial neural networks (SIANN) based on their shared-weight architecture and translation invariance characteristics.

CNNs are regularised variants of multi-layer perceptrons. Multilayer perceptrons typically mean completely linked networks, that is, each neuron in a single layer is connected to all neurons in the next layer. The "full-connectivity" of these networks makes them vulnerable to data over-fitting. Typical methods of regularisation include introducing some type of weight calculation to the loss function. CNNs take a different approach to regularisation i.e., they take advantage of the hierarchical structure in the data and assemble more complicated patterns using smaller and simpler patterns. As a consequence, on the scale of connectivity and complexity, CNNs are at the lower end.

Convolutional networks are influenced by biological processes in that the pattern of connectivity between neurons resembles the organisation of the animal visual cortex. Individual cortical neurons reply to stimuli only in a restricted area of the visual field known as the receptive field. The receptive fields of various neurons overlap partially in such a way that they occupy the entire visual field.

Pytorch

PyTorch is an open source machine learning library based on the Torch library, used for applications such as computer vision and natural language processing, primarily developed by Facebook's AI Research lab (FAIR).It is free and open-source software released under the Modified BSD license.

Although the Python interface is more polished and the primary focus of development, PyTorch also has a C++ interface.

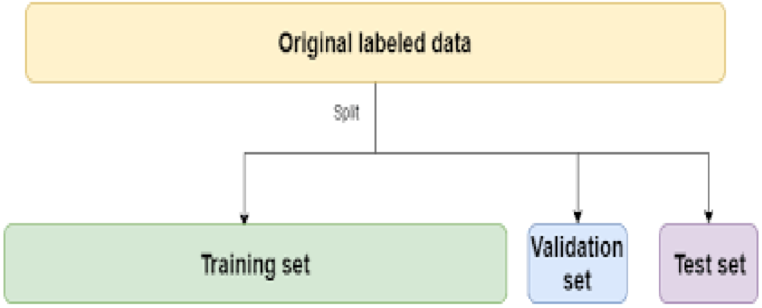
A number of pieces of Deep Learning software are built on top of PyTorch, including Tesla Autopilot, Uber's Pyro, HuggingFace's Transformers, PyTorch Lightning, and Catalyst.

PyTorch provides two high-level features:

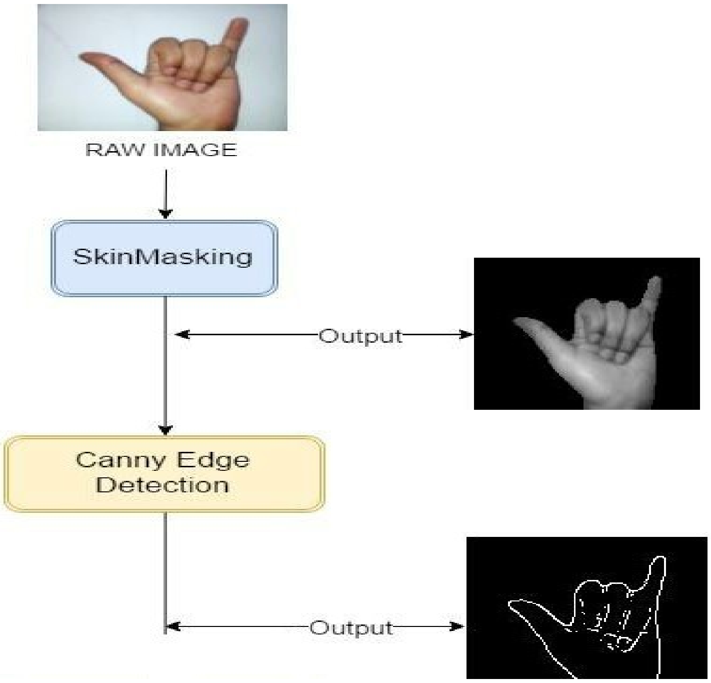
1) Tensor computing (like NumPy) with strong acceleration via graphics processing units (GPU).
2) Deep neural networks built on a tape-based automatic differentiation system.

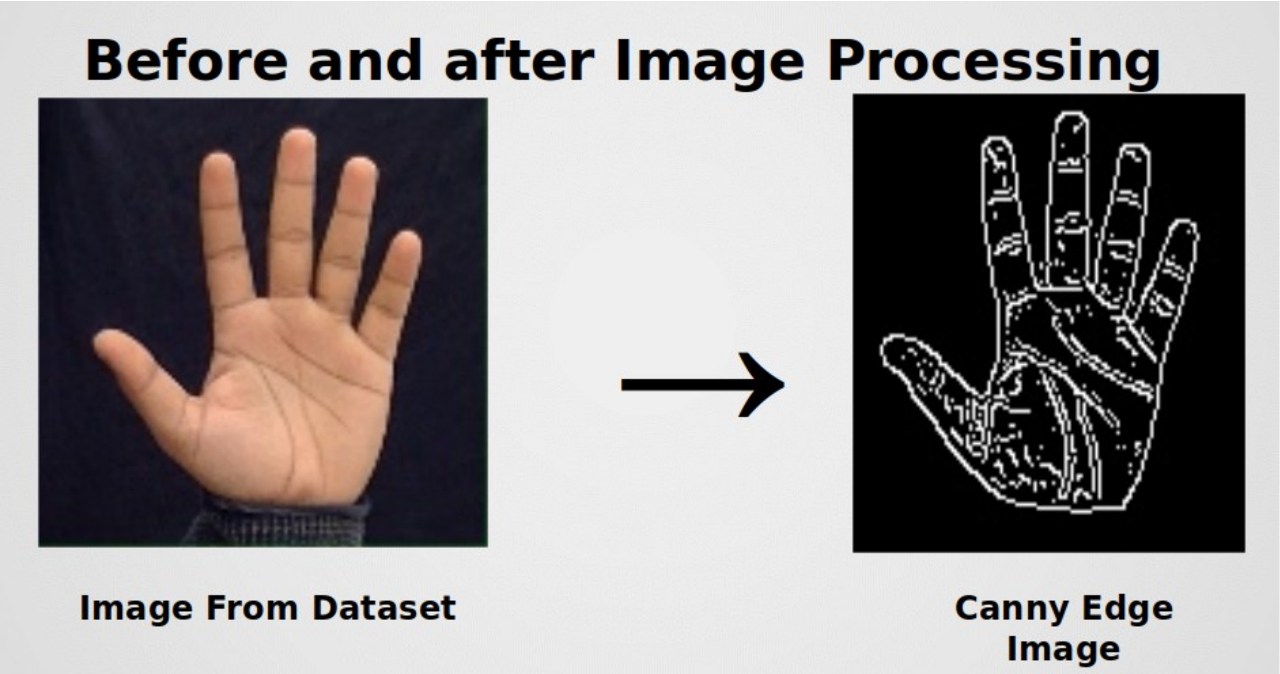
3.3 Methodology

For achieving the successful implementation the first step was to acquire the dataset and split the data set for training, validation and testing purposes. Splitting the dataset is essential for an unbiased evaluation of prediction performance.



After the data-set was acquired necessary preprocessing steps were to be performed in order to reduce noise, differentiate background form object, getting contours and highlight edges so that the model could predict correct target labels with great efficiency. For all this image processing algorithms were performed like BGR to Gray conversion, BGR to HSV conversion, skin masking, median blur and canny image detection.





All processed training data images were fed as input to the CNN classifier in the form of training data through the data loader in batch size of 64 and with transformations like resizing the image into (128x128) pixels, performing horizontal flip, batch normalization and in the end converting the image to tensor.

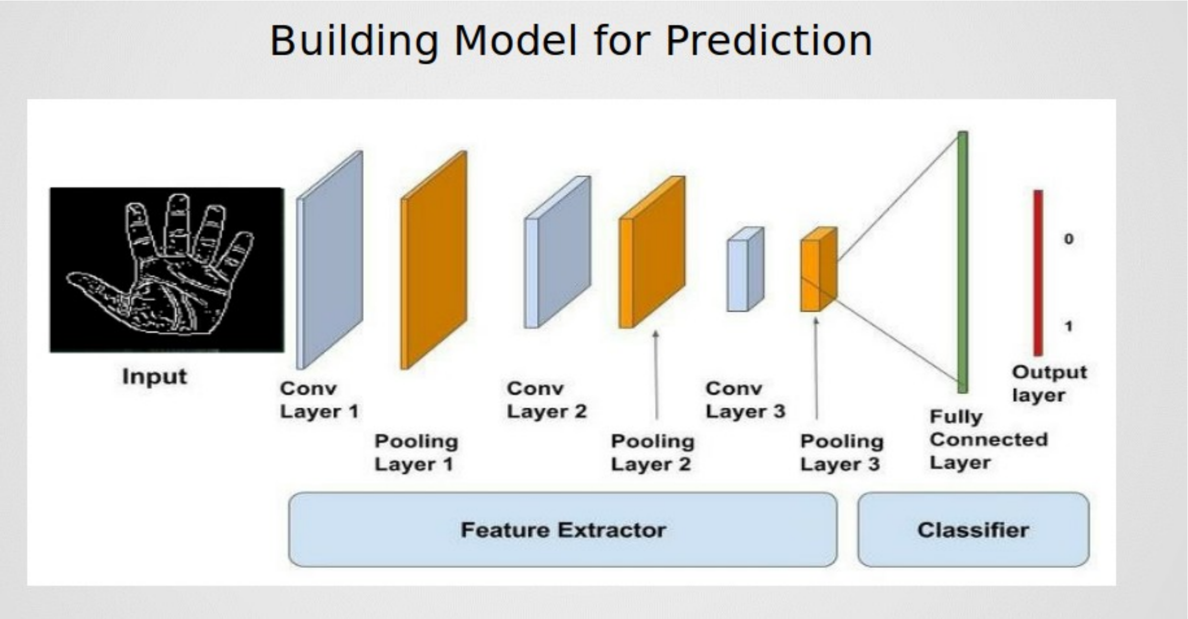


Fig 3.1 Working Model approach

The model training was trained for 35 epochs and every time the validation loss decreased model current state was saved so that the best model could be loaded later and be used for classification process.As Pytorch supports training the model over GPU (Graphics Processing Unit) whole training was done on the GPU for fast processing.

After the training part was done, the model was loaded and tested against the test data. The model gave an accuracy of 98 percentage which was quite good.

For the live video feed part OpenCv was used . All the pre-processing steps that were used before were used to convert the live feed .

Processed frames were given to the classifier such that if frame is a match and classified then it returns the corresponding alphabet or digit for that signed gesture.

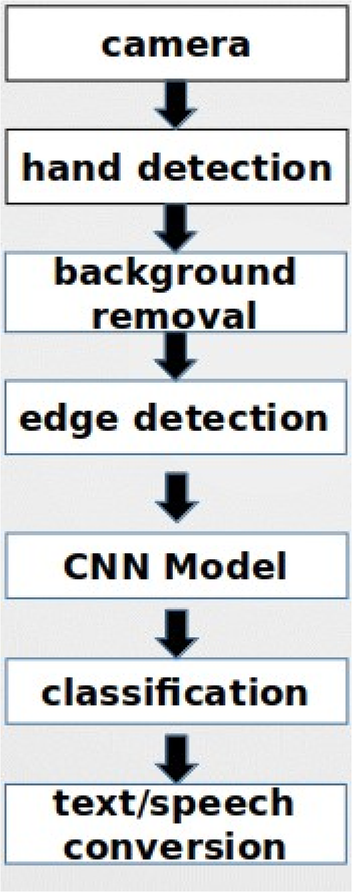


Fig 3.2 Process of Methodology