Conversion of Hand Gestures to Text and Speech in Multiple Languages

Saptarshi Dutta Gupta
Somisetty VR Dinesh
Jyoti M Angadi
Natasha P

Prof Preethi Sangamesh
Batch No. - 23

March 28, 2019

Problem Statement / Definition

Conversion of hand gestures of speech impaired people into a multitude of languages for reducing the communication gap between the differently abled and normal people with the help of image processing and pattern matching algorithms.

Motivation of the Work

For ages, communication has always been the medium for conveying expressions and information between individuals. However, there are people who are unfortunately born without the power of hearing or speaking. Communication for speech impaired people has always been a major challenge. For comprehending what they want to say, we need to understand their hand gestures or sign language. The proposed model aims at helping the differently abled communicate more effectively with people by converting their hand gestures into speech in different languages.

Literature Survey

- **A Hand Gesture Recognition using Feature Extraction**
- Acquires a user's hand gesture from common hardware support such as web-cams or mobile integrated cameras and classifies it based on the predefined hand gestures, stored in a database.
- The design of the system basically has two parts namingly preprocessing and classification.
- The binary image of the hand portion is converted using thresholding technique. The image so formed may contain some noise snd to remove these noises a median filter is used.

Advantages:

1. -Reduces complexity while comparing the image with the training image and also increases correctness of the system.

Disadvantages:

1. Presence of too much noise can cause hindrance to the recognition system.

Literature Survey

- **Intelligent Sign Language Recognition Using Image Processing**
- The project aims to determine human gestures by creating a HCI(human computer interface) and contains four modules: camera interfacing, image processing,pattern matching and text and audio generation.
- It takes the color image, converts it to a grayscale representation of its luminance by obtaining the value of its red, green, and blue (RGB) primaries making the required portion of image as white by using thresholding technique and garbage part is made as black.
- **Advantages:**
  1. Uses color differentiation so makes differentiating skin and markers from the background much easier.

  Disadvantages:

  1. It works only one way, that is, it does not convert speech or text to hand gestures.

Literature Survey

- **New Methodology for Translation of static sign Symbol to words in Kannada language**
- Taking 36 letters of Kannada language image acquisition and using Kannada Sign Language(KSL) and Hearing- Impaired(HI).
- Feature extraction phase uses histogram technique, Hough and segmentation to extract hand from the static sign. Classification phase uses neural network for training samples.
- **Advantages:**
  1. Usage of neural network will increase the speed of interpretation significantly.

  Disadvantages:

  1. Only using Kannada hand gestures and converting it into only Kannada language.

Literature Survey

- **Hand Gesture Recognition based on Digital Image Processing using MATLAB**
- Based on Digital Image Processing using Color Segmentation, Skin Detection, Image Segmentation and deals with American Sign Language(ASL)
- Uses Template Matching and combine feature detection with gesture detection very easily.
- **Advantages:**
  1. The design is very simple and the signer doesn't need to wear any type of hand glove. Also, this application can be run in an ordinary computer having a web camera

  Disadvantages:
  1. The boundaries of gesture have to be automatically detected.
  2. Delay in the processing execution can be occurred due to the large amount of high resolution images.

Literature Survey

- **Gesture to Speech Conversion in Hindi language for Hearing and Speech disabled person in INDIA**
- Taking 46 letters of Hindi language and measuring the depth of the image using gesture area segmentation
- Image extraction to black and white using fast indexing of SURF algorithm
- Classification using Neural Network and feature extraction using Principal Component Analysis. Speech Recognition using Windows text to speech system
- **Advantages:**
  1. Neural Networks for classification and PCA increases accuracy and efficiency at the same time increases complexity.

  Disadvantages:

  1. Only using hindi hand gestures and converting it into a single language

Literature Survey

- **Hand gesture recognition and voice conversion system for dumb people**
- The interpreter makes use of a glove based totally technique comprising of flex detector, instrument sensors.
- Makes use of flex sensors(measures the bent of fingers), data gloves and accelerometer sensor(measures the tilting of the hands)
- Amplification of signals are done when a gesture is made which is then captured by the microcontroller
- **Advantages:**
  1. Accuracy is very high because this method is taking into consideration all hand orientation and movements.

  Disadvantages:
  1. High cost of implementation

Methodology

There are five main stages in this project:

Image Acquisition

- Acquiring picture from webcam and cropping the hand portion of it.

Preprocessing Phase

- Converting picture from RGB to grayscale
- Image smoothing using Gaussian filter.
- Converting the smoothed image into binary format using Thresholding technique.
- Removing the noise using Morphological Technique namely Dilation and Erosion.

Methodology

Feature Extraction

- Detect image edges by Canny edge detection algorithm
- Obtain the feature vector from the algorithm output

Creation/Usage of the ML Model

- Train/Test using a Neural Network Model

Text to Speech Conversion

- Convert the obtained result from text to speech according to the user specified language

Methodology

A detailed discussion on the various methods:

Thresholding

- Thresholding is the simplest form of image segmentation, that converts a grayscale image to binary image.
- The simplest thresholding methods replace each pixel in an image with a black pixel if the image intensity I(i,j) is less than some fixed constant T (i.e. I(i,j)<T) or a a white pixel if the image intensity is greater than the constant.

Methodology

Morphological Techniques

- Morphological transformations are some simple operations based on the image shape. It is normally performed on binary images. It needs two inputs, one is our original image, second one is called structuring element or kernel which decides the nature of operation. Two basic morphological operators are Erosion and Dilation.
  1. Erosion: This erodes away the boundaries of foreground object when the kernel slides through the image (as in 2D convolution). A pixel in the original image (either 1 or 0) will be considered 1 only if all the pixels under the kernel is 1, otherwise it is eroded (made to zero).
  2. Dilation: Here, a pixel element is '1' if atleast one pixel under the kernel is '1'. So it increases the white region in the image or size of foreground object increases. Normally, in cases like noise removal, erosion is followed by dilation. Because, erosion removes white noises, but it also shrinks our object. So we dilate it. Since noise is gone, they won't come back, but our object area increases. It is also useful in joining broken parts of an object.

Methodology

Gaussian smoothing

- Gaussian smoothing is used as a pre-processing stage in computer vision algorithms in order to enhance image structures at different scales.
- The Gaussian blur is a type of image-blurring filter that uses a Gaussian function (which also expresses the normal distribution in statistics) for calculating the transformation to apply to each pixel in the image.

Methodology

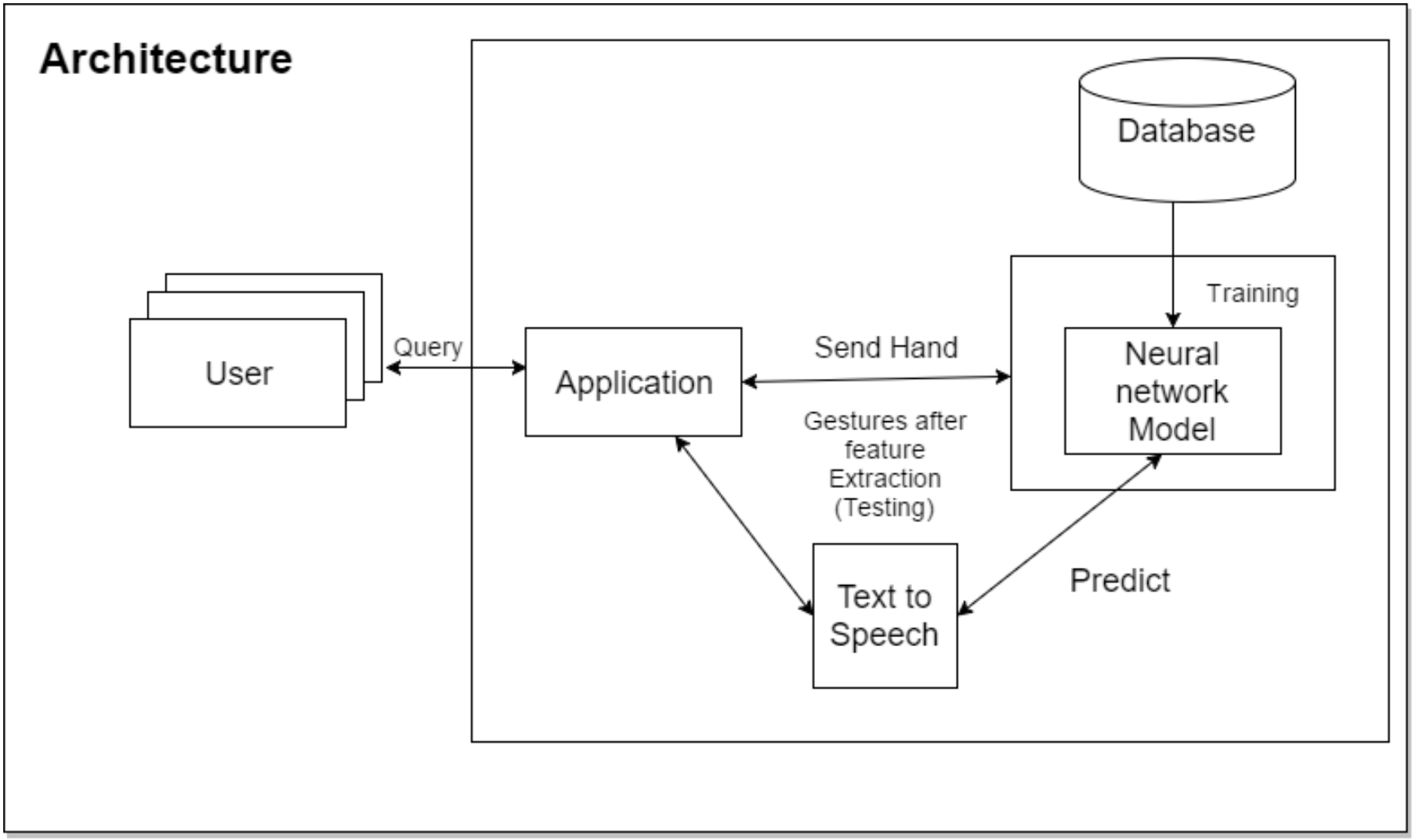
Canny Edge Detection Algorithm

- Canny Edge Detection is a multi-stage edge detection algorithm that follows the following steps:
  1. **Noise Reduction**: using a 5*5 Gaussian filter
  2. **Finding Intensity Gradient of The Image**: A Sobel kernel is used to find the derivative in both the horizontal(Gx) and vertical direction(Gy). From these the edge gradient and the direction for each pixel is determined.
  3. **Non-maximum suppression**: A full scan of image is done to remove any unwanted pixels which may not constitute the edge. For this, at every pixel, pixel is checked if it is a local maximum in its neighborhood in the direction of gradient.
  4. **Hysterisis Thresholding**: This stage decides which are all edges are really edges and which are not. For this, we need two threshold values, minVal and maxVal. Any edges with intensity gradient more than maxVal are sure to be edges and those below minVal are sure to be non-edges, so discarded.

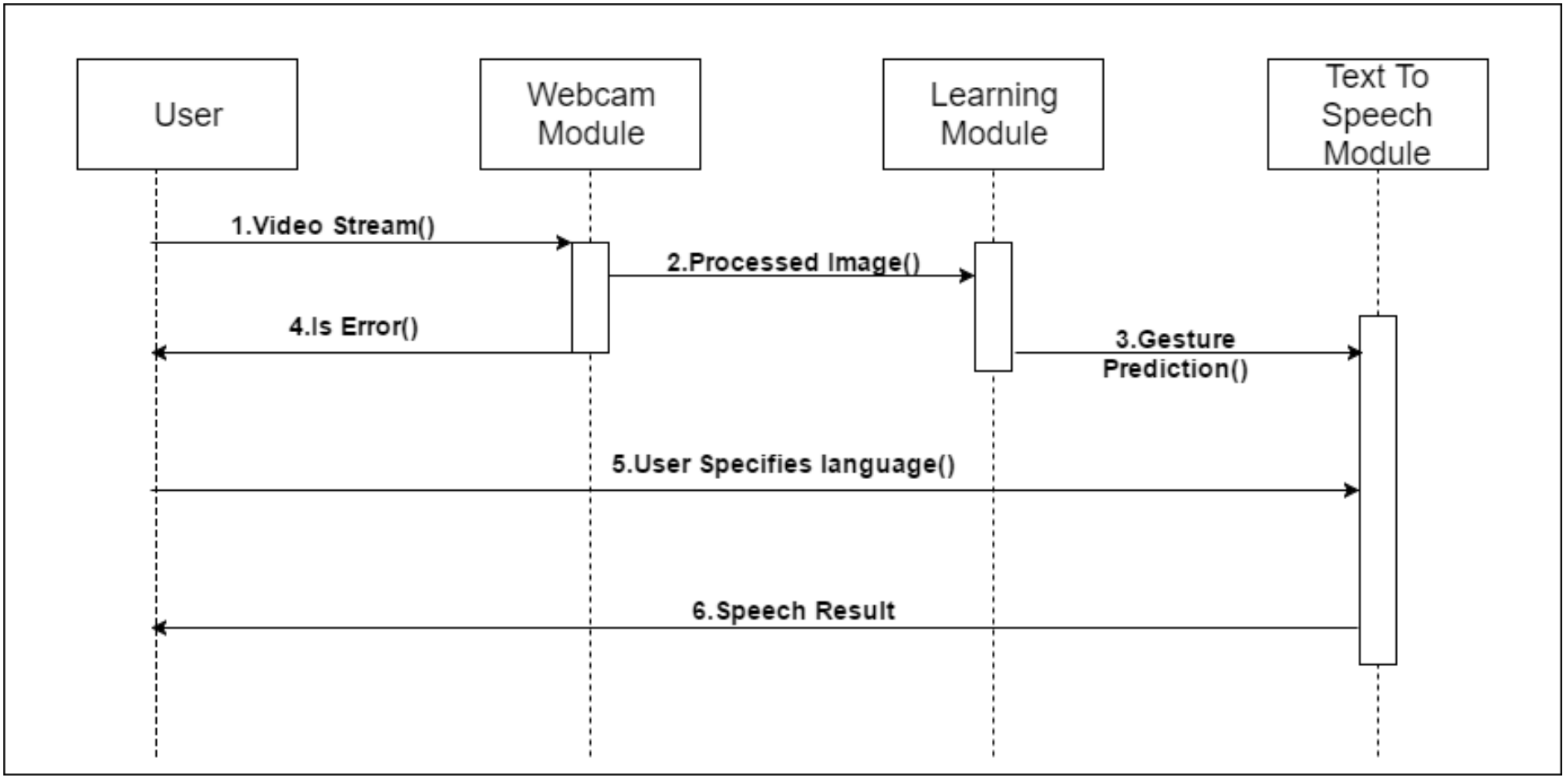
High Level Design

- The user queries the application for either entering a new gesture in the database or finding out the meaning of a already entered gesture.
- After image processing and feature extraction, the hand gesture is sent to the Learning Model which predicts what the gesture means
- Finally the prediction is converted to the required speech using the user specified language.

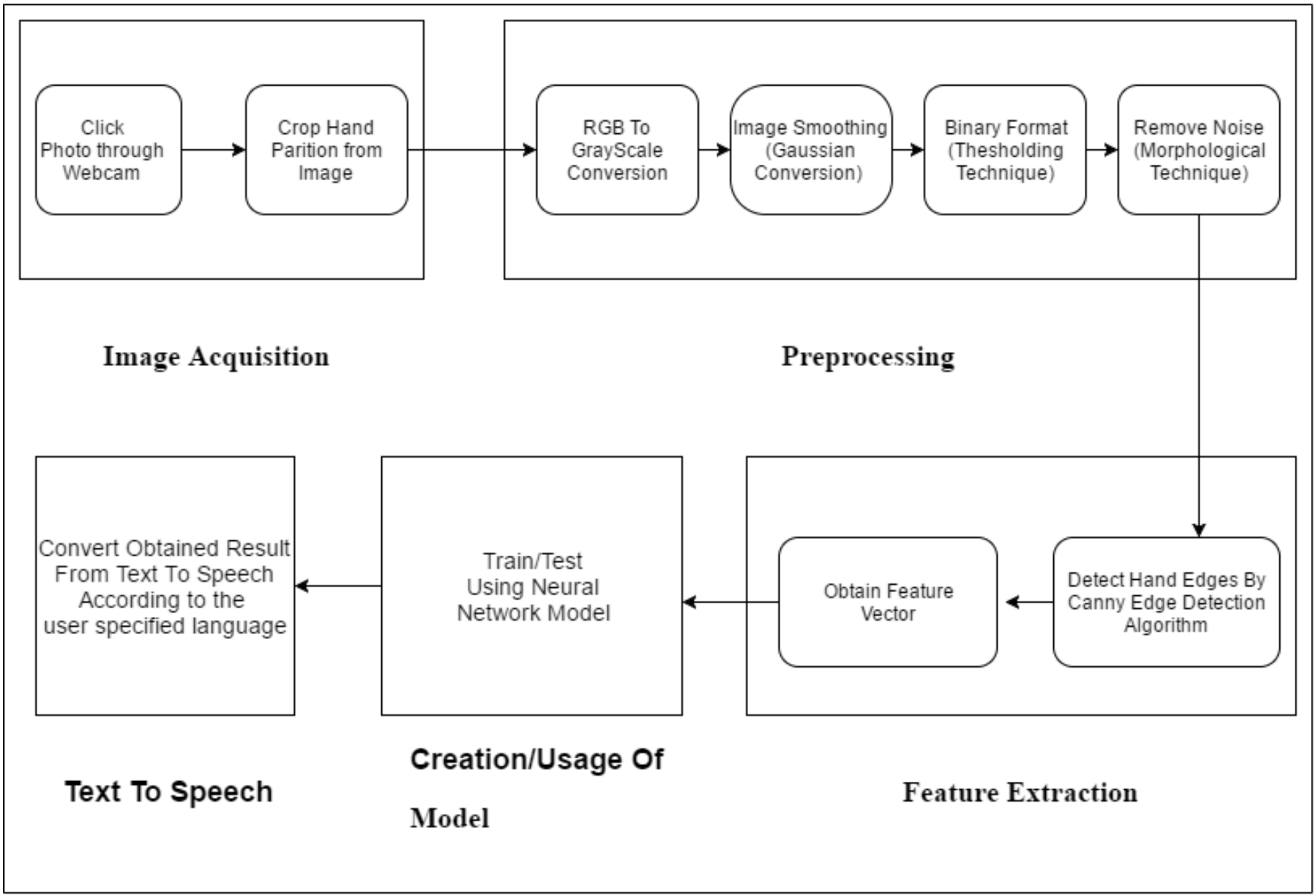
High Level Design



Sequence Diagram



Detailed Design



Dataset

- 36 classes from A-Z and 0-9.
- Each class has 250 images(total 9000)
- After artifically generating image data near about 58,000 images

Convolutional Neural Network

- A convolutional neural network is used for classifying images
- It is a deep feed forward neural network
- Four layered concept:
  1. Convolution
  2. Flattening
  3. Pooling
  4. Full Connections

Why Convolutions?

- Convolution is performed on an image to identify certain features in an image.

Example: classifier.add(Convolution2D(32, 3, 3, inputshape=(64, 64, 3), activation = 'relu'))

- Relu(Rectified Linear Unit): activation function to remove all the negative values from the convolution

Pooling Layers

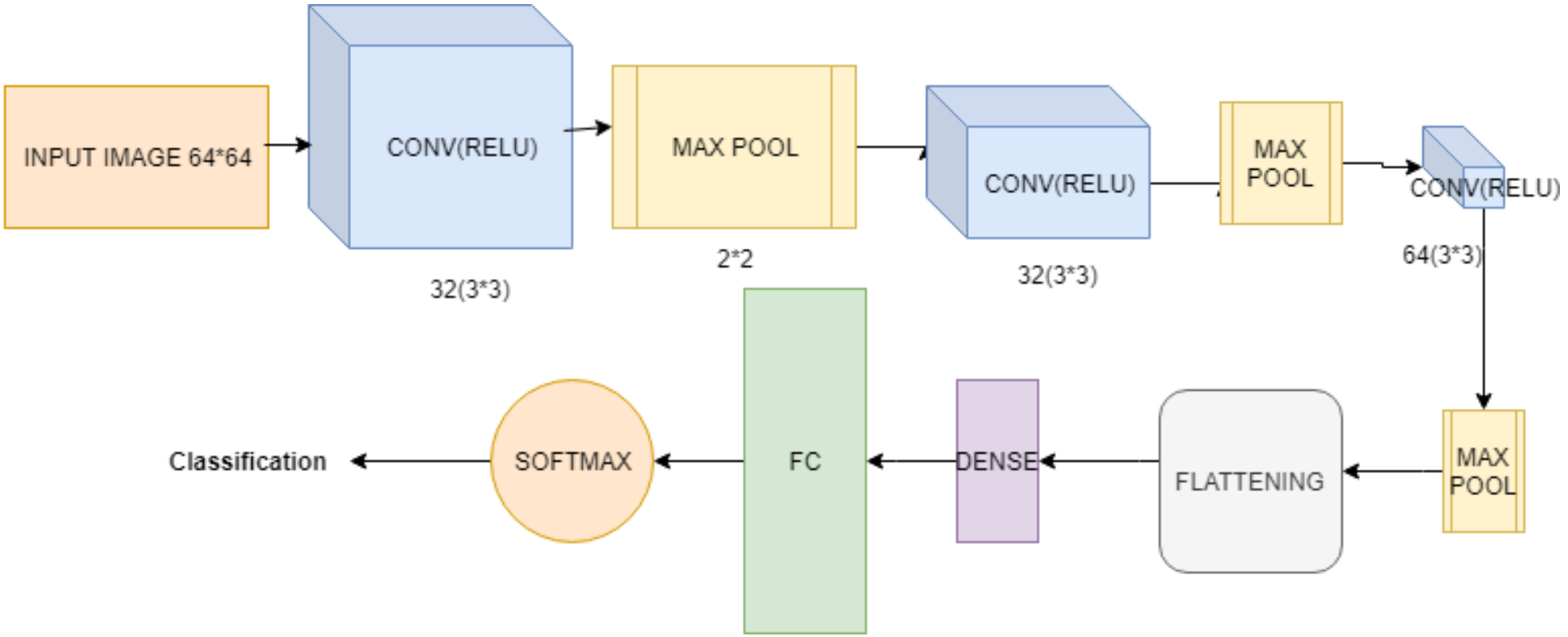
- We shrink the image stack into smaller size.
- Pooling consists of four steps:
  1. Pick a window size.
  2. Pick a stride
  3. Walk your windows across the filtered image.
  4. From each window take the maximum value.

Example: classifier.add(MaxPooling2D(poolsize =(2,2)))

Flattening and Fully Connected Layers

- **Flattening:** Flattening transforms a 2-D matrix of features into a vector of features that can be fed into a neural network or classifier.
- **Fully Connected Layer:** The neuron of preceding layers are connected to every neuron in subsequent layer. Example: classifier.add(Dense(26, activation = 'softmax'))
- **Softmax:**The softmax function is often used in the final layer of a neural network-based classifier. Such networks are commonly trained under a log loss (or cross-entropy) regime, giving a non-linear variant of multinomial logistic regression.

CNN Architecture

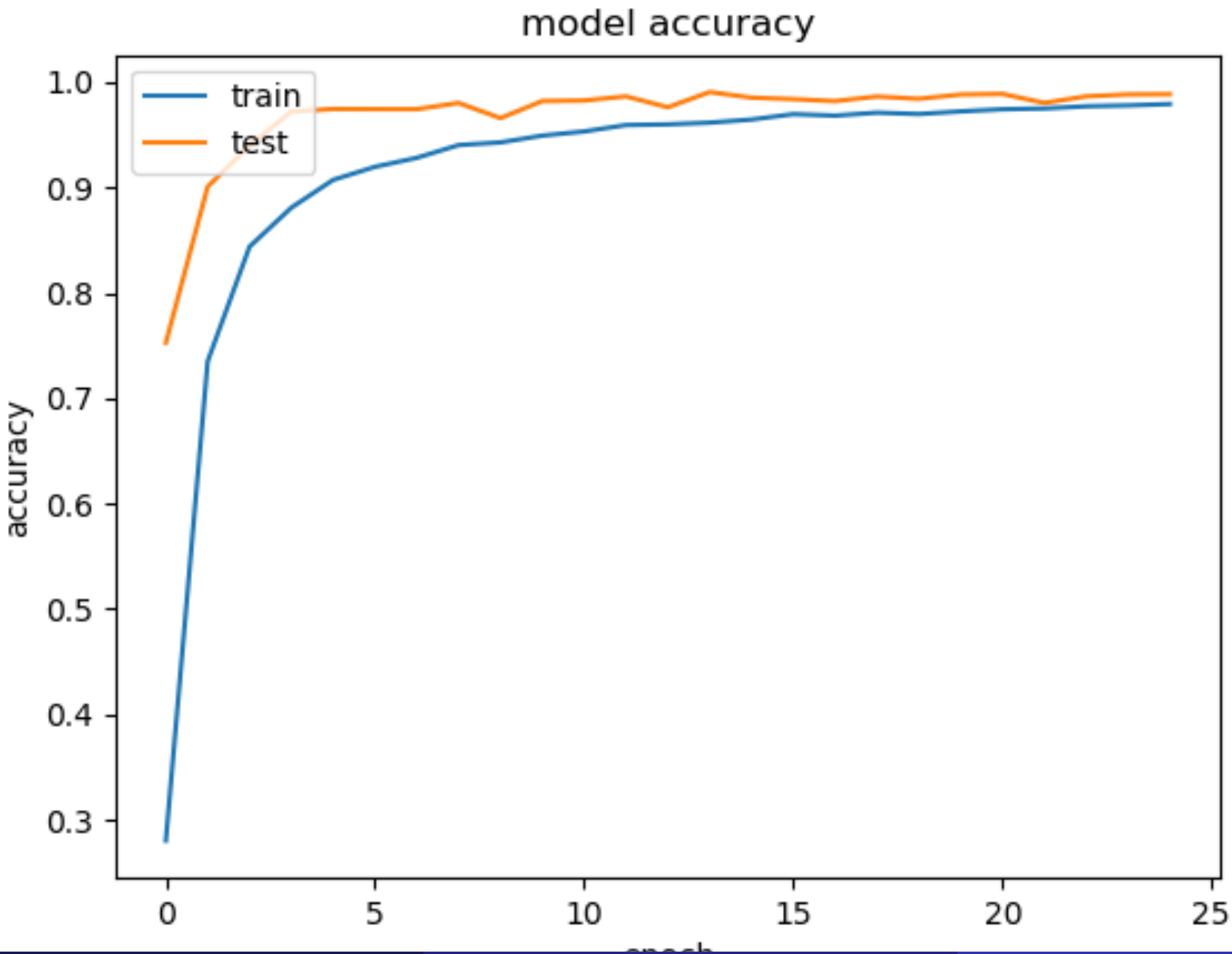


CNN Architecture

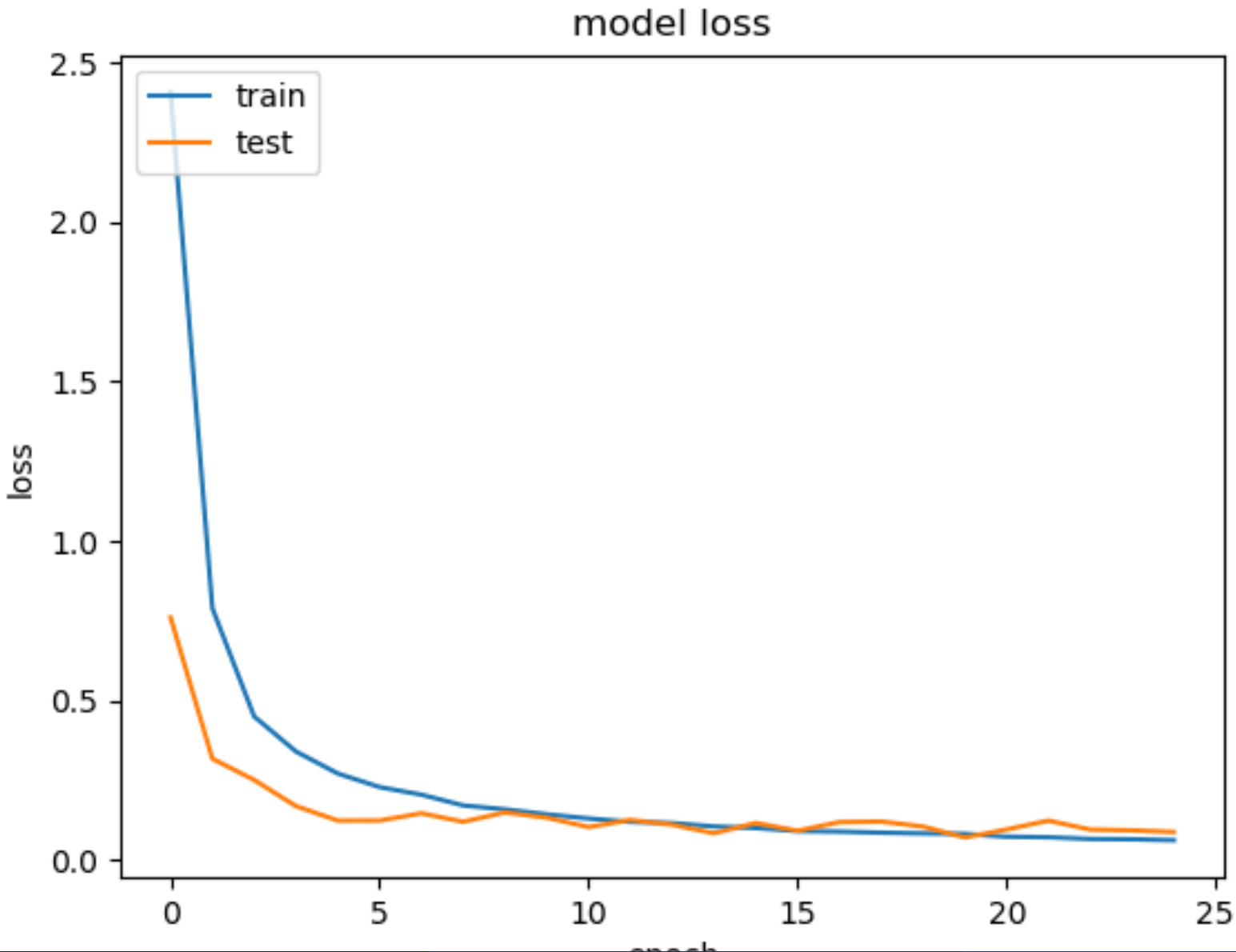
Other Implementation Details

1. **Image Data Generator:** Rescaling, Shearing and Zooming
2. **Target Image Size**= 64*64
3. **Batch Size**=32
4. **Number of epochs**=25
5. **Steps per epochs**= 8000
6. **Accuracy achieved**=0.979

Plot Of Model Accuracy



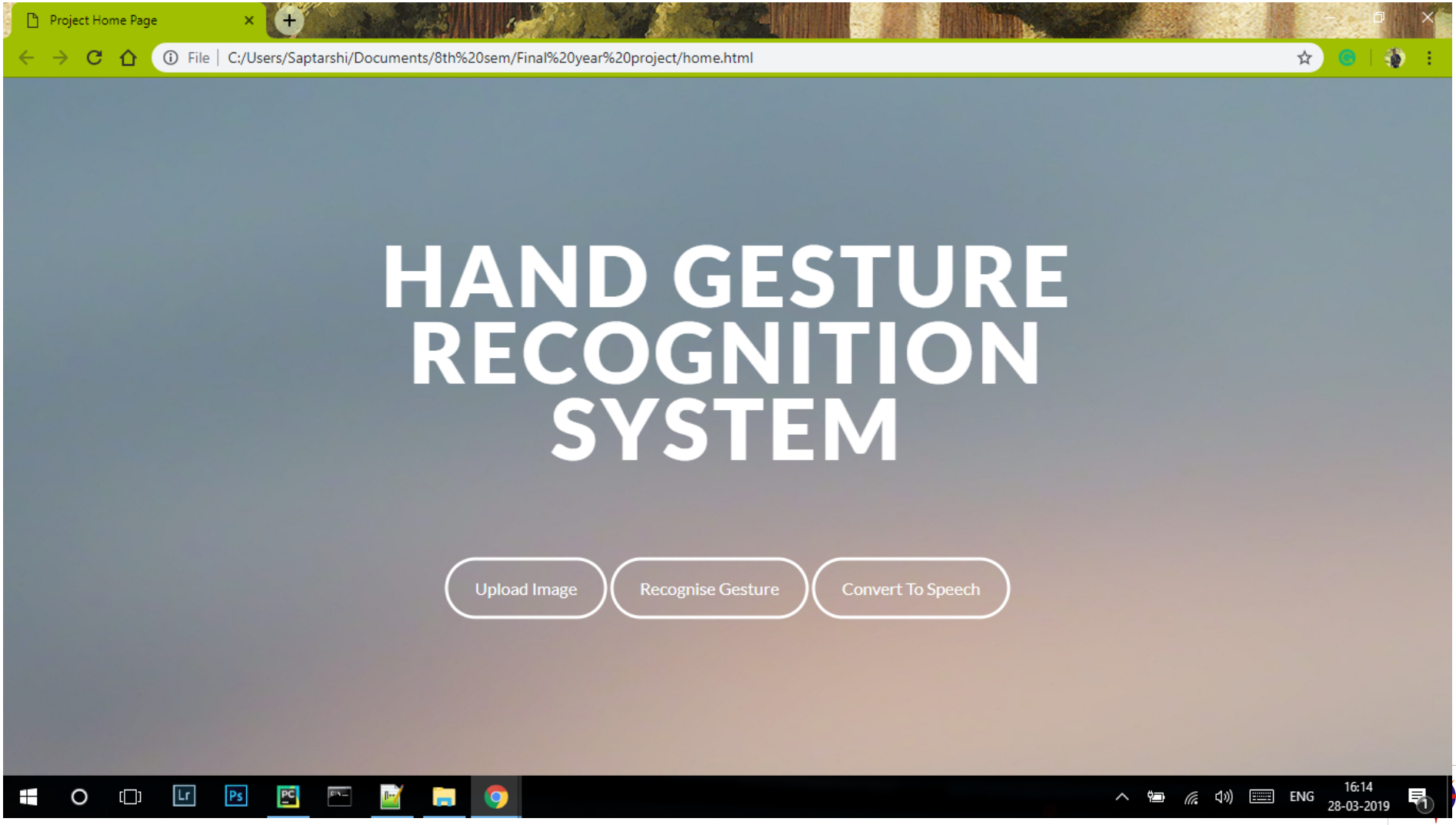
Plot Of Model Loss



User Interface Design

- A simple user friendly webpage using HTML,CSS and SCSS.
- Consists of three parts:
  1. Upload of the image
  2. Recognize the gesture
  3. Convert the gesture into speech

First glimpse of the UI



Speech Recognition and Translation

- The text generated from the classification is by default in English. Therefore we first need to convert the text into some other language according to the user choice.
- **Microsoft Azure TTS(Text To Speech)**:Text-to-speech from Azure Speech Services is a REST-based service that enables your applications, tools, or devices to convert text into natural human-like synthesized speech.
- Neural voices can be used to make interactions with chatbots and virtual assistants more natural and engaging

Hardware and Software Requirments

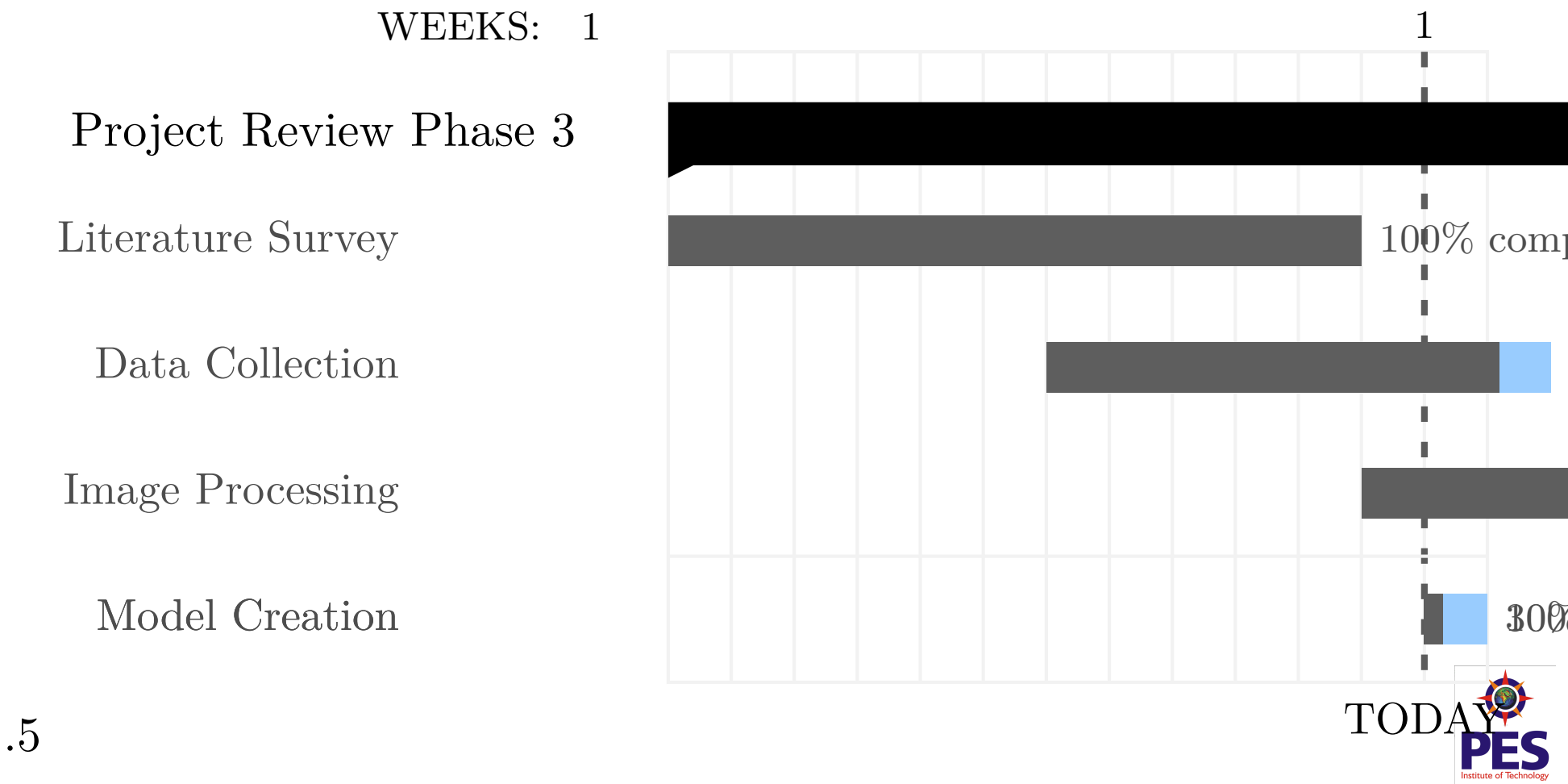
Hardware Requirements

- Processor: 2Ghz(Dual or Quad-core) or Faster processor
- Ram: 2GB(64-bit) or more
- Storage: minimum of 5GB hard disk space
- A high resolution web camera to capture images
- Speakers to get the voice output from device

Software Requirements

- Operating system: Windows 10
- Programming Languages: Python and OpenCV
- Packages used: Tensorflow, Keras, Numpy
- API: Microsoft Azure TTS

Time line of completion of project from August 2018-March 2018(Gantt Charts).



Expected Outcome/ Results

1.

Conversion of the hand gestures to the correct alphabet/symbol/meaning in the form of text.

2.

Further converting the text to speech in the language of the user's choice.

References

Ramesh M Kagalkar, Nagaraj H.N

New Methodology for Translation of Static Sign Symbol to Words in Kannada Language

International Journal of Computer Applications (0975 – 8887) Volume 121 – No.20, July 2015

Sawant Pramada,, Deshpande Saylee , Nale Pranita, Nerkar Samiksha, Archana S. Vaidya

Intelligent Sign Language Recognition Using Image Processing

IOSR Journal of Engineering (IOSRJEN) e-ISSN: 2250-3021, p-ISSN: 2278-8719 Vol. 3, Issue 2 (Feb. 2013), ||V2|| PP 45-51

V.Padmanabhan, M.Sornalatha

Hand gesture recognition and voice conversion system for dumb people

International Journal of Scientific Engineering Research, Volume 5, Issue 5, May-2014 427 ISSN 2229-5518

References

Anjali Singh,Balram Timande

Gesture to Speech Conversion in "HINDI" language for Hearing Speech disabled persons in INDIA

International Research Journal of Engineering and Technology (IRJET) Volume: 03 Issue: 06 | June-2016

Tahir Khan

Hand Gesture Recognition based on Digital Image Processing using MATLAB

International Journal of Scientific Engineering Research, Volume 6, Issue 9, September 2015 338 ISSN 2229-5518

Ashis Pradhana, M.K. Ghosea, Mohan Pradhana

A Hand Gesture Recognition using Feature Extraction

International Journal of Computer Applications (0975 – 8887) Volume 121 – No.20, July 2015

The End