Bengali Sign Language Recognition

Using Mediapipe

Project report submitted in fulfilment of the requirement for the

Degree

of

Bachelor of Technology

In

Computer Science and Engineering

By

- **ANAMIKA DEY(BT18CS016)**

Under Supervision of

- **Dr. Ranjita Das,**
  Faculty Computer Science and Engineering
  NIT Mizoram

Department of Computer Science and Engineering

National Institute of Technology Mizoram

(May 2022)

CERTIFICATE

This is to certify that the work contained in the project report titled "**Bengali Sign Language Recognition**" submitted by "**Anamika Dey**" in fulfilment of the requirement for the award of Bachelor of Technology Degree in Computer Science and Engineering at NIT Mizoram has been carried out under my supervision and that this work has not been submitted elsewhere for a degree.

Dr. Sandeep Kumar Dash
HoD & Asst. Professor
Computer Science and Engg.

Dr. Ranjita Daas
Supervisor and Asst. Professor
Computer Science and Engg.

DECLARATION

I declare that the following is my original work and I have made adequate citations and references to the original sources. I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violations can also evoke penal action from the sources which have not been properly cited or from whom permission has not been taken when needed and any violation of the above will be cause for disciplinary action by the Institute.

Anamika

Signature

Name : Anamika Dey

Enrollment No: BT18CS016

Date: 2.05.2022

Acknowledgement

The satisfaction and euphoria that accompany the successful completion of any task would be impossible without the mention of the people who made it possible, whose constant guidance and encouragement crowned our effort with success.

I have great pleasure in expressing my deep sense of gratitude to Mr. Lenin Laitonjam, our supervisor for our project for providing continuous support, guidance, and suggestions given to use in the course of our project.

I would like to thank the Department Computer Science and Engineering at the National Institute of Technology Mizoram for providing excellent education, a congenial environment and a research facility.

Finally, a note of thanks to the teaching and non-teaching staff of Dept. of Computer Science and Engineering, for their cooperation extended to me, and my friends, who helped me directly or indirectly in the course of the project work

Content

Abstract

Unfortunately, every research has its own limitations and are still unable to be used commercially. Some of the researches have known to be successful for recognizing sign language, but require an expensive cost to be commercialized. Nowadays, researchers have gotten more attention for developing Sign Language Recognition that can be used commercially. Researchers do their researches in various ways. It starts from the data acquisition methods. The data acquisition method varies because of the cost needed for a good device, but cheap method is needed for the Sign Language Recognition System to be commercialized. The methods used in developing Sign Language Recognition are also varied between researchers. Each method has its own strength compare to other methods and researchers are still using different methods in developing their own Sign Language Recognition. Each method also has its own limitations compared to other methods. The aim of this paper is to review the sign language.The aim of this Paper is to use Mediapipe, a cross-platform framework for building multimodal applied machine learning pipelines , and build a model which does not get disrupted due to the noise present in the background and shows an appropriate result

Chapter – 1

Introduction

1.1 Gesture Recognition

The gesture is known as a form of non-verbal communication or non-vocal communication where utilize of the body's movement that can convey a particular message originating from parts of the human body, the hand or face are the most commonly adopt. Gesture-based interaction introduced by Krueger as a new type of Human-Computer Interaction (HCI) in the middle 1970s has become a magnetic area of the research. In the Human-Computer-Interaction (HCI), building interfaces of applications with managing each part of the human body to communicate naturally are the great attention to do research, especially the hands as the most effective-alternative for the interaction tool, considering their ability

Through Human-Computer-Interaction (HCI), recognizing hand gestures could help achieve the ease and naturalness desired. When interacting with other people, hand movements have the meaning to convey something with its information. Ranging from simple hand movements to more complex ones. For example, we can use our hand to point something (object or people) or use different simple shapes of hand or hand movements expressed through manual articulations combined with their grammar and lexicon as well known as sign languages.

1.2 Sign Language

Sign language is an important part of communication for people with speaking and hearing disability. In the modern world, almost every spoken language has its own verified sign language which is parallel to the written language. It is an important part of linguistics for translating the sign language to the written form to communicate effectively with other audiences who do not know the signs . With the improvement of image recognition and computer speed, we can easily adopt the translation in real time and make the communication easy, convenient and fast. However, such a goal is not realized with the Bengali sign language. Considering this fact, we aim to build a Bengali sign language translator system. This will help a huge variety of audiences communicate more easily with the deaf and dumb community, allowing uniform advancement of the entire society

1.3 Sign Language Recognition

Sign Language Recognition systems are used to detect signs of numbers, alphabets, words or any other signs, e.g., hand gestures for traffic signals. Some researchers are working on real-time (video stream) sign detection and some others are working on static images. Recently, artificial neural network based methods have been employed for real-time American Sign Language (ASL) word and alphabet detection. In the paper, for hand gesture recognition using Microsoft Kinect sensor recognizes signs for two real life applications - arithmetic computation and rock-paper-scissors game showing mean accuracy above 90% for ASL. Deep learning has been used to detect signs for Indian Sign Languages, Arabic Sign Language, and other languages.

1.4 Related Works

Gesture recognition is an essential topic in computer science and builds technology that aims to interpret human gestures where anyone can use simple gestures to interact with the device without touching them directly. The entire procedure of tracking gestures to their representation and converting them to some purposeful command is known as gesture recognition [7]. Identify from explicit hand gestures as input then process these gestures representation for devices through mapping as output is the aim in hand gestures recognition.

● *High-Level Features-Based Approaches*: Aim to figure out the position of the palm and joint angles such as the fingertips, joint location, or anchor points of the palm. Whereas, effect collisions or occlusions on the image are difficult to detect after features are extracted, and sensitivity segmentation performance on 2D hand image are the problem that occurred frequently. The gestures are defined from the results with a set of rules and conditions from the vectors and joints of the hands.

● *Low-Level Feature-Based Approaches*: Utilized these features for could be extracted quickly for robust to noise. Zhou discovered recognition of the hand shape as a cluster-based signature using a novel distance metric called Finger Earth's Distance. Stanner determines the bounding region of the hand elliptically for implement hand recognition based on principal axes. Yang did research using the optical flow of the hand region as a low level feature. Low-Level Feature-Based is not efficient when cluttered background.

● *3D Reconstruction-Based Approaches*: Use the 3D model of features for achieving the construe of hand completely. Research showed that successfully segmenting the hand in skin color needs similarity and high contrast of the background related to the hand through structured light to bring in 3D of depth data. Another one uses a stereo camera to track numerous interest points of the superficies of the hand which results in difficulty for handle robust 3D reconstruction, despite data contains 3D has valuable information that can help dispose of vagueness. See for more 3D reconstruction-based approach. From kinds of literature, there are three Hand gesture recognition methods, as follow:

● *Machine Learning Approaches*: The resulting output came from the stochastic process and approach based on statistical modeling for dynamic gestures such as PCA, HMM, advanced particle filtering, and condensation algorithm.

● *Algorithm Approaches*: Collection of encoded conditions and restraints manually for defining as gestures in dynamic gestures. Galveia applied a 3rd-degree polynomial equation to determine the dynamic component of the hand gestures (create a 3rd-degree polynomial equation, recognition, reduced complexity of equations, and comparison handling in gestures library).

● *Rule-based Approaches*: Suitable for dynamic gestures either static gestures which are contained a set of pre-encoded rules and features inputs. The features of input gestures are extracted and compared to the encoding rules that are the flow of

1.5 Related works on Bengali Sign Language Recognition

Bengali Sign Language (BSL) recognition research came into focus in the last decade; different methods have been proposed to implement Bengali Sign Language Recognition system. In 2012, Karmokar et al. have proposed a method of BSL recognition using a neural network which has an accuracy of 93%. They used a dataset with unique skin colors and but with same background which allowed easy detection. A computer based Bengali sign language recognition has been

proposed by M. Rahman and others in 2014 . In 2015, Backpropagation method of ANN had been applied by Rahim et al. to detect signs of some common Bengali words. In 2015 and 2016 fingertip finder algorithms have been used for BSL by different researchers . Several methods using ANN have been proposed for detection of Bengali Sign Language. Most of the methods are showing promising accuracy. However, they have limitations of using small datasets, controlled background or scene and in some , they suffer from large error in lighting effects or skin color which they control to avoid further complicacy. The methods that do employ Neural Networks make use of complex pre-processing which may not be ideal for real-time applications.
In this paper, we focus on developing a manual user guide application with improving architecture application by applying hand gesture recognition using the MediaPipe framework.

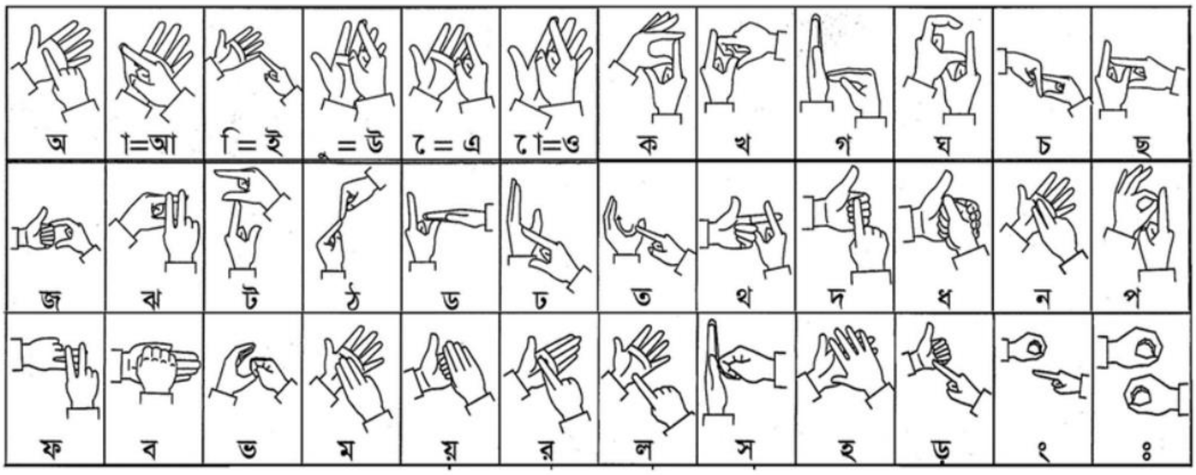


Fig1: Bengali Sign Language

Chapter -2

MediaPipe

2.1 Introduction

MediaPipe is a cross-platform framework for building multimodal applied machine learning pipelines

Currently, many frameworks or library machine learning for hand gesture recognition have been built to make it easier for anyone to build AI (Artificial Intelligence) based applications. One of them is MediaPipe. The MediaPipe framework is present by Google for solving the problem using machine learning such as Face Detection, Face Mesh, Iris, Hands, Pose, Holistic, Hair segmentation, Object detection, Box Tracking, Instant Motion Tracking, Objection, and KIFT. MediaPipe framework helps a developer focus on the algorithm and model development on the application, then support environment application through results reproducible across different devices and platforms which it is a few advantages of using features on the MediaPipe framework

The MediaPipe is a framework designed to implement production-ready machine learning that must build pipelines to perform inference over arbitrary sensory data, has published code accompanying research work, and build technology prototypes . In MediaPipe, graph modular components come from a perception pipeline along with the function of inference model function, media processing model, and data transformations. Graph of operations are used in others machine learning such as Tensor flow , MXNet, PyTorch, CNTK, OpenCV 4.0.

Using MediaPipe for hand gesture recognition has been researched by Zhang before, using a single RGB camera for AR/VR application in a real-time system that predicts a hand skeleton of the human. We can develop a combined MediaPipe using other devices. The MediaPipe implements pipeline in Figure 2.1. consists of two models for hand gesture recognition as follows:

1. A palm detector model processes the captured image and turns the image with an oriented bounding box of the hand,

2. A hand landmark model processes on cropped bounding box image and returns 3D hand key points on hand.

3. A gesture recognizer that classifies 3D hand key points then configuration them into a discrete set of gestures.

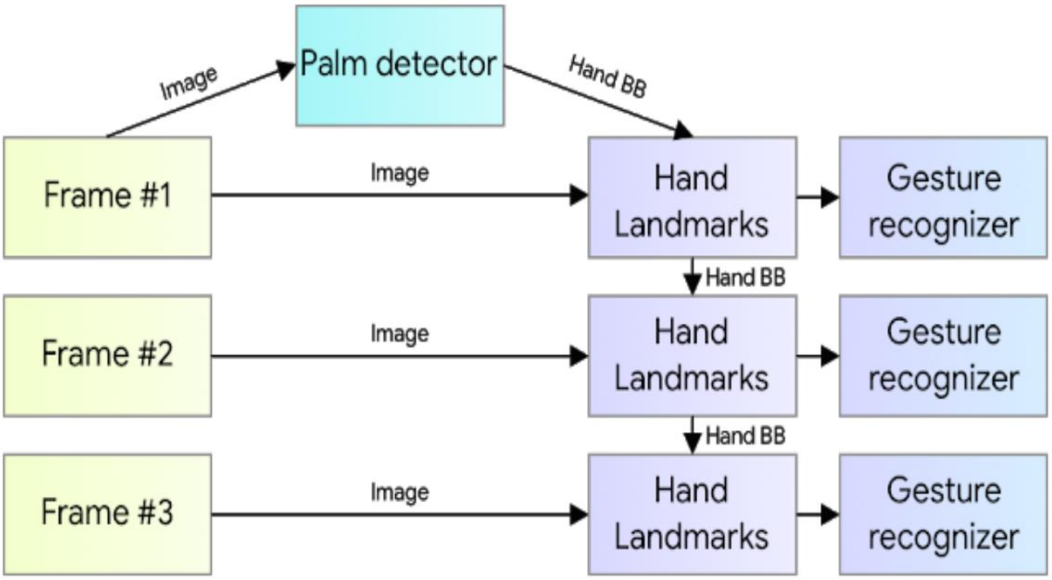


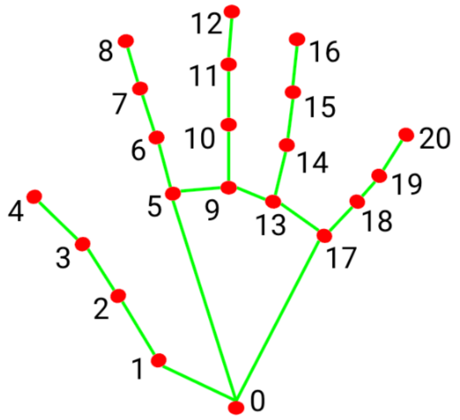
Fig.2.1 Hand Perception Pipeline Overview

2.2 Palm Detector Model

MediaPipe framework has built detect initial palm detector called BlazePalm. Detecting the hand is a complex task. Step one is to train the palm instead of the hand detector, then using the non-maximum suppression algorithm on the palm, where it is modeled using square bounding boxes to avoid other aspect ratios and reducing the number of anchors by a factor of 3-5. Next, encoder-decoder of feature extraction that is used for bigger scene context-awareness even small objects, lastly, minimize the focal loss during training with support a large number of anchors resulting from the high scale variance .

2.3 Hand Landmark

Achieves precise key point localization of 21 key points with a 3D hand-knuckle coordinate which is conducted inside the detected hand regions through regression which will produce the coordinate prediction directly which is a model of the hand landmark in MediaPipe, see in Figure 2.



0. WRIST
1. THUMB_CMC
2. THUMB_MCP
3. THUMB_IP
4. THUMB_TIP
5. INDEX_FINGER_MCP
6. INDEX_FINGER_PIP
7. INDEX_FINGER_DIP
8. INDEX_FINGER_TIP
9. MIDDLE_FINGER_MCP
10. MIDDLE_FINGER_PIP
11. MIDDLE_FINGER_DIP
12. MIDDLE_FINGER_TIP
13. RING_FINGER_MCP
14. RING_FINGER_PIP
15. RING_FINGER_DIP
16. RING_FINGER_TIP
17. PINKY_MCP
18. PINKY_PIP
19. PINKY_DIP
20. PINKY_TIP

Fig.2.2 Hand Landmarks (21 landmarks)

Each hand-knuckle of the landmark has coordinate is composed of x, y, and z where x and y are normalized to [0.0, 1.0] by image width and height, while z representation the depth of landmark. The depth of landmark that can be found at the wrist being the ancestor. The closed the landmark to the camera, the value becomes smaller.

2.4 Hand Landmark Model

After the palm detection over the whole image our subsequent hand landmark model performs precise keypoint localization of 21 3D hand-knuckle coordinates inside the detected hand regions via regression, that is direct coordinate prediction. The model learns a consistent internal hand pose representation and is robust even to partially visible hands and self-occlusions.

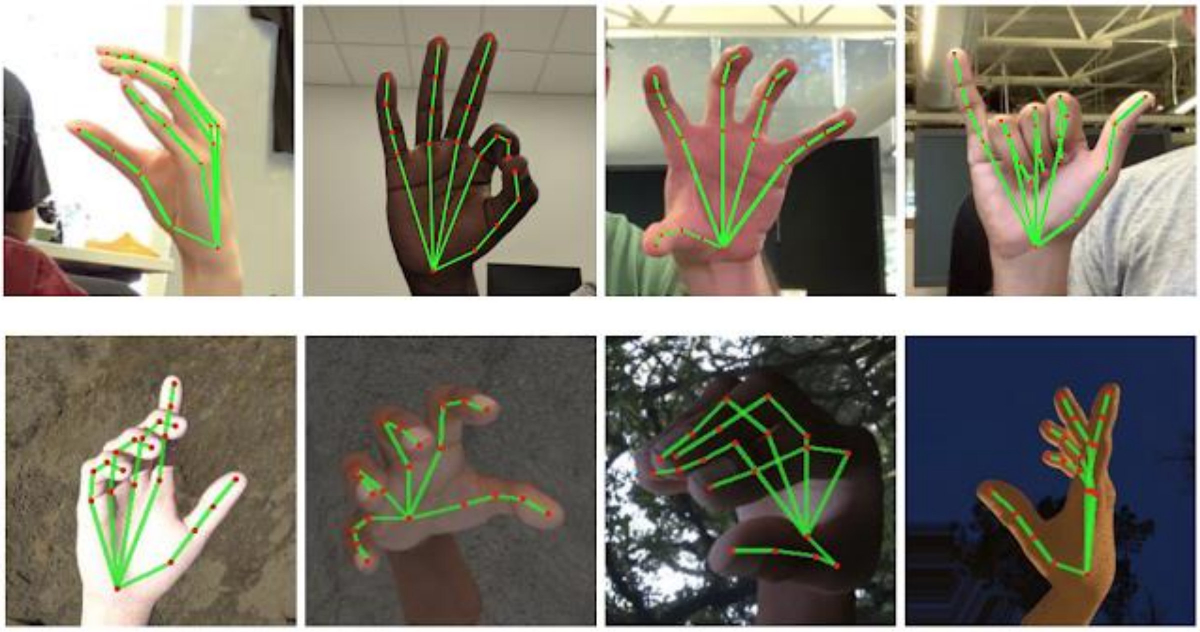


Fig. 2.3 Example of different hand posture

Chapter -3

Database Description

The Bengali Language has 51 letters ,which can be represented by 36 signs by grouping similar sounding alphabets into one single sign. The hand Sign of all the grouping are shown in Fig 1.

The Dataset comprises of 36 Hand Sign of Bengali Letters .Each Letters have a total of 40 videos ,where each videos is a collection of 50 frames. There are a Total of 1440 videos(36 letters * 40 videos each).

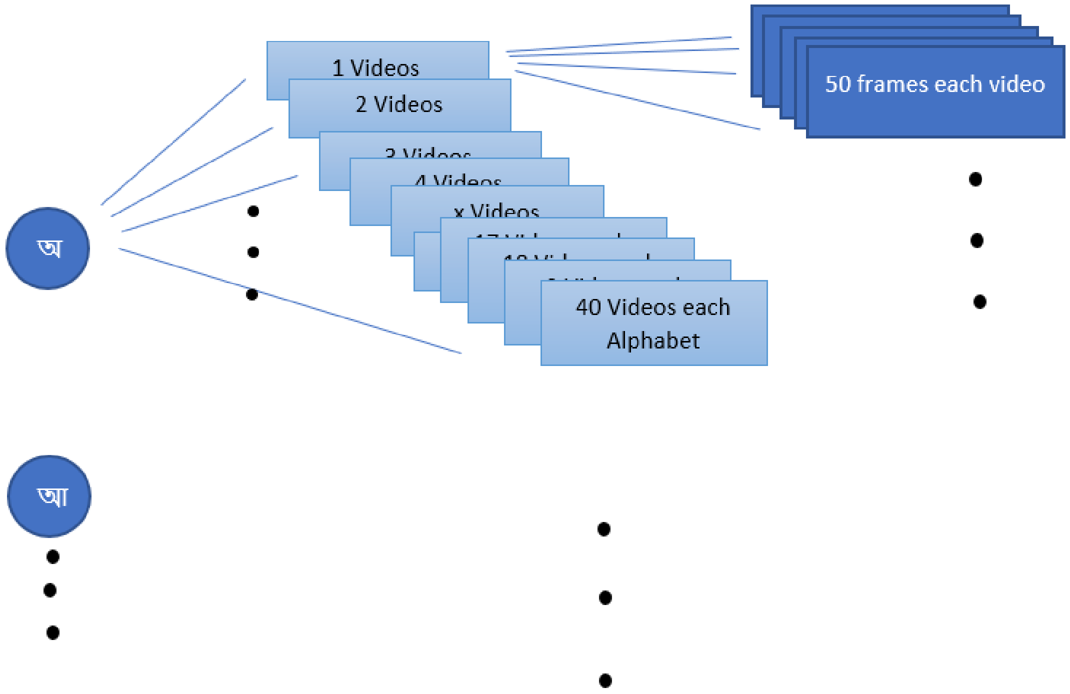


Fig. 3.1: Sample Structure of the database

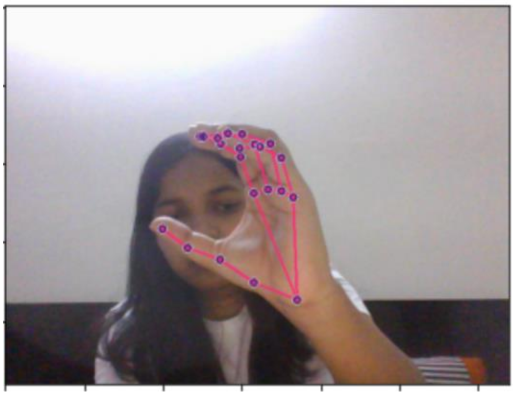
The Alphabets are labeled in the database as shown in the Table

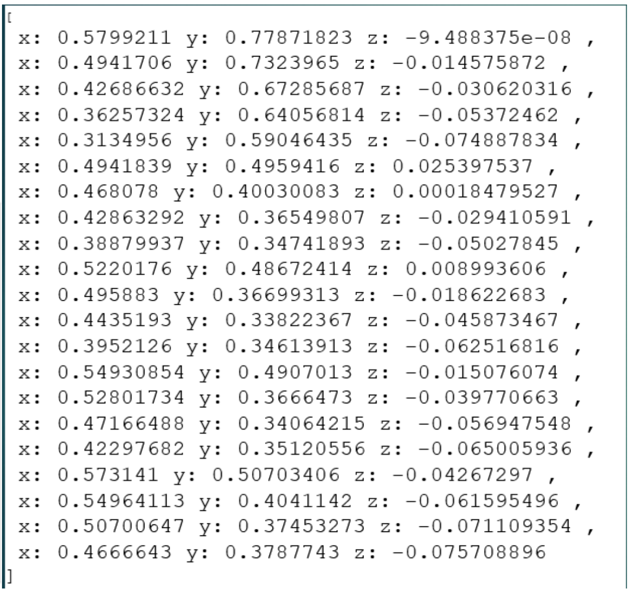
Table 1. The Alphabets and their Labels

Alphabets	Labels	Alphabets	Labels	Alphabets	Labels
অ	A	জ	Ja	ফ	Pha
আ	Aa	ঝ	Jha	ব	Ba
ই	I	ট	Tta	ভ	Bha
উ	U	ঠ	Ttha	ম	Ma
এ	E	ড	Dda	য	Ya
ও	O	ঢ	Ddha	র	Ra
ক	Ka	ত	Ta	ল	La
খ	Kha	থ	Tha	স	Sa
গ	Ga	দ	Da	হ	Ha
ঘ	Gha	ধ	Dha	ড়	Rra
চ	Ca	ন	Na	◌ং	Sign1
ছ	Cha	প	Pa	◌ঃ	Sign2

Since the project is based on Sign Language Recognition, MediaPipe Hand Landmarks are being used in

Each hand contains 21 landmarks as shown in fig . Thus for both the hands there will be a total of 42{21(left hand)+21(right hand)} landmarks. Each landmark contains values of x, y, z-axis.





```
[
 x: 0.5799211 y: 0.77871823 z: -9.488375e-08 ,
 x: 0.4941706 y: 0.7323965 z: -0.014575872 ,
 x: 0.42686632 y: 0.67285687 z: -0.030620316 ,
 x: 0.36257324 y: 0.64056814 z: -0.05372462 ,
 x: 0.3134956 y: 0.59046435 z: -0.074887834 ,
 x: 0.4941839 y: 0.4959416 z: 0.025397537 ,
 x: 0.468078 y: 0.40030083 z: 0.00018479527 ,
 x: 0.42863292 y: 0.36549807 z: -0.029410591 ,
 x: 0.38879937 y: 0.34741893 z: -0.05027845 ,
 x: 0.5220176 y: 0.48672414 z: 0.008993606 ,
 x: 0.495883 y: 0.36699313 z: -0.018622683 ,
 x: 0.4435193 y: 0.33822367 z: -0.045873467 ,
 x: 0.3952126 y: 0.34613913 z: -0.062516816 ,
 x: 0.54930854 y: 0.4907013 z: -0.015076074 ,
 x: 0.52801734 y: 0.3666473 z: -0.039770663 ,
 x: 0.47166488 y: 0.34064215 z: -0.056947548 ,
 x: 0.42297682 y: 0.35120556 z: -0.065005936 ,
 x: 0.573141 y: 0.50703406 z: -0.04267297 ,
 x: 0.54964113 y: 0.4041142 z: -0.061595496 ,
 x: 0.50700647 y: 0.37453273 z: -0.071109354 ,
 x: 0.4666643 y: 0.3787743 z: -0.075708896
]
```

Fig3.2 Key points of hand landmark corresponding in coordinate (x,y,z) in MediaPipe for one hand gesture

Chapter-4
Workflow Method

4.1 WorkFlow

The Sign Language Recognition Prototype is a real-time vision-based system whose purpose is to recognize the Portuguese Sign Language given in the alphabet of Fig. 1. The purpose of the prototype was to test the validity of a vision-based system for sign language recognition and at the same time, test and select hand features that could be used with machine learning algorithms allowing their application in any real-time sign language recognition systems. For that, the user must be positioned in front of the camera, doing the sign language gestures, that will be interpreted by the system and their classification will be based on the keypoints extracted in realtime.

Mediapipe helps us Collect the values of the Keypoints from each frames in the video and store it in the DataSet. MediaPipe will read the image that received in real-time, then on an image will do palm detection and make hand landmark that made return 3D hand key points and joint it make up like skeleton. 3D key points in the palm that has been marked in the image will be computed and initialized as a tool for reading pose hand and recognition based that will be conveyed information based on hand pose dataset . To identify the poses of the hand differently could be calculated using 42 key points on hand landmarks explained in Section 2.2 above. This identifying could be done with, firstly, determine it will determine the different position of the landmark of that particular sign , and then make the prediction which matches the maximum.

This Dataset is then split into a training and a Testing Dataset . In the Project we have used a 50 % of the dataset for Training purpose and the remaining dataset for Testing Purpose.

An LSTM model has been used in this project. LSTM model is special kind of recurrent neural network that is capable of learning long term dependencies in data. This is achieved because the recurring module of the model has a combination of four layers interacting with each other. It is a model that increases the memory of recurrent neural networks. Recurrent neural networks hold short term memory in that they allow earlier determining information to be employed in the current neural networks Let’s say while watching a video you remember the previous scene. Similarly RNNs work, they remember the previous information and use it for processing the current input.

The project is designed a develop a model for recognising the Bengli Sign Language using MediaPipe, for which the block diagram is shown in Figure .

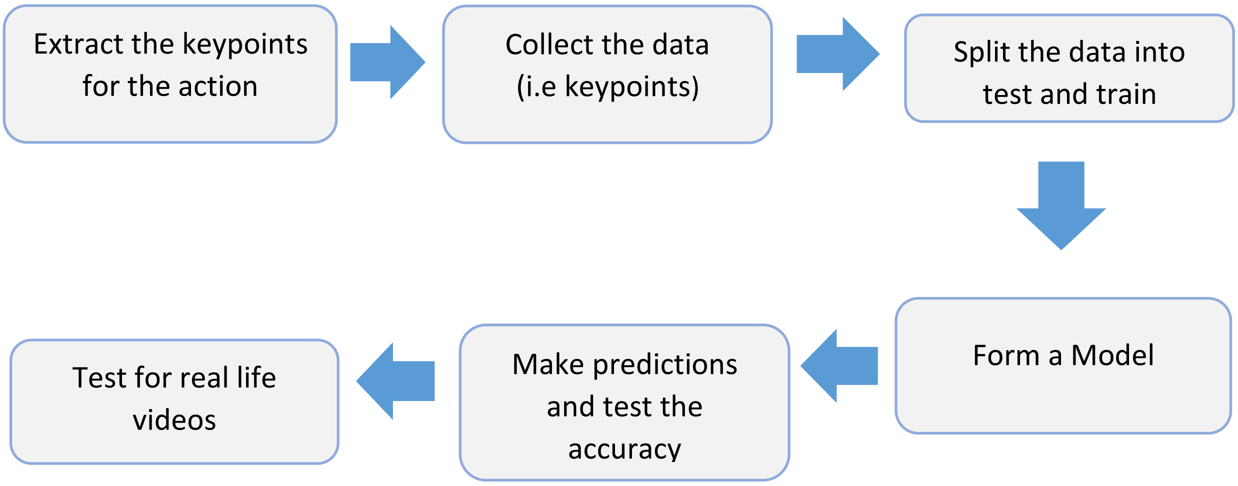

Extract the keypoints for the action
Collect the data (i.e keypoints)
Split the data into test and train
Form a Model
Make predictions and test the accuracy
Test for real life videos

Fig 4.1: Functional Block Diagram

4.2 Result

The measure of the performance on the model in machine learning used Confusion Matrix. In Python, we can use the library scikit-learn to develop a confusion matrix. Experiment datasets were obtained before we used them to predict the hand gestures. The Confusion Matrix was also used to observe an accuracy achieved for the model was made.

The graph for the categorical accuracy and the loss in each epoch is given below

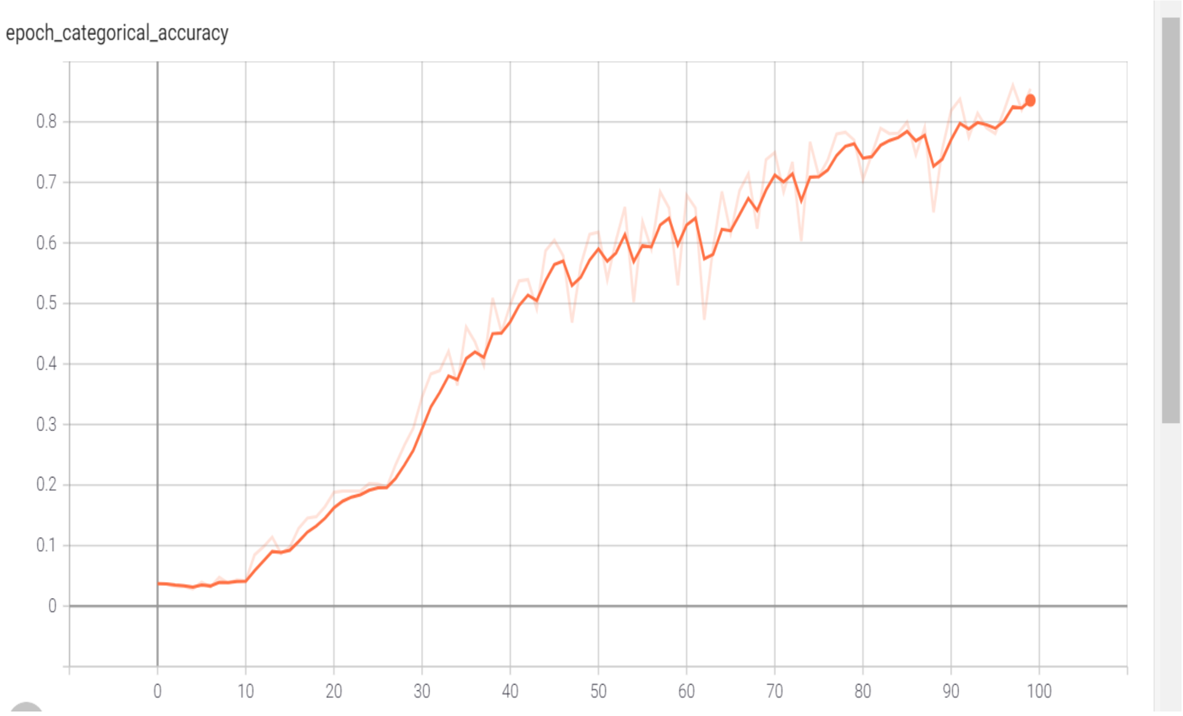


Fig 4.2 Epoch categorical accuracy

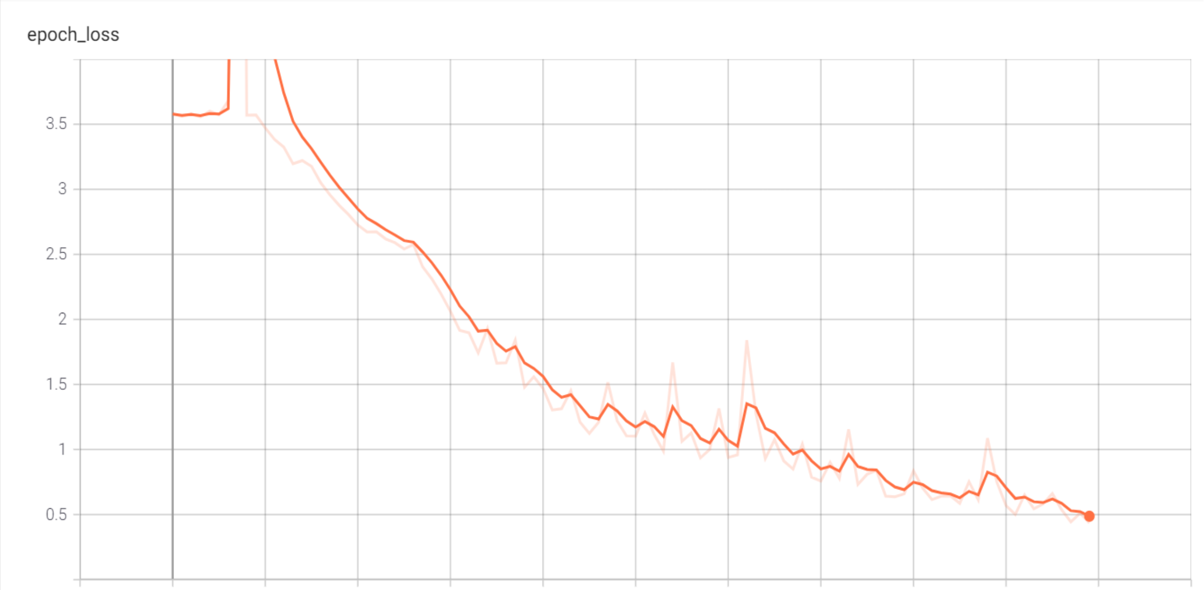


Fig 4.3 Epoch loss

Chapter -5

Conclusion

The Hand gesture recognition system has become an important role in building efficient human-machine interaction. Implementation using hand gesture recognition promises wide-ranging in technology industry. The MediaPipe as one framework based on machine learning plays an effective role in developing this application using hand gesture recognition, with the result has shown an accuracy performance of 83%.

Reference

[1] “Demand for Sign Language Interpreters Expected to Rise Nearly 50%”, A Sharp, http://wqad.com/2015/10/21/demand-for-sign-language-interpreters-expected-to-rise-nearly-50/, accessed 5/17/18

[2] Z. Ren, J. Yuan, J. Meng and Z. Zhang, "Robust part-based hand gesture recognition using kinect sensor," IEEE Trans. Multimedia, vol. 15, no. 5, pp. 1110–1120, August 2013.

[3] Padmavathi . S , Saipreethy.M.S, Valliammai.V, “Indian Sign Language Character Recognition using Neural Networks”, International Journal of Computer Applications, Special Issue on 45, RTPRIA 2013,

[4] N.Priyadharsini, N.Rajeswari ,“Sign Language Recognition Using Convolutional Neural Networks ”, International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 Volume: 5 Issue: 6, pp :625-628, June 2017

[5] M.A Hossen , Arun Govindaiah , Sadia Sultana , Alauddin Bhuiyan Bengali Sign Language Recognition Using Deep Convolutional Neural Network

[6] O Al-Jarrah, Ali Mohammd Shatnawi, Alaa Halawani, networks”, Artificial Intelligence and Soft Computing, Spain, August 28-30, 2006.

[7] B. C. Karmokar, K. M. R. Alam and M. K. Siddiquee, "Bangladeshi sign language recognition employing neural network ensemble," Int'l Journal of Computer Applications (IJCA), vol. 58, no. 16, pp. 43–46, November 2012.

[8] M. A. Rahaman, M. Jasim, M. H. Ali and M. Hasanuzzaman, "Real-time computer vision-based Bengali sign language recognition," in 17th Int'l Conf. on Computer and Information Technology, Daffodil International University, Dhaka, Bangladesh, December 2014.

[9] Md. Abdur Rahim, Tanzillah Wahid, Md. Khaled Ben Islam, Neural Network”, International Journal of Computer Applications, Volume 94 - Number 17, pp : 1-5 , 2014