VISION BASED REAL-TIME INDIAN SIGN LANGUAGE TRANSLATOR

[1]Sankalp Doshi, [2]Ruturaj Joshi, [3]Swaroop Chavan, [4]Pranav Burli, [5]Saroja Kulkarni

[1]Student, [2] Student, [3] Student, [4] Student, [5] Assistant Professor,
[1]Department of Information Technology,
[1]Vishwakarma Institute Of Information Technology, Pune, India

Abstract: A sign language translator in a real-time is an important climacteric in communication between the deaf-mute community and normal public. Speech-Impaired people communicate among themselves using Sign Language but to normal people, it is quite challenging to understand. Therefore, we hereby present the approach which is developed with the help of visual-based method. We utilize MobileNets (open source model for efficient On-Device vision) to apply transfer learning to the supervised machine learning model for image classification. The machine learning model is build using Tensorflow and image processing is done with the help of a python library named OpenCV. The purpose of this project is to develop a web application for Sign Language translation which will allow a normal person to translate Indian Sign Language (ISL) in a textual format which includes displaying most probable word from the series of output alphabet and vice versa.

KEYWORDS **- MobileNet, Tensorflow, Indian Sign Language, Machine Learning, Image Processing, OpenCV, Python.**

I. INTRODUCTION

Communication among two individuals is a basic part of the social fabric of the society. However some individual unfortunately have to achieve this the hard way. Among them are the people of the speech impaired community. However they also need to interact with the people who do not have this impairment. But there is a problem in this case in which a normal person will feel very hard to understand the speech impaired person's sign language which they use to communicate. Due to this there exists a possibility that these people may get isolated at workplaces or other public forums.

The main aim of the project is to identify the alphabet in Indian Sign Language (ISL) from the input video using a web camera. The Gesture recognition and sign language recognition is a well-researched topic for ASL, and very less research work is published regarding Indian Sign Language (ISL). It is not convenient to always use a glove or Kinect to communicate and hence this idea came up with solving this problem using computer vision and CNN.

This project includes four tasks to be done in real time:

1. Taking input of real-time video consisting of Indian Sign Language (input).
2. Distinguish each frame from the video input.
3. Predicting and displaying most likely word from series of alphabets.
4. Acquiring text input from the user and to display its respective ISL gestured animation.

II. RELATED WORK

A. Gesture Recognition

Sign Language recognition using gesture as input is old computer vision problem and can be solved now in this new era. Researchers working in this field have tried using various different approaches like data-glove approach which included different sensors in it, vision-based approach along with different classifiers which could be grouped into Bayesian networks, neural networks, and linear classifiers.

The way the data-glove approach works is that it has multiple sensors attached fabricated onto the glove which it utilizes to recognize the hand-gesture. The way it achieves this is by capturing the data based on hand orientation, finger movements, finger count and comparing it to the data on the existing hand gestures. Many commercially used sign language translator systems are using this as its implementation is easy and obtaining data on movements and 3-D projection of your hand. The framework is fast and requires minimum computational power, and continuous translation of hand gestures in the real-time is easier.

In the visual-based approach, real-time images of the hand gestures are captured by a camera and image processing is done on it to identify the most appropriate textual meaning corresponding to the hand gesture. In contrast to the data-glove approach, the visual-based approach is much more flexible in its pre-requisites in its framework for sign-language translation. With devices, today and their high-computational power and processing capabilities visual-based approach is a domain which is highly promising to explore.

Table 2.1: Analysis of different approaches to detect sign language

Approach	Hardware/ components	Gestures	Conditions	Accuracy
Data-glove approach	Motion sensor, Microcontroller, Accelerometer.	Fingerspelling (alphabets and numbers), Sign gestures.	No environmental concern.	Increase accuracy with the addition of sensors
Visual-based approach	Custom glove, Device camera.	Only for fingerspelling (alphabets and numbers)	Background condition, Camera position, and Intensity of light	Misinterpretation is possible
Virtual-based button approach	Optic sensor	Finger movements, Not suitable for moving gestures	No environmental concern.	Overall correctness 88.82%

Based on the above summarize study of different approaches we choose the Visual-based approach as it has several advantages as follows:

1. As discussed in the above few methods, one of the methods requires to wear custom hand-glove to gather data in real-time. This custom hand-glove includes expensive sensors and this will hamper its application and use in real-life.
2. The method which is selected won't hamper real-life application and work with great efficiency.

B. Image Classification Method

Researchers working in the field of image classification have used vision-based approach along with different classifiers which could be grouped into Bayesian networks, neural networks, and linear classifiers. It is very easy to work with Linear Classifier as they are comparatively simple models, they need highly developed for drawing out features and the method of preprocessing to become successful. Das and Singha have obtained an accuracy of about 96% of one hand gesture using Karhunen-Loeve Transforms with ten classes for images. Depending upon the movement of the axes and variance of data, a new coordinate system is built.

The transformation is used after removal of objects other than skin using HSV, cropping of hand and detecting an edge on the frames. The use of linear classifier is done to differentiate which include index finger pointing, thumbs up and digits. Sharma et al. use piece-wise classifiers (SVM and k-Means) to separate each colored finger after noise removal and background subtraction. The innovation came up with contour trace (a type of representation of hand contours). They accuracy achieved was of about 62.3% using a Support Vector Machine on the segmented finger model.

Bayesian networks, for example, Hidden Markov Models can also be considered. Unlike other models, even this model achieved good accuracy. But they need a very clearly defined model which needs to be defined prior to learning.

Suk et al. propose an idea to recognize hand gestures in a real-time continuous live video stream using a dynamic Bayesian network. The attempt was to classify Sign Language Translator. They acquired accuracy of over 99%, but it is not required that all sign has a unique gesture to get identified.

III. PROPOSED SYSTEM

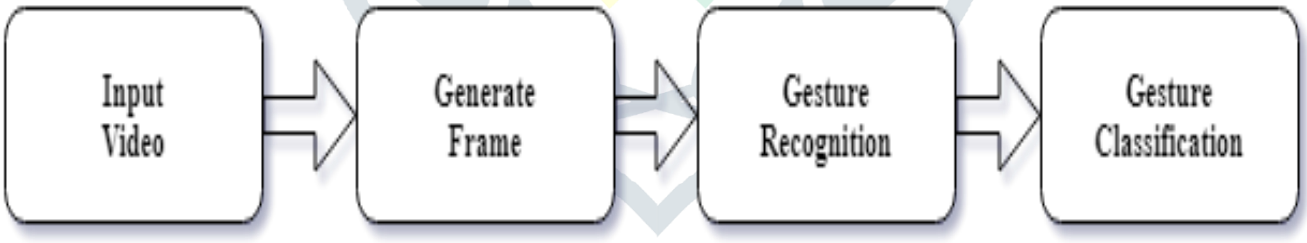

Fig 3.1: Block Diagram of Proposed System

A. Input Video

The input to the web application will be in video format. A webcam is used to acquire video frames from the signer in a type of still images and video streams in RGB (red-green-blue) shading.

B. Generate Frame

Using OpenCV (version 4.0.1), which is a python library, is used to convert the streaming video into an image of datatype uint8 (-256,256) type which can be further useful for gesture recognition.

C. Gesture Recognition

The frame is taken as input for this stage. The image containing background and unwanted noise in all included as input. This stage removes all the noise and converts the frame as a black and white image to get classified. Hand detection phase consists of four stages as follows:

i. **Background Subtraction**: This is the first stage of gesture recognition. The foreground moving object is only captured in this stage. The OpenCV creates a model where it relatively subtracts the background image from the input frame.

The Background Subtraction consists of two stages as follows: a] Background Initialization and b] Background Update. In Background Initialization, the initial frame in considered and saved in the background model. And then for all other frames coming as input, this model helps to subtract the input frame from the background frame and gives a foreground object.

ii. **Skin Detection:** After Background Subtraction step, HSV is used to eliminate the moving objects in the frame other than the hand. The HSV value of the skin color recognized ranges from 308, 89, and 32 to 322, 98, and 40 respectively.

iii. **Blur the Image:** By blurring, we create a smooth transition from one color to another and reduce the edge content. The colored frame of the skin will be blurred and converted to its respective grayscale.

iv. **Threshold the Image:** We use thresholding for image segmentation, to create binary images from grayscale images. The threshold value is considered, if the pixel value is over the threshold, the pixel will become white or else into the black.

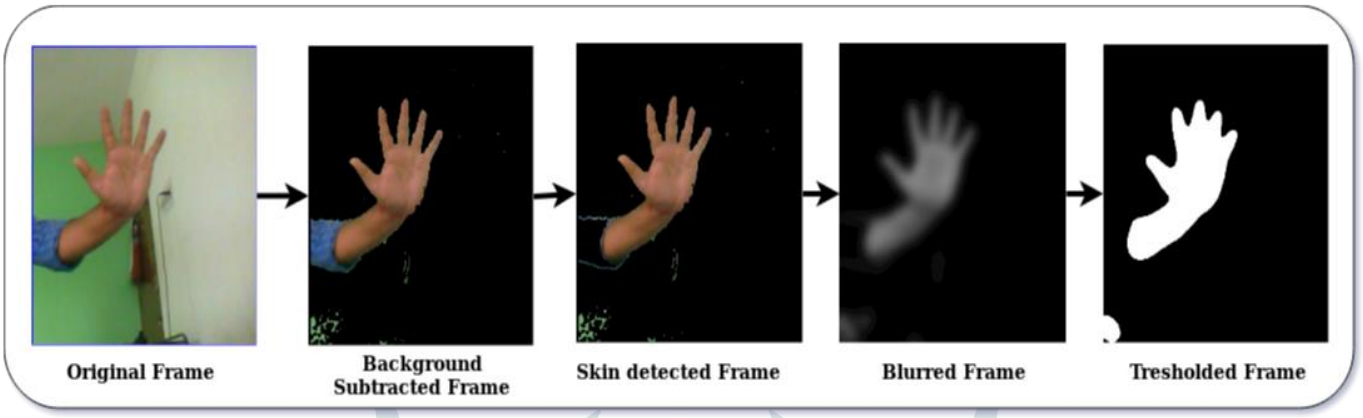

Fig 3.2: Stages of Gesture Recognition

D. Gesture Classification

Our Indian Sign Language classification of a letter is done with the help of transfer learning using a CNN (a machine learning algorithms which has achieved incredible results while handling tasks related to processing images and videos).
For classification, we will be using MobileNet (open source model for efficient On-Device vision) to apply transfer learning to this task. Mobile Net is a small efficient convolutional neural network.
MobileNet CNN has the following features:

i. This architecture utilizes depth-wise separable convolutions which significantly reduces the number of parameters when compared to the network with normal convolutions with the same depth in the networks
ii. The normal convolution is replaced by depth wise separable convolution.
iii. With the help of depth-wise separable convolutions, there is a slight decrease in the accuracy for a simple deep neural network.

IV. CHALLENGES

The research in Indian Sign Language (ISL) is much behind than that of American Sign Language because of the lack of standard datasets in ISL. Indian Sign Language (ISL) uses both hands which leads to feature obstruction. The Indian Sign Language (ISL) changes with the locality and hence there exists many signs for the same character. With the change in the area, the sign notation changes in India. The standard in dataset doesn't exist. Also, a problem is faced where a single gesture can be used to denote different letters (e.g. 2 and V can be denoted with the same finger gesture).

The challenges that will be faced during implementation according to the computer vision perspective are as follows:

i. **Environmental concerns** such as background, lighting sensitivity and camera position
ii. **Occlusions** like if Region of Interest i.e. if the hand is out of view, or some of the fingers out of view
iii. **The Boundary of Sign detection** (During the transition from one sign to other)
iv. **Co-articulation** (when there is an impact of succeeding or preceding sign)

V. DATASET

The initial dataset for the hand-gestures currently encompasses the alphanumeric system i.e. the letters from A-Z and numbers from 0-9. The images are contained within an individual subdirectory corresponding to a single alphanumeric element.

As there are not very relevant datasets related to Indian Sign Language (ISL) we decided to create one of our own. The dataset will be expanded as more complex hand-gestures are incorporated.

The user of the sign-language translator will have the functionality to add a label to a new hand-gesture not currently existing in the database. A new subdirectory will be created under the main directory for the user inputted label.

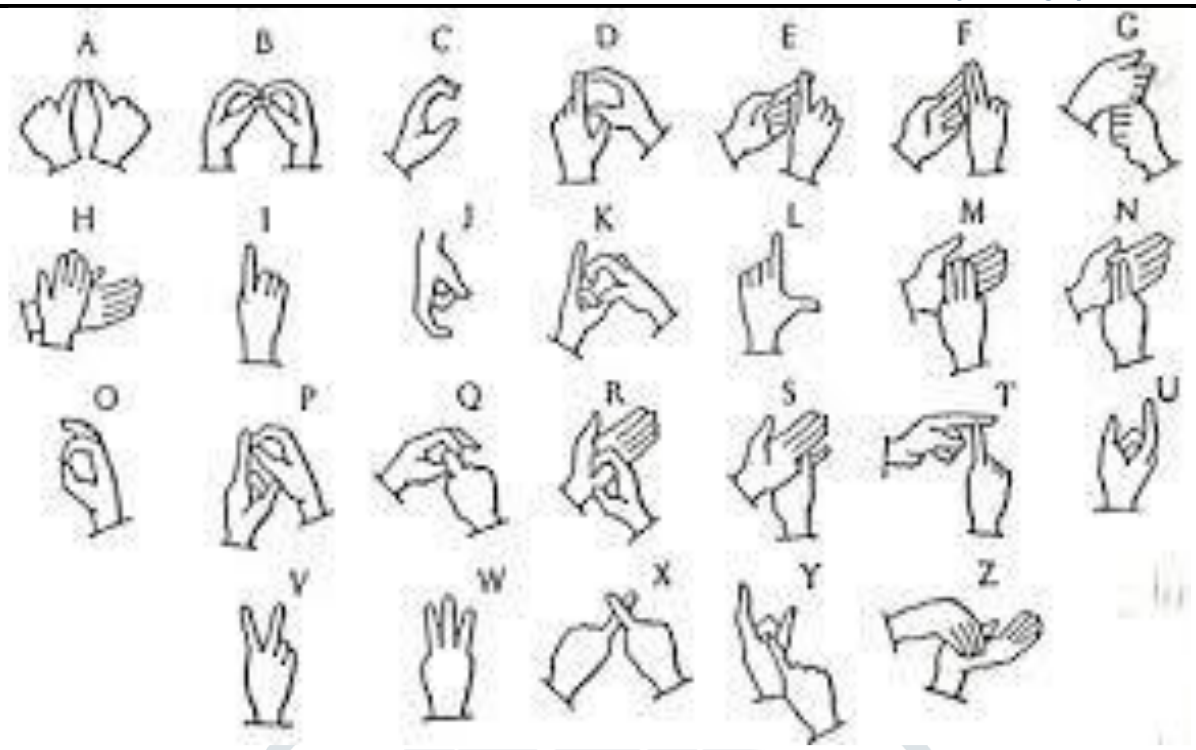

Fig 5.1: Indian Sign Language

VI. RESULT ANALYSIS

i. As per the above discussed, the model developed displays the most probable five letters. In the below fig:3.4, the input frame is provided denoting letter "C" of ISL and it recognized with the machine learning model with an accuracy of 95.32%.

ii. The time estimated to recognize the letter is of about 0.152 sec.

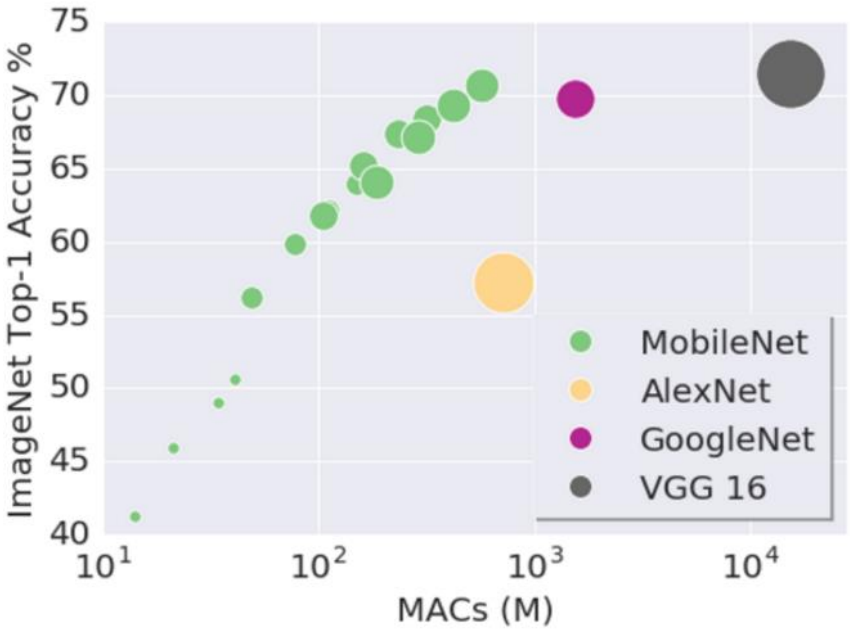

Fig 6.1: MobileNets trade off accuracy with other popular models

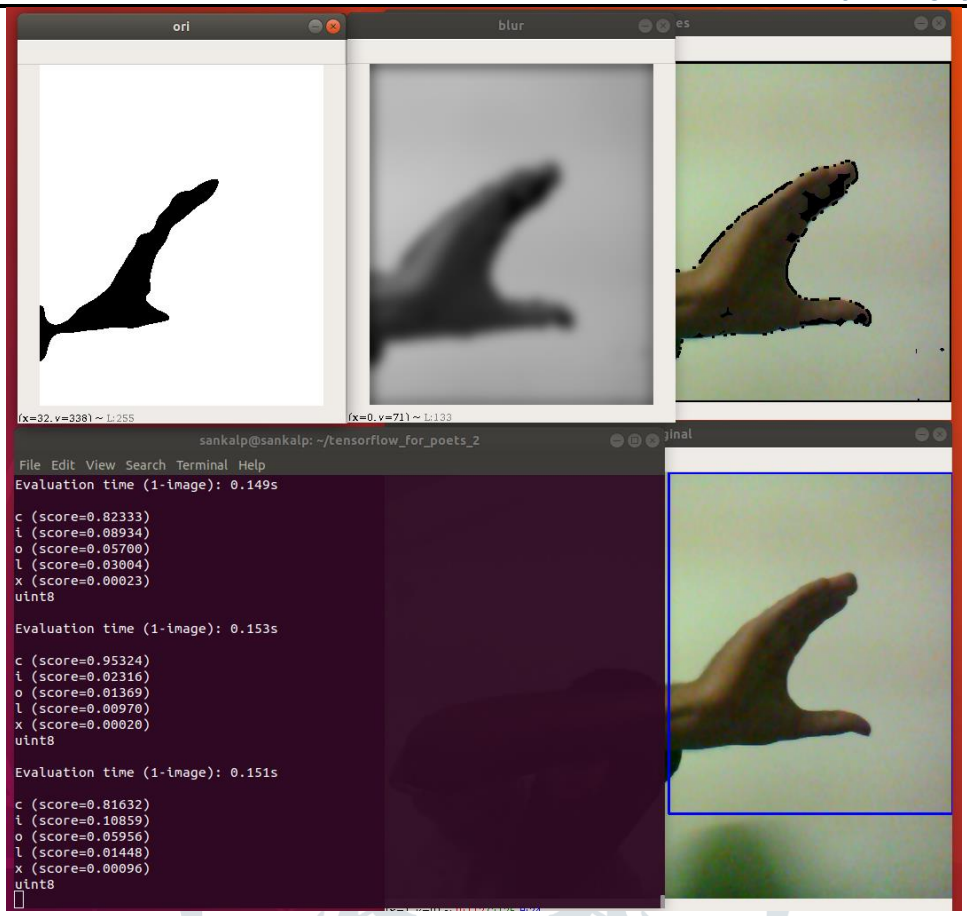

Fig 6.2: Result Analysis

VII. REFERENCES

[1] Surabhi S. Gatagat, Devyani D. More, Kalpana D. Varat, Janhavi S. Toshkhani, Shape Parameter-Based Recognition of Hand Gestures, International Journal of Innovative Research in Computer and Communication Engineering, October 2015

[2] Ahmad Zaki Shukor, Muhammad Fahmi Miskon, Muhammad Herman Jamaluddin, A New Data Glove Approach for Malaysian Sign Language Detection, 2015 IEEE International Symposium on Robotics and Intelligent Sensors (IRIS 2015)

[3] Dibyabiva Seth, Anindita Ghosh, Ariruna Dasgupta, and Asoke Nath, Real-Time Sign Language Processing System, Department of Computer Science, St. Xavier's College (Autonomous), Kolkata, India (2016)

[4] Research Article, Real-Time Hand Gesture Recognition Using Finger Segmentation: https://www.hindawi.com/journals/tswj/2014/267872/..

[5] Vision-based sign language translation device: https://www.researchgate.net/publication/261460857_Vision-based_sign_language_translation_device