VIVEKANAND EDUCATION SOCIETY'S INSTITUTE OF TECHNOLOGY

Department of Computer Engineering

Project Report on

Gesture Recognition and Language Interpretation for Indian Sign Language

In partial fulfillment of the Fourth Year, Bachelor of Engineering (B.E.) Degree in Computer Engineering at the University of Mumbai Academic Year 2018-2019

Submitted by

Sunmay Agharkar, D17A/01

Mukta Chandani, D17A/11

Kunal Jagasia, D17A/24

Nishant Shankar, D17A/67

Project Mentor

Mrs. Sujata Khandaskar

(2018-19)

VIVEKANAND EDUCATION SOCIETY'S INSTITUTE OF TECHNOLOGY
Department of Computer Engineering

Certificate

This is to certify that ***Sunmay Agharkar, Mukta Chandani, Kunal Jagasia, Nishant Shankar*** of Fourth Year Computer Engineering studying under the University of Mumbai have satisfactorily completed the project on "***Gesture Recognition and Language Interpretation for Indian Sign Language***" as a part of their coursework of PROJECT-II for Semester-VIII under the guidance of their mentor ***Mrs.Sujata Khandanskar*** in the year 2018-2019.

This thesis/dissertation/project report entitled ***Gesture Recognition and Language Interpretation for Indian Sign Language*** by ***Sunmay Agharkar, Mukta Chandani, Kunal Jagasia, Nishant Shankar*** is approved for the degree of ***B.E. Computer Engineering.***

Programme Outcomes	Grade
PO1,PO2,PO3,PO4,PO5,PO6,PO7, PO8, PO9, PO10, PO11, PO12 PSO1, PSO2	

Date:

Project Guide:

Project Report Approval
For
B. E (Computer Engineering)

This thesis/dissertation/project report entitled ***Gesture Recognition and Language Interpretation for Indian Sign Language*** by ***Sunmay Agharkar, Mukta Chandani, Kunal Jagasia, Nishant Shankar*** is approved for the degree of ***B.E Computer Engineering.***

Internal Examiner

External Examiner

Head of the Department

Principal

Date:
Place:

DECLARATION

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

---------- ----------

(Sunmay Agharkar, D17A-01) (Mukta Chandani, D17A-11)

---------- ----------

(Kunal Jagasia, D17A-24) (Nishant Shankar, D17A-67)

Date:

ACKNOWLEDGEMENT

We are thankful to our college Vivekanand Education Society's Institute of Technology for considering our project and extending help at all stages needed during our work of collecting information regarding the project.

It gives us immense pleasure to express our deep and sincere gratitude to Assistant Professor **Mrs. Sujata Khandaskar** (Project Guide) for her kind help and valuable advice during the development of project synopsis and for her guidance and suggestions.

We are deeply indebted to Head of the Computer Department **Dr.(Mrs.) Nupur Giri** and our Principal **Dr. (Mrs.) J.M. Nair ,** for giving us this valuable opportunity to do this project.

We express our hearty thanks to them for their assistance without which it would have been difficult in finishing this project synopsis and project review successfully.

We convey our deep sense of gratitude to all teaching and non-teaching staff for their constant encouragement, support and selfless help throughout the project work. It is great pleasure to acknowledge the help and suggestion, which we received from the Department of Computer Engineering.

We wish to express our profound thanks to all those who helped us in gathering information about the project. Our families too have provided moral support and encouragement at several times.

Computer Engineering Department
COURSE OUTCOMES FOR B.E PROJECT

Learners will be to,

Course Outcome	Description of the Course Outcome
CO 1	Able to apply the relevant engineering concepts, knowledge and skills towards the project.
CO2	Able to identify, formulate and interpret the various relevant research papers and to determine the problem.
CO 3	Able to apply the engineering concepts towards designing solution for the problem.
CO 4	Able to interpret the data and datasets to be utilized.
CO 5	Able to create, select and apply appropriate technologies, techniques, resources and tools for the project.
CO 6	Able to apply ethical, professional policies and principles towards societal, environmental, safety and cultural benefit.
CO 7	Able to function effectively as an individual, and as a member of a team, allocating roles with clear lines of responsibility and accountability.
CO 8	Able to write effective reports, design documents and make effective presentations.
CO 9	Able to apply engineering and management principles to the project as a team member.
CO 10	Able to apply the project domain knowledge to sharpen one's competency.
CO 11	Able to develop professional, presentational, balanced and structured approach towards project development.
CO 12	Able to adopt skills, languages, environment and platforms for creating innovative solutions for the project.

ABSTRACT

In India there are only 250 interpreters for a deaf population of around 18 million, which means that there is only 1 interpreter for 72,000 people. There is no official recognition for Indian Sign Language by the Government of India. This leads to a lack of awareness in the public consciousness about ISL, and it in turn results in a systemic apathy towards the education of the deaf community.

Our application is an interpreter of gestures from the Indian Sign Language and provides output in real time. The input is given in form of video, where the user makes gestures in front of a webcam. The application then interprets the gesture and provides the output as probabilities of the recognised word in English. It can also translate it to a regional language (Devanagari) for better understanding among the India's diverse population. The interpreter has been trained on word level data taken from the Indian Sign Language Research and Training Centre's YouTube channel, which has a number of videos of ISL gestures. This process makes use of not just the movement of the hands and their pose, but facial emotions as well. All these features are an essential part of the ISL gestures.

Through our project we strive to contribute to the efforts made to educate the deaf population, to make the hearing population of India cognizant of ISL and bridge the socio-economic gap.

TABLE OF CONTENTS

CHAPTER INDEX

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1:

INTRODUCTION

In this introductory chapter, we lay the groundwork for understanding our project and the motivations behind it. We describe the drawbacks persisting in the current system, and outline the problem statement that we have chosen. We discuss why tackling this problem is important, and also propose a methodology to do so.

1.1. Introduction to the project

Our project consists of making an application which can detect gestures from Indian Sign Language and translate it into English and other regional languages(Devanagari). Indian Sign Language, and its research and development are still in its nascency. We hope that our work will help deaf and hearing people to learn and communicate in Indian Sign Language. We are contributing to the efforts made by various groups in the country to educate, interpret, and disseminate Indian Sign Language throughout deaf communities, and to make the hearing population of India cognizant of ISL. Our application recognizes gestures through video, interprets the recognized ISL and gives text output in English as well as Devanagari.

1.2. Motivation for the project

"Persons with disabilities shall be entitled, on an equal basis with others, to recognition and support of their specific cultural and linguistic identity, including sign languages and deaf culture"

- *Article 30, Paragraph 4 of the United Nations Convention on the Rights of Persons with Disabilities*

As of 2018, there is no official recognition of Indian Sign Language. There is a lack of awareness in the public consciousness about ISL, and this results in a systemic apathy towards the education of the deaf community. Sign language in India is restricted to pocketed communities, and isn't formalized. Oralism, the system of teaching profoundly deaf people to communicate by the use of speech and lip-reading rather than sign language, has caused official establishment of sign language to be overlooked. There are only 250 interpreters for a population of 18 million deaf people in India. This means that there is only 1 interpreter for 72,000 people.

1.3. Problem Definition

The research problem that our project addresses is that of trying to interpret Indian Sign Language from a continuous video stream. Gestures and facial emotions are vital components of every sign, and any interpreter must take the signer's gestures and expressions into account while interpreting. The aim is to create an application that can capture video feed of a signer, and interpret the signs in real-time.

1.4 Existing systems and its comparison table

Paper	Year	Algorithm used	Dataset	Language	Performance
DeepASL	2018	Hierarchical bidirectional neural network and probabilistic framework based on Connectionist Temporal Classification (CTC)	7306 samples of 56 words and 100 sentences	American Sign Language	Word level translation accuracy : 94.5% Sentence level translation accuracy: 8.2% word error rate.
Video based Sign Language Recognition without Temporal Segmentation	2018	Hierarchical Attention Network and a two stream three dimensional Convolutional Neural Network	25000 labelled videos with 50 signers	Chinese Sign Language	83% accuracy
Real time American Sign Language	2018	GoogleNet architecture	ISLRVC2012 dataset : Letter level	American Sign Language	98% of accuracy for letters a-e and 74% accuracy for

Recognition with Convolutional Neural Networks			data		letters a-k
Selfie video based continuous Indian sign language recognition system	2018	Artificial Neural Networks	18 different signed by 10 different signers.	Indian Sign Language	Accuracy = 90%

Table 1: Existing systems and their comparison

1.5 Lacuna of the existing systems

The current research in Indian Sign language is more focused on hand movements whereas ISL is heavily face reliant. The movements of face and the emotions it depicts play a vital role in ISL. Research is more focused on character recognition. Translation into other regional languages has also not been carried out and it is essential to do so. An end-to-end application is not yet available.

1.6 Relevance of the Project

With an increase in the number of ways people interact with technology, software should be equipped with different methods of handling these interactions. This project has social and cultural ramifications, in that it can bridge communication barriers between different communities. Indian Sign Language does not have any constitutional status at the time of writing, but recently a PIL has been filed seeking legal status for ISL. Additionally, our project is

also relevant in terms of Human Computer Interaction (HCI), and it enables an alternative to voice interaction with a computer.

1.7 Methodology employed for deployment

Agile Methodology:

We have followed continuous iterations of development and testing throughout the development lifecycle of the project. Both development and testing activities are concurrently practiced. Agile is a form of incremental and iterative approach to software design. Any error occurring during any phase of the project could be rectified immediately. Every module once developed was tested before moving on to the next phase. The entire team worked for every phase of the project.

CHAPTER 2:

LITERATURE SURVEY

In this chapter, we discuss the work done in the fields of gesture recognition, sign language recognition and emotion recognition. We also explain why this research is relevant to our project, and make inferences that will prove to be useful to our work.

2.1 Interaction with domain experts

To be better acquainted with how Indian Sign Language works and is used to educate and communicate with deaf people, we visited Rochiram Thadani High School for hearing disabled, at Chembur, Mumbai. Mrs. Bhagyashree Vartak, the head faculty member having a profound knowledge and years of experience in ISL threw light on the fact that there are huge limitations in ISL grammar. Distinct words have distinct signs and the meaning of what is being conveyed is interpreted by the signs. She also mentioned that she is one of the very few interpreters in the city and that a real time application would be of great use to interpret, and disseminate Indian Sign Language throughout deaf communities, and to make the hearing population of India cognizant of ISL.

2.2 Research Papers Referred

1. DeepASL: Enabling Ubiquitous and Non-Intrusive Word and Sentence-Level Sign Language Translation
    a. Abstract

        DeepASL uses infrared light as its sensing mechanism to non-intrusively capture the ASL signs. They use a Hierarchical bidirectional deep recurrent neural network (HB-RNN) and a probabilistic framework based on Connectionist Temporal Classification (CTC) for word-level and sentence-level ASL translation respectively. 7,306 samples from 11 participants were collected, covering 56 commonly used ASL words and 100 ASL sentences They achieve 94.5% word-level translation accuracy and 8.2% word error rate on translating unseen ASL sentences

    b. Inference

        Existing translation tech is intrusive or not resilient to lighting changes. Also, they translate one sign at a time, requiring a pause between the signs. These limitations

significantly slow down face-to-face conversations, making those sign language translation systems much less useful in daily-life communication scenarios. This paper uses Leap Motion to non intrusively capture skeletal joint information. It also uses an HB-RNN to model spatial and temporal dynamics of these extracted ASL characteristics. This modelling can be applied in our project as well, to capture spatio-temporal concepts. They do not pre segment sentences into words, and then build a sentence. Instead, this paper performs end to end sentence generation. They have a labelled dataset that consists of sentences. Applying this method as it is can prove to be difficult for us as our data has phrases, not entire sentences.

2. Video-based Sign Language Recognition without Temporal Segmentation.

   a. Abstract

   This paper focuses on recognizing Chinese Sign Language by using a 2 stream 3 dimensional Convolutional Neural Network(CNN). The 2 streams are used for hand gesture detection and recognition. One stream contains features of the hand's global location, and the other consists of local hand movements. They also utilise a Hierarchical Attention Network (HAN) for continuous sign language recognition without temporal segmentation. HAN is an extension to an LSTM by considering structure information and attention mechanism. Video is represented with the proposed global-local features, while each word of the annotated sentence is encoded with a "one-hot" vector. The goal of the latent space is to construct a space to bridge semantic gaps. HAN decodes the hidden vector representation to a sentence word-by-word. They also incorporate a Latent Space model to explicitly exploit the relationship between visual video and text sentence. An accuracy of 83% is achieved using this method.

   b. Inference

   The above model focuses on sentence level recognition of sign language, by continuously feeding video input to the HAN. Video is captured non intrusively using a Kinect. This is an improvement over many current models which are primarily for character or word level recognition. Also, sentence level parsing is

carried out continuously, without segmenting it into individual words. This significantly speeds up the process, and makes it more practical for real time recognition, as is the need for our project.

3. Recent Advances of Deep Learning for Sign Language Recognition

   a. Abstract

   This is a survey paper,providing info about the current scope of research in Sign Language Recognition. They have started with ML techniques like using feature extraction (HOG, SIFT etc.) in SVMs and Hidden Markov models. 2D feature extraction is not really sufficient because of lack of depth perception in the images. There are 2 types of sensor based systems - touch based and untouched/vision based. Touch based uses sensors placed on the user's hands, which can prove to be cumbersome and uncomfortable. Vision based sensors like the Kinect or Google Tango, can provide much more accurate data and is less prone to errors due to external changes in illumination, and the huge colour and texture variability induced by clothing, hair, skin and background. Various datasets like The American Sign Language Lexicon Video Dataset (ASLLVD), The MSR Gesture3D dataset, Auslan Signbank, Indian Sign Language dataset along with useful, relevant information about them has been provided.

   b. Inference

   The Auslan Signbank dataset is very similar to the Indian Sign Language dataset as both rely on a visual means of communication since their signs are not just limited to hand movement, but have facial expressions as well. In Auslan, Each sign is made up of 5 five main parts; handshape, orientation, location, movement and facial expression.

   Deep Learning techniques have been proven to be better at sign language recognition than Machine Learning. Rioux-Maldague applied their technique to American Sign Language fingerspelling classification using a Deep Belief Network (DBN). DBN has 3 Restricted Boltzmann Machines (RBM) (with size of 1500, 700 and 400 units) and one translation layer. This achieved an accuracy of

99%. More importantly, their method is also capable of real-time sign classification and is adaptive to any environment or lightning intensity.

Huang provided the 3D coordinates of finger joints in real time. They also tried an approach with HMM, but the accuracy of a 3D CNN is better. They have also used a Kinect to provide data, colour is in RGB format. To boost the performance, multi-channels of video streams, including colour information, depth clue, and body joint positions, were used as input to the 3D CNN in order to integrate colour, depth, and trajectory information.

4. Real time American Sign Language Recognition with Convolutional Neural Networks

a. Abstract

This paper talks about an end to end web application which was developed to convert ASL to their corresponding letters. They try to provide a replacement for expensive motion sensor gloves with a scalable solution. They are making use of GoogLeNet architecture trained on ISLRVC2012 dataset. Their algorithm consistently classifies a-e whereas satisfies remaining letters in majority of cases. Their major problems were the background, sign boundary and some occlusions.

b. Inference

This model can be useful as our proposal is to create a web application which is scalable and provides accurate results. This model can be used for the translation of finger movements into their corresponding letters. Their preprocessing of data to resize all the images and removing the unnecessary background can be useful to our project.

5. Selfie video based continuous Indian sign language recognition system

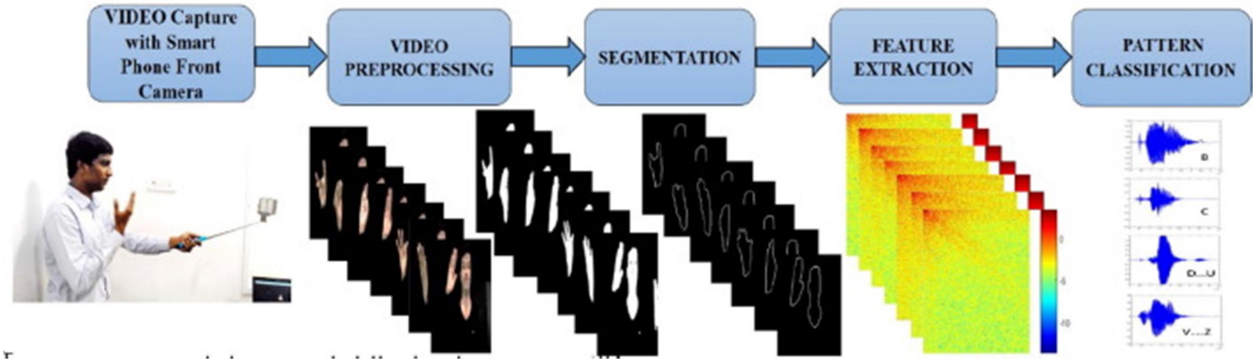
a. Abstract

This paper proposes a methodology which uses selfie stick and front camera to capture videos. As a result, only one hand can be used to make gestures. Here recognition and interpretation depend upon 5 parameters such as hand and head recognition, hand and head orientation, hand movement, shape of hand and location of hand and head. They also depend upon the background. In the training

dataset, only 18 different signs are worked upon by 10 different people with the same background. Object detection (segmentation) is done by applying gradient masking to the image. It divides the videos into 220 frames.

b. Inference

The methodology proposed in this paper, is quite useful and similar to what we intend to use with just the limitation of only one hand being able to make gestures. The feature sign matrix inputs a classifier. Since speed is the prime constraint during mobile implementation, it will be reasonable to use minimum distance classifier (MDC). The model of 3 layered artificial neural network is presented. Euclidean, normalized Euclidean and Mahalanobis distance metrics classify sign features. Mahalanobis distance reached an average word matching score of around 90.58% consistently when compared to the other two distance measures for the same train and test sets. ANN 2 and 4 hidden layered ANN are used for recognizing the sign language. With 50 × 50 feature matrix per frame and an average number of frames per video at 220 frames, the feature matrix for the considered 18 signs is a stack of 50 × 50 × 220 matrix. Initiating a multi-dimensional feature matrix of this size takes longer execution periods. Hence PCA treats each frames 50 × 50 energy features by computing Eigenvectors and retaining the principle components to from a 50 × 1 vector per frame.



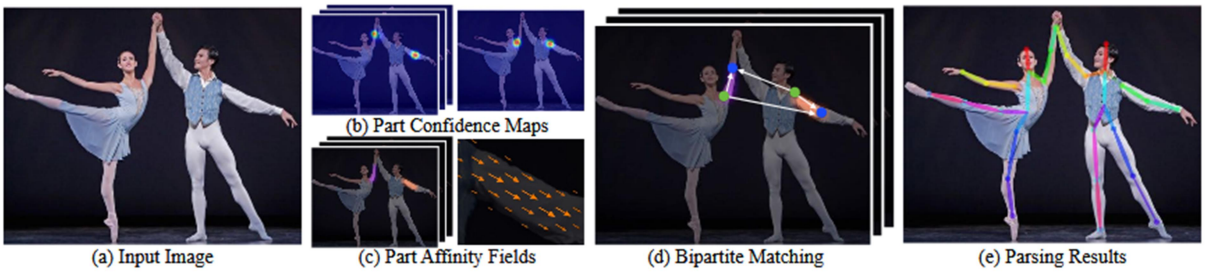
6. OpenPose: Realtime Multi-person 2D Pose Estimation using Part Affinity Field

a. Abstract

OpenPose is used for realtime multi-person 2D pose estimation. It uses Part Affinity Fields (PAFs), a set of 2D vector fields that encode the location and orientation of limbs over the image domain to learn associate body parts with the individuals in image. OpenPose overcomes the problem of using different system for estimating facial and body. It consists of three different blocks : body + foot keypoints detection, hand keypoints detection and face detection. This method has been evaluated on three datasets : MPII human multi-person dataset which consists of data multiple interactive individuals in highly articulated poses with 14 body parts, COCO keypoint challenge dataset which requires identifying 12 keypoints on human body and 5 facial keypoints for every individual, and a foot dataset created from the COCO keypoint dataset. OpenPose is the first real-time multi-person system to jointly detect human body, foot,hand, and facial keypoints (in total 135 keypoints) on single images.

b. Inference

This framework can be used in our system to detect the necessary keypoints oon the body for every frame of the video. This keypoints for every frame would provide a information about the position and orientation of the hand and the face and can be fed in to the neural network for identifying the gesture. The keypoints can be restricted only to upper body to make the system more scalable as most of the videos in the dataset involve movement of only the upper body.



7. Optical Flow based Real-time moving object detection in unconstrained scene

a. Abstract

This paper uses optical flow to detect moving object in a dynamic environment. To detect a moving object, this paper takes frames of video as input. It estimates the values of optical flow utilizing the value of two frames to establish a equation which is then solved to obtain a homography matrix. This homography matrices are used to obtain the background model . The foreground mask is generated by using adaptive threshold which is not dependent on the zooming situation. This paper gives a success rate of 0.9 for 40 iterations.

b. Inference

The methodology proposed in this paper can be used in our system to reduce the time of processing. Optical Flow can be used to detect the hand movement in the video. This can be used to establish a relation between the frames. As optical flow can be used in an unconstrained environment, it can be used for our system as the environment of the video is not static and it generates the matrix for hand movement irrespective of the zooming situation in different frames.

8.Recognizing American Sign Language Gestures from within Continuous Videos

a. Abstract

The authors propose a hybrid model to recognize ASL gestures from continuous videos. This hybrid model, the 3D Recurrent Convolutional Neural Network, consists of a 3D Convolutional Neural Network and a Fully Connected Recurrent Neural Network. With this model, they aim to capture both spatio-temporal and sequential information of the sliced clips. They believe this improves the power of the final feature representations. They also collect a new ASL dataset which contains videos of people signing specific ASL words and sentences. This dataset is different from previous datasets, as they include multiple modalities (facial movements, hand gestures, and body pose) and multiple channels (RGB, Optical flow, and depth). They fully annotate each semantic region, and the videos contain multiple input channels. Their method achieves a 69.2% accuracy on the word videos for 27 ASL words.

b. Inference

Their model architecture is powerful, in that it combines both a spatial model (the CNN) and a temporal model (RNN), and our model will also be inspired by this hybrid architecture. The question of multimodality is interesting, and deriving

features from these different modes and then concatenating them can create a richer feature representation for our data. Our model also aims to capture facial emotion and concatenate that to the gesture features. Another aspect of this work that is important to mention is their dataset collection methodology. Our dataset can be improved upon by some measures this work has taken.

CHAPTER 3: REQUIREMENT GATHERING

3.1 Definition of requirement gathering

In this chapter, we discuss the functional and nonfunctional requirements of our project. We also address the constraints of the problem, and the hardware and software requirements for our project. We discuss the tools and techniques relevant to our project and the algorithms and frameworks used.

3.2 Functional Requirements

- Pose Estimation - To recognize complicated gestures, system should be able to predict an approximation of the skeletal joints. This is done to feed downstream model accurate gesture data.
- Detecting movement - Motion detection is necessary to understand gestures, especially in spatio-temporal data.
- Emotion recognition - ISL and sign languages in general are heavily reliant on facial expressions. Identifying the mood of a speaker is essential to understanding the context, and to an extent the structure, of a word.
- ISL Dictionary - Since there is a chance users may not be familiar with ISL, a dictionary is provided so that it can act as a learning tool and as a reference.
- Translation of ISL into regional languages (Devanagari)

3.3 Non-Functional Requirements

- Speed - Translations should be quick
- Video sampling rate - Algorithm should work at better frame rates
- Operability - Application should work in different environments (web/mobile)
- Adaptability - ISL is not yet completely standardized and thus is still growing. Our system should be generalized enough to support changes in the future.

3.4 Constraints

- No formalized grammar for ISL
- Lack of sentence level data and only one video per sign in case of phrases
- Minimal computing power for training model
- Cloud VMs are randomly assigned and run out of RAM occasionally

3.5 Various Technology and tools used

Technologies : Openpose, Numpy, caffe, Flask, pandas, keras, tensorflow, etc.
Tools : Jupyter Notebook, VS Code, Google Colab

CHAPTER 4 :

PROPOSED DESIGN

In this chapter, we lay out the design of our project from different perspectives. We discuss the overall system design, the block diagram, and a data flow representation of our system. We also present a flowchart of the system. To conclude this section, we provide a scheduling chart of our methodology.

4.1 Block diagram of the system

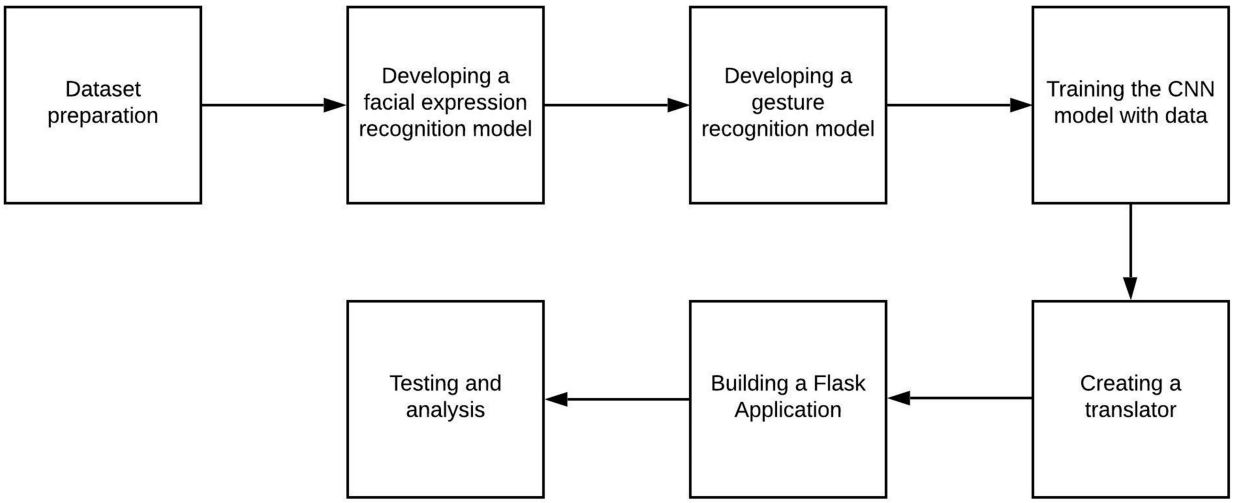


Figure 4(a) : Block Diagram of our Project

This block diagram represents the flow of our methodology in this project. Indian Sign Language is highly dependent on facial expressions, as they may change the meaning of the word. Therefore, along with recognising hand gestures, we also have to consider the facial expression of the speaker, and combine those features to accurately determine the sign. Native sign language speakers in India have grown up understanding their signs in their regional languages and not necessarily in English. So, a translator has also been provided. This project will be built as a Flask application.

4.2 Modular design of the system

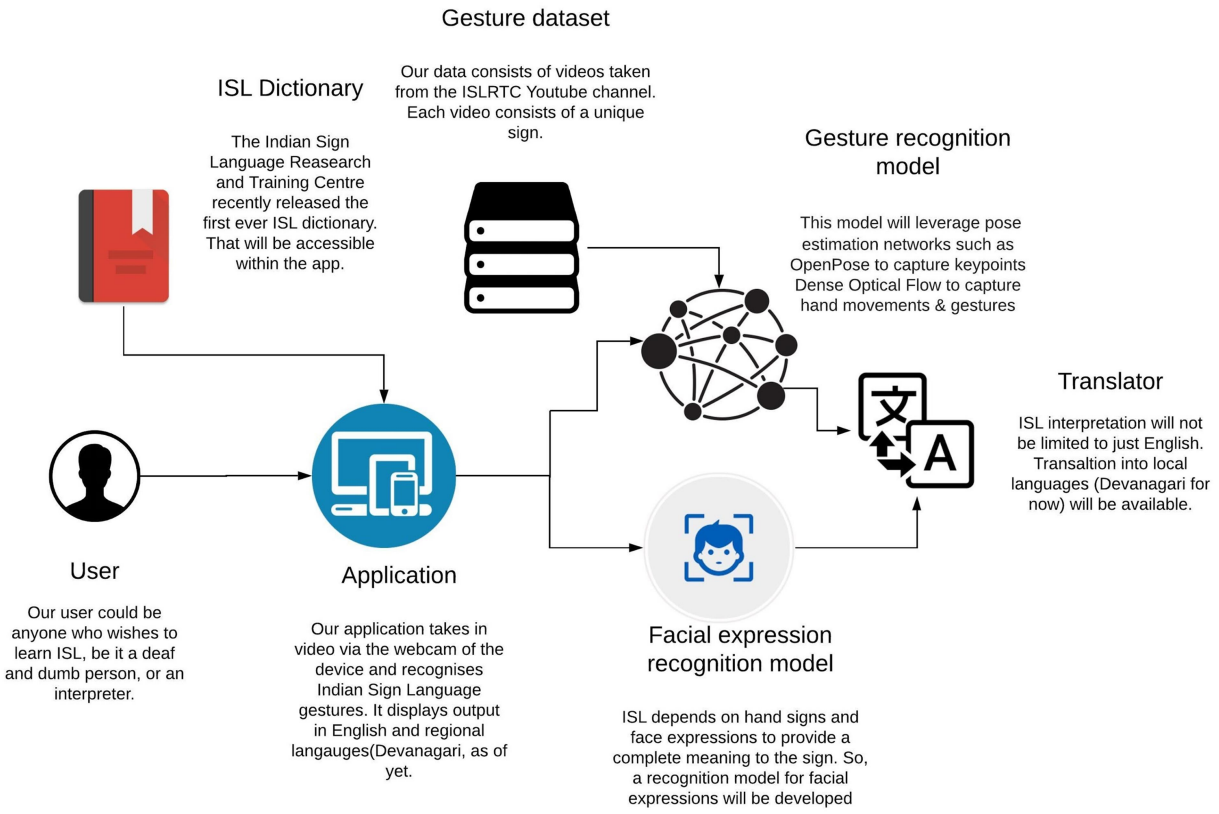


Figure 4(b): Modular Design of the System

Our application can be used by anyone who wishes to speak/learn Indian Sign Language. We are developing 2 models for recognition of a sign, a hand gesture recognition model and a facial expression recognition model. Hand gestures are the basis of any sign language, whereas facial expressions are special feature of ISL, since they can provide extra meaning to a sign. So, for correct interpretation, we are going to combine the features extracted from these models. These models will be trained on data like the videos from the ISLRTC Youtube channel, and the IITA-ROBITA hand gesture databank. Native ISL users might not have grown up with an English interpretation of the signs. Thus, we have also developed a translator to Devanagari for easier understanding. The ISL dictionary has been a recent and significant development in the ISL field and that too finds a place in our application.

4.3 Detailed design

4.3.1 Data flow diagrams

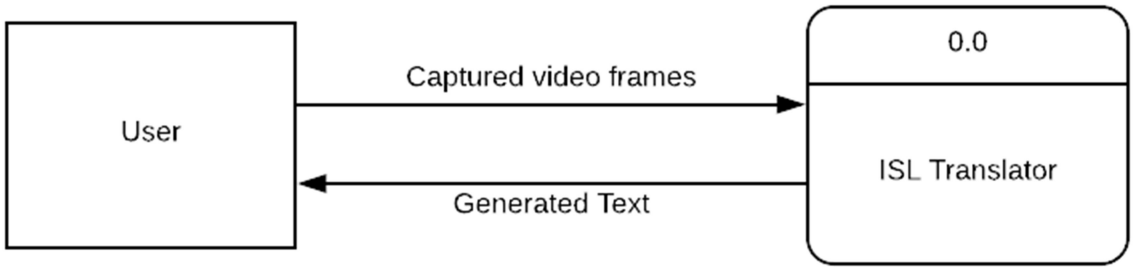


Figure 4(c) : DFD Level 0

User provides input to the ISL translator in the form of captured video frames of people communicating in sign language. The ISL translator generates output in the form of text, stating what the person in the input video is trying to convey.

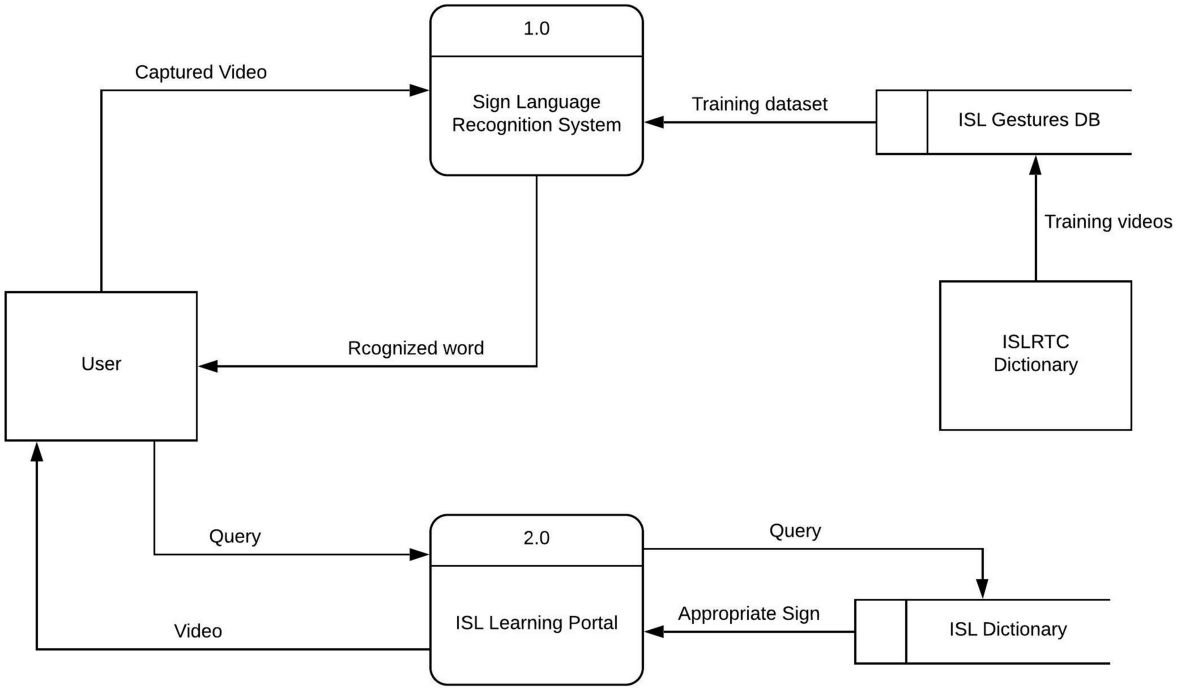


Figure 4(d) : DFD Level 1

User inputs captured video frames to the sign language recognition system, which converts various interpreted gestures into words. The data set used in this project is from a youtube channel, ISLRTC, from where training videos are stored into the ISL gestures database.

A user can use the ISL learning portal, to learn about the appropriate signs for a particular word/phrase. The ISL dictionary stores the exact sign used for respective words.

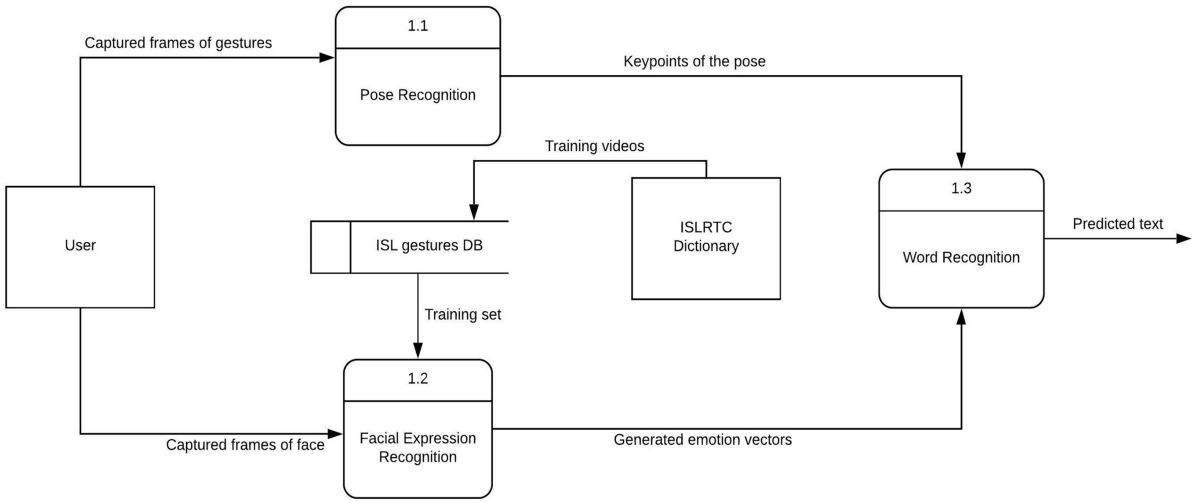


Figure 4(e) : DFD Level 2

To accurately predict text for sign language, we take into consideration both gestures and facial expressions. Phrase and emotion vectors generated on recognition are used for word recognition which in turn gives the predicted text as output.

4.3.2 Flowchart

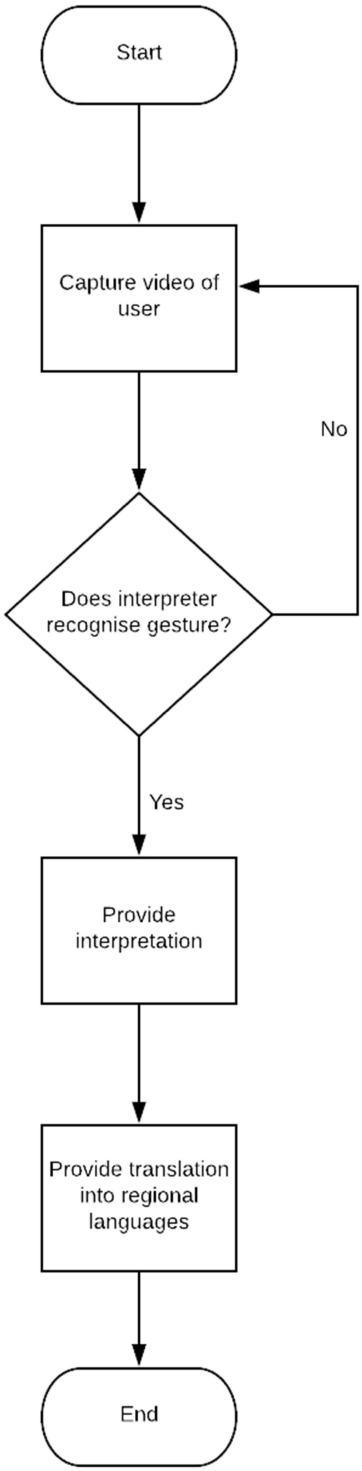


Figure 4(f): Flowchart

The above flowchart depicts the overall process flow in our application. The app will capture a video of the user making the sign, and in real time, provide an interpretation of the sign. There will also be a facility to translate the interpreted text into regional languages.

4.4 Project Scheduling & Tracking using Timeline / Gantt Chart

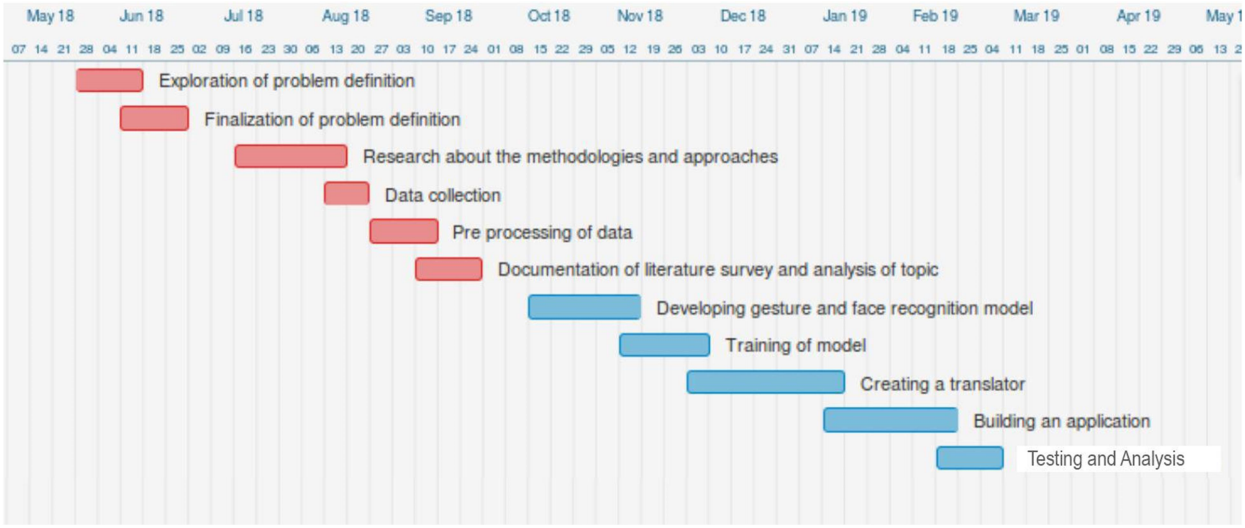


Figure 4(g) : Gantt Chart

CHAPTER 5: IMPLEMENTATION DETAILS

In this chapter we provide the various algorithms applied in our system and discuss in detail how they can be used collectively to provide the solution to our problem statement.

5.1 Algorithms developed for respective modules

5.1.1 OpenPose for Pose Estimation

OpenPose is real time pose estimation algorithm developed by CMU. It is the first algorithm to jointly extract keypoints of the face, both hands and body of multiple individuals in a frame totalling upto 135 keypoints per individual. OpenPose employs Part Affinity Fields to encode the location and orientation of limbs over the image domain. It can work for both 2D and 3D poses, however we have utilized only 2D pose estimation for our problem statement. It can take in input as images, video or webcam and provides output as an image with the keypoints highlighted and the actual features in another file.

In our application we are using Openpose to extract features of the face and body of the user. In our algorithm we extract 6 keypoints of the face and 9 of the body to detect movement across the frames. So, for every frame in the video we get a feature vector of 15 * 2, where 15 is the number of keypoints and 2 consists of the x,y coordinates in the frame. All these vectors are then stacked together to get 2D array of the features of the entire video.

```
/content/openpose/build/examples/tutorial_api_python
CURRENT IMG -  ../../../../data/Absent_s2/frame0.jpg
SIGN-    (15, 2)
CURRENT IMG -  ../../../../data/Absent_s2/frame1.jpg
SIGN-    (15, 2)
CURRENT IMG -  ../../../../data/Absent_s2/frame2.jpg
SIGN-    (15, 2)
CURRENT IMG -  ../../../../data/Absent_s2/frame3.jpg
SIGN-    (15, 2)
CURRENT IMG -  ../../../../data/Absent_s2/frame4.jpg
SIGN-    (15, 2)
CURRENT IMG -  ../../../../data/Absent_s2/frame5.jpg
SIGN-    (15, 2)
CURRENT IMG -  ../../../../data/Absent_s2/frame6.jpg
SIGN-    (15, 2)
CURRENT IMG -  ../../../../data/Absent_s2/frame7.jpg
SIGN-    (15, 2)
CURRENT IMG -  ../../../../data/Absent_s2/frame8.jpg
SIGN-    (15, 2)
CURRENT IMG -  ../../../../data/Absent_s2/frame9.jpg
SIGN-    (15, 2)
CURRENT IMG -  ../../../../data/Absent_s2/frame10.jpg
SIGN-    (15, 2)
```

Figure 5(a): Feature extraction on our dataset

Figure 5(b): OpenPose keypoints

5.1.2 Optical Flow for Hand Movement

- Farneback's algorithm

  Farneback's algorithm provides motion estimation between 2 consecutive frames. It uses polynomial expansion to approximate some neighborhood of each pixel with a polynomial. Using this expansion, the algorithm maps out a displacement field across the entire image, based on the detected motion between 2 frames.

  This algorithm is especially useful to segregate movements between every frame in the video. Detected motion between frames is highlighted using color i.e RGB values, and the portion of the image without any motion is left dark. This segregation makes it easier to capture the subtle motion between the hands, as only those parts of the image with movement detected are highlighted.

  The input provided is a video from our curated dataset and the output we get is an sequence of images or video with detected movement, and a feature vector of size 60 * 80 * 3. 60 * 80 are the number of pixels in each frame and 3 represents the (r,g,b) values per pixel.

```
Constructing pyramid...done!
Pyramid level 3
Pyramid level 2
Pyramid level 1
Pyramid level 0
```

Figure 5(c): Optical flow running

Figure 5(d): Optical flow output

5.1.3 Convolutional Neural Network for Facial Emotion Recognition

Facial emotions can be divided into six broad categories - Anger, Disgust, Fear, Happiness, Sadness, Surprise and neutral. The CNN would classify the frames into these six basic emotions. The CNN model would be trained on FER2013 dataset. There are 35,888 images in this dataset These images are gray-scale images and are in the form of a 48*48 matrix with each point representing the intensity of the pixel. So the captured frame from the videos are fed into the CNN by converting them into 48*48 gray-scale matrices. Keras is used to create a Sequential Convolutional Network with the following components: Convolutional Layers, Activation Functions, Pooling Layers, Dense Layers, Dropout Layers and Batch Normalization.

CHAPTER 6:

TESTING

In this chapter we have described the various types of testing which are performed on the system. The various test cases considered for testing and the inference drawn from those test cases is also described.

6.1 . Definition of testing

Software testing is a process, to evaluate the functionality of a software application with an intent to find whether the developed software met the specified requirements or not and to identify the defects to ensure that the product is defect free in order to produce the quality product. Testing assesses the quality of the product. Software testing is a process that should be done during the development process. In other words software testing is a verification and validation process.

6.2. Types of tests

- Unit Testing - It focuses on smallest unit of system design. It tests an individual unit or group of interrelated units. It is often done by programmer by using sample input and observing its corresponding outputs.
- Integration Testing - The objective is to take unit tested components and build a program structure that has been dictated by design. Integration testing is testing in which a group of components are combined to produce output. Integration testing is of two types:
  a)Black Box Testing
  b)White Box Testing
- Regression Testing - Every time new module is added leads to changes in program. This type of testing make sure that whole component works properly even after adding components to the complete program.
- Alpha Testing - This is a type of validation testing.It is a type of acceptance testing which is done before the product is released to customers. It is typically done by QA professionals.
- Beta Testing - The beta test is conducted at one or more customer sites by the end-user of the software. This version is released for the limited number of users for testing in real time environment

- System Testing - In this software is tested such that it works fine for different operating system. It is covered under the black box testing technique. In this we just focus on required input and output without focusing on internal working.
- Stress Testing-In this unfavorable conditions are given to the system and check how it performs in those condition.

6.3. Type of Testing considered with justification

- Unit testing - All the units i.e ISL dictionary, pose estimation, emotion recognition etc. were tested individually before integration. This ensured that every unit functioned properly without any errors.
- Integration testing - Once all the units were developed they were integrated together and tested for any errors. All the individual units developed were integrated using flask framework. The modules work properly to recognize the sign.
- System testing - The project was tested for operating system compatibility. For this it was tested on both Windows and Linux(Ubuntu) operating systems. This ensured that a user having either of the two operating systems would be able to use the developed system without any difficulty.
- Stress testing - The system was stress tested by varying the number of videos, from 20 to 300, given for training the model. This would ensure smooth functioning of the system on varied inputs.

6.4 Various test case scenarios considered

- Using multiple operating systems for project demonstration viz. Windows(32 and 64 bit) and Ubuntu.
- Changing the number of videos used for model training ranging from 20 to 300.
- Checking for system compatibility with multiple browsers like Google Chrome, Firefox, IE,etc.
- The classification of gestures is independent of the background and the lighting.

6.5. Inference drawn from the test

This showed that the length of the review was not a constraint for prediction. It was also able to learn from many reviews as training data. The system was found to be compatible with both the operating systems considered. It also displayed the exact same user interface on all the browsers considered.

CHAPTER 7 :

RESULT ANALYSIS

7.1. Module(s) under consideration

While discussing the results of our project and our contributions, we consider the Sign Classification module and define certain metrics that we use for evaluation. The sign classification module is a downstream module, and hence depends on the accuracy of optical flow and pose estimation. The ISL Dictionary module provides a front end interface to access the dictionary and as such doesn't require quantitative evaluation.

7.2. Parameters considered

- Important metrics of Sign Classification module.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

- Robustness of our application under different real-world settings
    a. Lighting
    b. Pose Variance
    c. Occlusions
- Runtime/memory/time

7.3. Screenshots of User Interface (UI) for the respective module

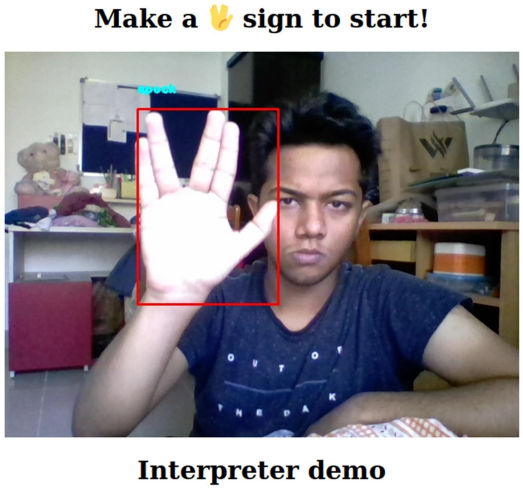
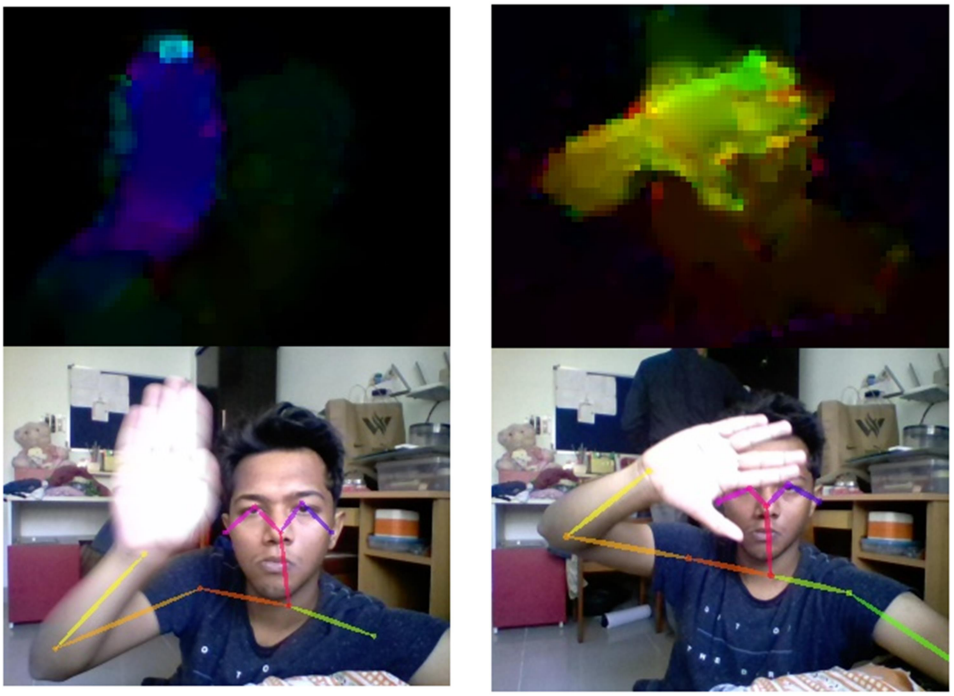


Fig 7(a) : User signing delimiter gesture



7(b) : Dense optical flow and Openpose Pose estimation visualization

```
layer     filters    size              input                output
    0 conv     16  3 x 3 / 1   416 x 416 x   3   ->   416 x 416 x  16  0.150 BFLOPs
    1 max          2 x 2 / 2   416 x 416 x  16   ->   208 x 208 x  16
    2 conv     32  3 x 3 / 1   208 x 208 x  16   ->   208 x 208 x  32  0.399 BFLOPs
    3 max          2 x 2 / 2   208 x 208 x  32   ->   104 x 104 x  32
    4 conv     64  3 x 3 / 1   104 x 104 x  32   ->   104 x 104 x  64  0.399 BFLOPs
    5 max          2 x 2 / 2   104 x 104 x  64   ->    52 x  52 x  64
    6 conv    128  3 x 3 / 1    52 x  52 x  64   ->    52 x  52 x 128  0.399 BFLOPs
    7 max          2 x 2 / 2    52 x  52 x 128   ->    26 x  26 x 128
    8 conv    256  3 x 3 / 1    26 x  26 x 128   ->    26 x  26 x 256  0.399 BFLOPs
    9 max          2 x 2 / 2    26 x  26 x 256   ->    13 x  13 x 256
   10 conv    512  3 x 3 / 1    13 x  13 x 256   ->    13 x  13 x 512  0.399 BFLOPs
   11 max          2 x 2 / 1    13 x  13 x 512   ->    13 x  13 x 512
   12 conv   1024  3 x 3 / 1    13 x  13 x 512   ->    13 x  13 x1024  1.595 BFLOPs
   13 conv    256  1 x 1 / 1    13 x  13 x1024   ->    13 x  13 x 256  0.089 BFLOPs
   14 conv    512  3 x 3 / 1    13 x  13 x 256   ->    13 x  13 x 512  0.399 BFLOPs
   15 conv     18  1 x 1 / 1    13 x  13 x 512   ->    13 x  13 x  18  0.003 BFLOPs
   16 yolo
   17 route  13
   18 conv    128  1 x 1 / 1    13 x  13 x 256   ->    13 x  13 x 128  0.011 BFLOPs
   19 upsample            2x    13 x  13 x 128   ->    26 x  26 x 128
   20 route  19 8
   21 conv    256  3 x 3 / 1    26 x  26 x 384   ->    26 x  26 x 256  1.196 BFLOPs
   22 conv     18  1 x 1 / 1    26 x  26 x 256   ->    26 x  26 x  18  0.006 BFLOPs
   23 yolo
```

7(c) : YOLO v3 architecture for delimiter detection

7.4. Evaluation of the developed system

- **Accuracy**

No. of signs	No. of videos per sign	Architecture	Accuracy (in %)
10	1	RNN-LSTM	2.0
10	10	RNN-LSTM	9.7
10	20	RNN-LSTM	24.3
10	30	RNN-LSTM	**40.2**
15	20	RNN-LSTM	35.6

Table 2: Accuracy table

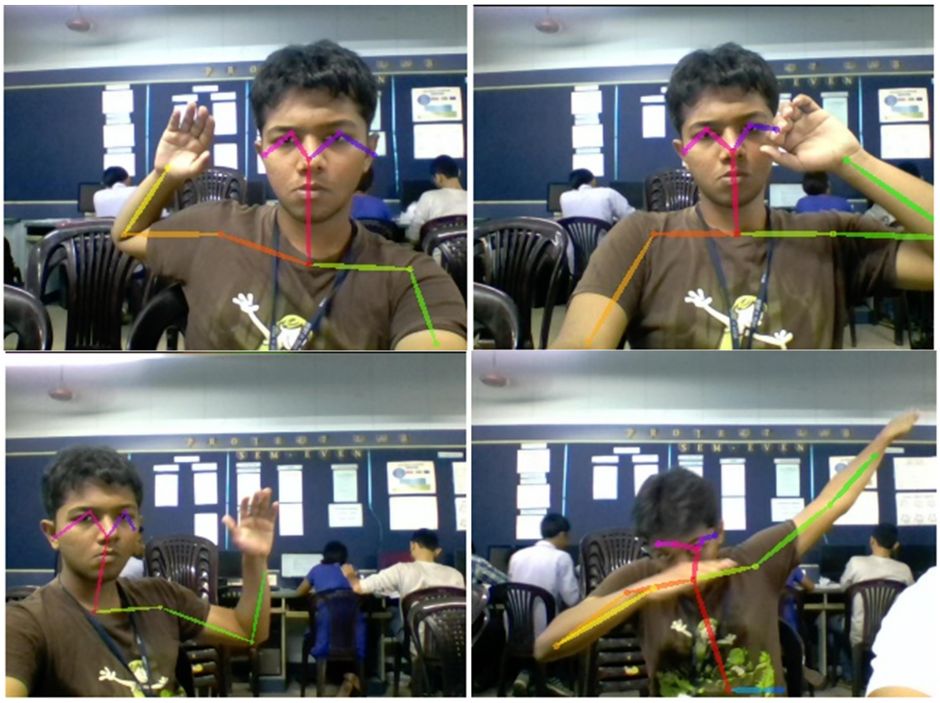
- **Robustness of our application under different real-world settings**
- Lighting

  The delimiter detection using YOLO v3 falters at times in conditions of extreme lighting, but this can be rectified with a larger training dataset consisting of such lighting variations. Presently, the YOLO model is satisfactorily working in predominant lighting conditions.

7(d): Interpreter demo

- Pose Variance

  OpenPose is able to efficiently estimate body pose keypoints in a variety of settings and postures. For our application, we require only the keypoints associated with the upper body and hence we utilize 15 keypoints out of the 25 keypoints that the COCO model detects.



7(e):Pose Variance 1

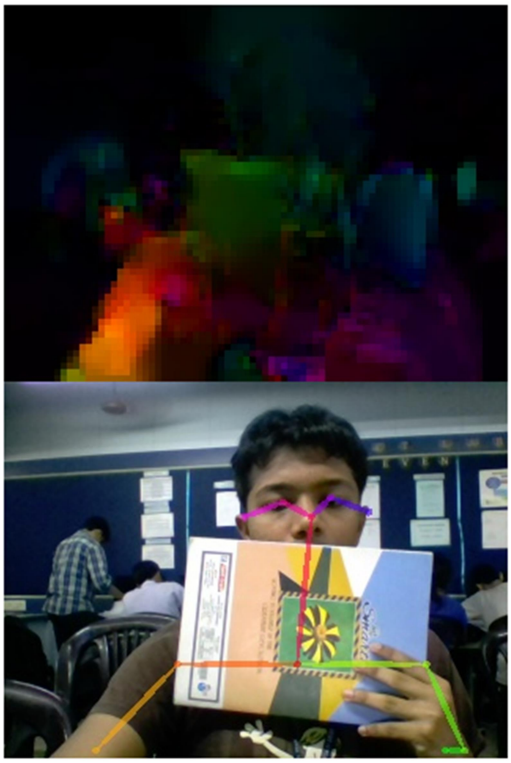
Though Openpose is able to capture a large variety of poses, due to the small pixel space we provide as input, it's accuracy is somewhat hampered at times. In the context of an image stream across multiple frames, these errors shouldn't matter.

7(f): Pose Variance 2

- Occlusions

  Pose estimation works even in the case of partial occlusions of body and face. We use Openpose to perform pose estimation which has been trained using a varied dataset

consisting of different poses and occlusions. Entire coverage of face and/or hands leads to poor accuracy as expected.



7(g): Occlusions

Runtime/memory/time

- Delimiter detection latency : ~10ms (almost instantaneously)
- Frame rates
    a. Openpose Pose Estimation : 10-12 fps on average
    b. Dense optical flow : 8 fps on average
- Model details:
    a. Training time (RNN) : 2-3 hours for saturation of loss
    b. Memory (RNN) : 4 GB RAM
    c. Memory (Conv LSTM) : 10 GB RAM
    d. Size (RNN) : ~125 MB

CHAPTER 8:

CONCLUSION

In this concluding chapter, we take a broad look at our system and discuss its merits and demerits. We also expand on how this project can be improved in the future.

8.1 Limitations

- Since only one video per sign is available for training on the ISLRTC YouTube channel, we created a custom dataset consisiting of additional videos for 20 signs. Since we are not native signers, this data itself will have certain errors associated with signing but we hope that this error accounts for variations in signing a particular word or phrase.
- To minimize training time and prototyping, fewer signs were considered for classification than are present in the dataset.
- A stopword or a delimiter is used to delineate signing of different phrases. In natural conversation, the usage of this delimiter could impede effective communication.
- On average, each phrase or word we looked at took less than 6 seconds to sign so we have constrained our signing time to 6 seconds. There may be outliers to this limit, as some signs are agglomerative and consist of subwords or sub phrases.

8.2 Conclusion

In this project, we have attempted to build an environment to support classification of Indian Sign Language. We collected data from the ISLRTC Youtube channel, but realizing it isn't sufficient, recorded a custom dataset with signs that we recorded ourselves. Ultimately, we organized a dataset consisting of 273 videos, for 10 signs. We also recorded videos for more signs, but did not focus on them while training our model. Flask seemed most suitable for our application, and we integrated different modules to make sign classification work in effectively real-time conditions. Our feature extraction is based on two techniques, namely pose estimation using Openpose and optical flow calculation using a Python wrapper of Coarse2fine optical flow. We achieved an accuracy of 40.2% with our best model. Given the limitations of our dataset and the difficulty of capturing spatio-temporal concepts while performing classification, we believe this is a promising result albeit the accuracy is relatively low.

8.3 Future Scope

- With a sentence level dataset, a multiclass classifier for interpreting sentences can be built.
- Accuracy could be improved with better hardware and a standardised dataset with more videos.
- Optical flow could be optimized for providing better results.
- More architectural models like multimodal networks could be tested.

REFERENCES

1. Articles referred

N. Sunavala, “The fight to make Indian Sign Language official - Times of India ►,” *The Times of India*, 08-Oct-2018. [Online]. Available: https://timesofindia.indiatimes.com/home/sunday-times/the-fight-to-make-indian-sign-language-official/articleshow/66101577.cms.

P. of View, “Oralism is the huge, entirely avoidable barrier that Deaf people are forced to face,” *Medium*, 15-Aug-2017. [Online]. Available: https://medium.com/skin-stories/oralism-is-the-huge-entirely-avoidable-barrier-that-deaf-people-are-forced-to-face-91e0421e533b.

S. Mukherjee, “This New Initiative Helps You Give The Deaf Community A Voice, Offers Sign language Courses,” *indiatimes.com*, 09-Jul-2016. [Online]. Available: https://www.indiatimes.com/news/india/this-new-initiative-helps-you-give-the-deaf-community-a-voice-offers-full-time-sign-language-courses-257992.html.

“Gloves Convert Sign Language to Speech [video],” *Health Tech Insider*, 03-May-2016. [Online]. Available: https://healthtechinsider.com/2016/05/03/gloves-convert-sign-language-speech-video/.

2. Research Papers Referred

[1] B. Fang, J. Co, and M. Zhang, “DeepASL,” *Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems - SenSys '17*, 2017.

[2] D. Rempel, M. J. Camilleri, and D. L. Lee, “The design of hand gestures for human–computer interaction: Lessons from sign language interpreters,” *International Journal of Human-Computer Studies*, vol. 72, no. 10-11, pp. 728–735, 2014.

[3] Z. A. Ansari and G. Harit, “Nearest neighbour classification of Indian sign language gestures using kinect camera,” *Sadhana*, vol. 41, no. 2, pp. 161–182, 2016.

[4] A. Nandy, J. S. Prasad, S. Mondal, P. Chakraborty, and G. C. Nandi, “Recognition of Isolated Indian Sign Language Gesture in Real Time,” *Communications in Computer and Information Science Information Processing and Management*, pp. 102–107, 2010.

[5] C. F. Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, “EmotioNet: An Accurate, Real-Time Algorithm for the Automatic Annotation of a Million Facial Expressions in the Wild,” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[6] C. Zimmermann and T. Brox, “Learning to Estimate 3D Hand Pose from Single RGB Images,” *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017.

[7] Huang, Jie, Wengang Zhou, Qilin Zhang, Houqiang Li and Weiping Li. “Video-based Sign Language Recognition without Temporal Segmentation.” *CoRR* abs/1801.10111 (2018): n. pag.

[8] L. Zheng, B. Liang, and A. Jiang, “Recent Advances of Deep Learning for Sign Language Recognition,” *2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, 2017.

[9] C. Corneanu, F. Noroozi, D. Kaminska, T. Sapinski, S. Escalera, and G. Anbarjafari, “Survey on Emotional Body Gesture Recognition,” *IEEE Transactions on Affective Computing*, pp. 1–1, 2018.

[10] Garcia, Brandon and Viesca , Sigberto. Real-time American Sign Language Recognition with Convolutional Neural Networks. In *Convolutional Neural Networks for Visual Recognition* at Stanford University, 2016.

[11] K. Tripathi and N. B. G. Nandi, “Continuous Indian Sign Language Gesture Recognition and Sentence Formation,” *Procedia Computer Science*, vol. 54, pp. 523–531, 2015.

[12] Yuancheng Ye, Yingli Tian, Matt Huenerfauth, Jingya Liu; The *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2018, pp. 2064-2073

[13] Rao, G. Ananth, and P.v.v. Kishore. “Selfie Video Based Continuous Indian Sign Language Recognition System.” *Ain Shams Engineering Journal*, 2017, doi:10.1016/j.asej.2016.10.013.

[14] Cao, Zhe, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields." arXiv preprint arXiv:1812.08008 (2018).

PROJECT PROGRESS REVIEW SHEETS

Inhouse/ Industry:

Class: D17 A/~~B/C~~
Group No.: 19

Project Evaluation Sheet 2018 - 19

Title of Project: GESTURE RECOGNITION AND LANGUAGE INTERPRETATION

Group Members(sign): NISHANT SHANKAR(67) KUNAL JAGASIA(24) MUKTA CHANDANKI(51) SUNNAY ACHARKAR(01)

	Engineering Concepts & Knowledge (5)	Interpretation of Problem & Analysis (5)	Design / Prototype (5)	Interpretation of Data & Dataset (3)	Modern Tool Usage (5)	Societal Benefit, Safety Consideration (2)	Environment Friendly (2)	Ethics (2)	Team work (2)	Presentation Skills (3)	Applied Engg &Mgmt principles (3)	Life - long learning (3)	Professional Skills (5)	Innovative Approach (5)	Total Marks (50)
Review of Project Stage 1	5	5	4	2	4	2	2	2	2	2	2	3	4	5	44

Comments: Well Conceived project. Needs to adhere to schedule.

Name & Signature Reviewer1

	Engineering Concepts & Knowledge (5)	Interpretation of Problem & Analysis (5)	Design / Prototype (5)	Interpretation of Data & Dataset (3)	Modern Tool Usage (5)	Societal Benefit, Safety Consideration (2)	Environment Friendly (2)	Ethics (2)	Team work (2)	Presentation Skills (3)	Applied Engg &Mgmt principles (3)	Life - long learning (3)	Professional Skills (5)	Innovative Approach (5)	Total Marks (50)
Review of Project Stage 1	5	5	4	2	4	2	2	2	2	2	2	3	4	5	44

Comments: well explained, very good, slightly lagging behind schedule.

Date: 9th Feb, 2019

Name & Signature Reviewer2

Inhouse/ Industry:

Class: D17 A/~~B/C~~
Group No.: 19

Project Evaluation Sheet 2018 - 19

Title of Project: Gesture Recognition and Language Recognition of ISL

Group Members(sign): Sunnay Agharkar(01), Nishant Shankar, Kunal Jagasiya, Mukta Chandanki (51)

	Engineering Concepts & Knowledge (5)	Interpretation of Problem & Analysis (5)	Design / Prototype (5)	Interpretation of Data & Dataset (3)	Modern Tool Usage (5)	Societal Benefit, Safety Consideration (2)	Environment Friendly (2)	Ethics (2)	Team work (2)	Presentation Skills (3)	Applied Engg &Mgmt principles (3)	Life - long learning (3)	Professional Skills (5)	Innovative Approach (5)	Total Marks (50)
Review of Project Stage 1	4	5	4	3	4	2	2	2	2	2	3	3	4	4	44

Comments: Good work, Need to work on Accuracy.

Name & Signature Reviewer1

	Engineering Concepts & Knowledge (5)	Interpretation of Problem & Analysis (5)	Design / Prototype (5)	Interpretation of Data & Dataset (3)	Modern Tool Usage (5)	Societal Benefit, Safety Consideration (2)	Environment Friendly (2)	Ethics (2)	Team work (2)	Presentation Skills (3)	Applied Engg &Mgmt principles (3)	Life - long learning (3)	Professional Skills (5)	Innovative Approach (5)	Total Marks (50)
Review of Project Stage 1	4	4	4	3	5	2	2	2	2	3	3	3	5	4	46

Comments:

Date: 6th March, 2019

Name & Signature Reviewer2

APPENDIX

Paper 1

Gesture Recognition and Language Interpretation for Indian Sign Language

Nishant Shankar[1], Sunmay Agharkar[2], Kunal Jagasia[3], Mukta Chandani [4] and Prof. Sujata Khandaskar [5]

Department of Computer Engineering, University of Mumbai
Vivekanand Education Society's Institute of Technology, Mumbai, India
[1] 2015nishant.shankar@ves.ac.in
[2] 2015sunmay.agharkar@ves.ac.in
[3] 2015kunal.jagasia@ves.ac.in
[4] 2015mukta.chandani@ves.ac.in
[5] sujata.khandaskar@ves.ac.in

Abstract. Indian Sign Language (ISL) is the main method of communication for millions of hearing and speech impaired people in India. In this work, we address the research problem of sign language recognition from continuous video sourced from an online repository such as YouTube. We discuss the importance of pose estimation and facial expression in learning semantic cues in signs, and to this effect propose an ISL recognition pipeline that takes advantage of the aforementioned techniques. To model video sequences effectively, we explore different sequence modelling networks that capture spatiotemporal concepts meaningfully. In addition to classification of signs, we also propose to parse a collection of words into pre-defined grammatical components of a sentence.

Keywords: Deep Learning, pose estimation, facial expression recognition, convolutional neural networks, sign language, sequence modeling

1 Introduction

Sign language is used as the primary medium of communication by millions of hearing and speech impaired people in the world. Sign language consists of words and phrases denoted mainly by hand gestures. However, non-verbal cues like facial expressions and body movement are crucial for its correct interpretation, as they can affect the meaning of the words or phrases.

As a rule of thumb, a lot of countries have their own languages with unique signs. Indian Sign Language (ISL) is practiced mainly in the Indian subcontinent with a variety of dialects based on the region it is used in, due to widespread diversity in local culture and customs.

There is a lack of awareness in the public consciousness about ISL, and this results in a systemic apathy towards the education of the deaf community. Oralism, the system of teaching deaf people to communicate by the use of speech and lip-reading rather than sign language, has caused formalization of sign language to be overlooked. Lack of infrastructure and support for teaching, training and promotion of ISL hinders the entire deaf community from gaining acceptance in society.

There is, however, an increasing amount of support being gathered for the further development and dissemination of ISL in India. The Government of India set up the Indian Sign Language Research and Training Centre in 2015. They also released the first ever ISL dictionary comprising of over 3000 words in 2018.

To enable easier communication between speakers of ISL and persons not well versed in ISL, we propose to develop an application which can detect and interpret ISL in real time. Additionally, our application will also carry out the translation of ISL into local languages.

In this paper, we will look at research conducted across areas such as gesture recognition and facial expression recognition. Both of these components are crucial for the correct recognition of ISL. We will also survey research dealing with recognition of American Sign Language (ASL), focusing primarily on sentence level detection and parsing. Similar studies will be conducted in the area of ISL. Based on the inferences gathered from the utilized techniques, we will propose our own solution to develop an application that interprets ISL.

2 Literature Survey

In this section, we discuss research regarding gesture recognition which involves pose estimation and facial expression recognition. We also review work done on recognizing sign languages in general, and ISL specifically. A detailed review of historical approaches is beyond the scope of this paper, so we restrict ourselves to recent advances made in these fields relevant to our proposed system.

2.1 Research in Sign Language Recognition

M. Zhang [2] and Jiu Huang et al. [3] use non-intrusive methods to capture motion data, and perform sentence level interpretation using deep learning techniques.

DeepASL [2] uses LeapMotion to capture video. In their architecture, they have employed a Hierarchical bidirectional deep recurrent neural network (HB-RNN) and a probabilistic framework based on Connectionist Temporal Classification (CTC) for word-level and sentence-level ASL translation respectively.

The HB-RNN helps model the spatial and temporal dynamics of the ASL characteristics very effectively. There is no pause while translating signs, which makes the application real-time. This modelling can be applied in our project as well, to capture spatio-temporal concepts.

Jiu Huang et al. [3] capture data using a Kinect and use a 2 stream 3-dimensional Convolutional Neural Network(CNN). The 2 streams are used for hand gesture detection and recognition. One stream contains features of the hand's global location and the other consists of local hand movements. They also utilise a Hierarchical Attention Network (HAN) for continuous sign language recognition without temporal segmentation.

2.2 Research on Gesture Recognition

Noroozi et al. [4] have defined a complete framework for automatic emotional body gesture recognition along with person detection and static and dynamic body pose estimation methods both in RGB and 3D. Multimodal approaches that combine speech or face with body gestures for improved emotion recognition are described.

The six basic emotions identified in this paper can be the basis for our emotion recognition model as well. The kinematic model of representing the human body inspires our hand pose estimation model. Deriving structural representation of the hand gestures is an important way of representing features. We also intend to use the emotion body gesture recognition system outlined in the paper, albeit our system will focus purely on RGB videos.

2.3 Research in ISL

Tripathi et al. [5] propose a continuous Indian Sign Language (ISL) gesture recognition system where both the hands are used for performing any gesture. Recognition of sign language from continuous gestures is done by using gradient based key frame extraction method. The features of pre-processed gestures are extracted using Orientation Histogram (OH) with Principal Component Analysis (PCA) is applied for reducing the dimension of features obtained after OH.

Experimental results show that the designed method gives satisfactory results with Euclidean distance and correlation. Results are also tested using a normal webcam and get appropriate results. We can enhance this work by creating the dataset with different background and different illumination conditions. We can apply more appropriate features which incorporate the shape of the hand in the time of acting gestures, the speed of performing each gesture, etc.

Rao et al. [6] propose a methodology which uses selfie stick and front camera to capture videos. As a result, only one hand can be used to make gestures. Here recognition and interpretation depend upon 5 parameters such as hand and head recognition, hand, and head orientation, hand movement, the shape of hand and location of hand and head. They also depend upon the background. Object detection (segmentation) is done by applying gradient masking to the image. It divides the videos into 220 frames.

The feature sign matrix inputs a classifier. Since speed is the prime constraint during mobile implementation, it will be reasonable to use minimum distance classifier (MDC). The model of 3 layered artificial neural network is presented. Euclidean, normalized Euclidean and Mahalanobis distance metrics are used to classify sign features.

3 Dataset

There are relatively few datasets pertaining to ISL as opposed to American Sign Language (ASL) or German Sign Language (DSG). Our dataset consists of videos from the Indian Sign Language Research and Training Centre YouTube channel. [7] They have compiled around 1000 videos of multiple people making ISL signs of common words and proper nouns which can come up in a conversation. Other datasets such as IITA ROBITA [8] have been compiled, but our system aims to use datasets from a source such as YouTube in hopes that future data can be crowd sourced.

4 Proposed Model

We propose a web application which uses the webcam / front camera to capture video. The user makes the signs while facing the camera, and the interpreted text will be shown on the screen in English. For the user's convenience, he/ she will also be able to translate that text into any other local language of his/ her choice. Our solution consists of these parts -

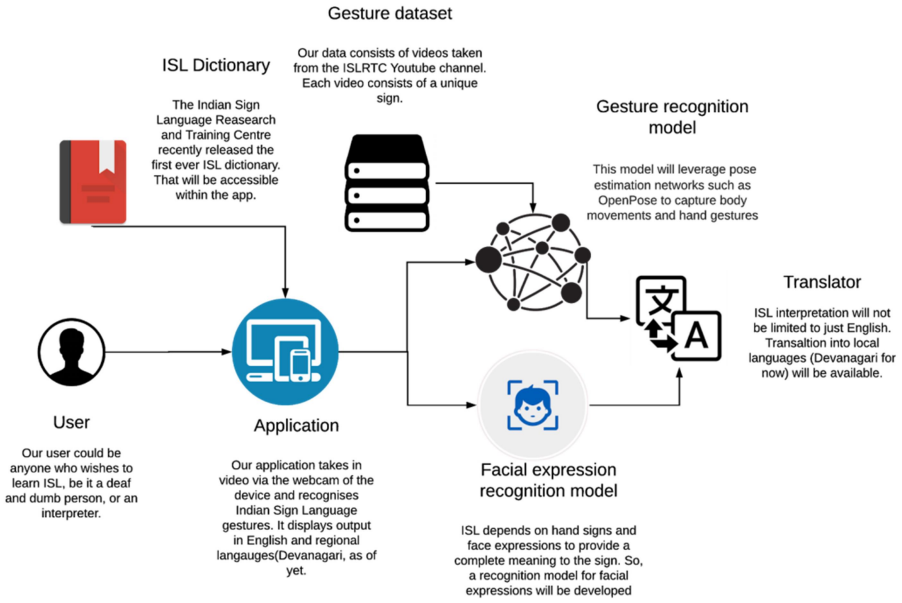


Fig. 1. System Architecture

4.1 Handling non-verbal parameters

Our application consists of two separate models to extract features from the hand gestures and the facial expressions respectively. These models will be trained on data from the ISLRTC YouTube channel [7]. The facial expression recognition model could be developed using a 3-dimensional Convolutional Neural Network. For hand gesture recognition, hand pose estimation networks seem promising and offer a relatively rich skeletal structure of the hands. Instead of just capturing hand gestures and facial emotions separately, systems like OpenPose [9] can be used which allow us to train pose estimation networks that predict subtler features such as facial key points and body language.

4.2 Sequence Modeling

The extracted features need to be modeled sequentially because a sequence of frames would constitute a word. So to provide a sequential flow, the features will be then passed through a Recurrent Neural Network (RNN) which would produce the output probability for each word present in the dataset.

There are multiple possibilities of modeling such a sequence, not just restricted to RNNs. In our proposed system, we will evaluate multiple networks such as 3D CNNs [10], convLSTMs [11] and Trellis networks [12]. The word with the maximum probability would be displayed on the screen and passed to the parser.

4.3 Grammar

We will make use of existing grammar structures prevalent in sign language communication to generate the sentence from the interpreted words, due to a lack of sentence data in our dataset.

A typical sentence in sign language consists of signs conveying time(tenses), topic(subject), comment(action) and referrals. We will generate a sentence using the data obtained from the sequence of signs using this structure.

This sentence will be translated into the regional language using Google Cloud Translation API. The application would also provide a built-in Indian Sign Language dictionary [7] for the users who wish to learn ISL. We aim to follow the REST API philosophy while building the application, and intend to expose the API to other developers as well.

5 Conclusion

Despite being used by a significant section of India's population every day, ISL remains relatively unknown to the masses. However, research in ISL has been gathering pace, and more data is available due to recent efforts by the Indian government and contributions of NGOs.

In this paper, we discuss sign language recognition from continuous video streams. We also review recent work in gesture recognition and facial expression recognition relevant to our proposal.

In our proposed solution, we use data from the ISLRTC YouTube channel, which consists of multiple speakers signing words in ISL as the training data for our classification model. Pose estimation and facial expression recognition will help us recognize non-verbal parameters and act as implicit grammatical markers. We also explore multiple possibilities in model spatiotemporal concepts effectively. Using known grammatical structures in the sentences while parsing a set of words will prove helpful for the hearing and speech impaired users to communicate with hearing persons. We also propose a translation facility that will increase our prospective user base, enabling interpretation of sign language among numerous communities.

References

1. ISL Dictionary Launch, http://www.islrtc.nic.in/isl-dictionary-launch
2. B. Fang, J. Co, and M. Zhang, "DeepASL," Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems - SenSys '17, 2017.
3. Huang, Jie, Wengang Zhou, Qilin Zhang, Houqiang Li and Weiping Li. "Video-based Sign Language Recognition without Temporal Segmentation." CoRR abs/1801.10111 (2018): n. pag.
4. C. Corneanu, F. Noroozi, D. Kaminska, T. Sapinski, S. Escalera, and G. Anbarjafari, "Survey on Emotional Body Gesture Recognition," IEEE Transactions on Affective Computing, pp. 1–1, 2018.
5. K. Tripathi and N. B. G. Nandi, "Continuous Indian Sign Language Gesture Recognition and Sentence Formation," *Procedia Computer Science*, vol. 54, pp. 523–531, 2015.
6. Rao, G. Ananth, and P.v.v. Kishore. "Selfie Video Based Continuous Indian Sign Language Recognition System." Ain Shams Engineering Journal, 2017, doi:10.1016/j.asej.2016.10.013.
7. ISLRTC New Delhi, https://www.youtube.com/channel/UC3AcGIlqVI4nJWCwHgHFXtg
8. Nandy, Anup, Soumik Mondal, Jay Shankar Prasad, Pavan Chakraborty and Gora Chand Nandi, "Recognizing & interpreting Indian Sign Language gesture for Human Robot Interaction." 2010 International Conference on Computer and Communication Technology (ICCCT) (2010): 712-717.
9. Cao, Zhe, Tomas Simon, Shih-En Wei and Yaser Sheikh, "Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields." 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017): 1302-1310.
10. Ji, Shuiwang, Wei Xu, Ming Yang and Kai Yu, "3D Convolutional Neural Networks for Human Action Recognition." IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (2010): 221-231.
11. Gammulle, Harshala, Simon Denman, Sridha Sridharan, and Clinton Fookes. "Two stream lstm: A deep fusion framework for human action recognition." In Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on, pp. 177-186. IEEE, 2017.
12. Bai, Shaojie, J. Zico Kolter and Vladlen Koltun,. "Trellis Networks for Sequence Modeling." CoRR abs/1810.06682 (2018): n. pag.

ICICCT-2019 submission 289 Inbox ×

ICICCT-2019 <icicct2019-0@easychair.org> Sat, Mar 30, 3:04 PM
to me

Dear authors,

We received your paper:

Authors : Sunmay Agharkar, Nishant Shankar, Kunal Jagasia, Mukta Chandani and Sujata Khandaskar
Title : Gesture Recognition and Language Interpretation for Indian Sign Language
Number : 289

The paper was submitted by Sunmay Agharkar
<2015sunmay.agharkar@ves.ac.in>.

Thank you for submitting to ICICCT-2019.

Best regards,
EasyChair for ICICCT-2019.