International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT217374

Gesture Based Real-time Indian Sign Language Interpreter

Akshay Divkar, Rushikesh Bailkar, Dr. Chhaya S. Pawar
Department of Computer Engineering, Datta Meghe College of Engineering Navi Mumbai, India

Article Info
Volume 7, Issue 3
Page Number: 387-394

Publication Issue :
May-June-2021

Article History
Accepted : 20 May 2021
Published : 29 May 2021

ABSTRACT

Hand gesture is one of the methods used in sign language for non-verbal communication. It is most commonly used by hearing & speech impaired people who have hearing or speech problems to communicate among themselves or with normal people. Developing sign language applications for hearing impaired people can be very important, as hearing & speech impaired people will be able to communicate easily with even those who don't understand sign language. This project aims at taking the basic step in bridging the communication gap between normal people, deaf and dumb people using sign language. The main focus of this work is to create a vision based system to identify sign language gestures from the video sequences. The reason for choosing a system based on vision relates to the fact that it provides a simpler and more intuitive way of communication between a human and a computer. Video sequences contain both temporal as well as spatial features. In this project, two different models are used to train the temporal as well as spatial features. To train the model on the spatial features of the video sequences a deep Convolutional Neural Network. Convolutional Neural Network was trained on the frames obtained from the video sequences of train data. To train the model on the temporal features Recurrent Neural Network is used. The Trained Convolutional Neural Network model was used to make predictions for individual frames to obtain a sequence of predictions. Now this sequence of prediction outputs was given to the Recurrent Neural Network to train on the temporal features. Collectively both the trained models i.e. Convolutional Neural Network and Recurrent Neural Network will produce the text output of the respective gesture.

Keywords : Language Interpreter, Convolutional Neural Network, Recurrent Neural Network

I. INTRODUCTION

Gestures are naturally performed by humans. Gestures are produced as part of deliberate actions, signs or signals, or subconsciously revealing intentions or attitude. They may involve the motion of all parts of the body, but the arms and hands, which are essential for action and communication, are often the focus of studies. Facial expressions are also considered gestures and provide an important role in communication. Gestures are present in most daily human actions or activities, and participate in human communication by either complementing speech or substituting themselves to spoken language in environments requiring silent communication (underwater, noisy environments, secret communication, etc.) or for people with hearing disabilities.

II. PROBLEM DEFINITION

The persistent problem in the present Indian Sign Language Recognition is that all the implemented systems rely on static gesture recognition which is very slow or not handy to be used by speech impaired people. In all the existing systems they use a single gesture for a single character which takes much time and isn't efficient working of the system. Speech impaired people use hand signs to communicate, hence normal people face problems in recognizing their language by signs made. Hence there is a need for systems which recognize the different signs and conveys the information to the normal people. Aim of this project is to develop a concept of virtual talking system without a sensor for people who are in need, this concept achieved by using image processing and human hand gesture input. This mainly helps people who can’t talk with other people

III. AIM AND OBJECTIVES

- Implementing a single gesture for a word and single gesture for a phrase so that it will be more natural.
- The problem statement revolves around the idea of a camera- based sign language recognition system, so that the end product will be Cost efficient.
- Objective of this project is to design a solution that is intuitive, simple and user friendly.
- Communication for normal society is not difficult. It should be the same way for the Speech impaired individuals.

IV. Findings

Sign Language Recognition Application Systems is developed in two steps, data acquisition and classification. There are two data acquisition methods that are often used by researchers, camera and Microsoft Kinect. Some use cameras for their Sign Language Recognition Systems. The main advantage from using a camera is that it removes the needs of sensors in sensory gloves and reduces cost from building the system. The camera is quite cheap and is available in almost all laptops. Some systems uses high specification cameras because of the blur caused by web cameras. But even though it is a high specification camera, it is still available on most smartphones. High specification cameras are used to acquire the detailed data they need. The disadvantage of using a web camera, or simply a camera, is that good image pre-processing of obtaining the feature is needed. The Microsoft Kinect is the other popular method used by researchers to acquire their data. Microsoft Kinect is getting more popular among researchers as it provides more data and it is needed by researchers. Kinect sensor gives an image with depth to acquire their data. The advantage of using Kinect is that it provides the depth data of the video stream. The depth data is very useful as it can easily distinguish the background and the signer. Furthermore, it can be used to distinguish hands and body as the signer usually performs sign language by hands in front of their body. The disadvantage is that the Microsoft Kinect device is costly and it should be connected to the computer. Another technique of simple camera and color gloves to differentiate both hands and ease the feature extraction process. Glows are using 3-axis accelerometers and flex sensors. All of these gloves are equipped with sensors attached to the gloves. The advantage is that it provides all the data needed more accurately as it also provides finger movement data. The disadvantages are that they are costly and are difficult to be used commercially. There are many existing systems of SLR most of which are based on static gesture recognition for various spoken languages and there are very few which are based on dynamic gestures but only for American Sign Language.

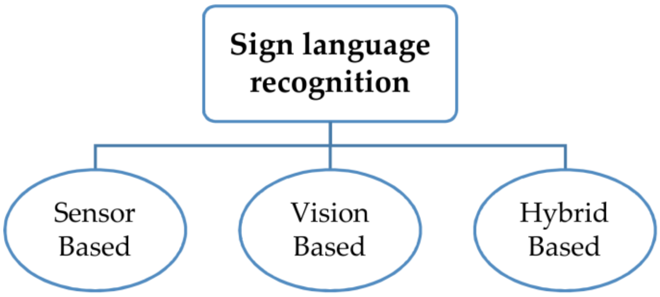

Figure 1 Sign Language Recognition Approaches

There are different implementations of Sign Language recognition. Which are based on the hardware used and the type of image sensor are used. The use of different hardware implementations results in different system architectures. So the methodologies are also changed:

A. Hardware Based

There are different implementations of Sign Language recognition. Which are based on the hardware used and the type of image sensor are used. The use of different hardware implementations results in different system architectures. So the methodologies are also changed.

i. Glows based (Sensor Based)

In the glove based system, sensors such as a potentiometer, accelerometers, etc. are attached to each of the fingers. Based on their readings the corresponding alphabet is displayed. It is expensive to implement, not handy to use, and requires mapping every time a new user wears it. A major advantage of glove-based systems over vision-based systems is that gloves can directly report relevant and required data (degree of bend, pitch, etc.) in terms of voltage values to the computing device, thus eliminating the need to process raw data into meaningful values.

ii. Image sensor based (Vision based)

Vision-based systems use cameras as primary tools to obtain the necessary input data. The main advantage of using a camera is that it removes the need for sensors in sensory gloves and reduces the building costs of the system. Cameras are quite cheap, and most laptops use a high specification camera because of the blur caused by a web camera. A well implemented system can give good results.

iii. Gloves and Image sensor based (Hybrid based)

The third method of collecting raw gesture data employs a hybrid approach that combines gloves and camera-based systems. This approach uses mutual error elimination to enhance the overall accuracy and precision. However, not much work has been carried out in this direction due to the cost and computational overheads of the entire setup. Nevertheless, augmented reality systems produce promising results when used with a hybrid tracking methodology.

B. Gesture Type Based

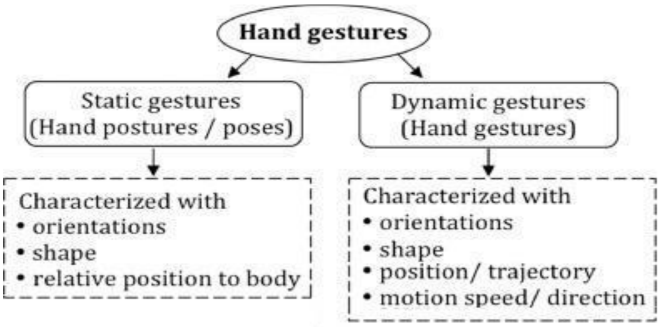
Hand gestures
Static gestures (Hand postures / poses)
Dynamic gestures (Hand gestures)
Characterized with
• orientations
• shape
• relative position to body
Characterized with
• orientations
• shape
• position/ trajectory
• motion speed/ direction

Figure 2 Gesture Types

i. Single Point Gesture Recognition (static)

Static gestures are those that only require the processing of a single image at the input of the classifier, the advantage of this approach is the lower computational cost. It is referred to as 2D gesture recognition. The working of a static gesture based system is shown in Figure 3.

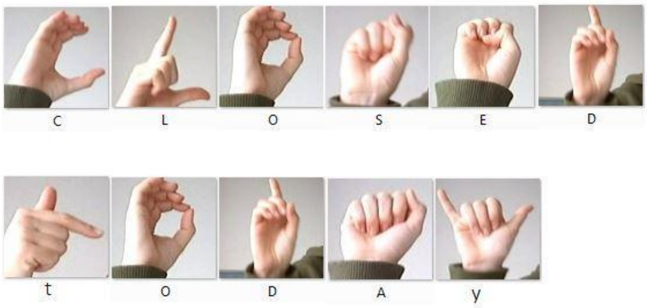

Figure 3 Static Gesture example

ii. Dynamic Gesture Detection

Dynamic gestures require the processing of image sequences and more complex gesture recognition approaches. It is referred to as 3D gesture recognition. The working of dynamic gesture based system is shown in Figure 4.

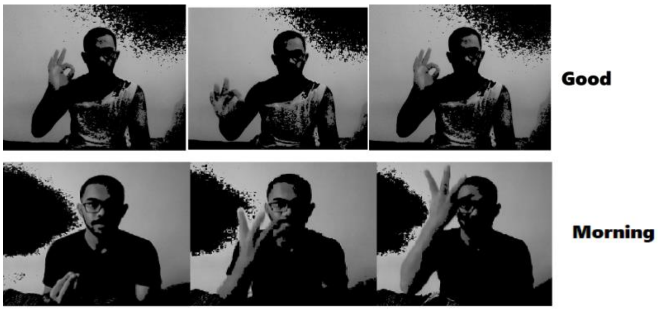

Figure 4 Dynamic Gesture Example

V. LITERATURE REVIEW

[1] Motionlets Matching With Adaptive Kernels for 3-D Indian Sign Language Recognition.

A model for recognizing gestures of Indian sign language 3D motion captured data is presented. The model builds a two phase algorithm that handles multiple attributes of 3D sign language motion data for machine translation. In phase–I, the unordered 3D sign database is restructured into a 4–class structured motionless database from the measured trajectories of motion segmented 3D joints. Each action in a signed frame is motion segmented into motion joints and non-motion joints. Phase–II extracts the shape and orientation of 3D motionless by applying joint relative distance and joint angle measurements respectively. Three feature kernels based on trajectories, finger shape, and their orientations are constructed, which measure the similarity between the query signs and the database signs. It is observed that the motionlet based adaptive kernel matching algorithm on 500 class 3D sign language data gives better classification accuracies compared to state–of–the–art action recognition models.[1]

[2] Real-Time Recognition of Indian Sign Language.

The system for recognizing real-time Indian Sign Language (ISL) portrays an impressive role in enhancing casual communication among people with hearing disabilities and normal persons. Though FCM is efficient, it requires more computation time than the others. Also, for high dimensionality datasets, most of the traditional algorithms suffer. Hence it is planned to extend the system by combining Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) to capture the spatial and temporal features. The developed system is capable of recognizing 40 words of ISL in real-time while similar systems have the capability to recognize static gestures only. For the static gesture system, FCM is more efficient and reliable.[2]

[3] The VirtualSign Channel for the Communication Between Deaf and Hearing Users

This article details the improvements and current structure of the VirtualSign platform, a bidirectional sign language to text translation tool in development.

The platform has two main components, sign to text and text to sign, that are both described. Translation from text to sign relies on a 3D avatar. Translation from sign to text relies on a set of data gloves. Two different data glows are used for testing, which is 5TD Data Glows and IFG Data Glows. The translator consists of two modes of workings, the "word mode", for the regular translation defined in previous paragraphs, and the "spelling mode". This mode is activated automatically when the user performs three distinct alphabet hand gestures. Results of the system show that the 5DT data gloves perform significantly better than the IFG data gloves. On the 20, 50, and 100 words dataset.[3]

By reviewing papers it is found out many solutions revolve around inefficient and high cost hardware infrastructure. With the intention to make hardware reach every user we must focus on reducing the cost of hardware infrastructure. There is No dataset available for dynamic recognition systems for Indian Sign Language. In India, there is no universal sign language. Though there exist many Sign Languages, normal people do not know about sign languages. Gesture recognition and sign language recognition has been a well-researched topic for American Sign Language but has been rarely touched for its Indian equivalent.

VI. METHODOLOGY

In this project, spatial features for individual frames were extracted using the inception model (CNN) and temporal features using RNN. Each video was then represented by a sequence of predictions made by CNN for each of their individual frames. This was given as input to the RNN. For every video corresponding to each gesture, frames were extracted and the background body parts other than hands were removed to get a grayscale image of hands which avoided color-specific learning of the model. Frames of the training set were given to the CNN model for training on the spatial features. The obtained model was then used to make and store predictions for the frames of the training and test data. The predictions corresponding to the frames of the training data were then given to the LSTM RNN model for training on the temporal features. The pool layer gives a 2048-dimensional vector that represents the convoluted features of the image, but not a class prediction. Once the RNN model was trained, the predictions corresponding to the frames of the test data were fed to it for testing.

The dataset for Indian Sign Language (ISL) gestures was not officially available, it was prepared by us. The prepared dataset contains around 456 videos for 38 gestures (labels) categories. Four non-expert subjects executed the 4 repetitions of each gesture, i.e., 12 videos per gesture. All of these video samples were used for training, and for testing, we had created a live camera feed as input. Thus, the training dataset had 456 videos and for the testing, we had run 5 samples per gesture of live video input.

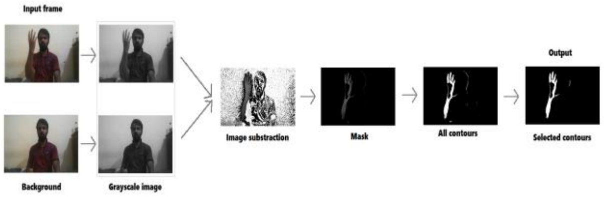

Figure 5 Image Processing

The input image is converted first converted to a grayscale image then the grayscale background image subtracted from the input image. For background subtraction Median based method is used. Median Filtering is a non-recursive approach, in which background is extracted based on the median value of pixels in the buffer. In median filtering, the correct selections of the buffer size (n) and frame time rate(Δt) are critical issues that affect the performance of median filtering.[4] By applying image thresholding to the grayscale image we obtain a binary image. The binary image is the resultant image of adaptive thresholding as it depicts the differences between different threshold values. The white region describes values less than the threshold and the black region describes values greater than the threshold value.[5] The median is used as a threshold with a tolerance of 20 for masking the input image. After the application of the threshold, we look for major contours in the image. we select maximum 3 contours from the image. After this, a segmented hand is obtained. This complete process in shown in fig. 6.

A. Algorithms Used

Video classification is a challenging problem as a video sequence contains both temporal and spatial features. Spatial features are extracted from the frames of the video, whereas temporal features are extracted by relating the frames of video in a course of time. We have used the pool layer approach to train our model on each type of feature. To train the model on spatial features, we have used CNN, and for the temporal features, we have used a recurrent neural network.

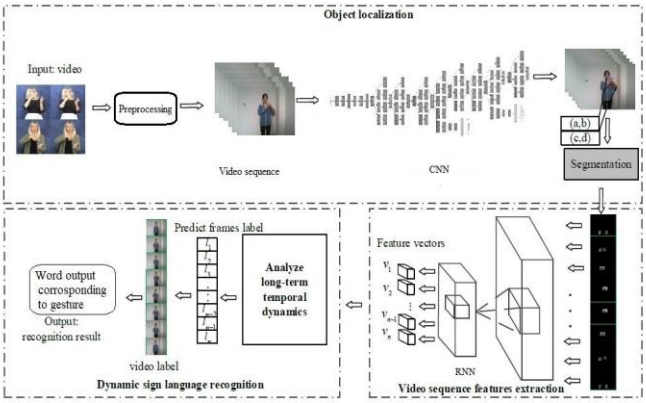

Figure 6 System Architecture

a. Convolutional Neural Network

Convolutional neural networks or ConvNets are great at capturing local spatial patterns in the data. They are great at finding patterns and then use those to classify images. ConvNets explicitly assume that input to the network will be an image. CNNs, due to the presence of pooling layers, is insensitive to rotation or translation of two similar images; i.e., an image and its rotated image will be classified as the same image. Due to the vast advantages of CNN in extracting the spatial features of an image, we have used a model of the TensorFlow library which is a deep ConvNet to extract spatial features from the frames of video sequences. Inception is a huge image classification model with millions of parameters for images to classify.

b. Recurrent Neural Network

The sequence itself has the information, and recurrent neural networks (RNNs) use this for the recognition tasks. The output from an RNN depends on the combination of current input and previous output as they have loops. One drawback of RNN is that, in practice, RNNs are not able to learn long-term dependencies. Hence, our model used Long Short-Term Memory (LSTM), which is a variation of RNN with LSTM units. LSTMs can learn to bridge time intervals in excess of 1000 steps even in case of noisy, incompressible input sequences.

VII. OUTPUT

It is seen that the more we train the model the accuracy is getting increased. 90% of the data was considered as train set and remaining was for validation, for 400 training steps. The final test accuracy achieved is 54.2%.

Training Steps	Train Accuracy	Cross Entropy	Validation
100	50.00%	2.576236	20.00%
200	60.00%	2.31399	20.00%

300	90.00%	1.555923	30.00%
400	90.00%	1.505622	50.00%

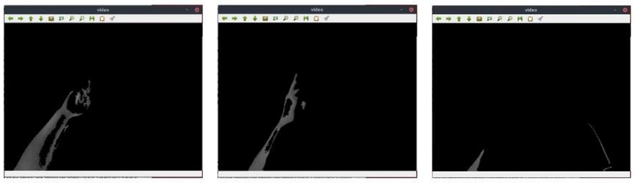

Figure 7 Gesture frames in sequence in input window(which are image processed)

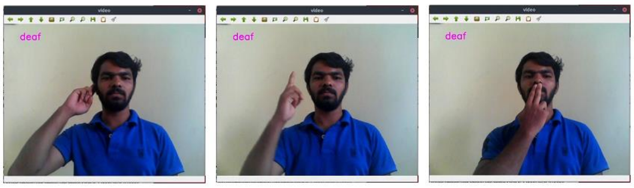

Figure 8 Gesture frames in sequence in output window

VIII. CONCLUSION

In this paper, we have proposed a vision-based system coupled with CNN and RNN to interpret isolated hand gestures from the Indian Sign Language. This work used a pool layer approach to classify spatial and temporal features. CNN was used to obtain the spatial features, whereas RNN was used to classify on the temporal features. The system reached to an accuracy of 54.2% by using a self-made Indian sign language dataset. This shows that CNN along with RNN can be successfully used to learn spatial and temporal features and classify sign language gesture videos.

There are several other directions for future work. First, this work can be further extended in recognizing continuous sign language gestures with better accuracy. This method for individual gestures can also be extended for sentence level sign language. Another promising research path is that the current process uses two different models, training CNN followed by training RNN. Future work can be focused on combining the two models into a single hybrid model. Also, other sequence learning approach can be developed, such as attention-based methods, to make better use of the temporal dependencies.

IX. ACKNOWLEDGMENT

Motivation and guidance are the keys towards success. I would like to extend my thanks to all the sources of motivation.

We would like to grab this opportunity to thank Dr. S. D. Sawarkar, principal for the encouragement and support he has given for our project.

We express our deep gratitude to Dr. A. P. Pande, head of the department who has been the constant driving force behind the completion of this project.

We wish to express our heartfelt appreciation and deep sense of gratitude to my project guide Dr. C. S. Pawar for her encouragement, invaluable support, timely help, lucid suggestions and excellent guidance which helped us to understand and achieve the project goal. Her concrete directions and critical views have greatly helped us in successful completion of this work.

We extend our sincere appreciation to all professors for their valuable inside and tips during the designing of the project. Their contributions have been valuable in so many ways that we find it difficult to acknowledge them individually.

We are also thankful to all those who helped us directly or indirectly in completion of this work.

X. REFERENCES

[1]. P. V. V. Kishore, D. A. Kumar, A. S. C. S. Sastry and E. K. Kumar, "Motionlets Matching With Adaptive Kernels for 3-D Indian Sign Language Recognition," in IEEE Sensors Journal, vol. 18, no. 8, pp. 3327-3337, 15 April15, 2018, doi: 10.1109/JSEN.2018.2810449.

[2]. H. Muthu Mariappan and V. Gomathi, "Real-Time Recognition of Indian Sign Language," 2019 International Conference on Computational Intelligence in Data Science (ICCIDS), Chennai, India, 2019, pp. 1-6, doi: 10.1109/ICCIDS.2019.8862125.

[3]. T. Oliveira, N. Escudeiro, P. Escudeiro, E. Rocha and F. M. Barbosa, "The VirtualSign Channel for the Communication Between Deaf and Hearing Users," in IEEE Revista Iberoamericana de Tecnologias del Aprendizaje, vol. 14, no. 4, pp. 188-195, Nov. 2019, doi: 10.1109/RITA.2019.2952270.

[4]. M. Hedayati, W. M. D. W. Zaki and A. Hussain, "Real-time background subtraction for video surveillance: From research to reality," 2010 6th International Colloquium on Signal Processing & its Applications, 2010, pp. 1-6, doi: 10.1109/CSPA.2010.5545277.

[5]. P. Roy, S. Dutta, N. Dey, G. Dey, S. Chakraborty and R. Ray, "Adaptive thresholding: A comparative study," 2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), 2014, pp. 1182-1186, doi: 10.1109/ICCICCT.2014.6993140.

Cite this article as :

Akshay Divkar, Rushikesh Bailkar, Dr. Chhaya S. Pawar, "Gesture Based Real-time Indian Sign Language Interpreter ", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 7 Issue 3, pp. 387-394, May-June 2021. Available at

doi : https://doi.org/10.32628/CSEIT217374

Journal URL : https://ijsrcseit.com/CSEIT217374