International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 05 | May - 2023 Impact Factor: 8.176 ISSN: 2582-3930

Indian Sign Language Recognition: A Survey

Utkarsh Jagtap[1], Vinayak Nangnurkar[2], Suhas Chalwadi[3], Neelam Jadhav[4]

Department of Computer Engineering
Genba Sopanrao Moze College of Engineering, Balewadi, Pune 45

Abstract - This paper discusses the literature review of sign language recognition systems, which is an important research area in the field of computer vision and machine learning. The focus of the review is on Indian Sign Language (ISL) recognition systems, which have been developed by various practitioners with varying scopes. The reviewed literature includes five papers that use different approaches for feature extraction and recognition, such as clustering and deep learning algorithms. The data sets used in the reviewed literature have distinct qualities from one another in terms of input modality, sample size, and words or signs included in the dataset. The paper provides a detailed analysis of the data presented in the reviewed literature. Overall, this review aims to provide a comprehensive understanding of the current state of research in ISL recognition systems and to highlight the areas that require further research.

Key Words: Mediapipe , CNN, Indian sign language , Deep learning, Clustering

1. INTRODUCTION

The use of sign languages in communication has been an important area of research in recent years. In this paper, we survey various literature on the topic of sign language recognition using machine learning techniques. We examine the scope of the task, the data sets used, and the approaches employed by different practitioners. The common goal of the surveyed literature is to recognize the alphabets of Indian Sign Language (ISL), which may serve as a common ground between signed language and spoken or literary language. However, the scope of the task varies from recognizing static signs to continuous sentences made up of dynamic signs. The data sets used in the literature also differ in terms of sample size per sign, input modality, and the inclusion of specific words or clothing. We analyze the features of the data sets and discuss their relevance to the task. While the technical aspects of the data collection process are disregarded in our analysis, we provide insights into the common approaches chosen by practitioners, such as clustering and deep learning models for prediction. The survey serves as a comprehensive overview of the research on sign language recognition using machine learning techniques, and provides a basis for future work in the field.

2.TAXONOMY

2.1 SCOPE

Various practitioners have performed this task with discrete scope, Shenoy et al.[3] were able to recognise 33 signs including 23 alphabets and 10 numbers, and can classify 13 gestures, the data set includes static images captured from webcam and smartphone, grid based method is used to extract features. k-NN is used to recognise static signs, to recognise dynamic signs Hidden Markov Model is used. Muthu Mariappan et al.[2] has used fuzzy clustering approach, for 80 words that can formulate 50 sentences, and has capability to recognise 40 dynamic signs in real time over its contemporary static gestures detection system. Regions of Interests are extracted from images and are used for prediction by clustering them. Teja Mangamuri et al.[5] has presented their goal as to creating a ISL dataset for benchmarking the classification of machine learning models, they have used their own dataset of 26 gestures made up of 350 images for per gesture with total of 9100 images. Main concern of this work is to recognise signs that require two hands. Mittal et al.[1] stated their goal as to recognise continuous signs that is to recognise sentences made of sequence of gestures instead of static gestures, that are standalone, they have tested their proposed

model on 942 sentences made up of 35 different signed words. They have used hardware accessory called Microsoft Kinetic for capturing images of hands, then the features that are gathered are passed to the deep learning algorithms to predict the signs. Sridhar et al.[4] aims to address the lack of publicly available dataset of ISL, they have proposed dataset consisting of total 4000 video of 263 signs from 15 word categories, this is relatively the largest dataset amongst the compared literature in terms of number of signs. Open Pose from OpenCV library is used for extracting features from each frame. Then the deep learning algorithm is used to recognise these signs.

2.1.1 DISCUSSION

The scope of surveyed literatures varies from each other, the common intention that we have seen is to recognise alphabets of ISL, the efforts are made to recognise alphabets might have motive to bridge the gap between signed language and spoken or literary language, alphabets seems to be common ground between these two types of Languages. It is general knowledge that the ISL signers do not communicate by formulating the words by alphabets with their respective signs, instead they rely on actual sign or gestures that are associated with the intention or thought they want to communicate. This has been taken into consideration by [1], their proposed system intended to recognise setences made up of continuous signs. The dataset used by [4], also consist of dynamic gestures of words, over the alphabets. Clustering is the common approach chosen by [2,3,5] and [1,4] have used deep learning models for prediction.

2.2 DATASET

The data sets used in the surveyed literature has rather distinct qualities from one another. Following section describe the nature of data used in the surveyed literature and its features like number of samples per sign followed by a brief discussion.

Following is more concrete analysis of the data presented in table 1. Upon the closer inspection on the above data, we have gathered following features.

The data presented in the tables above have been curated with the common features amongst the all the data sets in the surveyed literature. Considering the scope of the task, practitioners have used type of the data that would match with their needs, [2,3,5] have collected in such a way that the task of feature extraction would be easy. [5] has proposed their work as to set up benchmarking, even though the dataset they have used lacks the variance. The dataset used by [2] has great variance, it is created by 10 distinct signers, though sample size per sign is 10 which is lowest amongst the surveyed literature. [1,4] has strikingly similar sample size per word, both of them have used deep learning approach for this task, though they both have used different feature extraction techniques.

Year	Reference	Input Modality	Dataset	Words	Sample Size (No. of frames)
2019	Muthu Mariappan et al.[2]	RGB dynamic	Own	80	800
2018	Shenoy et al. [3]	RGB static and dynamic	Own	65	24624
2019	Teja Mangamuri et al.[5]	RGB static	Own	26	9100
2019	Mittal et al.[1]	RGB dynamic	Own	35	3150
2020	Sridhar et al.[4]	RGB dynamic	Own	263	4283

Table 1: Overview of Data sets

Reference	Words	Has alphabets	Signer is wearing specific cloths	Sample size	Samples per Sign or Word
Muthu Mariappan et al.	80	No	Yes	800	10
Shenoy et al.	65	Yes	Yes	24624	378
Teja Mangamuri et al.	26	Yes	Yes	9100	350
Mittal et al.	35	No	No	3150	90
Sridhar et al.	263	No	No	24624	93

Table 2: Insights of Data sets

2.2.1 DISCUSSION

There are few technical aspects of these data set that are disregarded in the tables above such as the type of equipment that were used while collecting data, lighting conditions of environment at which the data is collected, number of frames that were captured in one second, number of signers that has contributed to data set and their physical nature that might have added additional benefit to the system that could work in different scenarios given that the sample size has variance. This information is disregarded because it was not common amongst all the literature that has been surveyed, only few authors considered it to be of an importance. The common aspect that we have gathered is that the data sets are curated such that the subsequent tasks such as feature extraction and prediction would be easier.

2.3 FEATURE EXTRACTION

The task of extracting the feature is of a great importance for solving Computer Vision tasks. It requires practitioners to generalize the methods such that the features are extracted deliberately. These features are fundamental information that needs to learned and predicted by the models. The following section will have brief description of various feature extraction approach taken for this task, followed by a discussion.

2.3.1 SKIN COLOR SEGMENTATION AND GRID BASED FEATURE EXTRACTION

Shenoy et al.[3], has used this method. First they have eliminated the face from the frames, because face pixel have skin colors, for eliminating face they have used HOG(Histogram Oriented Gradient) method with SVM(Support Vector machine) classifier, then they have converted the image from RGB color space to YUV color space using equation 1 to easily identify the skin color. To classify if the given pixel is skin colored pixel or not they have used 2 after classifying the pixels into skin colored and non-skin colored, they have converted these frames into binary image, to easily identify the largest contour as Hand Region. Once the hand region is identified it is then divided into 4×4 cells, and by using the equation 3 they have calculated feature value for a given sign in the frame.

$$\begin{pmatrix} Y \\ U \\ V \end{pmatrix} = \begin{pmatrix} 0.299 & 0.587 & 0.114 \\ -0.147 & -0.289 & 0.436 \\ 0.615 & -0.515 & -0.100 \end{pmatrix} . \begin{pmatrix} R \\ G \\ B \end{pmatrix} + \begin{pmatrix} 0 \\ 128 \\ 128 \end{pmatrix} \quad (1)$$

$$\begin{Bmatrix} 80 < U < 130 \\ 136 < V < 200 \\ V > U \\ R > 80 \wedge G > 30 \wedge B > 15 \\ |R - G) > 15 \end{Bmatrix} \quad (2)$$

$$\text{Feature Value} = \frac{\text{Area of Hand contour}}{\text{Area of fragment}} \quad (3)$$

2.3.2 HAND SEGMENTATION AND FEATURE EXTRACTION FROM CONTOUR

Muthu Mariappan and Gomathi[2] has in their work used this approach of feature extraction. Initially preprocessing is started on BGR image frames which are converted into HSV color space in order to extract colored region objects easily, at the end of preprocessing, binary images are obtained where white colored pixels represent skin color, and black colored pixels represent the rest. Morphological operation are then performed to reduce the noise from the image. Now the contour from the binary image is easy to identify, they are referred as outline of an object, they are represented as array of (x,y) co-ordinates of boundary point of the object. From the image area of all contour is calculated, and top three contour are selected these contour represent Face, Left and Right Hand, and finally features such as orientation amongst the contour, distance from center to each finger are extracted from Regions of Interest.

2.3.3 HAND SEGMENTATIONAND HOG

This approach is used by Teja Mangamuri et al.[5] , similar approach of converting RGB image to HSV is followed, and images in dataset are deliberately captured in such background, that the background elimination process would be easier, once background is removed the image is dilated to remove black spots. Subsequently, the image is converted into Grayscale image. Now for extracting features HOG(Histogram of Gradient) is used, it does not check characteristic of whole image instead it divide the image in cells, here the cell size is set to 4×4. HOG is calculated by calculating magnitude of gradient for each pixel in the cell, and their angle of gradient. these two values are then classified in 9 bins, each representing range of angles of gradient, according to the angle the gradient calculated for each pixel, it is classified into bins of angles, these bins are collectively called feature vector of a cell, feature vectors of such 4×4 cells are concatenated, this concatenation is performed across the whole image and final feature vector is calculated, it has all the HOG values calculated for each 4×4 cell.

2.3.4 FEATURE EXTRACTION USING OPENPOSE

Features extracted in the work of Sridhar et al.[4] has three primary types.

- *Key Point Vector:* A vector of x- and y-coordinates of each of the key-points for each frame.
- *Pose video:* A frame-wise pixel map of all limbs.
- *PAF(Part Affinity Field) video:* A frame-wise pixel map with PAF aggregations.[4]

OpenPose model from OpenCV library is used to for extracting the features above. It estimates the pose for each given frame, by using bottom up approach, it represents each frame as PAFs, It is two-dimensional vector, that have directions of one part of limb to another part, additionally it provides 135 keypoints for each frame, out of which author has used 96 keypoints, discarding facial and leg keypoints, as Key Point Vector. these 96 keypoints represent position of face, left and right hand in the frame. Resolution of Pose Videos and PAF videos is same as resolution of frames in the videos, although they are significantly sparse. zeroed out pixels in Pose Video are 96.5% and in the PAF videos they are 79.7%.

2.3.5 FEATURE EXTRACTION USING LEAP MOTION DEVICE

This method is used by Mittal et al.[1], Leap Motion device is a hardware accessory, it used for tracking hand movements, it has Infrared camera that can capture 120 frames per second, the device is placed at the surface and from the captured images the, keypoints are generated for the different region of hands, these co-ordinates have 3 axis x,y and z. Once the videos are captured the keypoints from each frame are extracted, here [1]have discarded redundant information such as joints of fingers and wrist,

and finally feature vector of size 12 is extracted. Each of these 12 points, represent keypoints of both hands, 6 points each, including position of 5 fingers and position of palm.

2.3.6 DISCUSSION

The methods of feature extraction varies from each literature, this variety is due to the models that they have used for classification, Muthu Mariappan and Gomathi , Shenoy et al. , Teja Mangamuri , Jain , and Sharmay have used methods that directly deal with fundamental aspects of an image, such as calculating feature values based on some operation performed directly on pixel values. Approach of Mittal et al. , Sridhar et al. is more of top down where they have extracted positional values of region of interests, this approach directly deals with the context of hands, disregarding nature of color in the frame. The data they have collected is subtle and corresponds actual regions of hands.

2.3.7 FIGURES

Following is the list of figures, to better understand the feature extraction techniques used by surveyed practitioners.

Reference	Features	Techniques
Muthu Mariappan and Gomathi[2]	Hands, Face	ROI from contour
Shenoy et al.[3]	Hands	Grid based
Teja Mangamuri, Jain, and Sharmay[5]	Hands	HOG
Mittal et al.[1]	Hands	Leap Motion device
Sridhar et al.[4]	Hands	OpenPose

Table 3: Techniques of Feature extraction

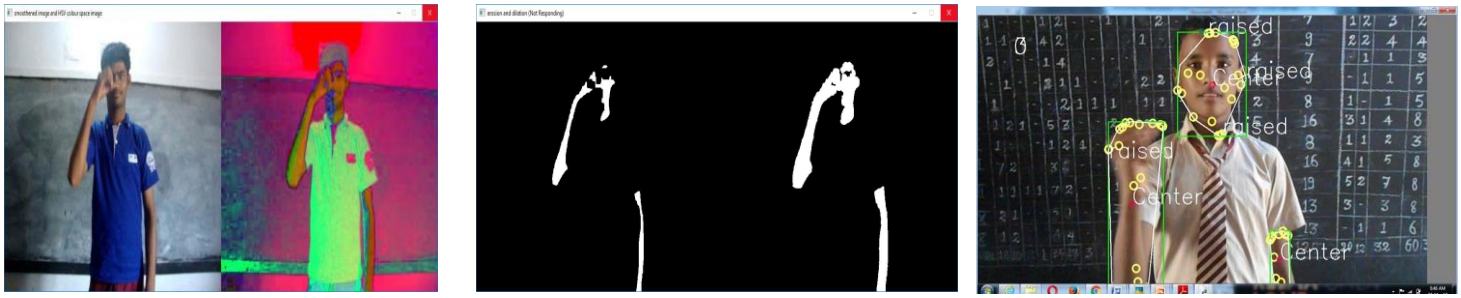

Figure 1: Hand Segmentation and Feature Extraction from Contour [3]

Figure 2: Skin color Segmentation and Grid based feature extraction [2]

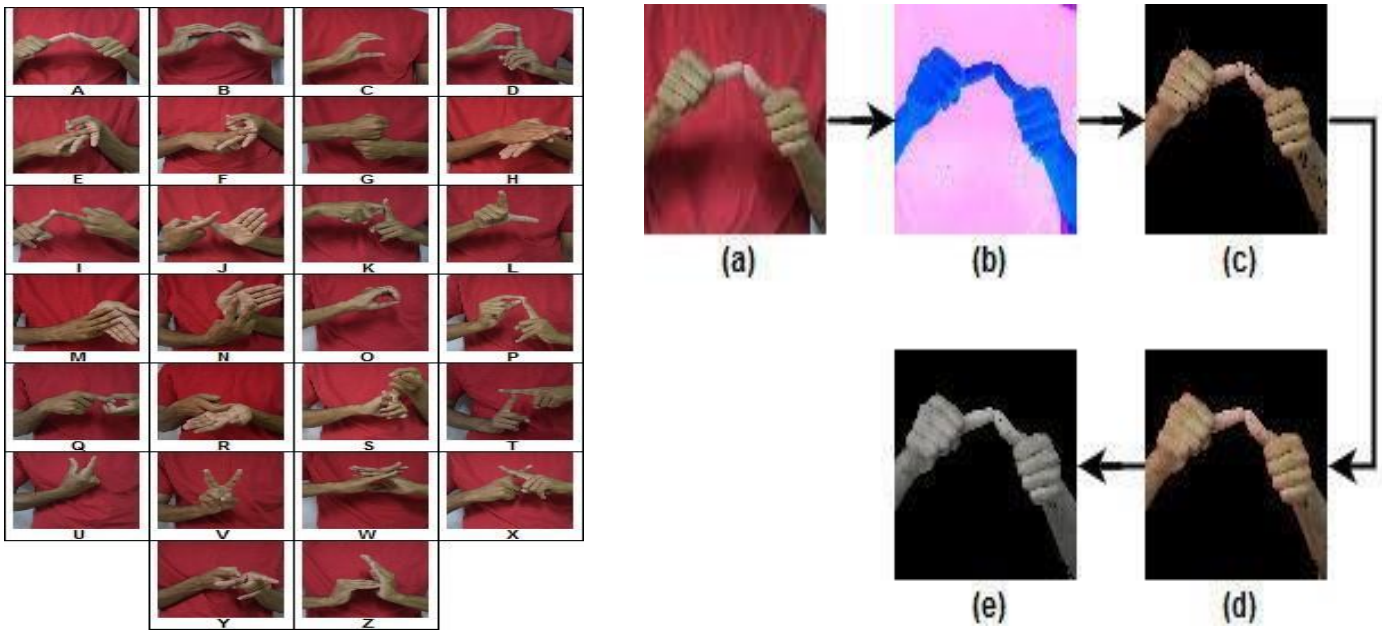

Figure 3: a) Input RGB image, b) HSV image, c) Background removed, d) Black holes removed, e) Grayscale image.[5]

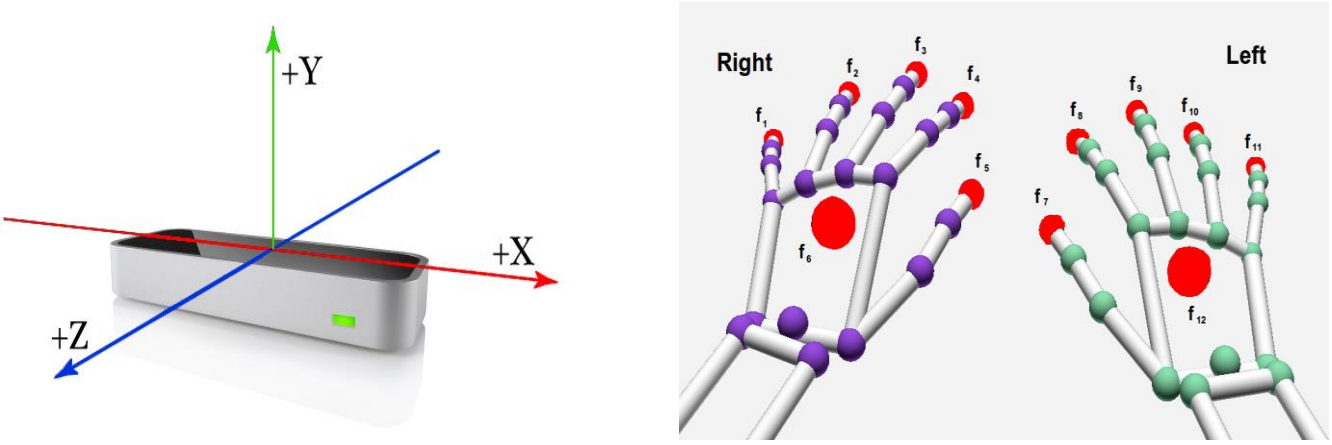

Figure 4: Feature extraction using Leap Motion device [1]

2.4 MODEL

Models used by surveyed practitioners are described in this section followed by a brief discussion. There are two primary classifiers which are used, these are either clustering techniques or deep learning techniques.

2.4.1 CLUSTERING MODELS

Muthu Mariappan et al.[2] have used the Fuzzy clustering model(FCM) which has 75% accuracy. The FCM algorithm groups similar data items by assigning membership values to each data point based on their proximity to the cluster centers. The closer a data point is to a cluster center, the higher its membership value. The sum of membership values for all data points is one. The algorithm updates the cluster centers and membership values iteratively until convergence is achieved, and then returns the resulting cluster centers and membership values for each data point. Using the resulting cluster centers and membership values, the Fuzzy c-means prediction algorithm can classify new data items. The algorithm assigns the new data items to the cluster with the highest membership value for the corresponding data points, which is then used as the gesture ID. Gesture identifications are made using this ID[2].Shenoy et al. have also used clustering model that is k-NN nearest neighbour with the accuracy 99.7%, this accuracy is achieved when they have used 10×10 grid for feature extraction. see Figure 5.

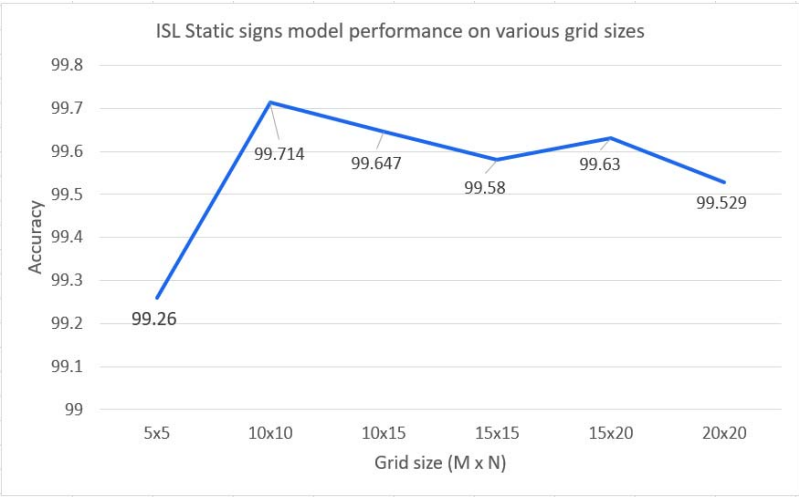

Figure 5: Comparison of accuracy of k-NN classification on features extracted using various grid sizes on hand poses data.

Clustering methods along with various other machine learning techniques are used by Teja Mangamuri et al.[5] out of which SVM has given the highest accuracy of 99.27%.

Model Name	Accuracy
SVM	99.27%
Naive Bayes	96.15%
kNN	99.03%
Decision Tree	91.4%

Table 4: Comparison of Machine Learning models used by [5]

2.4.2 DEEP LEARNING MODELS

Following is the Overview of deep learning model used by Sridhar et al.[4] , see Figure 6 Each frame is passed through OpenPose and the Pose and PAF videos are obtained. A channel-wise concatenation is done and fed through MobileNetV2 model to extract features. The extracted features are fed to a BiLSTM. The hidden states from LSTM cells are flattened and passed through a fully connected layer and a softmax layer for classification.

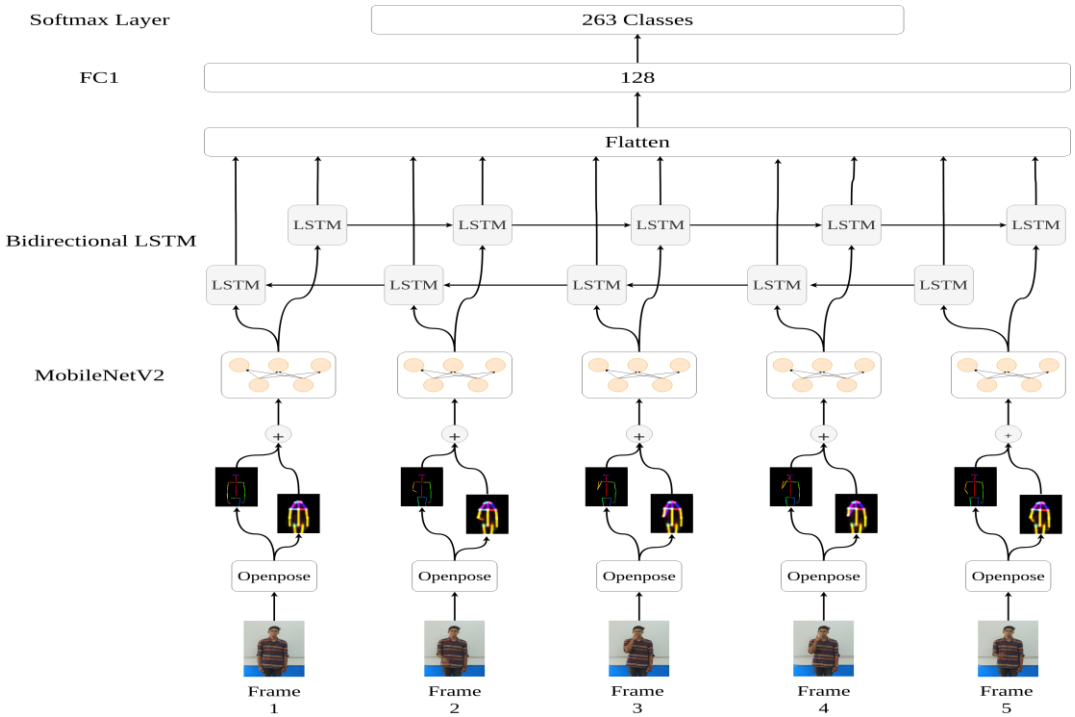

Figure 6: Structure of Model used by [4]

The model used by Mittal et al.[1] is a Modified Long Short-Term Memory (LSTM) neural network. LSTMs are a type of recurrent neural network (RNN) that can process sequential data and are commonly used in natural language processing and speech recognition. The Modified LSTM model used in this paper has two main components: an encoder and a decoder. The encoder takes the input motion data from the Leap motion device and processes it into a compact representation, which is then fed into the decoder. The decoder then generates the corresponding sign language gesture. The encoder component of the model consists of three layers: a 1D convolutional layer, a Bidirectional LSTM layer, and a MaxPooling layer. The 1D convolutional layer is used to extract features from the input motion data, the Bidirectional LSTM layer processes the features in both forward and backward directions to capture contextual information, and the MaxPooling layer reduces the dimensionality of the output. The decoder component of the model consists of two layers: an LSTM layer and a fully connected layer. The LSTM layer takes the output of the encoder and generates a sequence of intermediate representations, which are then passed through the fully connected layer to produce the final output, which is the predicted sign language gesture. The model was trained on a dataset of sign language gestures captured using the Leap motion device. The authors used a combination of mean squared error and categorical cross-entropy as loss functions during training.

2.4.3 DISCUSSION

Both clustering and deep learning models have been used for gesture recognition, achieving high accuracies in various studies. The choice of the model depends on the dataset and problem requirements.

Reference	Sample variation	Model	Accuracy
Muthu Mariappan and Gomathi[2]	10	Fuzzy Clustering	75%
Shenoy et al.[3]	378	k-NN	99.7%
Teja Mangamuri, jain and, Sharmay[5]	350	Various	91.72%
Mittal et al.[1]	90	CNN-LSTM	89.50%
Sridhar et al.[4]	93	CNN-LSTM	85.6%

Table 5: Summarized Accuracy of Models

3. CONCLUSIONS

In conclusion, the field of Indian Sign Language recognition is rapidly evolving with a growing number of practitioners working towards developing more accurate and efficient models. Through this literature review, we have analyzed and compared the data sets and techniques used by various researchers to recognize ISL gestures. The scope of the surveyed literature varies from recognizing static gestures to continuous sign sentences, with varying degrees of complexity and accuracy. It is evident that there is a need for a standardized and larger dataset of ISL gestures that can be used as a benchmark for the development of future models. The data sets used in the surveyed literature are limited in sample size and variance, with some lacking dynamic gestures and certain sign categories. Future work in this area should focus on creating larger and more diverse datasets, exploring new techniques for feature extraction and recognition, and developing models that can be used in real-world applications to benefit the hearing-impaired community.

REFERENCES

[1] Anshul Mittal et al. "A Modified LSTM Model for Continuous Sign Language Recognition Using Leap Motion". In: *IEEE Sensors Journal* 19.16 (2019), pp. 7056–7063. doi: 10.1109/ JSEN.2019.2909837.

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

[3] Kartik Shenoy et al. "Real-time Indian Sign Language (ISL) Recognition". In: *2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. 2018, pp. 1–9. doi: 10.1109/ICCCNT.2018.8493808.

[4] Advaith Sridhar et al. "INCLUDE: A Large Scale Dataset for Indian Sign Language Recognition". In: *Proceedings of the 28th ACM International Conference on Multimedia*. MM '20. Seattle, WA, USA: Association for Computing Machinery, 2020, pp. 1366–1375. isbn: 9781450379885. doi: 10.1145/3394171.3413528. url: https://doi.org/10.1145/3394171.3413528.

[5] Leela Surya Teja Mangamuri, Lakshay Jain, and Abhishek Sharmay. "Two Hand Indian Sign Language dataset for benchmarking classification models of Machine Learning". In: *2019 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT)*. Vol. 1. 2019, pp. 1–5. doi: 10.1109/ICICT46931.2019.8977713.