International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 12 Issue: 05 | May 2025 www.irjet.net p-ISSN: 2395-0072

Indian Sign Language Recognition using Full Body Pose Estimation

Sandesh Dandge[1], Pranav Indore[2], Vinayak Vairat[3], Mahesh Wagh[4], R. B. Murumkar[5]

[1]Department of Information Technology, SCTR's PICT Dhankawadi, Pune
[2] Department of Information Technology, SCTR's PICT Dhankawadi, Pune
[3] Department of Information Technology, SCTR's PICT Dhankawadi, Pune
[4] Department of Information Technology, SCTR's PICT Dhankawadi, Pune
[5] Assistant Professor, Department of Information Technology, SCTR's PICT Dhankawadi, Pune

---

Abstract - *This survey paper presents a comprehensive analysis of recent advancements in Indian Sign Language (ISL) detection systems, emphasizing the role of various machine learning techniques for both static and gesture-based sign recognition. It investigates the performance of multiple model architectures, including Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, Support Vector Machines (SVM), Inception models, Recurrent Neural Networks (RNN), Hidden Markov Models (HMM), and Random Forest classifiers, in accurately classifying sign gestures from video data. The study highlights preprocessing methodologies such as key point extraction and dataset preparation to enhance model training effectiveness. Additionally, the paper discusses the evaluation metrics employed to measure model performance and underscores the importance of selecting models tailored to the unique characteristics of ISL gestures. The findings highlight the necessity for improved recognition systems that are culturally relevant and capable of accommodating the distinctive features of Indian Sign Language, ultimately paving the way for more accessible communication solutions for the Deaf and Hard-of-Hearing (DHH) community in India.*

Key Words: **Indian Sign Language, Gesture Recognition, Machine Learning, Convolutional Neural Networks, Long Short-Term Memory, Random Forest, Hidden Markov Models**.

1. INTRODUCTION

Communication is a fundamental human right, yet individuals within the Deaf and Hard-of-Hearing (DHH) community often face significant barriers in accessing and utilizing technology tailored to their needs. Indian Sign Language (ISL) serves as the primary mode of communication for millions in India, yet existing technological solutions frequently overlook the nuances inherent in this language. Recent advancements in machine learning have opened new avenues for enhancing ISL recognition systems, facilitating more effective communication between the DHH community and broader society.

This survey paper aims to synthesize findings from seven recent research papers focused on developing and evaluating ISL detection systems. We explore methodologies employed in recognizing both static signs, which involve fixed hand shapes, and dynamic gestures that convey meaning through movement. By examining the strengths and limitations of various machine learning models---including CNNs, LSTMs, SVMs, Inception models, RNNs, HMMs, and Random Forest classifiers---we provide insights into the effectiveness of different approaches in accurately recognizing sign language gestures.

Moreover, this paper highlights essential preprocessing techniques for preparing datasets for model training and emphasizes the importance of evaluating model performance through various metrics. Through this comprehensive review, we aim to underscore the significance of developing culturally relevant and accessible technology that bridges communication gaps for the DHH community in India. Our research not only contributes to the existing body of knowledge in the field but also paves the way for future advancements prioritizing inclusivity and accessibility in technology for all individuals.

2. LITERATURE SURVEY

2.1 Indian Sign Language Recognition Using Random Forest Classifier [1]

This paper introduces a system to bridge the communication gap for speech-impaired individuals by recognizing Indian Sign Language (ISL) gestures. The system uses sensor equipped gloves integrated with various sensors (flex sensors, IMU, touch sensors) to capture hand gestures. The machine learning algorithm, specifically a Random Forest classifier, processes this data to recognize gestures and convert them into speech through a mobile app.

- Models/Requirements:

  Hardware: Arduino Nano microcontrollers, RF, Bluetooth modules, Flex sensors, and IMU sensors.

  Machine Learning Model: Random Forest Classifier.

The Random Forest classifier was chosen because of its ability to handle missing data, quick training, and high accuracy in gesture classification.

- Workflow:

  The system uses sensor gloves equipped with various sensors that capture hand and finger movements. Flex sensors detect the bending of fingers, while touch sensors detect whether a finger is fully bent. IMU sensors capture angular velocity and acceleration along X, Y, and Z axes. The collected sensor data is sent to the Random Forest classifier, which processes and classifies the gestures. The classification results are sent via Blue- tooth to a mobile app, where the recognized gesture is converted to speech.

- Dataset Used and Accuracy Achieved:

  The dataset consists of 10 basic gestures with data collected by performing each gesture 50 times to create a training set. The model achieved an accuracy of 96.66%, proving its effectiveness in recognizing different gestures despite their similarities.

- Future Direction and Improvements:

  Future work could focus on increasing the dataset to include more gestures, words, and phrases to improve the system's generalization.

  Additionally, improving the hardware design of the gloves by making them lighter and integrating printed circuit boards (PCBs) could improve usability. Two-way communication could be established to allow gesture-to-speech conversion for the impaired user and speech-to-text conversion for the normal user.

2.2 Hybrid InceptionNet Based Enhanced Architecture for Isolated Sign Language Recognition [2]

This paper presents a hybrid deep learning architecture using InceptionV4 and Vanilla CNN for recognizing isolated sign language gestures from video frames. The goal is to improve recognition accuracy by combining ensemble learning and deep learning techniques. The system is tested on the IISL-2020 dataset of isolated Indian Sign Language gestures.

- Models/Requirements:
  Ensemble Model: Combining Vanilla CNN and InceptionV4. InceptionV4 uses convolutional filters of different sizes (1x1, 3x3, 5x5) to capture multi-scale features, while the Vanilla CNN provides baseline efficiency. Ensemble learning is applied to improve overall prediction accuracy.

- Workflow:
  Input frames of gestures are processed through both Vanilla CNN and InceptionV4 models in parallel. The InceptionV4 model captures complex spatial patterns using multi-scale filters, while Vanilla CNN focuses on extracting simpler features. The outputs of both models are combined using ensemble learning with a weighted average to improve accuracy. Softmax activation is applied to classify the gesture into one of the predefined sign classes.

- Dataset Used and Accuracy Achieved:
  The dataset used is the IISL-2020 dataset, consisting of 11 Indian sign gestures such as "Hello", "Good", "Morning", etc., with 15 videos per class. The proposed ensemble model achieved an accuracy of 98.46%, outperforming other state-of-the-art models for isolated sign language recognition.

- Model Size Optimization:

  The proposed model is computationally heavy. Reducing model size and computational cost can make it suitable for real-time applications. Dataset Expansion: Adding more gestures, including dynamic gestures, could improve the system's real-world applicability. Data Augmentation: Techniques could help to generalize the model to unseen signs or gestures.

2.3 Comparative Analysis on Datasets for Sign Language Detection System [3]

This paper focuses on developing a Sign Language Detection System using Convolutional Neural Networks (CNN) to recognize gestures and convert them into text and image outputs. The research analyzes different input data formats—images and CSV files—to evaluate their impact on model efficiency and accuracy. The goal is to improve communication for people with hearing and speech impairments by providing an effective, real-time sign language recognition system.

- Models/Requirements:
  CNN is the primary model used due to its strong image classification capabilities. The CNN architecture includes convolutional layers for feature extraction, pooling layers for dimensionality reduction, and fully connected layers for classification. A VGG16-based model is used for feature extraction because of its efficiency in object recognition. RMSprop optimizer is applied for training on the image dataset, and Adam optimizer for the CSV dataset.

- Workflow:
  The system uses two types of datasets: a CSV dataset with alphabet-based sign language data from the Sign Language MNIST dataset and an image dataset with number-based sign language images. The CNN is trained separately on both datasets. For the image dataset, RGB segmentation and contour-based image segmentation are applied to improve feature extraction. The model has a sequential architecture with three convolutional layers followed by pooling and fully connected layers. After training, the system predicts sign gestures from a live camera feed, displaying output as both text and images.

- Dataset Used and Accuracy Achieved:
  The CSV dataset contains 27,455 training and 7,172 testing samples representing alphabets (except "J" and "Z" which require motion) in CSV format from grayscale 28x28 pixel images, achieving 93.8% accuracy. The image dataset includes 1,500 training and 300 testing JPEG images of sign language numbers (0-9) converted to RGB, with 87% accuracy. The CSV dataset outperformed the image dataset in accuracy and training time, achieving a validation accuracy of 98.77%, while the image dataset faced challenges due to class imbalance and lower resolution.

- Future Direction and Improvements:
  Future work involves incorporating dynamic gesture recognition using models like LSTM to handle motion-based gestures such as "J" and "Z," integrating facial expression analysis to capture emotions for enhanced communication, expanding the dataset to include more gestures and expressions using data augmentation to improve generalization, and optimizing the model for faster real-time processing through pruning, quantization, or efficient architectures like MobileNetV2.

2.4 American Sign Language Real-Time Detection Using TensorFlow and Keras in Python [4]

This paper presents a system for real-time American Sign Language (ASL) detection using TensorFlow and Keras, aimed at enhancing communication for individuals with hearing impairments by detecting ASL gestures and converting them into textual and spoken language. The system employs convolutional neural networks (CNNs) for real-time recognition and integrates a text-to-speech (TTS) system for vocal output.

- Models/Requirements:
  A CNN architecture built with Keras processes thresholded images of hand gestures to predict corresponding ASL signs. The system uses Pyttsx3, a text-to-speech library, to convert recognized text into speech. The model is implemented in Python, utilizing OpenCV for image processing and TensorFlow as the backend for deep learning.

- Workflow:
  Real-time video input is captured frame-by-frame and preprocessed by thresholding to remove the background. Image augmentation techniques such as flipping and blurring are applied to increase data variability. The thresholded images are split into training and testing datasets, and the CNN model is trained to recognize different ASL signs. In real time, the system predicts ASL signs from input frames and converts predictions into speech using Pyttsx3 to improve accessibility.

- Dataset Used and Accuracy Achieved:
  The study uses a custom dataset capturing images of individuals performing various ASL signs. Data augmentation through flipping and blurring enhances training diversity. The model achieves an accuracy of 97%, demonstrating high effectiveness in real-time ASL gesture recognition.

- Future Direction and Improvements:
  Future work includes integrating LSTM to enable dynamic gesture recognition by processing the temporal sequence of hand movements. Emotion and expression detection can be added through facial expression recognition using a sequential model combining LSTM for motion capture, Gaussian Hidden Markov Model (GHMM) for modeling gesture and expression transitions, and Random Forest for final classification. This approach improves accuracy and context-awareness by combining gesture and facial emotion data. Expanding the dataset with more complex gestures and facial expressions will enhance generalization for dynamic, real-world scenarios.

2.5 Pose Detection Using OpenCV and Media Pipe [5]

This paper presents a system for pose detection using OpenCV and Media Pipe, designed to assist users in maintaining correct posture during exercises by providing real-time feedback. The system identifies body landmarks, computes joint angles, and corrects posture to prevent injuries, with applications extending to healthcare for monitoring hand movements.

- Models/Requirements:
  OpenCV is used for image processing and real-time webcam feed handling. Media Pipe provides pose estimation to detect body landmarks. A k-NN algorithm classifies posture based on joint coordinates.

- Workflow:
  Live video input is captured via webcam while users perform exercises. Media Pipe analyzes video frames to detect key body joints and computes angles between them. The k-NN algorithm is trained to recognize correct postures from joint coordinates and provides real-time feedback to the user. The system displays the number of repetitions completed along with any posture corrections needed.

- Dataset Used and Accuracy Achieved:
  Data was collected from 10 volunteers performing pull-ups, push-ups, and squats, with each exercise repeated 10 times. Accuracy achieved was 92% for pull-ups, 83% for push-ups, and 78% for squats, with lower accuracy in push-ups and squats attributed to greater variation in users' postures.

- Future Direction and Improvements:
  Future enhancements include integrating LSTM for dynamic pose recognition to better handle temporal variations in exercises, incorporating emotion recognition via facial expressions for context-aware feedback using a sequential model (LSTM → GHMM → Random Forest), expanding exercise variety to include lunges and jumping jacks for versatility, enabling integration with fitness tracking devices for comprehensive monitoring, and improving accuracy by applying face super-resolution techniques to enhance pose estimation under low-light or low-resolution conditions.

2.6 Sign Language Detection Using LSTM Deep Learning Model and Media Pipe Holistic Approach [6]

This paper proposes a system combining Media Pipe for hand tracking with an LSTM deep learning model to recognize sign language gestures, aiming to improve communication accessibility for deaf and hard-of-hearing individuals by accurately detecting signs represented by letters from 'A' to 'Z'.

- Models/Requirements:
  Media Pipe is used for hand tracking and extracting 21 key palm coordinates. The LSTM model processes sequential data to capture long-term dependencies in hand movement. NumPy handles data representation and processing of extracted coordinates.

- Workflow:
  Frames are captured via OpenCV from a video feed of a person signing. Media Pipe detects hand movements and extracts key points converted into x, y, z coordinates. These coordinates are transformed into a NumPy array and input into the LSTM model. The model predicts the corresponding sign, outputting an array of predicted values. The system's accuracy is evaluated against other methods using a sign language dataset.

- Dataset Used and Accuracy Achieved:
  Although specific dataset details are not provided, the approach achieves an overall accuracy of 99% in recognizing sign language gestures, demonstrating high performance and reliability.

- Future Direction and Improvements:
  Future work includes optimizing the LSTM architecture and hyperparameters for better performance, evaluating the model on larger and more diverse datasets for generalization, integrating additional modalities like facial expressions or body posture to enhance accuracy, and deploying the system in real-world scenarios to assess usability and refine it based on user feedback.

2.7 Sign Language to Text Conversion Using RNN-LSTM [7]

This paper presents a system to bridge the communication gap between speech-impaired individuals and the general population by recognizing both static and dynamic Indian Sign Language (ISL) gestures and translating them into text using Media Pipe and LSTM networks, achieving 100% accuracy for 26 ISL motions.

- Models/Requirements:
  Media Pipe is used for hand gesture recognition and extracting key coordinates. The RNN-LSTM model processes sequential data to learn long-term dependencies in sign gestures. OpenCV is utilized for capturing video and image processing.

- Workflow:
  Sign language gestures are collected via OpenCV video capture. Media Pipe detects hand movements and extracts 21 key points. The x, y, and z coordinates are converted into a NumPy array and input into the LSTM model to predict signs. Performance is evaluated using a confusion matrix and statistical analysis.

- Dataset Used and Accuracy Achieved:
  Though dataset specifics are not detailed, the model achieves 100% accuracy in classifying 26 ISL gestures, effectively recognizing both static and dynamic signs.

- Future Direction and Improvements:
  Future work may integrate GHMM and Random Forest with LSTM to improve classification and sequence handling. Expanding the dataset for better generalization, incorporating facial expression and

emotion recognition for context-aware translation, and optimizing the system for real-time, user-friendly applications aimed at deaf and hard-of-hearing individuals are suggested.

2.8 Vision-Based Continuous Sign Language Spotting Using Gaussian Hidden Markov Model [8]

This paper tackles the challenges of continuous sign language recognition (SLR), particularly handling movement epenthesis (me)—non-sign motions between meaningful signs—by using a vision-based system with a Gaussian Hidden Markov Model (HMM) to accurately spot signs in continuous sequences, achieving an 83% spotting rate.

- Models/Requirements:
  Gaussian Hidden Markov Model (HMM) is employed for decoding and recognizing sign sequences. H.264/AVC compression is used for efficient video feature extraction. Face and hand detection is performed with Colo based segmentation techniques to isolate hand movements by removing the signer's face.

- Workflow:
  Sign sequences are recorded via an RGB camera. The signer's face is detected and removed to focus on hand movement. Skin colour segmentation isolates hands. Spatial and temporal features are extracted and used to train the HMM. The Viterbi algorithm decodes the state sequence, allowing the system to spot signs by distinguishing start/end frames and separating signs from movement epenthesis.

- Dataset Used and Accuracy Achieved:
  The American Sign Language Lexicon Video Dataset (ASLLVD) from Boston University, containing thousands of ASL signs recorded from various angles, is used. The system achieves an 83% spotting rate, outperforming methods without PCA (78% accuracy).

- Future Direction and Improvements:
  Future research could integrate LSTM models for improved sequential data handling and accuracy. Incorporating contextual information such as facial expressions may enhance robustness. Extending the system to recognize full sign language sentences and real-world deployment for feedback and iterative improvements is also suggested to benefit deaf and hard-of-hearing users.

3. CONCLUSION

This survey paper synthesizes key findings from the analysis of seven research papers focused on Indian Sign Language (ISL) detection and recognition. The advancements in machine learning and computer vision have provided significant opportunities for improving communication within the deaf and hard-of-hearing (DHH) community. The use of various models, including Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, demonstrates the potential of these technologies in effectively recognizing both static and dynamic signs.

The comparative analysis of different models revealed that CNNs are well-suited for static sign recognition, achieving high accuracy, while LSTMs excel in recognizing gesture-based signs due to their ability to process temporal data. This highlights the importance of selecting appropriate models based on the characteristics of the sign language gestures being analyzed. Furthermore, the integration of preprocessing techniques and feature extraction methodologies, such as MediaPipe and OpenCV, has enhanced the accuracy and reliability of the systems developed.

Overall, this research underscores the critical need for technology that accommodates the unique requirements of the DHH community in India. By leveraging machine learning, computer vision, and user-centric design, these systems can significantly improve accessibility and communication for individuals who rely on sign language.

4. FUTURE SCOPE

Despite the promising advancements outlined in this survey, several areas require further research and development to enhance the effectiveness of sign language recognition systems:

- Augmented and Virtual Reality (AR/VR) Applications:

  Future studies should explore the potential of AR and VR technologies to create immersive learning environments for ISL. This could facilitate enhanced training and practice opportunities for users, making the acquisition of sign language more interactive and engaging.

- Algorithm Improvement:

  Research could focus on developing advanced machine learning algorithms tailored specifically for ISL. This includes optimizing current models to improve accuracy and robustness, particularly in recognizing less common signs and accommodating regional variations within the language.

- Diverse Data Collection:

  Expanding the datasets used for training models is crucial. Future research should prioritize the collection of a comprehensive and diverse range of sign language gestures, ensuring that the systems are trained on a representative sample that includes various dialects and regional differences.

- User-Centric Design:

  Investigating the social and cultural factors affecting the DHH community's adoption of technology is essential. Understanding user perceptions, preferences, and barriers can inform the development of more accessible and relevant systems.

- Real-Time Recognition Systems:

  Developing real-time sign language recognition systems that can seamlessly integrate into everyday applications, such as communication apps, is vital. This would allow for practical use in various contexts, enhancing the utility of sign language technology.

- Two-Way Communication:

  Further research should focus on creating systems that support bi-directional communication, enabling both speech-to-sign and sign-to-speech translation. This is crucial for facilitating comprehensive interactions between the DHH community and the hearing population.

- Cross-Language Recognition:

  Investigating the feasibility of adapting sign language recognition systems globally can enhance the accessibility of communication for the DHH community worldwide. By pursuing these avenues, future research can build upon the findings of this survey and contribute to the ongoing development of innovative and effective solutions for sign language recognition, ultimately improving the quality of life for individuals in the DHH community.

REFERENCES

[1] A. S, A. Potluri, S. M. George, G. R and A. S, "Indian Sign Language Recognition Using Random Forest Classifier," 2021 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), Bangalore, India, 2021, pp. 1-6, doi: 10.1109/CONECCT52877.2021.9622672.

[2] D. R. Kothadiya, C. M. Bhatt, H. Kharwa and F. Albu, "Hybrid InceptionNet Based Enhanced Architecture for Isolated Sign Language Recognition," in IEEE Access, vol. 12, pp. 90889-90899, 2024, doi: 10.1109/ACCESS.2024.3420776.

[3] Nikitha, J., Keerthana, S., Balakrishnan, Sahithya, & Sathya, S. (2022). Comparative Analysis on Datasets for Sign Language Detection System. 1652-1657. 10.1109/ICSCDS53736.2022.9761026.

[4] A. M, H. S. Sree, Jayashre, K. Muthamizhvalavan, N. Gummaraju and P. S, "American Sign Language Real Time Detection Using TensorFlow and Keras in Python," 2024 3rd International Conference for Innovation in Technology (INOCON), Bangalore, India, 2024, pp. 1-6, doi: 10.1109/INOCON60754.2024.10511469.

[5] D. Rai, Anjali, A. Kumar and A. Baghel, "Pose Detection Using OpenCV and Media Pipe," 2024 International Conference on Integrated Circuits, Communication, and Computing Systems (ICIC3S), Una, India, 2024, pp. 1-6, doi: 10.1109/ICIC3S61846.2024.10603040.

[6] M. Deshpande et al., "Sign Language Detection using LSTM Deep Learning Model and Media Pipe Holistic Approach," 2023 International Conference on Artificial Intelligence and Smart Communication (AISC), Greater Noida, India, 2023, pp. 1072-1075, doi: 10.1109/AISC56616.2023.10085375.

[7] A. Seviappan, K. Ganesan, A. Anbumozhi, A. S. Reddy, B. V. Krishna and D. S. Reddy, "Sign Language to Text Conversion using RNN-LSTM," 2023 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI), Chennai, India, 2023, pp. 1-6, doi: 10.1109/ICDSAAI59313.2023.10452617.

[8] A. K. Talukdar and M. K. Bhuyan, "Vision-Based Continuous Sign Language Spotting Using Gaussian Hidden Markov Model," in IEEE Sensors Letters, vol. 6, no. 7, pp. 1-4, July 2022, Art no. 6002304, doi: 10.1109/LSENS.2022.3185181.