International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 10 | October - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

A Survey of Sign Language Recognition

Vaishnavi Karanjkar[1], Rutuja Bagul[2], Raj Ranjan Singh[3], Rushali Shirke[4]

Department of Computer Engineering, Smt. Kashibai Navale College of Engineering, Pune

karanjkarv4@gmail.com, rutujabagul04@gmail.com, rajranjan9080@gmail.com, srushali054@gmail.com

---

Abstract - Sign Language is mainly used by deaf (hard hearing) and dumb people to exchange information between their own community and with other people. It is a language where people use their hand gestures to communicate as they can't speak or hear. The goal of sign language recognition (SLR) is to identify acquired hand motions and to continue until related hand gestures are translated into text and speech. Here, static and dynamic hand gestures for sign language can be distinguished. The human community values both types of recognition, even if static hand gesture recognition is easier than dynamic hand gesture recognition. By creating Deep Neural Network designs (Convolution Neural Network designs), where the model will learn to detect the hand motions images throughout an epoch, we are using Deep Learning Computer Vision to recognize the hand gestures. After the model successfully recognizes the motion, an English text file is created that can subsequently be translated to speech. The user can choose from a variety of translations for this paragraph. This application can be used without an internet connection and is entirely offline. With this model's improved efficiency, communication will be easier for the deaf (hard of hearing) and disabled people. We shall discuss the use of deep learning for sign language recognition in this paper.

Key Words: sign language, convolutional neural network, computer vision.

1. INTRODUCTION

The application of Sign language to multilingual text and voice output is an innovative nexus of technology, linguistic accessibility, and inclusivity is created by the incorporation of sign language into multilingual text and voice output. For Deaf and hard of hearing people, sign language is a crucial means of communication that opens up the outside world to them. However, this particular language has frequently encountered difficulties when dealing with spoken and written languages, posing obstacles in daily life and the worlds of education and the workplace.

Innovative solutions have surfaced in response to these issues, utilizing technology to close the communication gap. A varied, multilingual world will benefit from these applications' increased accessibility, comprehension, and inclusivity of sign language. These apps are redefining how sign language is incorporated into our global culture by utilizing cutting-edge advancements in natural language processing, computer vision, and machine learning. Various techniques and methods for sign language recognition was developed by different researchers.

One example of this is the use of Recurrent Neural Networks (RNN) which are commonly used for sign language recognition systems that rely on sequential data [1]. One of the most common types of RNN that is used for sign language recognition is the Long Short-term memory (LSTM) which is created to solve the vanishing gradient problem that can occur in traditional RNNs, where the gradient becomes too small to be useful during backpropagation, resulting in poor training and performance [1]. The study of [1] in which used the LSTM model for the system to recognize Indian sign language, presents a high accuracy result.

It is a system which uses a camera to sense the information that has been obtained through finger motions. It is the most commonly used visual-based method. It has been a tremendous effort and has been gone into the development of vision-based sign recognition systems through worldwide [8].

In recent years, there has been an increasing interest in deep learning applied to various fields, and it has contributed to technological improvement [10]. There have recently been numerous studies in the field of sign language recognition using deep learning to classify images or videos.

The reason we chose sign language recognition is because it has both the characteristics of motion recognition and the characteristics of a time series language translation. Deep learning models that classify images have low complexity compared to models that classify videos [10].

A multilayer perceptron (MLP) is a deep, artificial neural network in which the first layer, that is, the input layer is used to receive the signal and the last layer, the output layer, predicts the class of the input. Between these two layers, there consists an arbitrary number of hidden layers that is the true computational engine of the MLP [7].

Researchers have used several picture-capturing tools to classify photos. This technology includes a camera or webcam, a data glove, a Kinect, and jump controls. Contrary to data glove-based systems, a camera or webcam is the instrument that most researchers employ since it offers better and more natural interaction with no need for extra equipment. Data gloves have shown to be more accurate in data collection, despite being relatively pricey and cumbersome [4].

An overview of [9] is with three main modules including the feature extraction module, the processing module, and the classification module. The feature extraction module uses MMDetection to detect hand or body bounding boxes depending on the dataset's characteristics. If the dataset has full-body images, the body bounding boxes are extracted. On the other hand, the hand bounding boxes are extracted from the only-hand dataset. After that, the detected bounding boxes will be forwarded to HRNet, a CNN based model, to determine the key points normalized in the processing module. In addition,

with the whole-body dataset, the hand bounding boxes are evaluated from the farthest key-point to the left, the farthest key-point to the right, the farthest key-point on the top, and the farthest key-point to the bottom in the processing module. The hand gestures are identified in the classification module that uses key points and hand bounding boxes as inputs [9].

A.MediaPipe

Google has created an open-source framework called MediaPipe that allows developers to build machine learning and computer vision pipelines for multimedia applications. It provides pre-built components and tools for processing, analyzing, and visualizing multimedia data. The framework's modular architecture enables the proponents to create pipelines for gestures of static and dynamic Filipino Sign Language recognition [1].

First step is to generate dataset as there is no publicly available datasets for words. Signs are captured from webcam and dataset was generated. The commonly used words in which only one hand represents a particular word are 'okay', 'yes' 'peace', 'thumbs up', 'call me', 'stop', 'live long', 'fist', 'smile', 'thumbs down', 'rock', and words which uses both hands are 'alright', 'hello', 'good', 'no'. After capturing 2536 images by stable camera, they are converted into frames. The dataset images are divided into 75% for training and 25% for testing. Second step is to pass video frames to MediaPipe framework. Google's MediaPipe Hands is a solution for accurate hand and finger tracking. It uses machine learning (ML) to deduce 21 3D hand landmarks from a single frame. Various existing state-of the-art approaches rely on desktop environments for inference whereas the proposed approach achieves real time performance even on a mobile phone and scales to multiple hands [2].

B.LSTM

Long Short-Term Memory is a kind of recurrent neural network. LSTM was designed by Hochreiter & Schmidhuber. It tackled the problem of long-term dependencies of RNN in which the RNN cannot predict the word stored in the long-term memory but can give more accurate predictions from the recent information. As the gap length increases RNN does not give an efficient performance. LSTM can by default retain the information for a long period of time. It is used for processing, predicting, and classifying on the basis of time-series data.

Long Short-Term Memory (LSTM) is a type of Recurrent Neural Network (RNN) that is specifically designed to handle sequential data, such as time series, speech, and text. LSTM networks are capable of learning long-term dependencies in sequential data, which makes them well suited for tasks such as language translation, speech recognition, and time series forecasting.

LSTMs can be employed to process sign language video sequences, which are essentially sequential data frames. Each video frame can be considered a time step, and the LSTM network can analyze the temporal patterns and dependencies between these frames. For example, it can capture the dynamic movement of hands and facial expressions in sign language gestures.

LSTMs can be used for feature extraction from the video data. They can learn to represent important temporal features from the video frames, such as the trajectory of hand movements, handshapes, and the order of signs in a sentence. These features can then be used as input to your overall recognition model.

LSTMs can be employed for recognizing sign language gestures as sequences. As a user signs a phrase or sentence, the LSTM network processes each video frame sequentially and maintains context. This allows it to make predictions about the sign language signs being performed in real time.

LSTMs are often used in combination with CNNs in a hybrid architecture. CNNs are suitable for processing static visual features in individual frames, while LSTMs excel in handling temporal sequences. The output of the CNNs can be fed as sequences into the LSTM network, allowing the model to consider both spatial and temporal information for sign language recognition.

The LSTM network is trained using labeled sign language data, where the sequences of video frames are associated with specific sign language signs or phrases. The LSTM learns to capture the dynamics and context for accurate recognition.

Once trained, the LSTM model can be used for real-time sign language recognition. It takes in video frames as input and provides predictions on the signs being performed as the user signs. The model's ability to maintain context and consider temporal dependencies is particularly valuable for this application.

C. CNN

CNNs are primarily used for image feature extraction. In this system, each frame of the sign language video can be considered an image. CNNs are employed to capture and analyze the spatial features within these images. They can identify handshapes, facial expressions, and the position of hands in the frame. CNNs consist of convolutional layers that apply filters to the input images. These filters detect various patterns and features, such as edges, corners, and textures in the sign language video frames. The network learns to recognize the most relevant features for sign language recognition.

Pooling layers downsample the feature maps, reducing the spatial dimension while preserving the essential features. This helps in reducing computational complexity and enhances the model's invariance to small variations in hand positions or orientations.

The convolutional filters can be trained to identify key aspects of sign language gestures, such as the shapes made by fingers and the positions of the hands relative to the face. This helps the CNN in understanding the visual characteristics of signs. CNN often requires preprocessing techniques, such as image resizing, normalization, and data augmentation. Preprocessing ensures that the input data is appropriately prepared for the network. Data augmentation techniques can help increase the robustness of the model by introducing variations in the training data.

CNNs are often used in combination with Long Short-Term Memory (LSTM) networks in a hybrid architecture. While CNNs capture spatial features, LSTMs handle the temporal aspects of sign language gestures. The output of the CNNs can be passed as sequences to the LSTM network, allowing the model to consider both spatial and temporal information for recognition. The CNN model is trained using labeled sign language image data. The model learns to recognize important visual features and patterns associated

with sign language gestures. Training includes optimization techniques and the adjustment of model parameters to minimize recognition errors.

2. METHODOLOGY

In this study, during the model training, the LSTM network learns to tune the weights and biases of the gates and memory cells through backpropagation over time. Moreover, in this study the model training is set to stop at the nth epoch. Models that have lower loss than the previous lowest will be saved.

After the model training, the generated models are evaluated and visualize it through a confusion matrix for the basis of the evaluation through their accuracy score. TP corresponds to "True Positive" wherein the model prediction is correct. FP corresponds to "False Positive" wherein the model prediction is incorrect.

After the evaluation of the models, the model with the highest accuracy score will be used as the final model. This will go through inferencing to determine if the model correctly predicts each sign language phrase.

Table -1: SUMMARY OF RELATED WORK/GAP ANALYSIS

Ref No	Paper Name	Algorithm & Accuracy	Limitations
1	Long Short-Term Memory-based Static and Dynamic Filipino Sign Language Recognition	LSTM Neural Network Accuracy – 98%	The system is not able to recognize rare or complex gestures.
2	Hand Gesture Based Sign Language Recognition Using Deep Learning	AlexNet Classifier Accuracy – 99%	It is only able to recognize fingerspelling, which is a subset of sign language.
3	Deep Learning based Sign Language Recognition robust to Sensor Displacement	CNN, LSTM Accuracy – 95.6%	This is only working with large amount of data.
4	A Review of Segmentation and Recognition Techniques for Indian Sign Language using Machine Learning and Computer Vision	CNN	The current datasets used for ISL recognition and segmentation are limited in size and diversity.
5	Speech to Sign Language Translation for Indian Languages	Wavelet based MFCC & LSTM Accuracy – 80%	Accuracy for voice input is very less and with lot of noise.
6	Sign Language to Text Conversion using Deep Learning	CNN Accuracy – 99%	The model accuracy is less with less dataset.
7	Sign Language Translator using ML	KNN, Decision Tree Classifier, Neural Network Accuracy – 97%	The system does not take into account the context of the signs. This means that the system may not be able to correctly classify a sign if it is performed in a different context.
8	Sign Language to speech translation using ML	CNN Accuracy – 90%	If the user places the sensor in a different location, the system's performance may degrade.
9	An improved hand gesture recognition system using key-points and hand bounding boxes	CNN Accuracy – 95%	The proposed method is computationally expensive, especially for large images.
10	Dataset Transformation System for Sign Language Recognition Based on Image Classification Network	STmap, CNN-RNN Accuracy – 99%	The system converts the skeleton data into an image, called an STmap, before training the image classification network. This conversion process may lead to some loss of information

3. CONCLUSION AND FUTURE WORK

Sign language recognition has greatly benefited from advancements in machine learning, particularly in computer vision and natural language processing. Deep learning models, such as convolutional neural networks (CNNs) and Long Short Term Memory (LSTM), have demonstrated impressive results in recognizing signs accurately.

Sign language recognition technology is poised to make a significant impact on the lives of Deaf and hard of hearing individuals. This report underscores the importance of ongoing research and collaboration between experts in machine learning, computer vision, and the Deaf community to drive innovation and make sign language recognition more accurate, accessible, and inclusive. One of the main important future works to be done is to improve the response time and accuracy of the textual and speech outputs. With further advancements and increased awareness, we can look forward to a future where communication barriers are significantly reduced for the Deaf community.

ACKNOWLEDGEMENT

The present world of competition there is a race of existence in which those who have the will to come forward succeed. Project is like a bridge between theoretical and practical work. With this willing we joined this particular project. First of all, we would like to thank the supreme power the Almighty God who is obviously the one who has always guided us to work on the right path of life. We sincerely thank Prof. R. H. Borhade sir, Head of the Department of Computer Engineering of Smt Kashibai Navale college of engineering, for all the facilities provided to us in the pursuit of this project.

We are indebted to our project guide Prof. P. V. Bhaskare, Department of Computer Engineering of Smt. Kashibai Navale college of engineering. We feel it's a pleasure to be indebted to our guide for his valuable support, advice and encouragement and we thank him for his superb and constant guidance towards this project.

We are deeply grateful to all the staff members of the computer department, for supporting us in all aspects. We acknowledge our deep sense of gratitude to our loving parents for being a constant source of inspiration and motivation.

REFERENCES

[1] Carmela Louise L. Evangelista, Criss Jericho R. Geli, Marc Marion V. Castillo: Long Short-Term Memory-based Static and Dynamic Filipino Sign Language Recognition (2023)

[2] Roli Kushwaha, Gurjit Kaur, Manjeet Kumar: Hand Gesture Based Sign Language Recognition Using Deep Learning (2023)

[3] Rinki Gupta, Roohika Manodeep Dadwal: Deep Learning based Sign Language Recognition robust to Sensor Displacement (2023)

[4] Subhangi Kumari, Ernest Tarlue, Aissatou Diallo, Megha Chhabra, Gouri Shankar Mishra, Mayank Kumar Goyal: A Review of Segmentation and Recognition Techniques for Indian Sign Language using Machine Learning and Computer Vision (2023)

[5] Jashwanth Peguda, V Sai Sriharsha Santosh, Y Vijayalata, Ashlin Deepa R N, Vaddi Mounish: Speech to Sign Language Translation for Indian Languages (2022)

[6] Dr. Aruna Bhat, Vinay Yadav, Vishesh Dargan, Yash: Sign Language to Text Conversion using Deep Learning (2022)

[7] Jaya Nirmala: Sign language translator using machine learning (2022)

[8] Mrs.Aerpula Swetha, Vamja Pooja, Vundi Vedavyas, Challa Datha Venkata Naga Sai Kiran, Sadu Sravan: SIGN LANGUAGE TO SPEECH TRANSLATION USING MACHINE LEARNING (2022)

[9] Tuan Linh Dang a,∗ , Sy Dat Tran a , Thuy Hang Nguyen a , Suntae Kim b , Nicolas Monet b: An improved hand gesture recognition system using keypoints and hand bounding boxes (2022)

[10] Sang-Geun Choi , Yeonji Park and Chae-Bong Sohn : Dataset Transformation System for Sign Language Recognition Based on Image Classification Network (2022)

[11] E. B. Villagomez, R. A. King, M. J. Ordinario, J. Lazaro and J. F. Villaverde, "Hand Gesture Recognition for Deaf-Mute using FuzzyNeural Network," 2019 IEEE International Conference on Consumer Electronics - Asia (ICCE-Asia), 2019, pp. 30- 33, doi: 10.1109/ICCEAsia46551.2019.8942220.

[12] G. K. R. Madrid, R. G. R. Villanueva and M. V. C. Caya, "Recognition of Dynamic Filipino Sign Language using MediaPipe and Long ShortTerm Memory," 2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2022, pp. 1-6, doi: 10.1109/ICCCNT54827.2022.9984599.

[13] M. B. D. Jarabese, C. S. Marzan, J. Q. Boado, R. R. M. F. Lopez, L. G. B. Ofiana and K. J. P. Pilarca, "Sign to Speech Convolutional Neural Network-Based Filipino Sign Language Hand Gesture Recognition System," 2021 International Symposium on Computer Science and Intelligent Controls (ISCSIC), Rome, Italy, 2021, pp. 147-153, doi: 10.1109/ISCSIC54682.2021.00036.

[14] K. E. Oliva, L. L. Ortaliz, M. A. Tobias and L. Vea, "Filipino Sign Language Recognition for Beginners using Kinect," 2018 IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology,Communication and Control, Environment and Management (HNICEM), Baguio City, Philippines, 2018, pp. 1-6, doi: 10.1109/HNICEM.2018.8666346.

[15] M. Allen Cabutaje, K. Ang Brondial, A. Franchesca Obillo, M. Abisado, S. Lor Huyo-a and G. Avelino Sampedro, "Ano Raw: A Deep Learning Based Approach to Transliterating the Filipino Sign Language," 2023 International Conference on Electronics, Information, and Communication (ICEIC), Singapore, 2023, pp. 1-6, doi: 10.1109/ICEIC57457.2023.10049890.

[16] A. S. M. Miah, J. Shin, M. A. M. Hasan, and M. A. Rahim, "BenSignNet: Bengali Sign Language Alphabet Recognition Using Concatenated Segmentation and Convolutional Neural Network," Applied Sciences 2022, Vol. 12, Page 3933, vol. 12, no. 8, p. 3933, Apr. 2022, doi: 10.3390/APP12083933.

[17] D. Li, X. Yu, C. Xu, L. Petersson, and H. Li, "Transferring Cross-Domain Knowledge for Video Sign Language Recognition," Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 6204–6213, 2020, doi: 10.1109/CVPR42600.2020.00624.

[18] L. Pigou, A. van den Oord, S. Dieleman, M. van Herreweghe, and J. Dambre, "Beyond Temporal Pooling: Recurrence and Temporal

Convolutions for Gesture Recognition in Video," Int J Comput Vis, vol. 126, no. 2–4, pp. 430–439, Apr. 2018, doi: 10.1007/S11263-016-0957-7.

[19] S. Adhikary, A. K. Talukdar, and K. Kumar Sarma, "A Visionbased System for Recognition of Words used in Indian Sign Language Using MediaPipe," Proceedings of the IEEE International Conference Image Information Processing, vol. 2021-November, pp. 390–394, 2021, doi: 10.1109/ICIIP53038.2021.9702551.

[20] O. Nafea, W. Abdul, G. Muhammad, M. Alsulaiman. "Sensor-based human activity recognition with spatio-temporal deep learning." Sensors, vol. 21, no. 6, p. 2141, 2021.