JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Sign Language To Text Conversion Using CNN Model

[1]Girija V, [2]Akshay Kumar Singh, [3]Nayab Sahil, [4]Shibu Singh, [5]Tulika Paul

[1]Assistant Professor,
Computer Science and Engineering,
Cambridge Institute of Technology,
Bengaluru, India

***Abstract*— A key communication tool for those with hearing loss is sign language, facilitating their interaction with the world. Convolutional Neural Networks have become an effective tool for a number of image processing applications, such as the recognition of sign language. This study suggests an innovative CNN-based method for translating sign language to text. The suggested CNN model is designed to efficiently extract spatial and temporal characteristics from video sequences containing sign language. It employs convolutional layers to extract hierarchical features and pooling layers to reduce spatial dimensions while retaining crucial information. A large collection of images in sign language is used to train the model, enabling robust representation learning for precise translation. Test results show that the suggested CNN model is effective at translating sign language movement into text. The model achieves high accuracy and outperforms existing approaches on American Sign Language datasets. Overall, the proposed CNN-based sign language to text translation system offers a solution for addressing the communication gap between those who use sign language and those who do not. This technology can improve accessibility and inclusivity for the deaf community in a variety of contexts, including ordinary communication, healthcare, and education, by offering real-time translation capabilities. This study advances assistive technologies and encourages more equality and integration for people with hearing impairments.**

Keywords—Sign language recognition, Image processing, Deaf communication, Gesture-to-text conversion.

I. INTRODUCTION

Deaf people use sign language, a visual-gestural language, to communicate. It is a rich and expressive a language with unique syntax and grammar. Deaf individuals often face communication barriers, especially with those who do not understand sign language. Bridging this communication gap is important for their integration and participation in society. Technology can play a significant role in addressing these challenges. Systems for recognizing and converting sign language motions into text or speech can be developed, making it easier for non-signers to understand. Recognizing the pressing need to surmount these barriers, this paper introduces a pioneering project focused on the development of a real-time sign language to text translation system. CNNs are a powerful technology for image classification. They enable the automatic extraction of features for classification by applying a number of layers to input images. In the first layer, convolutional layers scan the image with learnable filters to detect local features like edges and textures. Pooling layers then downsample these features, retaining essential information. After being flattened, the feature maps are fed into fully connected layers for analysis of global relationships and trends. Non-linearity is introduced by activation functions, and the network's output is converted into probability scores for various classes by a softmax layer. During training, the model's parameters are adjusted to minimize a loss function. Once trained, the CNN can be evaluated on new data, achieving high accuracy in image classification tasks. Additionally, pre-trained CNN models are commonly used for quick adaptation to specific tasks, making CNN technology a cornerstone of contemporary machine learning and computer vision. Figure 1 shows the CNN model.

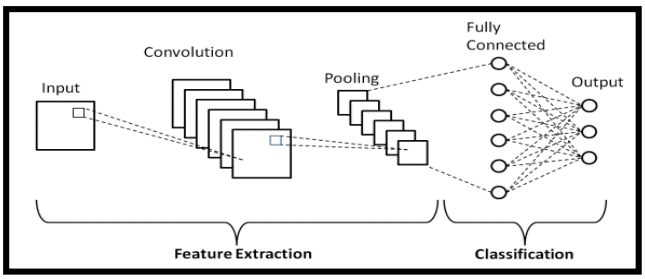

Fig 1: CNN Model

By harnessing CNNs, This approach aims to close the gap in communication between people who use sign language and those who do not, thereby fostering greater inclusivity and engagement within society. Central to this project's scope is the meticulous design of a CNN architecture tailored to process sign language gesture images or video frames. In the subsequent sections, this paper delineates the methodology underpinning the CNN-based sign language translation system, elucidating its architectural intricacies, training methodologies, and performance evaluation metrics. Artificial Neural Network is a connection of neurons, replicating the structure of human brain. Each connection of neuron transfers information to another neuron. Inputs are fed into first layer

of neurons which processes it and transfers to another layer of neurons called as hidden layers. After processing of information through multiple layers of hidden layers, information is passed to final output layer.

Additionally, we explore the broader societal implications of this system and its capacity to engender transformative change in communication accessibility for the deaf community. Through this research endeavor, we endeavor to advance the frontiers of inclusive technology and champion the cause of linguistic equity and accessibility for all.

II. LITERATURE SURVEY

Through a thorough examination of a multitude of research papers and studies in the area of sign language and finger spelling recognition, we embarked on a journey to synthesize this wealth of knowledge into a practical application. The outcome of this process is an interface that holds the potential to close the communication gap between the deaf communities, making sign language more accessible and inclusive.

A "Real-time Dynamic Hand Gesture Recognition using Hidden Markov Models" was proposed by M.M. Gharasuie [1]. They have presented a system that uses a Hidden Markov Model to identify dynamic hand movements for English digits 0–9. Because the likelihood of the hidden states plays a major role in HMM, learning more parameters requires a lot of time. The isolated and dynamic gesture recognition is done using the Hidden Markov Model, which has average recognition rates of 99.167% and 93.84%, respectively.

P. Vijayalakshmi [2] suggested converting sign language to speech. Flex sensors are used in a sensor-dependent gesture interpretation system that has been designed to detect hand movements. By recording the shift in resistor values, the flex sensor in the suggested system is implicitly utilized to calculate the angular tilt to which the finger is bent.

Bantupalli Kshitij [3] suggested utilizing machine vision and deep learning to recognize American sign language. A micro-proposed technique makes advantage of the author's data-set, which consists of a restricted set of motions with the most often used words. Together with recurrent neural networks (RNNs), convolutional neural networks (CNNs) are used to classify individual tr gestures and image sequences. CNN was constructed using code from the Inception Model. The pool layer's accuracy was approximately 58 LDR percent, while the SoftMax layer's accuracy was nearly 90%.

Dabre Kanchan [4] presented a machine learning model that uses webcam photos to understand sign language. It focuses on translating sign language, particularly Indian Sign Language, at the word level to text and then voice. In the classification stage, each static gesture is trained using 500 positive samples, 500 negative samples, and 50 test picture samples. The gesture is then interpreted using the Haar Cascade classification method. The last phase turns the text into speech. The results show that the accuracy is 92.68 percent accurate.

Aditya Das [5] suggested utilizing deep learning to recognize sign language from static gesture photographs that have been specially processed. Cropping, scaling, and flipping are methods used in data augmentation to make sure the neural network is not restricted to a certain kind of image. The photos are divided into testing, validation, and training sets using a bespoke algorithm that takes as parameters the percentages of testing and validation. Both the accuracy rate and the accuracy of the validation are more than 90%.

Kadam Kunal [6] suggested hiring a US sign language interpreter. A glove made of flex sensors is developed. Flex sensors, an LCD, an accelerometer, and a keypad make up the system. In addition to bridging the communication gap, the project aims to create a self-learning system that enables individuals to learn American Sign Language. There are two modes: the teaching mode and the learning mode. When using the teaching mode, a database is built by executing various gestures and stored in the microcontroller's EEPROM.

Nobuhiko MUKAI[7] Proposed a “Japanese Fingerspelling Recognition based on Classification Tree and Machine Learning”, NICOGRAPH International, 2017. Japanese sign language incorporates fingerspelling that are rooted in the American alphabet system, with additional elements influenced by Japanese characters, gestures, numbers, and specific meanings. The researchers sought to overcome this particular problem given by the language form's complexity by incorporating cutting-edge machine learning techniques. The core of their approach involved the utilization of a classification tree, a decision tree-based structure commonly employed in pattern recognition tasks. This tree-based model was designed to discern and categorize the intricate patterns inherent in fingerspelling, allowing for the accurate identification of individual signs. The researchers reported a noteworthy achievement of 86% accuracy rate for their proposed recognition method. This success underscored the effectiveness of the SVM-based model in deciphering the nuanced and diverse aspects of fingerspelling in Japanese sign language. The high accuracy implies the system's capability to reliably recognize and classify a significant portion of the complex linguistic expressions present in this form of sign language, technologies.

Adithya V[8] Proposed a “Artificial Neural Network Based Method for Indian Sign Language Recognition”. Image acquisition, hand segmentation, feature extraction and then classification based on supervised feed forward backpropogation algorithm was used by Adithya, Vinod and Usha Gopalkrishnan for hand feature extraction having average recognition rate of 91.11%.

M. Mohandes[9] Proposed a "Arabic sign language recognition using the leap motion controller". To recognize Arabic sign language, they employed a Leap motion controller. Using a jump motion sensor, 10 samples of each of the 28 letters were gathered for this system. Twelve of the 23 attributes that the Leap Motion Controller returned for every frame of data were deemed most pertinent for additional processing. They used Nave Bayes classifier and Multilayer Perceptron (MLP) to classify 28 letters in Arabic sign language. A correct recognition rate of 99.1% achieved using Multilayer Perceptron and 98.3% using Nave Bayes classifier.

Cao Dong[10] Proposed a "American Sign Language alphabet recognition using Microsoft Kinect" . This study used Microsoft’s Kinect sensor to obtain depth data. The Hand segmentation is done using per-pixel classification method. Random Forest (RF) gesture classifier was implemented to recognize ASL signs using the joint angles.

The system considered 24 static alphabets and achieved accuracy of 92%.

Ball, Jonathan [11] suggested using CNNs for sign language translation and recognition. It presents a novel method for using CNNs to recognize and translate American Sign Language (ASL). The authors' thorough research, which made use of a sizable dataset of ASL motions, produced encouraging results for both recognition and translation tasks. CNNs are used here to highlight the importance of deep learning methods for computer vision and language processing.

Cippitelli, Daniele [12] DeepASL: Enabling Pervasive and Non-Intrusive Mobile Sign Language Recognition was proposed. In the field of sign language recognition, it is a noteworthy contribution. This study addresses the need for practicality in real-world applications by emphasizing the creation of a deep CNN architecture designed for mobile devices.

Thad Starner[13] Proposed a Sign Language translation with Microsoft Kinect. It was conducted by Starner et al. stands as a pioneering endeavor in the realm of sign language recognition. Employing the Microsoft Kinect sensor, this project predates the deep learning era, yet its significance is undeniable. It serves as a foundational milestone that paved the way for subsequent research and innovations in sign language recognition.

Alex Graves[14] Proposed a Sign Language translation Using a Convolutional Neural Network. It presents a noteworthy contribution to the field of sign language recognition. The authors designed and implemented a Convolutional Neural Network based system specifically designed for recognizing Australian Sign Language (Auslan) gestures. By harnessing the capabilities of CNNs to process and test the visual cues inherent in sign language, they achieved remarkable accuracy.

Juyoung Shin[15] proposed a Sign Language Translation using Wearable Myoelectric Sensors. It represents an innovative approach to sign language recognition and translation. Although it places a greater emphasis on utilizing myoelectric sensors rather than conventional CNNs, it makes a noteworthy addition to the field. This work investigates a novel approach to sign language gesture recognition using wearable myoelectric sensors.

E. Assogba[16] Proposed a Deep Learning for Sign Language translation. They provide a thorough synopsis of the most recent advancements in sign language interpretation. Through the utilization of deep learning methodologies, such as CNNs, the writers explore the most recent developments in the domain. They offer a comprehensive analysis of several models and datasets used in research on sign language. This study sheds light on the capabilities and trends in the field of sign language technology, making it a useful resource for scholars, practitioners, and enthusiasts. It emphasizes how deep learning has the ability to change communication for the deaf and hard of hearing community by increasing accessibility to sign language.

Oscar Koller [17] Proposed a Neural Machine Translation for Sign Language. It presents an in-depth exploration of applying neural machine translation techniques to sign language, providing a thorough synopsis of the area. While its focus is not exclusively on CNNs, it provides valuable insights into the broader landscape of sign language translation. By examining various methods and strategies, this work addresses the critical challenge of bridging the gap between spoken and sign languages. It highlights the significance of advanced machine learning techniques in making sign language more accessible, facilitating communication, and promoting inclusivity for the deaf community, within the broader context of neural machine translation.

Hrishikesh Kulkarni[18] Proposed a Deep Learning-Based American Sign Language (ASL) Recognition System. It introduces an innovative deep learning system for the translation of American Sign Language (ASL) gestures. The authors leverage combination of CNNs and Long Short-Term Memory networks (LSTMs) to achieve precise and robust recognition results. Importantly, their work extends beyond theoretical application by proposing a practical model designed for real-time usage. This development holds good promise for the deaf community, as it paves the way for accessible and effective communication in real-world scenarios, underlining the transformative potential of deep learning in making ASL more accessible and inclusive.

Chien-Wei Wu[19] Proposed a Sign Language Recognition Using 3D CNN Convolutional Neural Networks. They represents a significant advancement in the area of sign language recognition. This work adopts a novel approach by incorporating 3D Convolutional Neural Networks (CNNs), which are capable of capturing spatiotemporal data, making them perfect for understanding gestures in sign language. By accounting for both the spatial and temporal aspects of signing, this research introduces a more holistic and nuanced understanding of sign language recognition.

Siawpeng Er, Jie Zhang,[20] Proposed a Sign Language Recognition and Translation: A Multimodal Deep Learning Approach. It represents an innovative approach to sign language translation. This study combines visual and depth data obtained from RGB-D sensors, thus leveraging a multimodal approach. CNNs are employed to process the visual data, allowing for a more comprehensive understanding of sign language movements. This research is unique in that it focuses on integrating various modalities and using depth and visual information to increase recognition accuracy. It recognizes the complexity of sign language and the necessity for a multimodal approach by combining several data sources. The deaf and hard of hearing people could benefit greatly from this study in terms of accessibility and communication.

III. METHODOLOGY

Data collection, preprocessing, feature extraction, training/testing, and gesture categorization were the stages.

Dataset Creation

In this model, data collection is carried out as images depicting various signs at different angles, covering the sign letters A to Z, utilizing the OpenCV library. A total of 180 raw images representing the alphabet from A to Z are captured, meeting the requirements for ASL (American Sign Language) representation. Figure 2 illustrates all the English alphabets alongside their corresponding sign conventions utilized in ASL.

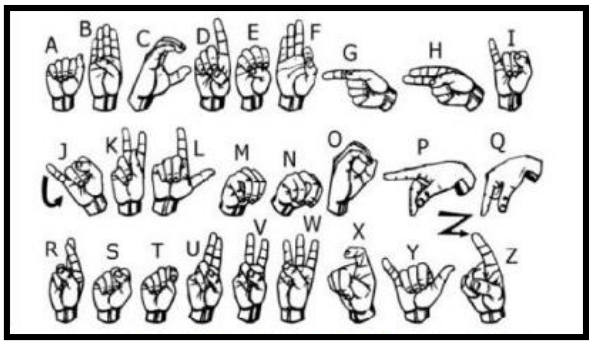

Fig 2: Sign conventions of alphabets in ASL

Data pre-processing and Feature extraction

In this hand detection method, the mediapipe library which is used for image processing is initially considered for the purpose of detecting hands from webcam-captured images. The region of interest from the webcam image is the hand; it is cut from the picture and, after applying Gaussian blur, is converted to a grayscale image using the OpenCV package. The OpenCV library, commonly referred to as the Open Computer Vision Library, makes it simple to apply the filter. Figure 3 displays the image data obtained from the conversion of the grayscale image to a binary image using threshold and adaptive threshold techniques.

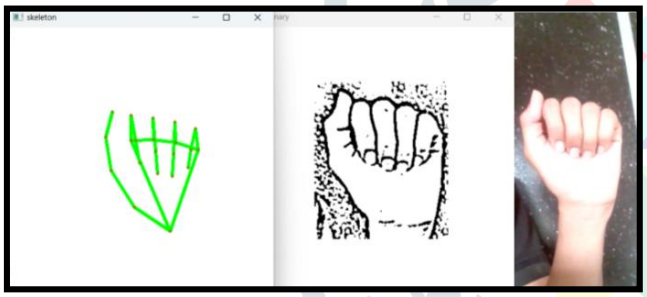

Fig 3: Data preprocessing

There are numerous flaws in this method, such as the requirement that the hand be ahead of a smooth, clean background and that it be under optimal lighting circumstances in order for it to produce correct results. However, in the actual world, obtaining both of these conditions is challenging. In order to get around this problem, a variety of strategies are investigated. Eventually, an intriguing one is discovered, in which the hand is first identified from the frame using Mediapipe, and the hand landmarks of the current hand are then drawn on a plain white image. Ultimately, the opencv library is used to draw these landmark points on a plain white background.

Training and testing

The procedure entails converting RGB input photos to grayscale and using Gaussian blur to reduce superfluous noise. After that, the hand is separated from the backdrop using adaptive thresholding. Ultimately, 128 x 128 pixel resolution is achieved by resizing the photos. After completing all required data preprocessing procedures, the input images post-preprocessing are fed into the model for training and testing. The prediction layer makes an educated guess as to which class the image will belong to. In order to ensure that the total of all the values in each class adds up to 1, the output is normalized between 0 and 1. The model used the softmax function to accomplish this. At first, the prediction layer's output could differ from the real value. The networks are trained with labeled data in order to increase accuracy. A popular performance metric for classification problems is cross-entropy. It is a continuous function that produces values that are exactly zero when the prediction matches the label and produces positive values when the prediction deviates from the labeled value. The goal of optimization is to minimize the cross-entropy as near to zero as feasible. The neural network weights are changed in the network layer to accomplish this. A cross-entropy calculation function is incorporated into TensorFlow. Once the cross-entropy function has been determined, it is optimized through the use of Gradient Descent, particularly using the Adam Optimizer one of the best gradient descent optimizers.

Gesture Classification

The method predicts the user's final symbol using two levels of algorithms. Once the features have been extracted from the frame captured with OpenCV, apply the Gaussian blur filter and threshold to obtain the processed image. After this image has been processed, it is sent into the CNN model for prediction. If a letter is identified for more than 60 frames, it is printed and used to build the word. Using the blank symbol, spaces between words are taken into consideration. The CNN layers that are used in the model are as follows:.

1. 1st Convolution Layer: The resolution of the input image is 128 × 128 pixels. 32 filter weights (3x3 pixels each) are used in the first convolutional layer to analyze it initially. A 126X126 pixel image will be produced as a consequence, one for each filter-weight.
2. 1st Pooling Layer : Using max pooling of 2x2, or keeping the highest value in the 2x2 square of the array, we down sample the images. Consequently, our image has been down-sampled to 63x63 pixels.
3. 2nd Convolution Layer: The second convolutional layer now uses this 63 x 63 from the first pooling layer's output as an input. 32 filter weights (3x3 pixels each) are used in the second convolutional layer to process it. This will produce a picture with 61x61 pixels.
4. 2nd Pooling Layer: The final photos are downsampled once more utilizing the maximum pool of 2x2, and their resolution is lowered to 30 x 30.
5. 1st Densely Connected Layer: These images are now fed into a second convolutional layer, which restructures the output into an array of 30x30x32 = 28800 values. The first layer is a fully connected layer with 128 neurons. This layer receives an array of 28800 values as input. The second densely connected layer receives the output from these layers. To avoid overfitting, we are utilizing a dropout layer with a value of 0.5.
6. 2nd Densely Connected Layer: A completely linked layer with 96 neurons now receives the output from the first densely connected layer as an input.
7. Final layer: The last layer gets its input from the output of the second densely connected layer, and the no. of neurons in that layer corresponds to the no. of classes (alphabets plus a blank sign) that need to be classified.

The figure 4 shows the convolution and max pooling layers in the model.

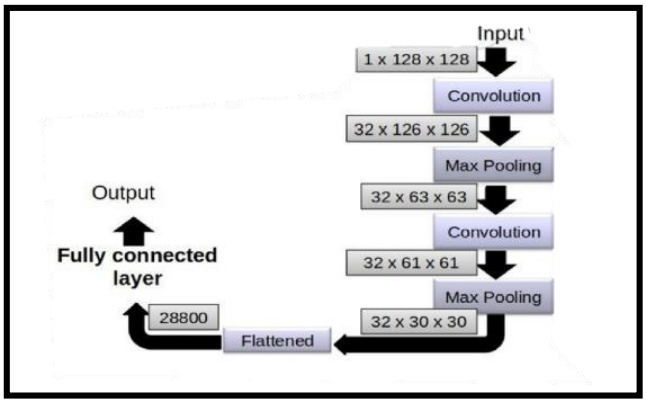

Fig 4: CNN layers used in the model

IV. RESULTS AND DISCUSSIONS

A 97% accuracy rate is attained utilizing this method, demonstrating consistent performance across various conditions, including scenarios with clean and without a clean background, as well as under optimal lighting conditions. The developed sign language to text translation system, leveraging CNNs, exhibited promising results in real-time interpretation of sign language gestures. Through comprehensive testing and evaluation, the system demonstrated robustness and efficiency in accurately translating a predefined vocabulary of sign language into text or speech output. The CNN architecture, adept at capturing spatial and temporal features from sign language video sequences, contributed significantly to the system's performance. By employing convolutional layers for feature extraction and recurrent layers for capturing temporal dependencies, the model achieved high accuracy in interpreting complex sign language gestures. The utilization of pooling layers facilitated dimensionality reduction while preserving essential information, enhancing the model's efficiency and computational speed.

Moreover, advanced techniques such as data augmentation, preprocessing, and transfer learning played a pivotal role in improving the system's generalization capabilities. By augmenting the training data with variations in lighting, backgrounds, and signer orientations, the system demonstrated resilience to real-world environmental factors, ensuring consistent performance across diverse scenarios. The real-time functionality of the model was rigorously evaluated under varying conditions, including different lighting environments, backgrounds, and signer orientations. The system consistently delivered prompt and accurate interpretations, demonstrating its potential to facilitate smooth communication in real-world situations between sign language users and those who do not.

However, it is essential to acknowledge some limitations and areas for future improvement. While the system achieved high accuracy in recognizing a predefined vocabulary of sign language gestures, scalability to accommodate a broader range of gestures and dialects remains a challenge. Furthermore, variables like signer variability, occlusions, and complex hand movements may affect the system's performance, requiring additional study and adjustment.

Overall, the developed sign language to text translation system is an important step in the right direction when it comes to communication barriers faced by the deaf and community. By harnessing the capabilities of deep learning and image recognition techniques, the system offers a promising solution for enhancing communication accessibility and fostering greater inclusivity in society. Continued research and development efforts in this domain are warranted to advance the state-of-the-art in sign language recognition technology and promote equitable communication opportunities for individuals of all linguistic backgrounds.

V. CONCLUSION

In summary, the creation of a real-time system that converts sign language to text using convolutional neural networks is a major step in removing communication barriers that the deaf community must contend with. Through the utilization of deep learning and image recognition algorithms, the system provides a workable means of reducing the communication gap between non-signers and sign language users, thus promoting more accessibility and inclusivity.

The successful implementation and evaluation of the CNN-based sign language recognition system underscore its potential to revolutionize communication accessibility for individuals with hearing impairments. Beyond its immediate applications, the system holds promise for integration into various communication devices and educational tools, thereby empowering people with hearing impairments to engage more fully in social, educational, and professional domains.

Moving forward, continued research and development efforts in this domain are warranted to further refine and optimize sign language translation systems. By advancing the state-of-the-art in sign language recognition technology, we can collectively work towards building a more inclusive and equitable society, where individuals of all linguistic backgrounds have equal opportunities for communication and engagement.

REFERENCES

[1] M.M.Gharasuie, H.Seyedarabi, proposed a Real-time Dynamic Hand Gesture Recognition using Hidden Markov Models, 8 th Iranian Conference on Machine Vision and Image Processing(MVIP), IEEE, 2013.

[2] P Vijayalakshmi and MAarthi, proposed a Sign language to speech conversion. In 2016 International Conference on Recent Trends in Information Technology (ICRTIT), IEEE, 2016.

[3] Kshitij Bantupalli and Ying Xie, proposed American sign language recognition using deep learning and computer vision. In 2018 IEEE International Conference on Big Data (BigData). IEEE, 2018.

[4] Kanchan Dabre and Surekha Dholay, proposed a machine learning model for sign language interpretation using webcam images. In 2014 International Conference on Circuits, Systems, Communication, and Information Technology Applications (CSCITA). IEEE, 2014.

[5] Aditya Das, Shantanu Gawde, KhyatiSurat wala, and Dhananjay Kalbande, proposed a Sign language recognition using deep learning on custom processed static gesture images. In 2018 International Conference on Smart City and Emerging Technology (ICSCET). IEEE, 2018.

[6] Kunal Kadam, Rucha Ganu, Ankita Bhosekar, and SD Joshi, proposed a American sign language interpreter. In 2012 IEEE Fourth International Conference on Technology for Education. IEEE, 2012.

[7] Nobuhiko MUKAI, Naoto HARADA, Youngha CHANG, proposed a “Japanese Fingerspelling Recognition based on

Classification Tree and Machine Learning", NICOGRAPH International, 2017.
[8] Adithya V., Vinod P., Usha Gopalakrishnan, proposed a "Artificial Neural Network Based Method for Indian Sign Language Recognition", IEEE Conference on Information and Communication Technologies (ICT 2013), JeJuIsland April 2013.
[9] M. Mohandes, S. Aliyu and M. Deriche, proposed a "Arabic sign language recognition using the leap motion controller," 2014 IEEE 23rd International Symposium on Industrial Electronics (ISIE), Istanbul, 2014.
[10] Cao Dong, M. C. Leu and Z. Yin, proposed a "American Sign Language alphabet recognition using Microsoft Kinect," 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Boston, MA, 2015.
[11] Jonathan Ball, Brian Price, proposed a Sign Language Recognition and Translation with CNNs, IEEE 2016.
[12] Daniele Cippitelli, Davide Cipolla, proposed a DeepASL: Enabling Ubiquitous and Non Intrusive Mobile Sign Language Recognition, IEEE 2018.
[13] Thad Starner, Mohammed J. Islam, proposed a Sign Language Recognition with Microsoft Kinect, IEEE 2013.
[14] Alex Graves, Santiago Fernández, proposed a Sign Language Recognition Using a Convolutional Neural Network, IEEE 2018.
[15] Juyoung Shin, Joo H. Kim, proposed a Sign Language Translation and Recognition using Wearable Myoelectric Sensors, IEEE (2017).
[16] E. Assogbaand P. H. S. Amoudé, proposed a Deep Learning for Sign Language Recognition and Translation, IEEE 2019.
[17] Oscar Koller, David Ney, proposed a Neural Machine Translation for Sign Language: A Survey, IEEE 2020.
[18] Hrishikesh Kulkarni, Suchismita Saha, proposed a Deep Learning-Based American Sign Language (ASL) Recognition System, IEEE 2020.
[19] Chien-Wei Wu, Eugene Lai, proposed a Sign Language Recognition Using 3D Convolutional Neural Networks, IEEE 2019.
[20] Siawpeng Er, Jie Zhang, proposed a Sign Language Recognition and Translation: A Multimodal Deep Learning Approach, IEEE 2020.