A Review on Sign Language Recognition Using CNN

Meena Ugale(✉), Odrin Rodrigues Anushka Shinde, Kaustubh Desle, and Shivam Yadav

Department of Information Technology, Xavier Institute of Engineering, University of Mumbai, Mumbai, Maharashtra, India

meena.u@xavier.ac.in

Abstract. In sign language, hand gestures are used as one type of non-verbal communication. Individuals with hearing or speech problems typically use it to communicate with others or among themselves. Many makers around the world have created numerous sign language systems. The software that shows a system prototype capable of automatically recognizing sign language to assist deaf and dumb individuals in communicating with each other or regular people more successfully. The study demonstrates that there is ongoing research in the field of vision-based hand gesture recognition, with various studies being undertaken and a large number of publications appearing every year in journals and conference proceedings. Data acquisition, data environment, and hand gesture representation are the three main areas of concentration in publications on the hand gesture recognition system. In terms of recognition precision, we have also analyzed how well the recognition system performs. The recognition accuracy for the signer dependent spans from 69% to 98%, with an average of 88.8% among the chosen experiments.

Keywords: Sign Language Recognition · Communication · Data Acquisition · Convolution Neural Network · Deep Learning

1 Introduction

The lack of communication between people with hearing or speaking impairments and the rest of the world. This can be frustrating and isolating for people with hearing or speaking impairments. When spoken communication is impossible or undesirable, sign language is used to communicate through bodily movements, particularly those of the hands and arms as shown in Fig. 1. Though sign languages are very effective in communication, they are not commonly used or known. This creates a communication barrier.

The American Sign Language (ASL), British Sign Language (BSL), Indian Sign Language (ISL), and others are all different sign languages. Similar to how spoken languages have a vocabulary of words, sign languages too have a vocabulary of signs. The grammar of sign languages varies from country to country

S. Tamane et al. (Eds.): ICAMIDA 2022, ACSR 105, pp. 251–259, 2023.
https://doi.org/10.2991/978-94-6463-136-4_23

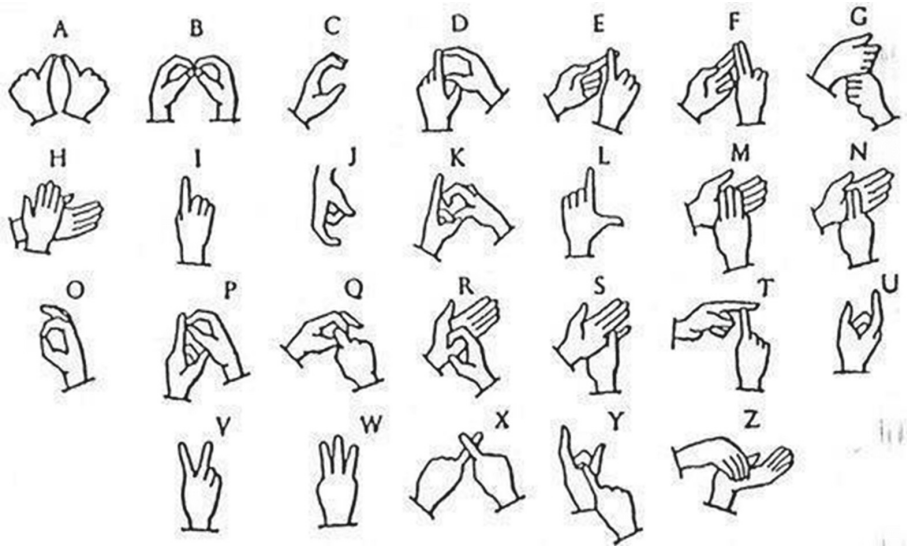

Fig. 1. Alphabetic Hand Sign [11].

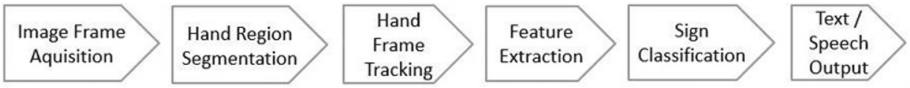

Fig. 2. Block Diagram of Sign Language Recognition.

and is not standardized or universal. A manual sign language interpreter is not always a good idea and frequently intrudes on the subject's right to privacy. This problem can be resolved by using an automated sign language translator that can translate sign language into spoken or written language as shown in Fig. 2. The hearing- and vocally- impaired people will benefit from an accurate automatic sign language translator since it will allow them to live independently and in close contact with others for the rest of their lives.

2 Convolution Neural Network Concept

Machine learning includes convolutional neural networks (CNNs). It is a subset of the several artificial neural network models that are employed for diverse purposes and data sets. A CNN is a particular type of network design for deep learning algorithms that is utilized for tasks like image recognition and pixel data processing [13,14]. The structure of a CNN is comparable to the connection structure of the human brain. Similar to how the brain has billions of neurons, CNNs also have neurons, but they are structured differently. A CNN performs better with image inputs and voice or audio signal inputs compared to the earlier networks.

2.1 CNN Layers

A deep learning CNN is composed of three layers: a convolutional layer, a pooling layer, and a fully connected (FC) layer. The first layer is the convolutional layer, while the final layer is the FC layer. The complexity of the CNN grows from the convolutional layer to the FC layer. The CNN is able to identify increasingly larger and more intricate aspects of an image until it successfully recognizes the complete thing as a result of the rising complexity.

Convolution Layer: The convolutional layer, the central component of a CNN, is where most computations take place. The first convolutional layer may be followed by a subsequent convolutional layer. A kernel or filter inside this layer moves over the image's receptive fields during the convolution process to determine whether a feature is present. The kernel traverses the entire image over a number of iterations. A dot product between the input pixels and the filter is calculated at the end of each iteration. A feature map or convolved feature is the result of the dots being connected in a certain pattern. In this layer, the image is ultimately transformed into numerical values that the CNN can understand and extract pertinent patterns from.

Pooling Layer: The pooling layer similarly to the convolutional layer sweeps a kernel or filter across the input image. Contrary to the convolutional layer, the pooling layer has fewer input parameters but also causes some information to be lost. Positively, this layer simplifies the CNN and increases its effectiveness.

Fully Connected Layer: Based on the features extracted in the preceding layers, picture categorization in the CNN takes place in the FC layer. Fully connected in this context means that every activation unit or node of the subsequent layer is connected to every input or node from the preceding layer. The CNN does not have all of its layers fully connected because that would create an excessively dense network. It would cost a lot to compute, increase losses, and have an impact on output quality.

2.2 Working

As shown in Fig. 3 each layer trains the CNN to recognize the many aspects of an input image. Each image is given a filter or kernel to create an output that gets better and more detailed with each layer. The filters may begin as basic characteristics in the lower layers. In order to check and identify features that specifically reflect the input item, the complexity of the filters increases with each additional layer. As a result, the partially recognized image from each layer's output, or convolved image, serves as the input for the subsequent layer. The CNN recognizes the image or object it represents in the final layer, which is an FC layer. The input image is processed through a number of different filters during convolution. Each filter performs its function by turning on specific aspects of

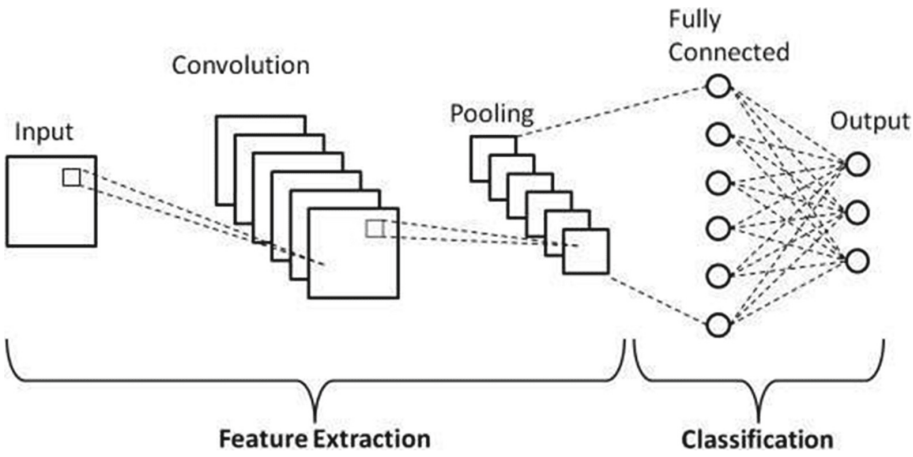

Fig. 3. Convolution Neural Network Architecture [12].

the image, after which it sends its output to the filter in the subsequent layer. The operations are repeated for dozens, hundreds, or even thousands of layers as each layer learns to recognize various features. Finally, the CNN is able to recognize the full object after processing all the picture data through its many layers.

3 CNN-Based Approach for Sign Language Recognition

According to paper [1], the model is constructed with input layer, four convolutional layers, five rectified linear units (ReLu), two stochastic pooling layers, one dense and one SoftMax output layer. The CNN design employs convolutional layers with various pooling sizes, an activation function and a rectified linear unit handeling non-linearities. Through supervised learning, the network is trained to learn the characteristics of each symbol. By switching from ANN to deep ANN, the identification accuracy is further enhanced, with a claimed 5% improvement in recognition rate. Hence, CNN's are a suitable tool for simulating sign language recognition on mobile platforms [1].

In proposed paper [2], Hierarchical Attention Network with Latent Space (LS-HAN) is used to translate signing videos sentence-by-sentence. By using a sliding window method, each video is split into frame segments. Upon encoding, the Hierarchical Attention Network (HAN) is given the start symbol "#Start," which marks the start of sentence prediction. For each decoding timestamp, the word with the highest probability after the soft max is selected as the predicted word, and its representation is sent to (HAN) for each succeeding timestamp until end flag "#End" is raised [2].

The author [3] approach was designed to function with one-handed gestures. Convolution, Max-Pooling, ReLU, Dropout, Fully Connected, and SoftMax layers form up the deep learning network. The neural network employs a stack of

layers to perform classification using the features that are collected from the convolution layers. Using the pre-trained network, the features required for classification are extracted once from the dataset. The validation accuracy of the authors' proposed model was 84.68%, with a validation loss of 0.3523 [3].

Kshitij Bantupalli, Ying Xie [4] mentioned in the paper about CNN model extracted temporal features from the frames which was used further to predict gestures based on sequence of frames. Two methods were employed to classify the signs: a. using outputs from the Softmax layer, and b. derived the results from the Global Pool layer. The global pool results in a 2048-sized vector, which allowed features to be analyzed by the RNN. These characteristics are transferred to Long-Short-Term Memory (LSTM), allowing for larger time dependencies. The vanishing/exploding gradient problem is handled with LSTMs, enabling improved accuracy on larger data sets. On the training set, the model scored high accuracy of 99%. The LSTM utilized sequence data to classify the gesture segments that CNN recognised and processed into one of the gesture classes. The system was able to achieve 93% accuracy with Softmax Layer rather than Pool Layer [4].

The paper [5] presented a vision based deep learning architecture for signer independent Indian sign language recognition system. Total 24 ISL static alphabets were trained using CNN, with training accuracy of 99.93% and testing and validation accuracy of 98.64%. The acquired recognition accuracy exceeds the majority of the techniques [5].

The author of paper [6] proposed a model where the network uses a Stochastic gradient descent optimizer as its optimizer to train the network having a learning rate of 1×10^{-2} . With a batch size of 500, the network was trained over the period of 50 epochs. The image size for training and evaluation was (50, 50, 1). To improve outputs, the Keras and CNN architecture is used, which comprises a number of layers for data processing and training. The CNN layers included more 64 filters. The fully connected layer is being specified by the dense layer along with rectified linear activation [6].

Ankita Wadhawan, Parteek Kumar [7] proposed sign language recognition system includes four major phases that are data acquisition, image preprocessing, training and testing of the CNN classifier. The model training is based upon convolutional neural networks. Preprocessed sign pictures were fed into the classifier, which assigns them to the appropriate category. The dataset of several ISL gestures is used to train the classifier. The system achieved training and validation accuracy of 99.76% and 98.35%, respectively, using RMSProp and it has been found that the SGDoptimizer outperformed Adam, RMSProp and other optimizers with training and validation accuracy of 99.90% and 98.70%, respectively, on gray scale image dataset [7].

The paper [8] used a modified version of JoeyNMT to implement the Sign Language Transformers. All the components of network were built using the PyTorch. 8 heads in each layer and 512 hidden units were used to construct the framework. Used a batch size of 32 to train the networks using the Adam optimizer. Network is evaluated at every 100 iterations [8].

Boundary identification of the sign presented the biggest hurdle in CSLR. The paper indicates Use of transfer learning to solve this issue. The authors developed a two-step solution model and a post- processing methodology. The hand-crafted SVD utilized to feed features to LSTM Network extracted from the features using the prediction model. Both the SVD feature extractor and the hand pose estimator receives a window size of 50 frames. Then, the many-to-one LSTM Network maps the matching SVD feature sequence. Into a single vector up to 50 frames. After that, a Fully Connected (FC) layer receives this vector. Finally, the FC outputs are covered with a Softmax layer. The suggested methodology employed a specified threshold, 0.51, to approve or reject a recognized class for the separation of the isolated signs in a continuous sign video. There are various difficulties with the comparable signs. In the sliding window that is open. There were some difficulties in the placards that said "Congratulation", "Excuse", "Upset", "Blame", "Fight" and "Competition". In order to learn more powerful features to better describe sign categories and decrease miss-classifications as a result, adding more samples could to title t inter-class variation [9].

Paper suggested an architecture based on CNN, containing dense, max-pooling, dropout, and many convolutional layers (fully-connected). Which performs convolution on input with various filter and kernel sizes to map feature. Model correctly learns features from three main blocks, each of which has a different parameter configuration and uses ReLU as an activation function. The flattening layer transforms input into a vector before connecting to a group of the fully connected layer. Authors evaluated the model on new data in a later stage and reported accuracy of 99.67% [10] (Table 1).

4 Discussion and Conclusion

Numerous authors with expertise in the deep learning field offered novel approaches to the Sign Language Recognition challenge. The dataset is the most crucial prerequisite for a sign language recognition system. On the internet, there exist numerous datasets for various sign languages, including ASL, American Sign Language, CSL, etc. For training and testing, authors [3] [4] [5] [6] [7] manually created their own dataset on the system. When implementing a model with a pooling layer, authors in [4] experienced poor accuracy (58%), authors in [9] encountered conflicts with other signs having the same hand movements, resulting in the wrong classification, and authors in [8] used a pre-trained CNN+LSTM+HMM setup followed by RELU function. To achieve good accuracy, the majority of the authors seeded their input layer using extracted features and retrained the model on average over 38–45 iterations. Among the other proposed systems for static signs, the author [7] achieved the best performance accuracy of 99.90%. The Softmax layer [4] technique produced a CSLR accuracy of 93%.

Table 1. Summary of CNN Based Approaches Applied for Sign Language Recognition.

Year	Authors	Title	Architecture	Performance/ Results
2019	G.Anantha Rao, K.Syamala, P.V.V.Kishore, A.S.C.S.Sastry [1].	"Deep Convolutional Neural Networks for Sign Language Recognition"	CNN	Models average recognition rate was 92.88%.
Apr 2018	Jie Huang, Wengang Zhou, Qilin Zhang, Houqiang Li, Weiping Li [2].	"Video-Based Sign Language Recognition without Temporal Segmentation"	Two-stream 3D CNN	Proposed method LS-HAN achieves 82.7 % accuracy which is more than LSTM-E 76.8%.
June 2018	M.A Hossen, Arun Govindaiah, Sadia Sultana, Alauddin Bhuiyan [3].	"Bengali Sign Language Recognition Using Deep Convolutional Neural Network"	Pre-trained VGG16, CNN	Recognition rate - validation loss of 0.3523 and validation of 84.68% achieved.
2018	Kshitij Bantupalli, Ying Xie [4].	"American Sign Language Recognition using Deep Learning and Computer Vision"	CNN, RNN, Machine learning, HMM.	Instead of using Pool Layer, the system was able to attain 93% accuracy with SoftMax Layer.
Apr 2019	Sruthi C. J and Lijiya A [5].	"Signet: A Deep Learning based Indian Sign Language Recognition System"	CNN	Training accuracy = 99.93% and validation accuracy was equal to 98.64%.
Dec 2019	Lean Karlo S. Tolentino, Ronnie O. SerfaJuan, August C.Thio-ac, Maria Abigail B.Pamahoy, Joni Rose R. Fortezaz and Xavier Jet O. Garcia [6].	"Sign language identification using Deep Learning"	CNN	The system's accuracy was on average 93.667%. The testing has a 90.04% letter recognition accuracy, a 93.44% number recognition accuracy, and a 97.52% static word identification accuracy.
Jan 2020	Ankita Wadhawan, Parteek Kumar [7].	"Deep learning- based sign language recognition system for static signs"	CNN	A performance accuracy rate of 99.90% was calculated by the authors.
2020	Necati Cihan Camg,Oscar Koller, SimonHadfifield and Richard Bowden [8].	"Sign Language Transformers: Joint End-to-end Sign Language Recognition and Translation"	RNN-based attention architectures, Connectionist Temporal Classification (CTC).	Compared to earlier approaches, the authors Language Transformers outperform both their recognition and their translation effectiveness with a 2% reduction in word error rate.
13 Apr 2022	Razieh Rastgoo, Kourosh Kiani, Sergio Escalera [9].	"Word separation in Continuous sign language using isolated signs and post- processing"	CNN, LSTM	The proposed model obtains an average of recognized Softmax outputs equal to 0.98 and 0.59.
30 Apr 2022	Abdul Mannan, Ahmed Abbasi, Abdul Rehman Javed, Anam Ahsan, Thippa Reddy Gadekallu, Qin Xin [10].	"Hypertuned Deep Convolutional Neural Network for Sign Language Recognition"	CNN	On test data the suggested Deep CNN model has a 99.67% accuracy rate when recognising ASL alphabets.

Each research suggested a virtually identical architecture, including image augmentation, feature extraction, and fully connected layers for the subsequent classifier outcomes. This research examined the challenges, advancements, and probable future directions of the vision-based hand gesture recognition system. It seems that the publications we read emphasized the value of data collection, features, and the training data's context. It was also noted that the majority of databases utilised in hand gesture recognition research were from a constrained context, underlining the need for sign language databases that are less constrained and incorporate data from diverse environments. The conclusion of this study is thus more attention needs to be paid to the uncontrolled environment, setting to build vision-based gesture recognition system for practical use. Because it can provide researchers a chance to enhance the system's capability to recognize hand gestures in any form of environment. Data collection is a core procedure that has been stressed and placed in the spotlight of many vision-based hand gesture recognition studies.

References

1. G. Anantha Rao, K. Syamala, P.V.V. Kishore, A.S.C.S. Sastry, "Deep Convolutional Neural Networks for Sign Language Recognition." 2018.
2. Jie Huang, Wengang Zhou, Qilin Zhang, Houqiang Li, Weiping Li, "Video-Based Sign Language Recognition without Temporal Segmentation", AAAI-18, April 2018.
3. M.A Hossen, Arun Govindaiah, Sadia Sultana, Alauddin Bhuiyan "Bengali Sign Language Recognition Using Deep Convolutional Neural Network" June 2018.
4. Kshitij Bantupalli, Ying Xie, "American Sign language identification using Deep Learning and Computer Vision." IEEE 2018.
5. Sruthi C. J and Lijiya A. Member, "Signet: A Deep Learning based Indian Sign Language Recognition System." IEEE, April 2019.
6. Lean Karlo S. Tolentino, Ronnie O. Serfa Juan, August C. Thio-ac, Maria Abigail B. Pamahoy, Joni Rose R. Fortezaz and Xavier Jet O. Garcia. "Sign language identification using Deep Learning." IJMLC, December 2019.
7. Ankita Wadhawan, Parteek Kumar, "Deep learning-based sign language recognition system for static signs", Jan 2021.
8. Necati Cihan Camg, Oscar Koller, Simon Hadfifield and Richard Bowden, "Sign Language Transformers: Joint End-to-end Sign Language Recognition and Translation", 2020.
9. Razieh Rastgoo, Kourosh Kiani, Sergio Escalera, "Word separation in continuous sign language using isolated signs and post-processing", 13 Apr 2022.
10. Abdul Mannan, Ahmed Abbasi, Abdul Rehman Javed, Anam Ahsan, Thippa Reddy Gadekallu, and Qin Xin "Hypertuned Deep Convolutional Neural Network for Sign Language Recognition", 30 April 2022.
11. Priyanka Rawat "All about Indian Sign Language", 11 Feb 2021. https://1specialplace.com/2021/02/11/all-about-indian-sign-language/
12. Van Hiep Phung and Eun Joo Rhee, "A High-Accuracy Model Average Ensemble of Convolutional Neural Networks for Classification of Cloud Image Patches on Small Datasets", 23 October 2019.https://www.researchgate.net/figure/Schematic-diagram-of-a-basic-convolutional-neural-network-CNN-architecture-26_fig1_3368059

13. Meena Ugale, Rohit Nalawde, Apoorv Nagap, Lakhan Jindam "Agriculture Field Monitoring and Plant Leaf Disease Detection", 3rd International Conference-IEEE, CSCITA-2020, April 2020. https://doi.org/10.1109/CSCITA47329.2020.9137805.
14. Gonsalves, T., Upadhyay, J., "Integrated deep learning for self-driving robotic cars", Artificial Intelligence for Future Generation Robotics, 2021, pp. 93–118.