A Comparative Analysis of ISLRS Using CNN and ViT

S. Renjith and Rashmi Manazhy

Abstract Indian Sign Language Recognition System (ISLRS) aims at recognizing and interpreting the hand gestures and movements in Indian Sign Language (ISL), in order to facilitate smooth communication between the hearing-impaired individuals and the normal population. This research aims at comparing ISLR System using a custom convolutional neural network (CNN) architecture as well as Vision Transformer (ViT). From the ISL alphabet dataset consisting of 36 classes, 26 classes corresponding to the English alphabets are considered in this analysis. The analysis showed that for the dataset, ViT outperforms CNN in terms of performance metrics considered.

Keywords Sign language · Deep learning · ISLRS · CNN · ViT

1 Introduction

Sign Language Recognition Systems (SLRSs) aim to translate sign language into written or spoken language. These systems use various technologies including image processing, deep learning, natural language processing, and computer vision for interpreting the gestures and movements used in sign language and convert them into meaningful text or speech. The main focus of SLR lies in bridging the communication gap between the hearing and deaf community. SLRS is designed to assist in various fields, such as education, health care, and social interactions, by enabling hearing-impaired individuals to communicate more effectively with others. Despite

S. Renjith (✉)
Department of Computer Science and Engineering, Amrita Vishwa Vidyapeetham, Amritapuri, India
e-mail: srenjith@am.amrita.edu

R. Manazhy
Department of Electronics and Communication Engineering, Amrita Vishwa Vidyapeetham, Amritapuri, India
e-mail: manazhyrashmi@am.amrita.edu

P. P. Joby et al. (eds.), *IoT Based Control Networks and Intelligent Systems*, Lecture Notes in Networks and Systems 789, https://doi.org/10.1007/978-981-99-6586-1_1

the challenges involved in developing robust and accurate Sign Language Recognition Systems, recent advances in machine learning and computer vision have led to significant progress in this field. Also, researchers continue to explore new trends in image processing and computer vision to improve the usability of these systems in real time.

The potential benefits of Sign Language Recognition Systems for the deaf and hard-of-hearing community are numerous, as they can enhance communication and accessibility in various contexts such as education, employment, and social interactions. However, developing accurate and reliable recognition systems presents several challenges, including variability in sign language gestures and differences in sign language dialects.

This research aims at comparing Indian Sign Language Recognition System (ISLRS) using two types of deep learning architectures, viz. a custom CNN architecture and ViT implemented using ResNet-50. The study discusses similar literatures in Sect. 2, methodology in Sect. 3, and results in Sect. 4. Section 5 concludes the work with further scope toward future direction.

2 Literatures

An extensive review of research on ISL recognition, covering topics such as data collection, preprocessing, feature extraction, and classification was given by Kumar et al. [1]. The authors also discussed the challenges of ISL recognition, including the complexity of sign language, variations between signers, and the lack of standardization. Amal et al. [2] focused on the challenges in developing ISL Recognition Systems, such as the lack of standardization, limited availability of annotated data, and the need for robust feature extraction and classification techniques. The authors also discussed various approaches for ISL recognition, i.e., vision-based and sensor-based methods. Ghotkar et al. [3] focused specifically on the challenges and opportunities for ISL recognition in India. Various approaches of ISL recognition, including template-based and machine learning-based methods, were discussed.

The use of deep learning frame works to hand gestures recognition for ISL was explored in [4, 5]. The work compared the performance of various deep learning models and proposed a new model for ISLRS. A CNN framework was utilized in ISL recognition in [6]. A large dataset consisting of hand gestures pertaining to sign language was used in order to achieve high accuracy in recognition. Kishore et al. [7] used various ML algorithms for ISL and compared the performance. ISL gesture dataset was used for the analysis. Joy et al. [8] worked on a hybrid approach for ISLR in real time. A rule-based system was combined with a machine learning model to achieve high accuracy and speed in recognition.

Another approach for ISLR in real time was presented in [9]. The authors proposed a methodology to capture the signs using an inexpensive data glove. The captured data was processed, and a support vector machine (SVM) classifier was used to recognize the signs. The real-time system could recognize 40 different ISL signs.

A sensor-based glove was used to capture the hand gestures and hence used as a dataset. The testing of the above system was carried out by 40 different ISL signs performed by 10 different people, and an accuracy of 91.2% was achieved. A user interface for the system, which displays the recognized sign on a computer screen, was also created in this work. This interface can be used to communicate with deaf and dumb people who understand ISL.

Rokade et al. [10] proposed an ISLR system for using a computer vision-based approach. The authors used a database of 26 Indian Sign Language gestures performed by 10 different users. The system consists of several stages: hand region extraction, hand contour detection, feature extraction, and gesture recognition. The hand region extraction stage involves segmentation of skin color. For detecting contours in the hand, morphological operations were applied to the extracted hand region. In the feature extraction stage, the authors used the Hu moments and Zernike moments as feature descriptors from the detected hand contour. These features were used to represent the shape and texture parameters of the corresponding hand gestures. Finally, in gesture recognition stage, a SVM classifier was used to recognize the 26 Indian Sign Language gestures. The system achieved an average recognition rate of 94.23%.

Recognition of ISL gestures by combining machine learning (ML) with computer vision techniques was done by Dixit et al. [11]. The method was based on hand shape, movement, and location. The dataset consisted of 50 ISL signs captured using a Kinect sensor. It also included annotations of the signs in terms of hand shape, movement, and location. A classification model based on the K-nearest neighbor (KNN) algorithm was used to recognize ISL signs from these features. Evaluation of the system's performance on the ISL dataset shows that the proposed approach achieves an accuracy of 86.7% in recognizing ISL signs.

A framework for ISL gesture recognition was proposed by Deora et al. [12], which involves three main steps, viz. acquisition of images followed by feature extraction and classification. In the first step, images were acquired by capturing video sequences of the signer's hand movements using a camera. The feature extraction step involves analyzing the video sequences to extract relevant features such as hand shape, movement, and trajectory. Finally, the classification step involved using ML-based algorithms to recognize the gesture. Authors tested the system on a dataset of 300 ISL gestures performed by ten different signers. A recognition accuracy of 86.67% was achieved using KNN classifier and 90% accuracy using SVM classifier. The work also discussed some of the challenges involved in ISL gesture recognition, such as variations in hand shape and movement, lighting conditions, and background clutter. The authors suggested that future work could involve developing more robust feature extraction techniques and explore deep learning algorithms for gesture recognition.

ISL using SVM was proposed by Raheja et al. [13]. For this work, the dataset of 1800 images for 10 ISL gestures was collected from 18 different signers to ensure a wide range of variation in terms of appearance, background, lighting, and signer characteristics. The authors proposed a feature extraction method that extracts local binary pattern (LBP) features from the gesture images. LBP is a texture descriptor that captures the local structure of an image. The authors trained SVM classifiers for each of the 10 gestures using the LBP features. The work experimented with different kernel functions and parameter settings to find the best performing SVM classifier. An overall recognition rate of 92.78% was achieved for the 10 ISL gestures.

ISL gestures using artificial neural network (ANN) were experimented by Adithya et al. [14]. A dataset of 1650 gestures for 26 letters and 10 numerals was created in ISL. The dataset was preprocessed to remove noise and segment the hand region. In order to extract features, a combination of intensity and texture-based image processing techniques was used. The input to the ANN was the extracted features, and the output was the corresponding gesture. The method achieved an accuracy of 98.67% for recognition of ISL gestures. The method was compared with other existing techniques, and it was showed that the proposed method outperformed other methods in terms of accuracy.

The following section explains the methodology adopted in this research.

3 Methodology

This research work proposes a comparative analysis of Indian Sign Language Recognition (ISLR) System using a custom convolutional neural network (CNN) architecture as well as Vision Transformer (ViT). The open-source ISL alphabet dataset which consists of 26 alphabet classes is considered in this analysis. The ensuing subsections give the explanations on the dataset used and the deep learning methodology adopted.

3.1 Dataset

The ISL alphabet dataset consists of 36 classes, which corresponds to the 26 letters of the English alphabet and 10 numerals (0–9). Each class contains approximately 1000 images of the corresponding sign, taken from multiple signers [15]. In this work, only 26 English alphabets are used for analysis. The color images available in the dataset are converted into grayscale and resized to resolution of 200 × 200 pixels. Figure 1 shows the grayscale preview of dataset of Letters C, F, B, and M.

3.2 Custom CNN Model

The architecture of CNN [16–21] typically consists of an input layer, hidden convolutional layers, max pooling layers, and a fully connected layer followed by the output layer. In this work, a custom CNN model with three-layer architecture is considered. Since the ISL dataset contains grayscale images, the input layer has only single

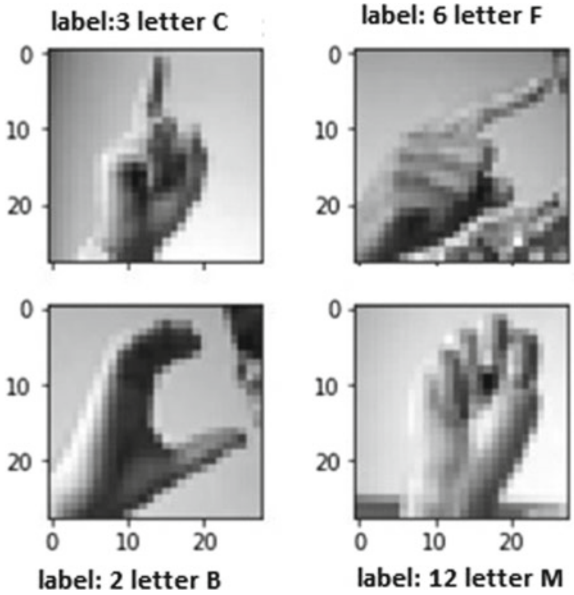


Fig.1 Preview of dataset

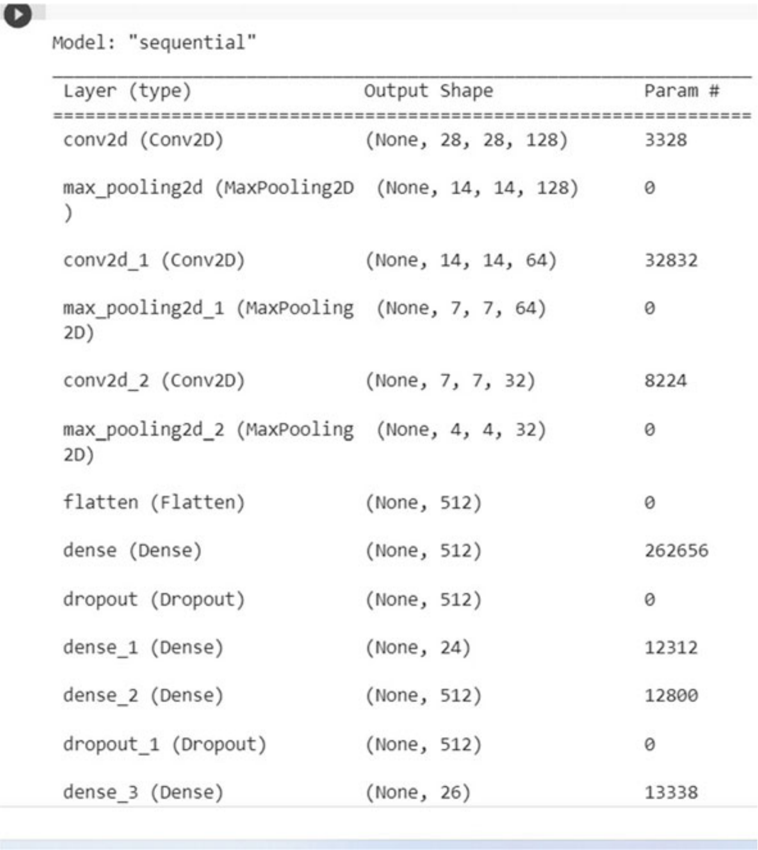
channel. The output layer has 26 neurons which are equal to the number of classes. Mathematical formula for the architecture can be expressed as

- Input layer: Size 200 × 200 × 1, where 200 × 200 denotes the input image dimension.
- Hidden layer 1 (convolutional): Size 3 × 3 × 64, where 3 × 3 is the size of the filter and 64 is the number of filters.
- Hidden layer 2 (convolutional): Size 3 × 3 × 32, where 3 × 3 is the size of the filter and 32 is the number of filters.
- Max pooling layer of 2 × 2.
- Dense layer of 4096 neurons.
- Output layer of 26 neurons.

The convolution layer used ReLU activation function, and the output layer used sigmoid activation function. Figure 2 shows the custom CNN model architecture.

3.3 Vision Transformer

The Vision Transformer (ViT) [22] is a deep learning architecture for image classification tasks that was introduced in 2020 by researchers at Google Brain. Traditionally, CNNs have been the dominant architecture for image classification tasks, but ViT offers a promising alternative. ViT divides an image into fixed-size patches and then applies self-attention mechanisms to these patches to extract features. These features



```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv2d (Conv2D)             (None, 28, 28, 128)       3328

 max_pooling2d (MaxPooling2D  (None, 14, 14, 128)      0
 )

 conv2d_1 (Conv2D)           (None, 14, 14, 64)        32832

 max_pooling2d_1 (MaxPooling  (None, 7, 7, 64)         0
 2D)

 conv2d_2 (Conv2D)           (None, 7, 7, 32)          8224

 max_pooling2d_2 (MaxPooling  (None, 4, 4, 32)         0
 2D)

 flatten (Flatten)           (None, 512)               0

 dense (Dense)               (None, 512)               262656

 dropout (Dropout)           (None, 512)               0

 dense_1 (Dense)             (None, 24)                12312

 dense_2 (Dense)             (None, 512)               12800

 dropout_1 (Dropout)         (None, 512)               0

 dense_3 (Dense)             (None, 26)                13338
```

Fig. 2 Custom CNN model architecture

are then processed by a series of fully connected layers to make the final classification. One of the key advantages of ViT over CNNs is its ability to handle long-range dependencies between image patches, which can be important for certain tasks. The Vision Transformer used in this work is a ResNet-50-based architecture.

4 Results and Discussion

ISL dataset is applied to CNN-based architecture. The CNN was trained for 10 epochs. The learning rate for the system is chosen as 0.0001. The system obtained an accuracy of 98%. The precision and recall for the system are 92% and 94%, respectively. Figure 3 shows the accuracy vs. epochs plot of ISL data on custom CNN. ISLRS on ViT-based model achieved an overall accuracy of 99%. Figure 4 shows the accuracy plot of ISL data on ViT-based model.

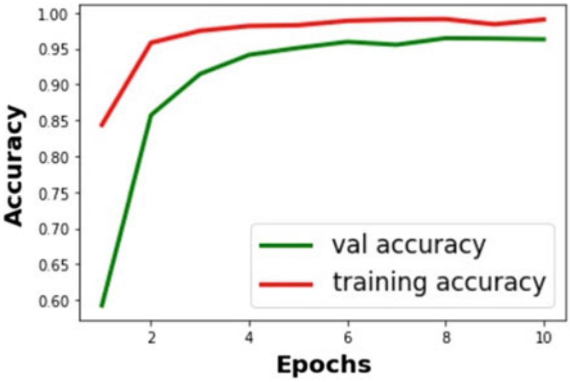


Fig. 3 Accuracy versus epochs plot of CNN model

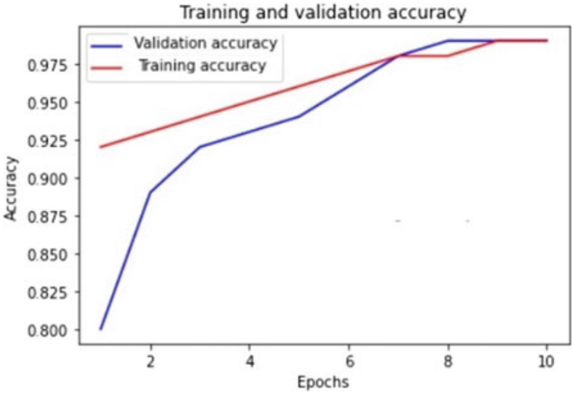


Fig. 4 Accuracy versus epochs plot of ViT model

Table1 illustrates the performance comparison of CNN and ViT. Based on the table, it can be found that the Vision Transformer model has a higher accuracy than the convolutional neural network model.

Since the models were trained and tested on the same dataset and with the same hyperparameters, it can be concluded that the ViT model is more effective in learning the features and patterns in the data and making accurate predictions. This may be due to the patch-based processing and self-attention mechanism used in the ViT model, which allows the model to attend to various regions of the input image, enabling it to capture both global and local information.

Table 1 Performance metrics comparison of CNN and ViT models

Model	Accuracy (%)	Precision (%)	Recall (%)
CNN	98.023	92.091	85.025
ViT	99.014	94.142	86.032

5 Conclusion and Scope of Future Work

This research introduces a recognition system for Indian Sign Language (ISL) that uses both CNN-based and ViT-based approaches. Based on these findings, the system can be applied to various applications, such as communication devices for individuals with hearing impairments. Even though vast amount of research works have been carried out in ISL systems with alphabet dataset, the real-time implementation of these systems needs word-level or sentence-level datasets. This area of research is still in the budding stage. Future work aims at analysis of ISL datasets which carry dynamic word/sentence-level representations. Finally, feasibility of real-time implementation of the ISL system on low-power devices such as smartphones and tablets will be investigated.

References

1. Kumar EK, Kishore PVV, Kumar DA, Kumar MTK (2021) Early estimation model for 3D-discrete indian sign language recognition using graph matching. J King Saud Univ-Comput Inf Sci 33(7):852–864
2. Amal H, Reny RA, Prathap BR. Hand kinesics in Indian sign language using NLP techniques with SVM based polarity
3. Ghotkar AS, Khatal R, Khupase S, Asati S, Hadap M (2012) Hand gesture recognition for indian sign language. In: 2012 international conference on computer communication and informatics. IEEE, pp 1–4
4. Sharma A, Sharma N, Saxena Y, Singh A, Sadhya D (2021) Benchmarking deep neural network approaches for Indian sign language recognition. Neural Comput Appl 33:6685–6696
5. Gupta R, Kumar A (2021) Indian sign language recognition using wearable sensors and multi-label classification. Comput Electr Eng 90:106898
6. Zomaya A, Wadhai V, Principal MIT, Kamilah A, Koeppen M (2012) Hybrid intelligent systems (HIS)
7. Kishore PVV, Kumar DA, Sastry ACS, Kumar EK (2018) Motionlets matching with adaptive kernels for 3-d indian sign language recognition. IEEE Sens J 18(8):3327–3337
8. Joy J, Balakrishnan K, Sreeraj M (2019) SignQuiz: a quiz based tool for learning fingerspelled signs in indian sign language using ASLR. IEEE Access 7:28363–28371
9. Rajam PS, Balakrishnan G (2011) Real time Indian sign language recognition system to aid deaf-dumb people. In: 2011 IEEE 13th international conference on communication technology. IEEE
10. Dixit K, Jalal AS (2013) Automatic Indian sign language recognition system. In: 2013 3rd IEEE international advance computing conference (IACC). IEEE
11. Rokade YI, Jadav PM (2017) Indian sign language recognition system. Int J Eng Technol 9(3):189–196
12. Deora D, Bajaj N (2012) Indian sign language recognition. In: 2012 1st international conference on emerging technology trends in electronics, communication & networking. IEEE
13. Raheja JL, Mishra A, Chaudhary A (2016) Indian sign language recognition using SVM. Pattern Recogn Image Anal 26:434–441
14. Adithya V, Vinod PR, Gopalakrishnan U (2013) Artificial neural network based method for Indian sign language recognition. In: 2013 IEEE conference on information & communication technologies. IEEE
15. Raghuveera T, Deepthi R, Mangalashri R, Akshaya R (2020) A depth-based Indian sign language recognition using microsoftkinect. Sādhanā 45(1):1–13

16. Charan MGKS, Poorna SS, Anuraj K, Praneeth CS, Sumanth PS, Gupta CVSP, Srikar K (2022) Sign language recognition using CNN and CGAN. In: Inventive systems and control: proceedings of ICISC 2022. Springer Nature Singapore, Singapore, pp 489–502
17. Charan MGKS, Poorna SS, Anuraj K, Praneeth CS, Sumanth PS, Gupta CVSP, Srikar K (2022) Comparative study of conditional generative models for ISL generation. In: IoT based control networks and intelligent systems: proceedings of 3rd ICICNIS 2022. Springer Nature Singapore, Singapore, pp 171–189
18. Aloysius N, Geetha M (2017) A review on deep convolutional neural networks. In: 2017 international conference on communication and signal processing (ICCSP), Chennai, India, pp 0588–0592. https://doi.org/10.1109/ICCSP.2017.8286426
19. Aloysius N, Geetha M (2020) Understanding vision-based continuous sign language recognition. Multimedia Tools Appl 79:22177–22209. https://doi.org/10.1007/s11042-020-08961-z
20. Al Mossawy MMT, George LE (2022) A digital signature system based on hand geometry-survey: basic components of hand-based biometric system. Wasit J Comput Math Sci 1(1):1–14
21. Sharma S, Singh S (2022) Recognition of Indian sign language (ISL) using deep learning model. Wirel Pers Commun: 1–22
22. Zhao H, Jiang L, Jia J, Torr P, Koltun V (2020) Point transformer. arXiv preprint arXiv:2012.09164