Real-time Recognition Framework for Indian Sign Language using Fine-tuned Convolutional Neural Networks

Rajat Soni, Anshul Vijay, Aakash Khandelwal, Radhika Vijay, Vipin Yadav, Deepak Bhatia

Rajasthan Technical University, Kota, Rajasthan, India

Corresponding author: Rajat Soni, Email: rajatsoni9549@gmail.com

Sign language is a very crucial aspect in the lives of those who cannot speak or listen and to those around them. People with these disabilities have difficulty communicating with the outside world and they feel left behind. Much research is ongoing to create a better way of communicating for these people. This work establishes interaction between hearing or speech impaired with the world by recognizing the 33-hand pose and gestures of Indian Sign Language (ISL). This framework can recognize alphabets and numbers in real-time and also generate gestures in real-time for the given alphabets and numbers. The fine-tuned Convolutional Neural Network (CNN) model is explored for the recognition of alphabets and numbers in real-time. A GUI is developed for an easy-to-use interface and immediate visual feedback. Data acquisition software is also developed to create a database. A database of 74,200 images of 33 static signs is captured and used in this work. The results are evaluated on different CNN architectures and learning rates. Accuracy, precision, recall, and F-score are used as performance metrics. The proposed work accomplished the most noteworthy training precision of 99.97% and a validation accuracy of 99.59%.

Keywords: Deep Learning, Sign language, CNN, Data acquisition.

2021. In Prashant Singh Rana, Deepak Bhatia & Himanshu Arora (eds.), *SCRS Proceedings of International Conference of Undergraduate Students*, 95–106. Computing & Intelligent Systems, SCRS, India. https://doi.org/10.52458/978-81-95502-01-1-10

1. Introduction

In this quickly developing world, there is a critical craving to energize the challenged section of society. Individuals with language incapacities impart in gesture-based communication and therefore have intricacy associating with able bodies. Consequently, there is a correspondence hole that is difficult to shut in standard society. The Indian Deaf Association appraises that around 1,000,000 populaces have some sort of utilitarian hearing misfortune [1].

People have adjusted communication via gestures to convey since verifiable occasions. Hand signals are just about as authentic as human civilization itself. Hand signals are particularly suitable for expressing any word or feeling that necessitates to be communicated. Therefore, despite the creation of writing conventions, population around the world continuously use hand signals to express themselves [2]. Communications via gestures use the visual-manual methodology to pass on importance. Communications via gestures are communicated through manual explanations related to non-manual components.

Much research has lately been carried out into the growth of systems that are capable of classifying signs in different sign languages. Such frameworks have discovered applications in games, computer generated reality conditions, robot controls, and normal language correspondence. Currently, the Indian communication via gestures frameworks is in the progression stage and there is no gesture-based communication acknowledgment framework open for perceiving signs progressively [2]. In this way, there is a necessity to create a total recognizer that distinguishes Indian Sign Language characters.

2. Literature Review

Gesture-based communication is the correspondence interaction for the meeting impeded in the public arena. Different researchers have dealt with gesture-based communications from their individual nations, including American, Chinese, Finnish, British, Italian, Ukrainian, and Arabic, to make a superior world for the hearing impaired [1]. G. Anantha et al. [3] they propose the recognition of Indian sign language gestures with the help of convolutional neural networks (CNN). Continuous sign language video in selfie mode is the recording method used in this work. Jaya Prakash et al. [4], who has been working on a PCA-based reduced deep CNN function is proposed for the acknowledgment of static hand signal pictures. The profundity highlights are separated from completely associated layers of pre-prepared AlexNet. The experiments are done with 36 ASL motion stances with LOO CV and Holdout CV test. Shadman et al. [5] Reason for perceiving a finger spell interpreter for American Sign Language (ASL) in light of skin segmentation and machine learning algorithms. They are a programmed shading data-based calculation to portion human skin utilizing the YCbCr shading space. Then, at that point, the Convolutional Neural Network (CNN) is applied to separate elements from the pictures, and the profound learning technique is utilized to prepare a classifier to perceive communication via gestures. Siming [6] worked on a hand location network based on the Faster R-CNN. The feature extraction based on 3D-CNN and the recognition process based on the recurring neural network LSTM (long and short-term memory). Ka Leong et al. [11] propose a fully convolution network (FCN) for online SLRs to at the same time take in spatial and worldly components from feebly explained video successions with just recorded explanations at the sentence level. A Gloss Enhancement Module (GFE) will be acquainted with the proposed network to empower better succession arrangement learning.

Ashish et al. [1] systematically looked at between three promising deep learning-based methodologies: the pre-prepared VGG16 model, the normal language-based output network, and the hierarchical network for recognizing ISL signals. The hierarchical network outperforms the other two models with a precision of 98.52% for one-handed and 97% for two-handed gestures. Javed et al. [7] used a state-of-the-art large and deep neural network (NN) that consolidates

convolution and max-pooling (MPCNN) for directed element learning and the grouping of hand signals from people to versatile robots with hued gloves to focus on an ongoing hand motion-based HRI interface for portable robots. Versatile robots with an ARM 11 533 MHz processor accomplish constant signal acknowledgment execution with a classification rate of 96%. E. Kiranet al. [12] proposed a two-stream CNN design that utilizes two shading coded pictures, the geological descriptor of normal distance (JDTD) and geographical descriptor of normal point (JATD) as information. They gathered and fostered the informational index of 50,000 sign recordings in Indian communication through signing and accomplished an exactness of 92.14%. Lucas et al. [8] developed a new approach to feature extraction for hand posture detection using depth and intensity images captured by a Microsoft Kinect sensor. They applied this method to the classification of finger spelling in American Sign Language utilizing a Deep Belief Network, which our feature extraction method is custom fitted. The aftereffects of a multi-client informational collection with two situations: one with all known users achieves 99% memory and precision and one with an unknown user achieves 77% memory and 79% precision.

In light of the above necessities, this paper expects to foster a total framework dependent on profound learning models to perceive 33 static Indian Sign Language signs gathered from various users. It is a successful method to perceive Indian communication via gestures digits and letter sets that are utilized in day-by-day life. The profound learning-based convolutional neural organization (CNN) design comprises layers of convolution followed by different layers. A webcam-based informational index of static characters was caught under various ecological conditions. The presentation of the proposed framework was surveyed utilizing different profound learning models, streamlining agents, accuracy, review, and F-score.

The paper is structured as follows. Section 3 describes the data acquisition. Section 4 illustrated the system design and architecture. Section 5 describes the methodology. Section 6 describes experiments and results. Finally, the conclusion and future scope in section 7.

3. Data Acquisition

The colour images are retrieved from the camera using data acquisition software developed by us and these images are then passed on to the image pre-processing module. The dataset comprises of the assortment of RGB pictures for different static characters. The dataset contains 74,188 pictures of the static characters. There are 33 distinctive person classes that incorporate 23 English letter sets and 0 - 9 digits. The dataset comprises of static sign pictures of various sizes, colours and recorded under various ecological conditions to help the better speculation of the classifier.

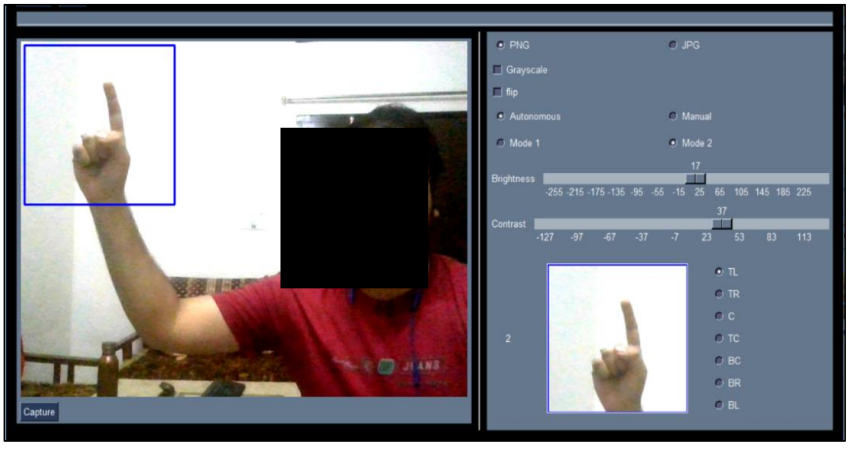

Fig. 1: GUI of data acquisition software

The first phase is data collection phase: Software was developed using python for collection of images for the dataset. We collected 39600 RGB images of static characters that include 23 English alphabets and 0 - 9 digits. The Software interface asks the user about the name of character for which the data will be collected using web-cam of laptop, the number of images to be taken, and time interval between the two successive snaps. The user can adjust the brightness and contrast of the image from user interface. The user is needed to frame hand gestures of that specific character keeping the hands inside the boundaries of a square being shown on the camera window. Each captured image is immediately resized to 100x100 image before saving so as to feed directly into the model. This dataset was collected with different backgrounds and lighting conditions. This collected dataset was saved in hierarchical folder/file structure. Data augmentation is performed on this dataset like flipping the image. After data augmentation and removing redundant images the final dataset comprises of 74,188 images which now can be used for better generalization of the classifier.

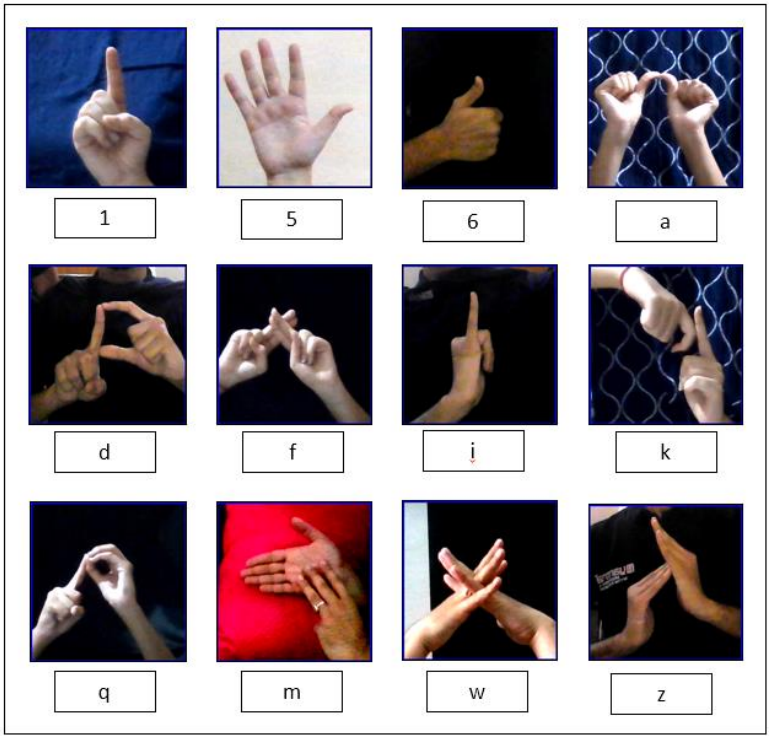

Fig. 2: Sample Data set

4. System design and architecture-

The proposed system through signing acknowledgment framework includes three primary stages, in particular data acquisition, training, and testing of the CNN classifier. The accompanying figure depicts the information stream graph that addresses the functioning model of the framework. The first stage is the data capture stage, in which the RGB information from static signs is caught with a camera. The gathered sign pictures are then pre-handled, i.e., Flip picture, and so forth. These pictures are saved in memory for some time later. In the following stage, the proposed framework is prepared with the CNN classifier and afterward, the prepared model is utilized for testing. The last

stage is the trying stage, which fine-tunes the parameters of the CNN architecture until the outcomes are just about as precise as wanted.

We likewise utilized Keras Tuner to calibrate the network configuration to track down the best values of filters, kernels, learning rate and FC layers for our dataset. The last plan of the CNN architecture utilized in the proposed framework is portrayed in Table 1.

Table 1: Proposed model specifications

Layer (type)	**Output**	**Parameters**
conv2d_9 (Conv2D)	(98, 98, 256)	7168
dropout_3 (Dropout)	(98, 98, 256)	0
conv2d_10 (Conv2D)	(96, 96, 176)	405680
max_pooling2d_6 (MaxPooling2)	(48, 48, 176)	0
conv2d_11 (Conv2D)	(44, 44, 192)	844992
max_pooling2d_7 (MaxPooling2)	(22, 22, 192)	0
flatten_3 (Flatten)	(92928)	0
dense_9 (Dense)	(330)	30666570
dense_10 (Dense)	(250)	82750
dense_11 (Dense)	(33)	8283
Total parameters: 32,015,443 Trainable parameters: 32,015,443 Non-trainable parameters: 0		

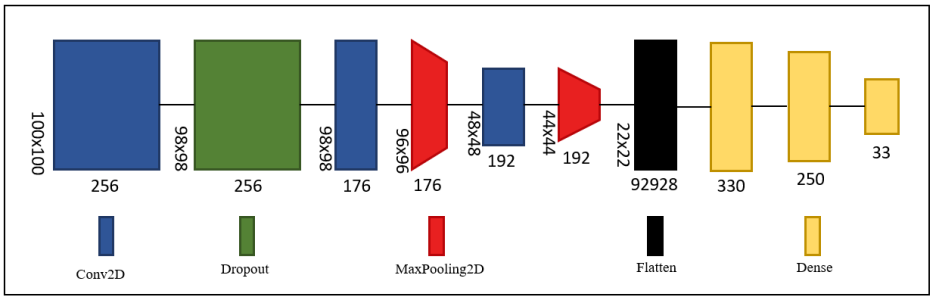

Fig. 3: Visualization of CNN architecture of proposed model

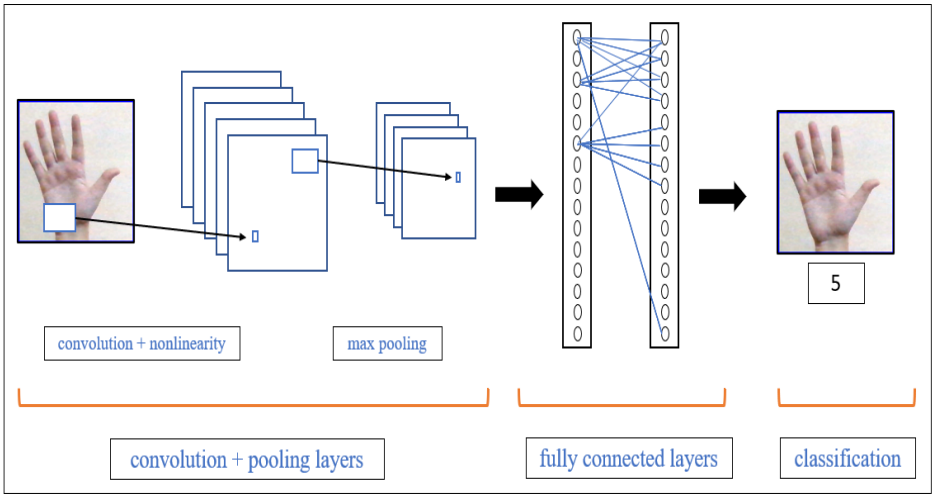

Fig. 4: General architecture of CNN model

5. Methodology

The proposed communication via gestures acknowledgment framework contains two principle stages, to be specific information securing, preparing, and testing of the last CNN classifier. The accompanying figure depicts the general framework stream chart that addresses the functioning model of the framework.

In the following stage is preparing and testing the CNN classifier: The proposed model is prepared with the NVIDIA GeForce GTX 1650 Graphical Processing Unit (GPU), 4 GB of RAM, 16 GB of irregular access memory (RAM) and 1000 GB of strong state drive (SSD). Around 30 different models were developed with different combination of convolutional layers, pooling layers, flatten, dropout (for avoiding overfitting of model) and dense layers which were then tuned for getting best hyperparameter values using keras tuner.

Table 2: Test results regarding parameters

S. No.	Number of layers	Number of Filters	Training accuracy (%)	Validation accuracy (%)	Validation Loss
1	8 (5CL,3FC)	64	98.25	97.55	2.57
2	7 (4CL,3FC)	32	98.55	95.39	2.99
3	6 (3CL,3FC)	256	98.98	97.91	0.1237
4	6 (3CL,3FC)	256	99.97	99.59	0.0297

Some conclusions were drawn when different combinations of layers were used with different filter sizes. We saw that the exactness of the proposed model increments as we decline the quantity of layers of the CNN architecture. Training and validation accuracy is increased to 99.97% and 99.59%, respectively, when the levels are reduced from 8 to 6, with an improved validation loss of 0.0297. On the other hand, if we change the number of filters from 64 to 256 filters, the accuracy is increased. The RMSprop optimiser is utilized to change the parameters or weights of the model, which assists with limiting the loss and to foresee results as precisely as could be expected, as displayed in Table 3.

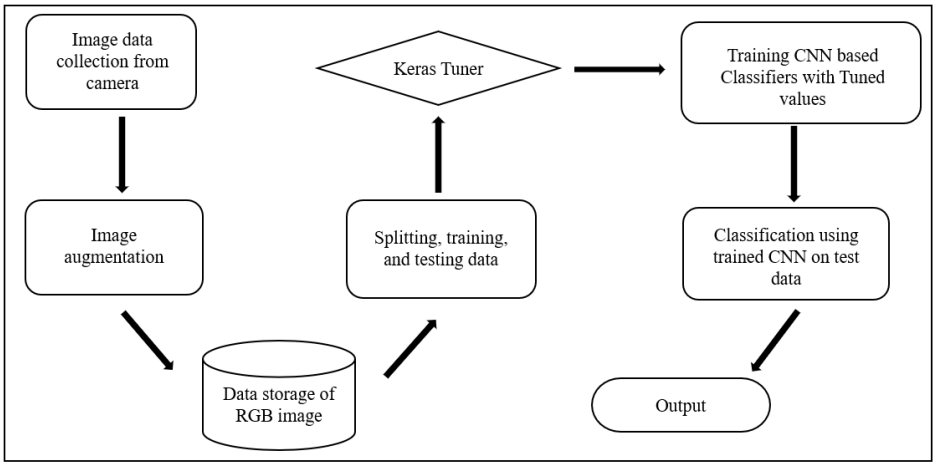

Fig. 5: System Flowchart proposed Model

Table 3: Test results regarding data split ratio

Data Split Ratio (Train: Test)	Training accuracy	Validation accuracy	Validation Loss
3:7	99.99	98.76	0.1468
5:5	99.91	99.50	0.0463
7:3	99.97	99.59	0.0297

The 3CL, 3FC model was finally used which is then again used to derive another conclusion regarding the test train split ratio. We observed that the accuracy and validation loss of the proposed model change when we change the data split ratio; i.e., the correlation between training and test data. We noticed that best outcomes in terms of training accuracy, validation accuracy, and validation loss were attained when the data split ratio was 7:3. The proposed model also performed well when the proportion of training data was kept low (3: 7), although the validation loss was sacrificed to 0.1468, as shown in Table 4. Combining the dataset is critical to add irregularity to the technique of neural organization preparing, which keeps the organization from being slanted as far as specific boundaries. The last proposed model design utilizes RMSprop optimizer utilized to prepare the model for a limit of 30 epochs with the loss function as categorical cross entropy.

6. Experimentation and Results

The performance of the Indian Sign Language acknowledgment framework is assessed based on two unique analyses. First and foremost, the parameters utilized in preparing the model are adjusted in which the number of layers and number of filters have been changed. In the second experiment, the performance of the trained model is evaluated data split ratio. The normal accuracy, recall, F1-score, and precision of the ISL acknowledgment framework have additionally been processed.

Precision is defined as; Precision = True Positive/ (True Positive + False Positives)

The Recall is defined as; Recall = True Positive/ (True Positive + False Negative)

The F1-score is defined as; F1-score = (2 * Precision * Recall)/ (Precision + Recall)

The best results in terms of training accuracy, validation accuracy and loss of validation were achieved with a data split ratio of 7: 3 (training data: test data) after 30 epochs. The completion loss and accuracy were- loss: 0.0011; accuracy: 0.9997. The validation loss and accuracy were- loss: 0.0297; accuracy: 0.9959.

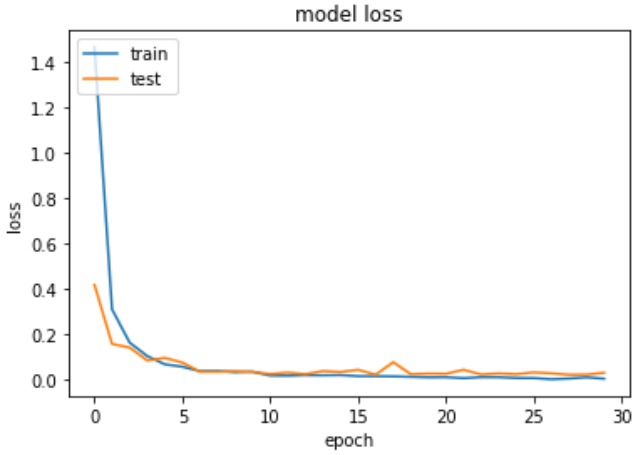

Fig. 6: Model Loss of proposed system

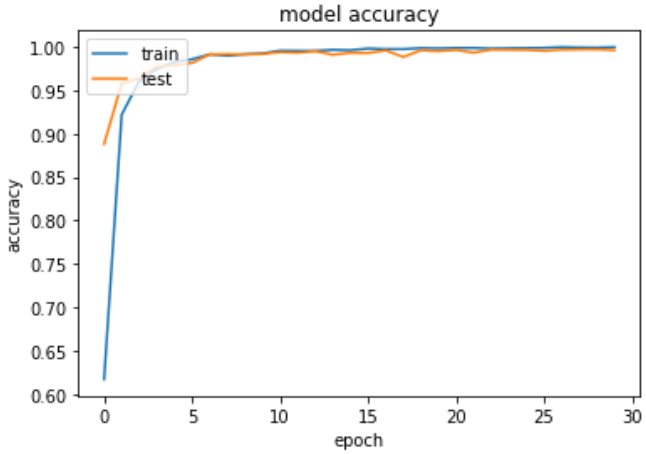

Fig. 7: Model Accuracy of proposed system

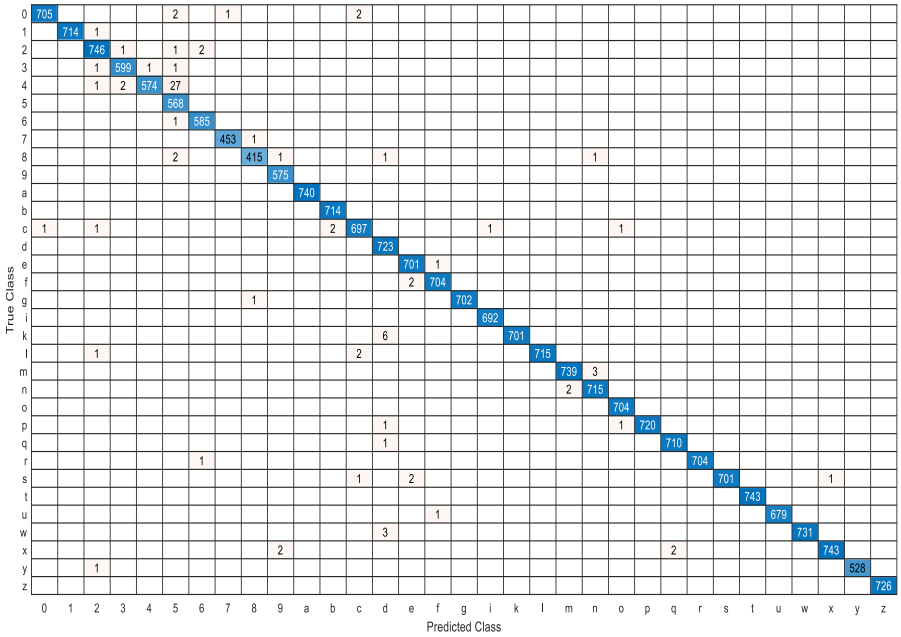

Fig. 8: Confusion Matrix of proposed system

The classification performance for all static characters indicating precision, recall, and F1 score is shown in Table 2.

Table 4: Classification report of proposed system

Classes	Precision	Recall	F1- Score	Support
0	1.00	0.99	1.00	710
1	1.00	1.00	1.00	715
2	0.99	0.99	0.99	750
3	1.00	1.00	1.00	602
4	1.00	0.95	0.97	604
5	0.94	1.00	0.97	568
6	0.99	1.00	1.00	586
7	1.00	1.00	1.00	454
8	1.00	0.99	0.99	420
9	0.99	1.00	1.00	575
a	1.00	1.00	1.00	740
b	1.00	1.00	1.00	714
c	0.99	0.99	0.99	703
d	0.98	1.00	0.99	723
e	0.99	1.00	1.00	702
f	1.00	1.00	1.00	706
g	1.00	1.00	1.00	703
i	1.00	1.00	1.00	692

k	1.00	0.99	1.00	707
l	1.00	1.00	1.00	718
m	1.00	1.00	1.00	742
n	0.99	1.00	1.00	717
o	1.00	1.00	1.00	704
p	1.00	1.00	1.00	722
q	1.00	1.00	1.00	711
r	1.00	1.00	1.00	705
s	1.00	0.99	1.00	705
t	1.00	1.00	1.00	743
u	1.00	1.00	1.00	680
w	1.00	1.00	1.00	734
x	1.00	0.99	1.00	747
y	1.00	1.00	1.00	529
z	1.00	1.00	1.00	726
Accuracy			1.00	22257
macro avg	1.00	1.00	1.00	22257
weighted avg	1.00	1.00	1.00	22257

6.1 Comparison with Existing Systems

The comparative examination of the proposed Indian communication through signing acknowledgment framework with different classifiers utilizing our own dataset is displayed in Table 4. Our strategy, we proposed a framework for Indian gesture-based communication acknowledgment that utilizes a profound learning-based CNN technique. The proposed system for recognizing Indian sign language was observed to outperform all other existing ISL systems with training and validation accuracy of 99.97% and 99.59%, respectively. It has likewise been presumed that the CNN collapsing structure is utilized in huge informational indexes utilizing the backpropagation calculation which shows how a machine may change its parameters used to deliver the portrayal in each layer from the portrayal in the past layer to rate.The aftereffects of the proposed CNN-based communication through signing acknowledgment framework are best while exploring different avenues regarding various quantities of layers in the CNN architecture. The thorough examinations were additionally done to track down the optimal parameter's esteems (number of layers, kernel size) for executing the algorithm.

Table 5: Comparison with existing models

Author	**Technique used**	**Recognition rate (%)**
Ashish et al. [1]	Hierarchical network	97.76
G. Anantha et al. [3]	Artificial neural network	98
Javed et al. [7]	MPCNN Supervised feature learning	96
E. Kiran et al. [12]	Two stream CNN architecture	92.14

Our Model	CNN	99.59

7. Conclusion and Future Scope

This research presents a viable strategy for perceiving ISL digits and letter sets. The proposed CNN architecture is planned with convolution layers followed by dropout layers and max pooling layers. Using software developed in-house, a data set of 74,188 images of 33 static signs was created under various ambience. The proposed model design was tried on roughly 30 CNN models utilizing the RMSProp optimiser. The framework brings about the most noteworthy training and validation accuracy of 99.97% and 99.59%, respectively, in terms of parameter changes such as the number of layers and data sharing ratios. The outcome of the proposed framework was likewise assessed based on accuracy, recall, and F-score. It was tracked down that the framework beat other existing frameworks even with a lower number of data and epochs.

For future work, more data sets desire to be collected in order to refine the detection method. In addition, the trained CNN model is being experimented with to perceive characters progressively in real timeLikewise, the framework will be extended to perceive dynamic characters that require the assortment and development of a video-based dataset, and the framework will be tried utilizing the CNN design by separating the recordings into video frame. The frames of the training set are given to the CNN model for the training process. At last, the prepared model is utilized as a future reference to make expectations of the training and test data. The work will likewise be extended to foster a mobile-based application for perceiving different characters in real-time.

References

[1] Sharma, A., Sharma, N., Saxena, Y., Singh, A. and Sadhya, D., 2020. Benchmarking deep neural network approaches for Indian Sign Language recognition. Neural Computing and Applications, pp.1-12.

[2] Wadhawan, A. and Kumar, P., 2020. Deep learning-based sign language recognition system for static signs. Neural Computing and Applications, 32(12), pp.7957-7968.

[3] Rao, G.A., Syamala, K., Kishore, P.V.V. and Sastry, A.S.C.S., 2018, January. Deep convolutional neural networks for sign language recognition. In 2018 Conference on Signal Processing and Communication Engineering Systems (SPACES) (pp. 194-197). IEEE.

[4] Sahoo, J.P., Ari, S. and Patra, S.K., 2019, December. Hand gesture recognition using PCA based deep CNN reduced features and SVM classifier. In 2019 IEEE International Symposium on Smart Electronic Systems (iSES) (Formerly iNiS) (pp. 221-224). IEEE.

[5] Shahriar, S., Siddiquee, A., Islam, T., Ghosh, A., Chakraborty, R., Khan, A.I., Shahnaz, C. and Fattah, S.A., 2018, October. Real-time American Sign Language Recognition using skin segmentation and image category classification with convolutional neural network and deep learning. In TENCON 2018-2018 IEEE Region 10 Conference (pp. 1168-1171). IEEE.

[6] He, S., 2019, October. Research of a sign language translation system based on deep learning. In 2019 International Conference on Artificial Intelligence and Advanced Manufacturing (AIAM) (pp. 392-396). IEEE.

[7] Nagi, J., Ducatelle, F., Di Caro, G.A., Cireşan, D., Meier, U., Giusti, A., Nagi, F., Schmidhuber, J. and Gambardella, L.M., 2011, November. Max-pooling convolutional neural networks for vision-based hand gesture recognition. In 2011 IEEE International Conference on Signal and Image Processing Applications (ICSIPA) (pp. 342-347). IEEE.

[8] Rioux-Maldague, L. and Giguere, P., 2014, May. Sign language fingerspelling classification from depth and color images using a deep belief network. In 2014 Canadian Conference on Computer and Robot Vision (pp. 92-97). IEEE.

[9] Cheok, M.J., Omar, Z. and Jaward, M.H., 2019. A review of hand gesture and sign language recognition techniques. International Journal of Machine Learning and Cybernetics, 10(1), pp.131-153.

[10] Beena, M.V., Namboodiri, A. and Thottungal, R., 2020. Hybrid approaches of convolutional network and support vector machine for American Sign Language prediction. Multimedia Tools and Applications, 79(5), pp.4027-4040.

[11] Cheng, K.L., Yang, Z., Chen, Q. and Tai, Y.W., 2020, August. Fully Convolutional Networks for Continuous Sign Language Recognition. In European Conference on Computer Vision (pp. 697-714). Springer, Cham.
[12] Kumar, E.K., Kishore, P.V.V., Kumar, M.T.K. and Kumar, D.A., 2020. 3D sign language recognition with joint distance and angular coded color topographical descriptor on a 2–stream CNN. Neurocomputing, 372, pp.40-54.
[13] Albawi, S., Mohammed, T.A. and Al-Zawi, S., 2017, August. Understanding of a convolutional neural network. In 2017 International Conference on Engineering and Technology (ICET) (pp. 1-6). IEEE.
[14] Abhishek, K.S., Qubeley, L.C.F. and Ho, D., 2016, August. Glove-based hand gesture recognition sign language translator using capacitive touch sensor. In 2016 IEEE International Conference on Electron Devices and Solid-State Circuits (EDSSC) (pp. 334-337). IEEE.
[15] Ittisarn, P. and Toadithep, N., 2010, July. 3d animation editor and display sign language system case study: Thai sign language. In 2010 3rd International Conference on Computer Science and Information Technology (Vol. 4, pp. 633-637). IEEE.
[16] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R. and Fei-Fei, L., 2014. Large-scale video classification with convolutional neural networks. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (pp. 1725-1732).
[17] Hong, C., Yu, J., Wan, J., Tao, D. and Wang, M., 2015. Multimodal deep autoencoder for human pose recovery. IEEE Transactions on Image Processing, 24(12), pp.5659-5670.
[18] Ibrahim, M.S., Muralidharan, S., Deng, Z., Vahdat, A. and Mori, G., 2016. A hierarchical deep temporal model for group activity recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1971-1980).
[19] Shou, Z., Wang, D. and Chang, S.F., 2016. Temporal action localization in untrimmed videos via multi-stage cnns. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1049-1058).
[20] Zhu, G., Zhang, L., Shen, P. and Song, J., 2017. Multimodal gesture recognition using 3-D convolution and convolutional LSTM. Ieee Access, 5, pp.4517-4524.
[21] Wang, J., Yang, Y., Mao, J., Huang, Z., Huang, C. and Xu, W., 2016. Cnn-rnn: A unified framework for multi-label image classification. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2285-2294).
[22] Dong, C., Leu, M.C. and Yin, Z., 2015. American Sign Language alphabet recognition using MicrosoftKinect. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops (pp. 44-52).
[23] Tushar, A.K., Ashiquzzaman, A. and Islam, M.R., 2017, December. Faster convergence and reduction of overfitting in numerical hand sign recognition using DCNN. In 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC) (pp. 638-641). IEEE.
[24] Yang, S. and Zhu, Q., 2017, May. Video-based Chinese sign language recognition using convolutional neural network. In 2017 IEEE 9th International Conference on Communication Software and Networks (ICCSN) (pp. 929-934). IEEE.
[25] Oyedotun, O.K. and Khashman, A., 2017. Deep learning in vision-based static hand gesture recognition. Neural Computing and Applications, 28(12), pp.3941-3951.
[26] Pigou, L., Dieleman, S., Kindermans, P.J. and Schrauwen, B., 2014, September. Sign language recognition using convolutional neural networks. In European Conference on Computer Vision (pp. 572-578). Springer, Cham.
[27] Molchanov, P., Gupta, S., Kim, K. and Pulli, K., 2015, May. Multi-sensor system for driver's hand-gesture recognition. In 2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG) (Vol. 1, pp. 1-8). IEEE.
[28] Simonyan, K. and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
[29] CNN with keras, https://www.kaggle.com/amarjeet007/visualize-cnn-with-keras