Fig. 5 — Keyframe extraction: (a) Uniform, (b) Normalized

method, the video frames are normalized with respect to the time period. The frames that are at the beginning and end of the video have low intensity in the normalized curve as they contain incomplete information regarding the gesture class, as shown in Fig. 5(b). The mean obtained is used for computing the keyframes. The frames are taken at a constant distance from one another around the mean. Four frames are taken to the left of the mean time period and five frames to the right of the meantime period, along with the mean time period frame.

These keyframes are augmented translatively to improve the dataset count. The videos are reduced to $10 \times 32 \times 32 \times 3$ size for easier computation through the Deep CNN layers. Here, 10 represent the number of frames for the video and. $32 \times 32 \times 3$ represents the size of each frame.

Gesture Recognition

Training took place for a maximum of 15 minutes and included a set of complex operations performed using the GPU. Three performers recording five videos for fifteen gestures each, making up to 225 videos, which are augmented and used as a dataset for training the CNN model, where several optimizers are utilized to have the maximum accuracy and minimum test loss. The Adam optimizer is more efficient than other as results depicted in the Table 1. The frames of the videos were split into train and test with a random state. The videos were running over the model for about 250 epochs until the change in parameters became negligible. Every epoch was evaluated within 4 seconds. For every epoch, there was an improvement in accuracy along with a reduction in loss, which demonstrates that the dataset is trained in the right manner.

Table 1 — Comparison of test accuracies and loss over various optimizers

Optimizers/Parameters	Test accuracy	Test loss
SGD	0.77847113884	1.66007373262
RMSProp	0.67831513260	0.81501329488
Adamax	0.71903276131	1.17280815172
Adam	0.89903276131	0.93798099793

Table 2 — Grammar phrases represented as classes forming patterns to predict word 3 from word1 and word 2

S. No	Word 1	Word 2	Word 3
1	1	2	3
2	1	3	5
3	1	4	7
4	1	5	9
5	1	6	11
6	1	7	13
7	1	14	15

Sentence Validation

The sentence sequences consist of three classes comprising a grammar sequence. The three classes are separated as two training inputs and one next-word for the sequence in the same order. These grammar sequences are trained over an RNN- LSTM network with soft max classification to predict classes. There are 47 records of such grammar sequences that are being formed in patterns. These sequences represent the sequential order in which the sign language words are expected to appear in grammar, as depicted in Table 2. The dataset is being split into 37 training records and 10 testing records. The training period was ~10 minutes, with around 4 seconds for each epoch. Since there are only a few records available for training, it is being split with a batch size of 1. This data was trained for about 200 epochs, and the accuracy of the training model increases over epoch.

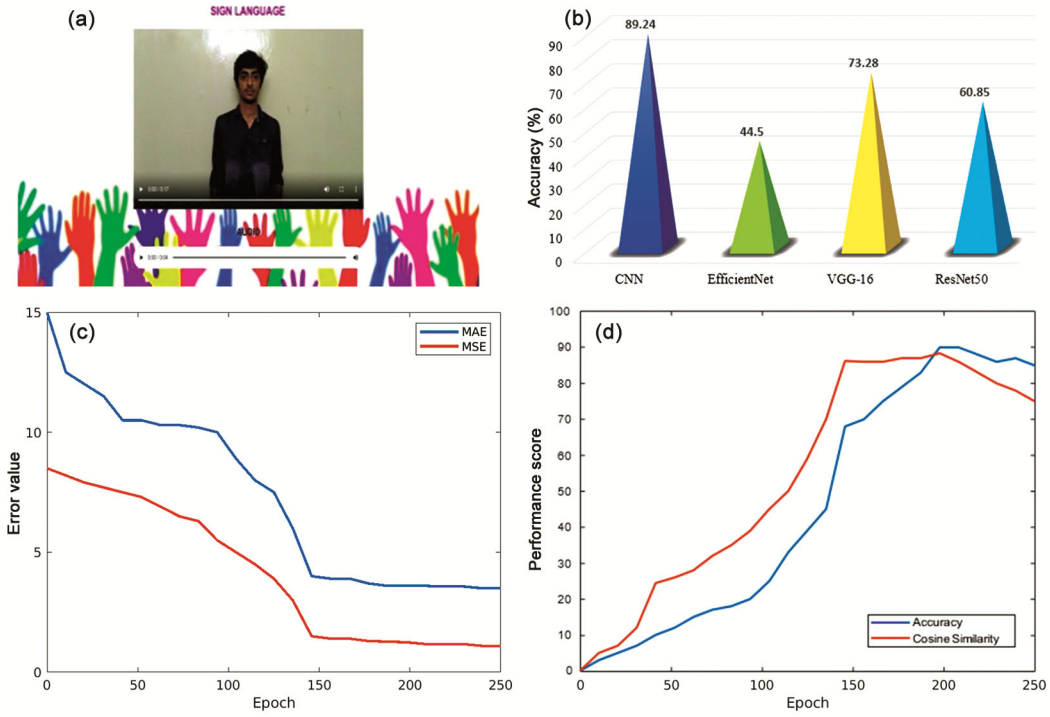

Fig. 6 — Performance analysis: (a) UID of sign language convertor, (b) Performance analysis of designed CNN model with pre-trained CNN models, (c) Error value of next-word prediction over epoch, (d) Accuracy and cosine similarity score over epoch

Table 3 — Confusion matrix sign language gesture recognition

S. No.	Labels	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	House	0.875	0.0	0.0	0.0	0.0	0.0	0.0	0.125	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	Neighbour	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	Address	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	Family	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5	Relative	0.0	0.0	0.0	0.0	0.941	0.059	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
6	Children	0.0	0.0	0.0	0.0	0.0	0.875	0.0	0.125	0.0	0.0	0.0	0.0	0.0	0.0	0.0
7	People	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
8	Person	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
9	Engagement	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.923	0.07	0.0	0.0	0.0	0.0	0.0
10	Baby	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.077	0.923	0.0	0.0	0.0	0.0	0.0
11	Marriage	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.923	0.077	0.0	0.0	0.0
12	Divorce	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
13	Enemy	0.125	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.875	0.0	0.0
14	Birth	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.091	0.0	0.0	0.0	0.0	0.091	0.818	0.0
15	Funeral	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0

The model, after reaching stability gives a test accuracy of 89.99%. The pre-recorded videos of signs are trained over Deep CNN to detect and classify the gestures. The set of sentence sequences is trained over the RNN-LSTM network to find the next-word aiding sentence formation and adhering to the grammar. An algorithm to evaluate every sentence sequence and modify the sentence in the case of wrong grammar has also been implemented. The video is converted to a list of gestures, and the list is modified to have proper grammar and is converted to its voice equivalent.

The User Interface (UI) of the proposed work with ease of interaction is shown in Fig. 6(a). The confusion matrix of 15 gestures being classified for the test data after 250 epochs is illustrated in Table 3. From this matrix, it is clear that the misclassification of data among the trained gestures is very low. Thus, a sequence of keyframes extracted from the recording after normalization is classified as having accuracy,

Table 4 — Comparative study of normalized and uniform keyframe extraction

Performance Metrics	Uniform	Normalized
Accuracy	0.797	0.892
Precision	0.823	0.894
Recall	0.797	0.807
F1 Score	0.787	0.770

precision, recall, and f1 score values of 89.2%, 89.4%, 80.7%, and 77% respectively. The metrics of the normalized keyframe extraction methodology in comparison to the uniform keyframe extraction mechanism are depicted in the Table 4. To compare the performance of the designed CNN with existing pre-trained models like EfficientNet, VGG-16, and ResNet50, they were trained and tested with normalised keyframe extraction. During the analysis, CNN outperformed the other three networks with a maximum accuracy of 89.2%, while the other pre-trained models ended up with an accuracy of 44.5%, 73.2%, and 60%, respectively and shown in the Fig. 6(b).

There is a significant increase in accuracy when detecting gestures using the novel method. Evaluation of these performance metrics for the CNN model indicates that extracting key frames closer to the mean outperforms extracting frames uniformly across the video. It is observed in Fig. 6(c) that the accuracy of the RNN-LSTM model for predicting the next word kept increasing over epochs and reached a maximum of 89.99% after 200 epochs. Training for 200 epochs provided a cut above any other option. The cosine similarity of the predicted next-word with the actual next-word also was increasing. The mean-square and mean-absolute errors of the model had a steep decrease while training for more epochs, plotted in Fig. 6(d). Dropout after every layer helped in avoiding over fitting for the model. This proved that the model was a good fit to predict the next-word and eventually complete the whole sentence with grammar.

Conclusions

The novel normalized key frames extraction from pre-recorded videos has proved advantageous by improving accuracy with a value of 84.2% using custom CNN. Also, RNN-LSTM, which was used to validate the sentence detected by CNN, ensures the reliability of the model. The voice output for the video gives us a clear picture of what is being conveyed by the hearing- and speech-impaired people. Thus, Sign Language Recognition can be accomplished by getting video samples of gestures and evaluating them against the known Sign Language. This can be put to use in real life, enhancing communication. This also helps us break the barrier between common people and hearing-impaired ones. The proposed system is not designed to recognize specific regional language signs, like Indian Sign Language signs. As the Sign Languages are prone to being regional, the application could be made to facilitate the needs of the users. At the outset, the proposed model is capable of recognizing sign gestures and predicting the sentence with an accuracy of 89.99% with the help of custom CNN-LSTM model.

Acknowledgement

This research work is supported by TIH-IoT CHANAKYA Group (PhD, PG & UG) Fellowship Program, 2021-2022, TIH Foundation for IoT and IoE, IIT Bombay (TIH- IoT), Mumbai, India. We are immensely thankful to MIT students for their help rendered to prepare the dataset of ISL relationship signs.

References

1 Kim S, Park G, Yim S, Choi S & Choi S, Gesture-recognizing hand-held interface with vibro tactile feedback for 3D interaction, *IEEE Trans Consum Electron*, **55** (2009) 1169–1177, DOI:10.1109/TCE.2009.5277972.

2 Soumya R M, Deepthi K, Goutam S & Anirban S, A feature weighting technique on SVM for human action recognition, *J Sci Ind Res*, **79(7)** (2020) 626–630, DOI:http://nopr.niscpr.res.in/handle/123456789/54986.

3 Jayanthi P, Ponsy R K B, Swetha K & Subash S A, Real time static and dynamic sign language recognition using deep learning, *J Sci Ind Res*, **81(11)** (2022) 1186–1194, DOI:https://doi.org/10.56042/jsir.v81i11.52657.

4 Jayanthi P & Sathia P R K B, Gesture recognition based on deep convolutional neural network, *Proc Int Conf Adv (IEEE)* 2018, 367–372, DOI:10.1109/ICoAC44903.2018.8939060.

5 Palak M, Pawanesh A & Parveen K L, Scene based classification of aerial images using convolution neural networks, *J Sci Ind Res*, **79(12)** (2020) 1087–1094, DOI:http:// nopr.niscpr.res.in/handle/123456789/55729.

6 Mohandes M, Deriche M & Liu J, Image-based and sensor-based approaches to Arabic sign language recognition, *IEEE Trans Hum Mach Syst*, **44** (2014) 551–557, DOI:10.1109/THMS. 2014.2318280.

7 Kritika N & Madhu S, Automated isolated digit recognition system: an approach using HMM, *J Sci Ind Res*, **70(4)** (2011) 270–272, DOI:http://nopr.niscpr.res.in/handle/123456789/11585

8 Elmezain M, Al-Hamadi A, Appenrodt J & Michaelis B, A Hidden Markov Model-based continuous gesture recognition system for hand motion trajectory, *Proc Int Conf Pattern Recognition* (Tampa, FL) 2008, 1–4, DOI:10.1109/ICPR.2008.4761080.

9 He K, Zhang X, Ren R & Sun J, Spatial pyramid pooling in deep convolutional networks for visual recognition, *Proc*

Comput Vis ECCV (Zurich) 2014, 346–361, DOI:https:// doi.org /10. 1007/978-3-319-10578-9_23

10 Kishore P V V, Prasad M V D, Prasad C R & Rahul R, 4-Camera model for sign language recognition using elliptical Fourier descriptors and ANN, *Proc Int Conf Signal Proc Commun Eng Syst* (Guntur, India) 2015, 34–38, DOI:10.13140/RG.2.1.4220.8803.

11 Starner T & Pentland A, Real-time American sign language recognition from video using Hidden markov models in motion-based recognition, *Comput Image Vis*, **12** (1997) 227–243, DOI:10.1109/ISCV.1995.477012.

12 Pankajakshan P C & Thilagavathi B, Sign language recognition system, *Proc Int Conf on Inno in Infor Embedded and Commun Syst* (Coimbatore, India) 2015, 2–5, DOI:10.1109/IC IIECS.2015.7192910

13 Dardas N H & Georganas N D, Real-time hand gesture detection and recognition using bag-of-features and support vector machine techniques, *IEEE Trans Instrum Meas*, **60** (2011) 3592–3607, DOI:10.1109/TIM.2011.2161140.

14 Adithya V, Vinod P R & Gopalakrishnan U, Artificial neural network based method for Indian sign language recognition, *Proc Int Conf Info & Commun Tech IEEE* (Thuckalay, Tamil Nadu, India) 2013, 1080–1085, DOI:10.1109/ CICT.2013.6558259.

15 Gaolin F & Wen G, A SRN/HMM system for signer-independent continuous sign language recognition, *Proc Int Conf on Automatic Face Gesture Recognition IEEE* (Washington-DC, USA) 2002, 312–317, DOI:10.1007/ 3-540-47873-6_8.

16 Haque P, Das B & Kaspy N N, Two-handed Bangla sign language recognition using principal component analysis (PCA) and KNN algorithm, *Proc Int Conf on Electr Comput Commun Eng* (Cox's Bazar, Bangladesh) 2019, 1–4, DOI:10.1109/ECACE.2019.8679185.

17 Bao P, Maqueda A I, Del-Blanco C R & Garcıa N, Tiny hand gesture recognition without localization via a deep convolutional network, *IEEE Trans Consum Electron*, **63** (2017) 251–257, DOI:10.1109/TCE.2017.014971

18 Kumar E K, Kishore P V V, Sastry A S C S, Kumar M T K & Kumar D A, Training CNNs for 3-D sign language recognition with color texture coded joint angular displacement maps, *IEEE Signal Process Lett*, **25** (2018) 645–649, DOI:10.1109/LSP.2018.2817179

19 Bantupalli K & Xie Y, American sign language recognition using deep learning and computer vision, *Proc Int Conf on Big Data IEEE* (Seattle, WA, USA) 2018, 4896–4899, DOI:10.1109/BigData.2018.8622141.

20 Islam M R, Mitu U K, Bhuiyan R A & Shin J, Hand gesture feature extraction using deep convolutional neural network for recognizing American sign language, *Proc Int Conf on Frontiers of Signal Proc* (Poitiers, France) 2018, 115–119, DOI:10.1109/ICFSP. 2018.8552044.

21 Molchanov P, Gupta S, Kim K & Kautz J, Hand gesture recognition with 3D convolutional neural networks, *Proc Conf Comput Vis. Pattern Recognit* (Boston, MA) 2015, 1–7, DOI:10.1109/CVPRW.2015.7301342.

22 Rung-Huei L & Ming O, A real-time continuous gesture recognition system for sign language, *Proc Int Conf on Automatic Face and Gesture Recognition IEEE* (Nara) 1998, 558–567, DOI:10.1109/AFGR.1998.671007.

23 Wang H, Leu M C & Oz C, American sign language recognition using multi-dimensional hidden Markov models, *J Inf Sci Eng*, **22(5)** (2006) 1109–1123.

24 Pradeep K, Himaanshu G, Partha P R & Debi P D, Coupled HMM based multi-sensor data fusion for sign language recognition, *Pattern Recognit Lett*, **86** (2017) 1–8, DOI: 10.1016/ j.patrec.2016.12.004

25 Mittal A, Kumar P, Roy P P, Balasubramanian B & Chaudhuri B B, A modified LSTM model for continuous sign language recognition using leap motion, *IEEE Sens J*, **19** (2019) 7056–7063, DOI : 10.1109/JSEN.2019.2909837.

26 Chuan C, Regina E & Guardino C, American sign language recognition using leap motion sensor, *Proc Int Conf on Mach Learn Appl* (Detroit, MI) 2014, 541–544, DOI:10.1109/ICMLA.2014.110

27 Kumar P, Roy P P & Dogra D P, Independent Bayesian classifier combination based sign language recognition using facial expression, *J Inform Sci*, **428** (2018) 30–48, DOI:10.1016/ j.ins.2017.10.046

28 Naglot D & Kulkarni M, Real time sign language recognition using the leap motion controller, *Proc Int Conf on Inventive Comput Tech* (Coimbatore, Tamilnadu) 2016, 1–5, DOI:10.1109/INVENTIVE.2016.7830097.

29 Hisham B & Hamouda A, Arabic sign language recognition using Ada-Boosting based on a leap motion controller, *Int J Inf Tecnol*, **13** (2021) 1221–1234, DOI: https://doi.org/ 10.1007/s41870-020-00518-5

30 Huang S, Mao C, Tao J & Ye Z, A novel Chinese sign language recognition method based on keyframe-centered clips, *IEEE Signal Process Lett*, **25** (2018) 442–446, DOI:10.1109/ LSP.2018.2797228

31 Zhu G, Zhang L, Shen P & Song J, Multimodal gesture recognition using 3-d convolution and convolutional LSTM, *IEEE Access*, **5** (2017) 4517–4524, DOI:10.1109/ACCESS. 2017.2684186.

32 Man G & Sun X, Interested keyframe extraction of commodity video based on adaptive clustering annotation, *Appl Sci*, **12** (2022) 1502, DOI:https://doi.org/10. 3390/app12031502.