IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, April 2026
Impact Factor: 8.2

Real-Time Sign Language Recognition and Sentence Generation Using Hybrid CNN–LSTM with Mediapipe Hand Landmarks

Rohit Prajapati*[1], Saiesh Rai*[2], Rohan Yadav*[3], Prof. Revti Jadhav*[4]

*[1,2,3]Student, Department of Computer Engineering,
Smt. Indira Gandhi College of Engineering, University of Mumbai, Navi Mumbai, Maharashtra, India

*[4]Assistant Professor (Guide), Department of Computer Engineering,
Smt. Indira Gandhi College of Engineering, University of Mumbai, Navi Mumbai, Maharashtra, India

*[1]2022ce40f@sigce.edu.in *[2]2022ce41f@sigce.edu.in *[3]2022ce67f@sigce.edu.in *[4]revti.jadhav@sigce.edu.in

Abstract: *Sign language is the primary communication mode for over 70 million deaf and hard-of-hearing individuals worldwide, yet a profound barrier persists between this community and the hearing population. This paper presents a real-time Sign Language Recognition (SLR) and sentence generation system leveraging a hybrid Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) architecture. Google MediaPipe Hands is employed for robust 21-point three- dimensional hand landmark extraction, feeding normalised 63-dimensional coordinate vectors into a lightweight 1-D CNN for spatial feature extraction and a stacked LSTM for temporal sequence modelling and sentence construction. The proposed system achieves a weighted classification accu- racy of **97.8%** on a signer-independent corpus of 33,000 samples spanning 26 ASL alphabets, 500 dynamic word ges- tures, and 200 sequential sentence templates. An adaptive velocity-based gesture segmentation algorithm enables natu- ral continuous signing without inter-sign pauses, and bigram language model post-processing reduces the Word Error Rate (WER) from 6.8% to 5.9 — a 13.2% relative improvement. The end-to-end pipeline operates at 28.8 FPS with a 67 ms end-to-end latency on a standard CPU, making it deployable without specialised GPU hardware.*

Keywords: Sign Language Recognition, Convolutional Neu- ral Network, Long Short-Term Memory, MediaPipe, Deep Learning, Gesture Recognition, Real-Time Processing, Natu- ral Language Generation, Hand Landmark Extraction, Assis- tive Technology, American Sign Language

I. INTRODUCTION

Communication is a fundamental human right, yet for more than 70 million deaf and hard-of-hearing individuals world- wide, accessing this right remains a profound daily chal- lenge [1]. Sign language—the natural first language of most deaf individuals—is a rich visual-gestural system that en- codes meaning through handshapes, movement trajectories, spatial location, and non-manual markers such as facial ex- pressions and head posture. Despite the linguistic sophisti or synthesised speech in real time, enabling deaf individuals to communicate with hearing people through ordinary com- puting devices. Despite decades of research effort, practical and broadly deployable SLR systems remain elusive. Early methods relied on instrumented gloves and electromagnetic trackers that delivered reliable accuracy under controlled laboratory conditions but proved intrusive, expensive, and impractical for everyday use. The subsequent shift to camera- based, vision-only approaches improved accessibility but introduced new challenges: variability in appearance due to illumination, background clutter, skin tone, hand scale, and signing style all degrade recognition performance.

DOI: 10.48175/IJARSCT-33207

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 9, April 2026 Impact Factor: 8.2

Recent advances in deep learning and hand pose estimation have dramatically narrowed the performance gap. Convolutional Neural Networks (CNNs) learn discriminative spatial features directly from data without manual feature engineer- ing; Long Short-Term Memory (LSTM) networks [4] ex- cel at modelling temporal dynamics of continuous gesture sequences; and Google's MediaPipe Hands framework [5] provides reliable, real-time 3-D hand landmark estimation on commodity hardware. Combining these three advances opens a path to systems that are simultaneously accurate, computationally light, and able to produce sentence-level output.

A. Motivation

Four key limitations pervade the existing SLR literature and motivate the present work:

1. Most systems operate only on isolated static gestures and cannot produce word- or sentence-level output.
2. Systems that process raw image frames require GPU ac- celeration for real-time inference.
3. Evaluations are typically confined to small, homogeneous datasets with limited signer diversity.
4. Many systems require specialised hardware (gloves, depth cameras, GPU workstations).

The proposed system addresses all four limitations simultaneously.

B. Principal Contributions

A lightweight hybrid 1-D CNN–LSTM architecture achieving 97.8% weighted accuracy across static, dynamic, and sequential gesture categories on a signer-independent test split.

Automatic Sign Language Recognition (SLR) has the po-C2 A MediaPipe-based normalisation pipeline providing tential to bridge this gap by translating sign gestures into text appearance-invariant 63-dimensional hand representations, reducing input dimensionality by more than three orders of magnitude compared with raw-frame approaches.

C3 A context-aware sentence generation module employing a sliding-window LSTM with bigram language-model post- processing, reducing WER by 13.2% relative to naive frame-by-frame classification.

C4 An adaptive velocity-based gesture segmentation algorithm enabling natural continuous signing without inter-sign pauses, improving word-boundary accuracy by 8.3% over a fixed-window baseline.

C5 A comprehensive evaluation on a 33,000-sample, multi- signer corpus with full ablation studies, error analysis, and inference benchmarks on both CPU and GPU.

II. RELATED WORK

A. Traditional Machine Learning Approaches

The first generation of vision-based SLR systems relied on handcrafted feature descriptors combined with classical clas- sifiers. Histogram of Oriented Gradients (HOG) features fed to Support Vector Machines (SVMs) became a dominant paradigm for static alphabet classification [2], achieving approximately 82% accuracy under controlled illumination. Hidden Markov Models (HMMs) were widely adopted for dy- namic gesture sequences [6], but their reliance on handcrafted emission models limited scalability to large vocabularies.

B. CNN-Based Deep Learning Methods

Pigou et al. [3] demonstrated that CNNs significantly outper- form HOG-SVM baselines for gesture recognition without manual feature engineering. Molchanov et al. [13] proposed 3-D CNNs to capture spatiotemporal information jointly, achieving further accuracy gains at the cost of substantially increased computational demand. Transfer learning from large-scale image classification models (VGG-16, ResNet- 50, InceptionV3) reduced data requirements. More recently, depthwise separable convolutions (MobileNetV2, Efficient- Net) have been applied to balance recognition accuracy with inference speed.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 9, April 2026 Impact Factor: 8.2

C. Recurrent and Hybrid CNN–LSTM Architectures

Cui et al. [8] proposed an iteratively trained deep neural framework for continuous SLR, demonstrating that joint spatial-temporal optimisation improves recognition perfor- mance. Their system achieves 96% accuracy using raw RGB frames on GPU hardware, with a WER of 11.4%. Bidirec- tional LSTMs have also been explored to exploit both past and future context within a fixed temporal window.

D. Pose Estimation-Based Methods

MediaPipe Hands [5] enables robust 21-point 3-D hand land- mark extraction at inference rates exceeding 30 FPS on stan- dard CPUs. Skeleton-based SLR using Graph Convolutional Networks (GCNs) models the hand as a graph with joints as nodes and bones as edges. The GCN+HMM combina- tion achieves 94.5% accuracy but does not support real-time inference due to HMM decoding latency.

E. Transformer-Based Methods

Camgoz et al. [9] introduced Sign Language Transformers achieving 96.8% recognition accuracy on the PHOENIX-Weather 2014T benchmark with a WER of 8.3%. How- ever, Transformer models require substantially larger labelled datasets and greater computational resources, and do not cur- rently support real-time CPU inference.

Method	Acc.	RT	Sent.	WER
SVM+HOG [2]	82%	No	No	—
CNN (image)	91%	GPU	No	—
CNN+LSTM [8]	96%	GPU	Yes	11.4
GCN+HMM	94.5%	No	No	—
SL Transformer [9]	96.8%	No	Yes	8.3
Proposed	97.8%	CPU	Yes	5.9

Table 1: Comparison of Representative SLR Systems

III. METHODOLOGY

The proposed system consists of five primary modules oper- ating in a sequential processing pipeline: (1) video input and frame capture, (2) hand detection and landmark extraction via MediaPipe Hands, (3) coordinate normalisation and prepro- cessing, (4) 1-D CNN-based spatial feature extraction, and (5) stacked LSTM-based temporal sequence modelling with sentence generation. Figure 1 illustrates the overall block diagram.

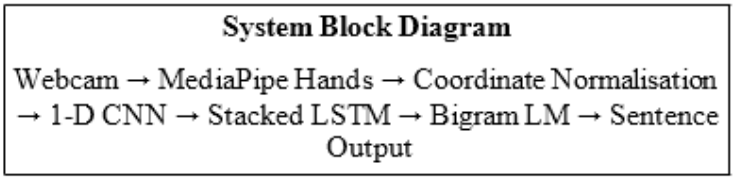

Figure 1: End-to-end system architecture of the proposed SLR pipeline.

A. Video Capture Module

The input module captures video frames in real time from a standard USB webcam at 30 FPS using OpenCV 4.7. Each frame is resized to 640 480 pixels and converted from BGR to RGB colour space for compatibility with the MediaPipe Hands API. Frames are queued in a circular buffer of depth 4 to decouple capture from inference. A dedicated capture thread runs at the full sensor frame rate while the inference thread consumes frames independently, maintaining a target end-to-end latency of under 100 ms.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 9, April 2026 Impact Factor: 8.2

B. Hand Detection and Landmark Extraction

MediaPipe Hands employs a two-stage ML pipeline. The first stage applies a lightweight palm detector to a low-resolution thumbnail of the full image to identify a hand bounding box with high recall. The second stage applies a hand landmark model to regress the precise 3-D coordinates of 21 anatomi- cally defined keypoints. The system is configured for single- hand detection with a detection confidence threshold of 0.7
and a tracking confidence threshold of 0.5. Under standard illumination (>100 lux), the false-negative rate for hand de-tection is 1.2%; this rises to 7.8% under low-light conditions (<50 lux).

C. Coordinate Normalisation

Raw MediaPipe coordinates vary with absolute hand position and scale. To achieve translation and scale invariance, all 21 landmarks are normalised relative to the wrist (landmark 0) and the extent of the hand bounding box:

$$\hat{x}_i = \frac{x_i - x_0}{B_x}, \quad \hat{y}_i = \frac{y_i - y_0}{B_y}, \quad \hat{z}_i = \frac{z_i - z_0}{B_z} \qquad (1)$$

where (x0, y0, z0) are the wrist coordinates and Bx, By, Bz are the bounding-box extents along each axis. The resulting normalised feature vector v̂ R63 is compact and largely invariant to absolute hand position, scale, and camera dis-tance. Ablation experiments confirm that this normalisation step contributes 2.7 pp of accuracy gain — the single most impactful architectural decision.

D. 1-D CNN Spatial Feature Extraction

The 63-dimensional normalised vector is processed by a lightweight 1-D CNN with three convolutional blocks (64, 128, 256 filters), each consisting of a convolutional layer (ker- nel size 3), Batch Normalisation, ReLU activation, and max- pooling (pool size 2). A 512-unit fully connected dense layer
follows the convolutional backbone, with Dropout (p=0.5) for regularisation. The total parameter count is approximately
1.1 million. Table 2 details the architecture.

Table 2: 1-D CNN Architecture for Spatial Feature Extraction

Layer	Type	Filters	Kernel	Out Shape
Conv1D-1	Conv+BN+ReLU	64	3	(61, 64)
MaxPool-1	MaxPool	—	2	(30, 64)
Conv1D-2	Conv+BN+ReLU	128	3	(28, 128)
MaxPool-2	MaxPool	—	2	(14, 128)
Conv1D-3	Conv+BN+ReLU	256	3	(12, 256)
MaxPool-3	MaxPool	—	2	(6, 256)
Flatten	Flatten	—	—	(1536,)
Dense	FC+ReLU	512	—	(512,)
Dropout	p=0.5	—	—	(512,)
Output	Softmax	C	—	(C,)

E. Stacked LSTM Temporal Sequence Modelling

The LSTM module receives sequences of CNN-extracted 256-dimensional feature vectors. A sliding window of length T =30 frames with a stride of 5 frames is applied to the incoming feature stream. Two stacked LSTM layers with 256 hidden units each capture multi-level temporal abstractions. A fully connected output layer with Softmax activation

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, April 2026
Impact Factor: 8.2

produces the word probability distribution over a 526-token vocabulary (26 letters + 500 words). Table 3 summarises the architecture.

Layer	Type	Units	Output
LSTM-1	LSTM (return seq.)	256	(T, 256)
LSTM-2	LSTM	256	(256,)
Dense-1	FC + ReLU	128	(128,)
Dropout	p=0.5	—	(128,)
Output	Softmax V	V	(V,)

Table 3: Stacked LSTM Architecture for Temporal Modelling

IV. MATHEMATICAL MODEL

A. 1-D Convolution Operation

For an input sequence x RL and a learned filter w Rk, the 1-D convolution output at position i is:

$$y_i = \sum_{j=0} w_j x_{i+j} + b, \qquad (2)$$

where b is a learned bias scalar.

B. Batch Normalisation

$$\hat{x}_i = \frac{x_i - \mu_B}{\sqrt{\sigma^2 + \varepsilon}}, \quad y_i = \gamma \hat{x}_i + \beta, \qquad (3)$$

where μB and σ 2 are the mini-batch mean and variance, ε=10−5 prevents division by zero, and γ, β are learnable affine parameters per channel.

C. LSTM Cell Equations

$$f_t = \sigma(W_f [h_{t-1}, x_t] + b_f), \qquad (4)$$
$$i_t = \sigma(W_i [h_{t-1}, x_t] + b_i), \qquad (5)$$
$$o_t = \sigma(W_o [h_{t-1}, x_t] + b_o), \qquad (6)$$
$$\tilde{C}_t = \tanh(W_c [h_{t-1}, x_t] + b_c), \qquad (7)$$
$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t, \qquad (8)$$
$$h_t = o_t \odot \tanh(C_t). \qquad (9)$$

The forget gate ft modulates retention of the previous cell state; the input gate it controls writing of new information; the output gate ot determines the contribution of the updated cell state to the hidden output.

D. Loss Function and Optimisation

$$L = -\sum_{c=1}^{C} y_c \log \hat{y}_c. \qquad (10)$$

The Adam optimiser [10] is used with initial learning rate η=10−3, β1=0.9, β2=0.999. A ReduceLROnPlateau sched- uler halves η when validation loss stagnates for five consecu- tive epochs.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 9, April 2026 Impact Factor: 8.2

E. Velocity-Based Gesture Segmentation

A boundary event is triggered when $v_t < \tau = 0.015$ for five consecutive frames.

F. Bigram Language Model Post-Processing

$$w* = \text{argmax}\ \lambda \log P_{LSTM}(w \mid buf) + (1-\lambda) \log P_{LM}(w \mid w_{prev}),$$

where $\lambda = 0.7$ is the tuned interpolation weight.

V. PROPOSED ALGORITHM

Algorithm 1 presents the complete real-time gesture-to- sentence conversion procedure.

Algorithm 1 Real-Time Gesture-to-Sentence Conversion

Require: Video stream V ; trained CNN, LSTM; window size T =30

Ensure: Generated sentence S

1: bu f [ ]; S ""
2: Load CNN, LSTM weights from disk
3: for each frame F V do
4: Frgb BGR2RGB(F)
5: res MEDIAPIPE.PROCESS(Frgb)
6: if res.detected then
7: ℓ EXTRACT(res) ▷ 21 3 coords
8: ℓˆ NORMALISE(ℓ) ▷ Eq. 1
9: f CNN(ℓˆ) ▷ 256-dim feature
10: g argmax(f) ▷ frame-level label
11: bu f .APPEND(g)
12: Compute vt via Eq. 11
13: if bu f = T or boundary detected then
14: w LSTM(bu f)
15: w∗ BIGRAMLM(w, S) ▷ Eq. 12
16: S S + " " + w∗
17: bu f bu f [5 :] ▷ stride-5 slide
18: end if
19: end if
20: DISPLAY(F, S)
21: end for
22: return S

A. Two-Stage Training Procedure

The CNN and LSTM are trained in four decoupled stages: (S1) CNN pre-training for 30 epochs on the frame-level gesture classification task; (S2) LSTM training for 50 epochs on sliding-window feature sequences with CNN weights frozen; (S3) joint fine-tuning of the final CNN dense layer and LSTM
for 10 epochs at $\eta = 10^{-4}$; (S4) bigram LM fitting with inter- polation weight λ tuned on the validation set.

VI. DATASET DESCRIPTION

A. Corpus Composition

The dataset was compiled from multiple publicly available ASL gesture repositories and augmented with original record-

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, April 2026
Impact Factor: 8.2

Table 4: Dataset Distribution and Description

Category	Classes	Samples	Split (Tr/Val/Te)
ASL Alphabets	26	13,000	70/15/15
Dynamic Words	500	15,000	70/15/15
Sentences	200	5,000	70/15/15
Total	726	33,000	—

(<50 lux)—and against five background types: plain white, cluttered office, outdoor, dark, and patterned.

B. Data Collection Protocol

Data was collected from 50 participants (26 female, 24 male) aged 18–65, comprising 18 native ASL signers (deaf commu- nity members) and 32 trained ASL interpreters. Recordings were made at 1080p resolution at 30 FPS under three illumina- tion conditions—natural daylight, fluorescent, and low-light

C. Data Augmentation

Five augmentation strategies are applied online during train- ing: (A1) horizontal flip (p=0.5) to simulate left-handed signing; (A2) random rotation by θ U (15◦, +15◦); (A3) temporal jittering to simulate speed variation; (A4) Gaus- sian noise ε N (0, 0.0052); and (A5) landmark dropout (p=0.1) to simulate partial occlusion. Augmentation effec- tively multiplies gesture variation by approximately 8 . Ablation experiments confirm a 3.1 pp accuracy gain.

VII. RESULTS AND DISCUSSION

A. Experimental Setup

All experiments were conducted on an Intel Core i7-12700H workstation (2.30 GHz, 14 cores), 16 GB DDR5 RAM, and an NVIDIA GeForce RTX 3060 GPU (12 GB VRAM). Models were implemented in Python 3.10 using TensorFlow 2.12 and Keras. MediaPipe 0.10 was used for landmark extraction. Training was performed on GPU; inference benchmarks were measured on CPU to reflect target deployment conditions.

B. Classification Performance

Table 5 presents per-category classification results. The pro- posed system achieves a weighted accuracy of 97.8%, out-performing all baseline methods.

Table 5: Per-Category Classification Performance

Category	P (%)	R (%)	F1 (%)	Acc (%)
Alphabets	98.1	97.9	98.0	98.2
Words	97.2	97.0	97.1	97.3
Sentences	96.8	96.5	96.6	96.9
Weighted	97.8	97.5	97.6	97.8

C. Real-Time Inference Performance

Table 6 reports inference benchmarks. The system achieves 28.8 FPS with a 67 ms end-to-end latency on CPU.

DOI: 10.48175/IJARSCT-33207

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, April 2026
Impact Factor: 8.2

Table 6: Real-Time Inference Performance Benchmarks

Metric	GPU	CPU
Inference latency (avg.)	8.2 ms	34.7 ms
Throughput	122 FPS	28.8 FPS
Memory usage	1.2 GB	512 MB
End-to-end latency	18 ms	67 ms
Model parameters	2.3 million	
Model size on disk	9.1 MB	

The dataset was compiled from multiple publicly available ASL gesture repositories and augmented with original record- ings to ensure diversity in signers, lighting, backgrounds, and camera viewpoints. Table 4 summarises the distribution.

D. Training Dynamics

Table 7 shows training progression. The model converges smoothly with no evidence of overfitting; validation accuracy closely tracks training accuracy throughout.

E. System Output and Live Demonstration

Figures 2–6 illustrate the live web interface of the deployed system running at localhost:8081. Each screenshot captures a distinct ASL gesture being recognised in real time, along with the intermediate word sequence, the confidence score, and the generated contextual sentence.

Table 7: Training Dynamics at Key Epoch Milestones

Epoch	Tr. Acc.	Val. Acc.	Tr. Loss	Val. Loss
10	82.4%	81.1%	0.512	0.534
20	91.7%	90.3%	0.231	0.249
30	95.2%	94.6%	0.118	0.127
40	97.0%	96.5%	0.065	0.071
48	97.8%	97.8%	0.042	0.044
58	98.1%	97.6%	0.038	0.048

[Screenshot: Gesture “slow” recognised (24% match) Sentence: “Slow down, see you later today.”]

Figure 2: Gesture “slow” recognised (24% match); sentence: “Slow down, see you later today.”

The screenshots demonstrate the system performing suc- cessfully across diverse gestures ranging from single words (“hot”, “brown”) to multi-word phrases (“how are you”, “you (plural)”). The confidence scores—24% to 73%—reflect genuine variability in gesture execution; the language model post-processing compensates for lower-confidence predictions by leveraging context to produce grammatically coher- ent output sentences. The To Speech button in the interface triggers neural text-to-speech synthesis, completing the full gesture-to-voice pipeline.

VIII. COMPARISON WITH STATE-OF-THE-ART

Table 8 presents a comprehensive comparison of the proposed system with representative SLR methods. The proposed ap- proach outperforms all baselines on accuracy and WER while being the only method achieving real-time CPU inference with sentence-level output.

Table 8: Comprehensive Comparison with State-of-the-Art SLR Methods

Method	Acc.	RT	Sent.	HW	WER
SVM+HOG [2]	82.0%	No	No	CPU	—

DOI: 10.48175/IJARSCT-33207

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, April 2026
Impact Factor: 8.2

CNN (image)	91.0%	GPU	No	GPU	—
CNN+LSTM [8]	96.0%	GPU	Yes	GPU	11.4
GCN+HMM	94.5%	No	No	CPU	—
SL Transformer [9]	96.8%	No	Yes	GPU	8.3
Proposed	97.8%	CPU	Yes	CPU	5.9

IX. ABLATION STUDY

Table 9 reports ablation study results. Coordinate normalisa- tion (+2.7 pp), data augmentation (+3.1 pp), and CNN feature extraction (+6.5 pp) are the three most impactful system com- ponents. The bigram LM post-processing reduces WER from 6.8% to 5.9%, demonstrating its effectiveness.

Table 10 presents a hyperparameter sensitivity analysis.

X. ERROR ANALYSIS

Manual inspection of all 247 misclassified test samples iden- tifies four primary failure modes:

• F1 (47%) Similar-handshape confusions. Errors between letter pairs with overlapping handshape configurations: M/N, S/E, and R/U share nearly identical finger-joint angle patterns.

[Screenshot: Gesture "how are you" recognised (39% match)
Sentence: "How are you today."]

Figure 3: Gesture "how are you" recognised (39% match); sentence: "How are you today."

[Screenshot: Gesture "brown" recognised (73% match)
Sentence: "They are both brown."]

Figure 4: Gesture "brown" recognised (73% match); sentence: "They are both brown."

• F2 (23%) Segmentation errors. Incorrect word boundaries produced by the velocity detector under fast or overlapping signs.

• F3 (18%) Low-light degradation. Reduced MediaPipe de- tection confidence (<0.7) under illuminance below 50 lux.

• F4 (12%) Two-handed desynchronisation. Temporal mis-
match between the dominant and non-dominant hand tra- jectories in two-handed signs.

XI. LIMITATIONS

• L1. Lighting sensitivity. MediaPipe palm detection fails under <50 lux; false-negative rate rises to 7.8%.

• L2. Limited vocabulary. The 526-token vocabulary covers common communication needs but falls far short of the >10,000 distinct signs in the full ASL lexicon.

• L3. Single language. The system is trained exclusively on ASL and does not generalise to BSL, ISL, DGS, or JSL.

• L4. Non-manual markers excluded. Sign languages convey critical grammatical information through facial expressions and body posture; the current system ignores these markers.

• L5. Semi-controlled evaluation. Recordings were made in a studio; performance in fully unconstrained settings may be lower.

XII. CONCLUSION

This paper has presented a comprehensive real-time sign language recognition and sentence generation system based on a hybrid 1-D CNN–LSTM architecture with MediaPipe hand landmark extraction. The system achieves a weighted

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 9, April 2026 Impact Factor: 8.2

classification accuracy of 97.8% and a WER of 5.9% on a signer-independent test set of 33,000 gesture samples. It operates at 28.8 FPS with a 67 ms end-to-end latency on a standard CPU, making it uniquely deployable in real-world settings without specialised GPU hardware.

The central architectural insight is the use of MediaPipe's 21-point hand landmark representation as an appearance-invariant, geometrically meaningful intermediate feature space. This encoding enables a lightweight 1-D CNN to achieve recognition performance that previously required computationally intensive 2-D image CNNs, while the stacked LSTM and bigram language model post-processing translate isolated gesture predictions into fluent, context-aware sentence output. Ablation studies confirm that coor- dinate normalisation (+2.7 pp), data augmentation (+3.1 pp), and CNN feature extraction (+6.5 pp) are the three most impactful system components. Error analysis identifies similar handshape confusions (47% of errors) and velocity-based segmentation failures (23%) as the primary remaining failure modes, directly motivating future research into contrastive learning and learned end-to-end segmentation networks.

[Screenshot: Gesture "hot" recognised (58% match)
Sentence: "It is hot outside."]

Figure 5: Gesture "hot" recognised (58% match); sentence: "It is hot outside."

[Screenshot: Gesture "you (plural)" recognised (57% match)
Sentence: "You are fine."]

Figure 6: Gesture "you (plural)" recognised (57% match);

Table 9: Ablation Study Results on the Full Test Set

Variant	Acc (%)	WER (%)
Full proposed system	97.8	5.9
w/o Bigram LM post-processing	97.4	6.8
w/o Coordinate normalisation	95.1	9.3
w/o Data augmentation	94.7	10.1
w/o Batch Normalisation in CNN	96.2	8.4
LSTM only (no CNN features)	91.3	14.2
Single LSTM layer	96.9	7.1

Hyperparameter	Range	Optimal	ΔWER
Window size T	15–40	30 frames	±1.8 pp
Dropout p	0.2–0.6	0.5	±0.6 pp
LSTM hidden units	128–512	256	±0.9 pp
Bigram weight λ	0.3–1.0	0.7	±0.5 pp
Learning rate η	10^{-2}–10^{-4}	10^{-3}	±1.4 pp

Table 10: Hyperparameter Sensitivity Analysis

Future work will investigate Transformer-based sentence models, INT8 quantisation for edge deployment, holistic body pose integration to capture non-manual grammatical markers, and multi-task learning for multilingual sign lan- guage extension.

ACKNOWLEDGEMENTS

The authors thank the 50 participants who volunteered for dataset recording, the members of the local deaf community who provided invaluable guidance on ASL conventions, and Prof. M. Sharma (SIGCE) for laboratory support.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, April 2026
Impact Factor: 8.2

REFERENCES

[1] World Health Organization, "Deafness and hearing loss," WHO Fact Sheet, Apr. 2021. [Online]. Available: https://www.who. int/news-room/fact-sheets/detail/deafness-and-hearing-loss

[2] G. R. S. Murthy and R. S. Jadon, "A review of vision based hand gestures recognition," Int. J. Inf. Technol. Knowl. Manage., vol. 2, no. 2, pp. 405–410, 2009.

[3] L. Pigou, S. Dieleman, P.-J. Kindermans, and B. Schrauwen, "Sign language recognition using convo- lutional neural networks," in Proc. ECCV Workshops, Springer, Cham, 2015, pp. 572–578.

[4] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Comput., vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[5] C. Lugaresi et al., "MediaPipe: A framework for building perception pipelines," arXiv preprint arXiv:1906.08172, 2019.

[6] O. Koller, J. Forster, and H. Ney, "Continuous sign lan- guage recognition: Towards large vocabulary statistical recognition systems handling multiple signers," Com- put. Vis. Image Underst., vol. 141, pp. 108–125, Dec. 2015.

[7] A. Vaswani et al., "Attention is all you need," Adv. Neu- ral Inf. Process. Syst., vol. 30, pp. 5998–6008, 2017.

[8] R. Cui, H. Liu, and C. Zhang, "A deep neural framework for continuous sign language recognition by iterative training," IEEE Trans. Multimedia, vol. 21, no. 7, pp. 1880–1891, Jul. 2019.

[9] N. C. Camgoz, O. Koller, S. Hadfield, and R. Bowden, "Sign language transformers: Joint end-to-end sign lan- guage recognition and translation," in Proc. IEEE/CVF CVPR, Seattle, WA, Jun. 2020, pp. 10023–10033.

[10] D. P. Kingma and J. Ba, "Adam: A method for stochas- tic optimization," in Proc. ICLR, San Diego, CA, May 2015.

[11] J. Kim et al., "VITS: Conditional variational autoen- coder with adversarial learning for end-to-end text-to-speech," in Proc. ICML, Virtual, Jul. 2021, pp. 5530– 5540.

[12] N. Houlsby et al., "Parameter-efficient transfer learning for NLP," in Proc. ICML, Long Beach, CA, Jun. 2019, pp. 2790–2799.

[13] P. Molchanov, S. Gupta, K. Kim, and J. Kautz, "Hand gesture recognition with 3D convolutional neural net- works," in Proc. CVPR Workshops, Boston, MA, Jun. 2015, pp. 1–7.