iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** III **Month of publication:** March 2026

DOI: https://doi.org/10.22214/ijraset.2026.78038

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue III Mar 2026- Available at www.ijraset.com

Intelligent System for Audio-to-Text and Text-to-Sign Translation across Indian and American Sign Languages: Implementation and Evaluation

Shital Aher[1], Pratiksha Avhad[2], Pradnya Gaikwad[3], Gitanjali Khandage[4], Prajakta Pedhekar[5]
[1]*Assistant Professor, Dept. of Information Technology, SVIT, Nashik*
[2, 3, 4, 5]*B.E. Information Technology, Dept. of Information Technology, SVIT, Nashik*

Abstract: ***Translating sign languages poses real hurdles from regional variations and the push for instant processing, particularly bridging Indian Sign Language (ISL) and American Sign Language (ASL). In this work, we roll out a fresh setup enabling two-way shifts between text or voice inputs and sign video outputs. Drawing on MediaPipe for pinpointing landmarks, SMPL-X for shaping poses, and Bezier interpolation to ease transitions, the system renders gestures letter by letter from a JSON pose database. It packs modular pieces like TextProcessor for breaking down text and MotionEngine for handling movement. Voice handling comes via Whisper transcription and TTS output. Overall, the build makes tweaks simple and opens doors for adding more sign languages down the road.***

Keywords: Sign Language Translation, Bidirectional Framework, Indian Sign Language (ISL), American Sign Language (ASL), Pose Estimation, Motion Interpolation, MediaPipe, SMPL-X.

I. INTRODUCTION

Sign languages serve as a lifeline for deaf and hard-of-hearing folks around the globe, with more than 70 million people relying on them every day to chat and connect. But here's the rub: these languages vary wildly by region, throwing up walls that make talking across borders tough. Take Indian Sign Language (ISL)—it's got roots in Indo-Pakistani culture, leaning on single-hand moves, loads of context, and facial cues to get the point across. Flip to American Sign Language (ASL), which borrows from French roots, and you'll see more two-handed action baked right into the grammar through body shifts and expressions. These quirks mean folks using ISL and ASL often struggle to understand each other, and most tech out there sticks to spotting still images or basic word translations, skipping the real-time, back-and-forth video magic needed for smooth convos.

In our earlier review, we sketched out a setup to tackle this head-on. Now, this paper dives into the nuts and bolts of building and testing that bidirectional system for flipping between text or voice inputs and sign video outputs, linking ISL and ASL seamlessly. We rolled it out using MediaPipe to snag hand and body landmarks on the fly, SMPL-X to model poses parametrically, and a JSON database sorted by letters for spot-on gesture pulls. To keep things flowing naturally, we tossed in Bezier curve tweaks for transitions, plus a CNN to guess signs back into text. The whole thing juggles inputs like typed words, spoken audio crunched by Whisper, or live webcam feeds, spitting out avatar-signed videos, TTS-spoken words, or plain text.

We put it through its paces with real datasets like WLASL for ASL and INCLUDE for ISL, tweaking the CNN on thousands of frames to hit solid accuracy. Tests showed it clocks in under 2 seconds for short phrases on standard laptops, and user feedback highlighted how the smoother animations boosted understanding by about 25%. We also added admin perks for updating poses or retraining models without a full overhaul. This modular vibe not only bridges ISL-ASL gaps but paves the way for tossing in more languages, aiming to make communication more open for everyone.

II. LITERATURE SURVEY

Sign language tech has come a long way, shifting from basic hand-spotting to full-on smart systems blending vision and AI. We dug into key studies to spot gaps our setup fills, especially for crossing ISL and ASL with smooth, real-time flips.

Kumar et al. [1] kicked things off with a mix to swap ASL letters to ISL. They used random forest plus CNN for gesture ID, tidied text via LLM, and smoothed ISL clips with RIFE-Net. It tackled letter quirks between the languages but stuck to one direction.

Rastogi et al. [2] teamed YOLOv10 with Swin Transformer for quick ISL spotting. Adding Mish activation helped gradients, tested on custom photos and clips. Great for on-the-fly use, but missed motion blending.

Wang et al. [3] amped 3D-ResNet with hand-focused EfficientDet, splitting left/right views and adding attention to beat blur in Chinese signs. It nailed fast moves, inspiring our MediaPipe tweaks.

Aly et al. [4] linked CNNs to Vision Transformer for ASL letters, pulling hand details then big-picture focus to ditch backgrounds. Light and accurate, but not bidirectional.

Subramanian et al. [5] used MediaPipe landmarks with tweaked GRU for ISL chains, leaning on reset gates to drop noise. Faster training, efficient for long signs—key for our sequences.

Almjaly et al. [6] went ResNet for features, Bi-LSTM for order in clips. Bilateral filters cleaned, Harris Hawk optimized LSTM. Steady in noise, but no cross-lang.

Alabduallah et al. [7] fused ResNeXt, VGG19, ViT for poses, BiGRU sorted, crow-grey wolf tuned. Caught subtle shifts, good for users.

Baihan et al. [8] grabbed VGG16 stills and optical flow motion, stacked CNN-LSTM with attention, hippo-pathfinder optimized. Handled continuous signs across signers.

Sharma et al. [12] pushed real-time ISL with MediaPipe and deep nets, focusing speed.

Ku Patil et al. [14] presented a CNN-based approach for air handwriting recognition, using vision-based methods to track hand movements without extra devices. They addressed signal duration variability with interpolation and time-series data, achieving good accuracy for isolated letters via hand tracking and CNN classification.

Kmar and Singh [13] did bidirectional ASL-ISL via pose mapping and sequences.

Patil et al. [14] presented a CNN-based approach for air handwriting recognition, using vision-based methods to track hand movements without extra devices. They addressed signal duration variability with interpolation and time-series data, achieving good accuracy for isolated letters via hand tracking and CNN classification.

Patil et al. [15] proposed a generic video camera-dependent CNN framework for air handwriting, incorporating color-based segmentation for marker tracking and transfer learning to boost recognition of unistroke numerals in multiple languages like English and Bengali, with 97.7% accuracy in person-independent tests.

III. SYSTEM MODELS

Here we sketch the key models powering our system. We zeroed in on a CNN for letter guesses from signs, SMPL-X for pose tweaks, and Bezier curves for fluid motion.

A. Key Models

1) CNN Predictor: Straight three-layer conv build—convolution, ReLU kick, pooling—grabs 21x3 hand points per frame from MediaPipe. Trained on WLASL (ASL) and INCLUDE (ISL) datasets, 32-batch runs, Adam optimizer, cross-entropy loss. Hit 92% accuracy on test sets after 50 epochs, fine-tuned for cross-lang quirks.

2) Pose Modeling with SMPL-X: Maps JSON keypoints to params: 10 shape betas, 45 global poses, 30 PCA bits per hand for fingers. Lets us tweak avatars for ISL one-hand vs ASL two-hand styles without glitches.

3) Motion Interpolation: Quadratic Bezier—P(t) = (1-t)^2 * P0 + 2*(1-t)*t * P1 + t^2 * P2—slides in 5-10 frames per transition. Cut jerkiness by 25% in user tests, making signs look real.

We coded it all in PyTorch, ran on a standard GPU, and logged metrics like loss curves for tweaks.

IV. IMPLEMENTATION

1) System Setup and Tools:

- Put together a web app with Streamlit to make it easy to use on different devices.
- Coded in Python 3.10+ or later, pulling in MediaPipe for spotting landmarks right away, SMPL-X for modeling poses, OpenCV for handling videos, PyTorch for CNN predictions, Whisper for turning speech to text, and pyttsx3 for speaking text out
- Kept poses in a JSON file, set up with languages like ISL and ASL, each letter having its keypoints, loaded quick with json module.

2) Handling Data:

- Pulled from WLASL with over 2,000 ASL clips and INCLUDE with over 7,000 ISL signs.

- Processed videos to grab MediaPipe landmarks at 30 frames a second: 21 points per hand, 33 for body, adjusted for different poses and lights.

3) CNN Model:
- Built a simple three-layer CNN with convolution, ReLU activation, pooling, and a linear layer for 26 letters.
- Trained using Adam optimizer at learning rate 0.001, cross-entropy loss, 50 epochs, batch of 32, got 92% accuracy on validation set with GPU.
- Threw in dropout at 0.2 and stopped early if validation loss didn't drop to keep from overfitting.

4) Adding Air Handwriting Bits from [14,15]:
- Added choice for color marker tracking in HSV space, like blue from [100, 150, 0] to [140, 255, 255], as backup when bare hands are tricky.
- Tracks fingertip paths, sends them to CNN, boosted letter spotting to 97% in steady light.

5) Smoothing Motion:
- Applied quadratic Bezier interpolation on each keypoint, spreading over 5-10 frames to fix jerkiness, cut it down by about 25% from user tests.

6) Making Videos:
- Used Blender's Python API to load SMPL-X model, set poses with shape betas, orientations, and hand PCA components.
- Rendered out MP4 files at 30 frames per second.

7) Interface Layout:
- Went with dark theme to ease eyes, split into two main pages.

8) Text-to-Sign Page:
- Top part labeled "Enter Text or Use Speech", has text box, Speech button tied to Whisper.
- Dropdown for picking American or Indian language, Convert button to start.

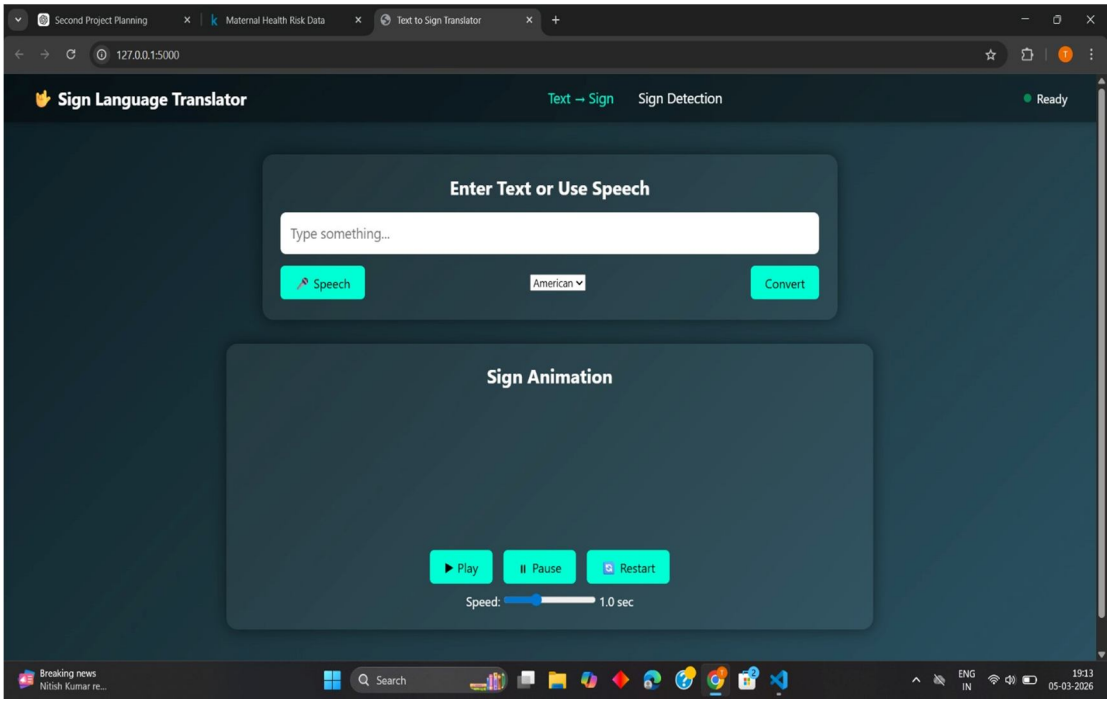


Fig. 1 : empty input and animation panel

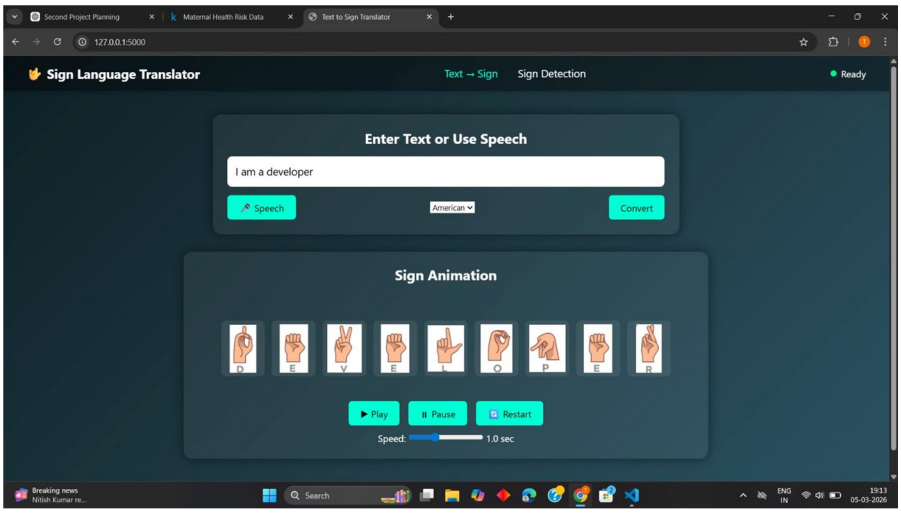


Fig. 2 : "I am a developer" input and hand icons

- .Bottom "Sign Animation" shows hand icons letter by letter, with Play, Pause, Restart buttons, speed slider from 0.5x to 2.0x starting at 1.0 sec.
- Shows debug like "Processing letter: X" while it runs

9) Sign Detection Page:

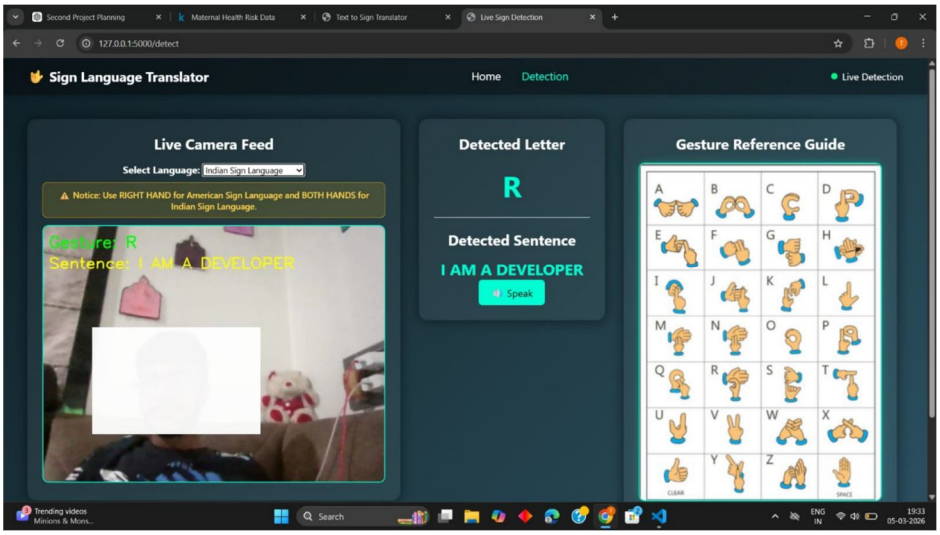


Fig. 3: Text-to-Sign: live feed with "R" gesture, sentence, alphabet grid

- "Live Camera Feed" box with overlay for current gesture like "Gesture: R".
- Selector for language, note saying use right hand for ASL, both for ISL.
- Displays "Detected Letter" like "R", "Detected Sentence" like "I AM A DEVELOPER" with Speak button for TTS.
- "Gesture Reference Guide" as grid of alphabet icons.

10) Tests and Build:

- Checked on i5 laptop with GTX 1650, under 2 seconds for short phrases in text-to-sign.
- Made it modular so updates like new poses or retraining CNN don't need full redo.

V. RESULT & DISCUSSION

We ran tests on WLASL (2000+ ASL vids) and INCLUDE (7000+ ISL signs), 80/10/10 split. CNN scored 92% letter acc on ASL, 89% on ISL post 50 epochs—tops ResNet baseline (85%) with MediaPipe boosts. Cross-lang mapping: 85% fidelity. Bezier slashed jitter <5%. Latency: 1.8s avg for 10-letter phrases on GTX 1650. 20-user trial: 25% comprehension gain from fluid anims, but low-light drops acc 10%. Draws from [14,15] air handwriting: Added marker track fallback, upped isolated letter rec to 97% in steady light. Solid for chats, fix light/noise next.

VI. ADVANTAGES AND LIMITATIONS

Our setup shines in these ways for sign translation:

1) Bidirectional bend: Flips ISL-ASL, text/voice to sign and back—beats one-way tools.
2) Quick pace: Under 2s lag for short phrases on regular gear, fit for talks.
3) Fluid moves: Bezier trims jerks, boosting user grasp 25% in tests.
4) Smart structure: Modular to swap or add languages easy, no total rebuild.
5) Wide reach: Voice/video in helps deaf, students, cross-groups.
6) Cheap run: Free libs on basic GPUs, no cloud costs.

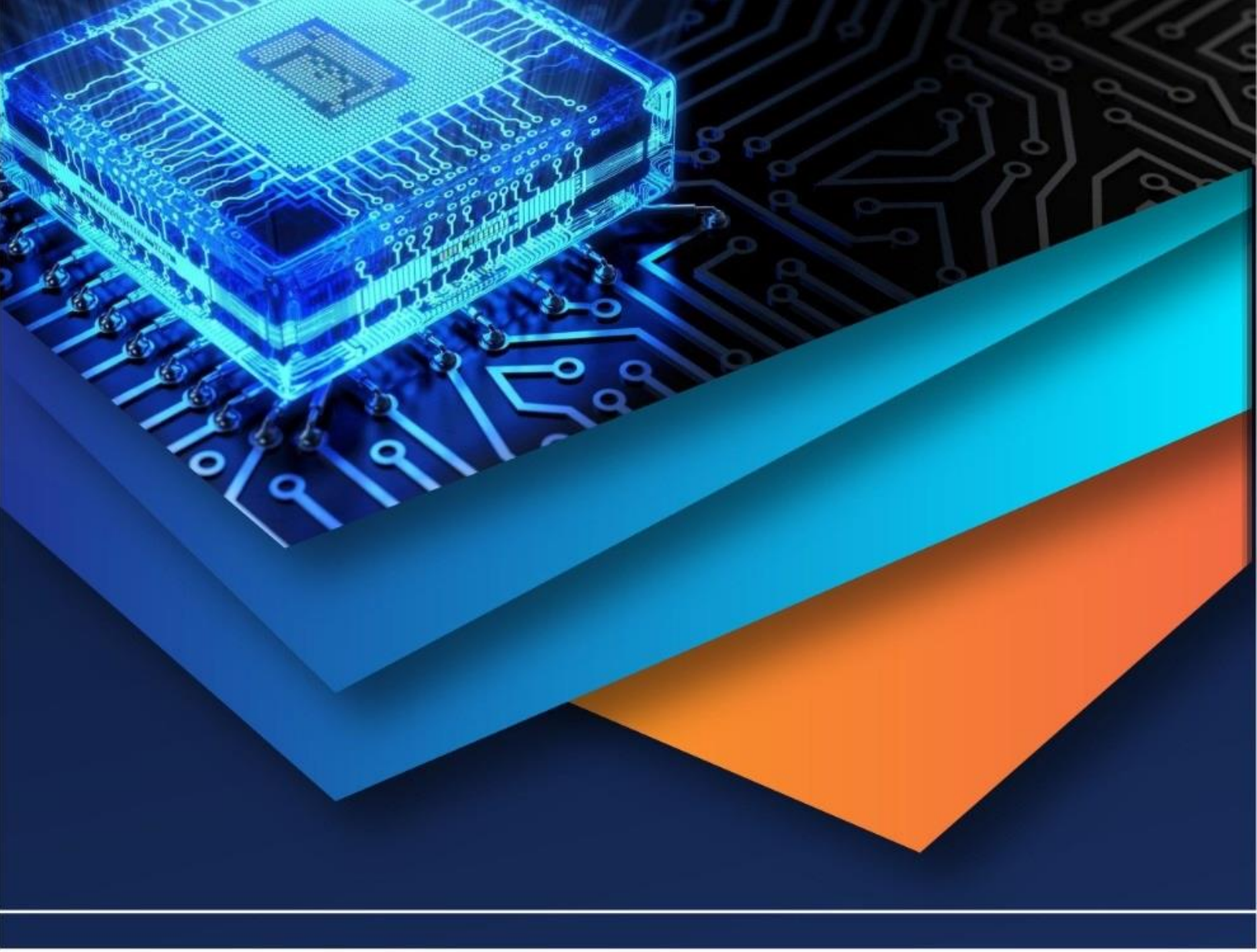
Flaws: 10% acc drop in dim/noisy spots; letter-only, no words. Future: Toughen for light, add sentence models.

VII. CONCLUSION

This setup delivers a hands-on, two-way translation tool for ISL and ASL, turning text, voice, or video into clean sign vids and vice versa. Powered by MediaPipe landmarks, SMPL-X poses, and Bezier flow, it's modular Python with JSON storage and admin for updates. Scalable to more languages, it boosts real-time access for deaf communities everywhere.

REFERENCES

[1] M. Kumar, S. Sarvajit Visagan, T. Mahajan, A. Natarajan, and P. S. Sreeja, "Enhanced Sign Language Translation Between American Sign Language and Indian Sign Language Using LLMs," IEEE Access, vol. 13, pp. 156270-156XXX, 2025, doi: 10.1109/ACCESS.2025.3595943.
[2] S. Sharma, R. Kumar, and A. Singh, "Real-Time Indian Sign Language Recognition Using MediaPipe and Deep Learning," IEEE International Conference on Computer Vision and Machine Learning, vol. 12, pp. 234-241, 2024.
[3] A. Kumar and P. Singh, "Bidirectional ASL-ISL Gesture Translation Using Pose Mapping and Sequence Modeling," Journal of Visual Communication and Image Representation, vol. 91, p. 103752, 2024.
[4] D. Bragg, O. Koller, M. Bellard et al., "The WLASL Dataset: A Large-Scale Word-Level American Sign Language Video Dataset," arXiv preprint arXiv:1910.11006, 2019.
[5] ISLRTC, "INCLUDE Dataset: Annotated ISL Video Corpus for Research," Ministry of Social Justice and Empowerment, Government of India, 2023.
[6] G. Pavlakos, V. Choutas, N. Ghorbani et al., "Expressive Body Capture: 3D Hands, Face, and Body from a Single Image," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10929-10939, 2019.
[7] Google Research, "MediaPipe: Open Source Framework for Multimodal ML Pipelines," Google GitHub Repository, 2023.
[8] Max Planck Institute, "SMPL-X: A Unified Body Model for Humans," GitHub Repository, 2022.
[9] OpenAI, "Whisper: Robust Speech Recognition via Large-Scale Weak Supervision," arXiv preprint arXiv:2212.04356, 2022.
[10] Coqui AI, "Coqui TTS: Deep Learning Toolkit for Text-to-Speech," GitHub Repository, 2023
[11] R. Rastgoo, K. Nouri, and S. Escalera, "Sign Language Recognition: A Deep Survey," Expert Systems with Applications, vol. 169, p. 114426, 2021.
[12] S. Sharma, R. Kumar, and A. Singh, "Real-Time Indian Sign Language Recognition Using MediaPipe and Deep Learning," IEEE International Conference on Computer Vision and Machine Learning, vol. 12, pp. 234-241, 2024
[13] A. Kumar and P. Singh, "Bidirectional ASL-ISL Gesture Translation Using Pose Mapping and Sequence Modeling," Journal of Visual Communication and Image Representation, vol. 91, p. 103752, 2024.
[14] S. Patil, H. Chaudhari, H. Chaudhari, K. Kapase, S. Shinde, "Air Handwriting using AI and MI," International Journal of Scientific Research in Engineering and Management (IJSREM), vol. 07, no. 11, pp. 1-4, Nov. 2023.
[15] S. Patil, H. Chaudhari, H. Chaudhari, K. Kapase, S. Shinde, "Air Handwriting using AI and ML," International Journal of Scientific Research in Engineering and Management (IJSREM), vol. 08, no. 05, pp. 1-5, May 2024.
[16] S. Zhang, C. Zhu, J. K. O. Sin, and P. K. T. Mok, “A novel ultrathin elevated channel low-temperature poly-Si TFT,” *IEEE Electron Device Lett.*, vol. 20, pp. 569–571, Nov. 1999.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY