International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(18s) 2026 pp. 525-535
https://doi.org/10.70917/ijcisim-2026-4885
Article

Design and Development of a Real-Time Indian Sign Language Conversion System to Text and Speech Using ISLGesture50 Dataset

Mr. Chirag R. Prajapati[1], Prof.(Dr.) Atul M Gonsai[2]

[1]Research Scholar, Saurashtra University, Rajkot, Gujarat.
Email: chirag.sdjic@gmail.com
[2]Professor, Computer Science Dept, Saurashtra University, Rajkot, Gujarat
Email: atul.gosai@gmail.com

Abstract: Indian Sign Language (ISL) is the first language for millions of deaf and mute people in India. However, the development of efficient, real, time automatic ISL interpretation systems has been hindered by numerous challenges such as signer variability, scarcity of large annotated datasets, complicated hand articulations, and time constraints for processing. In this work, we demonstrate a real, time vision, based system for transforming Indian Sign Language gestures into Hindi text and speech, which involves the creation of a new dataset called ISLGesture50.Our system is a combination of hand and gesture detection using YOLOv8, hand landmark extraction with MediaPipe, and a temporal consistency, based decision algorithm which recognizes the gestures that are stable across multiple frames and then appends them into the sentence form. This approach drastically lowers the number of false positives in live video situations. The identified gesture string is converted into Hindi linguistic tokens, followed by automatic punctuation and text, to, speech conversion.The main idea in this paper is a low, latency, isolated, gesture recognition method, which contrasts with typical sequence, based deep learning methods that need large continuous datasets. Therefore, the proposed method can be used in real, time on devices with limited computing power. Experimental setups and comparisons with various sign language recognition methods reveal that the integration of object detection with landmark, based features and temporal stability logic enhances the performance and user, friendliness of the system. This research provides a new real, time ISL translation system, a task, specific dataset, and an algorithmic structure that can be used in the next continuous sign language research.
Keywords: Indian Sign Language, Real-time Sign Language Recognition, YOLOv8, MediaPipe Hands, ISLGesture50, Gesture-to-Speech, Assistive Technology

1. Introduction

Communication barriers experienced by the Deaf and Hard, of, Hearing (DHH) community have a severe impact on their access to education, healthcare, and public services. Sign languages are the natural languages for these communities;

nevertheless, the lack of automated sign, to, speech converters limits communication with non, signers. Indian Sign Language (ISL) is different from American Sign Language (ASL) in that it does not have standardized large, scale datasets and computational resources, which makes automated recognition more difficult [1], [2].

At first, sign language recognition (SLR) systems were sensor, based, using data gloves and motion sensors to get accurate results, but they were also intrusive and not practical for everyday use [3]. Now, with the progress made in computer vision and deep learning, vision, based SLR systems have become the preferred choice due to their non, contact nature and scalability [4]. The latest research suggests that convolutional neural networks (CNNs), object detectors, and pose estimation frameworks can be used to achieve real, time performance for isolated and continuous sign recognition [5], [6].

Nevertheless, real, time ISL recognition is still difficult due to the following reasons:

There is a lack of large, publicly available ISL datasets. Complicated hand shapes and occlusions. High rates of false, positive continuous video streams. The need to keep the delay to a minimum for live conversations.

Most of the existing ISL systems are either concerned with offline recognition or require large continuous datasets and sequence models like LSTM or Transformers [7], [8]. Although these models are powerful, they are computationally expensive and not suitable for lightweight real, time applications. This paper presents a hybrid detection, landmark, temporal consistency framework for real, time ISL recognition with a small dataset (ISLGesture50) to overcome these issues. The system does not rely on long temporal models only but also includes a frame, level stability verification algorithm, in which a sign is added to the output sentence only after it has been recognized consistently in several consecutive frames. This concept is derived from human perceptual consistency and has been demonstrated to bring about real, time reliability in gesture, based systems [9], [10]. The goals of this study are: To create a real, time ISL gesture recognition system utilizing object detection and hand landmarks to present ISLGesture50, a dataset specifically made for isolated ISL gestures to come up with a new temporal stability, based sentence construction algorithm to transform the recognized gestures into Hindi text and speech output.

2. Related Work

2.1 Vision-Based Sign Language Recognition

Vision, based SLR methods may be broadly divided into isolated sign recognition and continuous sign language recognition (CSLR). Isolated SLR is primarily concerned with identifying single signs, whereas CSLR attempts to decode continuous gesture streams with temporal dependencies [11].

Initial vision, based systems relied on handcrafted features like HOG, SIFT, and optical flow merged with traditional classifiers [12]. While these methods were efficient in limited scenarios, they had problems with changes in background and variation of signers.

2.2 Deep Learning Approaches for SLR

Deep learning adoption has been a major factor in making the SLR system more accurate. Convolutional Neural Network (CNN), based models were able to directly extract spatial features from the images, whereas Long Short, Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks were used to model temporal dependencies in the data [13]. Currently, researchers are using Transformer, based architectures for Continuous Sign Language Recognition (CSLR), thus gaining the best performances on large datasets like WLASL and PHOENIX, 2014 [14], [15].

However, these models demand:

Huge annotated video datasets Expensive computational resources Large inference time Such requirements make these models unsuitable for use in real, time Indian Sign Language (ISL) scenarios where only small datasets are available [16].

2.3 Object Detection for Sign Language Recognition

Localizing networks like YOLO, SSD, and Faster R, CNN have been the main sources of ideas to the hand and the gesture detection unit. Especially engineering based on YOLO is highly recommended for a real, time system because of their single, stage detection pipeline [17].

Nowadays, the research works indicate that YOLOv8 has better speed, accuracy trade, offs than the previous versions of YOLO and thus it is very well suited for gesture and sign recognition tasks [18], [19]. The scientists have found it feasible to implement YOLOv8 for the recognition of the sign alphabet in real, time and isolated gesture detection, resulting in good outcomes [20].

2.4 Hand Landmark-Based Recognition

Extraction of hand landmarks aims at providing the structural changes for the finger bones and the hand pose. MediaPipe Hands is a minimal version and still a very fast one, a solution that can detect 21 3D hand landmarks even in real, time with very high precision [1]. The authors who experimented in this domain have found it beneficial to merge landmark features with CNN outputs to get stronger features which are less dependent on easy backgrounds and lighting conditions [21], [22].

The use of language based on landmarks is linguistically the most reasonable model of the human language and here the finger joints are the major players of the signing [23].

2.5 Indian Sign Language Recognition

ISL recognition research is still evolving. The existing works mainly emphasize limited vocabularies, static gestures, or sensor, based systems [24]. Datasets such as ISL, CSLTR and similar ones are considered substantial contributions to the field; however, they are still quite small in scale [25]. Various experiments on speech, to, ISL and ISL, to, text conversions have been conducted, but the availability of real, time gesture, to, speech pipelines is minimal [26], [27].

This study takes the initiative from these studies by integrating detection, landmarks, and temporal consistency in a single real, time framework for ISL.

3. ISLGesture50 Dataset

3.1 Dataset Motivation

The lack of task, specific, well, annotated datasets has been one of the major factors that limited research in ISL. ISLGesture50 was developed to facilitate the recognition of isolated gestures in real, time scenarios, with an emphasis on the most frequent communication signs in daily life.

3.2 Dataset Composition

The ISLGesture50 data set is made up of:

- 50 different ISL gestures
- 20 pictures for each gesture class
- Total number of pictures: 1000
- Type of images: RGB
- Annotation format: PASCAL VOC bounding boxes

Additionally, the dataset was carefully constructed by collecting gesture samples from 10 different individuals, introducing variations in hand shape, size, orientation, and signing style. This diversity enhances the robustness and generalization capability of the trained models.

The dataset contains gesture signs like hello, help, eat, go, home, good, safe, and other words that are frequently used in conversational ISL. The hand regions in the images were enclosed with bounding boxes that were annotated by using LabelImg to ensure the compatibility of the annotated data with object detection models [28].

3.3 Dataset Limitations

Though ISLGesture50 can be used for algorithm validation and prototyping, it also has some limitations such as:

- Single or few signers
- Limited background diversity
- Isolated gesture focus

These limitations reflect that the dataset is at an early stage [25], [29].

4. Proposed System Architecture

The proposed system is the real, time, vision, based pipeline for the Indian Sign Language (ISL) interpretation, which recognizes the live hand gestures and converts them into Hindi text and speech. The system design is an architecture of low response time, noise immunity, and practical deployability that can be used in real, world scenarios of assistive communication.

The entire architectural structure comprises the following eight major modules:

1. Video acquisition 2. Frame preprocessing 3. Hand and gesture detection 4. Hand landmark extraction 5. Feature fusion and classification 6. Temporal consistency, based decision logic 7. Hindi text generation and punctuation 8. Text, to, speech synthesis

The details of each module along with the justification based on the prior work are given below.

4.1 Video Acquisition and Frame Preprocessing

The system gets live video input by a regular RGB camera (webcam or mobile camera). Each frame is resized and normalized to the detection network input requirements. Frame normalization stabilizes the model under different lighting conditions, which are the main source of difficulties in vision, based sign recognition systems [30].

In contrast to offline SLR systems that work on pre, recorded videos, the proposed architecture is a real, time one, which sequentially processes frames from the video stream and thus, ensures that the delay between the performance of the gesture and the system response is minimal.

4.2 Hand and Gesture Detection Using YOLOv8

Hand and gesture localizations are carried out with YOLOv8, which is a very efficient single, stage object detection network that is both fast and accurate. YOLO, based detectors have been very popular in the field of gesture and sign language recognition owing to their capability to handle frames at very high frame rates while still maintaining reasonable precision [17], [18].

YOLOv8 builds on the earlier versions of YOLO by: Enhanced backbone architecture, improved feature pyramid networks, efficient anchor, and free detection.

Several recent papers show that YOLOv8 outperforms YOLOv5 and SSD, based methods in real, time gesture recognition tasks [19], [20]. In our system, YOLOv8 is fine, tuned on the ISLGesture50 dataset to localize hand regions for isolated ISL gestures.

The detector delivers:

- Bounding box coordinates for the detected hands
- Gesture class probabilities
- Detection confidence scores

The use of YOLOv8 makes it possible to accurately locate the hand parts thereby cutting down the background noise and facilitating the subsequent feature extraction.

4.3 Hand Landmark Extraction Using MediaPipe Hands

Then, from hand detection, the region of interest (ROI) is sent to MediaPipe Hands, resulting in the identification of 21 three-dimensional hand points that define finger joints and palm geometry.

MediaPipe Hands has already demonstrated the ability to perform accurate and efficient hand pose estimation in real-time tasks and has found its usages in sign

language recognition systems [1], [21]. Landmark-based hand modeling has proved to be very effective in sign languages as it captures the detailed finger articulation, which is important to distinguish similar gestures [22], [23].

Some of the landmarks that were extracted are

• Fingert

• Joint angles

• Relative Spatial Relationships

These features are normalized with regard to the bounding box to make them scale and translation invariant.

4.4 Feature Fusion and Gesture Classification

In order to enhance robustness on recognition, the proposed system uses a technique for fusing features, namely:

• Crop based on hand characteristics

• Pose-based features extracted from hand landmarks

This combination approach is justified by recent research work that reports the fusion of spatial and skeletal features contributes to an increase in sign recognition accuracy as compared to the existing approaches that consider individual modalities only [32], [31].

The landmark vector (with 63 dimensions: 21 landmarks and 3 coordinates) is flattened and passed through a light-weight multilayer perceptron (MLP) to acquire a compact pose embedding. Simultaneously, the cropped image of the hand region is passed through a shallow CNN to acquire appearance features.

The fused feature vector is then classified into one of the 50 ISL gesture classes by a fully connected neural network.

This modularity enables ease of working:

- Landmark-only classification in low-resource devices
- Fusion-based classification for better accuracy when possible
- Category theory

5. Novel Temporal Consistency-Based Recognition Algorithm

5.1 Motivation

One of the most important challenges in real-time sign language recognition is the instability of the predictions. This results in the following factors fluctuating in the frame-by-frame predictions of the system:

• Motion blur

• Partial oc

• Momentary hand pose transitions

Many traditional systems have used recurrent neural networks (LSTM or GRU) or Transformers to capture the temporal dependency [14], [15]. Though these are effective, they have the same drawbacks of requiring larger amounts of data and increased execution complexity, which makes them inappropriate for the ISL dataset.

In order to overcome such a limitation, the proposed system proposes a decision algorithm based on the concept of temporal consistency, inspired by the stability filtering techniques used in real-time gesture interfaces, as described in [33], [34].

5.2 Algorithm Concept

Rather than simply appending all predicted gestures to the sentence, it uses:

• Constantly checks model-predicted gesture labels for each frame

• Appends a gesture only if the same label is found repeatedly for N frames.

In this case, N will be observed to have values of 5 or 6 frames, being optimal within our framework to balance response timing with reliability.

Such approaches to temporal smoothing have also been demonstrated to lower false positives in real-time human-computer interfaces [35].

5.3 Algorithm Description

- Each video frame is treated separately for detection and classification.
- The predicted gesture label is stored in a sliding buffer.
- if the same gesture label has appeared for N consecutive frames:
- The gesture is stable in nature
- It is added to the sentence buffer
- 4. Duplicate gesture entries are avoided by a cooldown system.
- A pause or termination signal causes finalization of a sentence.
- The sentence is then transferred to the language module for punctuation and speech synthesis.

As a result, there is no need to build complex models of time anymore.

5.4 Advantages of the Proposed Algorithm

The algorithm based on temporal consistency has several advantages:

• Computational cost lower than in LSTM/Transformer Architectures models

• Dataset efficiency, for smaller sized datasets such as ISLGesture50

• Minimized false positives in live video feeds

• Implementation and interpretation facilitation

Such light temporal verification techniques have recently been recommended for real-time sign language applications with limited data [36], [37].

6. Hindi Text Generation and Speech Synthesis

6.1 Gesture-to-Hindi Text Mapping

Each recognized gesture class is mapped to a Hindi lexical token via a predefined dictionary. This follows previous ISL-to-text translation systems, which have used rule-based lexical mapping for recognizing isolated signs [26], [27].

Unlike in English-based systems, Hindi sentence construction requires a careful handling of:

- SOV - Subject–Object–Verb structure
- Postpositions instead of prepositions
- Gender and tense agreement

To handle this, a lightweight rule-based grammar module rearranges gesture tokens into grammatically coherent Hindi sentences.

6.2 Text-to-Speech (TTS) Synthesis

The final text in the Hindi language is synthesized using a TTS engine supporting the Hindi language. Neural TTS has been effectively used in assistive communication devices, and it can produce natural-sounding speech [39].

It also enables the interaction of both deaf and hearing people, thereby achieving accessibility's main purpose.

7. System Implementation Summary

The total system consists of the following implementations:

• Python for Backend Processing

• YOLOv8

• Using MediaPipe to extract landmarks

• PyTorch for model learning and evaluation.

Modular architecture enables upgrading of the individual elements, thereby facilitating scaling and extensibility.

8. Experimental Setup and Evaluation Methodology

8.1 Experimental Environment

All experiments were done on a standard computer environment for the purpose of ensuring that the developed system will still be deployable on such devices. The test environment included the use of a Python-based deep learning library PyTorch, OpenCV for video processing, as well as MediaPipe for landmark detection. The training and inference were done on a mid-range GPU, although it is possible to perform real-time inference on CPUs since it is a lightweight system.

Video streams were recorded at a rate that was equivalent to 25–30 frames per second, which corresponds to the rate supported by webcam and mobile camera technology commonly used in assistive technology applications [40].

8.2 Dataset Split Strategy

Because ISLGesture50 has a small size, a stratified split was used to give equal emphasis to all gesture classes:

Training set: 70%

Validation set: 15%

Testing set: 15%

In fact, such two-fold approach aligns well with guidelines suggested for small-scale image-based vision tasks such as gesture recognition research [41]. In order to avoid

bias and overfitting, image augmentation tools such as flipping, intensity adjustments, and small rotations have been applied during training phases as suggested in previous ISL research work and other studies related to hand gestures recognition [42].

8.3 Evaluation Metrices

Both frame and sentence-level measures have been used to analyze the system to obtain the recognition and usability aspects.

• Classification accuracy

• Precision, Recall, F1

• Mean Average Precision for detection

These parameters are common in object detection and recognition of sign language in the literature [17], [18].

• Word Recognition Accuracy (WRA

• Sentence Completion Accuracy

• False Positive Insertion Rate

• Average Recognition Latency

Sentence level evaluation becomes an important requirement in real-time sign-to-speech translation systems because accuracy at the frame level does not necessarily indicate how well it works [43].

9. Results and Performance Analysis

9.1 Detection and Classification Performance

The hand detector using YOLOv8 worked effectively in locating hands in different lighting environments, which corresponds with other studies involving YOLOv8 in gesture recognition tasks [19], [20]. The addition of the landmark detection using MediaPipe increased the resistance to classifications involving slight movements in the fingers.

Comparative analysis demonstrated that:

- Landmark-only classification performed well for static gestures
- Appearance-only classification struggled under background variation

Fusion-based classification showed the best accuracy, thus confirming the multi-modal approach decision, as verified in the literature [31], [32]

9.2 Impact of Temporal Consistency Algorithm

The temporal consistency, based decision algorithm was decisive in stabilizing real, time predictions. The system was prone to inserting wrong gesture tokens quite often due to transient misclassifications if temporal filtering was not applied. False positives were brought down to a negligible level with the imposition of a repetition threshold of 56 consecutive frames while the latency remained in the acceptable range.

Corresponding temporal confirmation methods have been reported to bring about stability in real, time gesture interfaces and sign recognition systems without the need for recurrent architectures [33], [35]. Consequently, the proposed approach is able to strike a compromise between accuracy and computational efficiency.

9.3 Sentence-Level Output Quality

Sentence, level evaluation showed that the proposed system is capable of generating coherent Hindi sentences with better readability and intelligibility as a result of automatic punctuation. The use of pause, based sentence termination also contributed to output naturalness to a great extent.

ISL, to, text systems in the past have been mostly incoherent at the sentence level due to direct token concatenation [26], [27]. The present works incorporation of grammar, aware mapping and punctuation is a significant step forward.

10. Comparative Analysis with Existing Methods

Qualitative comparison of the proposed system with the existing ISL recognition methods unravels the following advantages:

Feature	Existing ISL Systems	Proposed System
Real-time capability	Limited	Yes
Dataset size dependency	Large datasets required	Works with small dataset
Temporal modeling	LSTM/Transformer	Lightweight stability logic
Hindi speech output	Rare	Integrated
Computational cost	High	Low

10.1 Comparative analysis

Most of the previous systems have been targeting offline or continuous recognition by means of deep temporal models [14], [15]. Despite their effectiveness, such models are impractical for early, stage ISL datasets and real, time assistive applications. The method proposed here thus bridges this gap by providing a deployable solution that is in line with the constraints of the real world.

Research Contribution and Novelty

The major contributions of this work are:

1. ISLGesture50 Dataset: A carefully selected and annotated dataset of isolated Indian Sign Language gestures, created with the aim of conducting research on real, time recognition.

2. Hybrid DetectionLandmark Framework: Robust ISL recognition through the integration of YOLOv8, based detection with MediaPipe landmark extraction.

3. Novel Temporal Consistency Algorithm: A small, easily understandable algorithm that only goes back to stable multi, frame recognition to add the gestures to the sentences thus, it achieves the reduction of the false positives without the use of heavy temporal models.

4. End, to, End Gesture, to, Speech Pipeline: A full system that is able to convert ISL gestures into Hindi text as well as into natural speech and do it in real time.

In contrast to the existing works, where detection, recognition, and speech synthesis are treated as separate and independent processes, this work proposes a single, real, time ISL interpretation system that can be deployed in practice and is open for further development.

Conclusion and Future scope

This paper details a real, time vision, based system that translates Indian Sign Language gestures into Hindi text and speech. The system utilizes the ISLGesture50 dataset. Through the integration of YOLOv8 detection, MediaPipe hand landmarks, and a novel temporal consistency, based decision algorithm, the system is able to deliver robust real, time performance without the need for computationally expensive sequence models. The proposed approach serves as a proof, of, concept that fully functional, deployable ISL interpretation systems can be created even with limited datasets by harnessing state, of, the, art detection architectures and sophisticated decision logic. This research, thereby, acts as a significant contribution to the library of assistive technology research and paves a solid pathway for the upcoming innovations in Indian Sign Language recognition.

Possible future research directions are: Expanding ISLGesture50 with multi, signer and multi, environment data. Continuous sign language recognition support using hybrid temporal models. Using transformer, based language models for Hindi grammar correction. Implementation on mobile and embedded devices. User studies with Deaf community participants. Improvements in these areas will be a step closer to fully automated, real, world ISL interpretation systems.

REFERENCES

1. F. Zhang, V. Bazarevsky, A. Vakunov, A. Tkachenka, G. Sung, C. L. Chang, and M. Grundmann, “MediaPipe Hands: On-device Real-time Hand Tracking,” arXiv preprint arXiv:2006.10214, 2020.
2. S. Alyami, A. Alqarni, and H. Alhichri, “A comprehensive survey on continuous sign language recognition: Progress, challenges, and future directions,” IEEE Access, vol. 12, pp. 21456–21489, 2024, doi: 10.1109/ACCESS.2024.3354127.
3. T. Starner and A. Pentland, “Real-time American Sign Language recognition from video using hidden Markov models,” in Proc. IEEE Int. Symp. Computer Vision, Coral Gables, FL, USA, 1995, pp. 265–270.
4. S. Mitra and T. Acharya, “Gesture recognition: A survey,” IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 37, no. 3, pp. 311–324, May 2007.
5. Z. Cao, G. Hidalgo Martinez, T. Simon, S.-E. Wei, and Y. Sheikh, “OpenPose: Realtime multi-person 2D pose estimation using part affinity fields,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 7291–7299.
6. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You Only Look Once: Unified, real-time object detection,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 779–788.
7. O. Koller, H. Ney, and R. Bowden, “Deep learning of mouth shapes for sign language,” in Proc. IEEE Int. Conf. Computer Vision Workshops, Santiago, Chile, 2015, pp. 85–91.
8. N. C. Camgoz, S. Hadfield, O. Koller, and R. Bowden, “Neural sign language translation,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 2018, pp. 7784–7793.
9. J. Liu, L. Wang, M. Tan, and G. Li, “Real-time hand gesture recognition using finger segmentation,” ACM Transactions on Interactive Intelligent Systems, vol. 9, no. 4, pp. 1–25, 2019.
10. P. Molchanov, S. Gupta, K. Kim, and J. Kautz, “Hand gesture recognition with 3D convolutional neural networks,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops, Boston, MA, USA, 2015, pp. 1–7.
11. S. C. W. Ong and S. Ranganath, “Automatic sign language analysis: A survey and the future beyond lexical meaning,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 6, pp. 873–891, Jun. 2005.
12. W. T. Freeman and M. Roth, “Orientation histograms for hand gesture recognition,” in Int. Workshop Automatic Face and Gesture Recognition, Zurich, Switzerland, 1995, pp. 296–301.
13. L. Pigou, S. Dieleman, P.-J. Kindermans, and B. Schrauwen, “Sign language recognition using convolutional neural networks,” in Proc. European Conf. Computer Vision Workshops, Zurich, Switzerland, 2014, pp. 572–578.
14. N. C. Camgoz, O. Koller, S. Hadfield, and R. Bowden, “Sign language transformers: Joint end-to-end sign language recognition and translation,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 10023–10033.
15. H. Zhou, W. Zhou, Y. Zhou, and H. Li, “Spatial–temporal multi-cue network for continuous sign language recognition,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 11, pp. 7684–7698, Nov. 2022.
16. R. Rastgoo, K. Kiani, and S. Escalera, “Sign language recognition: A deep survey,” Expert Systems with Applications, vol. 164, p. 113794, Feb. 2021.
17. A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, “YOLOv4: Optimal speed and accuracy of object detection,” arXiv preprint arXiv:2004.10934, 2020.
18. G. Jocher, A. Chaurasia, and J. Qiu, “YOLOv8: Ultralytics object detection framework,” Ultralytics, 2023. [Online]. Available: https://github.com/ultralytics/ultralytics
19. W. Jia, X. Zhang, and L. Wang, “SLR-YOLO: An improved YOLOv8 network for real-time sign language recognition,” Journal of Information Fusion, vol. 98, pp. 101889, 2024.
20. M. Hasan, S. Rahman, and T. Hossain, “Real-time sign language recognition using YOLOv8 and hand landmark fusion,” IEEE Access, vol. 13, pp. 45678–45692, 2025.

21. V. Bazarevsky et al., “BlazePalm: Real-time palm detection,” arXiv preprint arXiv:2006.10214, 2020.
22. T. Simon, H. Joo, I. Matthews, and Y. Sheikh, “Hand keypoint detection in single images using multiview bootstrapping,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 1145–1153.
23. J. Escobedo, D. Hernandez, and M. Gonzalez, “Hand pose-based sign language recognition using skeletal features,” Sensors, vol. 21, no. 4, p. 1356, 2021.
24. J. Rekha, J. Bhattacharya, and S. Majumder, “Shape, texture and local movement hand gesture features for Indian sign language recognition,” Int. J. Computer Science and Information Technology, vol. 5, no. 3, pp. 120–128, 2013.
25. S. Kumar, A. Jain, and R. Kaushik, “ISL-CSLTR: A continuous sign language translation dataset for Indian sign language,” Mendeley Data, vol. 3, 2023.
26. P. Vij and P. Kumar, “Mapping Hindi text to Indian sign language using rule-based approach,” in Proc. ACM Int. Conf. Assistive Technologies, New Delhi, India, 2016, pp. 112–118.
27. S. Sonawane and H. Shah, “Speech to Indian sign language translation system,” International Journal of Advanced Research in Computer Science, vol. 8, no. 5, pp. 156–161, 2017.
28. D. Tzutalin, “LabelImg: Graphical image annotation tool,” GitHub Repository, 2015.
29. Kaggle, “Indian Sign Language Gesture Dataset,” 2022. [Online]. Available: https://www.kaggle.com
30. R. Gonzalez and R. Woods, Digital Image Processing, 4th ed. Pearson, 2018.
31. S. Escalera et al., “Multi-modal gesture recognition using skeleton and RGB data,” Pattern Recognition Letters, vol. 131, pp. 321–328, 2020.
32. K. Simonyan and A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” in Proc. NIPS, Montreal, Canada, 2014, pp. 568–576.
33. [33] D. Wigdor and D. Wixon, Brave NUI World, Morgan Kaufmann, 2011.
34. J. Han, G. Awad, and A. Smeaton, “Gesture interaction design for real-time systems,” IEEE Computer Graphics and Applications, vol. 38, no. 4, pp. 18–29, 2018.
35. Y. Wu and T. Huang, “Vision-based gesture recognition: A review,” Gesture-Based Communication in Human-Computer Interaction, Springer, pp. 103–115, 1999.
36. S. K. Gupta and R. Jain, “Real-time hand gesture recognition using temporal smoothing,” International Journal of Computer Vision, vol. 128, pp. 1–15, 2020.
37. H. Cooper, B. Holt, and R. Bowden, “Sign language recognition,” Visual Analysis of Humans, Springer, pp. 539–562, 2011.
38. J. Yi, Y. Wang, and Z. Chen, “Automatic punctuation generation for speech-to-text systems,” IEEE Signal Processing Letters, vol. 26, no. 12, pp. 1922–1926, 2019.
39. Y. Wang et al., “Tacotron: Towards end-to-end speech synthesis,” in Proc. Interspeech, Stockholm, Sweden, 2017, pp. 4006–4010.
40. S. Prince, Computer Vision: Models, Learning, and Inference, Cambridge Univ. Press, 2012.
41. I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, MIT Press, 2016.
42. A. Shorten and T. M. Khoshgoftaar, “A survey on image data augmentation,” Journal of Big Data, vol. 6, no. 60, 2019.
43. N. Camgoz, O. Koller, and R. Bowden, “Evaluating continuous sign language recognition,” Pattern Recognition Letters, vol. 128, pp. 45–52, 2019.