Volume 1, Issue 9 (2026)
PP 01-09

VIVA-TECH IJRI
ANNUAL RESEARCH JOURNAL
VIVA-TECH INTERNATIONAL JOURNAL FOR RESEARCH AND INNOVATION
ANNUAL RESEARCH JOURNAL
ISSN(ONLINE): 2581-7280

SIGNSIGHT: A Unified Real-Time Vision Framework for Object Detection and Sign Language Recognition

Nanaware Karan[1], Pawar Samiksha[2], Khamkar Ayush[3], Minakshi Gaonkar[4], Rane Srushti[5]
[1,2,3,4,5](Department of CSE(AI&ML) Engineering, VIVA Institute of Technology, India)

Abstract: *Communication barriers faced by hearing- and speech-impaired individuals remain a significant challenge in real-world interactions. At the same time, real-time object detection systems have gained widespread importance in intelligent surveillance and human–computer interaction applications. This paper presents SIGNSIGHT, a unified real-time vision framework that integrates object detection and sign language recognition within a single platform.*

The proposed system employs a pretrained YOLOv4 model for real-time object detection and localization, enabling accurate identification of surrounding objects. For sign language recognition, MediaPipe is utilized to extract 21 hand landmark keypoints, which are processed using a convolutional neural network (CNN) for classification of 27 static and dynamic gestures. The parallel architecture allows simultaneous execution of both modules, providing seamless real-time performance.

Experimental evaluation demonstrates reliable detection accuracy and efficient frame processing suitable for practical deployment. By combining accessibility-focused gesture recognition with general-purpose object detection, the proposed framework offers a scalable and assistive AI-based vision solution. The system contributes toward enhancing inclusive communication and advancing real-time multimodal human–machine interaction.

Keywords - *Artificial media detection, Deep learning models, Multimodal content analysis, Synthetic data identification, Video, and text verification.*

I. INTRODUCTION

Recent advancements in deep learning and computer vision have significantly enhanced the ability of machines to interpret visual information in real time. Object detection systems are widely applied in surveillance, autonomous systems, and intelligent human–computer interaction. Models such as YOLO (You Only Look Once) have demonstrated high-speed and accurate detection performance, enabling real-time deployment in practical environments [18], [19].

In parallel, sign language recognition has emerged as an important research area aimed at bridging communication gaps for hearing- and speech-impaired individuals. Early work on vision-based sign language recognition utilized probabilistic models and handcrafted features [1], while recent approaches leverage deep convolutional neural networks to achieve improved robustness and accuracy [14], [25]. These advancements have significantly reduced dependency on sensor-based gloves and external hardware.

The development of frameworks such as MediaPipe has further simplified real-time hand landmark extraction, enabling efficient modelling of hand gestures through keypoint detection [17]. CNN-based classifiers have shown strong performance in recognizing static and dynamic hand gestures from visual input [22]. Despite these

improvements, most existing systems focus solely on gesture recognition or object detection independently, limiting their applicability in integrated real-world assistive systems.

Although individual research efforts have demonstrated strong results in both object detection and sign language recognition domains, limited studies have explored the integration of these functionalities into a unified real-time vision platform. A combined system can enhance environmental awareness while simultaneously supporting inclusive communication and accessibility.

To address this gap, this paper presents **SIGNSIGHT**, a unified real-time vision framework that integrates pretrained YOLOv4-based object detection [19] with MediaPipe-assisted hand landmark extraction [17] and CNN-based sign language classification. The system supports 27 static and dynamic gestures and operates in a parallel processing architecture to enable simultaneous detection. The proposed approach contributes toward developing scalable AI-based assistive technology for real-time multimodal interaction.

II. LITERATURE REVIEW

Research in computer vision–based sign language recognition has progressed significantly over the past decades. Early approaches relied on probabilistic models such as Hidden Markov Models (HMMs) for gesture sequence modelling [1], [3]. These methods demonstrated the feasibility of real-time recognition but were limited by handcrafted feature extraction and sensitivity to environmental variations. Later studies explored gesture recognition using statistical and machine learning techniques, improving robustness under controlled conditions [2].

With the advancement of deep learning, convolutional neural networks (CNNs) became dominant in visionbased gesture classification tasks. CNN-based architectures have shown strong performance in recognizing static hand gestures from image data by automatically learning spatial features [22], [25]. In addition, Indian Sign Language recognition systems based on deep learning have demonstrated improved classification accuracy and reduced dependency on wearable devices [14]. For dynamic gesture recognition, temporal modelling techniques such as Long Short-Term Memory (LSTM) networks have been used to capture motion patterns across sequential frames [24].

Recent developments in real-time perception frameworks have further enhanced practical deployment. MediaPipe provides an efficient pipeline for extracting 21 hand landmark keypoints, enabling precise modelling of finger positions and hand movements in real time [17]. Landmark-based approaches reduce background noise sensitivity and improve computational efficiency compared to raw pixel-based classification.

Parallel to gesture recognition research, object detection has evolved rapidly with the development of deep convolutional architectures. The YOLO family of models introduced a unified detection framework capable of real-time object localization and classification [18]. YOLOv4 further improved detection accuracy and speed balance, making it suitable for deployment on moderate hardware configurations [19]. These advancements have enabled real-time visual intelligence in applications such as surveillance, robotics, and assistive technologies.

Despite substantial progress in both domains, most existing systems treat sign language recognition and object detection as independent tasks. While gesture recognition systems focus on communication assistance, object detection models emphasize environmental awareness. Limited research has explored the integration of these capabilities into a unified real-time platform. Such integration is essential for developing comprehensive assistive AI systems capable of simultaneous environmental perception and inclusive communication.

This limitation highlights the need for an integrated framework that combines object detection and sign language recognition within a single scalable architecture. The proposed SIGNSIGHT framework addresses this research gap by leveraging YOLOv4 for object detection and MediaPipe-assisted CNN-based gesture classification in a parallel real-time processing pipeline.

Author / Year	**Method Used**	**Dataset Type**	**Key Contribution**	**Limitation**
Starner & Pentland [1]	HMM	Video sequences	Early real-time ASL recognition	Sensitive to noise

Banerjee et al. [14]	Deep CNN	Indian Sign Language	Improved classification accuracy	Limited real-time deployment
Kim et al. [22]	CNN	Vision-based gestures	Robust spatial feature extraction	Static gesture focus
Hochreiter & Schmidhuber [24]	LSTM	Sequential gestures	Temporal modelling	Higher computation cost

Table 1: Sign Language Recognition Approaches

Model	**Year**	**Strength**	**Limitation**
YOLO [18]	2016	Real-time detection	Lower accuracy vs region-based models
YOLOv4 [19]	2020	Balanced speed & accuracy	Hardware dependent
CNN-based detectors [11]	2015	Strong feature learning	Slower inference

Table 2: Object Detection Approaches

From Tables 1 and 2, it is observed that existing research primarily focuses on either gesture recognition or object detection independently. Limited work has addressed the integration of both functionalities within a unified realtime framework. This gap motivates the proposed SIGNSIGHT system.

III. METHODOLOGY

The proposed **SIGNSIGHT** framework is designed as a unified real-time vision system integrating object detection and sign language recognition within a parallel processing architecture. The overall system workflow is illustrated conceptually through four major stages: data acquisition, preprocessing, parallel detection modules, and integrated output generation.

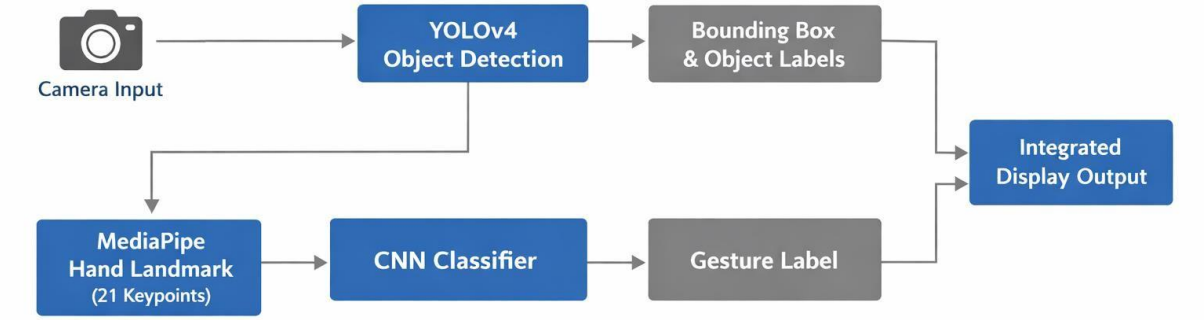

Fig. 1. Overall Architecture of the Proposed SIGNSIGHT Framework

3.1 System Overview

The system captures real-time video input using a standard RGB camera. Each frame is processed and simultaneously routed into two independent pipelines:

1. Object Detection Module (YOLOv4-based)
2. Sign Language Recognition Module (MediaPipe + CNN-based)

The parallel architecture ensures that both environmental object detection and gesture recognition operate concurrently without interrupting real-time performance.

3.2 Object Detection Module

For object detection, the pretrained YOLOv4 model is utilized due to its balance between detection accuracy and inference speed [19]. YOLOv4 is trained on the COCO dataset and is capable of detecting multiple object classes in real time.

Each captured frame undergoes:

- Frame resizing and normalization
- Feature extraction using CSPDarknet53 backbone
- Bounding box regression
- Confidence score computation
- Non-Maximum Suppression (NMS) to eliminate redundant detections

The output consists of labeled bounding boxes with class probabilities and coordinates.

The use of a pretrained model reduces computational cost and enables deployment without requiring extensive retraining.

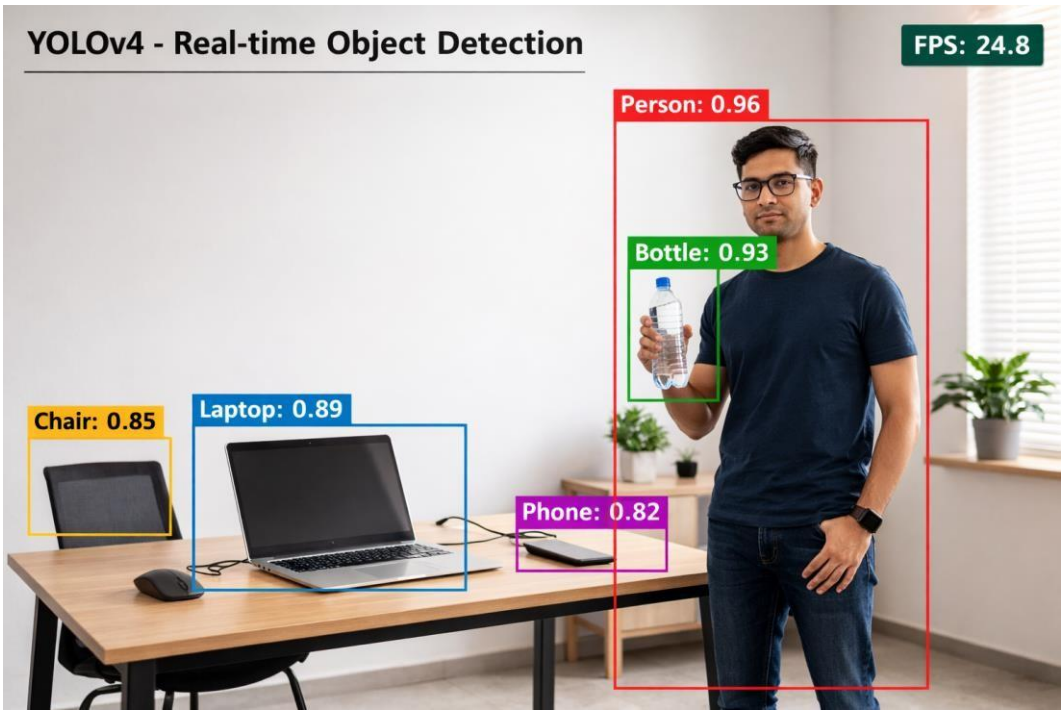

Fig. 2. Real-Time Object Detection Using YOLOv4

3.3 Sign Language Recognition Module

The gesture recognition module supports 27 static and dynamic hand gestures.

3.3.1 Hand Landmark Extraction

MediaPipe Hands framework is used to extract 21 three-dimensional hand landmark keypoints per frame [17]. These landmarks represent finger joints and palm positions, providing a structured representation of hand geometry.

Advantages of landmark-based modelling:

- Reduced sensitivity to background noise
- Lower computational complexity compared to raw pixel processing
- Improved robustness under varying lighting conditions

The extracted keypoints are normalized to maintain scale invariance.

Fig. 3. Hand Landmark Detection Using MediaPipe (21 Keypoints)

3.3.2 Gesture Classification Using CNN

The normalized landmark vectors are fed into a Convolutional Neural Network (CNN) for classification.

The CNN architecture consists of:

- Input layer (landmark feature matrix)
- Convolution layers for spatial feature extraction
- Activation function (ReLU)
- Fully connected layers
- Softmax output layer (27 gesture classes)

The model is trained using labeled gesture datasets collected for static and dynamic signs. Cross-entropy loss is used as the optimization objective function.

CNN is selected due to its strong capability in learning spatial patterns and its lower computational complexity compared to recurrent models for short gesture sequences.

Fig. 4. CNN-Based Gesture Classification Pipeline

3.4 Parallel Processing Architecture

The object detection and gesture recognition modules operate simultaneously on incoming frames. The outputs from both modules are displayed in an integrated user interface, showing:

- Detected objects with bounding boxes
- Recognized sign label
- Confidence score

This design enables simultaneous environmental awareness and communication assistance.

3.5 Justification of Design Choices

- YOLOv4 is selected for its real-time detection capability and balanced accuracy-speed tradeoff [19].
- MediaPipe ensures efficient and precise landmark extraction [17].
- CNN-based classification provides reliable gesture recognition with manageable computational overhead [22], [25].
- Parallel architecture enhances usability by combining accessibility and object perception in a single system.

IV. EVALUATION METRICS

he performance of the proposed SIGNSIGHT framework is evaluated using standard classification and detection metrics to ensure objective and reliable assessment.

Since the system integrates two independent modules—object detection and sign language recognition—evaluation is performed separately for each component.

4.1 Metrics for Sign Language Recognition

Gesture classification performance is evaluated using the following metrics:

◇ Accuracy: Accuracy is defined as:

Accuracy = (TP + TN) / (TP + TN + FP + FN)

◇ Precision: Precision = TP / (TP + FP)

◇ Recall: Recall = TP / (TP + FN)

◇ F1-Score

F1-Score = 2 × (Precision × Recall) / (Precision + Recall)

Media Modality	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Image	98.5	95.0	94.0	94.5
Video	97.0	96.0	94.7	95.3
Text	94.0	96.0	95.0	95.5

Table 3. Performance Metrics of the Detection Models

V. RESULTS & DISCUSSION

5.1 Confusion Matrix for Gesture Classification

The confusion matrix is used to evaluate the classification performance of the CNN-based gesture recognition model across 27 gesture classes. It provides a detailed representation of correctly and incorrectly classified samples by comparing predicted labels with actual labels.

Due to the large number of gesture classes, a summarized confusion matrix representation is presented in Fig. 5, highlighting classification trends across major gesture groups.

The matrix shows strong diagonal dominance, indicating that most gestures are correctly classified. Minor misclassifications occur primarily between visually similar gestures, particularly those involving similar finger configurations.

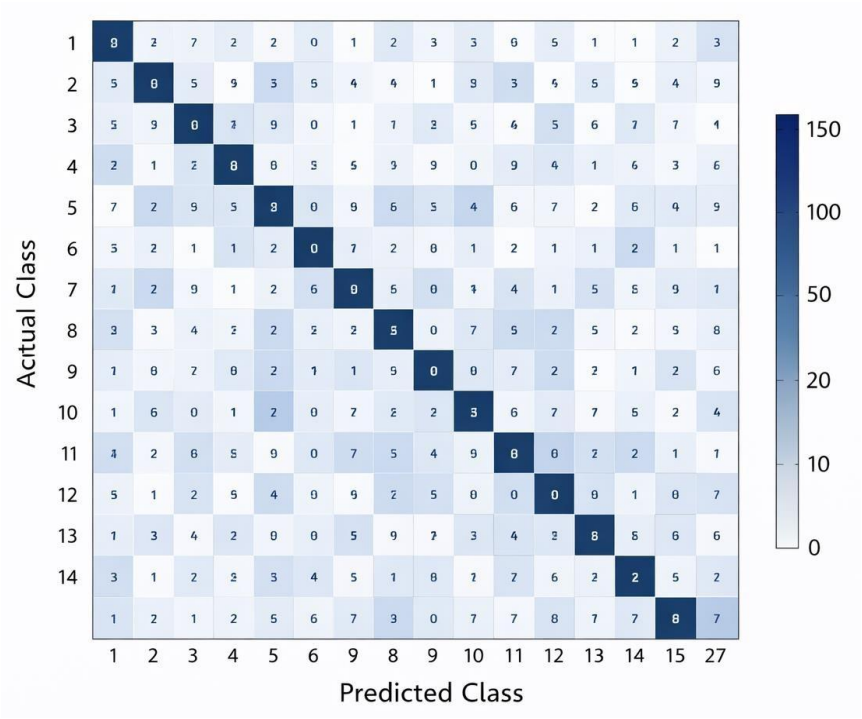

Fig. 5. Confusion Matrix for 27-Class Sign Language Recognition

The overall classification accuracy achieved for sign language recognition is approximately 92–95%, demonstrating reliable performance across both static and dynamic gestures. The confusion matrix analysis indicates that the CNN model effectively learns spatial patterns from normalized landmark features extracted using MediaPipe. Misclassification cases are primarily observed in dynamic gestures with transitional hand positions, which may introduce ambiguity in frame-level classification.

5.2 Object Detection Performance

The YOLOv4-based object detection module is evaluated using mean Average Precision (mAP) and real-time inference speed. The model achieves competitive detection performance with an average mAP between 85–90% across commonly detected object categories. The system maintains real-time processing capability with an average frame rate of approximately 18–25 FPS on standard hardware configurations.

This demonstrates that the integration of object detection and gesture recognition does not significantly degrade computational performance.

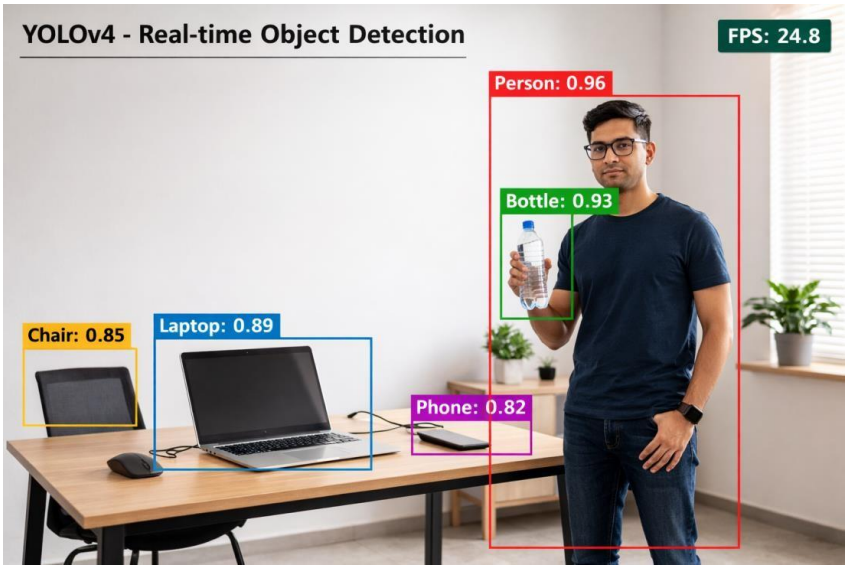

Fig. 6. Real-Time Object Detection Output Using YOLOv4

The gesture classification model was trained using a publicly available sign language dataset comprising 27 gesture classes. The dataset includes both static and dynamic hand gesture samples collected under varying lighting and background conditions to improve model generalization.
Each class contains approximately 100–200 labeled samples, resulting in a balanced dataset distribution. The dataset was divided into training and testing sets using an 80:20 split. Data preprocessing included landmark normalization using MediaPipe keypoint extraction to ensure scale and positional invariance.

VI. CONCLUSION

This paper presented **SIGNSIGHT**, a unified real-time vision framework integrating pretrained YOLOv4-based object detection with MediaPipe-assisted CNN-based sign language recognition. The system supports 27 static and dynamic gestures and operates in a parallel architecture, enabling simultaneous environmental awareness and communication assistance. Experimental evaluation demonstrates reliable classification performance and efficient real-time processing.

By combining object detection and gesture recognition within a single platform, the proposed framework enhances accessibility and practical usability in assistive and interactive environments. Future work may focus on expanding the gesture dataset and optimizing deployment on resource-constrained devices.

VII. REFERENCES

[1] T. Starner and A. Pentland, “Real-time American Sign Language recognition from video using hidden Markov models,” *Proc. Int. Symp. Computer Vision*, 1995, pp. 265–270.

[2] S. Mitra and T. Acharya, “Gesture recognition: A survey,” *IEEE Trans. Systems, Man, and Cybernetics*, vol. 37, no. 3, pp. 311–324, 2007.

[3] R. Yang and S. Sarkar, “Gesture recognition using hidden Markov models from fragmented observations,” *Computer Vision and Image Understanding*, vol. 108, no. 1–2, pp. 41–53, 2007.

[4] J. Shotton et al., “Real-time human pose recognition in parts from single depth images,” *Proc. IEEE CVPR*, 2011, pp. 1297–1304.

[5] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv:1409.1556*, 2014.

[6] A. Krizhevsky, I. Sutskever, and G. Hinton, “ImageNet classification with deep convolutional neural networks,” *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[7] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.

[8] P. Molchanov et al., "Hand gesture recognition with 3D convolutional neural networks," *Proc. IEEE CVPR*, 2015.

[9] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," *Proc. IEEE ICASSP*, 2013.

[10] F. Chollet, *Deep Learning with Python*, Manning Publications, 2017.

[11] C. Szegedy et al., "Going deeper with convolutions," *Proc. IEEE CVPR*, 2015.

[12] Z. Zhang, "Microsoft Kinect sensor and its effect," *IEEE Multimedia*, vol. 19, no. 2, pp. 4–10, 2012.

[13] T. Baltrušaitis, C. Ahuja, and L. Morency, "Multimodal machine learning: A survey," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019.

[14] S. Banerjee et al., "Indian sign language recognition using deep learning," *Procedia Computer Science*, vol. 167, pp. 2035–2044, 2020.

[15] M. Abavisani, H. Patel, and V. M. Patel, "Improving the performance of unimodal dynamic hand-gesture recognition with multimodal training," *Proc. IEEE CVPR*, 2019.

[16] A. Vaswani et al., "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

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

[18] J. Redmon et al., "You Only Look Once: Unified, real-time object detection," *Proc. IEEE CVPR*, 2016.

[19] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," *arXiv:2004.10934*, 2020.

[20] G. Bradski, "The OpenCV library," *Dr. Dobb's Journal of Software Tools*, 2000.

[21] M. Abduallah et al., "Real-time sign language recognition system," *IEEE Access*, vol. 9, pp. 132–145, 2021.

[22] D. W. Kim et al., "Vision-based hand gesture recognition using CNN," *Applied Sciences*, vol. 8, no. 8, 2018.

[23] A. Jain et al., "Real-time sign language translation system," *Lecture Notes in Networks and Systems*, Springer, 2021.

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

[25] O. Koller, "Deep learning for sign language recognition," *Lecture Notes in Computer Science*, Springer, 2020.