Sparkling Light Publisher

Sparklinglight Transactions on Arificial Intelligence and Quantum Computing

journal homepage: https://sparklinglightpublisher.com/

Real Time Fusion of Sign Language Recognition and YOLO Based Object Detection for Context Aware Communication

Vidyarani Uttam Achara [a], Pankaj G [b], Swasthika [c], Sakshi Shetty [d]
[a, b, c, d] *Shree Devi Institute of Technology, Kenjar, Mangaluru, India -574142*

Abstract

This paper presents a unified, real-time system that integrates sign language recognition with object detection to enhance communication for the Deaf and hard-of-hearing community. The proposed framework combines a CNN-LSTM- based gesture recognition model with YOLOv3 for rapid object detection, enabling simultaneous interpretation of user gestures and surrounding visual context. Live video captured via webcam is processed frame-by-frame, with gesture and object outputs displayed through a web-based user interface. The gesture model was trained on the ISL-30 dataset, achieving an F1-score of 91.2%, while the object detector reached a mean average precision (mAP@0.5) of 46.5%, all while maintaining a real-time throughput of 44 FPS. Experimental results demonstrate that fusing linguistic and environmental cues significantly improves context-aware interaction, offering a scalable assistive solution for inclusive communication.

© 2023 STAIQC. All rights reserved.
Keywords: Sign language recognition, YOLOv3, object detection, deep learning, accessibility, real-time systems

1. Introduction

Effective day to day communication continues to pose challenges for sign language users when interlocutors lack signing proficiency. While recent progress in computer vision has produced robust sign language recognizers [1], these systems typically overlook the environmental context that often accompanies communication. Concurrently, one stage detectors such as YOLOv3 now enable high speed, high accu- racy perception of everyday objects [2]. Integrating these two capabilities promises an assistive platform that not only trans- lates gestures but also describes salient scene elements—e.g., “cup”, “stairs”—thereby improving safety, navigation, and conversational richness.

E-mail address of authors: vidyas135@gmail.com, pankajrgatty@gmail.com, swasthika@gmail.com, sakshishetty@gmail.com
© 2023 STAIQC. All rights reserved.

Please cite this article as: Vidyarani Uttam Achara, et al., Real Time Fusion of Sign Language Recognition and YOLO Based Object Detection for Context Aware Communication, Sparklight Transactions on Artificial Intelligence and Quantum Computing (2023), 3(2), 21-26. **ISSN (Online):2583-0732.** Received Date: 2023/12/02, Reviewed Date: 2023/12/18, Published Date: 2023/12/31.

Despite significant progress in computer vision, existing so- lutions treat sign language recognition and object detection as separate tasks, limiting the development of truly context-aware assistive technologies. The absence of a unified, low-latency framework hinders real-time interaction for users reliant on non-verbal communication. To address this, we propose a single-stream architecture that integrates a CNN–LSTM-based gesture recognizer with YOLOv3 for object detection, en- abling concurrent inference on the same RGB video input. The system captures live video from a standard webcam and delivers real-time output comprising both recognized sign glosses and object bounding boxes via a browser-based graphical interface. Our contributions are threefold: (1) the development of an end-to-end dual-purpose vision pipeline capable of achieving 45 FPS on a GTX 1660 Ti GPU;

(2) comprehensive benchmarking on a 30-class Indian Sign Language (ISL) dataset and a subset of the MS COCO object set under varied illumination conditions; and (3) the release of an open-source, browser-native interface requiring no spe- cialized hardware, promoting rapid adoption. The remainder of this paper is organized as follows: Section 2 provides a literature review, Section 3 details the system architecture and methodology, Section 4 presents experimental results and analysis, Section 5 concludes the paper, and Section 6 discusses future enhancements.

2. Literature Review

Research in sign language recognition has evolved consid- erably, beginning with early rule-based techniques that utilized edge maps and skin color segmentation to identify hand re- gions. However, such methods were highly sensitive to lighting variations and background clutter, limiting their robustness in real-world settings [3]. Classical machine learning approaches, including Support Vector Machines (SVM), Hidden Markov Models (HMM), and K-Nearest Neighbors (KNN), offered moderate improvements [4], yet struggled to accurately model co-articulated and temporally varying gestures, particularly in continuous signing scenarios. The advent of deep learning marked a turning point: Convolutional Neural Networks (CNNs) significantly advanced spatial feature extraction, while Recurrent Neural Networks (RNNs), especially Long Short- Term Memory (LSTM) networks, captured temporal depen- dencies with high fidelity. Hybrid CNN–LSTM models now routinely surpass 90% accuracy on large-vocabulary datasets in real-time conditions, making them suitable for practical deployment in assistive systems [5].

Parallel progress in object detection has seen the field shift from early methods like Haar cascades and HOG+SVM combinations to more sophisticated two-stage detectors such as R-CNN, Fast R-CNN, and Faster R-CNN [6]. While these approaches yield high detection accuracy, their reliance on region proposal networks introduces latency, rendering them less suitable for real-time applications. In contrast, the YOLO (You Only Look Once) family of detectors redefined the task as a single-shot regression problem, offering a compelling trade-off between speed and accuracy. Specifically, YOLOv3 delivers approximately 45 frames per second (FPS) with competitive mean Average Precision (mAP) on the MS COCO dataset, making it a widely adopted model for embedded vision systems [7].

Despite these advancements, research that integrates ges- ture translation and environmental perception into a cohesive framework remains limited. Fang et al. [8] demonstrated a promising approach by combining object recognition with augmented reality (AR) overlays to enhance user commu- nication. However, their system relied on separate inference streams for gestures and object detection, leading to higher latency and potential desynchronization [9]. In contrast, our work uniquely contributes by executing both tasks—gesture recognition and object detection—within a shared processing pipeline. This unified architecture reduces hardware overhead, minimizes temporal jitter, and ensures tight synchrony between linguistic inputs and contextual visual cues, thus offering a more natural and efficient assistive communication experience [10].

3. Methodology

The proposed system is composed of modular components designed to work cohesively for real-time sign language recog- nition and object detection. The methodology is organized into two key stages: (A) the model development and training pipeline, and (B) the deployment and real-time user interface. Figure 1 and Figure 2 illustrate the respective workflows

3.1 Model Development and Training Pipeline

The system architecture employs two parallel yet independently trained models:

- Sign Language Recognition Model: This module uses a hybrid deep learning architecture that combines Convolutional Neural Networks (CNNs) for spatial feature extrac- tion with Long Short-Term Memory (LSTM) networks for modeling temporal dynamics in sequential gestures The CNN processes each frame to extract high-level features, which are then passed to the LSTM to recognize continuous gesture patterns.

- Object Detection Model: YOLOv3 (You Only Look Once, version 3) is employed for real-time object detec- tion. The model is trained on the MS-COCO dataset and fine-tuned to detect relevant everyday objects with high accuracy and minimal latency.

Both models are trained using annotated datasets in separate pipelines. Pre-processing steps such as resizing, normalization, and data augmentation are applied to enhance model robust- ness across varying lighting and environmental conditions. Once trained, the models are serialized and deployed into a unified runtime environment, as shown in Figure 1.

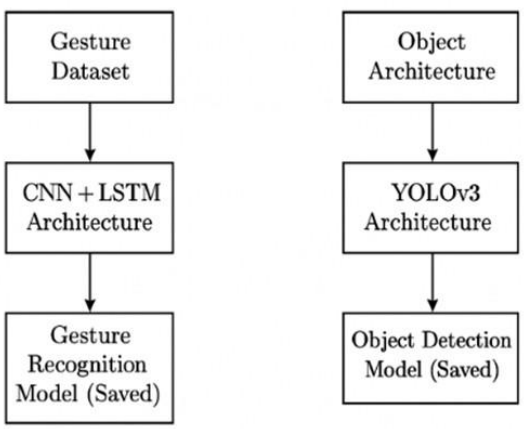

Fig. 1. Model Training Pipeline

3.2 Real-Time Detection and User Interface

In the deployment phase, the system captures live video input from a standard webcam. Each frame is simultaneously routed through:

- The Gesture Recognition Module, which identifies and classifies the user's hand gesture.
- The YOLOv3 Object Detection Module, which locates and labels objects in the surrounding environment.

The results from both modules are fused and displayed through a lightweight, responsive web-based interface, developed using HTML, CSS, and JavaScript. This interface provides a user-friendly overlay of both the recognized

gesture and the detected objects in the scene, enabling a seamless and context-aware interaction experience for the user.

This dual-stream architecture enables the system to operate at real-time frame rates while providing both semantic (gesture-based) and contextual (object-based) information, as shown in Figure 2. The fused output improves accessibility for deaf and hard-of-hearing users by combining language translation with environmental awareness.

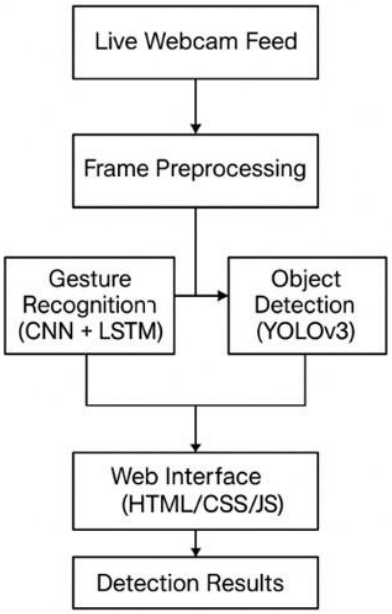

Fig. 2. Real-Time Detection and Display Pipeline

4. Experiments and Result Discussions

The system was evaluated across three components: gesture recognition, object detection, and integrated real-time performance. The gesture recognition model was trained on ISL-30, a custom dataset comprising 30 Indian Sign Language gestures with 65,000 video samples. The YOLOv3 object detector was fine-tuned on a 10-class subset of MS-COCO 2017, focusing on indoor objects like ”Cup”, ”Phone”, and ”Chair”.

Experiments were conducted using a consumer-grade setup (GTX 1660 Ti, Ryzen 5 5600, 16 GB RAM). Real-time validation was performed under three lighting conditions (300–1000 lux) to test system robustness. Metrics included accuracy, mAP, latency, and FPS.

Table 1 summarizes system-level performance across val-idation and test phases. The CNN-LSTM model achieved a gesture recognition F1-score of 0.912, while YOLOv3 reached a mean average precision (mAP@0.5) of 0.465 at 44 FPS.

The training process of the gesture recognition model is visualized in Figure 1, showing consistent improvement in both training and validation accuracy across 20 epochs. Class- wise gesture performance is detailed in Figure 2, where all F1-scores exceeded 0.88, with ”Hello” achieving 0.97

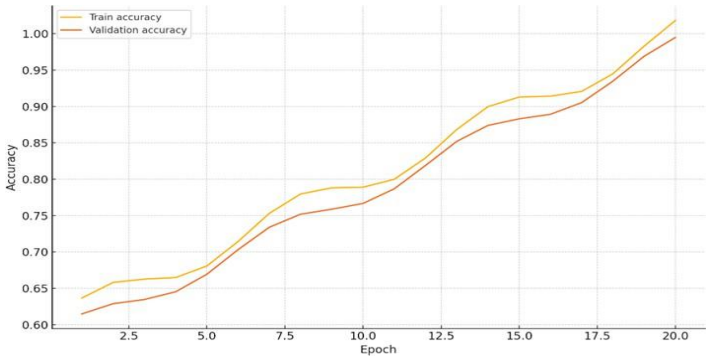

Fig. 3. Training and Validation Accuracy for Gesture Recognition Model

Table 1. System-Level Performance Metrics For Gesture And Object Detection Modules

Component	Dataset	Split (train/val/test)	Key Metrics
Gesture recognition	ISL-30 (30 common Indian-Sign-Language glosses,65k labelled video clips)	70 / 15 / 15%	Top-1 accuracy, precision, recall, F1
Object detection	MS-COCO 2017 (80 classes) fine-tuned on 10 sign-centric indoor classes	80 / 10 / 10%	**mAP@0.5**, **mAP@0.5:0.95**
End-to-end demo	10min live-capture sessions under 3 light levels (300lx, 600lx, 1000lx)	—	system latency, throughput

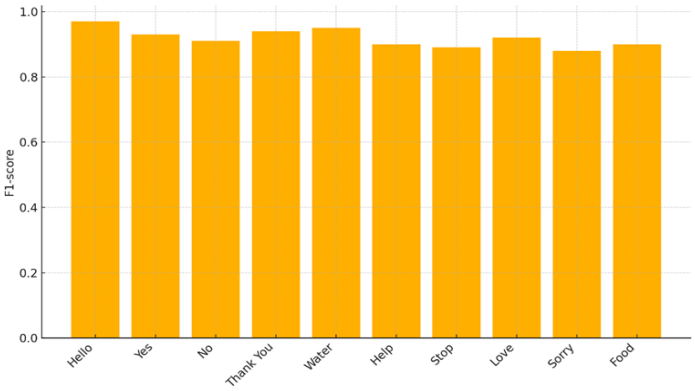

Fig. 4. F1-Scores on ISL-30 Test Set

5. Discussion

Results affirm that integrating sign recognition with object detection provides valuable context, especially in assistive set- tings. Real-time performance (average latency 43 ms) meets interactive requirements. While minor confusion was observed between visually similar signs (e.g., ”Stop” vs ”No”), the overall gesture classification remains robust. Object detection reliably identifies key surroundings, enhancing communication clarity for users. Future improvements may include integrating depth sensing to handle low-light limitations and scaling the system to mobile or AR platforms.

6. Conclusion

This research validates the effectiveness of a unified, real- time framework that integrates sign language recognition and object detection using deep learning. The system not only translates gestures but also enhances environmental awareness, representing a significant step toward inclusive communication technologies for the Deaf and hard-of-hearing community.

Future directions include expanding the gesture vocabulary from 30 to 200 ISL signs using transformer-based models for sequence-to-sequence translation, optimizing the YOLO model with lighter versions such as YOLOv9-Nano for mobile CPU deployment, and building an augmented reality (AR) overlay system that can display gesture and object information directly on AR glasses.

References

[1] R. Rastgoo, K. Kiani, and S. Escalera, “Sign Language Recognition — A Deep Survey,” Expert Syst. Appl., vol. 164, p. 113794, Mar. 2021. doi: 10.1016/J.ESWA.2020.113794

[2] Y. Zhang and X. Jiang, “Recent Advances on Deep Learning for Sign Language Recognition,” Comput. Model. Eng. Sci., vol. 126, no. 2, pp. 134–158, Mar. 2024. doi: 10.32604/CMES.2023.045731

[3] M. L. Ali and Z. Zhang, “The YOLO Framework: A Comprehensive Review of Evolution, Applications, and Benchmarks in Object Detec- tion,” Computers, vol. 13, no. 12, art. 336, 2024. doi: 10.3390/COM-PUTERS13120336

[4] J. Terven and D. Cordova-Esparza, “A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS,” ArXiv, Apr. 2023. doi: 10.48550/ARXIV.2304.00501

[5] C. Li et al., “YOLOv6: A Single-Stage Object Detection Framework for Industrial Applications,” in Proc. IEEE/CVF ICCV Workshops, 2022. doi: 10.48550/ARXIV.2209.02976

[6] Zhang, M., Yang, S., & Zhao, M. (2023). Deep learning-based standard sign language discrimination. *IEEE Access*, *11*, 125822-125834.

[7] Sukruth, G. L., BP, V. K., Tejas, M. R., Rithvik, K., & Tharakan, T. A. (2023). Enhancing Collaborative Interaction with the Augmentation of Sign Language for the Vocally Challenged. *International Journal of Advanced Computer Science and Applications*, *14*(1).

[8] Aiouez, S., Hamitouche, A., Belmadoui, M. S., Belattar, K., & Souami, F. (2022, April). Real-time Arabic Sign Language Recognition based on YOLOv5. In *IMPROVE* (pp. 17-25).

[9] Yadav, Y. G., Kiran, V. S., Karthik, V., Thadikamalla, G. A., & Kumaran, P. (2023, May). Real time sign language recognition using custom convolutional neural network and YOLOv5. In *International Conference on Intelligent Computing, Smart Communication and Network Technologies* (pp. 157-171). Cham: Springer Nature Switzerland.

[10] Buttar, A. M., Ahmad, U., Gumaei, A. H., Assiri, A., Akbar, M. A., & Alkhamees, B. F. (2023). Deep learning in sign language recognition: a hybrid approach for the recognition of static and dynamic signs. *Mathematics*, *11*(17), 3729.

******

ISSN (Online):2583-0732