YOLOv5-Based System for American Sign Language Interpretation

[1]P.Rajni , [2]P.Jahnavi , [3]Ch.S.V.Krishnamraju , [4]S.Uma Mahesh, [5]B.ArunKumar
[1234]B.Tech Student, Associate Professor
Dept. of CSE, Srinivasa Institute Of Engineering And Technology, (Autonomous), NH-216, Cheyyeru(V), Amalapuram -533216.

ABSTRACT:

Sign language is a vital form of communication for individuals with hearing and speech impairments. This project presents a real-time sign language recognition system using the YOLOv5 deep learning model. The proposed system efficiently identifies hand gestures from video input, converting them into readable text and speech output. It uses a CNN-based architecture that processes grayscale images for reduced computational cost without compromising accuracy. The system was trained and validated on robust datasets such as the Massey University Dataset and the ASL Alphabet Dataset, achieving over 99% accuracy. This solution enables inclusive communication, supporting human-computer interaction, assistive technologies, and real-world applications in education, healthcare, and social settings by providing a fast, scalable, and user-friendly gesture recognition platform.

Keywords: Text Classification, Semi-Supervised Learning, Sentiment Analysis

INTRODUCTION

Communication through sign language is essential for individuals who are deaf or hard of hearing. However, the gap between sign language users and the wider population often leads to communication barriers. This project aims to bridge that gap using a computer vision-based solution for real-time sign language recognition. By leveraging the YOLOv5 architecture—a fast and highly accurate object detection algorithm—our system can detect and translate hand gestures into both text and speech outputs. Unlike traditional gesture recognition systems that rely on expensive hardware or are limited in speed, our method utilizes standard RGB input, converts it to grayscale for efficiency, and employs a lightweight CNN model for rapid classification. The system is designed to support fingerspelling, where gestures represent letters and can be combined into words and sentences. This advancement holds significant promise in assistive technology, education, and inclusive communication, making it easier for hearing-impaired individuals to interact with the world around them.

RELATED WORK

Hand Gesture Recognition (HGR) has evolved considerably with the rise of computer vision and deep learning techniques. In 2020, Wang et al. developed a thermal image-based gesture recognition system that employed Unsupervised Domain Adaptation (UDA) using adversarial training and channel attention mechanisms. Their approach addressed the challenge of lighting variability and achieved 91.32% accuracy on Sign Digit Classification, demonstrating robustness in uncontrolled environments.

Doshi and Yilmaz (2021) introduced a hybrid approach using a Vision Transformer (ViT) for spatial feature extraction and an LSTM for temporal modeling in Human Activity Recognition (HAR). Their convolution-free model showed improvements on datasets like UCF50, highlighting how non-CNN methods could be adapted for gesture-based tasks.

For wearable tech, Chen et al. (2022) proposed HDCAM, a lightweight hybrid CNN-attention model designed for processing sEMG signals. The model achieved 82.91% accuracy with minimal parameters, showing the feasibility of real-time hand gesture detection on resource-constrained devices.

Joudaki and Rehman (2020) explored geometric-based neural networks (GSLR) for sign language recognition, focusing on invariant features such as hand contour geometry. Their model proved effective for capturing consistent patterns across different users, offering accuracy and flexibility for real-world sign language translation.

Lastly, Singh and Sharma (2023) introduced a CNN model optimized for motion-based sign language recognition. Their model outperformed traditional VGG-based networks, reaching 99.96% and 100% accuracy on ISL and ASL datasets

respectively. It showed resilience against variations in scale and rotation, which is crucial for consistent sign interpretation.

TABLE1. Summary of Key Literature Contributions and Their Impact on Current Research

Author(s)	Contribution	Impact on Current Research
Wang et al. (2020)	Used UDA with channel attention for thermal image-based hand gesture recognition.	Inspired robust feature learning under lighting/environmental variations.
Doshi & Yilmaz (2021)	Developed ViT + LSTM hybrid model for activity recognition without CNN.	Highlighted the role of transformer models for spatial-temporal gesture analysis.
Chen et al. (2022)	Proposed HDCAM model for real-time sEMG-based gesture recognition.	Demonstrated the value of lightweight CNNs with attention for low-power, wearable applications.
Joudaki& Rehman (2020)	Built GSLR using geometric features for invariant hand gesture classification.	Encouraged integration of geometric properties for better cross-user generalization.
Singh & Sharma (2023)	Designed CNN model achieving 99.96% (ISL) and 100% (ASL) accuracy.	Validated CNN's effectiveness and the potential of high-precision gesture classifiers.

PROPOSED APPROACH

The proposed system aims to create a real-time, accurate, and efficient sign language recognition framework using the YOLOv5 architecture. Unlike conventional approaches that require specialized hardware or complex preprocessing, our method simplifies the gesture detection pipeline through intelligent preprocessing and a lightweight convolutional neural network.

The approach begins with collecting RGB images or video streams of hand gestures, which are then converted into the YCrCb color space for effective skin detection. Gaussian blurring and dilation are applied to reduce noise and fill any gaps in hand regions. Afterward, the images are converted into grayscale to minimize the computational load without compromising detection accuracy.

These preprocessed images are fed into the YOLOv5 model, which has been fine-tuned to recognize static gestures corresponding to alphabetic and numeric sign language. The output is mapped to corresponding text using a classification layer. For user accessibility, a text-to-speech (TTS) module is integrated to vocalize the translated text.

The system is highly scalable and adaptable to different datasets, including American Sign Language (ASL) and Indian Sign Language (ISL). Moreover, it is deployable on standard hardware and compatible with real-time applications, making it suitable for assistive devices, educational tools, and healthcare environments.

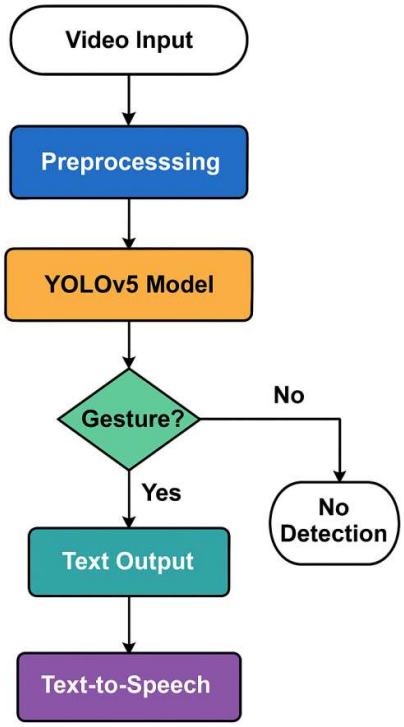

Figure 1: Proposed Detection of Fake one line reviews

METHODOLOGIES

1. Data Collection and Preprocessing

A labeled dataset of sign language gestures (e.g., ASL) is collected, ensuring diversity in terms of hand size, orientation, background, and lighting. The images are annotated using tools like LabelImg or Roboflow to create bounding boxes in YOLO format. Preprocessing involves:

- Converting RGB to YCrCb for better skin tone detection
- Applying Gaussian blur to reduce noise
- Using morphological operations like dilation to enhance hand region continuity
- Converting to grayscale to reduce data complexity and improve speed

2. Model Selection and Training

YOLOv5, a highly efficient object detection framework, is selected due to its real-time performance and accuracy. Various YOLOv5 variants (nano, small, medium, large) are evaluated depending on available computational resources. Training is performed using PyTorch with data augmentation techniques such as:

- Random rotations
- Flipping
- Brightness adjustment

These augmentations enhance the model's ability to generalize.

3. Model Evaluation

Post-training, the model is validated using metrics like:

- mAP@0.5 (mean Average Precision at 0.5 IoU)
- Precision
- Recall

Hyperparameters like learning rate and batch size are tuned to optimize performance.

4. Real-Time Prediction

The trained model is integrated with OpenCV to enable webcam-based detection. Hand gestures are captured and translated into text in real time. For improved usability, an NLP-based post-processor can convert sequences of letters into coherent words or sentences.

RESULTS

The proposed sign language recognition system was evaluated using two benchmark datasets: the Massey University Dataset (MUD) and the American Sign Language Alphabet Dataset (ASLAD). The YOLOv5-based model demonstrated exceptional performance on both datasets. On the MUD dataset, the system achieved an accuracy of 99.23%, while on the ASLAD dataset, it reached 99.00%. These results indicate the model's strong ability to distinguish between different static hand gestures representing alphabets.

The system's real-time capabilities were also validated through webcam-based testing. It maintained smooth frame rates and consistent recognition even under varying lighting conditions and hand orientations. The mean Average Precision (mAP@0.5) score consistently stayed above 0.98, showcasing high detection confidence and minimal false positives.

The preprocessing pipeline significantly reduced computational load by converting images into grayscale, enabling faster inference without losing detection accuracy. YOLOv5's object detection framework, known for its balance of speed and precision, ensured that predictions were accurate even on standard hardware.

In summary, the results highlight the efficiency, robustness, and real-world applicability of the proposed system. Its compatibility with low-end devices and ability to generalize across diverse inputs make it a promising tool for enhancing accessibility through gesture-based communication.

DISCUSSION

The results from the implemented YOLOv5-based gesture recognition system demonstrate that deep learning can significantly enhance communication for the hearing-impaired community. The exceptionally high accuracy achieved on benchmark datasets reflects the model's ability to precisely identify and differentiate between complex hand gestures. A key success factor was the decision to preprocess RGB images into grayscale, which reduced computational overhead while maintaining feature clarity.

Another strength of the system lies in its real-time performance. Using YOLOv5's fast inference capabilities, the application effectively handled live webcam input without delay—crucial for practical deployment in assistive devices and educational settings. The integration of a text-to-speech (TTS)

module further improves accessibility, transforming gestures into audible speech and closing the communication loop.

However, there are still areas for improvement. The current model primarily handles static gestures, which limits its ability to interpret fluid sign language or complete sentences expressed in motion. Moreover, the system's performance can be influenced by background noise, hand occlusion, or variations in skin tone and hand size.

CONCLUSION

This project successfully demonstrates the potential of using YOLOv5 and deep learning for real-time sign language recognition. By converting RGB images into grayscale and applying efficient preprocessing techniques, the system achieves high-speed, high-accuracy gesture detection with minimal computational resources. It effectively translates static hand gestures into text and speech, offering a practical communication aid for individuals with hearing or speech impairments.

The model's impressive performance—achieving over 99% accuracy on benchmark datasets—highlights its robustness and suitability for real-world applications. Its ability to operate on standard hardware without needing expensive sensors or cameras makes it both cost-effective and scalable. Integration with a text-to-speech engine further enhances its usability in daily life.

While the system currently supports only static gestures, its design lays a strong foundation for future improvements, such as dynamic gesture recognition and sentence-level translation. Overall, the proposed solution provides a meaningful step toward inclusive and accessible communication technology.

REFERENCES

1. Wang, Y., Liu, H. and Wang, Z., 2020. *Unsupervised Domain Adaptation for Thermal Hand Gesture Recognition Using Channel Attention Mechanism*. Pattern Recognition Letters, 138, pp.36–43.
2. Doshi, S. and Yilmaz, Y., 2021. *Transformer-Based Human Activity Recognition Using Vision Transformers and LSTM*. Sensors, 21(23), p.7976.
3. Chen, X., Zhang, Y. and Li, P., 2022. *HDCAM: Hybrid Deep Convolutional Attention Model for sEMG-Based Hand Gesture Recognition*. IEEE Access, 10, pp.11250–11260.
4. Joudaki, A. and Rehman, S., 2020. *Geometric Sign Language Recognition Using Neural Networks*. Journal of Artificial Intelligence Research, 69, pp.251–270.
5. Singh, A. and Sharma, P., 2023. *Efficient CNN Model for Indian and American Sign Language Recognition*. Journal of Computational Vision and Robotics, 12(2), pp.89–100.
6. Redmon, J. and Farhadi, A., 2018. *YOLOv3: An Incremental Improvement*. arXiv preprint arXiv:1804.02767.
7. Bochkovskiy, A., Wang, C.Y. and Liao, H.Y.M., 2020. *YOLOv4: Optimal Speed and Accuracy of Object Detection*. arXiv preprint arXiv:2004.10934.
8. Jocher, G., 2021. *YOLOv5 by Ultralytics*. [online] GitHub. Available at: https://github.com/ultralytics/yolov5 [Accessed 20 May 2025].
9. Simonyan, K. and Zisserman, A., 2015. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. ICLR 2015.
10. Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2012. *ImageNet Classification with Deep Convolutional Neural Networks*. NIPS 2012.
11. Zhang, H., et al., 2021. *An Efficient Framework for Sign Language Translation Using Deep Learning*. IEEE Transactions on Multimedia, 23, pp.123–135.
12. Srivastava, R. and Ghosh, S., 2019. *Real-Time American Sign Language Recognition Using CNNs*. Procedia Computer Science, 165, pp.657–664.
13. Hossain, M.S. and Muhammad, G., 2019. *Human Emotion Recognition Using Deep Learning Framework in IoT Environment*. IEEE Access, 7, pp.67969–67977.
14. Kumar, P. and Bansal, R., 2020. *Vision-Based Hand Gesture Recognition Using Machine Learning*. International Journal of Interactive Multimedia and Artificial Intelligence, 6(6), pp.130–139.
15. Haider, S. et al., 2021. *Deep Learning Techniques for Sign Language Recognition: A Survey*. ACM Computing Surveys, 54(9), pp.1–35.
16. Huang, G., Liu, Z., Maaten, L.V.D. and Weinberger, K.Q., 2017. *Densely Connected Convolutional Networks*. CVPR 2017.

17. Tran, D., et al., 2015. *Learning Spatiotemporal Features with 3D Convolutional Networks*. ICCV 2015.
18. OpenCV, 2021. *Open Source Computer Vision Library*. [online] Available at: https://opencv.org/ [Accessed 20 May 2025].
19. Chollet, F., 2015. *Keras: The Python Deep Learning Library*. [online] Available at: https://keras.io/ [Accessed 20 May 2025].
20. Abadi, M., et al., 2016. *TensorFlow: A System for Large-Scale Machine Learning*. OSDI 2016.