Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Sign Language Interpreter for American Sign Language

Anjali Yeole[1], Priyanka Amrute[1], Saylee Gharge[1], Vaibhav Narwane[2]

[1]Vivekanand Education Society's Institute of Technology, Mumbai, India
[2]K J Somaiya college of Enginering, Mumbai, India
Email: anjali.yeole@ves.ac.in

Sign language serves as a visual method of communication that relies on hand gestures, facial expressions, and other visual cues. It is primarily used by individuals with hearing or speech impairments as their main mode of interaction. Additionally, people with conditions such as autism spectrum disorder may find sign language a valuable tool for enhancing communication. This paper presents a deep learning-based system for recognizing and classifying hand gestures in American Sign Language (ASL). This system aims to bridge the communication gap by utilizing models such as YOLOv5 for real-time detection and MobileNet for alphabet classification. The proposed approach involves data preprocessing, transfer learning, and object detection, achieving an overall accuracy of 87.5%. Despite challenges in lighting conditions and gesture variability, the paper demonstrates potential for real-world application and sets the foundation for future work in sentence formation.
Keywords: sign language detection, YOLOv5, MobileNet, object detection, ASL, deep learning

1. Introduction

Over 5% of the world's population – or 430 million people – require rehabilitation to address their disabling hearing loss (including 34 million children). It is estimated that by 2050 over 700 million people – or 1 in every 10 people – will have disabling hearing loss.[1] Deaf individuals have created a form of communication that is specifically suited to their needs, known as sign language, which provides them with an effective and accessible way to interact. Non-verbal manner conveys and communicates our views, emotions, and thoughts visually through sign language. Deaf and hard-of-hearing people use sign languages for their communication with other people. Sign languages are also used for the communication between deaf and non-deaf people, including different types of hand gestures and facial

expressions for communication and emotional expression[2]. Sign language can also be beneficial for individuals with disabilities, such as Autism and Down Syndrome, to aid in communication. There are over 300 distinct sign languages, which differ across countries. Even nations that share the same spoken language may have different sign languages. For example, American, British, and Australian Sign Languages are all unique forms of English. A sign typically involves specific hand gestures, facial expressions, lip movements, shapes, or body movements. This work thus helps people who are suffering from deaf and dumb. The major concern of this paper is to connect them with the real world with great esteem[3].Sign language recognition involves translating gestures into words or letters of established spoken languages. Therefore, using an algorithm or model to convert sign language into text can help bridge the communication gap between individuals with hearing or speech impairments and the wider world.

Sign language recognition is a complex and challenging task, offering numerous research opportunities with the current advancements in artificial intelligence technology. It involves various elements, such as datasets, input modalities, features, classification, computational resources, and applications. The primary features involved in sign language recognition include hand movements, facial expressions, and body movements. Hand gesture recognition can be approached in two main ways: vision-based methods and data glove methods. Vision-based gesture recognition is a rapidly developing field within computer vision and machine learning. Since hand gestures are a natural form of human interaction, many researchers are exploring ways to enhance human-computer interaction (HCI) without the need for additional devices. The primary objective of gesture recognition research is to develop systems that can accurately identify specific human gestures, which can then be used to convey information. To achieve this, vision-based systems must have fast and reliable hand detection and gesture recognition capabilities, operating in real-time. This particular research aims to create a vision-based system capable of performing real-time sign language recognition, as it offers a more straightforward and intuitive communication method between humans and computers.

The paper utilizes a Deep Learning model for sign language recognition. Deep Learning, a subset of machine learning, involves neural networks with three or more layers. It focuses on algorithms that are inspired by the structure and function of the human brain, known as artificial neural networks. These networks aim to replicate the brain’s ability to process and learn from large volumes of data. By combining data inputs, weights, and biases, Deep Learning neural networks attempt to simulate the brain's processes and improve recognition performance.

In this study, we have employed the YOLOv5 (You Only Look Once version 5) algorithm, a state-of-the-art object detection model, to recognize hand gestures in sign language[4]. To tailor the model for sign language recognition, we fine-tuned it using a custom dataset specifically designed for American Sign Language (ASL). This dataset includes diverse hand gestures with variations in orientation, lighting, and background to improve the robustness of the model. The fine-tuning process involved adapting the pre-trained YOLOv5 weights to the ASL dataset, enabling the model to accurately detect and classify ASL gestures in real time. This approach combines the efficiency of YOLOv5 with the specificity of a domain-focused dataset, achieving high accuracy and making it suitable for practical applications in bridging the communication gap[5].YOLOv5 has been widely recognized as an effective model for sign

language recognition, with various studies highlighting its capability to detect American Sign Language gestures in real-time [6].

2. Literature Review

Sign language detection and recognition have garnered significant research interest due to their potential to bridge communication gaps between hearing and hearing-impaired individuals. Various methodologies and techniques have been proposed to enhance the performance and applicability of such systems. Pathak et al. (2022) proposed a system for communication between normal and deaf individuals using hand gestures, which employs a webcam or in-built camera for real-time recognition. While their system performs well under controlled conditions, it faces challenges with lighting and background variability, which is a common issue in sign language recognition. Our paper resonates with these findings, as we also face accuracy issues influenced by lighting conditions, which affect the model's performance. To mitigate these issues, we have implemented advanced preprocessing techniques such as noise reduction and image normalization. These preprocessing methods ensure that the model generalizes better and performs effectively in diverse environments, much like the controlled setup Pathak et al. suggested, but with more adaptability[7].

Shirbhate et al. (2021) conducted a comprehensive review of feature extraction techniques for sign language recognition, highlighting the use of machine learning algorithms like Neural Networks, Support Vector Machines (SVM), and Hidden Markov Models (HMM). These methods, while effective, do not offer the same level of real-time performance that modern object detection frameworks can. Their paper diverges by using YOLOv5, a state-of-the-art deep-learning model designed for object detection. This choice enables us to detect ASL gestures in real-time, as opposed to traditional methods that typically sequentially process gestures, often reducing processing speed. The review emphasizes the importance of feature extraction, which is similarly critical in our paper, especially in how YOLOv5 extracts complex features from hand gestures to recognize ASL alphabets accurately[8].

PCA-based techniques for hand gesture recognition, as explored by several studies, utilize skin colour models and template matching for static gesture recognition. While PCA effectively reduces dimensionality, it lacks the depth required for recognizing intricate gestures with high accuracy. In contrast, our paper uses YOLOv5, which combines convolutional layers and multi-scale processing to learn deeper, more complex features from the images. This method allows the model to not only recognize static hand gestures but also handle variations in the size and position of gestures within the image. Our approach extends beyond the limitations of PCA by using CNNs to capture both low- and high-level features, making the system more robust for real-time recognition[9].

Phong and Ribeiro (2022) focused on dynamic hand gesture recognition for American Sign Language, utilizing DenseNet and LSTM models to handle sequential data. While our paper currently addresses static ASL gestures, the insights from this study open up the potential for future work, where we could incorporate temporal information to enhance dynamic gesture recognition. If we extend our model to dynamic gestures, techniques such as LSTM and Temporal Convolutional Networks (TCN) could be used to capture the sequential nature of

sign language, improving accuracy for continuous signs[10].

Telugu Sign Language (TSL) recognition is also an underexplored area, and the study by Reddy et al. addresses this gap by leveraging the YOLOv5 algorithm for gesture detection and classification like the one explored in this paper. The authors curated a custom dataset of TSL gestures, including preprocessing and labelling steps to prepare the data for training. YOLOv5, a real-time object detection algorithm, was implemented to identify and localize hand gestures in video frames. The model demonstrated high accuracy and efficiency in recognizing TSL gestures, outperforming traditional methods in terms of precision and speed. The work provides a significant contribution by introducing a TSL-specific dataset and a robust recognition framework, paving the way for future advancements in regional sign language recognition and its integration into assistive technologies[5].

Chavan et al. (2022) explored sign language detection using traditional feature extraction methods like ORB combined with CNN architectures. They demonstrated the effectiveness of these methods but also highlighted the challenge of manually designing features for gesture recognition. Our paper builds upon their work by utilizing YOLOv5, which automates feature extraction and classification in a single, end-to-end pipeline. This integration of feature extraction and classification into one model makes YOLOv5 more efficient and accurate for detecting ASL gestures in real time, reducing the need for complex manual feature engineering[11].

A study on 3D Convolutional Neural Networks (3DCNN) for hand gesture recognition explores dynamic gesture recognition by analyzing the temporal aspect of sign language gestures. While our current system primarily focuses on recognizing static ASL gestures, we are aware of the growing need for dynamic gesture recognition. The use of 3DCNNs in this context provides an exciting future direction for our paper. By integrating dynamic gesture recognition capabilities, we can improve the system's versatility, enabling it to recognize more complex sequences of hand movements in real time[12].

In conclusion, the literature highlights various methodologies for sign language recognition, each contributing valuable insights into feature extraction, preprocessing, and model architecture. Our project draws from these studies by employing YOLOv5 for real-time ASL detection, leveraging preprocessing techniques to handle environmental challenges, and focusing on static gesture recognition. While our approach offers improvements over traditional methods in terms of speed and accuracy, we recognize the potential for future enhancements, such as dynamic gesture recognition, which could further elevate the performance and usability of the system.

3. Methodology

3.1 Data Collection and Preprocessing

The dataset we used is for American Sign Language letters, containing 1728 images with bounding boxes and labels for each letter of the alphabet. The images have a resolution of 416 x 416 pixels and are in the JPG format. Each bounding box specifies the coordinates of the top-left and bottom-right corners of the region containing the letter, as well as the label corresponding to the letter[13].

3.1.1 Dataset Collection

The data collection process focused on curating a custom dataset to suit the system's requirements. Images of hand gestures representing the 26 alphabets of ASL were gathered using controlled setups to ensure consistency in lighting, background, and gesture clarity. These images were labelled with bounding boxes to highlight regions of interest, such as hand contours, using the Roboflow platform. This labelling process ensured that the dataset was both comprehensive and well-structured, catering to the specific requirements of YOLOv5 and MobileNet.

3.1.2 Dataset Preprocessing

To enhance the quality and uniformity of the dataset, the following preprocessing techniques were applied:

1. Grayscale Conversion: Each image was converted to grayscale to reduce complexity and focus on the structural features of the hand gestures.

2. Normalization: Pixel values were scaled to the range of [0,1] to enhance the training process and improve convergence.

3. Data Augmentation: Methods like rotation, flipping, and zooming were applied to increase the diversity of the dataset and enhance the model's ability to generalize.

4. Noise Reduction: Gaussian filtering was applied to reduce noise and enhance image clarity.

3.1.3 Annotation and Labeling

The images were annotated using bounding boxes to indicate the hand regions and labelled with the corresponding alphabet (A-Z). Each image was carefully annotated for ASL alphabets to ensure accurate and consistent labelling for model training.

3.2 Model Architecture

The proposed model for recognizing American Sign Language (ASL) alphabets is built upon the YOLOv5 (You Only Look Once version 5) object detection framework, optimized for real-time performance and accuracy[14]. The architecture as illustrated in Figure 1, begins with the CSP (Cross-Stage Partial) Backbone, which is responsible for extracting low-level and high-level features from the input image. This backbone is composed of a series of convolutional layers and CSP modules, where the CSP design partitions feature maps into two sections, processes them through bottleneck residual blocks, and merges them. This architecture ensures efficient gradient flow, reduced computational overhead, and the preservation of both spatial and semantic information.The convolutional layers in the early stages are responsible for detecting fundamental features like edges and textures, whereas the deeper layers focus on capturing more intricate patterns, such as hand shapes and gesture specifics.A pivotal component of the architecture is the Spatial Pyramid Pooling-Fast (SPPF) layer, which aggregates spatial information across multiple scales, enhancing the detection of ASL gestures regardless of size or position in the image. The architecture also incorporates upsampling and concatenation layers, which integrate features from different levels of the network to ensure multi-scale object detection. This design allowed the model to identify intricate ASL gestures

effectively. After a sequence of upsampling, concatenation, and convolutional operations, the network ended with a detection head that generated bounding boxes and class probabilities for each gesture.[15].

Transfer learning was employed by initializing the model with pre-trained YOLOv5 weights, which accelerated the learning process and improved performance on the ASL-specific dataset. Training was conducted with hyperparameter tuning to optimize the learning rate, batch size, and IoU thresholds for a balance between precision and recall. The spatial pyramid pooling, feature aggregation, and detection head work in unison to deliver high performance in detecting and classifying gestures. The modularity of the architecture also allows for further enhancements, such as the integration of additional layers or features for dynamic gesture recognition, which could form the basis for future work.

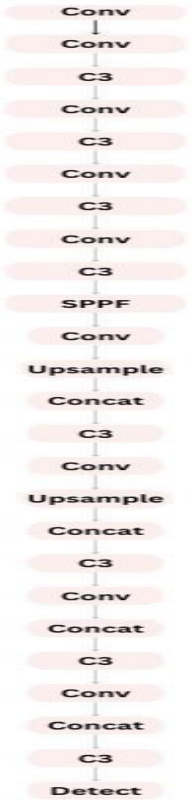

Figure 1: Model Architecture

4. Result

After training the YOLOv5 model on the selected dataset and optimizing it, the model's performance was evaluated by testing its accuracy on individual alphabets in the sign language dataset. This comprehensive evaluation helped to assess the model's overall effectiveness. The results revealed that the model achieved an overall accuracy of 87.5%, marking a significant

improvement over the initial accuracy of 78%.

Table 1 presents the accuracy achieved by the YOLOv5 model for each alphabet. The analysis highlights a variation in accuracy across different alphabets. The highest accuracy, 0.94, was recorded for the alphabet 'V', while the lowest accuracy, 0.43, was observed for the alphabet 'O'. Certain alphabets, such as 'A', 'B', 'L', and 'W', demonstrated high accuracy scores above 0.90, indicating strong performance in recognizing these letters. Conversely, other alphabets, including 'E', 'G', 'H', 'J', 'N', 'O', and 'Z', exhibited accuracy scores below 0.70, suggesting areas for further improvement. These findings emphasize the model's strengths and weaknesses, offering insights into potential avenues for optimization.

The confusion matrix, shown in Figure 2, offers a comprehensive visual representation of the model's performance in classification. It highlights the instances of correct and incorrect predictions for each alphabet, further validating the quantitative results.

Table 1: Accuracy using YOLOv5 model

Alphabets	Accuracy	Alphabets	Accuracy
A	0.92	N	0.51
B	0.92	O	0.43
C	0.80	P	0.70
D	0.76	Q	0.81
E	0.45	R	0.90
F	0.86	S	0.90
G	0.56	T	0.85
H	0.50	U	0.66
I	0.90	V	0.94
J	0.45	W	0.92
K	0.91	X	0.60
L	0.92	Y	0.85
M	0.70	Z	0.65

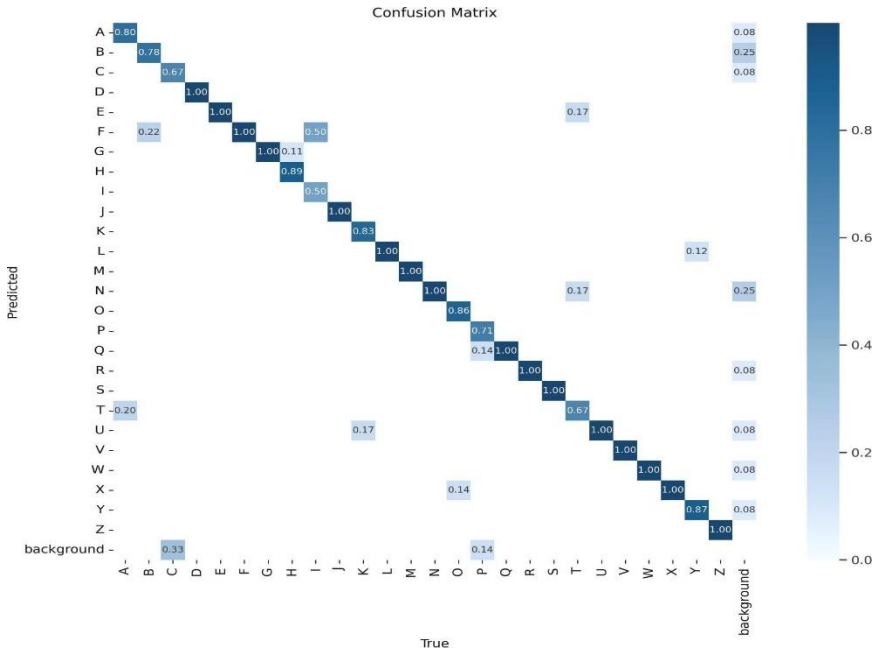

Figure 2: Confusion Matrix Graph

The model was tested with real interactions involving human hand gestures to evaluate its performance in dynamic, practical scenarios. Figure 3 showcases live accuracy testing, displaying images of ASL alphabets processed by the YOLOv5 model. For each tested image, the model predicts the corresponding ASL letter and provides a confidence score for the prediction, further highlighting its ability to recognize most letters accurately while identifying areas for refinement.

Figure 3: Live Accuracy Testing

The model was then deployed in a web application, as shown in Figure 4, successfully facilitating real-time predictions by leveraging the YOLOv5 model. Users can upload an image or provide a video feed via their webcam, which is processed to detect and display the predicted sign language alphabet along with the associated confidence score. However, it was observed that lighting conditions significantly impacted the model's accuracy, highlighting the need for advanced preprocessing techniques to improve performance. The system has been designed with scalability in mind, allowing for the inclusion of additional sign language alphabets and features in future iterations.

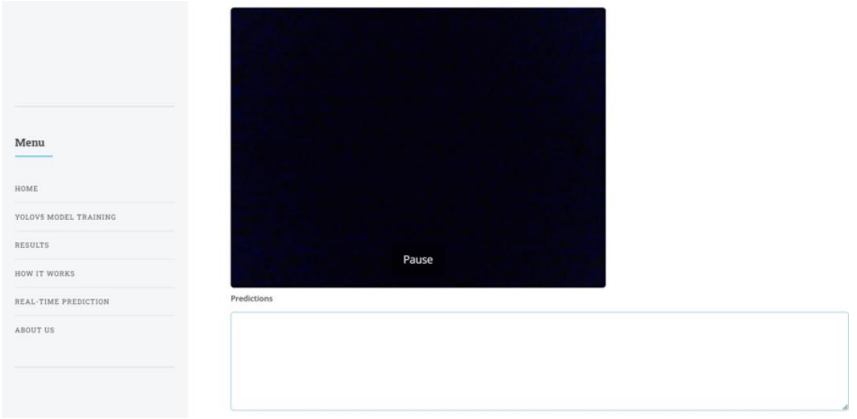

Figure 4: Real-time Prediction Page

5. Conclusion

In conclusion, this study presents a proof-of-concept for real-time ASL alphabet recognition using YOLOv5 and MobileNet. Despite challenges in lighting and gesture variability, the system demonstrates potential for improving communication accessibility for the hearing and speech impaired. The paper also showcases the potential of deep learning in addressing real-world problems and highlights the importance of continuous improvements and developments in the field.

For future work, several areas could be explored to further enhance the performance and capabilities of the system. One potential avenue for improvement is the addition of more data to the training set, as this could improve the model's ability to recognize sign language alphabets in various contexts and settings. Another area of improvement could be the integration of natural language processing (NLP) techniques to enable the system to interpret full sentences in sign language. This could potentially allow for more complex and nuanced communication between users.

In addition, the system could also benefit from the development of a mobile application, as this would enable users to access the interpreter on the go and in real-world settings. Another potential area of development is the incorporation of more advanced computer vision techniques, such as object tracking and segmentation, which could improve the model's ability to recognize and track the movement of sign language gestures.

In summary, the paper has established a solid foundation for future progress in sign language interpretation and deep learning. With ongoing advancements and enhancements, the system holds the potential to greatly impact the lives of the deaf and hard-of-hearing community, allowing them to communicate more efficiently and engage fully in society.

Acknowledgement

We extend our heartfelt gratitude to our students, Disha Bhat, Shivam Gupta, and Mehvish Khan, for their valuable assistance in executing the required programs.

References

1. World Health Organization, "Deafness and hearing loss," 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss
2. M. Papatsimouli et al., "Real Time Sign Language Translation Systems: A review study," 2022 11th International Conference on Modern Circuits and Systems Technologies (MOCAST), Bremen, Germany, 2022, pp. 1-4, doi: 10.1109/MOCAST54814.2022.9837666.
3. P. Dubey and P. Shrivastav, "Sign language conversion flex sensor based on IoT," International Journal for Research in Applied Science and Engineering Technologies, vol. X, no. Y, pp. 1-10, Year. [Online]. Available: https://www.ijraset.com/research-paper/sign-language-to-text-conversion-using-flex-sensors
4. J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 779-788, doi: 10.1109/CVPR.2016.91.
5. Reddy, V. P., Reddy, V. V. B., & Sukriti. (2024). Sign Language Recognition based on YOLOv5 Algorithm for the Telugu Sign Language. arXiv preprint arXiv:2406.10231. https://arxiv.org/abs/2406.10231.
6. Tyagi, Shobhit & Upadhyay, Prashant & Fatima, Hoor & Jain, Sachin & Sharma, Avinash. (2023). American Sign Language Detection using YOLOv5 and YOLOv8. 10.21203/rs.3.rs-3126918/v1.
7. Pathak, A., Kumar, A., Priyam, Gupta, P., & Chugh, G. (2022). Real-time sign language detection. International Journal for Modern Trends in Science and Technology, 8, 32–37.
8. Shirbhate, R. S., Shinde, V. D., Metkari, S. A., Borkar, P. U., & Khandge, M. A. (2021). Sign language recognition using machine learning algorithm.
9. Ahuja, M. K., & Singh, A. (2015). Hand gesture recognition using PCA. International Journal of Computer Science Engineering and Technology (IJCSET), 5(7), 267–271.
10. Phong, N. H., & Ribeiro, B. (2022). Action recognition for American Sign Language. arXiv Preprint. https://doi.org/10.48550/arXiv.2205.12261
11. Deshmukh, S., Fernandes, F., & Chavan, A. (2022). Sign language detection. arXiv Preprint. https://doi.org/10.48550/arXiv.2209.03578
12. Raghuveera, T., Deepthi, R., Mangalashri, R., & Akshaya, R. (2020). A depth-based Indian sign language recognition using Microsoft Kinect.
13. Roboflow, "American Sign Language Letters Dataset," [Online]. Available: https://public.roboflow.com/object-detection/american-sign-language-letters
14. Thuan, D. Evolution of Yolo Algorithm and Yolov5: The State-of-the-Art Object Detection Algorithm. 2021. Available online: https://www.theseus.fi/handle/10024/452552
15. Jocher, A. (2020). YOLOv5 [GitHub repository]. Retrieved from https://github.com/ultralytics/yolov