Hand Sign Translator Using Yolov8, OpenCV and Machine Learning

[1] Sharon Paul B, [2] S Sanjay Kumar, [3] K Senthamizh Selvan, [4]Dr. P. Dinesh Kumar
[5]Dr. F. Antony Xavier Bronson
[1,2,3]Final Year students, Department of CSE,
Dr. M. G. R Educational and Research Institute, Maduravoyal, Chennai-95
[4,5] Associate Professor, Department of CSE,
Dr. M. G. R Educational and Research Institute, Maduravoyal, Chennai-95
Corresponding Author: Sharon Paul. B, sharonpaul.engineer@gmail.com

Abstract--There are people who cannot communicate in the same way as others. Deaf and hard-of-hearing individuals use sign languages to communicate with others. Sign languages facilitate communication between deaf and non-deaf people, employing various hand gestures and facial expressions to convey messages and emotions. However, there has historically been a communication gap between deaf and non-deaf individuals. Fortunately, advancements in technology have significantly reduced this gap. Our research focuses on developing real-time sign language translators. These translators aim to convert sign language into text and text into sign language, enabling seamless communication between deaf and non-deaf individuals. As technology continues to evolve, we can expect even more innovative solutions to emerge, further bridging the communication gap and fostering a more inclusive world.

Keywords: Sign Language Translations, Machine Learning, Computer Vision.

I. INTRODUCTION:

The goal of this project was to build a machine learning model that will be able to classify which letter of the American Sign Language (ASL) alphabet is being signed, given an image of a signing hand and some commonly used words in day-to-day life. This project aims to build a possible sign language translator, which can take communications in sign language and translate them into text. Loneliness and depression exist at higher rates among the deaf population, especially when they are immersed in a hearing world. Sign languages are made up of two groups, namely static gesture, and dynamic gesture. The static gesture is used for alphabet and number representation, whereas the dynamic gesture is used for specific concepts. Dynamic also includes words, sentences, etc.

In this project, our primary focus is on creating a model that can recognize hand movements and combine each motion to form a whole word & then convert that predicted text into speech. Few gestures that are practiced are displayed in the image below

Fig. 1: American Sign Language

II. RELATED WORK:

Sign languages have been studied by the computer vision community for the last three decades. The end goal of computational sign language research is to build translation, that are capable of translating sign language videos to spoken language sentences and vice versa, to ease the daily lives of the deaf. Many approaches have been proposed to address the constraints represented by hand detection. Those methods broadly fall into four categories, namely, skin segmentation, depth-based detection, hand detection based on hand-crafted features, and CNN-based approaches.

Many algorithms for hand detection rely on skin colour segmentation to detect and extract hands from the background. *Dardas N.H, Georganas N.D: Real-time hand gesture detection and recognition [17]* using bag-of-features and support vector machine techniques proposed a thresholding method to segment hands in the hue, saturation, and value (HSV) colour space after extracting other skin regions, such as the face. *Mittal A, Zisserman A, Torr P: Hand detection using multiple proposals [18]*, used a skin-based detector to generate hand hypotheses for the first stage of their hand detection algorithm. To improve the robustness of skin segmentation, *Stergiopoulou E, Sgouropoulos K, Nikolaou N, Papamarkos N, Mitianoudis N: Real time hand detection in a complex background[19]*, used a skin-probability map (SPM) in the HSV colour space for skin colour classification, along with extra information, such as motion and morphology weights of hands. Combining skin detection with deep learning object detectors has also been proposed. *Roy K, Mohanty A, Sahay R. R: Deep Learning Based*

Hand Detection in Cluttered Environment Using Skin Segmentation [20], proposed two architectures (patch-based CNN and regression-based CNN for skin segmentation). Their main purpose was to reduce the occurrence of false positives resulting from the estimated bounding boxes of recent object detectors.

The recent development of emerging colour-depth camera-based sensing techniques, such as the Microsoft Kinect TM, has solved many problems related to hand gesture recognition, including hand extraction using depth data. *Keskin C, Kiraç F, Kara Y. E, Akarun L: Hand pose estimation and hand shape classification using multi-layered randomized decision forests [21]* extracted scale invariant shape features from depth images then fed them into a per-pixel randomized decision forest (RDF) classifier. Inspired by the recent success of convolutional neural networks, researchers have proposed numerous methods for object detection and recognition based on CNN. Consequently, these methods have been developed and used for hand detection. *Roy K, Mohanty A, Sahay R.R: Deep Learning Based Hand Detection in Cluttered Environment Using Skin Segmentation [20],* proposed a two-stage hand detector based on the region-CNN (R-CNN) and Faster R-CNN frameworks. Initially, they used an object detection algorithm to generate hand regions and then a CNN-based skin segmentation was used to reduce occurrences of false positives during hand detection. *Huang Y, Liu X, Zhang X, Jin L: A Pointing Gesture Based Egocentric Interaction System Dataset, Approach and Application [24],* proposed an egocentric interaction system using Faster R-CNN to locate and recognize static hand gestures. Their system achieved better performance on a challenging dataset under challenging conditions.

Recently, deep learning-based methods have emerged and advanced the research in this area. *Chevtchenko S.F, Vale R.F, Macario V, Cordeiro F.R: A convolutional neural network with feature fusion for real-time hand posture recognition [26],* presented a novel approach based on combining traditional hand-crafted features with a CNN. They evaluated their approach on depth and grayscale images, where the background has already been removed using depth data by considering the hand as the closest object to the camera. *Liang C, Song Y, Zhang Y: Hand gesture recognition using view projection from point cloud [27],* utilized CNNs as feature extractors from point clouds captured by a depth sensor.

Despite the success of the abovementioned methods, there is still a lack of a highly accurate approach to recognize hands in uncontrolled scenarios, such as those where hands are subjected to

high occlusion, noise, poor illumination, etc. In this paper, we attempt to address these challenging problems by proposing a yolo learning-based system to localize hands and recognize hand gestures under both real-life and uncontrolled situations.

III. PROPOSED WORK:

We will be using computer vision to build an American Sign Language Translator using our device webcam as the input parameter to show output in a web application, to achieve this we use open cv, and ultralytics modules. There are two main modules proposed in our project, first one is translating hand sign gestures into text and the other is translating text into hand sign language gestures.

Necessary Resources: To successfully conduct this research, we require a good dataset with a high-quality and relevant dataset, a device equipped with a powerful GPU, a webcam to capture real-time data for analysis and processing and a development environment for Python 3.

Methodology

The research process involves several steps: Data Collection, where the required data is either manually captured or imported from online repositories; Preprocessing, which includes annotating and labelling the data for analysis; Feature Extraction, to obtain relevant features for the model; Model Training, using the YOLOv8 model to develop an accurate system; and Model Testing, to evaluate the performance and accuracy of the trained model.

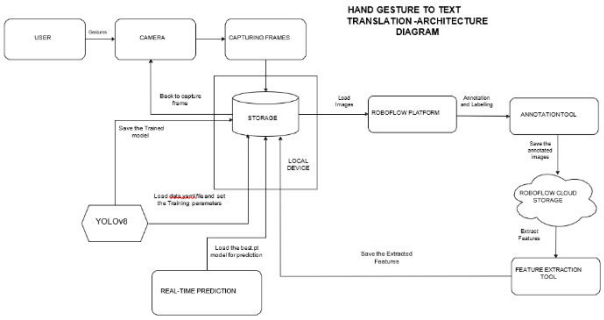


Fig. 2: System Architecture diagram

Modules:

1. Hand Sign Gesture to Text Translation:

This module captures real-time hand sign gestures using a webcam and translates them into corresponding text. The process includes several components: the Image Capture Module, the Preprocessing Module for annotating and labelling images, the Feature Extraction Module for extracting relevant features, the Training Module for training the YOLOv8 model, and the Gesture Recognition Module for recognizing and translating hand sign gestures into text output based on the trained model.

2. Text to Hand Sign Gesture Translation:

This module converts text into corresponding hand sign gestures. This module consists of three components: the Text Input Module, which captures the user's input text; the Image Retrieval Module, which fetches labelled images from the designated folder; and the Image Display Module, which displays the retrieved hand sign gesture images to the user.

3. Web Application Development:

This module integrates the above functionalities into a web-based platform, making it accessible and user-friendly. Frontend Development involves using HTML, CSS, and JavaScript to create a responsive user interface. Backend Development focuses on implementing server-side logic to handle data processing and model integration. Finally, deployment involves setting up the application on a web server to make it accessible online.

IV. RESEARCH AND DEVELOPMENT:

In the study of object-detection models, we encountered a problem in the preprocessing phase, where huge datasets are used to train the model. These large datasets need to be manually annotated and labelled, leading to extended training periods. Additionally, a high number of epochs are often used in the hope of better training the model. However, this approach carries a high risk of overfitting, where models perform well on training data but struggle with real-world applications. To address these challenges, our research focused on developing a more efficient and accurate method for American Sign Language (ASL) hand gesture recognition. Instead of relying on a huge dataset, we created a custom dataset with diverse data. In the context of our model, this means images. We decided to collect images based on four parameters:

1. Images with background noise

2. Images with good lighting

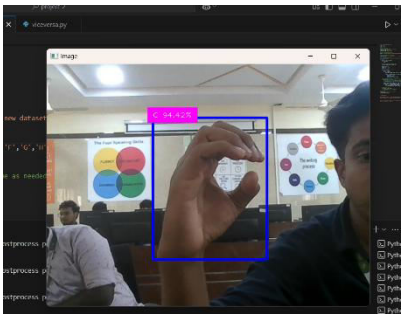
3. Images with bad lighting

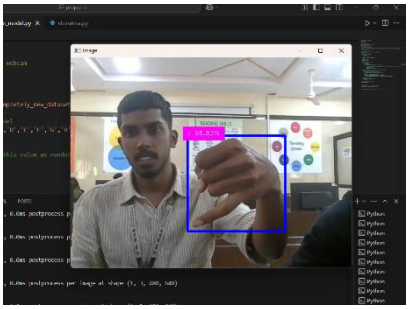
4. Images with no colour (black and white)

We trained this diverse dataset with a comparatively lower number of epochs to see how well this approach makes a difference. This approach allowed us to achieve higher accuracy compared to previous models while reducing the number of epochs needed for training.

IMPLEMENTATION:

To train and test our YOLOv8 model for ASL hand gesture recognition, we started with data preparation. We utilized the "American Sign Language Letters Object Detection Dataset" from Roboflow Universe, consisting of 720 images, and manually collected an additional 530 images for diversity. Using the Roboflow platform, we accurately annotated and labelled the hand sign gestures, organizing them into separate folders based on labels. During preprocessing, we applied auto-orientation, resized images to 640x640 pixels, normalized them, and used data augmentation techniques like horizontal flipping. Feature extraction was carried out on Roboflow, which provided us with train, test, and validation folders, each containing subfolders for images and labels, along with a data.yaml file. For training, we used the YOLOv8m model, training it for 100 epochs with a batch size of 8, achieving a balance between accuracy and computational efficiency. Finally, we tested the model's performance in real-time using the best.pt file, evaluating its effectiveness in recognizing ASL hand gestures.





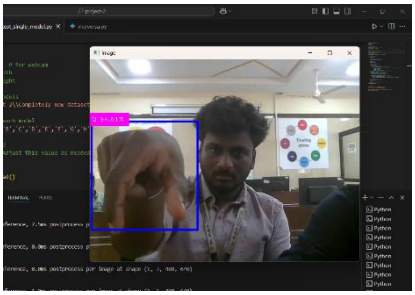


Fig. 3: Predictions and confidence score in real-time testing

V. RESULTS AND DISCUSSION:

In evaluating our scheme, we utilized criteria such as mAP50, mAP50-95, Weighted-Averaged Precision, Weighted-Averaged recall, and Weighted-Averaged F1-score.

***Training and Validation Losses**:* Over the course of 100 epochs, we observed a significant decrease in training and validation losses. The training box loss started around 3.0 and steadily decreased to approximately 0.2, while the validation box loss began at 2.0 and also dropped to around 0.2. These trends indicate effective learning and model optimization.

Performance Metrics: The performance metrics showed substantial improvement. The precision metric increased from 0.2 to 0.9, and the recall metric rose from 0.2 to 0.8 by the 100th epoch. The mAP50 metric improved from 0.2 to 0.9, and the mAP50-95 metric increased from 0.2 to 0.8, demonstrating the model's enhanced accuracy.

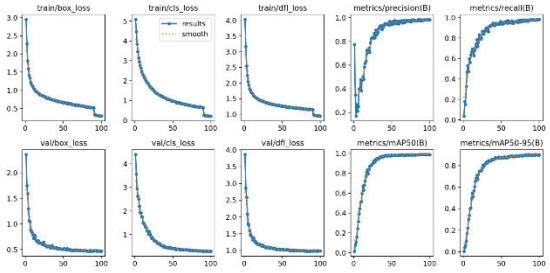


Fig. 4: Performance metrics

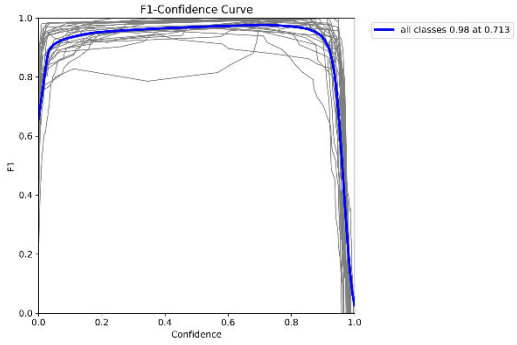


Fig. 5: F1-Confidence Curve

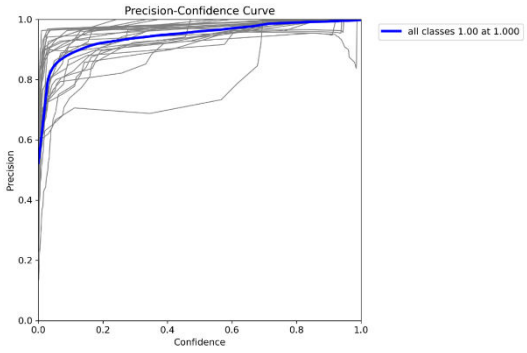


Fig. 6: Precision-Confidence Curve

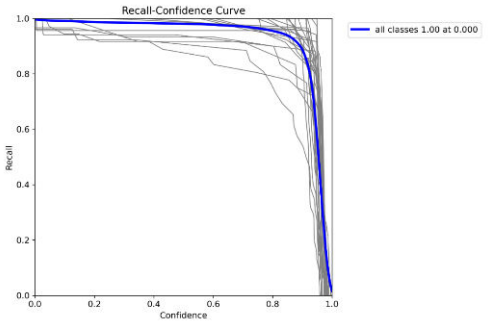


Fig. 7: Recall-Confidence Curve

Table -1
Comparison of model performance metrics:

Ref	Year	Model	Images	Precision	Recall	F1 Score	mAP 50	mAP 50-95
[13]	2021	YOLOv 5	2515	95	97	98	98	98
[14]	2022	YOLOv 4	8000	96	96	96	98.01	-
[16]	2024	YOLOv 8	29820	98	98	99	98	93
Our s	2025	YOLOv 8	2600	98.2	98.2	98.24	98.7	89

DISCUSSION:

The goal of this project was to create an efficient American Sign Language (ASL) translator using computer vision techniques. Our approach included the use of YOLOv8 for accurate hand gesture recognition. The results showed that our custom dataset, trained with fewer epochs, achieved higher accuracy compared to previous models. This improvement was reflected in the substantial increase in precision, recall, mAP50, and mAP50-95 metrics over the training period. The significant reduction in training and validation losses further demonstrated the effectiveness of our model.

Our findings indicate that a diverse dataset with images containing various conditions (background noise, different lighting, black and white) can greatly enhance the performance of an ASL hand gesture recognition model. By using fewer epochs, we were able to mitigate the risk of overfitting, leading to a more robust model that performs well in real-world applications.

Despite the success of our model, there are some limitations to consider. The custom dataset, while diverse, may still not capture all variations in hand signs that could occur in real-world scenarios. Additionally, the performance of the model in different environments and with different users has yet to be thoroughly tested. Future research should focus on expanding the dataset and evaluating the model in various real-life conditions to ensure its robustness and reliability.

VI. CONCLUSION:

In summary, this project successfully achieved its objective of developing an effective American Sign Language (ASL) translator using computer vision techniques. By utilizing a diverse dataset and reducing the number of epochs, we created a model that demonstrates high accuracy and reliability in recognizing hand signs. This ASL translator holds significant potential to address and eliminate serious communication barriers faced by the deaf and mute community, enhancing their ability to engage in day-to-day interactions and reducing feelings of isolation. Throughout the process, challenges such as background noise, lighting variations, and hand occlusions were addressed to ensure accurate and reliable performance.

REFERENCES:

[1] Real Time Sign Language Translation Systems: Maria Papatsimouli, Konstantinos- Filippos Kollias, Lazaros Lazaridis, George Maraslidis, Herakles Michailidis, Panagiotis Sarigiannidis and George F. Fragulis Department of Electrical and Computer Engineering, University of Western Macedonia, Kozani, Greece.

[2] A Survey of Advancements in Real-Time Sign Language Translators: Maria Papatsimouli, Panos Sarigiannidis, George F. Fragulis -Department of Electrical and Computer Engineering, University of Western Macedonia.

[3] Time Series Neural Networks for Real Time Sign Language Translation: Sujay S Kumar, Tenzin Vangyal, Varun saboo, Ramamoorthy Srinath-State University At New York.

[4] Real Time Translation of Sign Language to Text: Arun Chakaravathy, Ruby Mythili, Mansi K Vachhani-KGISL Institute of Technology.

[5] Design of low cost and efficient sign language interpreter for the speech and hearing impaired- Shanthi K. G, Manikandan A, Sesha Vidhya S, Venkatesh Perumal Pranay Chandragiri, Sriram T. M and Yuvaraja K. -R. M. K College of Engineering and Technology, Anna University, India.

[6] Techno-talk: An American Sign Language (ASL) Translator: Arslan Arif, Syed Tahir Hussian –Politecnico di Torino, Iqra Jawaid-Drexel University, Muhammad Walled-Riphah International University.

[7] Virtual Assistant with Sign Language: Dr. D. S. Hirolikar, Hari Shelar, Prajwal Berad, Mohini Gawade, Yogesh Pawar, Department of Information technology, PDEA's College of Engineering, Pune, India.

[8] Recognition of Tamil Sign Language Alphabet using Image Processing to aid Deaf-Dumb People: R. Subha Rajam, G. Balakrishnan.

[9] Literature survey on hand gesture techniques for sign language recognition- Ms Kamal Preet Kour, Dr. (Mrs) Lini Mathew, Department of Electrical Engineering, NITTTR, Chandigarh (India).

[10] Computer vision Trends and Challenges- Jorge Bernal, David Vazquez, Autonomous University of Barcelona.

[11] Deep Learning for Computer Vision: A Brief Review - Athanasios Voulodimos, Nikolaos Doulamis Anastasios Doulamis and Eftychios Protopapadakis -Department of Informatics, Technological Educational Institute of Athens, Athens, Greece -National Technical University of Athens,Athens, Greece.

[12] Computer Vision and Image Processing: A Paper Review - Victor Wiley, Thomas Lucas

[13] T.F. Dima, M.E. Ahmed, Using YOLOv5 algorithm to detect and recognize American sign
language, in: 2021 International Conference on Information Technology, ICIT, IEEE, 2021.

[14] A. Al-shaheen, M. Çevik, A. Alqaraghulı, American sign language recognition using yolov4 method, Int. J. Multidiscip. Stud. Innov. Technol. (2022)

[15] A. Imran, M.S. Hulikal, H.A. Gardi, Real-time American sign language detection using YOLO-v9, 2024.

[16] Transfer learning with YOLOV8 for real-time recognition system of American Sign Language Alphabet Alsharifa, Easa Alalwanyc, Mohammad Ilyasa ,2024.

[17] Dardas N.H, Georganas N.D: Real-time hand gesture detection and recognition using bag-of-features and support vector machine techniques.

[18] Mittal A, Zisserman A, Torr P: Hand detection using multiple proposals.

[19] Stergiopoulou E, Sgouropoulos K, Nikolaou N, Papamarkos N, Mitianoudis N: Real time hand detection in a complex background.

[20] Roy K, Mohanty A, Sahay R. R: Deep Learning Based Hand Detection in Cluttered Environment Using Skin Segmentation.

[21] Keskin C, Kiraç F, Kara Y. E, Akarun L: Hand pose estimation and hand shape classification using multi-layered randomized decision forests.

[22] Kang B, Tan K.H, Jiang N, Tai H.S,Treffer D, Nguyen T: Hand segmentation for hand-object
interaction from depth map.

[23] Deng X, Zhang Y, Yang S, Tan P, Chang L, Yuan Y, Wang H: Joint Hand Detection and Rotation Estimation Using CNN.

[24] Huang Y, Liu X, Zhang X, Jin L: A Pointing Gesture Based Egocentric Interaction System Dataset, Approach and Application.

[25] Le T.H.N, Zhu C, Zheng Y, Luu K, Savvides M: Robust hand detection in Vehicles.

[26] Chevtchenko S.F, Vale R.F, Macario V, Cordeiro F.R: A convolutional neural network with feature fusionfor real-time hand posture recognition.

[27] Liang C, Song Y, Zhang Y: Hand gesture recognition using view projection from point c cloud.

[28] Oyedotun O.K, Khashman A: Deep learning in vision-based static hand gesture recognition.

[29] Li Y, Wang X, Liu W, Feng B: Deep attention network for joint hand gesture localization and recognition using static RGB-D images.