Sign Language Recognition System for Service-Oriented Environment

Mohamed Mamdouh Mohamed Eldesouky Elnamla
Faculty of Electrical Engineering
Universiti Teknologi Malaysia
Johor Bahru, Malaysia
m.mohamed@graduate.utm.my

Nor Hisham Haji Khamis
Faculty of Electrical Engineering
Universiti Teknologi Malaysia
Johor Bahru, Malaysia
hisham@fke.utm.my

Nasrul Azizi Bin Nor Hisham
Faculty of Information Sciences and Technology
Management and Science University
Shah Alam, Malaysia
nasrulazizi2002@gmail.com

2024 IEEE International Conference on Advanced Telecommunication and Networking Technologies (ATNT) | 979-8-3503-5350-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ATNT61688.2024.10719250

Abstract—**This paper presents a real-time sign language recognition and translation system, designed to improve communication for the deaf and hard-of-hearing community. The system uses YOLOv5 and Convolutional Neural Networks (CNNs) to translate American Sign Language (ASL) gestures into text. Trained on 2369 images and validated with 342 images covering 49 ASL signs, the model achieved high accuracy in real-time tests. Intended for service-oriented environments, this system enhances accessibility and inclusivity by facilitating communication between sign language users and those unfamiliar with ASL.**

Keywords— Convolution Neural Network (CNN), American Sign Language (ASL), You Only Look Once (YOLO), Artificial Intelligence (AI), Artificial Neural Network (ANN).

I. Introduction

Communication barriers between the deaf and hard-of-hearing community and those unfamiliar with sign language pose significant challenges. While various assistive technologies exist, there is a need for efficient, real-time systems that accurately interpret sign language gestures into text or speech. This research addresses this gap by developing a real-time sign language recognition and translation system using advanced computer vision techniques [1].

American Sign Language (ASL) is a complex visual language, and its interpretation by non-signers is limited. It is a natural language that serves as the predominant sign language of deaf communities in the United States and many parts of the world. ASL is also widely learned as a second language, serving as a lingua franca. Existing solutions often require specialized equipment or lack real-time capabilities. This proposed system uses YOLOv5 and Convolutional Neural Networks (CNNs) for robust, real-time ASL recognition, leveraging YOLOv5's speed and accuracy. The methodology involves data collection, model training, and validation. Trained on a diverse dataset of ASL gestures, the system can recognize 49 different signs with high accuracy, making it suitable for practical use in environments like stores, restaurants, and reception areas to enhance accessibility and inclusivity.

This research contributes to:

1) Developing a real-time ASL recognition system using YOLOv5 and CNNs.
2) Implementing a practical model to improve communication for the deaf and hard-of-hearing community.

This work aims to facilitate better interaction and contribute to a more inclusive society by addressing sign language recognition challenges.

II. Literature Review

A. Importance of Sign Language Recognition

Sign language is the primary means of communication for the deaf and hard-of-hearing community, employing a combination of hand gestures, facial expressions, and body movements to convey meaning. The World Federation of the Deaf estimates that there are over 70 million deaf individuals globally, utilizing more than 300 different sign languages [1]. Sign language recognition (SLR) technology aims to bridge the communication gap between sign language users and those who do not understand it, thereby fostering inclusive communication and aiding in the social and professional integration of deaf individuals [2].

B. Current Technologies and Methods

Advancements in Artificial Intelligence (AI), particularly in Machine Learning (ML) and Deep Learning (DL), have significantly contributed to the development of SLR systems [2]. These technologies enable the automatic recognition and translation of sign language into text or speech, enhancing accessibility and communication for the deaf community [3].

C. Convolutional Neural Networks (CNNs)

CNNs are a type of deep learning algorithm particularly effective in image recognition tasks. They have been widely used in SLR for detecting and classifying hand gestures and facial expressions. CNNs consist of multiple layers that process and extract features from images, making them suitable for recognizing complex patterns in sign language [3

D. You Only Look Once (YOLO)

YOLO is a real-time object detection system that divides images into a grid and predicts bounding boxes and class probabilities for each grid cell. YOLOv5, an iteration of this algorithm, is known for its high speed and accuracy in detecting objects, making it ideal for real-time SLR applications. YOLOv5's ability to process images quickly and accurately is crucial for translating dynamic sign language gestures [4].

E. Techniques in Sign Language Recognition

SLR systems employ various techniques to capture and interpret the gestures involved in sign language [2]. These include:

979-8-3503-5350-1/24/$31.00 ©2024 IEEE

1) 2D and 3D Modeling:

- 2D Modeling: Utilizes two-dimensional images to recognize hand shapes and movements. While effective, it can struggle with recognizing gestures in complex backgrounds or varying lighting conditions.
- 3D Modeling: Incorporates depth information, providing a more accurate representation of gestures. This approach is more robust in dynamic environments but requires more computational resources.

2) Machine Learning Approaches:

- Supervised Learning: Involves training models on labeled datasets where the correct output is known. This approach is effective for tasks like hand shape recognition but requires extensive annotated data.
- Unsupervised Learning: Learns patterns from unlabeled data, useful for discovering new gestures and patterns without prior knowledge [5].
- Reinforcement Learning: Models learn by interacting with the environment, and receiving feedback on their actions, which is particularly useful for dynamic and continuous gesture recognition [5].

F. *Challenges and Limitations*

Despite the advancements, several challenges remain in developing effective SLR systems:

1) Data Scarcity:

Large, annotated datasets are crucial for training accurate models. However, the availability of such datasets for sign language is limited, hindering the development and evaluation of SLR systems [6].

2) Real-Time Processing:

Achieving real-time performance is essential for practical applications but remains challenging due to the computational complexity of recognizing and translating gestures accurately and quickly.

3) Variability in Sign Language:

Different regions use different sign languages, and even within the same language, there can be significant variation in how signs are performed. SLR systems must be adaptable to these variations to be effective in diverse settings [2].

III. Methodology

A. *Data Collection and Labeling*

The dataset for this project consists of 2369 training images and 342 validation images, encompassing 49 distinct American Sign Language (ASL) signs. The data collection process involved capturing images of various hand gestures representing numbers (0-9), the English alphabet, and 13 common words. To ensure consistency and accuracy, images were taken under controlled lighting conditions using a standard camera setup. For labeling the dataset, the “makesense” platform was utilized. Each image was categorized and labeled to create a comprehensive dataset for training the recognition model.

B. *Model Development*

The model development process involved organizing labeled images into training and validation sets, training YOLOv5 model on Google Colab using these datasets, and continuously evaluating the model's performance using validation metrics such as loss values, precision, and recall. The training parameters were optimized to enhance the model’s accuracy and efficiency in real-time sign language recognition. The stages of this process are shown in Figure 1.

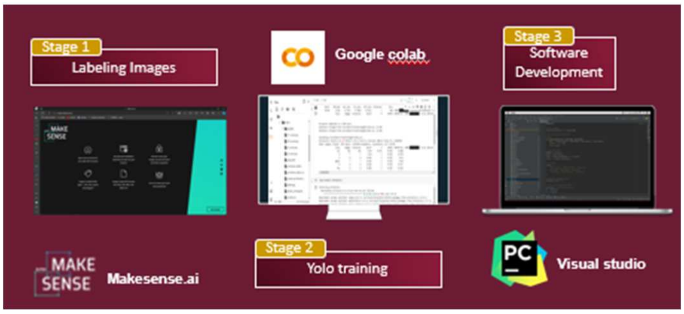

Fig. 1. Model development stages.

C. *System Implementation*

The implemented system comprises several interconnected components: a webcam to capture real-time video feeds, a YOLOv5 model to detect and recognize hand gestures from the video frames, and a text-to-speech engine to translate recognized gestures into readable text and spoken words. The system is designed for real-time processing, ensuring minimal latency between gesture detection and translation output.

D. *Tools and Platforms*

Several tools and platforms were essential in the development of the sign language recognition system:

- Visual Studio: Used for software development and code management.
- Google Colab: Provided the computational resources for training the YOLOv5 model.
- Makesense.ai: Facilitated the annotation and labeling of the dataset.
- YOLOv5: The core object detection algorithm used for recognizing hand gestures.
- Python: The primary programming language used for developing the recognition and translation system.

E. *Flowchart*

The overall workflow of the system can be summarized in the following flowchart:

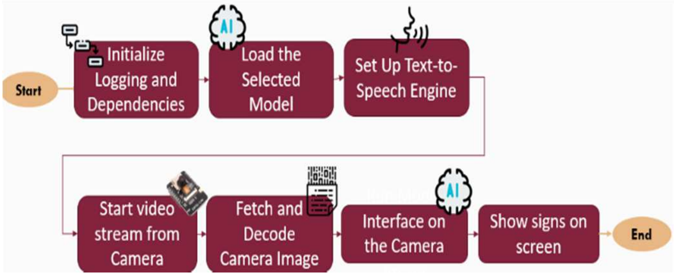
Start
Initialize Logging and Dependencies
AI
Load the Selected Model
Set Up Text-to-Speech Engine
Start video stream from Camera
Fetch and Decode Camera Image
AI
Interface on the Camera
Show signs on screen
End

Fig. 2. System workflow.

Figure 2 shows the system workflow for real-time sign language recognition and translation. The process starts with initializing logging and dependencies, loading the YOLOv5 model, and setting up the text-to-speech engine. The system captures the video stream, processes each frame, and uses the YOLOv5 model to detect hand gestures. Recognized signs are displayed on the screen and translated into text and speech, ensuring real-time recognition and translation of sign language gestures.

IV. Results

A. *Model Performance Metrics*

Figure 3 illustrates the key performance metrics of the YOLOv5 model, including training and validation loss curves, as well as precision and recall curves. During the training phase, the model's performance was monitored using various loss metrics. The train/box_loss measures the error in predicting the coordinates of bounding boxes during training, while the train/obj_loss represents the error in predicting the presence of an object within a bounding box. Additionally, the train/cls_loss indicates the error in predicting the class of an object within a bounding box. These images correspond to 49 different signs, resulting in 49 unique classes.

For validation, similar metrics were used to assess the model's performance. The val/box_loss measures the error in predicting bounding box coordinates during validation, the val/obj_loss represents the error in predicting the presence of an object within a bounding box, and the val/cls_loss indicates the error in predicting the class of an object within a bounding box. These metrics collectively highlight the model's robust performance in recognizing and classifying ASL gestures, demonstrating its potential for practical applications.

Precision and recall curves further evaluate the model's effectiveness. Precision is the ratio of correctly predicted positive observations to the total predicted positives, reflecting the accuracy of positive predictions. Recall is the ratio of correctly predicted positive observations to all actual positives, indicating how well the model captures all relevant instances of a particular class. These metrics are critical for evaluating the model's effectiveness in recognizing and classifying ASL gestures.

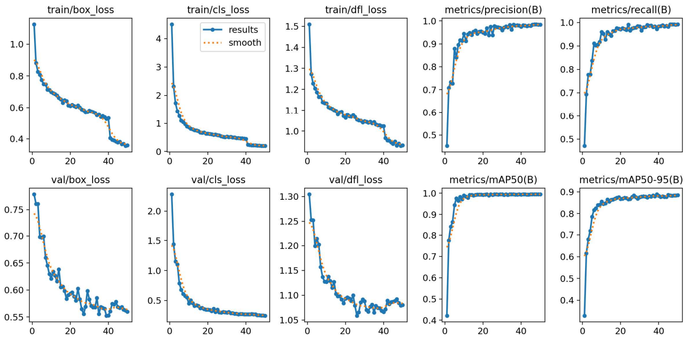

Fig. 3. Model Performance Metrics

The results shown in Figure 3 indicate that the consistency of losses is optimal around epoch 45 which indicates the training is adequate for the system to predict correctly the hand gestures. This is also reflected in the metrics graphs which show the high accuracy of the model which can detect objects with at least 50% overlap with the ground truth, also ensuring the model's robustness across different Intersection over Union (IoU) thresholds.

B. *Confusion Matrix*

Figure 4 shows the confusion matrix for the YOLOv5 model, providing insights into the model's performance across different classes. The matrix highlights true positives (TP) where gestures are correctly recognized, false positives (FP) where gestures are incorrectly recognized, true negatives (TN) where non-gesture inputs are correctly rejected, and false negatives (FN) where gestures are missed. The confusion matrix helps in identifying specific classes where the model excels or needs improvement, guiding further refinements and training adjustments.

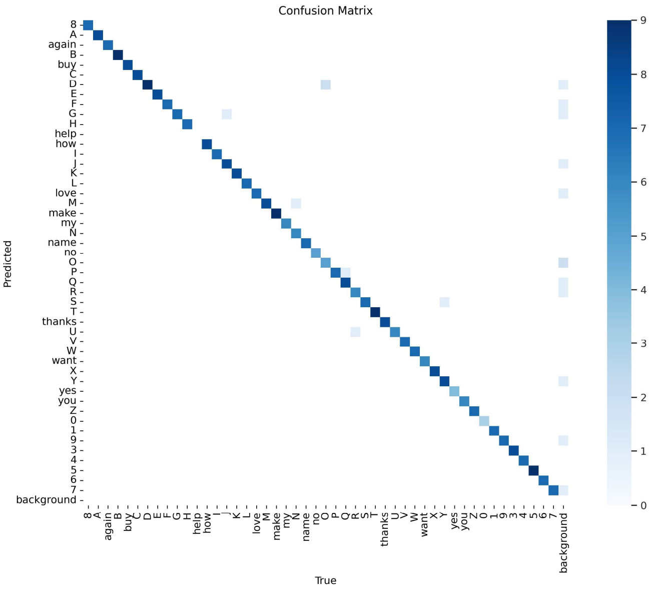

Fig. 4. Confusion Matrix

Figure 4 illustrates the output confusion matrix for the trained YOLO model, demonstrating exceptional performance across all 49 classes. The high number of true positives (TP) indicates the model's precise and accurate identification and classification of objects. The minimal occurrence of false positives (FP) and false negatives (FN) highlights the model's ability to avoid misclassifications and accurately detect objects present in the ground truth. The presence of true negatives (TN) further underscores the model's capability to correctly identify areas without objects. The high scores achieved by each class underscore the reliability and robustness of the YOLO model, showcasing its effectiveness in accurately classifying and precisely locating objects across a diverse range of categories

C. *YOLO Training Outputs*

Figure 5 displays the output images from the YOLO training process on Google Colab. These images show the detection and recognition of various ASL gestures by the trained model. Each gesture is identified and bounded with boxes, indicating the model's ability to accurately recognize different signs in real time. This visual confirmation of the model's performance is crucial for validating its effectiveness in practical applications.

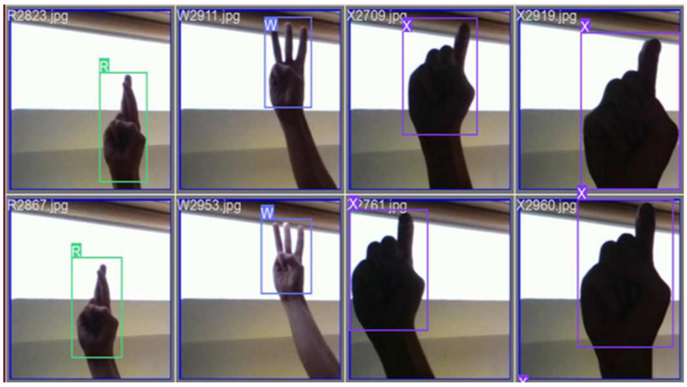

Fig. 5. YOLO training Output

V. Discussion

The results demonstrate the effectiveness of the YOLOv5 model in real-time American Sign Language (ASL) recognition. High precision and recall rates indicate robust performance in correctly identifying and classifying hand gestures, essential for practical applications to enhance communication for the deaf and hard-of-hearing community.

Training and validation loss curves show minimal error in predicting bounding box coordinates, object presence, and class labels, highlighting the model's efficient learning and generalization capabilities. The confusion matrix reveals that while the model performs well, certain classes may benefit from further dataset augmentation or parameter fine-tuning to improve accuracy.

The YOLO training outputs visually confirm the model's ability to detect and recognize ASL gestures accurately, demonstrating its practical applicability in real-world environments like stores, restaurants, and reception areas.

Challenges such as variability in lighting conditions and limited training data need to be addressed in future work. Expanding the dataset and incorporating other sign languages can enhance the system's versatility. Integrating the system with wearable devices or mobile applications, along with user feedback, can further improve its usability and functionality.

In conclusion, the YOLOv5-based sign language recognition system shows promising results. With further refinements, it can significantly aid in bridging communication gaps for the deaf and hard-of-hearing community.

VI. Conclusion

The YOLOv5-based real-time sign language recognition system demonstrates robust performance in recognizing and translating American Sign Language (ASL) gestures into text and speech. High precision and recall rates, along with minimal errors in training and validation loss, highlight the model's effectiveness and practical applicability. Future work will focus on addressing variability in real-world conditions, expanding the dataset, and incorporating additional sign languages to further enhance the system's versatility and usability. This system has the potential to significantly improve communication for the deaf and hard-of-hearing community, fostering greater inclusivity and accessibility.

REFERENCES

[1] World Federation of the Deaf. 2018. Our Work. (2018). Available: http://wfdeaf.org/our-work/. Accessed 2019-03-26.

[2] Zhang, L., Tjondronegoro, D., & Chandran, V. (2019). A comprehensive review of recent advances in gesture recognition. Journal of Computer Science and Technology, 29(4), 641-659. doi:10.1007/s11390-019-1952.

[3] K. G. Kim, "Book Review: Deep Learning," Healthc Inform Res, vol. 22, no. 4, p. 351, 2016, doi: 10.4258/hir.2016.22.4.351.

[4] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," Apr. 2018, [Online]. Available: http://arxiv.org/abs/1804.02767.

[5] B. Mahesh, "Machine Learning Algorithms," International Journal of Science and Research, 2018, doi: 10.21275/ART20203995.

[6] Rastgoo, R., Kiani, K., & Escalera, S. (2020). Sign language recognition: A deep survey. Expert Systems with Applications, 158, 113545.

[7] Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. Available: http://arxiv.org/abs/1409.1556.

[8] Murthy, G. R. S., & Jadon, R. S. (2020). A review of vision-based hand gestures recognition. International Journal of Information Technology, 12, 731–743. doi:10.1007/s41870-019-00373-3.

[9] Thuan, D. (2021). Evolution of Yolo algorithm and Yolov5: The State-of-the-Art object detection algorithm.