Translation of American Sign Language to English using a 3-Model Architecture

Neil Rojan, Yaseen Mohemed, Atul Sunny Thomas, Mohmmed Shameem, Mr. Ajith Jacob
Information Technology, Rajagiri School of Engineering & Technology
Asst. Professor, Information Technology, Rajagiri School of Engineering & Technology

Abstract — There are 70 million people worldwide who are hearing impaired or deaf. Unlike most of us, their first language is Sign Language. One of the most common ones is ASL. American Sign Language is used as a standard in most places and many sign languages borrow from it. However, the longstanding challenge of communication accessibility faced by the deaf and hard-of-hearing community has not been addressed. In an increasingly digital world, equitable communication remains a fundamental concern. Our project aims to bridge the communication gap by harnessing the power of machine learning and computer vision. Our project's goal is to translate ASL to English, using a 3-model architecture to break up the translation process into 3 processes. The first model uses object detection to identify hands, then Image Classification to identify the letter used, and then finally Natural Language Processing to string together the letters to sentences. Our project has several User profiles. The ASL user, who we are translating, The English Speaker who we are translating for, and the Admin who manages the application. We plan to use several technologies Pytorch for model creation, specifically the Object detection, Image Classification and Natural Language Processing Suites.

Index Terms — Convolutional Neural Networks, American Sign Language, Machine Learning, Natural Language Processing, Computer Vision, Sign Language Translation, Object Detection, ResNet-50

I. INTRODUCTION

Sign language is a vital communication method for millions of Deaf and hard-of-hearing people, yet translating it into English remains complex. This paper introduces a novel 3-model architecture designed to bridge this communication gap by translating American Sign Language (ASL) into English.

The first model uses Convolutional Neural Networks (CNNs) and computer vision to detect and isolate hand movements from video input. This step focuses on identifying relevant gestures, laying the groundwork for accurate translation. The second model applies image classification, relying on advanced architectures like ResNet-50, to identify specific ASL signs, enhancing accuracy and reducing errors.

The third model leverages natural language processing (NLP) to convert recognized signs into coherent English text. It integrates custom correction maps and grammar-checking frameworks to ensure accurate and meaningful translations.

Our paper details the design of this architecture and discusses the training process, data preparation, and experimental results. We also address challenges such as variability in ASL expressions and outline the steps taken to overcome them. This approach shows promise in improving communication between ASL users and English speakers, with potential applications in education, healthcare, and customer service.

II. PROBLEM DEFINITION

The communication barrier between the deaf and hearing communities persists, posing significant challenges in accessibility, inclusivity, and understanding. With an estimated 70 million people worldwide relying on American Sign Language (ASL), there's a growing need for effective translation tools to facilitate cross-community communication. Traditional solutions struggle with accuracy and scalability, often leading to misinterpretations. Our project addresses this gap by introducing a 3-model architecture that combines object detection, image classification, and natural language processing to accurately translate ASL to English. By breaking down the translation process into distinct stages, we aim to create a more

accessible communication bridge for the deaf community.

III. PROPOSED SOLUTION

Our project aims to bridge the communication gap between the deaf and hard-of-hearing community and those who use spoken and written languages. The proposed solution involves a three-model architecture[1] to translate American Sign Language (ASL) into coherent English text. This architecture combines advanced computer vision, image classification, and natural language processing (NLP) techniques to create a comprehensive ASL-to-English translation system.

1. Three-Model Architecture

Hand Detection Model:
The first model uses convolutional neural networks (CNNs)[2] to detect hands in ASL videos. It applies object detection algorithms to identify hand positions and movements within the video frames. This model's output is the location of hands, which is essential for subsequent processing.

ASL Letter Classification Model:
The second model takes the output from the hand detection model and classifies the ASL letters. It uses CNNs and ResNet-50[3] architectures to identify specific hand gestures and their corresponding letters. This model translates visual hand movements into individual ASL letters, forming the basis for further translation.

Natural Language Processing (NLP) Model:
The third model focuses on transforming the classified ASL letters into coherent English sentences. It employs NLP techniques[4] to correct spelling, address grammatical issues, and create meaningful sentences from the classified letters. The final output is a structured English text representing the translation of ASL signs.

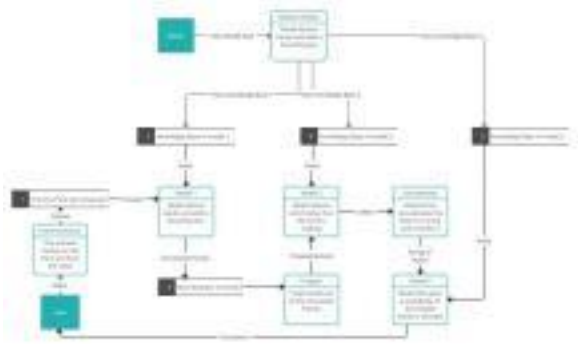

Fig .1. Data Flow Diagram of the proposed solution

2. Data Flow Diagram

The data flow diagram (DFD) illustrates the sequence of data processing steps in the proposed solution. It begins with video input containing ASL signs, processed through the hand detection model, followed by the ASL letter classification model, and culminating in the NLP model to produce English text. Here's a high-level outline of the DFD:

1. Video Input: ASL videos are uploaded to the system for processing.
2. Hand Detection: The first model detects hand positions and movements in the video frames.
3. ASL Letter Classification: The second model classifies the detected hand movements into specific ASL letters.
4. Text Transformation: The NLP model converts the classified ASL letters into meaningful English text.
5. Output: The final output is the translated English text representing the ASL signs.

The proposed solution's architecture effectively addresses the complexities of ASL translation, offering a scalable and adaptable framework for bridging the communication gap. By integrating computer vision, image classification, and NLP, the system provides a robust approach to translating ASL into English, fostering greater accessibility and inclusivity.

IV. METHODOLOGY

The methodology for this project revolves around a 3-model architecture designed to translate American Sign Language (ASL) to English. Each model contributes distinct functionalities, combining computer vision, image classification, and natural

language processing to achieve accurate translation. The following sections explain the methodology for each model and the techniques used to implement them.

1. Model 1 & 2: Computer Vision for Hand Detection and ASL Letter Recognition

The first and second model focus on detecting and identifying hand movements from video input and classifies the detected hands into ASL letters or numbers.. Model 1 employs a convolutional neural network (CNN) architecture with a Faster R-CNN framework to extract hand positions and generate bounding boxes around them. Model 2 uses image classification techniques to determine which ASL sign corresponds to the detected hand.

Techniques:

- Convolutional Neural Network (CNN): This technique uses multiple convolutional layers to detect patterns in the image. It is responsible for learning features like edges and shapes, which are then used to identify hand movements.
- Region-based Convolutional Neural Network (Faster R-CNN): This method integrates a Region Proposal Network (RPN) to generate candidate regions (bounding boxes) for detected hands.
- ResNet-50: The backbone architecture for the CNN, designed to process images with a high level of accuracy and depth.
- Data Augmentation: Data augmentation is used to enhance the generalization of hand detection in American Sign Language (ASL). This approach involves various techniques such as rotations, flips, and other transformations to simulate different hand positions and orientations. By augmenting the training dataset with these variations, the model can better generalize to real-world conditions, improving its accuracy when detecting hands in diverse contexts.
- Data Labelling: Each image is manually labelled with the corresponding ASL sign, forming the training dataset.

Code Features:

- Data Preprocessing: Video frames are extracted and pre-processed to standardize dimensions and normalize pixel values.
- Training and Inference: The model is trained on a labelled dataset of hand positions, with regularization techniques to prevent overfitting. Inference is used to generate bounding boxes around detected hands.
- Image Cropping: Based on the output from Model 1, the detected hand regions are cropped for classification.
- Model Training: The model is trained with labelled data, incorporating data augmentation to simulate various hand gestures and positions.

2. Model 3: Natural Language Processing for Sentence Construction

The third model constructs coherent English sentences from the classified ASL letters. This model employs natural language processing (NLP) techniques to translate sequences of ASL signs into structured text.

Techniques:

- Correction Map: Model 3 employs a correction map to address common errors in text translation. The correction map acts as a dictionary containing typical misspellings, contractions, and other frequent language errors, allowing the model to correct known issues before advanced spell-checking and grammar correction. The correction map[5] effectively addresses common spelling errors and homophones, reducing the risk of misinterpretation and ensuring accurate output.
- SpellChecker: This module checks and corrects individual word spelling, ensuring accurate translation.
- Language Tool Python: This open-source library is used to check grammar, tense, and punctuation to create coherent sentences.

Code Features:

- Text Processing: The output from Model 2 is processed to identify word sequences and correct common errors.
- Grammatical Correction: Using Language Tool Python, the model checks for grammatical consistency and refines the sentence structure.
- Sentence Construction: This process converts the classified ASL letters into coherent English sentences.
- Final Output: The constructed sentences can be displayed as text.

V. RESULTS

The application was designed to accurately interpret American Sign Language (ASL) signs and convert them into coherent English text. This section discusses the results of our testing process and highlights the key findings from each stage of the translation process.

The initial stage of the application involves detecting hand positions and motions in the uploaded ASL video. This is achieved using Faster Region-based Convolutional Neural Network (Faster R-CNN) with ResNet-50 as the backbone architecture. The use of this object detection model ensures a high level of precision in identifying hands, even with subtle gestures or movements.

During testing, the hand detection model exhibited over 95 percent probability of hand occurrence in various frames, indicating a robust ability to capture hand positions across a range of ASL expressions. This high accuracy in detecting hands is crucial for the subsequent stages of the translation process.

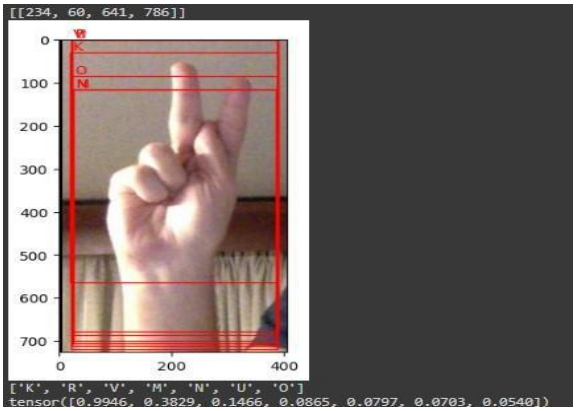

Fig.2. Example of Hand Detection with Resultant Letter Probabilities

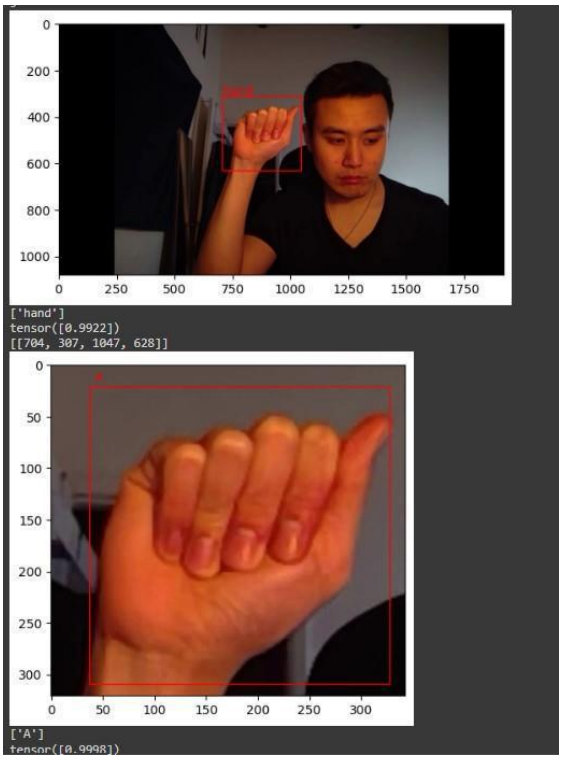

Fig .3. Example of Hand Detection with Resultant Letter Probabilities

These examples demonstrate how the model can accurately isolate hands from the video frames and assess the potential ASL letters based on the detected gestures.

After detecting the hands, the model proceeds to classify the ASL signs using an image classification process. This stage involves using a Convolutional Neural Network (CNN) based on the ResNet-50 architecture. The model assigns probabilities to each possible ASL letter or number based on the detected hand gestures.

The application is able to provide multiple probabilities for the likely ASL letter as can be seen from Fig. 2 and Fig. 3. This approach accounts for ambiguities in hand gestures and allows the system to consider various possible interpretations before determining the final output. The output from the hand detection model serves as the input for this classification process, ensuring continuity and accuracy.

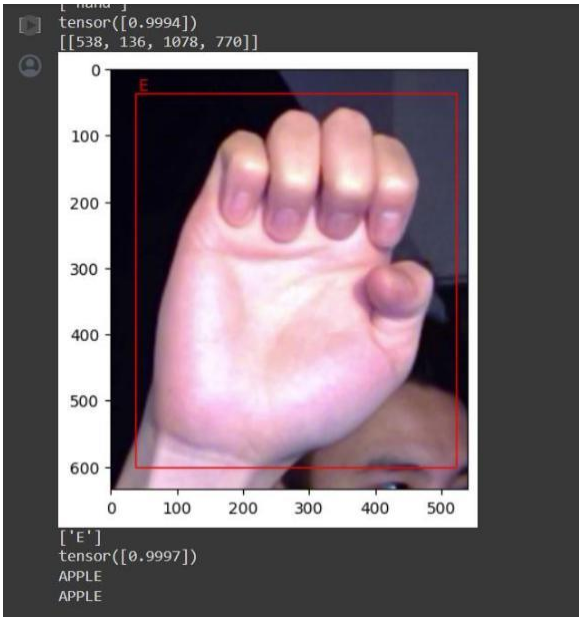

Fig.4. Example of word formation from test input.

The final step in the application involves translating the identified ASL letters into coherent English sentences. This is where the natural language processing (NLP) model plays a critical role. It uses a combination of correction maps and grammar-checking frameworks to ensure accurate and meaningful translations. The NLP model takes the output from the image classification stage and constructs sentences that reflect the intended message of the ASL user as can be seen from Fig.4.

VI. APPLICATIONS & FUTURE SCOPE

1. Applications

The versatility of this, opens doors to a wide range of applications, positively impacting various aspects of daily life. Below are some key areas where it can be utilized:

- Education: Facilitating communication between deaf or hard-of-hearing students and educators, creating an inclusive learning environment.
- Workplace Communication: Helping bridge communication gaps among colleagues, fostering a more inclusive workplace culture.
- Healthcare Services: Enhancing communication between healthcare professionals and patients with hearing impairments, ensuring comprehensive and accurate patient care.
- Customer Service: Improving customer interactions by enabling seamless communication between service providers and customers with hearing disabilities.
- Emergency Services: Supporting effective communication during emergencies, providing assistance to individuals with hearing impairments.
- Entertainment Industry: Offering sign language interpretation for live events, broadcasts, or streaming services, broadening accessibility for the deaf and hard-of-hearing community.
- Public Transportation: Facilitating communication between passengers and staff, particularly for those with hearing impairments.
- Legal Interactions: Assisting communication among legal professionals, clients, and witnesses with hearing disabilities, ensuring equitable participation in legal processes.

These applications demonstrate the transformative potential of this application, illustrating how it can impact various industries and contribute to a more inclusive society.

2. Future Scope

The future scope presents numerous opportunities for further development and enhancement. The following areas could be explored to expand the capabilities:

- Real-Time Translation: Implementing real-time translation capabilities for dynamic interactions, enhancing the versatility of the application.
- Multilingual Support: Adding support for multiple sign languages to extend the application's reach to different linguistic and cultural contexts.
- Integration with Wearable Devices: Integrating with smart glasses or wristbands for a portable, hands-free experience, improving convenience and usability.
- Continuous Machine Learning Advancements: Investing in ongoing research to improve the accuracy and efficiency of the machine learning models,

keeping pace with evolving sign language gestures.

- Gesture Customization: Allowing users to customize sign language gestures, providing adaptability to individual preferences and dialects.
- Cybersecurity Measures: Ensuring robust cybersecurity to safeguard user data, protecting privacy and security within the platform.

These future developments offer a promising outlook for this application, demonstrating its potential to evolve and adapt to changing needs and technologies.

VII. CONCLUSION

In conclusion, the "ASL to English Translation using 3 Model Architecture" project represents a significant advancement in bridging communication gaps for the deaf and hearing-impaired community. By employing a three-model architecture that integrates hand detection, ASL letter identification, and natural language processing, the project offers a comprehensive and innovative solution for translating ASL videos into coherent English text.

The project's architecture proved effective in accurately processing uploaded ASL video frames and converting them into meaningful English sentences. This advancement has the potential to enhance communication accessibility for an estimated 70 million people worldwide who rely on sign language. Additionally, the project's emphasis on addressing challenges such as variability in ASL expressions and processing constraints demonstrates a thoughtful and inclusive approach to ASL-to-English translation.

While there is still work to be done, particularly in real-time processing and user interface development, this project provides a promising foundation for future advancements in ASL translation technology. By continuously refining the existing models and exploring additional enhancements, the project can contribute significantly to a more inclusive and accessible world for those who depend on ASL for communication.

ACKNOWLEDGMENT

The authors thank and express gratitude towards the support provided by our mentor who guided us in all respects and helped us in achieving the desired outcome.

REFERENCES

[1] N. C. Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden, "Neural sign language translation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[2] Y. Pei, Y. Huang, Q. Zou, X. Zhang, and S. Wang, "Effects of Image Degradation and Degradation Removal to CNN-Based Image Classification," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 4, pp. 1239–1253, Apr. 2021.

[3] X. Bi, J. Hu, B. Xiao, W. Li, and X. Gao, “IEMask R-CNN: Information-Enhanced Mask R-CNN,” IEEE Transactions on Big Data, vol. 9, no. 2, pp. 688–700, Apr. 2023.

[4] P. Danenas and T. Skersys, “Exploring Natural Language Processing in Model-To-Model Transformations,” IEEE Access, vol. 10, pp. 116942–116958, 2022.

[5] J. Hu, Y. Liu, K.-M. Lam, and P. Lou, “STFE-Net: A Spatial-Temporal Feature Extraction Network for Continuous Sign Language Translation,” *IEEE Access*, vol. 11, pp. 46204–46217, 2023.