Generating High Quality Egocentric American Sign Language Data

John Houser
Department of Computer Science
University of California, Los Angeles
Los Angeles, USA
johnhouser@cs.ucla.edu

Abstract—**American Sign Language (ASL) is a natural language that is widely used in the Deaf and Hard-of-Hearing community, but not in the hearing community. As such, a gap remains between the two communities. While many AI/ML researchers are actively working on ASL – English translations with the aim of bridging the gap, much of these efforts face the challenge of data scarcity. To address this challenge, we have developed a solution to convert existing ASL instruction videos into ML datasets with the aim of fueling ongoing research. Our pipeline employs a pose estimation module through Google's Mediapipe to extract key points from ASL instruction videos, and then perform various augmentation techniques to prepare the data for training. Specifically, with the aid of computer vision, we identify the gestures required to sign common words in ASL and apply error correction techniques to improve data quality. Additionally, we address the issues resulting from the multiple perspectives found in the signing datasets by implementing a workflow to automatically convert third-person perspective data into egocentric ones. Finally, we apply various fixes that address inaccuracies in the pose estimation model to provide more accurate and precise training data. We hope that the work done in this capstone project provides a strong foundation for building models for ASL – English translation and bolsters ongoing research in this field.**

I. Introduction

American Sign Language (ASL) is a popular communication method in the United States and Canada for those with hearing disabilities. It allows users to express themselves and interact with others. In the United States alone, more than a half a million people use ASL as their primary natural language [1]. The growing interest in bridging the communication gap between those in the ASL community and those who are not has led to significant advancements in technology aimed at translating ASL into spoken or written English and vice versa. However, one of the primary challenges in developing reliable ASL translation systems is the scarcity of high-quality, annotated datasets. Traditional methods of data collection and annotation for ASL are labor-intensive and time-consuming, which limits the availability of datasets needed for training robust machine learning models.

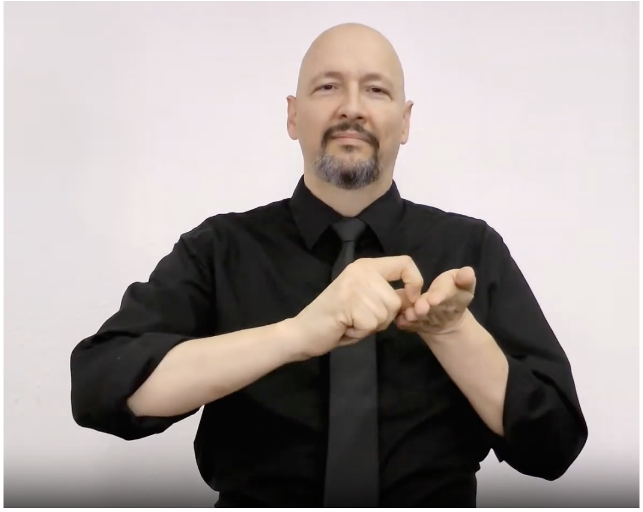

Fig. 1. An example of an ASL instruction video for the word "write" [2].

To address this challenge, our project leverages existing ASL instructional videos to create a rich dataset that can fuel ongoing research and development in ASL-to-English translation technologies. By employing tools such as Google's Mediapipe for pose estimation, we extract detailed key points from these videos, focusing on capturing the intricacies of ASL gestures. Additionally, we apply data augmentation techniques to mitigate inaccuracies in Mediapipe's pose estimation. This approach not only enhances the quantity of available data but also ensures that the data is precise and relevant, providing a strong foundation for the development of effective ASL translation models.

II. Background

A. American Sign Language

ASL is not closely related to English and not a direct translation of English. Rather, ASL is an independent natural language more closely related to spoken Japanese or Navajo. In contrast to most languages, ASL is a visual language, which relies on body movements and signs instead of spoken words to convey messages [1]. Additionally, ASL is not just a collection of gestures but a fully developed natural language with its own pronunciation, word order, and word formation, enabling complex and nuanced communication. Its visual-spatial modality is distinct from the auditory-oral modality of spoken languages, which often presents a barrier between

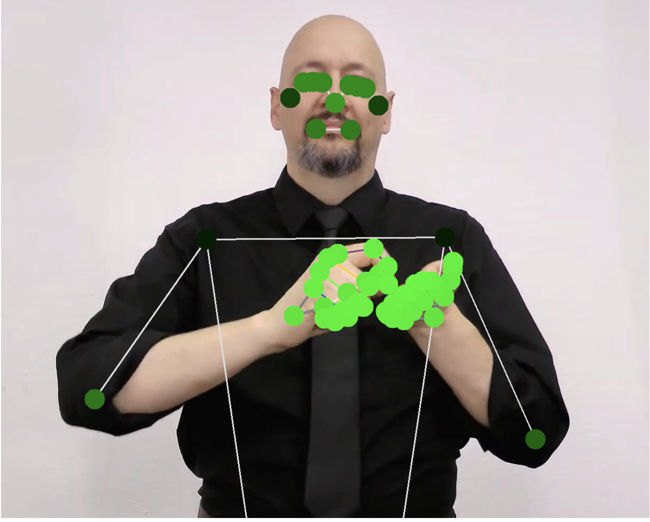

Fig. 2. Applying Google's Mediapipe pose estimation on an ASL instruction video for the word "write."

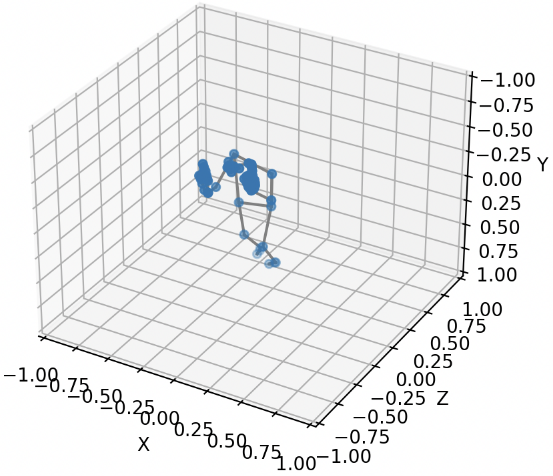

Fig. 3. A 3D mapping of the coordinates to world space for the word "write."

the Deaf and Hard-of-Hearing community and the hearing population [3].

B. Instruction Videos

The ASL community benefits from a substantial collection of instructional videos available online. Internet users have created both premium and free ASL content to facilitate language learning. Platforms such as YouTube host numerous free instructional videos, making ASL accessible to a wide audience. For this capstone project, we utilized some of these videos as the source of our data. For instance, Dr. Bill Vicars and other content creators have uploaded videos demonstrating the signs for many words in the ASL vocabulary. Figure 1 shows a frame from one of Dr. Vicars' videos used in our research [2]. By leveraging these instructional videos, we were able to generate comprehensive training data for a large number of ASL words, supporting the development of ASL-to-English translation models.

C. Google's Mediapipe

Google's Mediapipe is an open-source framework designed to support the development and deployment of multimodal machine learning pipelines. Mediapipe offers a robust platform for building real-time perception applications by providing a comprehensive set of tools and libraries for processing and analyzing video data. The framework supports various functionalities, including face detection, hand tracking, pose estimation, and object detection, making it an invaluable resource for developers and researchers in the computer vision and machine learning domains [4].

One of the key strengths of Mediapipe is its ability to perform real-time pose estimation with high accuracy and efficiency. The pose estimation module can detect and track key points on the human body, face, and hands, providing detailed information about the positions and movements of different body parts. This capability is particularly beneficial for applications that require precise gesture recognition and tracking, such as those involving ASL.

In the context of our project, Mediapipe's pose estimation module plays a critical role in converting existing ASL instruction videos into machine learning datasets. By extracting key points from the videos, we can capture the essential gestures and movements required for ASL. The accuracy and real-time processing capabilities of Mediapipe ensure that the extracted data is relevant for training machine learning models. Additionally, the framework's flexibility allows us to implement various data augmentation techniques, enhancing the diversity and robustness of our datasets.

III. Methods

In order to generate high fidelity ASL data, we utilize Google's Mediapipe to track the gestures of the subject. Using this method, we can get the position of keypoints on the subject including their eyes, shoulders, fingers, and various other points. Since the videos we used were taken facing the signer, the extracted points are from a third person perspective. During some of the videos, the hands become obstructed causing Mediapipe to mislabel them. We apply some augmentation techniques to correct this error. Then, we apply a rotation to the extracted points to make them egocentric. Finally, we apply these steps to all of the videos to get a comprehensive set of ASL data in an egocentric view.

A. Pose Estimation with Mediapipe

To use Mediapipe with an input video, a video file is entered into the Mediapipe pipeline. The framework processes each frame of the video, applying its pose estimation models to identify and track specific landmarks on the human body, face, and hands. This real-time analysis facilitates the extraction of precise positional data, which is essential for understanding the gestures being performed. The resulting data can then be

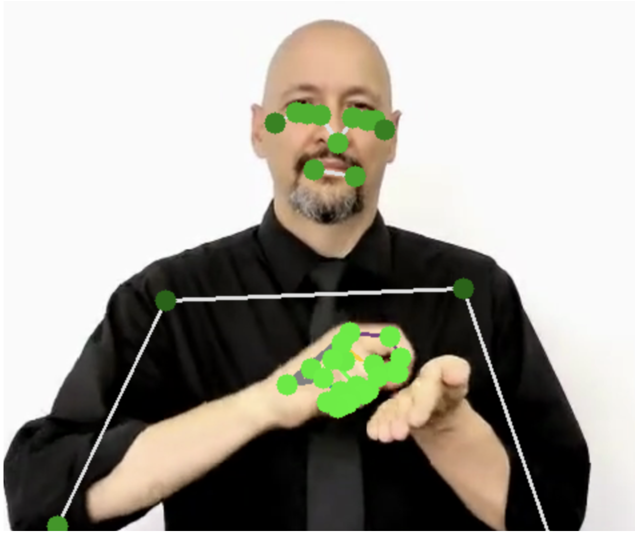

Fig. 4. Google's Mediapipe pose estimation failed to identify the left hand in this frame on the ASL instruction video for the word "write."

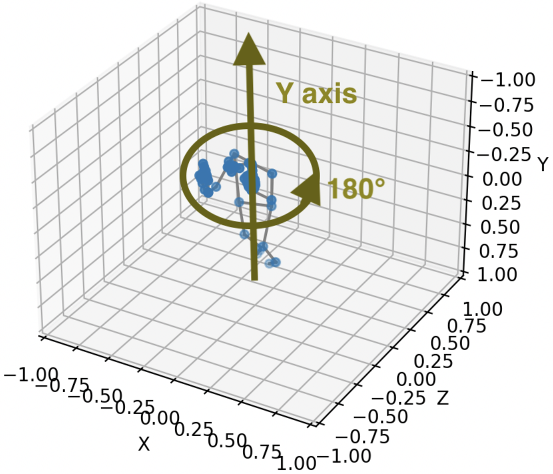

Fig. 5. To make the dataset egocentric, we had to apply a 180 degree rotation around the y axis.

used for various purposes, such as training machine learning models or performing further gesture analysis.

In the context of pose estimation, Mediapipe distinguishes between "Landmarks" and "WorldLandmarks." Landmarks refer to the coordinates of specific points on the body, face, or hands within the image frame. The x and y coordinates are normalized between 0 and 1. They represent a map of the positions of various body parts relative to the image with the origin being the top left most point. The z coordinate is the landmark depth where the origin is the midpoints of the hip. Smaller z values indicate that the landmark is closer to the camera. In Figure 2, we show an example of running the pose estimation on the video for the word "write" using landmarks.

Alternatively, WorldLandmarks use a 3D space with a origin at the hips of the character. Like Landmarks, WorldLandmarks also provides a depth from the camera, z, with the midpoints of the hip as the origin. To calculate the z coordinate, the model uses a scaled orthographic projection with a fixed average depth. This weak-perspective approximates perspective closely in many cases [5]. Using the WorldLandmarks, we can plot the various points of a subject in a 3D space like in Figure 3. Additionally, in the analysis of ASL gestures, WorldLandmarks can help in understanding the depth and visualization of hand movements, which is essential for determining the accuracy of our model.

We ultimately decided to use WorldLandmarks over Landmarks. This decision was based on a number of factors including that we wanted to be able graph the points in a 3D plot. While both coordinate systems are 3D, WorldLandmarks, as denoted in its name, is designed to scale to the coordinate system in the real world. The coordinates use meters for units so that the values mimic measurements from the actual video [6]. Since we were interested in maintaining the scale and relative positioning of items in the video, we chose to use WorldLandmarks.

In our project, we decided to combine the pose and hand landmarkers to enhance the accuracy and comprehensiveness of our ASL gesture recognition. While the pose landmarker provides a broad overview of the body's key points, including the torso, arms, and face, it lacks the fine-grained detail necessary for precise hand and finger movements, which are critical in ASL. By integrating the hand landmarker, we can capture the intricate details of hand shapes and positions, essential for accurately interpreting sign language. Mediapipe even provides a model that combines these two landmarkers into one, called a Holistic landmarker. This combined approach allows us to create a more detailed and holistic representation of each gesture, ensuring that our datasets capture the full complexity of ASL signs. This integration not only improves the precision of our data but also enhances the robustness of our machine learning models, making them better suited to handle the nuances of ASL translation.

B. Addressing Pose Estimation Inaccuracies

While Google's Mediapipe has provided significant advantages in pose estimation and landmark detection, our project encountered several challenges when utilizing this technology for ASL gesture recognition. One notable issue arose when the hands came too close to each other or were obstructed from view. In such instances, Mediapipe often failed to label the hands accurately. We show an example of this issue in Fig. 4. This limitation posed a significant problem because many ASL gestures involve intricate hand movements that can bring the hands into close proximity or partially obscure them. This symptom of Mediapipe led to incomplete or inaccurate data in some cases, which could undermine the effectiveness of machine learning models.

Another challenge came from the use of Mediapipe's Holistic model, which is designed to combine pose, face, and hand landmark detection into a single, integrated solution.

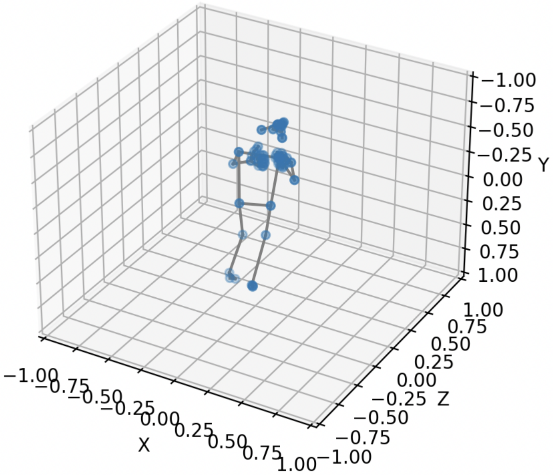

Fig. 6. An egocentric 3D mapping of the coordinates to world space for the word “write.”

While the Holistic model provides comprehensive coverage and can simplify the workflow by handling multiple types of landmarks simultaneously, it has a critical limitation that it does not include world landmarks for the hands to our knowledge. World landmarks provide 3D coordinates, which are essential for capturing the depth and spatial relationships of hand movements in a three-dimensional space. Without these world landmarks, the data from the Holistic model lacks the depth information necessary for accurately interpreting the complex gestures involved in ASL. This omission in the Holistic model significantly impacted our ability to create detailed and accurate datasets. The lack of 3D hand data meant that our models could not fully understand the spatial positioning and movements of the hands, leading to potential inaccuracies in gesture recognition.

Instead of the Holistic model, we switched to use a combination of the hand and pose estimators. Now, we run the pose estimation and hand estimation models independently and combine their results. Since the hand models' coordinates are relative to the midpoint of the hand, we can move the wrist to the origin and then add these points to the pose estimation. This compromise allowed us to capture the world landmarks of all points in the subject. Without the Holistic model, we had to independently determine which wrist a hand belonged with. Additionally, we noticed that Mediapipe occasionally misclassified the left hand from the right hand. When this occurred, it would cause the hands to mistakenly switch wrists.

To address both the issues of obscured and mistaken hands, we apply information from previous frames for correction. Since the last frame that the hand was visible was likely the most similar to the actual position of the hand, this would likely give us the best approximation. We can check if a hand was obscured in the current frame by determining if Mediapipe was able to detect the hand. If a hand is missing in the current frame, we check if it was used in a previous frame. If it was, then we reuse the previous frame's hand coordinates in the current frame.

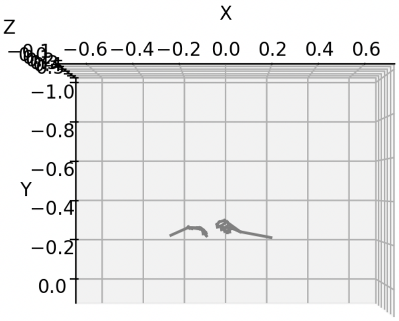

Fig. 7. Arms and hands used to sign the word “write.” The lines are displayed only to increase visibility.

Similarly, if a hand is suddenly mistaken for the other wrist, we check which hand resembles the current one more. This comparison is done by subtracting the current hand's coordinates from the previous left and right hands' coordinates. Then, we take the magnitude of both vectors and choose the one with a smaller magnitude. Geometrically, we are comparing how different of a position the hand is compared to previous left and right hands. This will indicate whether to use the left or right hand.

To account for errors where a hand should actually be hidden, we apply a two pass solution where we count the number of contiguous frames where the hand is absent. If this number exceeds a certain threshold, then we do not reuse a hand from a previous frame. This way, hands that return to the subject's sides will not appear anymore. To determine this threshold, we experimentally tried different values. Using larger thresholds will reduce the number of frames where the hand is absent, but it can also lead to cases where the hand is visible in the animation but not in the source footage.

Using these methods for correcting hidden and mistaken hands, we were able to fix a majority of the cases. We noticed that the method for addressing hidden hands gave us the best results for the most instances. This technique was able to address many of the frames that did not contain one or both hands. Additionally, determining which wrist a hand belonged to removed many problems where the hands would suddenly swap wrists. These augmentations are necessary to ensure the data is temporally consistent enough to be used in a machine learning model.

C. Modifying the Dataset

In developing our ASL dataset, we wanted to make several adjustments to our source data that required careful effort. One major adjustment was that the dataset consisted primarily of videos featuring subjects signing words from a third-person perspective. While these videos are correct depictions of the ASL words, they do not align with our ideal format. Instead, our end application would rely on an egocentric perspective, where the signs are viewed as if the observer were signing with their own hands. This perspective is crucial for accurately

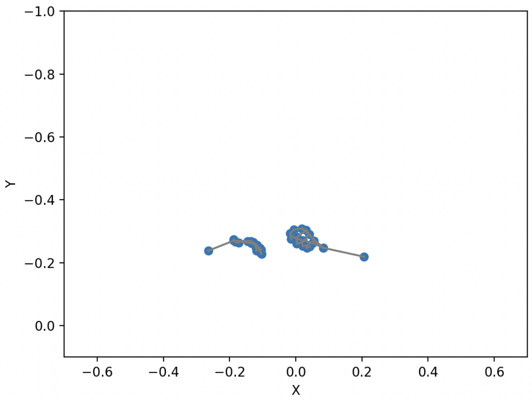

Fig. 8. Mapping the coordinates to a 2D plot for the word “write.”

training models to understand and replicate ASL gestures from the signer’s viewpoint. To convert the third-person perspective data into egocentric views, we had to take the world coordinates of the subject and apply a 180-degree rotation around the y-axis. Fig. 5 shows the rotation that we applied to the world coordinates. After applying the rotation, our subject will be more closely aligned with a user’s perspective. In Fig. 6, we show the plot after applying the rotation.

D. Removing Irrelevant Parts

To refine the dataset for more accurate ASL gesture recognition, one crucial step was to remove irrelevant parts of the subject, particularly everything except the arms and hands. The lower body does not typically play a significant role in ASL signing, and including it in the data could introduce unnecessary noise and complexity. Additionally, our end application will not be able to view the eyes or facial expressions. By focusing solely on the hands and arms, we are able to concentrate on the critical components of ASL gestures—primarily the movements of the hands and arms. Fig. 7 shows this plot without these excessive parts. This selective focus helped to streamline the dataset, making it more relevant and manageable for machine learning models designed to recognize and interpret ASL signs.

Additionally, in order to verify the authenticity of our results, we chose to visualize the projection of the hands on a 2d plot. This projection helped us better identify the correctness of the final animation. Fig. 8 is an example of one such plot for the word “write.” To do this, we disregarded the z coordinate from the world coordinates and graph the x and y coordinates. Using this method, we were able to identify that many of the problems in the animation came from inaccuracies in the pose estimation.

IV. Results

Our comprehensive efforts in refining the WLASL 300 dataset and enhancing its applicability for ASL gesture recognition culminated in the creation of a substantial and high-quality dataset [7]. We successfully generated ASL data spanning 900 videos, encompassing over 300 unique words. Each video was processed to ensure it adhered to our standards, eliminating irrelevant parts and modifying it to be from an egocentric viewpoint. This attention should make our dataset an accurate representation of ASL signs from the user’s perspective.

The transformation of third-person perspective videos into egocentric views, combined with the removal of any videos where the subject turned to the side, significantly improved the dataset’s usefulness. By applying a 180-degree rotation around the y-axis, we ensured that the signs were viewed as if performed by the viewer. This consistency is crucial for training machine learning models, as it provides a stable and clear reference for learning the gestures. Furthermore, our focus on the hands and arms ensured that the models trained on this data would have a high degree of accuracy in recognizing and interpreting ASL gestures. Additionally, removing frames where hands disappear or appear on the other wrist, greatly increased the fidelity of our dataset.

V. Conclusion

Overall, the resulting dataset is a valuable resource for ongoing and future research in ASL-to-English translation. The extensive collection of video data, each curated and processed, provides a solid foundation for developing sophisticated machine learning models capable of accurately translating ASL. This dataset not only addresses the challenges of data scarcity in this field but also offers a robust tool for enhancing the quality and reliability of ASL translation models. Hopefully, this can ultimately contribute to bridging the communication gap between the Deaf and Hard-of-Hearing community and the hearing world.

Acknowledgment

The author thanks Professor Songwu Lu, and his students Ziyue Dang and Yunze Long for their guidance in this project.

References

[1] Rhode Island Commission on the Deaf and Hard of Hearing. https://cdhh.ri.gov/information-referral/american-sign-language.php. Accessed: 2024-06-02.

[2] Bill Vicars. https://www.youtube.com/billvicars. Accessed 2024-06-02.

[3] National Institute on Deafness and Other Communication Orders. https://www.nidcd.nih.gov/health/american-sign-language. Accessed 2024-06-02.

[4] Viso AI. https://viso.ai/computer-vision/mediapipe. Accessed 2024-06-02.

[5] Github Issue 742. https://github.com/google-ai-edge/mediapipe/issues/742.Accessed2024-06-04.

[6] Mediapipe Hand Landmarks. https://ai.google.dev/edge/mediapipe/solutions/vision/hand_landmarker/python. Accessed 2024-06-04.

[7] Li, Dongxu, Cristian Rodriguez, Xin Yu, and Hongdong Li. ”Word-level Deep Sign Language Recognition from Video: A New Large-scale Dataset and Methods Comparison.” In *The IEEE Winter Conference on Applications of Computer Vision*, pp. 1459-1469, 2020.