Novel AI Powered Sign Language Translator

Nand Vinchhi

January 26, 2022

Abstract

Although we see many advancements in voice-enabled technologies, there is an acute disparity between speech-based human-machine interface technologies (e.g. speech to text) and those designed for the speech and hearing impaired. This paper proposes a novel camera vision powered translator for American Sign Language for real-time use. It incorporates technologies such as Mediapipe pose estimation and Dynamic Time Warping for two separate sign language translators - one for alphanumeric characters and the other for words and phrases. This novel method does not require intensive training, and has demonstrated an accuracy of over 90 percent on a test data-set.

1 Introduction

Over 70 million people globally need sign language for communication. Although we see many advancements in voice-enabled technologies, the disparity between advances in these general-purpose communication interfaces (e.g. speech to text) and those designed for the speech and hearing impaired is acute. No automated and scalable technology exists for the translation of sign language. Such a platform would revolutionize the areas of digital communication in the form of AI-assisted closed captioning of videos and conference calls. In addition, sign language translation technology would create an immeasurable impact in the field of education. Current self-learning solutions for sign language as a beginner are largely ineffective due to a lack of interactivity. Creating an engaging platform, similar to Duolingo, with AI-powered practice quizzes and feedback is an extremely compelling use case of such technology. The objectives of this project are:

- To implement a camera vision powered translator for American Sign Language (ASL).
- To optimize the design and algorithms for real-time use.

Existing implementations of ASL translators use Long Short Term Memory RNN models, and have the following drawbacks:

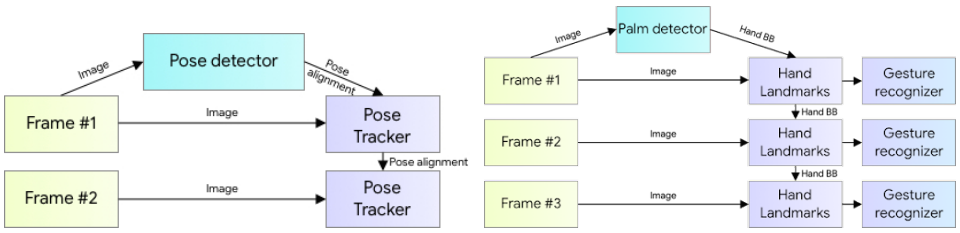
- The over-fitting problem leads to inaccuracies in real-world application.
- Massive data-sets and compute resources are required to facilitate such deep learning models.
- They often rely on external hardware such as depth sensors, which are not readily available in standard computers and mobile phones.

2 Pose Estimation Models

The core feature extraction from raw video data is implemented by leveraging pre-trained Mediapipe pose estimation machine learning models. Two separate models are used - one for body pose estimates and the other for hand pose estimates.

2.1 Body Pose Tracking

Mediapipe Pose utilizes a two-step detector-tracker ML pipeline. [BGR+20] Using a detector, the pipeline first locates the person/pose region-of-interest (ROI) within the frame. The tracker subsequently predicts the pose landmarks and segmentation mask within the ROI using the ROI-cropped



frame as input. Note that for video use cases the detector is invoked only as needed, i.e., for the very first frame and when the tracker could no longer identify body pose presence in the previous frame. For other frames the pipeline simply derives the ROI from the previous frame's pose landmarks. The pipeline is implemented as a MediaPipe graph that uses a pose landmark subgraph from the pose landmark module and renders using a dedicated pose renderer subgraph. The pose landmark subgraph internally uses a pose detection subgraph from the pose detection module.

2.2 Hand Pose Tracking

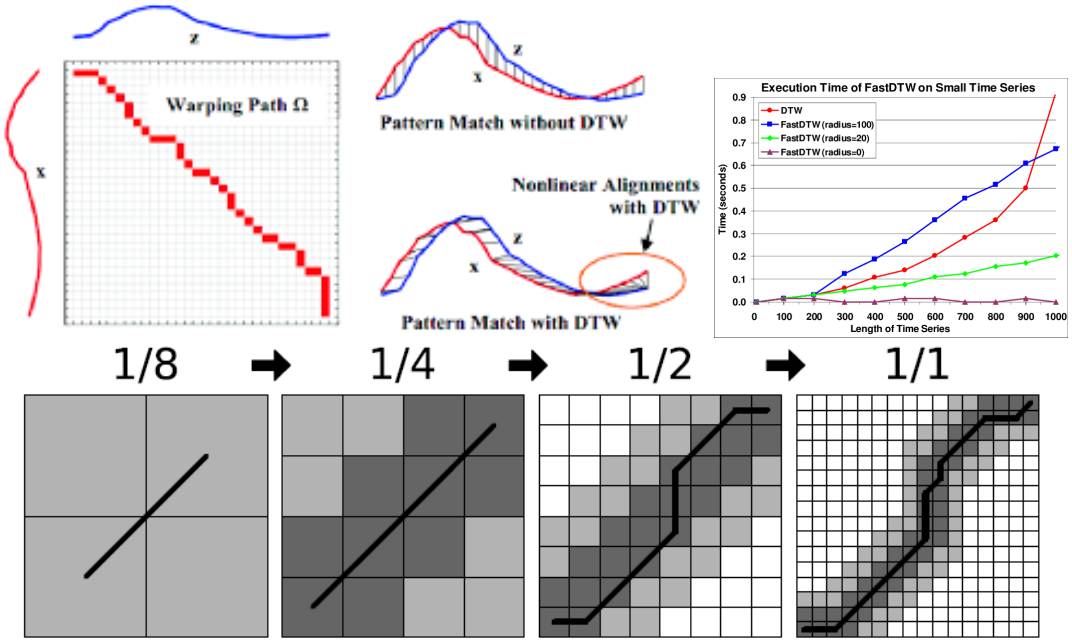
MediaPipe Hands utilizes an ML pipeline consisting of multiple models working together: [ZBV+20] A palm detection model that operates on the full image and returns an oriented hand bounding box. A hand landmark model that operates on the cropped image region defined by the palm detector and returns high-fidelity 3D hand keypoints. Providing the accurately cropped hand image to the hand landmark model drastically reduces the need for data augmentation (e.g. rotations, translation and scale) and instead allows the network to dedicate most of its capacity to coordinate prediction accuracy. In addition, the crops can also be generated based on the hand landmarks identified in the previous frame, and only when the landmark model could no longer identify hand presence is palm detection invoked to relocalize the hand. The pipeline is implemented as a MediaPipe graph that uses a hand landmark tracking subgraph from the hand landmark module, and renders using a dedicated hand renderer subgraph. The hand landmark tracking subgraph internally uses a hand landmark subgraph from the same module and a palm detection subgraph from the palm detection module.

3 Dynamic Time Warping

In time series analysis, dynamic time warping (DTW) is one of the algorithms for measuring similarity between two temporal sequences, which may vary in speed and duration. For instance, similarities in walking could be detected using DTW, even if one person was walking faster than the other, or if there was acceleration and deceleration during the course of an observation. DTW has been applied to temporal sequences of video, audio, and graphics data — indeed, any data that can be turned into a temporal sequence can be analyzed with DTW. In this case, it was used to compare data from a live video feed with similar data extracted from reference videos.

3.1 FastDTW

The conventional DTW algorithm has an $O(N^2)$ time and space complexity, making it very slow for a large number of comparisons. Instead, I made use of the FastDTW algorithm, which is able to find an excellent approximation of the optimal distance between two time series. The FastDTW algorithm avoids the brute-force approach of the standard DTW algorithm by using a multilevel approach. The time series are initially sampled down to a very low resolution. A warp path is found for the lowest resolution and projected onto an incrementally higher resolution time series. The projected warp path is refined and projected again to a higher resolution. The process of refining and projecting is continued until a warp path is found for the full resolution time series. FastDTW achieves a time and space complexity of $O(N)$. [SC04]



4 Real-time Translation

The key inspiration was the observation that there is a variability in speed of sign language gestures. DTW aims at aligning two sequences of temporal feature vectors by warping the time axis iteratively until an optimal match is found. It has been used in the past to account for precisely such a variability in speech signals. [PHA19] The FastDTW variant, which achieves linear time complexity, is used to translate ASL words and phrases from a live webcam feed.

I start by extracting raw body and hand pose values from a dataset of reference videos using pre-trained Mediapipe machine learning models. I computed size-independent values from this raw data using certain feature extraction algorithms. I extracted these same features from a live feed and compared them with each reference word/phrase using FastDTW and a custom scoring function: $score = distance/frames_{avg}$. A word/phrase is detected when the score is below a fixed threshold.

Since single alphanumeric characters in ASL are static in nature, FastDTW is not appropriate. I used a simple RMSE score based thresholding algorithm to detect these characters.

4.1 Testing Procedure for Optimization

There were two leading contenders for size-independent feature extraction from raw pose values. The first one was using a reference point and distance to compute localized and rescaled Cartesian coordinates. The second was to compute angles at each joint (i.e. finger, shoulder and elbow).

$$\theta = arccos(\frac{x.y}{|x||y|})$$

Likewise, for the FastDTW distance metric, there were two leading contenders: Root mean squared error and Euclidean distance.

$$Euclidean = \sqrt{(\sum_{i=1}^{n}(b_i - a_i)^2)}$$

$$RMSE = \sqrt{(\frac{\sum_{i=1}^{n}(b_i - a_i)^2}{n})}$$

My testing procedure involved an exhaustive search over all possible combinations of these techniques on a test data-set of over 150 clips of 35 words/phrases (YouTube channel: PortaleTosign). The results of these experiments are summarized below:

Algorithm for Hands	Algorithm for Body	Distance metric for FastDTW	Accuracy
Angles	Angles	Euclidean	65.22%
Angles	Angles	RMSE	59.96%
Angles	Localized Coordinates	Euclidean	67.29%
Angles	Localized Coordinates	RMSE	66.04%
Localized Coordinates	Angles	Euclidean	90.41%
Localized Coordinates	Angles	RMSE	88.50%
Localized Coordinates	Localized Coordinates	Euclidean	79.72%
Localized Coordinates	Localized Coordinates	RMSE	82.37%

5 Conclusions and Applications

This project has successfully demonstrated ASL translation technology that yields very high accuracy detections, approximately 90%. Unlike deep learning models, this method does not require high performance compute resources or large datasets. Being independent of external hardware, it is economical, extremely scalable and can be made accessible on commonly available devices. Sign language translation technology will revolutionize several application areas. Automatic captioning of sign language will make video calls more inclusive for the speech and hearing impaired. An estimated 80% of people who need sign language have not learnt it. Sign language translation technology will democratize the access to sign language tutoring through a Duolingo-like interactive platform. This technology can also integrate with AI voice assistants such as Alexa. The core of this technology is not limited to sign language translation. It can have diverse applications in the fields of medical technology, gesture analysis for law enforcement and automatic AI physical fitness coaches.

References

[BGR+20] Valentin Bazarevsky, Ivan Grishchenko, Karthik Raveendran, Tyler Zhu, Fan Zhang, and Matthias Grundmann. Blazepose: On-device real-time body pose tracking. *CoRR*, abs/2006.10204, 2020.

[BX18] Kshitij Bantupalli and Ying Xie. American sign language recognition using deep learning and computer vision. pages 4896–4899, 2018.

[PHA19] Yurika Permanasari, Erwin H. Harahap, and Erwin Prayoga Ali. Speech recognition using dynamic time warping (DTW). *Journal of Physics: Conference Series*, 1366(1):012091, nov 2019.

[SC04] Stan Salvador and Philip Ka-Fai Chan. Fastdtw: Toward accurate dynamic time warping in linear time and space. 2004.

[ZBV+20] Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, and Matthias Grundmann. Mediapipe hands: On-device real-time hand tracking. *CoRR*, abs/2006.10214, 2020.