Synchronic Automatic Sign Language Recognition Project Proposal

Zidong Xu(zx2507), Caiwu Chen(cc4786)

Introduction and background

Automatic Sign Language Recognition (ASLR) has great potential in promoting communication between the hearing-impaired community and the wider world. For our final project, we will build upon the midterm project previously investigated by Zidong Xu, aiming to enhance the robustness and generality of the model while contributing to improved human interaction.

Our previous work demonstrated the capability to recognize sign language against a pure background using Convolutional Neural Networks (CNNs) and Multi-Layer Perceptrons (MLPs). The next step is to develop a program capable of identifying sign languages and translating them into text in real-time. The video can be split into frames, CNN can be used to extract frame features, and LSTM can be used for time series modeling. Since different countries have distinct sign language interpretations, we will focus on the Word-Level American Sign Language (ASL) recognition dataset to ensure a standardized approach.

Methodology

1. Data collection:
   We use the Word-Level American Sign Language (WLASL)[4] or the How2Sign dataset[5]. Considering the size of the training process, we will process it by normalization and augmentation image.s
2. Try to use a combination of multiple models to recognize sign language: Use CNN + LSTM, where CNN is responsible for extracting spatial features from each frame and LSTM is used to model gesture changes in time series. LSTM is a classic structure for processing dynamic sign language recognition in videos [1]. Or try CNN + Vision Transformer (ViT): ViT is better at modeling long-distance dependencies in image processing and can be used in combination with CNN [2,3].
3. Background noise suppression: It is planned to apply MediaPipe Hands or OpenPose to extract hand landmarks and use them as model input instead of directly inputting the original image, which can avoid the influence of most environmental noise.
4. The whole project is divided into two stages. The first is the training stage: training the model with a large amount of data and outputting .pt or .onnx files. Then there is the deployment stage, loading the trained model into the system, inputting real-time images from the camera, and quickly inferring and outputting the results.

Core Research Question

1. How can the sign language be identified from the noised background?

2. How to detect the changing of sign language among a series of images, i.e. video?
3. How to embed the sign language translator into a personal computer?
4. How can the robustness of an automatic sign language recognition model be improved for accurate translation?

Dataset

1. How2Sign Dataset
2. Word-Level American Sign Language (WLASL)

Expected Result

Our goal is to design a synchronic automatic sign language recognition system that takes a sequence of videos as the input while outputting a sequence of text related to the corresponding sign language. We will build on the system from zero but inspired from Zidong's midterm experiment. The basic architecture will be similar to the figure below:

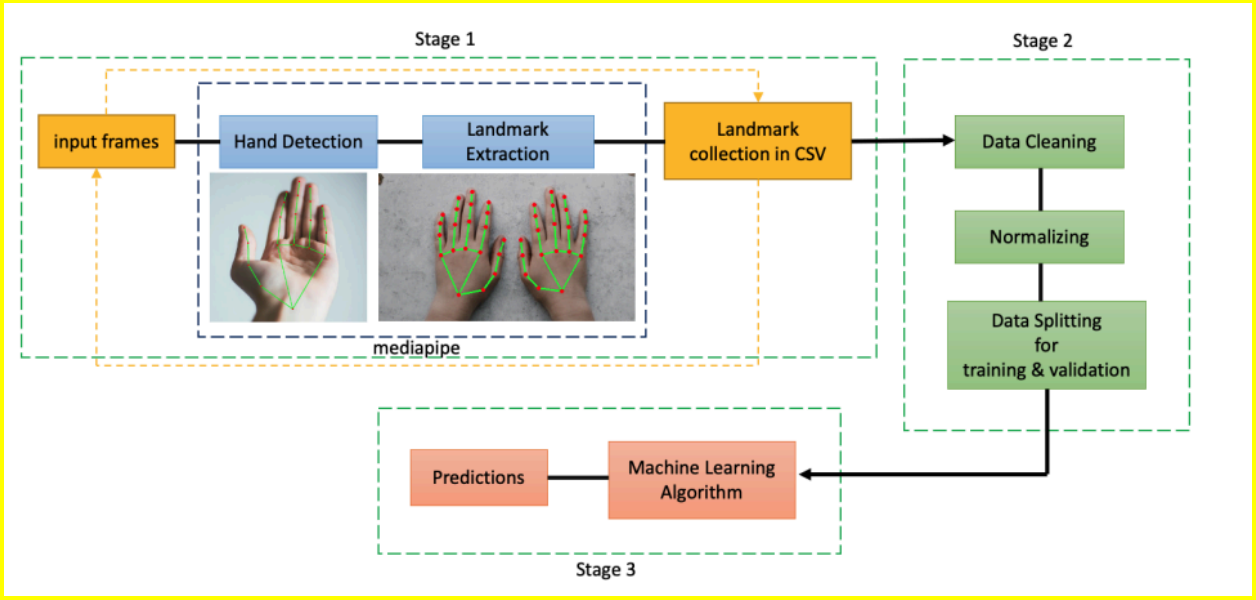

Fig 1, the architecture of ASLR, cited from [6]

Reference

1. S. Mhatre, S. Joshi and H. B. Kulkarni, "Sign Language Detection using LSTM," *2022 IEEE International Conference on Current Development in Engineering and Technology (CCET)*, Bhopal, India, 2022, pp. 1-6, doi: 10.1109/CCET56606.2022.10080705. Sign Language Detection using LSTM | IEEE Conference Publication | IEEE Xplore
2. Nimisha, KP, and Agnes Jacob. "A brief review of the recent trends in Sign language recognition." *2020 International Conference on Communication and Signal Processing*

(ICCSP), July 2021, pp. 186–190, https://doi.org/10.1109/iccsp48568.2020.9182351.

3. Shin, Jungpil, et al. “Korean sign language recognition using transformer-based deep neural network.” *Applied Sciences*, vol. 13, no. 5, 27 Feb. 2023, p. 3029, https://doi.org/10.3390/app13053029.
4. D. Li, C. R. Opazo, X. Yu, and H. Li, "Word-level Deep Sign Language Recognition from Video: A New Large-scale Dataset and Methods Comparison," *CoRR*, vol. abs/1910.11006, 2019. [Online]. Available: http://arxiv.org/abs/1910.11006.
5. A. C. Duarte, S. Palaskar, D. Ghadiyaram, K. DeHaan, F. Metze, J. Torres, and X. Giró-i-Nieto, "How2Sign: A Large-scale Multimodal Dataset for Continuous American Sign Language," *CoRR*, vol. abs/2008.08143, 2020. [Online]. Available: https://arxiv.org/abs/2008.08143.
6. A. Tayade and A. Halder, "Real-time Vernacular Sign Language Recognition using MediaPipe and Machine Learning," May 2021. DOI: 10.13140/RG.2.2.32364.03203. Available: https://www.researchgate.net/publication/369945035_Real-time_Vernacular_Sign_Language_Recognition_using_MediaPipe_and_Machine_Learning