Google - Isolated Sign Language Recognition

王浩
國立成功大學　資訊工程學系
F74082141
Howard.H.Wang.23@gmail.com

杜孟聰
國立成功大學　資訊工程學系
F74082028
F74082028@gs.ncku.edu.tw

Abstract

This paper aims to classify isolated American Sign Language (ASL) signs using a deep learning model. To achieve this, we trained the model on labeled landmark data extracted via the MediaPipe Holistic Solution. Our approach leverages the accuracy and reliability of this technology to enhance the robustness of the model and improve its performance. By accurately classifying ASL signs, our method has the potential to support individuals with hearing impairments and enable better communication between hearing and non-hearing populations.

1. Introduction

Deaf children born to hearing parents are at risk of Language Deprivation Syndrome if they don't have access to a fully developed language early in life. However, learning American Sign Language (ASL) can be time-consuming and resource-intensive for parents. Automatic sign language recognition technology can help overcome this barrier by recognizing and translating signs into text or speech.

In this paper, we propose a novel approach to isolated sign language recognition using computer vision techniques and machine learning algorithms. Our approach improves accuracy, robustness, and speed, as demonstrated on a publicly available dataset.

1.1. American Sign Language

American Sign Language (ASL) is a language used primarily by members of the Deaf community in North America. It is not a visual representation of English, but has its own unique grammar, syntax, and vocabulary. ASL is a visual-gestural language, meaning that it uses facial expressions, body language, and hand movements to convey meaning.

1.2. Isolated Sign Language Recognition

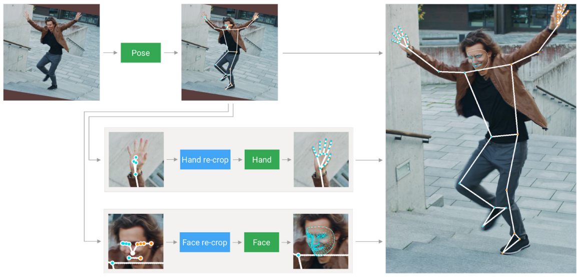
ISLR stands for Isolated Sign Language Recognition, which is the task of recognizing individual signs or tokens called glosses from a given segment of signing video clip. It is the process of recognizing sign language gestures performed by a person in isolation, without considering the context or the surrounding gestures. We will train a machine learning model that can accurately recognize isolated sign language signs and classify them into the correct sign category.

2. System framwork

We will use the raw data extracted from the MediaPipe Holistic Solution, which records every movement in video by landmark. Based on that data, we use an ensamble model to train a deep learning model for classifying isolated American Sign Language (ASL) signs.

2.1. Data Type

MediaPipe Holistic Solution [1] is a powerful, easy-to-use software tool that can detect and track multiple human body parts and gestures in real-time video streams.



The way that MediaPipe Holistic Solution records facial expressions, body language, and hand movements is by using landmarks.

Landmarks or keypoints are like dots that are placed on important areas of an object or a person's body. These dots help a computer to understand where these important areas are and how they are moving.

2.2. Model

Our proposed model leverages a simple linear architecture with added activation functions to form the foundation of our approach. To enhance the model's performance, we will modify the input data format, taking into consideration the unique properties of facial expressions, body language, and hand movements.

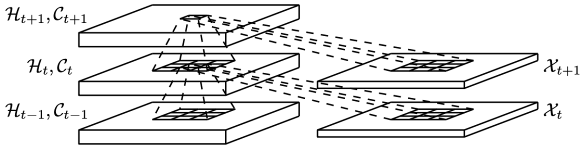

$\mathcal{H}_{t+1}, \mathcal{C}_{t+1}$
$\mathcal{H}_t, \mathcal{C}_t$
$\mathcal{H}_{t-1}, \mathcal{C}_{t-1}$
$\mathcal{X}_{t+1}$
$\mathcal{X}_t$

Figure 2: Inner structure of ConvLSTM

In addition, we will explore the use of time-series models and convolution-based models, such as ConvLSTM [2], to identify the optimal model for fitting the features present in our video-captured frame data. By evaluating multiple model architectures and comparing their effectiveness, we aim to select the most robust and efficient approach for our isolated American Sign Language (ASL) sign recognition task.

Our model development process is designed to enhance the accuracy and reliability of our approach, thereby improving the overall quality of our results. Through a comprehensive evaluation of various model architectures, we aim to establish a state-of-the-art ASL sign recognition system that can effectively support communication between deaf and hearing populations.

3. Expected Result

Our proposed state-of-the-art model, which utilizes the input data of every video frame's landmarks captured by the MediaPipe Holistic Solution, is expected to significantly improve the accuracy and speed of isolated sign language recognition. We anticipate that the model will achieve high recognition rates for individual signs, thus aiding in the communication and language acquisition process of individuals who are deaf or hard of hearing. We believe that our research will contribute to the development of more sophisticated and efficient sign language recognition technology, making communication more accessible for the Deaf community

References

[1] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. Mediapipe: A framework for building perception pipelines, 2019. 1

[2] Xingjian SHI, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun WOO. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. 2