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
secondauthor@i2.org

Abstract

In this paper, we classify isolated American Sign Language (ASL) signs. We create a deep learning model trained on labeled landmark data extracted using the MediaPipe Holistic Solution.

1. Introduction

Around 90% of which are born to hearing parents many of which may not know American Sign Language. Without sign language, deaf babies are at risk of Language Deprivation Syndrome. Learning American Sign Language takes time and resources, which many parents don't have. Training a sign language recognizer can improve the learning and confidence of parents who want to learn sign language to communicate with their loved ones.

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.

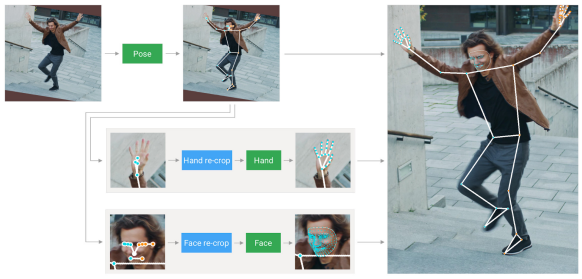
1.2. Isolated Sign Language Recognition

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 MediaPipe Holistic Solution which record every movement in video by landmark. Based on that data, we use ensamble model training 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 record facial expressions, body language, and hand movements is 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

First, we will use simple linear model with adding some activation functions to be our basic model. In order to get a better score, we would also tried to modified some input data fromat according to the properties of facial expressions, body language, and hand movements.

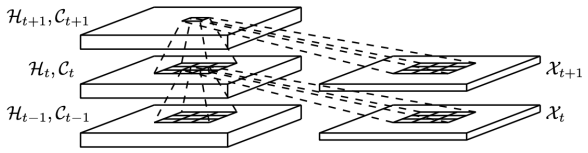


Figure 2: Inner structure of ConvLSTM

We will also try some time-series model and convolution-based model such as ConvLSTM [2] in order to find out the best model fitting the feature in our viddo-captured frame data.

3. Expected Result

We hope creating the state-of-the-art model to predict the American Sign Language (ASL) signs correctly with the input data of every video frame landmarks captured by MediaPipe Holistic Solution.

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