Preprocessing Mediapipe Keypoints with Keypoint Reconstruction and Anchors for Isolated Sign Language Recognition

LREC-COLING 2024, 11th Workshop on the Representation and Processing of Sign Languages

Kyunggeun Roh, Huije Lee, Eui Jun Hwang, Sukmin Cho, Jong C. Park

School of Computing, Korea Advanced Institute of Science and Technology (KAIST), South Korea
{rohbrian, angiquer, ehwa20, nelllpic, jongpark}@kaist.ac.kr

Summary

Goal

- Preprocess Mediapipe keypoints for better recognition
- Concentrate on the hands with diverse preprocessing methods and analyze results

Motivation

- The importance and difficulty of recognizing hands
  - Keypoints extracted from hands are highly noisy and are difficult to capture.
  - Hands contain dense information with minute movements, but are difficult to recognize.
- Necessity of preprocessing for Sign Language Recognition (SLR).
  - Previous work shows the importance of preprocessing keypoints.

Methods

- Anchor-based hand normalization
  - We separately normalize hands based on anchor-points on the palm.
- Hand keypoint reconstruction
  - We apply bilinear interpolation to reconstruct empty hands.
- Fixing Length
  - We fix the length of the input sequence by duplicating frames into a unified distribution.

Results

- Each method improves performance across different transformer architectures and datasets.
- We examine the results to show how each method improves the performance through ablations and to provide analysis with discussions.

Introduction

- Sign language is the visual means of communication and is composed with body and hand gestures.
- Pose-based SLR approaches have been showing the strengths of pose keypoints that are independent of backgrounds or signers, but are quite difficult to recognize.
- Previous researches have shown the importance of preprocessing by masking, normalizing, and augmenting the keypoint sequence.
- We apply preprocessing methods on the keypoints to improve the SLR performance by concentrating on the hands with anchor-based normalization, hand keypoint reconstruction, and fixing the input length.

Contributions

- We explore several preprocessing methods to recognize pose keypoints more effectively, especially by concentrating on the hands.
- The proposed preprocessing methods significantly increase the performance in ISLR and achieve the highest accuracy among pose-based approaches on the WLASL dataset.
- We analyze the results and show in what cases the preprocessing methods make the models more robust and accurate.

Methods

Anchor-based Hand Normalization

- To emphasize the hand shape, we set anchor points to help models learn better with a standard point and the relative distance.
- After normalizing the entire keypoints with an anchor on the chest, we normalize hands separately with anchors on the palm.

Hand Keypoint Reconstruction

- Pose estimation frameworks such as Mediapipe or Openpose easily fail to detect hands, making the task difficult.
- For frames with empty hand keypoints, we initialize the first and last frames with the average position and apply bilinear interpolation from the surrounding frames.

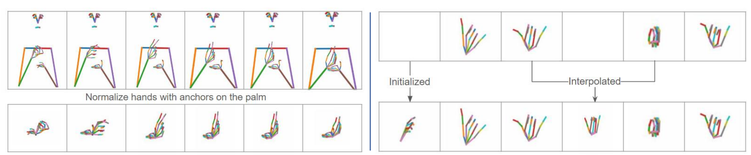

Figure 1. An illustration of the two preprocessing methods, (left) anchor-based hand normalization, and (right) hand keypoint reconstruction by bilinear interpolation.

Fixing Length

- As the lengths of input sequences highly diverse, we extend the length of the input sequence and fix it.
- We extend the length of the input sequence by a unified distribution.

Experiments

Models

- **Transformer Encoder**
  - Utilized a transformer encoder model that shows its strengths in recognition tasks.
  - Optimized with 4 layers, spatial embeddings and class tokens concatenated.
- **Transformer Encoder-Decoder (SPOTER)**
  - Employed the architecture to see if the proposed methods work well on a different architecture.
  - Composed with 6 layers and positional embeddings.

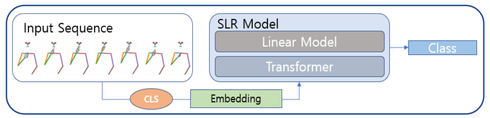

Figure 2. An illustration of the overall model pipeline.

Datasets

- WLASL-100
  A subset of the Word-Level American Sign Language dataset with 100 classes.
  High diversity of signers and backgrounds with low resolution videos.
- AUTSL
  Ankara University Turkish Sign Language dataset with 226 classes.
  Low diversity with high resolution videos.

Dataset	# Glosses	# Videos	# Signer	Detect %
WLASL	100	2,038	97	46.56
AUTSL	226	36,302	43	78.83

Table 1. Statistics of the two datasets.
• Detect % : Hand detection rate with Mediapipe

Experimental Results

Main Results

Dataset	Model	Method: Hand Normalize	Method: Hand Reconst.	Method: Fixing Length	Acc. (%)
WLASL	Transformer Encoder-Decoder (SPOTER)	✗	✗	✗	71.63
		✓	✗	✗	79.38
		✓	✓	✗	**80.31**
		✓	✗	✓	78.68
		✓	✓	✓	79.46
	Transformer Encoder-only (Baseline)	✗	✗	✗	76.12
		✓	✗	✗	79.85
		✓	✓	✗	80.62
		✓	✗	✓	81.16
		✓	✓	✓	**82.17**
AUTSL	Transformer Encoder-only (Baseline)	✗	✗	✗	90.40
		✓	✗	✗	90.76
		✓	✓	✗	90.77
		✓	✗	✓	90.95
		✓	✓	✓	**91.15**

Table 2. Accuracy scores with the proposed methods.
• Hand Reconst. : Hand keypoint reconstruction

- The proposed methods demonstrate the same tendency across different datasets.
- Most of the methods are model-independent, except for fixing the length.

Effectiveness of Anchor-based Normalization

Method	None	Bounding Box	Anchor-based
SPOTER	71.63	76.59	**79.38**
TF Encoder	76.12	78.06	**79.85**

Table 3. Ablation results with different methods of normalizing hands.

- Anchor-based hand normalization demonstrates better results than the bounding box approach with a higher accuracy score.

Analysis of Hand Keypoint Reconstruction

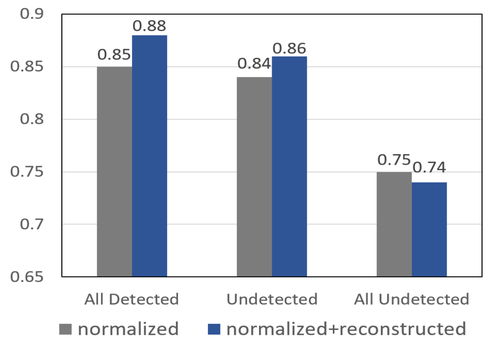

Figure 3. Effect of reconstruction on hand detected and undetected cases.

Case Study

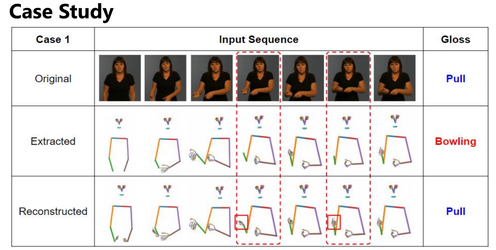

Figure 4. Hand reconstruction correcting the prediction.