LREC 20 Palma 26

12th Workshop on the Representation and Processing of Sign Languages: Language in Motion

Continuous Sign Language Recognition using Multimodal Input and Handshape-aware Boundary Detection

Mingyu Zhao*[1], Zhanfu Yang*[1], Yang Zhou*[1], Zhaoyang Xia*[1], Can Jin*[1], Xiaoxiao He*[1], Shuhang Ling*[1], Carol Neidle[2], Dimitris N. Metaxas[1]

{zhao.mingyu, zhanfu.yang, eta.yang, zx149, can.jin, xiaoxiao.he, shuhang.lin}@rutgers.edu, carol@bu.edu, dnm@cs.rutgers.edu

[1] Rutgers University

[2] Boston University

Abstract

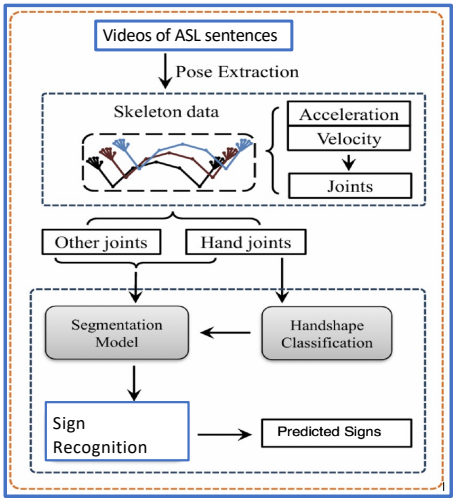
This paper employs a multimodal approach for continuous sign recognition by first using ML for detecting the start and end frames of signs in videos of American Sign Language (ASL) sentences, and then by recognizing the segmented signs. For improved robustness, we use 3D skeletal features extracted from sign language videos to take into account the convergence of sign properties and their dynamics that tend to cluster at sign boundaries. To further improve our recognition results, we incorporate information from 3D hand configuration for boundary detection. To detect handshapes expected at the beginning and end of signs, we pretrain a handshape classifier for detection of 87 linguistically defined canonical hand-shape categories using a dataset that we created by integrating and normalizing several existing datasets. A multimodal fusion module is then used to unify the pretrained sign video segmentation framework and handshape classification models. The estimated boundaries are then used for sign recognition, where the recognition model is trained on a large database containing both citation-form isolated signs and signs pre-segmented (based on manual annotations) from continuous signing—as such signs often differ a bit in certain respects. We evaluate our method on the ASLLRP corpus and demonstrate significant improvements over previous work.

Pipeline

We begin by extracting 2D skeletal keypoints from sign language videos using a pose estimation tool. We use velocity and acceleration information to augment the joint feature in each frame.

These features are then separated into two branches: a) Hand joints are fed into a pretrained handshape classification model; b) All joints are used as input to a segmentation model. The outputs of both branches are fused to improve segmentation boundaries.

For sign recognition, the segmented sign clips are further passed into a sign recognition model trained on both citation-form signs and signs segmented from continuous signing, to produce final sign predictions.



Our Contribution

We design a segmentation module based on a spatiotemporal convolutional network, incorporating velocity and acceleration features.

A transformer-based multimodal fusion module is then employed, to integrate features from a pretrained handshape classifier into the segmentation stream using a cross-attention mechanism with gating.

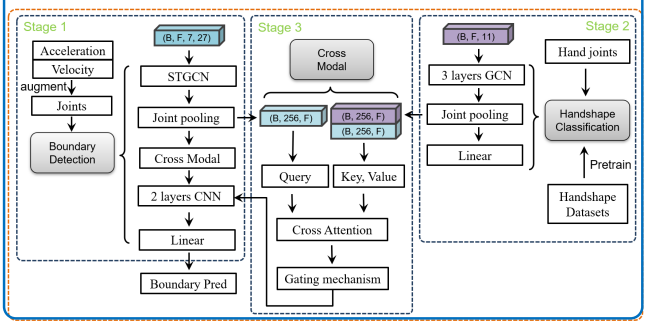
New CSLR pipeline: Integration of the segmentation model with the sign language recognition model. Experiments demonstrate the effectiveness of our segmentation framework.

Our Approach

Stage 1: We first pretrain the segmentation branch based on ST-GCN and CNN modules. This branch takes all skeletal joints together with their velocity and acceleration as input to model spatiotemporal motion patterns for boundary detection.

Stage 2: We then pretrain the handshape branch on curated handshape datasets using a 3-layer GCN. This branch learns frame-wise handshape representations that provide complementary local hand information for the boundary detection task.

Stage 3: Finally, we introduce a cross-modal attention module with a gating mechanism to fuse segmentation and handshape features for enhanced temporal boundary prediction.



Datasets*

Handshape recognition pretraining datasets

1. **ASLLVD dataset**
2. **ASLLRP Sentences (ASLLRP-S)**
3. **DSP dataset**
4. **NCSLGR handshape videos**

Sign recognition datasets

1. **ASLLVD dataset**
2. **ASLLRP Sentences (ASLLRP-S)**
3. **DSP dataset**
4. **WLASL dataset**
5. **RIT dataset**
6. **DSP Sentences (DSP_S)**

Boundary detection dataset

For boundary detection, we used all pre-segmented clips from ASLLRP-S. Start and end frame annotations were used as supervision for temporal boundary prediction.

*Data and further information available from **https://dai.cs.rutgers.edu/dai/s/signbank** and **https://www.bu.edu/asllrp/cslgr/pages/ncslgr-handshapes.html**.

Results

Table 1: Segmented signs recognition results.
– *Column 1:* Minimum # of sign occurrences in our sign recognition dataset.
– *Column 2:* Top-1 sign recognition accuracy.

Table 2: Ablation studies: demonstrating that our method outperforms existing methods. mF1B / mF1s means (mean F1 score of boundary / segments).

1.

min sign occ	Top-1 Accuracy
6	80.23%
10	80.86%
15	81.67%
20	82.17%
30	**83.30%**

2.

Method	mF1B	mF1S
I3D, TCN	71.50	52.78
MSTCN, HaMeR	76.22	50.18
Ours w/o handshape module	77.29	56.98
Ours with handshape module	**79.40**	**58.26**

This work was partially funded by grants 2212302, 2212301, and 2212303 from the National Science Foundation.