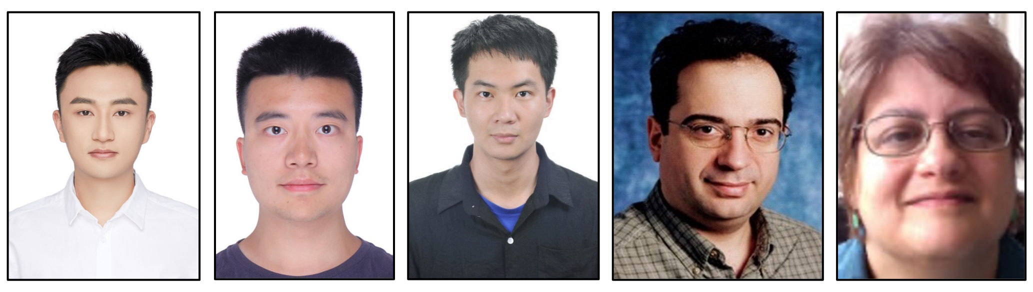
11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources

LREC-COLING 2024

A Multimodal Spatio-Temporal GCN Model with Enhancements for Isolated Sign Recognition



Yang Zhou*[1], Zhaoyang Xia*[1] , Yuxiao Chen*[1], Carol Neidle[2], Dimitris N. Metaxas[1]

[1] Rutgers University; [2] Boston University

{eta.yang, zx149}@rutgers.edu, yc984@cs.rutgers.edu, carol@bu.edu, dnm@cs.rutgers.edu

Abstract

We propose a multimodal network using skeletons and handshapes as input to recognize individual signs and detect their boundaries in American Sign Language (ASL) videos. Our method integrates a spatio-temporal Graph Convolutional Network (GCN) architecture to estimate human skeleton keypoints; it uses a late-fusion approach for both forward and backward processing of video streams. Our (core) method is designed for the extraction—and analysis of features from—ASL videos, to enhance accuracy and efficiency of recognition of individual signs. A Gating module based on per-channel multi-layer convolutions is employed to evaluate significant frames for recognition of isolated signs. Additionally, an auxiliary multimodal branch network, integrated with a transformer, is designed to estimate the linguistic start and end frames of an isolated sign within a video clip. We evaluated performance of our approach on multiple datasets that include isolated, citation-form signs and signs pre-segmented from continuous signing based on linguistic annotations of start and end points of signs within sentences.

We have achieved very promising results when using both types of sign videos combined for training, with overall sign recognition accuracy of 80.8% Top-1 and 95.2% Top-5 for citation-form signs, and 80.4% Top-1 and 93.0% Top-5 for signs pre-segmented from continuous signing.

Datasets *

Isolated, citation-form, sign datasets

1. **ASLLVD** **9,746** sign video clips
2. **WLASL** **19,666** sign video clips
3. **RIT** **12,197** sign video clips
4. **DSP** **2,935** sign video clips

Datasets of signs pre-segmented from continuous signing

5. **ASLLRP sentences** **17,222** sign video clips
6. **DSP sentences** **3,136** sign video clips

Total of 64,902 video clips. After imposing a requirement of at least 6 available example video clips per sign, we arrived at a total of **56,681 distinct video clips** corresponding to **2,377 distinct signs.**

*Data from **https://dai.cs.rutgers.edu/dai/s/signbank**

Results

Recognition of isolated signs trained on the combined isolated & pre-segmented datasets

DATASETS:	WLASL	ASLLVD	RIT	DSP	*Combined*
Top-1	81.32%	86.70%	75.31%	79.97%	**80.76%**
Top-5	95.41%	96.95%	93.38%	95.28%	**95.18%**

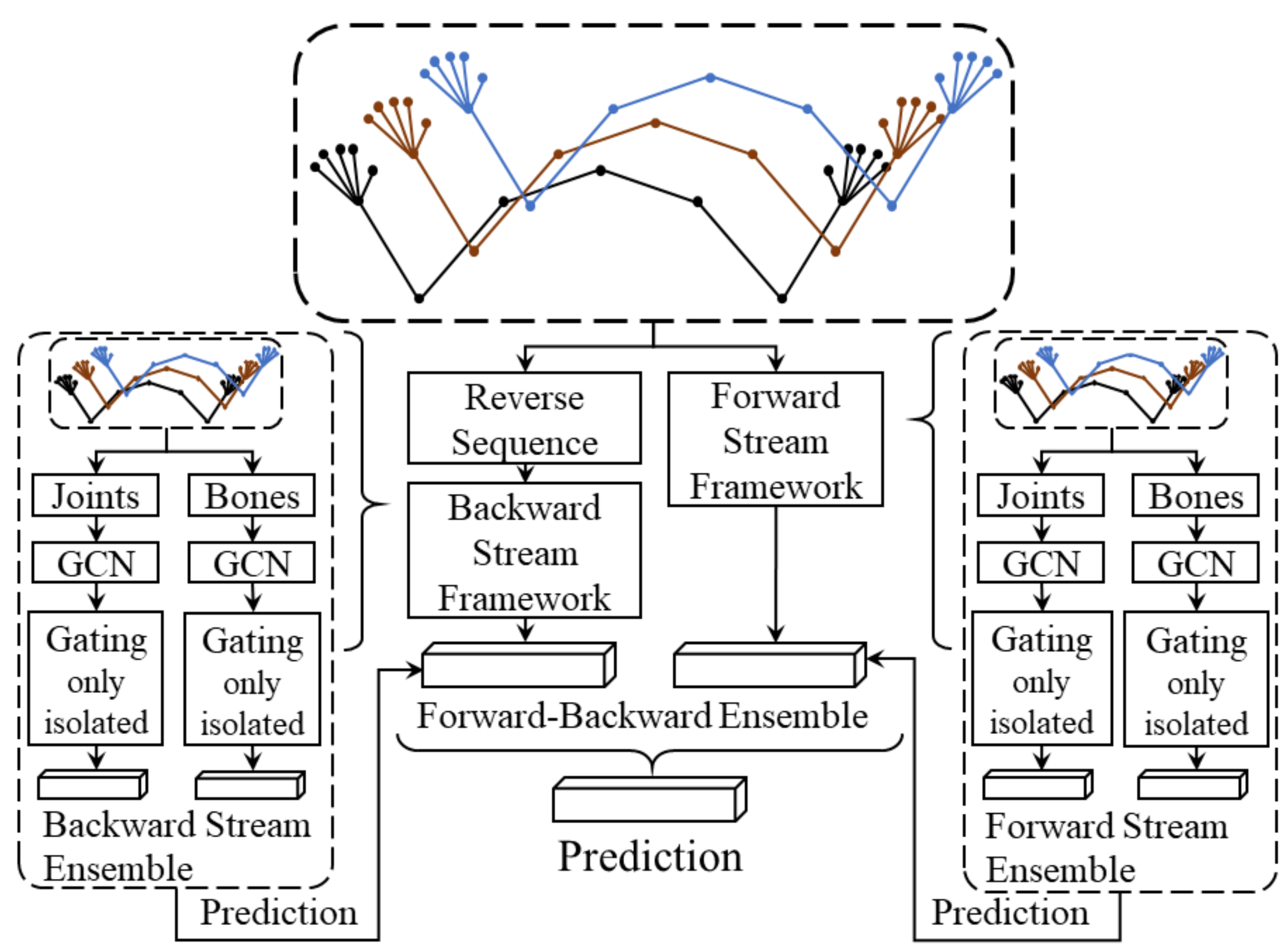
Recognition of pre-segmented signs trained on the combined isolated & pre-segmented datasets

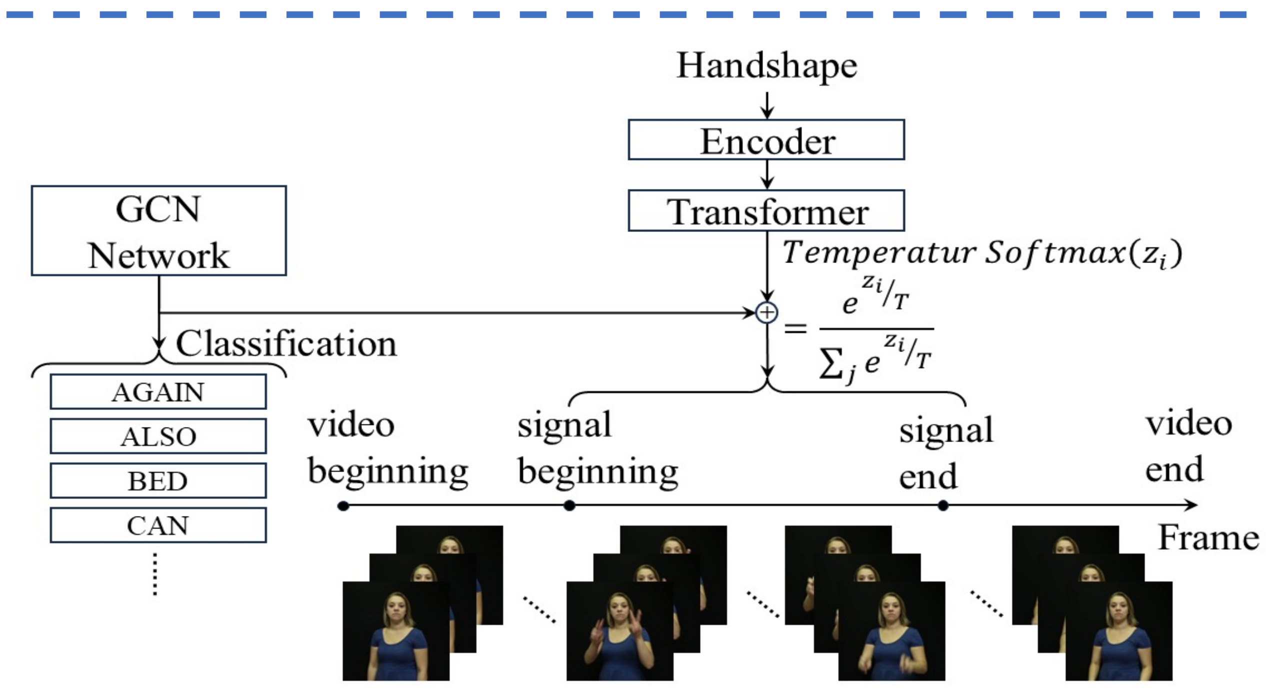
DATASETS:	ASLLRP	DSP_S	*Combined*
Top-1	81.58%	73.86%	**80.39%**
Top-5	93.39%	90.62%	**92.96%**

Our Approach

Our method analyzes a signer's arm and hand movements by extracting keypoints and bones. These are input into a **modified GCN** with spatio-temporal graph convolutions, which:

- Processes video frames in both forward and backward directions to enhance sign language recognition.
- Uses a Gating module with temporal attention to filter out non-informative frames to improve sign recognition accuracy.
- Uses an auxiliary multimodal branch network to further improve sign recognition by identifying the start and end frames of isolated signs using a transformer network.





Comparison

OVERVIEW from Xiao *et al.* 2023

Performance on WLASL dataset (isolated signs)

Method	Top-1	Top-5
Metric-based		
Matching Nets (Vinyals *et al.* 2016)	41.22%	50.26%
Prototypical Nets (Snell *et al.* 2017)	47.61%	65.13%
Relation Net (Sung *et al.* 2018)	45.26%	63.21%
Meta-based		
MetaLSTM (Ravi & Larochelle, 2016)	41.56%	60.38%
SNAIL (Mishra *et al.* 2017)	42.18%	53.77%
MAML (Finn *et al.* 2017)	46.21%	59.15%
MMNet (Cai *et al.* 2018)	52.13%	65.06%
Dynamic-Net (Gidaris & Komodakis, 2018)	54.21%	70.21%
Generation-based		
VERSA (Gordon *et al.* 2018)	49.11%	61.19%
Param Predict (Qiao *et al.* 2018)	55.36%	73.28%
wDAE (Gidaris & Komodakis, 2019)	55.05%	70.12%
Graph-based		
GNN (Garcia and Bruna, 2017)	52.02%	63.89%
CovaMNet (Li *et al.* 2019)	51.18%	66.39%
TPN (Liu *et al.* 2018)	52.15%	65.22%
SL-GCN (Xiao *et al.* 2023)	56.15%	73.26%
COMPARE WITH...		
Dafnis *et al.* 2022	77.43%	94.54%
Ours	**79.59%**	**95.32%**

This work was partially funded by grants from the National Science Foundation.