ONE-SHOT ISL WORD RECOGNITION VIA I3D & POSE FUSION

By R.Charan

MOTIVATION

- Indian Sign Language (ISL) has a very strong number of users but is relatively resource-poor.
- A small dataset makes training models, especially deep-learning models very hard just on ISL Resources.
- We take up the task of word-gloss traslation, that is,
- Given a video of a signer (performing gestures and actions), the task is to predict the corresponding gloss label(word)

INTRODUCTION

- We work with the publically available CISLR_v1.0- a dataset.

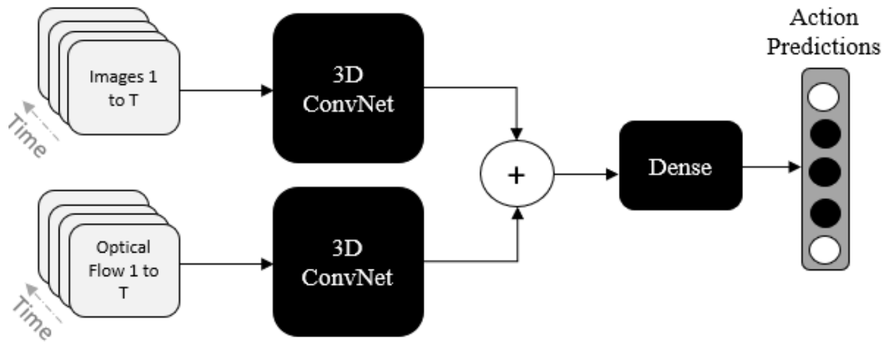
- We leverage both the extracted I3D as well as mediapipe pose data to accomplish the task

METHODS

- Although the final method is of one-shot matching as proposed in CISLR-2022, we try different preprocessings on the I3D and pose data.

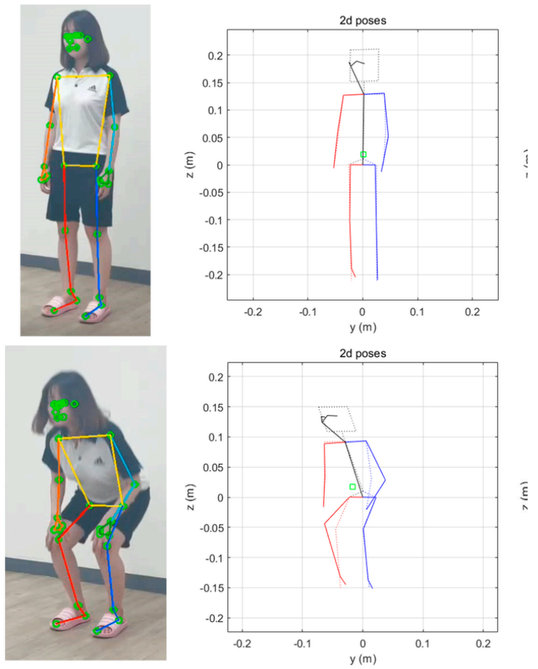
I3D (INFLATED 3D CONVNETS)

CISLR dataset uses I3D features extracted from I3D pre-trained on resource rich American Sign language.
I3D extends a 2D CNN (Inception v-1) into 3D to capture spatiotemporal features in video data.



MEDIAPIPE POSE

Uses 2D CNNs for pose estiation.
It has a 2-step pipeline - detector followed by landmarker.
There are 543 points consisting of 33 body-pose joints, 468 face-mesh vertices, and 21 hand joints per hand, each annotated with normalized 3D coordinates and confidence scores



INITIAL RESULTS

Method	Top-1	Top-5	Top-10
I3D Only	17.02%	20.88%	22.93%
Pose Only	0.31%	0.39%	0.83%
Pose-Guided I3D (segment-wgt)	16.54%	20.53%	22.23%
Velocity-weighted I3D	15.45%	19.12%	21.27%

NEXT STEPS

- Pose data seems to be too noisy and degrades pure I3D features.

- We now explore methods that work purely with I3D but processes the extracted (1, 1024, S, 1, 1) feature set of each gloss better.

FIRST STEPS

Branch	Top-1	Top-5	Top-10
GeM-I3D	19.04%	24.33%	27.44%
PCA-whiten I3D	18.99%	24.64%	27.66%
Attn-Velocity	18.86%	24.29%	27.22%
Mahalanobis I3D	17.37%	21.53%	23.37%

FUSING THE BEST

- We can see that GeM, PCA and Mahalanobis on I3D features alone gave a significant boost in the results. Now we try to fuse the 3.

Method	Top-1	Top-5	Top-10
Fused Score (α=0.6, β=0.4, γ=0.0)	19.21%	25.16%	28.18%
PCA-Whitened GeM-I3D (1024d)	19.69%	24.64%	27.22%

FUSING THE BEST

- Initial CISLR-2022 results:

Dataset	# Test Samples	Top-1	Top-5	Top-10
CISLR v1.0-a	2285	16.81	20.04	22.58

- Final results on CISLR_v1-a dataset:

Method	Top-1	Top-5	Top-10
Fused Score (α=0.6, β=0.4, γ=0.0)	19.21%	25.16%	28.18%
PCA-Whitened GeM-I3D (1024d)	19.69%	24.64%	27.22%

OTHER WORKS

- **Data scraping :**

Created a semi-automated python script that scraps data from a publically available youtube channel - DEF.
The script is robust to small format variations and automatically segments and saves data in the needed format - word(gloss) and sentence wise.
It also automatically appends the newly scrapped info into a csv file. (videos are uniquely labelled)

- Scraped data for 6 years (2019-2024). Can be augmented to the existing dataset, to increase its size.

AKNOWLEDMENT AND CREDITS

Professor Ashutosh Modi
Guide (PhD) Sanjeet Singh

2022-CISLR Corpus for Indian Sign Language Recognition
https://huggingface.co/datasets/Exploration-Lab/CISLR
https://github.com/sign-language-processing/pose

THANK YOU