Fact sheet: CVPR 2021 ChaLearn Looking at People Large Scale Signer Independent Isolated SLR Challenge

This is the fact sheet's template for the CVPR 2021 ChaLearn Looking at People Large Scale Signer Independent Isolated SLR Challenge [1]. Please fill out the following sections carefully in a scientific writing style. Then, send the compressed project (in `.zip` format), i.e., the generated PDF, `.tex`, `.bib` and any additional files to `juliojj@gmail.com`, and put in the Subject of the email "CVPR 2021 SLR Challenge / Fact Sheets", following the schedule and instructions provided in the Challenge webpage [1] "*Wining solutions (post-challenge), Fact Sheets*". Note, if you participated in both track, you will need to send one fact sheet per track.

I. Team details

- **Challenge Track** (RGB or RGB+D): RGB
- Team leader name: Amit Moryossef
- Username on Codalab: AmitMY
- Team leader affiliation: Google, Bar-Ilan University
- Team leader address: Israel
- Team leader phone number: +972 54-3333-207
- Team leader email: amitmoryossef@gmail.com
- Name of other team members (and affiliation):
  - Ioannis Tsochantaridis (Google)
  - Annette Rios (UZH)
  - Mathias Müller (UZH)
  - Sarah Ebling (UZH)
  - Joe Dinn (University Of Surrey)
  - Necati Cihan Camgöz (University Of Surrey)
  - Richard Bowden (University Of Surrey)

II. Contribution details

A. Pose Estimation for SLR

Human poses are explainable, representative, person independent, low dimension representation for the state of a person and how it changes over time. In this submission we explored the viability of using state of the art pose estimation for person independent sign language recognition. We performed two independent studies by two teams, with two different pose estimators to understand if existing pose estimation tools perform well enough for sign language, and made these poses publically available (`https://github.com/sign-language-processing/datasets`).

B. Introduction and Motivation

Our motivation was the evaluation of pose estimation systems for their immediate applicability in sign language processing. Poses are somewhat privacy perserving, as they abstract the features of the human body while reducing identifying characteristics. These reduced characteristics are a feature of this work, as it is then easier for the model to work for unseen people.

C. Representative image / workflow diagram of the method

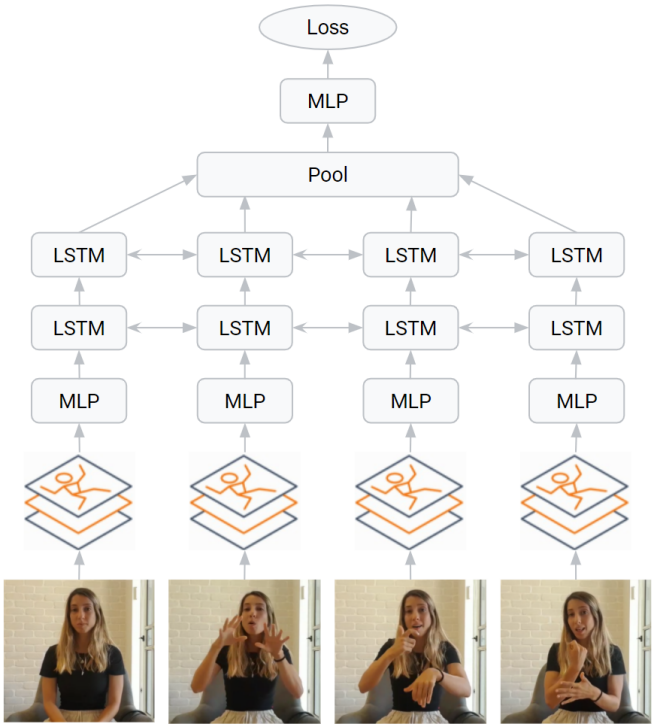

Fig. 1. Model workflow diagram

D. Detailed method description

1) Data Loading: We load the AUTSL dataset [2] including OpenPose [3] and Mediapipe Holistic pose estimation using the sign language datasets library [4].

2) Model: We use a standard sequence classification architicture. For each frame in the original video, we take the pose as a flat vector representation, concatenated with the 2D angle and length of every limb in the body, using the *pose-format*[1] library. To that representation, we apply 20% dropout, 1D batch normalization, and project it from its size to a vector with 512 dimensions using an MLP. We feed the projected vectors into a 2 layers BiLSTM with hidden dimension 256 and apply max pooling on the output to obtain a single representation vector per video. Finally,

[1] `https://github.com/AmitMY/pose-format`

we project that vector from 512 dimensions to 226, where each dimension represents a class.

3) Training: We train the network using the cross entropy loss with a default parameters Adam optimizer, and a batch size of 512 on a single GPU. We apply *no* data augmentation or frame dropout at training time.

E. Challenge results and final remarks

We did not perform a submission of our final system in the validation phase, and so the number we report from the public leaderboard is outdated. Instead, our approach reachs 82.89% accuracy on the validation set.

TABLE I
LEADERBOARD: RESULTS OBTAINED BY THE PROPOSED APPROACH.

Phase	Track	Rank position	Rec. Rate
Development	RGB	30	76.57
Test	RGB	19	81.93

III. ADDITIONAL METHOD DETAILS

Please reply if your challenge entry considered (or not) the following strategies and provide a brief explanation.

- **Did you use any kind of depth information (directly, such as RGBD data, or indirectly such as 3D pose estimation trained on RGBD data), either if during training or testing stage?** (X) Yes, ( ) No
  If yes, please detail:
  We used MediaPipe Holistic pose estimation, which is a 3D pose estimation tool.
- **Did you use pre-trained models?** ( ) Yes, (X) No
  If yes, please detail:
- **Did you use external data?** ( ) Yes, (X) No
  If yes, please detail:
- **Did you use other regularization strategies/terms?** ( ) Yes, (X) No
  If yes, please detail:
- **Did you use handcrafted features?** (X) Yes, ( ) No
  If yes, please detail:
  In addition to the pose keypoints locations, we used the 2D angle and length of every limb.
- **Did you use any face / hand / body detection, alignment or segmentation strategy?** ( ) Yes, (X) No
  If yes, please detail:
- **Did you use any pose estimation method?** (X) Yes, ( ) No
  If yes, please detail:
  We used MediaPipe Holistic 3D pose estimation, and OpenPose single-network pose estimation [3].
- **Did you use any fusion strategy of modalities?** ( ) Yes, (X) No
  If yes, please detail:
- **Did you use ensemble models?** ( ) Yes, (X) No
  If yes, please detail:
- **Did you use any spatio-temporal feature extraction strategy?** ( ) Yes, (X) No
  If yes, please detail:
- **Did you explicitly classify any attribute (e.g. gender)?** ( ) Yes, (X) No
  If yes, please detail:
- **Did you use any bias mitigation technique (e.g. rebalancing training data)?**
  ( ) Yes, (X) No
  If yes, please detail:

IV. CODE REPOSITORY

`https://github.com/AmitMY/ChaLearn-AUTSL-Challenge`

REFERENCES

[1] ChaLearnLAP. CVPR 2021 ChaLearn Looking at People Large Scale Signer Independent Isolated SLR Challenge. [Online]. Available: http://chalearnlap.cvc.uab.es/challenge/43/description/

[2] O. M. Sincan and H. Y. Keles, "Autsl: A large scale multi-modal turkish sign language dataset and baseline methods," *IEEE Access*, vol. 8, pp. 181 340–181 355, 2020.

[3] G. Hidalgo, Y. Raaj, H. Idrees, D. Xiang, H. Joo, T. Simon, and Y. Sheikh, "Single-network whole-body pose estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6982–6991.

[4] A. Moryossef, "Sign language datasets," https://github.com/sign-language-processing/datasets, 2021.