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

I. Team details

- **Challenge Track** (RGB or RGB+D): RGB
- Team leader name: Jose Luis Alba
- Username on Codalab: jalba
- Team leader affiliation: atlanTTic Research Center, University of Vigo
- Team leader address: E.E.Telecomunicación, Campus Universitario de Vigo, 36310, Spain
- Team leader phone number: +34986812680
- Team leader email: jalba@gts.uvigo.es
- Name of other team members (and affiliation): Laura Docío, Eduardo R. Banga, Manuel Vázquez, Soledad Torres, Ania Pérez, Carmen García (same affiliation)
- Team website URL (if any): http://gtm.uvigo.es/

II. Contribution details

A. Learning Isolated Signs with Multi-Scale Graph Convolutional Networks

Isolated Sign recognition fits perfectly in the domain of problems that can be handled by graph-structured spatial-temporal algorithms, like Graph Convolutional Networks. A recent variation, MS-G3D [1], that leverages semantic connectivity among no-neighbor nodes of the graph in a flexible temporal scale has resulted in an improved performance in classical HAR datasets. In this challenge we have used a skeleton graph that includes body and finger joints, so, that specific property seems to be crucial to capture the internal relationship among distant nodes semantically connected in the sign dynamics.

B. Introduction and Motivation

State of the art methods for sign recognition based on extracting features directly from raw RGB data without any kind of previous feature extraction, are quite data hungry. After testing the contest dataset with I3D [2] and S3D [3], with good accuracy on validation set ($\sim$ 88%), we thought that skeleton-based approaches could offer complementary and robust information to fuse with. So we resort to a skeleton-based technique using spatial-temporal graphs, MS-G3D [1]. One important feature of the sign language is the semantic connection between both hand configurations in bi-manual signs and between one hand and other body parts in monomanual signs. MS-G3D introduces a flexible mechanism to learn the connected variation among nodes from any part of the graph in a spatial and temporal predefined scale. Signs contained in the AUTSL dataset have many similar signs that are only distinguishable by hand and finger configurations. Tests with I3D and S3D seemed not to be able to capture enough discriminative detail, probably because the amount of blurriness in many videos and also the scarcity of examples of enough quality. The MS-G3D technique, with some slight regularization tweaks, and fusing results of a network trained with joints and a network trained with bones, resulted in a 95.51% accuracy in the validation set, an improvement of 7% over RGB-based approaches. Given that the computational requirements for training I3D/S3D and also the larger size of the model, the option of fusing both approaches seemed not worth it. It is obvious that the performance of the model depends on the performance of the keypoint extraction (using OpenPose [4] in our case). So we could consider a drawback this dependency, but during development we have tested that the model is slightly robust to missing keypoints, so we implemented an augmentation technique for the model to learn to increase this robustness to prevent a bad location of keypoints during inference. This technique resulted in a 1-2% gain in our tests.

C. Representative image / workflow diagram of the method

An overview of the proposed SLR system is presented in Figure 1, and the set of keypoints (joints and bones) in Figure 2.

D. Detailed method description

In order to be able to reproduce the results you will need to follow the explanations already in the github repository. In this section we give the details of the main building blocks and specific details on the training hyperparameters.

1) **Keypoint extraction using Openpose.** The input video is processed with openpose through this parameters:
   –net_resolution "-1x512" –display 0 –scale_number 4 –scale_gap 0.25 –hand –hand_scale_number 6 –hand_scale_range 0.4 –render-pose=0'
2) **Selection of the subset of joints and bones.** Only keypoints and bones from the upper-body are kept, as indicated in Figure 2, where white/black numbers represent joint labels and orange labels represent bones.
3) **Graph Network definition.** The physically connected joints are represented in the Adjacency matrix, both for joints and bones.
4) **Data augmentation.** The original set is augmented in training time (each mini-batch) by mirroring left-right, adding location and size noise and random remove of keypoints weighted by openpose confidence score.
5) **Training MS-G3D.** All models are trained with SGD with Nesterov's accelerated gradient, momentum 0.9,

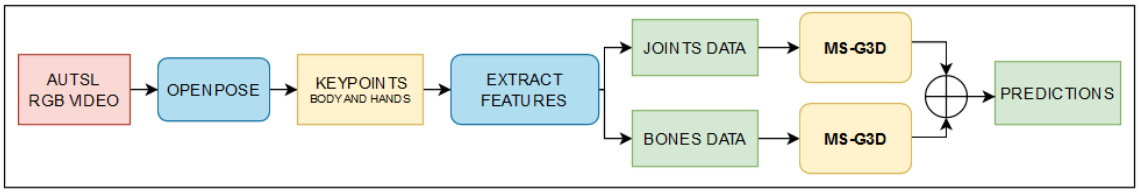

Fig. 1. Block diagram of the proposed approach.

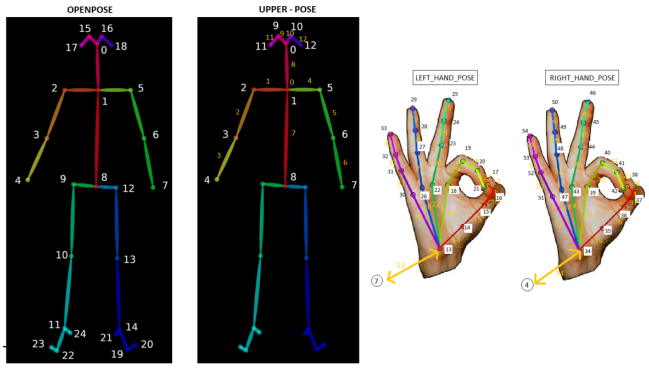

Fig. 2. Skeleton and hand joints.

batch size 64, weight decay 0.0003 and an initial learning rate 0.1 with step LR decay with a factor of 0.1 at epochs 45 and 55. All skeleton sequences are padded to T = 157 frames. For the disentangled aggregation scheme used for multiscale learning, we set a number of scales of 8 in the G3D and GCN blocks

6) **Test Time Augmentation.** Make predictions for six different versions of validation and test set, and then average their results.

E. Challenge results and final remarks

TABLE I
LEADERBOARD: RESULTS OBTAINED BY THE PROPOSED APPROACH.

Phase	Track	Rank position	Rec. Rate
Development	RGB	8*	95.51%
Test	RGB	7	96.15%

* This is the rank of our best development submission by the time of closing that phase.

III. ADDITIONAL METHOD DETAILS

- **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?** ( ) Yes, (X) No
  If yes, please detail:

- **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?** (X) Yes, ( ) No
  If yes, please detail: We used data augmentation by mirroring frames, slight scale change and eliminating low-confidence keypoints randomly. During inference, the test sample is also augmented 10 times and the estimated probabilities averaged

- **Did you use handcrafted features?** ( ) Yes, (X) No
  If yes, please detail:

- **Did you use any face / hand / body detection, alignment or segmentation strategy?** (X) Yes, ( ) No
  If yes, please detail: It is used implicitly on the 3rd party OpenPose keypoints detection

- **Did you use any pose estimation method?** (X) Yes, ( ) No
  If yes, please detail: We use OpenPose

- **Did you use any fusion strategy of modalities?** ( ) Yes, (X) No
  If yes, please detail:

- **Did you use ensemble models?** (X) Yes, ( ) No
  If yes, please detail: We have trained the MS-G3D model using Openpose keypoints and another model trained with the Joints (difference of connected keypoints, or bones), the we averaged the estimated class posteriors. The final submission uses a simple output average of the ensemble of 2 models trained with different seed and mini-batch/data augmentation schedule

- **Did you use any spatio-temporal feature extraction strategy?** (X) Yes, ( ) No
  If yes, please detail: The model itself, MS-G3D, implicitly extracts spatial-temporal features

- **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

Code repository: `https://github.com/ManuelGTM/ChaLearn_2021_Looking_at_People_Large_Scale_Signer_Independent_Isolated_SLR_CVPR_Challenge.git`

REFERENCES

[1] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, "Disentangling and unifying graph convolutions for skeleton-based action recognition,"

in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[2] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4724–4733.

[3] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Springer International Publishing, 2018, pp. 318–335.

[4] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.