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: Wenbin Wu
- Username on Codalab: wenbinwuee
- Team leader affiliation: **Netease Games**
- Team leader address: Kingold International Financial Center, No. 62 Jinsui Road, Tianhe District, Guangzhou
- Team leader phone number: +86-13660213575
- Team leader email: wenbinwuee@163.com
- Name of other team members (and affiliation): Xiang Gao(**SCUT**), Shuying Liu(**Netease Games**)
- Team website URL (if any):

II. Contribution details

A. Title of the contribution

We use RGB, optical flow and human segmentation data to train several models and ensemble these models to get the final results.

B. Introduction and Motivation

We use SlowFast [2], SlowOnly [2] and TSM [3] to train the several models, we believe that ensemble the strongest models can get more stonger results. And we use optical flow data, segmented data to train our model for we think that the feature extracted from the dynamic ROI is important for sign language recognition.

C. Representative image / workflow diagram of the method

Please see Fig 1 for our workflow. For details, please refer to the README of our code. `https://github.com/Koooko96/Chalearn2021code`.

D. Detailed method description

Our team believe that dynamic action cues for sign language recognition, in addition to RGB data, we use optical flow sequences and segmented data(getting from removing the backgrond by human segmentation) to train our model for not only reducing the possibilities of overfitting to backgrond, but leading the network to focus on the dynamic ROI. We consider that high performance can be obtained by a ensemble of various SOTA models, so we use the above datasets to train SlowFast [2], SlowOnly [2], and TSM [3] and fuse all the network prediction scores to get the final results, one important thing is that the pretrained model on Kinects-400 can impove the training results. The basic model used in this method is Resnet 50, and we also use non local network [4] and multigrid [5] to train our network. The training stage was carried out on a 8 x 3090Ti GPUs with 24GB GPU memory footage for each GPU.

E. Challenge results and final remarks

Fill Table I with your obtained results, shown in the leaderboard of the challenge associated to the **Challenge Track** you defined in Section I (RGB[1] or RGB+D[2]). Note that if you joined the challenge in the test phase, keep the "development" row blank.

TABLE I
LEADERBOARD: RESULTS OBTAINED BY THE PROPOSED APPROACH.

Phase	Track	Rank position	Rec. Rate
Development			
Test	RGB	5	0.9655

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: optical flow

- **Did you use pre-trained models?** (X) Yes, ( ) No
  If yes, please detail: slowfast models pretrained on

[1]RGB: `https://competitions.codalab.org/competitions/27901#results`
[2]RGB+D: `https://competitions.codalab.org/competitions/27902#results`

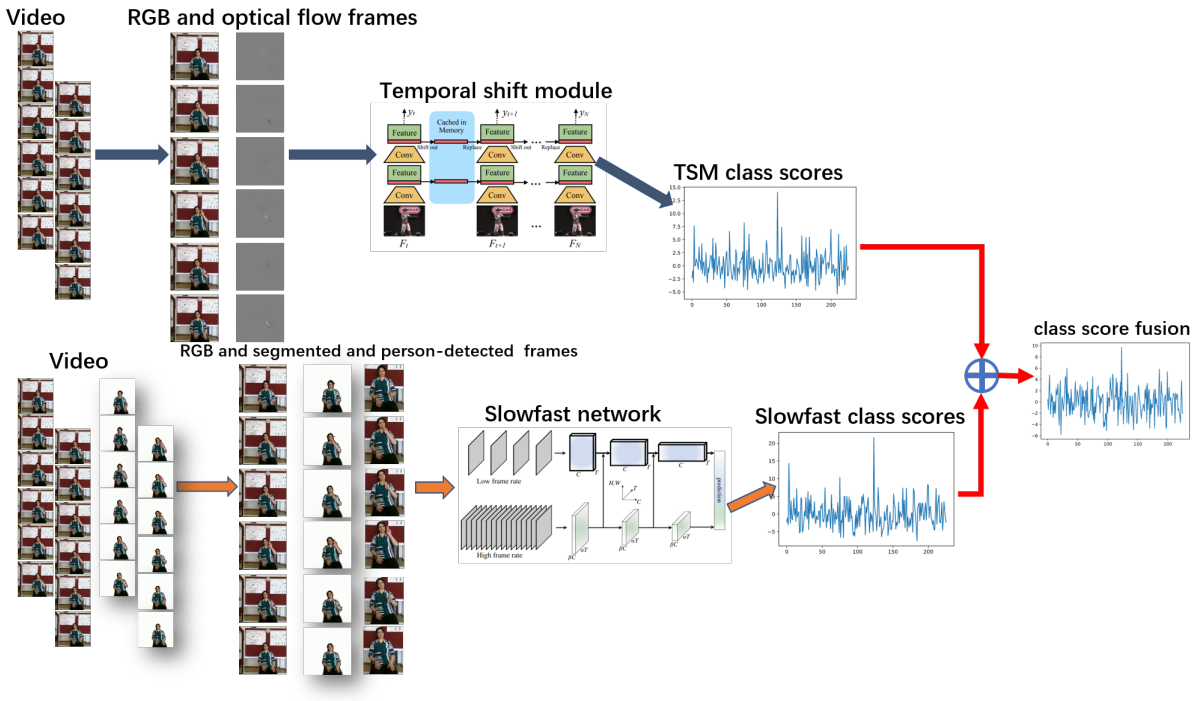

Fig. 1. workflow of our method

Kinects400

- **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?** ( ) 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: body segmentation

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

- **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: read our README on github

- **Did you use any spatio-temporal feature extraction strategy?** (X) Yes, ( ) No
  If yes, please detail: slowfast, slowonly, TSM

- **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

Link to a code repository with complete and detailed instructions so that the results obtained on Codalab can be reproduced locally. This includes a list of requirements, pre-trained models, and so on. Note, training code with instructions is also required. This is recommended for all participants and mandatory for winners to claim their prize. **Organizers strongly encourage the use of docker to facilitate reproducibility**.

Code repository: `https://github.com/Koooko96/Chalearn2021code`

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] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE Computer Society, 2019, pp. 6201–6210.

[3] J. Lin, C. Gan, and S. Han, "Tsm: Temporal shift module for efficient video understanding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7083–7093.

[4] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” 2017.
[5] C.-Y. Wu, R. Girshick, K. He, C. Feichtenhofer, and P. Krähenbühl, “A Multigrid Method for Efficiently Training Video Models,” in *CVPR*, 2020.