ECCV 2022 Sign Spotting Challenge

Fact sheet

This is the fact sheet's template for the ECCV 2022 Sign Spotting Challenge. 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, following the schedule and instructions ("Wining solutions (post-challenge)", Fact Sheets) provided in the Challenge webpage.

I. Team details

- Team leader name: Weichao Zhao
- Username on Codalab: th
- Team leader affiliation: University of Science and Technology of China
- Team leader email: saruka@mail.ustc.edu.cn
- Name of other team members (and affiliation): Hezhen Hu, Landong Liu, Kepeng Wu, Wengang Zhou, and Houqiang Li (University of Science and Technology of China)
- Team website URL (if any): `https://ustc-slr.github.io/`
- Competition track (mark with X one single option)[1]:
  - (X) Track 1: **MSSL** (multiple shot supervised learning).
  - ( ) Track 2: **OSLWL** (one shot learning and weak labels).

II. Contribution details

A. Title of the contribution

MSSL Track aims to spot the query signs in the continuous videos with their precise time stamps. In this challenge, we construct a two-stage framework, which consists of feature extraction and temporal sign action localization. Our simple framework ranked 2nd in the test phase of this track.

B. Representative image / workflow diagram of the method

The overview of our method is illustrated in Fig. 2.

C. Detailed method description

Our framework contains two important stages. Firstly, we utilize multiple modalities to extract robust spatio-temporal feature representation depicting the continuous sign video. Then we adopt a Transformer backbone to identify actions in time and recognize their categories.

Preprocessing. Since the signer only occupies a relatively small size of the original video, we utilize MMDetection [1] to first detect the signer spatial location, which is followed by MMPose [2] to extract the body and hand poses. Then, the detected pose is utilized to crop the upper-body patch of the signer. Besides, the flow is calculated by the TV-L1 algorithm. Finally, three types of data are generated, *i.e.,* RGB, flow and pose modalities.

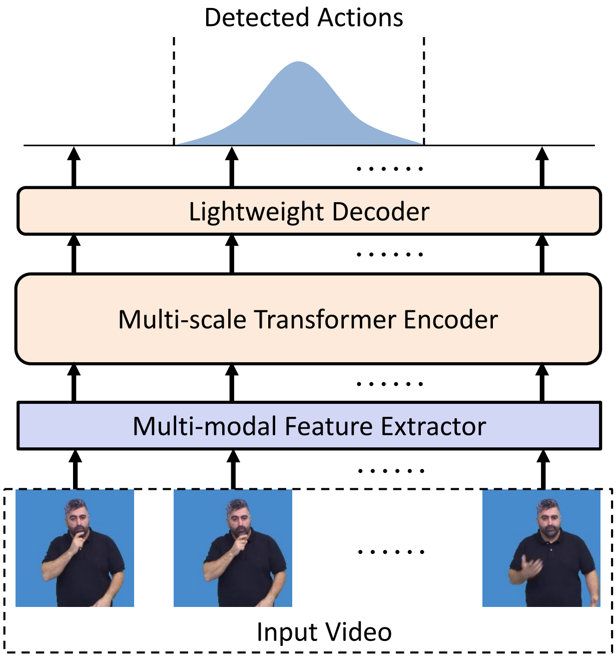
Detected Actions
Lightweight Decoder
Multi-scale Transformer Encoder
Multi-modal Feature Extractor
Input Video

Fig. 1. The workflow diagram of the method.

Feature extraction. For different modalities, we utilize different backbones for their complementary representations. All representations are extracted in the clip level with the receptive field and stride as 8 and 2, respectively. **RGB** modality is fed into the BSL-1k [3] pre-trained I3D [4] backbone. **Flow** modality is processed by the AUTSL [5] pre-trained I3D [4] backbone. For the **Pose** modality, we utilize pre-trained GCN [6], [7] to extract the body and hand cues, which contain the main meaning of the sign language. The extracted features from these modalities are concatenated for the next localization stage.

Temporal sign action localization. We leverage the strong modeling capability of Transformer to perform localization similar to [8]. Specifically, it combines a multi-scale feature representation with local self-attention, and uses a light-weighted decoder to classify every moment

[1] If you participated in more than one competition track, you need to share with the organizers one fact sheet per track.

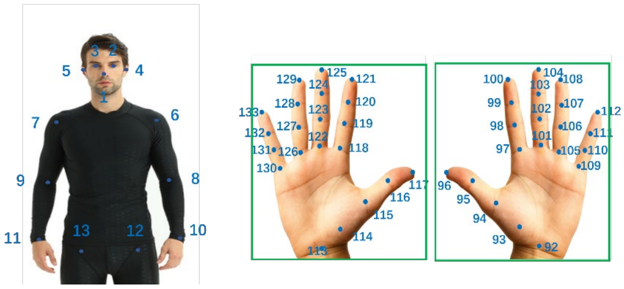

Fig. 2. The view of utilized mmpose skeleton.

in time and estimate the corresponding action boundaries. This framework is simple yet effective without using action proposals or relying on pre-defined anchor windows.

Training and inference settings. In our experiment, all networks are implemented by PyTorch and trained on NVIDIA RTX 3090. During training, the focal loss for sign action classification and generalized IoU loss for distance regression are adopted. We utilize the Adam optimizer with the peak learning rate as 2e-4. The training lasts 100 epochs, with 10 epochs as the warm-up.

During inference, we take the output for every time step, which is organized as the triplet sign action confidence score, onset and offset of the action. These candidates are further processed via NMS to remove highly overlapping instances, which leads to the final localization outputs.

D. Challenge results

Fill Table I with your obtained results, shown in the leaderboard of the challenge.

TABLE I
RESULTS FROM LEADERBOARD (TEST PHASE) OBTAINED BY THE PROPOSED APPROACH.

Rank position	avg F1
2	0.566752

E. Final remarks

Our method leverages the strong capability of the Transformer backbone to extract multi-modal information from the continuous sign video. Further exploring more effective multi-modal fusion strategies is desirable.

III. Additional Method Details

Please, reply if your challenge entry considered (or not) the following strategies and provide a brief explanation. For each question, mark with X one single option.

- **Did you use pre-trained models?** (X) Yes, ( ) No
  If yes, please detail:
  I3D (RGB); I3D (Flow); GCN (Pose).
- **Did you use external data?** ( ) Yes, (X) No
  If yes, please detail:
- **Did you use any kind of depth information (e.g., 3D pose estimation trained on RGBD data)?** ( ) Yes, (X) No
  If yes, please detail:
- **At the final phase, did you use the provided validation set as part of your training set?** (X) Yes, ( ) 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:
  We utilize MMDetection to localize the bounding boxes of signers.
- **Did you use any pose estimation method?** (X) Yes, ( ) No
  If yes, please detail:
  We utilize MMPose to extract the body and hand poses.
- **Did you use any spatio-temporal feature extraction strategy?** (X) Yes, ( ) No
  If yes, please detail:
  We utilize the I3D backbone for extracting features from RGB and flow modalities. Besides, GCN is adopted for feature representation of the pose modality.
- **Did you explicitly classify any attribute (e.g., gender/handedness)?** ( ) 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:
- *Just answer the following question if this fact sheet is associated with an **OSLWL track** submission.*
  Did you use MSSL train/val data and/or OSLWL val data as a means of unsupervised model training?
  ( ) 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://mailustceducn-my.sharepoint.com/:f:/g/personal/alexhu_mail_ustc_edu_cn/EsXmzk8AqEpPpytPmfl5zQQBegaAcFvZKIOuozg3EPBsWQ?e=3gw7tO`

References

[1] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, Z. Zhang, D. Cheng, C. Zhu, T. Cheng, Q. Zhao, B. Li, X. Lu, R. Zhu, Y. Wu, J. Dai, J. Wang, J. Shi, W. Ouyang, C. C. Loy, and D. Lin, "MMDetection: Open mmlab detection toolbox and benchmark," *arXiv*, 2019.

[2] M. Contributors, "Openmmlab pose estimation toolbox and benchmark," https://github.com/open-mmlab/mmpose, 2020.

[3] S. Albanie, G. Varol, L. Momeni, T. Afouras, J. S. Chung, N. Fox, and A. Zisserman, "BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues," in *ECCV*, 2020, pp. 35–53.

[4] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *CVPR*, 2017, pp. 6299–6308.

[5] 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.

[6] Y. Cai, L. Ge, J. Liu, J. Cai, T.-J. Cham, J. Yuan, and N. M. Thalmann, "Exploiting spatial-temporal relationships for 3D pose estimation via graph convolutional networks," in *ICCV*, 2019, pp. 2272–2281.

[7] H. Hu, W. Zhao, W. Zhou, Y. Wang, and H. Li, "SignBERT: Pre-training of hand-model-aware representation for sign language recognition," in *ICCV*, 2021, pp. 11 087–11 096.

[8] C. Zhang, J. Wu, and Y. Li, "Actionformer: Localizing moments of actions with transformers," *arXiv preprint arXiv:2202.07925*, 2022.