ECCV 2022 Sign Spotting Challenge

Fact sheet

I. Team details

- Team leader name: Xin Yu
- Username on Codalab: Mikedddd
- Team leader affiliation: University of Technology Sydney and Netease
- Team leader email: xin.yu@uts.edu.au
- Name of other team members (and affiliation): Beibei Lin, Xingqun Qi, Chen Liu, Hongyu Fu, Lincheng Li
- Team website URL (if any):
- Competition track (mark with X one single option)[1]:
  - ( ) Track 1: **MSSL** (multiple shot supervised learning).
  - (X) Track 2: **OSLWL** (one shot learning and weak labels).

II. Contribution details

A. Multi-modal Fusion via a Combined Sign Spotting Loss

The advantages of the proposed method are summarized as follows:

- We propose a multi-modal framework which extracts sign language features from RGB images, 2D-pose based and 3D angle-axis pose based joints to pursue high spotting accuracy in the challenge.
- We introduce a novel sign spotting loss function by combining the triplet loss and cross-entropy loss to obtain more discriminative feature representations. The triplet loss is used to maximize the inter-class distances and minimize the intra-class distances, and cross-entropy is applied for the sign classification.
- we propose a Top-K transferring technique to address the domain gap between the gallery set and the query set to further improve the sign retrieval performance.

B. Detailed method description

The pipeline of the proposed framework is shown in Figure 1. We employ three models, including 2D-Pose based SL-GCN, 3D-Pose based SL-GCN, and I3D-MLP, to extract sign language features. The input of 2D-Pose based SL-GCN is 2D key points that are generated by HRNet [1]. The input of 3D-Pose based SL-GCN is angle-axis 3D pose joints that are generated by Frankmocap [2]. The input of I3D-MLP is RGB images.

During the training phase, we exploit a triplet loss and a cross-entropy loss as a combined loss to train our models [3]. Note that each model is trained separately. In the inference phase, we employ our three models to extract visual features and then fuse these multi-modality features for the sign spotting task.

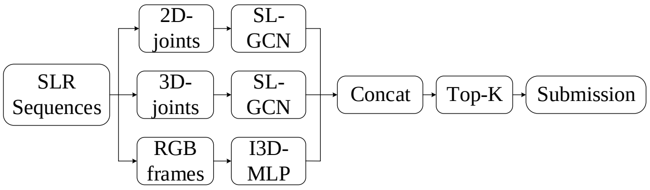

Fig. 1. The inference pipline of our framework.

The competition dataset contains two subsets: the gallery set (20 videos, labeled data) and the query set (607 videos, unlabeled data). This track aims to locate the start and end frames of certain signs from the query videos. The signs are specified by the videos' names and they are from the gallery set with only one exemplar available. To achieve this goal, we implement our method in the following steps:

First, we observe there is an obvious domain gap between isolated signs and continuous signs. Therefore, we trim the isolated sign videos by removing the video frames before hands up and after hands down. Then, we feed the gallery sign videos into our model to generate feature representations. As a result, we obtain 20 feature representations in total, each of which represents an isolated sign.

Second, for each video from the query set, we utilize a sliding window to crop video clips from the beginning to the end of a video. Then, we input all clips into the proposed model to extract feature representations. Here, the stride of the sliding window is 1.

Third, all feature representations from this video are used to calculate the Cosine Distance with respect to the corresponding feature representation of the sign provided in the gallery set. From the video name, we also know which sign to be spotted in a certain video. The clip that has the maximal similarity with the gallery sign representation are regarded as our retrieval result. Here, the sliding window size is set to 16 frames for the sign spotting task in all our experiments. In other words, we found that using 16 frames, we can achieve the best performance on the validation set.

Fourth, we propose a Top-K transferring technique to address the domain gap between the gallery set and the query set. After the second step, we will obtain many retrieved results for each isolated sign. We sort the distances of all retrieved results and find the most similar K clips. Then,

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

each feature representation from the gallery set will be updated by the average of the feature representations of the top-K retrieved clips. Iteratively, the updated feature representation for a certain sign will be used to retrieve signs from continuous sign videos again.

In this paper, the length of the sliding window is 16 frames. The parameter K is set to 5 and we use the Top-K transferring technique once.

C. Challenge results

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

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

Rank position	avg F1
2	0.559295

D. Final remarks

Please identify the pros and cons (if any) of the proposed approach.

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?** Yes.
  In this paper, we adopt two networks, including SL-GCN [4] and I3D-MLP [5], to extract features. The pretrained model of SL-GCN is provided by [4], while I3D-MLP is pretrained on the BSlDict dataset [6].

- **Did you use external data?** No.

- **Did you use any kind of depth information (e.g., 3D pose estimation trained on RGBD data)?** Yes.
  In this project, we employ the pre-trained 3D pose extraction model Frankmocap [2], [7] to obtain the 3D joints as training data for our networks.
- **At the final phase, did you use the provided validation set as part of your training set?** Yes.
  During the final phase, we combine both the dataset from the track 1 (60 classes) and the validation set (20 classes) to train our network.

- **Did you use other regularization strategies/terms?** No.

- **Did you use handcrafted features?** No.

- **Did you use any face / hand / body detection, alignment or segmentation strategy?** Yes.
  We design an alignment strategy to pre-process data. Firstly, we use HRNet [1] to get 133 key points of human pose, and then select the nose as the central landmark to crop the video frames. In this way, signers are in the center of images of 512 × 512 pixels.
- **Did you use any pose estimation method?** Yes.
  We use the HRNet [1] to estimate 133-point whole-body key points from the RGB videos [4]. Then, we select 27 key points to construct a skeleton graph. Besides, we leverage the Frankmocap [2] to acquire the 3D angle-axis based on 54 whole-body joints from the RGB videos. Afterward, 39 joints are selected to construct the 3D skeleton graph.
- **Did you use any spatio-temporal feature extraction strategy?** Yes.
  As aforementioned, we adopt both SL-GCN and I3D-MLP to extract features. The input of SL-GCN is a spatio-temporal skeleton graph [4]. SL-GCN utilizes a decoupled spatial convolutional layer (Decouple SCN) and a temporal convolutional layer (TCN) to model spatial and temporal dynamics. I3D-MLP [5] extracts spatio-temporal information by 3D convolutions.

- **Did you explicitly classify any attribute (e.g., gender/handedness)?** No.

- **Did you use any bias mitigation technique (e.g. rebalancing training data)?** No.

- *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? No.

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://drive.google.com/drive/folders/1bjlRq1dJlVhSuqKG-GqnvzDRVD5L-GcQ?usp=sharing`

REFERENCES

[1] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5693–5703.

[2] Y. Rong, T. Shiratori, and H. Joo, "Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration," in *IEEE International Conference on Computer Vision Workshops*, 2021.

[3] B. Lin, S. Zhang, and X. Yu, “Gait recognition via effective global-local feature representation and local temporal aggregation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 648–14 656.

[4] S. Jiang, B. Sun, L. Wang, Y. Bai, K. Li, and Y. Fu, “Skeleton aware multi-modal sign language recognition,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3413–3423.

[5] J. Carreira and A. Zisserman, “Quo vadis, action recognition? a new model and the kinetics dataset,” in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[6] L. Momeni, G. Varol, S. Albanie, T. Afouras, and A. Zisserman, “Watch, read and lookup: learning to spot signs from multiple supervisors,” in *Proceedings of the Asian Conference on Computer Vision*, 2020.

[7] H. Joo, N. Neverova, and A. Vedaldi, “Exemplar fine-tuning for 3d human pose fitting towards in-the-wild 3d human pose estimation,” *3DV*, 2021.