ECCV 2022 Sign Spotting Challenge

Fact sheet

I. TEAM DETAILS

- Team leader name: Ryan Wong
- Username on Codalab: ryanwong
- Team leader affiliation: University of Surrey
- Team leader email: rwong@surrey.ac.uk
- Name of other team members (and affiliation):
  Necati Cihan Camgoz (University of Surrey)
  Richard Bowden (University of Surrey)
- Team website URL (if any):
- Competition track:
  - ( ) Track 1: **MSSL** (multiple shot supervised learning).
  - (X) Track 2: **OSLWL** (one shot learning and weak labels).

II. CONTRIBUTION DETAILS

A. Feature Based Sign Spotting

We introduce a method for sign spotting using existing I3D models pretrained on sign language datasets. We first show how these models can be used for identifying important frames from isolated sign dictionaries. Then we demonstrate the use case of I3D models for feature extraction on both the isolated sign videos and co-articulated sign videos for sign spotting based on the cosine similarity of extracted features.

B. Method

1) Finding important frames in isolated sign videos: Since the isolated sign videos have frames where the signer is in their resting pose or frames with irrelevant information. We needed to identify a method to remove these frames. While previous approaches used keypoint based methods to remove resting pose signs [1] we demonstrate an alternative, feature based approach using existing pretrained sign models.

We use an I3D model [2] pretrained on the WLASL dataset [3] as our feature extractor. The features are extracted after spatio-temporal global average pooling. For each frame in the isolated sign we repeat the frame 16 times and feed it into the I3D model which outputs a feature vector of size 1024. We then compute the cosine similarity between each of the frame features across all of the isolated sign videos, which creates a cosine similarity matrix $z \in \mathbf{R}^{n \times n}$ where n is the total number of frames in the isolated sign video dictionary.

The sum of the cosine similarity across the second axis $z_{sum} \in \mathbf{R}^{n}$ is calculated. z_{sum} is used to determine how common the sign frame is within the sign video dictionary, where higher values of z^{i}_{sum} at index i indicate a higher frequency of the sign frame. In the case of isolated sign videos the most common frames are resting pose, we therefore threshold z_{sum} with a value of $max(z_{sum}) - std(z_{sum}) * 2$ keeping frames for each isolated sign sequence which are below the threshold. For each of the isolated sign videos we select the minimum and maximum index frames which satisfies the above condition as the start and stopping point of the isolated sign label.

2) Identifying similarities between isolated and continuous sign videos: For each isolated query video we randomly select 8 frames (sorted by indices) of the important frames identified in section II-B.1. We apply randomly applying augmentation such as colour jitter, random cropping, grayscale, horizontal flipping and resizing to 224×224. This query sequence is used as input into an I3D model for feature extraction of a vector $q \in \mathbb{R}^{1 \times 1024}$.

Similarly the co-articulated sign video is used as input into the I3D model for feature extraction with a stride of 1 and sequence length of 8, which also undergoes randomly applying augmentations, such as, colour jitter, random cropping, grayscale, horizontal flipping and resizing to 224×224, where output sequence of feature vectors $k \in \mathbb{R}^{t \times 1024}$ is obtained (t is the number of frames in the co-articulated sign video).

The cosine similarity between q and k is then calculated to obtain the similarity matrix $s \in \mathbb{R}^{t \times 1}$.

We repeat this process with 64 different combinations of query sequences (random data augmentation and random frame selection) and 64 different co-articulated sign video (random data augmentation), obtaining 4096 similarity scores such that $s_{all} \in \mathbb{R}^{t \times 4096}$. The mean of all similarity scores at each time step is then calculated to obtain the final similarity score $s_{final} \in \mathbb{R}^{t}$.

3) Computing Sign Matches: Using the similarity score s_{final}, we compute the normalise similarity score s_{norm} by making the assumption that there exists at least 1 occurrence of the isolated sign in the co-articulated sequence by dividing s_{final} by the maximum value in the sequence. Any indices in s_{norm} greater than a threshold 0.9 is considered a match between the isolated sign video and continuous sign segment.

Since we given that around 10% of the OWLSL dataset do not have matching spotted signs we remove spotting predictions where maximum of s_{final} is not greater than 0.36, which eliminates around 10% of the test predictions.

For each time index with a matched spotting we include the 8 subsequent frames as spotting matches and combine

matching spottings if they are in range of 10 frames of each other.

4) Ensemble: Using the above setup we use 2 models from [4], 1 pretrained on WLASL [3] and the other on MSASL [5], taking the average between the normalised cosine similarities s_{norm}.

C. Challenge results

Table I shows the 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
3	0.514309

D. Final remarks

The advantages of the proposed approach are that no training is required for sign spotting and the use of existing pretrained sign models is used as feature extractors. A disadvantage is that ability for sign spotting is highly reliant on the features of the pretrained models ability to generalise to the sign spotting dataset.

III. ADDITIONAL METHOD DETAILS

- **Did you use pre-trained models?** (X) Yes, ( ) No
  Pretained models from [4] on WLASL and MSASL are used as feature extractors.
- **Did you use external data?** ( ) Yes, (X) No
- **Did you use any kind of depth information (e.g., 3D pose estimation trained on RGBD data)?** ( ) Yes, (X) No
- **At the final phase, did you use the provided validation set as part of your training set?** ( ) Yes, (X) No
- **Did you use other regularization strategies/terms?** ( ) Yes, (X) No
- **Did you use handcrafted features?** ( ) Yes, (X) No
- **Did you use any face / hand / body detection, alignment or segmentation strategy?** (X) Yes, ( ) No
  We used OpenPose [6] as a body detector to crop signer region.
- **Did you use any pose estimation method?** ( ) Yes, (X) No
- **Did you use any spatio-temporal feature extraction strategy?** (X) Yes, ( ) No
  Features are extracted from pretrained I3D models based on the output of the final convolutional layer after spatial temporal global average pooling. The full details are described in section II.
- **Did you explicitly classify any attribute (e.g., gender/handedness)?** ( ) Yes, (X) No
- **Did you use any bias mitigation technique (e.g. rebalancing training data)?**
  ( ) Yes, (X) No

  Did you use MSSL train/val data and/or OSLWL val data as a means of unsupervised model training?
  ( ) Yes, (X) No

IV. CODE REPOSITORY

Code repository: `https://github.com/ryanwongsa/ECCV22_Chalearn-OSLWL`

Data link: `https://drive.google.com/file/d/14YLgo-HwXt7TRX4PD4r5sBIULZRfp1tm/view?usp=sharing`

REFERENCES

[1] T. Jiang, N. C. Camgöz, and R. Bowden, "Looking for the signs: Identifying isolated sign instances in continuous video footage," in *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*. IEEE, 2021, pp. 1–8.

[2] 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.

[3] D. Li, C. Rodriguez, X. Yu, and H. Li, "Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2020, pp. 1459–1469.

[4] 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 *European conference on computer vision*. Springer, 2020, pp. 35–53.

[5] H. R. V. Joze and O. Koller, "Ms-asl: A large-scale data set and benchmark for understanding american sign language," *arXiv preprint arXiv:1812.01053*, 2018.

[6] 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.