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

I. TEAM DETAILS

- **Challenge Track** (RGBD):
- Team leader name: Hui Shen
- Username on Codalab: wz
- Team leader affiliation: Baidu, INC.
- Team leader address: No. 10 Shangdi 10th Street, Haidian District, Beijing 100085, The People's Republic of China
- Team leader phone number: +86 13161308878
- Team leader email: jhonjoe.c@gmail.com
- Name of other team members (and affiliation): DuVis
- Team website URL (if any):

II. CONTRIBUTION DETAILS

A. Abstract

We treat the challenge as a video classification problem. To this end, we propose two different but complementary models. Since the signer language recognition is highly correlated with the motion of human skeletons, our first model extracts human keypoints with off-the-shelf keypoint estimation methods[1], [2], and then utilizes the estimated skeletons for action recognition. This model can be seen as a two-stage framework, which divides the challenging video recognition problem into two relative simpler sub-tasks. However, such cascaded system highly relies on the outputs of its first stage, i.e. the quality of skeletons, which is a challenging task itself under the scenarios with rapid motion or severe occlusions. To address this issue, our second model directly feeds the entire video frames to the network and performs classification with the generated features. Finally, our entire system combines the skeleton model and video model in a principled way. As a result, we multi-model framework achieves 97.55 on the validation set of AUTSL dataset, and 98.34% on the test set of AUTSL dataset, ranking 2nd place in the challenge.

B. Introduction and Motivation

Sign Language Recognition, a primary communication tool for the deaf community, is an important task in computer vision. This task is challenging due to the intrinsic linguistic structures, subtle visual difference in spatial and temporal domain, as well as motion blur of human skeletons. In this challenge, we treat the problem as a video classification problem. To make our final model various and complementary, we create different modalities from raw video frames. Specifically, our utilizes three modalities in our method: the first one is human skeletons, including body skeletons, hand keypoints, and facial keypoints, the second one is optical flow, and the last one is raw video frames. Since sign language recognition is highly correlated with human skeletons, our first model is a skeleton-based classification network. We choose the GCN based methods[3], [4] due to their effectiveness of dependency modeling. Different from traditional video-based methods that treat optical flow symmetry to RGB frames, we enhance the skeletal data with the optical flow of the corresponding keypoints. In this way, each skeleton is aligned with predicted motion, and hence the model is more robust to noisy skeletons. Our second model is a video-based model[5], [6], which directly estimates classes using the entire RGB frames. We notice that human instances are relative stable during performing signs, so we first detect human instances of each frame using a human detector, and then for each sign video, we generate a union bounding box of the signer by the union of detected boxes of each frame. We then perform spatial data cropping for the pre-croped human instances. Finally, we fuses the skeleton-based model and video-based model in a score level fusion. The experiments show that our multi-modality model achieves 98.34% on the test set of the AUTSL dataset, ranking 2nd place of the challenge.

C. Multi-Media Sign Language Recognition

As we can see from the pipeline Fig.1, our method consist of two main source, human skeleton and rgb-video (rgbd-video).

1) skeleton-based action recognition: We utilize three pretrained models to extract human skeletons from the rgb video, specifically, Openpose[2], Alphapose[1] and our refined human skeleton detection model, we construct the human skeleton with total 78 points, including 18 points on body, 40 points on hands and 20 points on face respectively, besides we extract optical flow[7] from rgb video and generate 3D points from the detected 2D-keypoints with SemGCN[8], our data has 8 channels at last, with [i, j, s, fx, fy, x, y, z], where i and j is coordinate of keypoints on the original image, s is corresponding score of keypoints at location (i,j), fx and fy is the optical flow in x and y direction respectively, x, y and z is the coordinates generated by SegmGCN. Cause the video length is various, we padding the short video to 96 frames uniformly, for each video, we generate an $R^{96\times78\times8}$ matrix as input data. Then we define the adjacency matrix and adopt MS-G3D[3] to extract features, different from the original MS-G3D architecture, we add an SE[9] module to STGCN Module, except for joint and bone stream, we also train a complement graph[10] of joints stream. We also adopt the 2S-AGCN [4] and

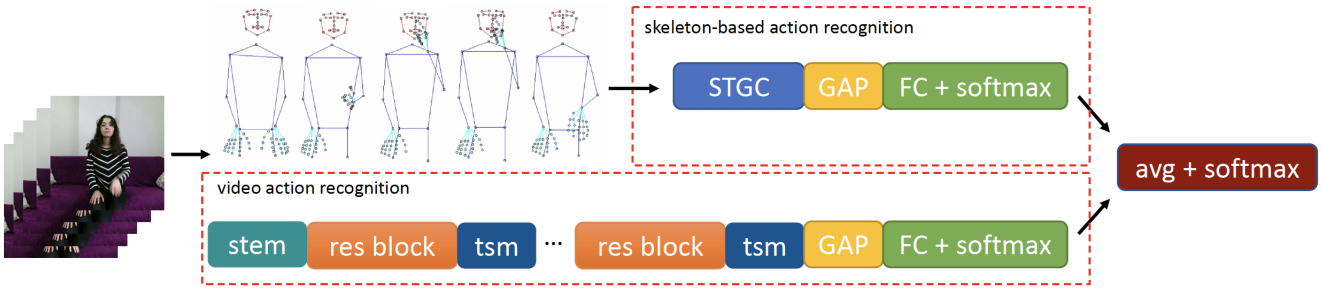

Fig. 1. the pipeline for multi-media action recognition

DGNN[11] extracting features to extent the diversity of our skeleton-based model.

For RGBD track, we simply append the normalized depth information to location and flow, the data have 6 channels with [i, j, s, fx, fy, d], where d is the normalized depth information at location (i, j) on depth image.

2) video action recognition: We employee TSM[5] and MVFNet[6] for video action recognition. Cause the video data is limited, our model is slightly over fitting on the training data, therefore we apply multiply data augmentations on the input video, such as color jitter, temporal jitter and random crop, furthermore, cause the background will interfere the recognition, we detect person bounding box by human skeleton and crop person out from the original image, which suppress the interference from background. Besides, we modify the MVFNet[6] slightly for extracting robust representation, instead of inserting shift module at last two stages, we insert the shift module in each stage at the beginning of the block.

D. Experiments setup

1) skeleton-based action recognition: We implement our method on Tesla V100 x8 GPUS, the batch size is set to 96, total 80 training epochs, and SGD optimizer with momentum 0.9 is selected. The initial learning rate was set to 0.01, after 600 iterations linearly warm up, the learning rate rise to 0.25, and decays by 0.1 at 55 and 75 epoch, to prevent over fitting, we set the weight decay to 5e-4. Compared with training from pretrained model on NUT-RGBD[12] and Kinetics-400[13], training from scratch achieve nearly identical results. So all our model is trained directly with random initialized parameters.

2) video action recognition: The video action recognition is implemented on Tesla V100 x8 GPUS, the batch size is set to 64, total 140 training epochs with SGD. blueWe perform uniform sampling with 32 clips and 24 clips in our experiments. We set the beginning learning rate nearly to 0.01, with 700 iterations warm up, the learning rate rise to 0.01, and decays by 0.1 at 80 and 130 epochs. The momentum is set to 0.9, weight decay is 1e-3, we apply the FP16 precision which can improve the final result slightly. The input resolution is set to 280x280 for keep more details on hands. In addition, we trained pretrained model on WLASL dataset[14], we can improve the clssification precision significantly.

E. Challenge results and final remarks

Table I is the final results on the test set, our method achieve a top-1 accuracy of 98.34% on test set.

TABLE I

LEADERBOARD: RESULTS OBTAINED BY THE PROPOSED APPROACH.

Phase	Track	Rank position	Rec. Rate
Development	RGBD		
Test	RGBD	2	98.34

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?** ( ) Yes, ( ) No
  If yes, please detail:
  We simply append the normalized depth information to skeleton data as mentioned in section II-C.1
- **Did you use pre-trained models?** ( ) Yes, ( ) No
  If yes, please detail:
  For RGB video action recognition, we utilize the WLASL[14] train a model as pretrained model.
- **Did you use external data?** ( ) Yes, ( ) No
  If yes, please detail:
  We have extra annotated keypoints data, besides, we refine the keypoints detection model with the predicted keypoints on Chanllenge dataset[15]
- **Did you use other regularization strategies/terms?** ( ) Yes, ( ) No
  If yes, please detail:
  We employee L2-Norm with weight decay 1e-3 on RGB video recognition and 5e-4 on skeleton-based action recognition
- **Did you use handcrafted features?** ( ) Yes, ( ) No
  If yes, please detail:
  No
- **Did you use any face / hand / body detection, alignment or segmentation strategy?** ( ) Yes, ( ) No

If yes, please detail:
As mentioned before, we refine the skeleton detection model on predicted results with high confidence

- **Did you use any pose estimation method?** ( ) Yes, ( ) No
  If yes, please detail:
  We adopt three keypoints detection model, Openpose[2], Alphapose[15] and our refined keypoints detection model
- **Did you use any fusion strategy of modalities?** ( ) Yes, ( ) No
  If yes, please detail:
  We employee xgboost at final stage to enhance the results further
- **Did you use ensemble models?** ( ) Yes, ( ) No
  If yes, please detail:
  Our last submitted results ensemble 7 models, including 5 skeleton-based models and 2 rgb video models
- **Did you use any spatio-temporal feature extraction strategy?** ( ) Yes, ( ) No
  If yes, please detail:

- **Did you explicitly classify any attribute (e.g. gender)?** ( ) Yes, ( ) No
  If yes, please detail:
  No
- **Did you use any bias mitigation technique (e.g. rebalancing training data)?** ( ) Yes, ( ) No
  If yes, please detail:
  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://github.com/CodesFarmer/MS-G3D-SLR.git`

References

[1] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *ICCV*, 2017.

[2] 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.

[3] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, "Disentangling and unifying graph convolutions for skeleton-based action recognition," in *CVPR*, 2020.

[4] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *CVPR*, 2019.

[5] J. Lin, C. Gan, and S. Han, "Tsm: Temporal shift module for efficient video understanding," in *ICCV*, 2019.

[6] W. Wu, D. He, T. Lin, F. Li, C. Gan, and E. Ding, "Mvfnet: Multi-view fusion network for efficient video recognition," in *AAAI*, 2021.

[7] B. K. Horn and B. G. Schunck, "Determining optical flow," *Artificial Intelligence*, vol. 17, no. 1, pp. 185–203, 1981.

[8] L. Zhao, X. Peng, Y. Tian, M. Kapadia, and D. N. Metaxas, "Semantic graph convolutional networks for 3d human pose regression," in *CVPR*, 2019.

[9] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *CVPR*, 2018.

[10] K. Cheng, Y. Zhang, C. Cao, L. Shi, J. Cheng, and H. Lu, "Decoupling gcn with dropgraph module for skeleton-based action recognition," in *ECCV*, 2020.

[11] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with directed graph neural networks," in *CVPR*, 2019.

[12] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+d: A large scale dataset for 3d human activity analysis," in *IEEE Conference on Computer Vision and Pattern Recognition*, June 2016.

[13] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The kinetics human action video dataset," *CoRR*, 2017.

[14] 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 *The IEEE Winter Conference on Applications of Computer Vision*, 2020, pp. 1459–1469.

[15] ChaLearnLAP. CVPR 2021 ChaLearn Looking at People Large Scale Signer Independent Isolated SLR Challenge. [Online]. Available: http://chalearnlap.cvc.uab.es/challenge/43/description/