Fully Convolutional Networks for Continuous Sign Language Recognition
Supplementary Material

Ka Leong Cheng[1], Zhaoyang Yang[2], Qifeng Chen[1], and Yu-Wing Tai[1,3]

[1] The Hong Kong University of Science and Technology
{klchengad,cqf}@ust.hk
[2] Tencent
yangzhaoyang6@126.com
[3] Kwai Inc.
yuwing@gmail.com

1 Online Recognition

To better demonstrate the advantages of our proposed FCN method over the previous LSTM based method, we build a back-end server system to simulate the online recognition process of sign language. The server (receiver) receives frame packages of signing videos sent by the client (sender) and returns the recognition results as responses to the client. The receiving and recognition processes are ongoing parallelly. The system for the online recognition simulation is tested in the Linux system on Alienware, and the recognition model runs on an NVIDIA's GeForce GTX 1080 Graphics Card, with a memory size of 8 GB.

The system uses slightly different implementations for the two methods in order to make the recognition online, where the system can provide intermediate results for the received video frames. Since the LSTM based model needs all the information up to the current frame to give intermediate recognition, the memory has to cache all the in-taken frames for the LSTM based model. Differently, the proposed FCN method only requires the memory to cache a small window size of frames to provide the recognition results at a specific frame step. The window size is decided by the accumulated receptive field of 1D-CNNs in the two-level gloss feature encoder.

We include some testing sample results for both the RWTH-PHOENIX-Weather-2014 (RWTH) dataset [2] and the Chinese Sign Language (CSL) dataset [1] in the supplementary video file "demo.mp4". In the video, we use red to denote the deletion error, green to denote the insertion error, and blue to denote the substitution error. We found that our method can give intermediate results word by word and provide a relatively better user experience. However, the LSTM based method usually gives some inaccurate intermediate results, which may somehow confuse the users of online recognition. At the same time, the memory usage of the LSTM based method accumulates significantly along time, almost occupying all the 8 GB of the GPU. In contrast, our method can do recognition with relatively small and constant memory size. A low memory usage rate is

significant for the deployment of online recognition, as the saved memory can be used to do recognition work for other users, which can help save the resources and reduce costs. From our observation, we can conclude that our method is more user-friendly and more efficient for online sign language recognition.

2 Qualitative Comparison

We show more qualitative comparisons on full videos in the RWTH dataset. A major feature of the RWTH dataset is that it is made up of unique sentences, so all the sentences in the testing sets are not seen during the model training process. Hence, the performance of the models on the RWTH dataset can better reveal their generalization qualities.

We compare results obtained from the LSTM based model and the proposed network. When training with LSTMs, we followed the designs in SF-Net [3]. Results on the testing set are shown in Figure 1.

For the LSTM based model, errors are usually clustered and more likely to caused by the contextual information. The clustering property of errors in the LSTM based model indicates that a predicted error has a much negative influence on the glosses around. Also, the LSTM based model tends to give more insertion and deletion errors to form ordered gloss sub-sequences that appear more frequently in the training set. In contrast, for the proposed network, errors are usually isolated, indicating that a predicted error has less impact on the recognition results of adjacent glosses. Overall, on the RWTH dataset, we can see that the proposed network shows much better generalization and recognition capability in recognizing unseen sign language videos.

References

1. Huang, J., Zhou, W., Zhang, Q., Li, H., Li, W.: Video-based sign language recognition without temporal segmentation. In: Proceedings of AAAI Conference on Artificial Intelligence. pp. 2257–2264 (2018) 1
2. Koller, O., Forster, J., Ney, H.: Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. Computer Vision and Image Understanding **141**, 108–125 (2015) 1
3. Yang, Z., Shi, Z., Shen, X., Tai, Y.W.: Sf-net: Structured feature network for continuous sign language recognition. arXiv preprint arXiv:1908.01341 (2019) 2

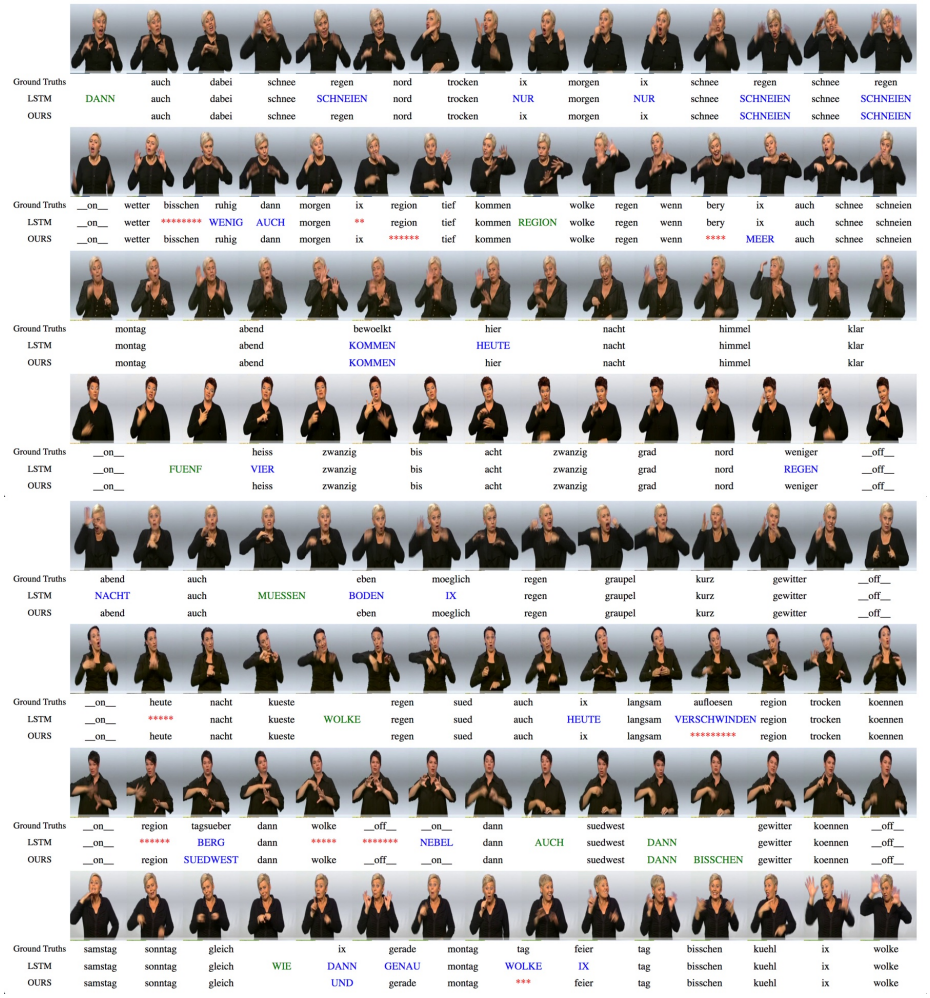

Fig. 1. Comparisons on full videos in the testing set of the RWTH dataset. Deletion, insertion, substitution errors are colored in red, green, blue, respectively. Glosses containing "_" are special flags that mark specific conditions in the signing sequences. "****" represents missing glosses