

Skeleton Aware Multi-modal Sign Language Recognition

Songyao Jiang[§], Bin Sun[§], Lichen Wang, Yue Bai, Kunpeng Li and Yun Fu
Northeastern University, Boston MA, USA

Abstract

Sign language is commonly used by deaf or speech impaired people to communicate but requires significant effort to master. Sign Language Recognition (SLR) aims to bridge the gap between sign language users and others by recognizing signs from given videos. It is an essential yet challenging task since sign language is performed with the fast and complex movement of hand gestures, body posture, and even facial expressions. Recently, skeleton-based action recognition attracts increasing attention due to the independence between the subject and background variation. However, skeleton-based SLR is still under exploration due to the lack of annotations on hand keypoints. Some efforts have been made to use hand detectors with pose estimators to extract hand key points and learn to recognize sign language via Neural Networks, but none of them outperforms RGB-based methods. To this end, we propose a novel Skeleton Aware Multi-modal SLR framework (SAM-SLR) to take advantage of multi-modal information towards a higher recognition rate. Specifically, we propose a Sign Language Graph Convolution Network (SL-GCN) to model the embedded dynamics and a novel Separable Spatial-Temporal Convolution Network (SSTCN) to exploit skeleton features. RGB and depth modalities are also incorporated and assembled into our framework to provide global information that is complementary to the skeleton-based methods SL-GCN and SSTCN. As a result, SAM-SLR achieves the highest performance in both RGB (98.42%) and RGB-D (98.53%) tracks in 2021 Looking at People Large Scale Signer Independent Isolated SLR Challenge. Our code is available at <https://github.com/jackysy/CVPR21Cha1-SLR>

1. Introduction

Sign language [14] is a visual language performed with the dynamic movement of hand gestures, body posture, and facial expressions. It is an effective and helpful approach for

Figure 1. Concept of our Skeleton Aware Multi-modal Sign Language Recognition Framework (SAM-SLR). All local and global motion information is extracted and utilized for final prediction.

deaf and speech-impaired people to communicate with others. Understanding and utilizing sign language requires a considerable time of learning and training which is not practical and feasible for the public. Moreover, sign language is affected by the language [59, 24, 67] (e.g., English and Chinese) and culture [35] which further limits its popularization potential. As machine learning and computer vision achieved great progress in the past decade, it is important to explore sign language recognition (SLR) which automatically interprets sign language and helps deaf-mute people communicate smoothly with others in their daily lives.

Compared with conventional action recognition, SLR is a more challenging problem. First, sign language requires both global body motion and delicate arm/hand gestures to distinctly and accurately express its meaning. Facial expression can be utilized to express emotions as well. Similar gestures can even impose various meanings depending on the number of repetitions. Second, different signers may perform sign language differently (e.g., speed, localism, left-hander, right-hander, body shape), making SLR more challenging. Collecting more samples from as many signers as possible is desired yet expensive.

Traditional SLR methods mainly deploy handcrafted features such as HOG [71] and SIFT [33]) associated with conventional classifiers like kNN and SVM [67, 11, 34]. As deep learning achieves significant progress, general video

[§]Equal contribution

This work was supported by the U.S. Army Research Office Award W911NF-17-1-0367.

