Abstract

Video-based sign language recognition is vital for improving communication for the deaf and hard of hearing. However, due to a lack of resources, creating and maintaining the quality of Thai sign language video datasets is challenging. To address this issue, we assess multiple models with a novel dataset of 90 signs, covering the full letters of alphabets, vowels, intonation marks, and numbers, as demonstrated by 43 signers. We investigate seven deep learning models with three distinct modalities for our analysis: video-only methods (including RGB-sequencing-based CNN-LSTM and VGG-LSTM), human body joint coordinate sequences (processed by LSTM, BiLSTM, GRU, and Transformer models), and skeleton analysis (using TGCN with graph-structured skeleton representation). A thorough assessment of these models is conducted across seven circumstances, encompassing single-hand postures, single-hand motions with one, two, and three strokes, and two-hand postures with static and dynamic point-on-hand interactions. The research highlights that the TGCN model is the optimal lightweight model in all scenarios. In single-hand pose cases, a combination of the Transformer and TGCN models of two modalities delivers outstanding performance, excelling in four particular conditions: single-hand poses, single-hand poses requiring one, two, and three strokes. In contrast, two-hand poses with static or dynamic point-on-hand interactions present substantial challenges, as the data from joint coordinates is inadequate due to hand obstructions stemming from insufficient coordinate sequence data and the lack of a detailed skeletal graph structure. The study recommends integrating RGB-sequencing with visual modality to enhance the accuracy of two-handed sign language gestures. Moreover, experimental results on our dataset show that our method outperforms previous state-of-the-art methods significantly in five out of seven conditional hand pose experiments, especially two-hand poses.

Keywords: Thai Finger Spelling, Sign Language Recognition, Deep Learning, Multimodal Learning, Benchmark Dataset.