ECE4880J Project Proposal

Group 19
Mohan Huang
Zhe Yu
Kaiqi Zhu

Real-Time Continuous Chinese Sign Language Recognition in Complex Environments

Objective

Nowadays, there are millions of deaf and hard-of-hearing individuals relying on sign language to communicate. However, they are likely to find it hard to communicate with those who are not familiar with this language, which can give rise to great difficulty in their daily life. Recent work (see related work) in computer vision and deep learning has offered solutions for this kind of sign language recognition. However, more optimization and investigation should still be implemented due to the lack of accuracy and ease of use (especially in Chinese sign language). Therefore, in this project, we aim to develop a continuous Chinese sign language (CSL) recognition system that takes short video clips as input and output sequences of glosses. Our primary goal is to improve the recognition accuracy under complex visual conditions. To achieve this, we plan to enhance the TFNet baseline by exploring stronger visual feature extractors or incorporating additional facial expressions and head motion to better capture the semantics of sign language.

Related Work

[1]:Yin, S., Camgoz, N. C., & Bowden, R. (2021). *Improving sign language translation with monolingual data by sign back-translation*. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 13104–13114. https://doi.org/10.1109/CVPR46437.2021.01291

[2]:Camgoz, N. C., Saunders, B., Sun, Y., & Bowden, R. (2020). *Sign language transformers: Joint end-to-end sign language recognition and translation*. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 10023–10033. https://doi.org/10.1109/CVPR42600.2020.01004

[3]:Zhu, Q., Li, J., Yuan, F., Fan, J., & Gan, Q. (2024). *A Chinese continuous sign language dataset based on complex environments*. arXiv. https://arxiv.org/abs/2409.11960

[4]:Zuo, R., Wei, F., & Mak, B. (2023). *Natural language-assisted sign language recognition*. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 14890–14900). https://doi.org/10.1109/CVPR52729.2023.01430

[5]:Jiao, P., Min, Y., Li, Y., Wang, X., Lei, L., & Chen, X. (2023). *CoSign: Exploring co-occurrence signals in skeleton-based continuous sign language recognition*. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 20676–20686). https://doi.org/10.1109/ICCV.2023.01867

Dataset

We will use the **CE-CSL dataset**, which contains **5,988 continuous Chinese sign language video clips** from **70+ real-life backgrounds**, addressing the generalization problem of previous CSL datasets. The dataset provides both **gloss-level annotations** and corresponding **full Chinese sentence translations**. It also includes an official **TFNet baseline model**.

At this stage, we do not plan to collect or use new datasets because different CSL datasets vary in size, annotation detail, and conditions, making integration difficult and potentially harmful. The CE-CSL dataset is already large and diverse enough to support effective training and evaluation. Focusing on it allows us to dedicate efforts to improving model accuracy without the extra complexity of handling multiple datasets.

Evaluation

We will evaluate our system using the standard **Gloss-level Word Error Rate (WER)**, which measures the difference between predicted gloss sequences and standard reference annotations in terms of substitutions, deletions, and insertions. A lower WER indicates better recognition accuracy and is our primary evaluation metric.

We expect to:

- Match the TFNet baseline in M1.
- Achieve higher gloss-level accuracy in M2.

Project Plan

Our project will proceed in two **stages**:

M1: Baseline Reproduction

Replicate the performance of **TFNet** on the CE-CSL dataset. In brief, TFNet is a combination of CNNs and a Bi-LSTM(Bi-directional Long and Short-Term Memory) module. First a CNN is deployed to extract features from original video (which can be seen as, say, 24 pictures per second), then the feature map is transmitted into two sequence feature extractors both in frequency and temporal domain. The extractors are both a 1D CNN plus a Bi-LSTM module. Afterwards, feature maps extracted from these two domains are fused together and then passed into a FC layer for recognition. The baseline model is mainly based on [3] in related works.

M2. Model Improvement

We have figured out three possible perspectives to improve this model, and some of them can be employed simultaneously.

- **Plan**:
    - To increase the accuracy of recognizing sign patterns using more powerful models in CV, as well as empowering the functionality of forming a sentence from a sequence of gloss by employing more powerful models and architecture.
    - To decrease the number of parameters and simplify the model architecture to be more lightweight as well as increasing responding time, so that instant translation is possible.
    - To adopt facial features as another clue for translation, since it can convey a lot of meaning by creating different context.

All team members will study the baseline model and jointly participate in training and evaluation.

In M2, individual contributions are:

- **Kaiqi Zhu**: Works on better visual feature extraction.
- **Mohan Huang**: Incorporates non-manual features like facial expressions and head motion.
- **Zhe Yu**: Optimizes model architecture to improve accuracy.