**Cost-Controllable Large-Scale Sign Language (CCLS) Dataset**

**Tianjin University**

Dataset Information:

Lacking a large-scale dataset is the major obstacle limiting sign language recognition (SLR). To solve above problem, we propose a novel cost-controllable way to construct a SLR dataset, called cost-controllable large-scale sign language (CCLS). We disassemble the sentence into a template part and an entity part. We collect template videos and entity videos, respectively. Then template videos and entity videos are reassembled to produce more combined videos, which serve as the training set. Meanwhile, we collect complete videos as a test set to verify the performance of the model.

The CCLS contains 130 entity words, 47 template sentences and 200 complete sentences. Each entity/template component is performed by 10 signers for 10 times. Therefore, the number of entity videos is 130×10×10=13,000, and the number of template videos is 47×10×10=4,700. For complete component, 5 signers participate in the collection, and each complete sentence is collected for 2 times. Hence, the number of complete videos is 200×5×2=2,000. All videos have the resolution of 1280 × 720 and a frame rate of 30 FPS (frames per second). The videos are collected under a natural background.

The contents of the dataset cover three scenarios, such as hotels, attractions, and restaurants. The entity words and template sentences are derived from the CrossWOZ corpus. The CCLS dataset contains 181 Chinese sign words. The entity videos and template videos are used to produce 20,000 combined videos, which are viewed as the training set. And the complete videos are viewed as the test set. The table 1 shows the details of the CCLS dataset.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Name | CCLS | Associated task | CSLR/SLT/ISLR | Vocabulary | 181 |
| Country | China | Language level | Word/Sentence | Signer | 10 |
| Entity video | 13,000 | Template video | 4,700 | Complete video | 2,000 |
| Total number of samples | 19,700 | Training videos (combined) | 20,000 | Testing videos | 2,000 |

Download:

The CCLS dataset is released to universities and research institutes for research purpose only. To request the access right to the data resources, please follow the link below:

**Baidu Netdisk:** https://pan.baidu.com/s/1W0sIgeKAhHIxy9-RDcv4QA [password: riqn]

**OneDrive:** https://tjueducn-my.sharepoint.com/:f:/g/personal/lqgao\_tju\_edu\_cn/EpBjyEcFnepJn32lrbEvtWkBvdnIFVFZ1TdXNH9-4bnEbg?e=EU4DIk

Contact:

If you have any questions about the dataset, please feel free to contact us:

Wei Feng, Professor, Tianjin University, wfeng@ieee.or

Liqing Gao, Ph.D Candidate, Tianjin University, lqgao@tju.edu.cn