Applying (3+2+1)D Residual Neural Network with Frame Selection for Hong Kong Sign Language(HKSL) Recognition

Zhenxing Zhou, King-Shan Lui, Vincent W.L. Tam and Edmund Y. Lam

Department of Electrical and Electronic Engineering

The University of Hong Kong

Background and Motivation

- In Hong Kong, more than 1.5 million residents suffer from hearing loss and rely on HKSL for daily communication
- But there are only 63 registered sign language interpreters in Hong Kong
- Collect a HKSL dataset and propose corresponding recognition method to address this specific social issue and facilitate the communication between the hearing impaired and other people

Proposed HKSL Dataset

In this dataset, there are 45 isolated sign words and at least 30 videos for each isolated sign word currently. In total, there are more than 1500 sign videos in this dataset, and we are still enlarging it by collecting more sign videos for different sign words. The technical details of the dataset include sample rate: 30fps, resolution: 480 × 640, duration: 6 to 10 seconds

Six Methods for Comparison

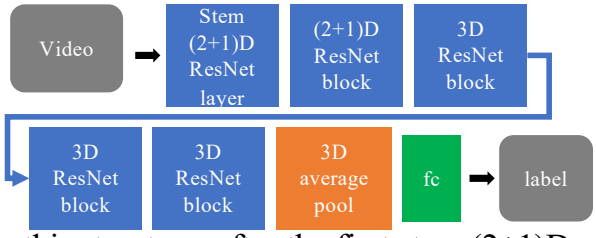
2D Approaches for HKSL recognition:

1. 2D HOG feature with LSTM.
2. 2D Pose Estimation with LSTM.
3. 2D Feature Extraction with LSTM.
4. Integrated Features with LSTM.

3D Approaches for HKSL recognition:

1. 3D ResNet
2. (2+1)D ResNet

Proposed (3+2+1)D ResNet Model



In this structure, after the first stem (2+1)D Residual layer, there are one (2+1)D ResNet block and three 3D ResNet blocks. In the (2+1)D ResNet block, there are four (2+1)D ResNet layers followed by the 3D batch-norm layer and ReLu activation layer while each 3D ResNet block consists of four 3D ResNet layers followed by the 3D batch-norm layer and ReLu activation layer.

Proposed Frame Selection Method

the blurry level of each frame in the video was calculated. Then, pick the frame with lowest blurry level in every K consecutive frames for down-sampling and denoising. Finally, all the selected frames will be used to construct video clips and the length of each video clip is set to be 16 frames.

Experimental Results

Different Methods Performance in HKSL	Accuracy
2D Pose Estimation with LSTM	66.7%
2D Feature Extraction with LSTM	71.1%
3D ResNet without Frames Selection	89.1%
(2+1)D ResNet without Frames Selection	89.7%
3D ResNet with Frames Selection	92.2%
(2+1)D ResNet with Frames Selection	93.5%
Proposed Hybrid ResNet model (hybrid_1_3)	**94.6%**