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

Catalogue

I. Introduction

- In Hong Kong, more than 1.5 million residents are suffering from hearing loss and most of them rely on HKSL for daily communication
- Only 63 registered sign language interpreters in Hong Kong

Objective: address this social issue and facilitate the communication between the hearing impaired and other people.

Catalogue

II. Proposed Hong Kong Sign Language Dataset

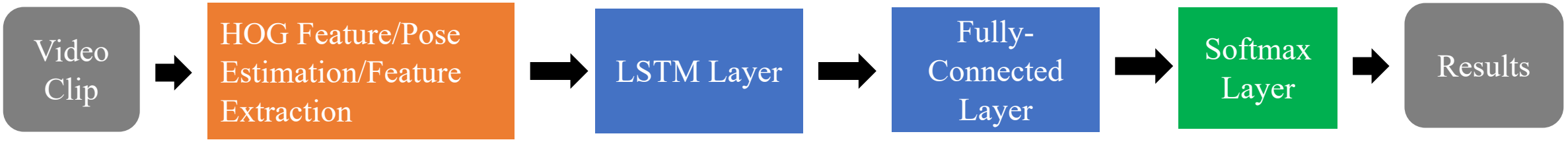
In this dataset, there are 45 isolated sign words and at least 30 videos for each isolated sign word currently. In total, there are more than 1500 sign videos in this dataset, and we are still enlarging it by collecting more sign videos for different sign words. The technical details of the dataset:

Sample rate: 30fps, Resolution: 480 × 640, Duration: 6 to 10 seconds

Catalogue

Baseline methods for HKSL recognition

- 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

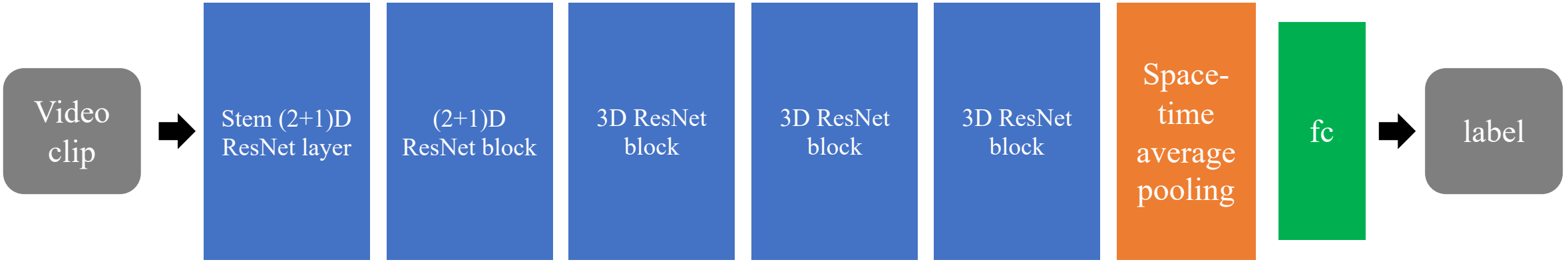


- 3D Approaches for HKSL recognition
  1. 3D ResNet for Sign Language Recognition
  2. (2+1)D ResNet for Sign Language Recognition

Catalogue

Proposed (3+2+1)D ResNet Model

After the first stem (2+1)D Residual layer, there are one (2+1)D ResNet block and three 3D ResNet blocks.



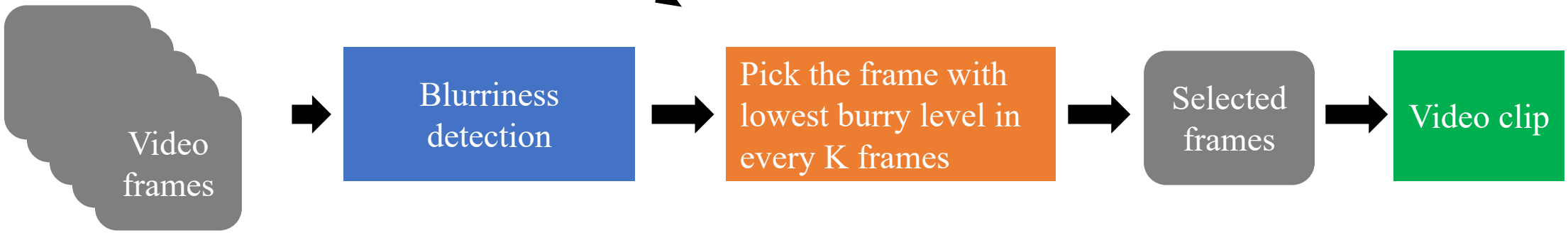
Reason for this Structure: The major task of the top deep learning layer is to extract some basic features from both temporal dimension and spatial dimension, such as edge detection. In this state, the weak ability of (2+1)D ResNet in considering the relation between these two dimensions will be probably over-weighted by its advantages in extracting features from different dimensions separately and increasing the complexity. Thus, we will benefit from the (2+1)D ResNet layer in this situation. However, when it comes to the deeper part of the model, the relation between temporal dimension and spatial dimension becomes more important than before. In this case, it would be better for us to adopt 3D ResNet layer as it can process both spatial and temporal information at the same time and consider the relation between them.

Catalogue

Constructing Video Clip with Frame Selection

$$K_f = \begin{pmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{pmatrix}$$

$$Equation\ I\colon\ B = -Variation(conv2d(frame, K_f))$$



Catalogue

Experimental Results

These two tables summaries the experimental results of different methods in two dataset. Among the 3D approaches, by adopting the proposed frame selection preprocessing methods, the accuracy of (2+1)D ResNet model is 3.8% higher than the one without it which successfully verify the significance of frame selection. Meanwhile, compared with other models, the proposed (3+2+1)D ResNet Model achieves the best performance and reaches the highest accuracy of 94.6% with frame selection. These experimental results strongly prove the effectiveness of both the frame selection method and the proposed (3+2+1)D ResNet model.

Different Methods in CSL	Accuracy
fc7-3DCNN+fc-LeNet	85.8%
LSTM_fc2	86.2%
eSC+HOG	92.0%
Proposed (3+2+1)D ResNet Model with Frame Selection	**96.0%**

Different Methods in HKSL	Accuracy
2D HOG feature with LSTM	62.7%
2D Pose Estimation with LSTM	66.7%
2D Feature Extraction with LSTM	71.1%
Integrated Features with LSTM	75.8%
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%
Reversal Hybrid ResNet model r_hybrid_1_3	90.7%
Reversal Hybrid ResNet model r_hybrid_2_2	90.3%
Reversal Hybrid ResNet model r_hybrid_3_1	91.1%
Hybrid ResNet Hybrid model hybrid_3_1	93.2%
Hybrid ResNet Hybrid model hybrid_2_2	93.8%
Proposed Hybrid ResNet model (hybrid_1_3)	**94.6%**

Thank you!