Deep Learning for Gesture Recognition in Smart TV Control

Zihang Xu, Zhen Xu, Hantao Fu, Siyi Hu
Rice University

1. Abstract

Gesture recognition is popular in smart TV applications. In this work, we use different deep learning models including CNN+RNN, Conv3D and Temporal Convolution Network to train a specific video dataset containing 5 gestures for operating TVs. Up to now we've found that Conv3D has the best training results, which lays a solid foundation for us to explore more advanced training models and pursue better results in the future.

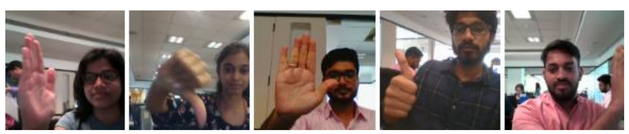
2. Background & Motivation

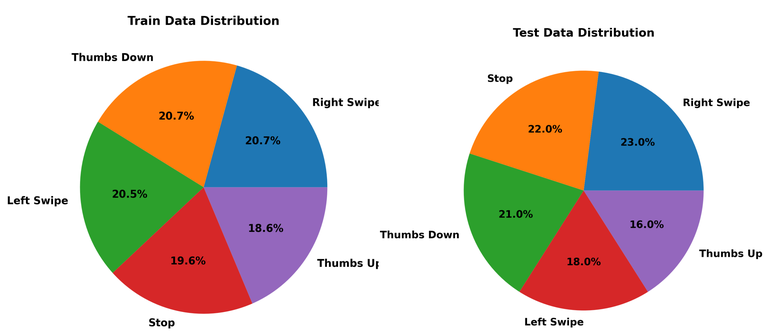
Gesture recognition has become an important component of human-computer interaction, especially in smart home applications. For smart TVs, hand gesture recognition allows users to perform commands to control the TV without physical contract. This improves the convenience and reduces the reliance on traditional remote controls.

Prior research has already explored gesture recognition by using different models. For example, hybrid CNN_RNN models applied to gesture recognition with EMG signals showed the robust performance and scalability. In our project, we would like to explore more about how to improve the accuracy of the gesture recognition using different model configurations. To address these challenges, our project compares the hybrid CNN and RNN model, Conv3D and Temporal Convolution model to find the most effective approach for smart TV gesture recognition.

3. Dataset

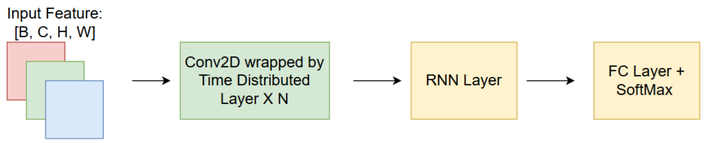
- This dataset has videos categorised into one of the **five** classes.
  - Stop, Right swipe, Left swipe, Thumbs down, Thumbs up
- Each video is divided into a sequence of **30** frames.
- Two types of dimensions - 360x360 && 120x160





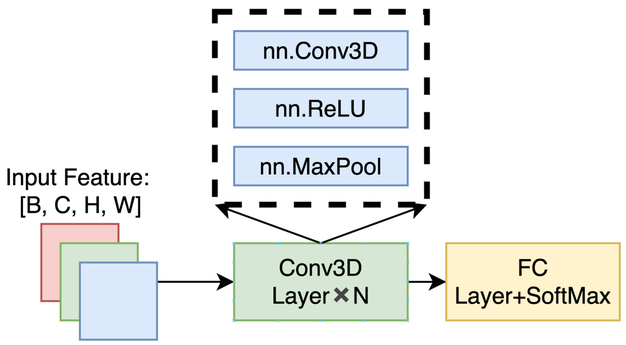
4.1. Method 1

Method 1 used a hybrid neural network model combining Convolutional Neural Networks (CNNs) with Recurrent Neural Networks (RNNs) using the Keras framework.



4.2 Method 2

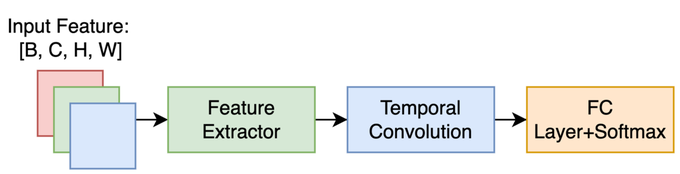
Method 2 leverages **3D Convolution** to extract spatial and temporal features. By extending traditional 2D convolutions into the temporal domain, 3D convolution is an ideal approach to analyzing spatial and temporal patterns in image sequences.



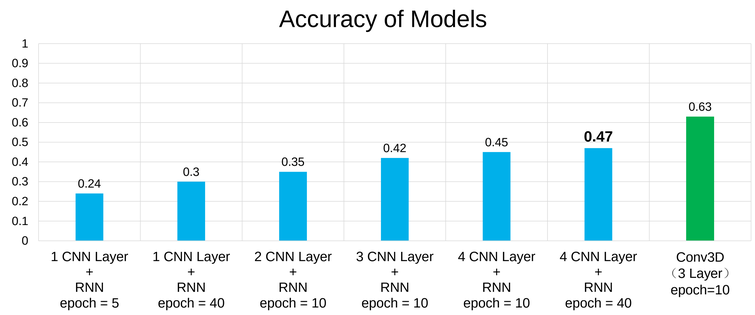
4.3 Method 3

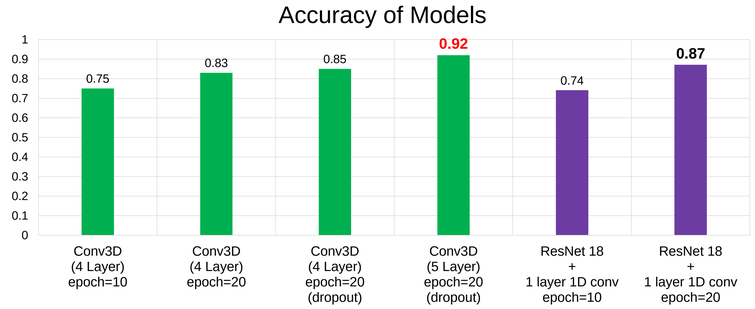
Method 3 integrates 2D and 1D convolution layers

- **2D Convolutions** are used to capture spatial features from the input data
- **1D Convolutions**, inspired by Temporal Convolution Network, is used to model temporal relationships across frames.



5. Results and Conclusion





- Currently, Conv3D (Method 2) has best performance: 92% Accuracy.
- 2D CNN + 1D Conv (Method 3) also has relatively high accuracy.
- Currently, CNN+RNN's (Method 1) performance is relatively low.

6. Future Work

- We will further improve the accuracy and efficiency of training from the following two aspects:
  1. Introducing more advanced methods for training gesture video datasets, including GANs, Transformer and Optical Flow Models
  2. Developing hybrid model combined with the characteristics of the dataset and advantages of each model implemented before.

7. Reference

1. Gesture Recognition Dataset:https://www.kaggle.com/datasets/abhishek14398/gesture-recognition-dataset/data
2. "Gesture Recognition with Hybrid Models." PLOS ONE, 2024, https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0264543.
3. Sapiński, Tomasz, et al. "Hybrid Deep Learning Models for Hand Gesture Recognition with EMG Signals." IEEE Xplore, 2024, https://ieeexplore.ieee.org/document/10582166.
4. Tran, Du, et al. "Learning spatiotemporal features with 3d convolutional networks." Proceedings of the IEEE international conference on computer vision. 2015.
5. Bai, Shaojie, J. Zico Kolter, and Vladlen Koltun. "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling." *arXiv preprint arXiv:1803.01271* (2018).