Gesture Classification From RGB and RGB-D Video Using Lucas-Kanade Optical Flow

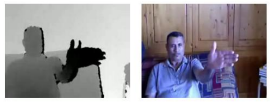
Kenny Leung
kenleung@stanford.edu

INTRODUCTION

Automated sign language recognition trails as much as thirty years behind speech recognition. A related variant of the sign language recognition problem is the **video gesture classification task**.

This study investigates whether **optical flow features** are beneficial for gesture classification. Optical flow is a technique for measuring the apparent motion of objects relative to an observer.

DATASET



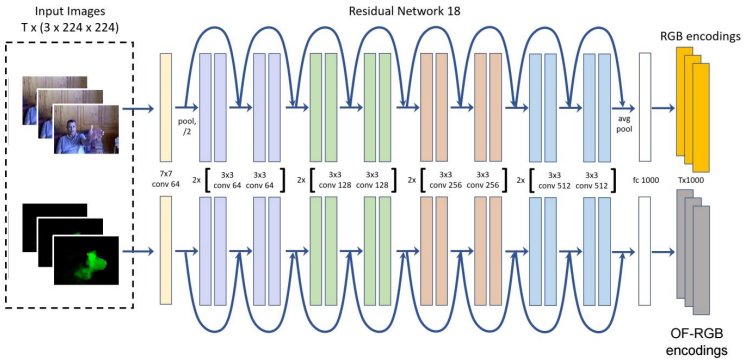
The dataset consists of **47,933 total videos** of **249 distinct gestures**, performed by 21 actors in varying conditions.

Each action is recorded in 240 x 360 **RGB** and Kinect grayscale **RGB-D** videos, as depicted above. Each video lasts between 5-15 seconds, and they are downsampled at 10 fps.

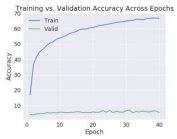
For the **data preprocessing** step, we train **18-layer residual networks** on individual frames from the training set videos, in order to learn discriminative encodings for each frame.

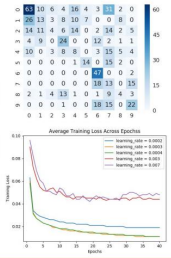
Then, we convert each frame into a 1000-dimensional tensor from the trained residual network, and pass the resulting sequence into an **LSTM network** with **512 hidden units**, followed by **3 fully connected layers** with **256 hidden units** and **ReLU activation**, and finally a **softmax** output layer. We use the **cross-entropy loss** function and **Adam optimization**.

ARCHITECTURE



RESULTS





The **combined RGBD-OFRGBD** (depth & optical-flow depth) input yields **2.16% validation accuracy** and **2.30% test accuracy**.

This constitutes a decrease in accuracy, when compared to the **combined RGB-RGBD** input to the LSTM, which achieves **7.24% test accuracy**.

The figures to the left depict the training curves and confusion matrix for the RGB-RGBD model. The RGBD-OFRGBD model only improves during the first iteration, and then immediately plateaus.

ANALYSIS

Optical flow features **did not improve classification accuracy** of the LSTM model.

RGB-D images are likely not useful for optical flow, because they violate both **pixel intensity** and **pixel neighborhood** assumptions made by Lucas-Kanade optical flow estimations.

Future directions for this project involve **improved hyperparameter** sweeping and **more frequent frame sampling**.

In the future, I would consider **training separate models for hand, face, and body detection** to isolate gesture-specific features.

REFERENCES

P. Wang, W. Li, S. Liu, Z. Gao, C. Tang, and P. Ogunbona, **"Large-scale isolated gesture recognition using convolutional neural networks."**
K. Simonyan and A. Zisserman, **"Two-stream convolutional networks for action recognition in videos."**
K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition,"
G. Farnebäck, **"Two-frame motion estimation based on polynomial expansion."**