SignNet: Recognize Alphabets in the American Sign Language in Real Time

Zeqiang Lai Kexiang Huang Zhiyuan Liang

School of Computer Science
Beijing Institute of Technology

Course of Computer Vision, December 2020

Introduction

- Sign Language Recognition
- 26 Letters && Space, Delete
- Real Time Demo with Common Commercial Camera

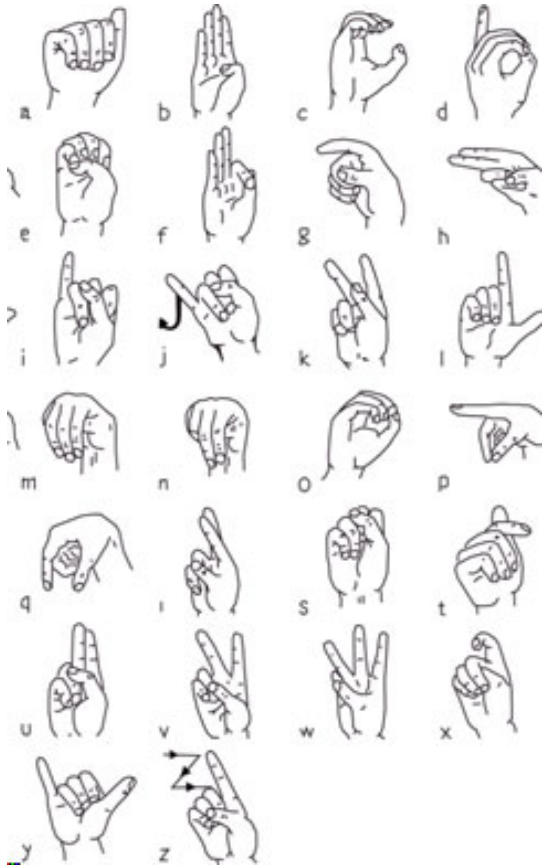

Figure: American Sign Language

Approach

1. Formulate the task as an **image classification** problem.
2. Use a modified version of **VGG** network for recogniton.
3. Trained on the **custom dataset** collected by our own.

Modified VGG

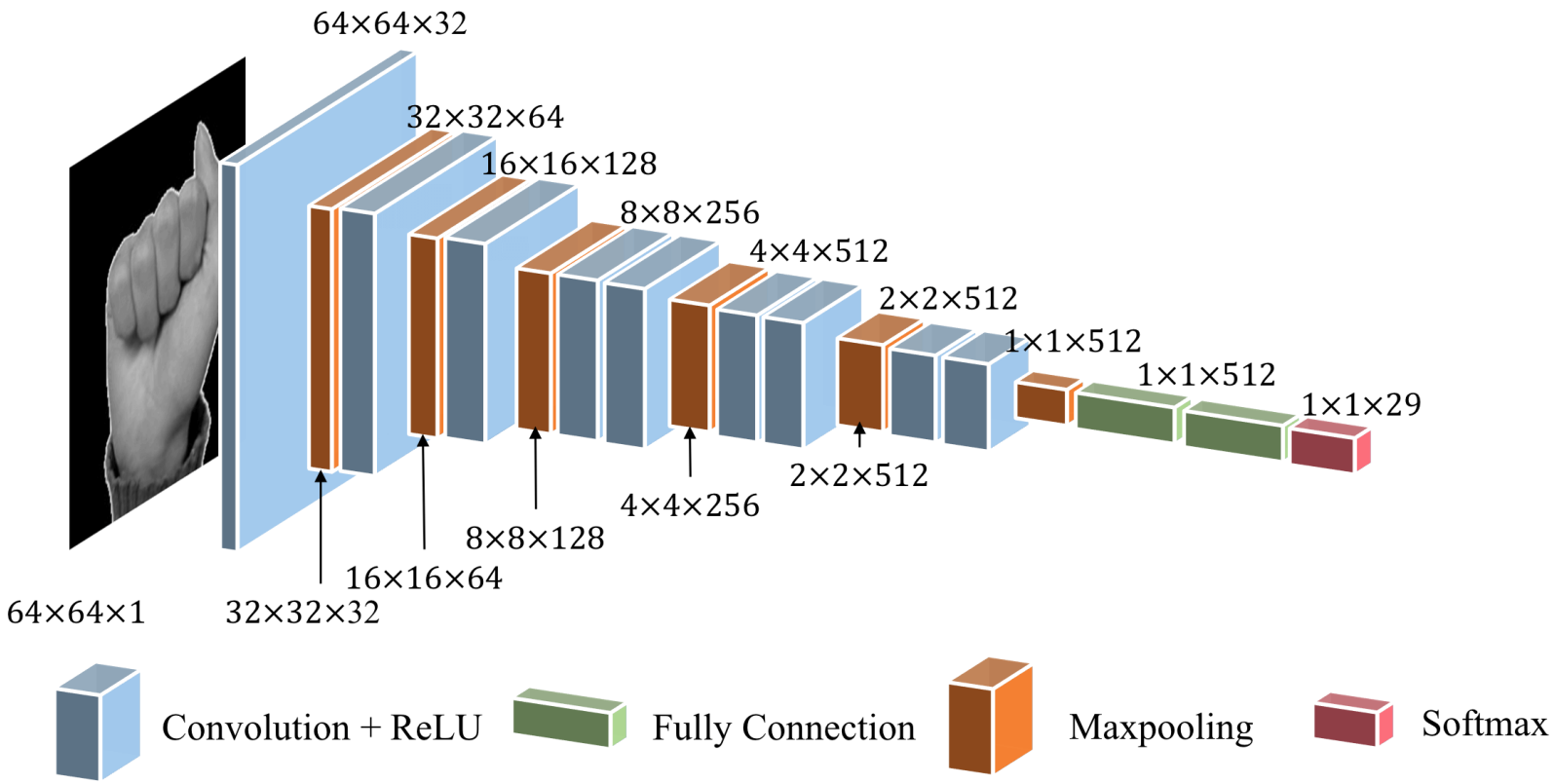

Figure: Network architecture

First Try: Train on ASL Dataset

1. ASL Alphabet is a public available dataset
   - 200×200 color image.
   - Training set contains 87000 images (29 classes, 3000 for each class).
   - Test set contains 870 images, 300 for each class.
2. Train with origin pictures firstly (Bad performance)
3. Train with cropped - gaussian blurred - grayscale version (Still Bad)

First Try: ASL Result

Dataset	Accuracy
ASL Alphabet Train	86906/87000 (100%)
ASL Alphabet Test	145/870 (17%)
Our ASL Test Set	1087/21418 (5%)

Table: Results on ASL dataset.

Analysis

Why our model works perfectly on training set, but fails on test set ?

- Obviously, **Overfit!**
- Training set lacks variation on the background.

Analysis

Improvement

- Record dataset with diverse backgrounds ?

Improvement

- ~~Record dataset with diverse background.~~
- Laborious and time-consuming.
- **Record a dataset without background directly**

Improvement: Custom Dataset

- Use average background subtraction algorithm to remove background.
- 29 classes (the same as ASL).
- Train set: 30s video (900 images) for each class.
- Test set: 10s for each class.

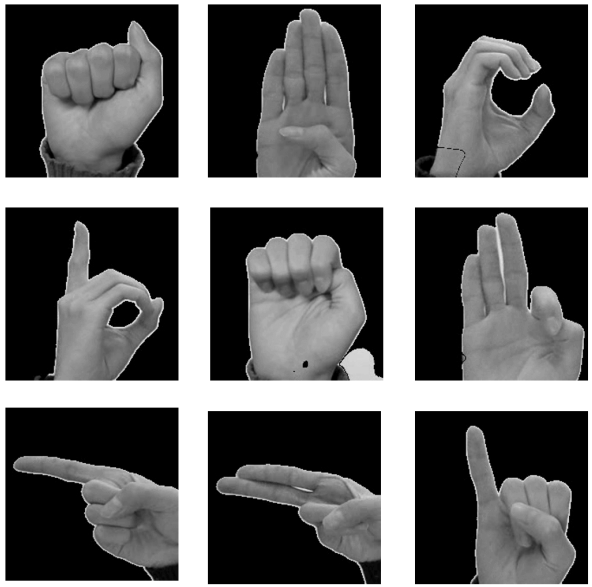

Figure: Samples from custom dataset

Improvement: Average Background Subtraction

1. Running average of first few frames as reference.
2. Subtract all the subsequent frames with respect to reference.
3. Pixels that exceed certain threshold are considered to be foreground, while the others are background.

Improvement: Experiment - Accuracy

- Train with the same setting as it is in ASL training.

Dataset	Accuracy
Custom Train	25804/25809 (100%)
Custom Test	7364/8700 (85%)

Table: Results on custom datasets

Preview of Our System

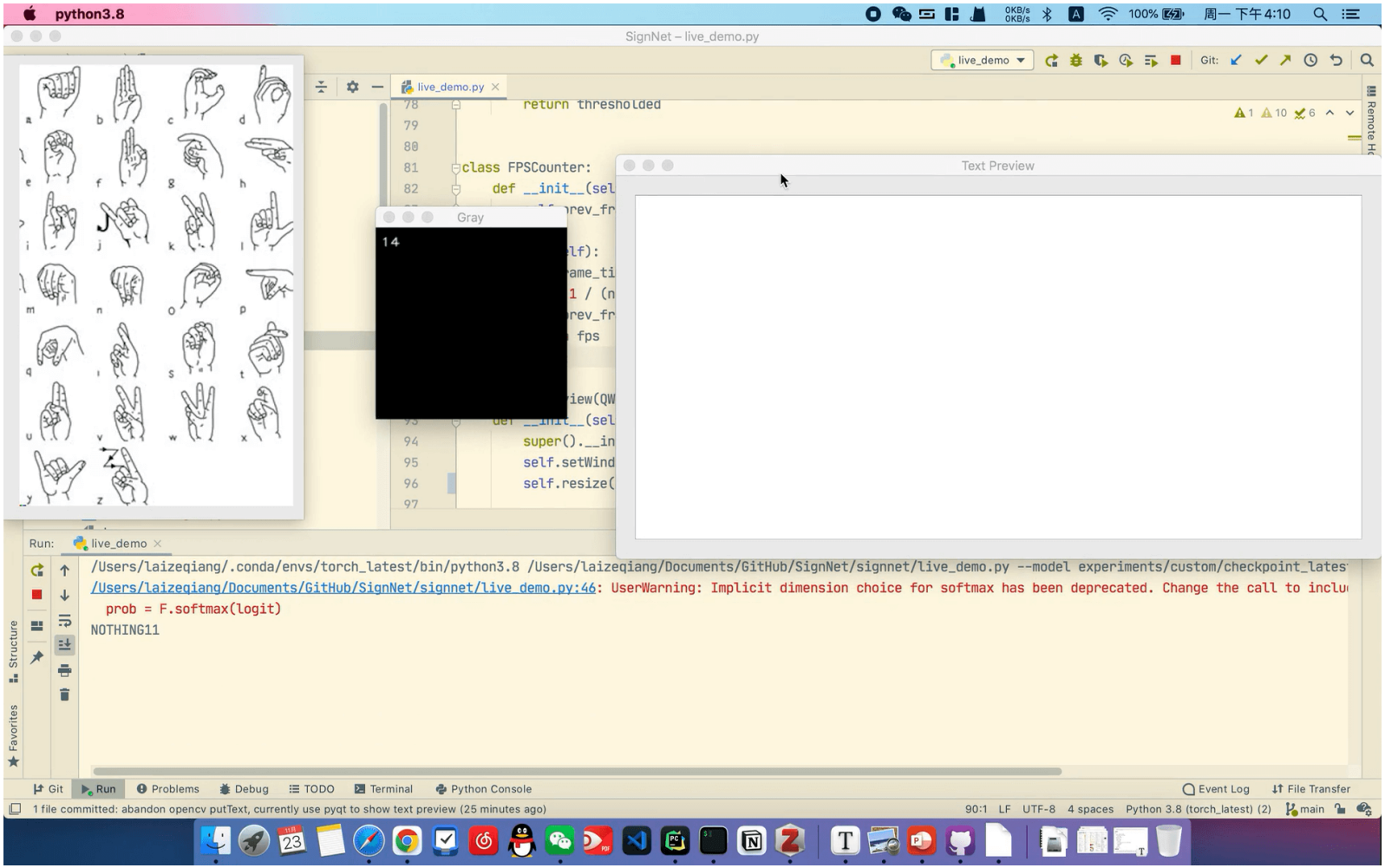

Figure: Preview

Improvement: Experiment - Confusion

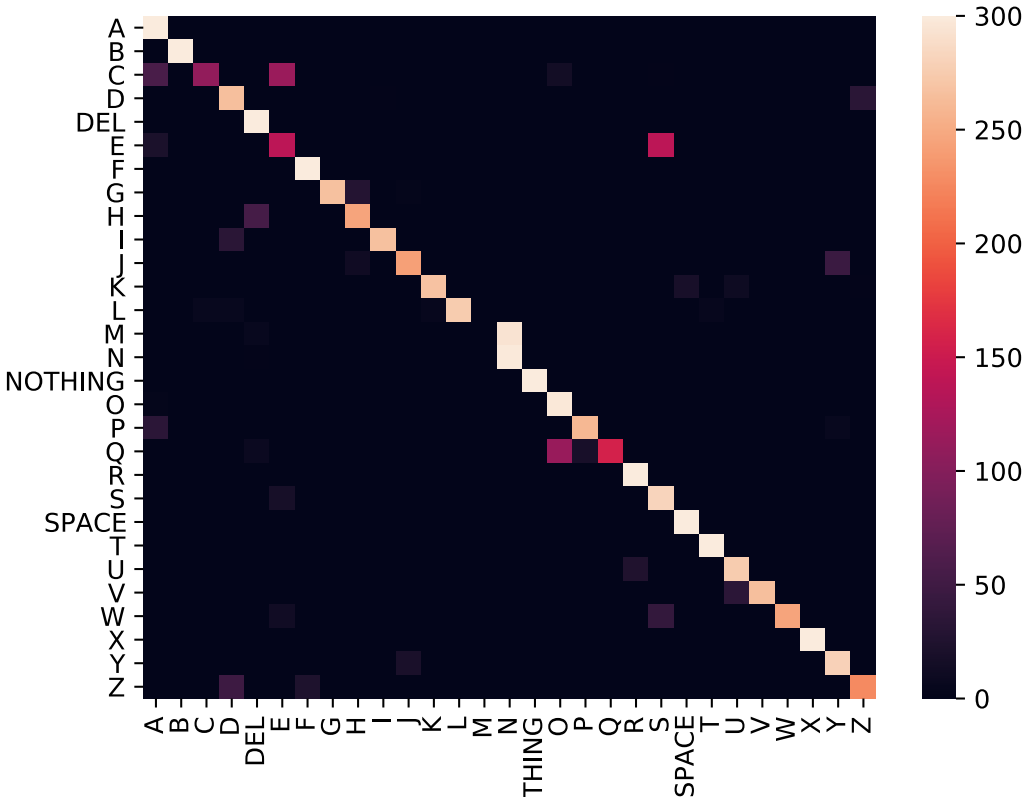

Figure: Confusion matrix on custom test set

Improvement: Experiment - FPS

Hardware	Platform	Type	FPS
Intel Core i7	macOS 10.15.7	live	15.73
		static	81.00
Nvidia RTX 2060	Windows 10	live	31.05
		static	534.12
Nvidia GTX 1070	Ubuntu 20.04.1	static	453.78

Table: Average FPS of our model in different settings. Type "live" means it is tested using our live_demo script which uses OpenCV to record video in real time, and type "static" means it is tested using static images.

Limitations

- Heavily rely on the performance of background subtraction.
- Poor performance on some extreme cases.
- Sensitive to the size, orientation of gestures.