Real Time Detection of Hand Gestures for Interface Extension using Convolutional Neural Networks

Scott Mathews, Satendra Varma, Shyam Narasimhan

Abstract

We present a method for real-time detection of hand gestures that employs a two phase image recognition pipeline, wherein hands are localized within the original input image, extracted, and fed into a CNN based classifier which outputs the detected gesture. Our results indicate that our method achieves reasonable accuracy for real time detection on a predefined set of hand gestures.

Introduction

As the power of machine learning techniques continues to grow, the number of interfaces between machine and man continue to grow. Convolutional neural networks have demonstrated their efficacy in computer vision tasks. We employ CNNs for creating a vision based interface, which enhances user interaction. Our goal with this project is to prototype a gesture based interface with computers, with the goal of being able to recognize a predefined set of hand based gestures from a live video feed.

Background and Related Work

Previous methods for hand gesture recognition have made use of basic image processing techniques as edge detection, segmentation based on color palette, and template matching. These models lack robustness, and are susceptible to the presence of secondary objects in the scene. Many existing techniques [1] rely upon the fact the hand gesture is the main and only subject within the scene, thus limiting real life use cases.

[1] arXiv:1503.03614 [cs.HC]

Methods

Gesture Classification Network Architecture (with 25 class output)

Hand Detection Network Architecture

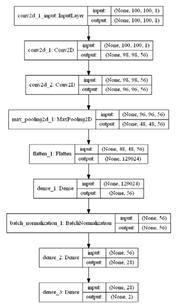
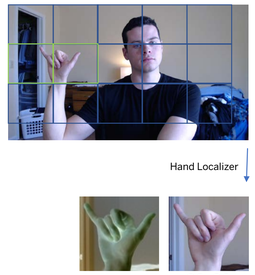


Image Processing Pipeline Illustration



Gesture classification architecture classifies images at ~166 frames per second (on i5-4300U @ 1.90 GHz) using Keras with a Tensorflow backend.

Hand Localization

Our technique is modeled after YOLO, with an even further emphasis on performance (aiming to run in real time on low powered CPUs). We segment the image into a grid of boxes, and perform binary classification on each box to indicate whether there is a hand in that box. Boxes which indicate presence of a hand are grouped into a singular box, giving a rough location of the hand in the image.

Gesture Classification

Post localization, the region of the image identified as a hand is passed to a CNN which is fed the image and its inverse, and takes the chosen class with the highest confidence (of original and inverse) To compensate for volatility in accuracy, we aggregate the classifications of 10 consecutive frames into a single decision. This improves perceived stability of classification

Data

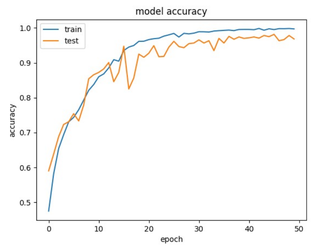
Samples from the localization dataset

A. Mittal, A. Zisserman, P. H. S. Torr
Hand detection using multiple proposals
British Machine Vision Conference, 2011

Samples from one of the gesture classification datasets used

tecperson
Sign Language MNIST
Kaggle, 2017

Results



Currently, validation accuracy for our gesture classifier is approximately 95% after 50 epochs of training.

Conclusion

The gesture detection framework we present is demonstrated to have utility in localization and classification of hand gestures from live video. The results we present indicate that, in the right conditions, our methodology works well for the gesture recognition task. Our method performs undesirably when presented with situations which our training data does not account for, such as abnormally dark conditions and high-noise backgrounds.