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

Scott Mathews
Indiana University
Bloomington, Indiana
scomathe@iu.edu

Shyam Narasimhan
shynaras@iu.edu

Satendra Varma
satvarma@iu.edu

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 a reasonable accuracy for real time detection on a predefined set of hand gestures.

1. 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 a system based on convolutional neural networks for creating a vision based interface, which provides a new avenue for user interaction with their machine. 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 gestures from a live video feed.

The space of voice assistants exploded recently with the advent of products from each major technology company aimed at providing a voice-based interface with computers. As desktop and phone processors continually grow more powerful, it is our belief that the next generation of interfaces with computers will utilize vision based interfaces.

Devices with built-in cameras would be able to utilize this interface by having an always-on camera, which performs actions on the system in response to certain visual inputs from the camera. One such visual input might be hand gestures. For example, a certain gesture might tell the camera to start taking commands, serving the same purpose as an activation phrase in current voice assistants. Following this "activation gesture" a series of additional gestures might be performed to indicate a command.

An interface similar to this one is explored in the paper by Haria et al. [1] In their interface, hand gestures give commands to the system directly. Specifically, they use the system to iterate through slides of a PowerPoint presentation.

Another intriguing use case for the hand-gesture based interface is with American Sign Language recognition. This interface could allow for text input using a purely visual interface, making use of the American Sign Language alphabet.

In order to facilitate these use cases, it is necessary for the algorithm powering the visual recognition to be lightweight enough to run in the background on low powered devices such as laptops and phones, while still providing enough performance to be useful.

Our goal in this project is to present a model for an efficient real time detection of hand gestures. Our results indicate that such interfaces are viable, given an appropriate performance conscious setup.

2. Background and Related Work

There has been previous work on the subject of real time hand gesture recognition. In the work of Chen et al. [2] a method utilizing segmentation of fingers was presented. In Xu's work [3] a method utilizing Kalman filters to preprocess images, followed by a convolutional neural network for recognition was employed.

3. Methods

We have seen multiple applications of gesture recognition in the past ,using depth sensors or simple images frame by frame. In the case of frame by frame gesture recognition, there are many systems which work when there is only a hand in the image frame. Our gesture recognition system uses a video feed as input, and learns to distinguish between gestures even when there are other objects in the frame such as faces or bodies.

We initially classified data from the "Sign Language MNIST" dataset using a custom deep convolutional neural

network, and were able to achieve 98% accuracy.

Next, we tried using the VGG-16 architecture to classify any given frame into the three gesture categories. We applied transfer learning, and added 2 layers before the output layer. The output layer (after softmax activation) consisted of three neurons corresponding to gesture 1, gesture 2, and no gesture. The maximum node was thresholded at 0.4, after which it was classified as that gesture, otherwise classified as no gesture.

We achieved 95% accuracy in classification, and the gesture recognition transferred well into real world usage, however it was not running in real time, since it took a long time for the network to preprocess the images and pass them through so many layers.

3.1. Gesture Classification

Finally, we converged on a custom smaller architecture, using a convolutional neural network. We built a dataset of each gesture, as well as no gesture frames, and achieved 87% accuracy in training and 68% accuracy in real classification, while classifying frames in real time.

3.2. Adding Localization

We used hand localization methods to detect and localize hand in the video frames, and crop out the rest of the image. We pass this cropped picture in the network which recognizes the different gestures.

Our method for localization involved dividing the input image into 100x100 blocks, and running a custom convolutional neural network to recognize if a given block had a hand in it. Blocks with hands recognized inside were then resized and passed to the gesture classification network.

Our method achieved 92% accuracy in localizing the hand in specific lighting conditions, and 90% in recognizing the gesture.

3.3. Future Work

In the future, we plan to use more advanced methods for hand localization, such as Mask RCNN or YOLO v2. We also plan to improve the quantity and diversity of our training data.

4. Results

4.1. Accuracy

We achieved in accuracy of 98% training our model on the augmented "sign language mnist" data. In real world scenarios, we found the model sensitive to lighting conditions. It worked best in the lighting of an enclosed room, and poorly in natural light or large rooms, such as in Luddy Hall. More discussion of classification results. Graphic of training.

We achieved an accuracy of 92% while training the model for hand detection. Due to the imbalanced nature of the dataset, this accuracy was actually quite bad (bordering on the performance of a trivial classifier). In practice, we found that the network would always output the negative class, as that was by far the more commonly occurring class. Nevertheless, the network would report a spike in probability for having recognized a hand when a hand was in the frame. To remedy this, we changed the prediction from a simple choice of higher probability, to a much lower threshold, around 7% confidence for there being a hand. This resulted in acceptable results for hand recognition tasks. More discussion of detection results. Graphic of training.

5. Conclusion

References

[1] A. Haria, A. Subramanian, N. Asokkumar, S. Poddar, and J. S. Nayak. Hand gesture recognition for human computer interaction. *Procedia Computer Science*, 115:367 – 374, 2017. 7th International Conference on Advances in Computing and Communications, ICACC-2017, 22-24 August 2017, Cochin, India.

[2] Z. hua Chen, J.-T. Kim, J. Liang, J. Zhang, and Y.-B. Yuan. Real-time hand gesture recognition using finger segmentation. *The Scientific World Journal*, 2014, 2014.

[3] P. Xu. A real-time hand gesture recognition and human-computer interaction system. *CoRR*, abs/1704.07296, 2017.