Hand gesture tracking

Arun Ganesan, Michael (Caoxie) Zhang

April 4, 2012
EECS 545

Abstract

1 Introduction

2 System

2.1 Overview

There are three main stages to our system. First is acquiring training samples using a colored glove. Second is training a classifier on the samples. Third is live prediction and pooling.

The approach is pixel-level classification, and future pooling. The motivation for pixel-level classification is the potential of parallelizing the task and using the GPU to achieve real-time prediction.

2.2 Acquiring training samples

Color glove approach. To simplify the data acquisition step, we used cropping, depth thresholding, and a single colored glove. We used 4 gestures, giving a total of 5 classes including the background. We collected 400 training samples evenly spread across the gestures.

2.3 Training classifier

We trained the classifier on different subsets of the 400 training samples to study the effect of the training sample size. Also, we set aside 50 samples for testing purposes. Because our prediction is at the pixel level, we sample 1000 pixels from each of the training image, trying to get a good balance between the background and the gesture.

We used a depth-invariant feature vector for each pixel in our training sample. Each feature in the feature vector is the difference in the depth value of two offset pixels from the current pixel. We experimented with different number of such offset pairs for the feature vector.

We trained two classifiers - an SVM classifier for baseline testing, and a random forest classifier with multiple configurations. The random forest classifier can have multiple trees. We explored the effect of different number of trees in the forest.

2.4 Prediction

We aimed for real-time prediction of the hand gesture. The trained model from the classifier is used for per-pixel classification. That is, each pixel is classified as belonging to one of the four gestures or to the background. From here we have to first identify the likely gesture, and then find the location of that gesture, a process we called 'pooling'.

2.4.1 GPU

Running the prediction algorithm using the CPU proved to be very slow. For a 640×480 image, our system took 2.5 minutes to classify all the pixels. Given that this problem is highly parallelizable, we turned to the GPU. Re-implementing the prediction algorithm with the GPU reduced the prediction time from 2.5 minutes to 400 milliseconds - a 99.7% speedup!

2.4.2 Pooling

- mean versus median – mean is prone to outliers messing up the location. median is more resistant to that.
- majority gesture – if the hand is far away, this leads to the noisy labels being
- k-meniods – can be used to support multiple hands at once. and can be used to filter out the noise.

3 Experimental results

3.1 GPU performance enhancements

3.2 Classification algorithm experiments

4 Related work

5 Conclusion

References

[1] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, A. Blake. Real-time human pose recognition in parts from single depth images. CVPR, 2011.

[2] R. Wang and J. Popović. Real-time hand-tracking with a color glove. In Proc. ACM SIGGRAPH, 2009.

[3] I. Oikonomidis, N. Kyriazis, and A. Argyros. Efficient model-based 3D tracking of hand articulations using kinect. In BMVC, Aug 2011.

[4] V. Lepetit, P. Lagger, and P. Fua. Randomized trees for real-time keypoint recognition. In Proc. CVPR, pages 2:775-781, 2005.

[5] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification Journal of Machine Learning Research 9(2008), 1871-1874.

[6] B. Schneiderman. The eyes have it: a task by data type taxonomy for information visualizations. Visual Languages, 1996.

[7] D.A. Keim. Information visualization and visual data mining. Visualization and Computer Graphics, IEEE Transactions. Vol 8, no.1, pp. 1-8, Jan 2002.

[8] W. Stuerzlinger, C Wingrave. The value of constraints for 3D user interfaces. Dagstuhl Seminar on VR, 2010.

[9] M. Hoffman, P. Varcholik, and J. LaViola. Breaking the status quo: improving 3D gesture recognition with spatially convenient input devices. IEEE VR, 2010.

[10] V. Bystritsky. ALGLIB. 14 Aug 1999. Web. `http://www.alglib.net`.