

Hand Detection and Pose Estimation using Convolutional Neural Networks

Degree Project in
Computer Science and Communication,
Second Cycle (DA226X)

Adam Knutsson

`adamkn@kth.se`

`kth.se/profile/adamkn`

Computer Vision and Active Perception Lab (CVAP)
School of Computer Science and Communication (CSC)
KTH Royal Institute of Technology

2015-08-31

Outline

- 1 Introduction
- 2 Hand Detection using CNN
- 3 Hand Pose Estimation using CNN
- 4 Conclusions
- 5 References & Further Reading

The Human Hand

An Abstract Concept

Complex Structure

- 27 bones.
- Large number of muscles and tendons.
- Wide set of constraints with high variability.
- 30 - 50 degrees of freedom.

Core Challenges

- Strong self occlusion.
- High variability in visual appearance.
- Difficult to impose model constraints.

Deep (Machine) Learning

The Concept of Deep Learning [1]

*"...branch of machine learning based on a set of algorithms that attempt to **model high-level abstractions** in data **by using** model architectures (...) composed of **multiple non-linear transformations**."*

Deep Learning and Neural Networks

- Deep Boltzmann Machines
- Deep Belief Networks
- **Convolutional Neural Networks**

Convolutional Neural Networks (CNN)

Convolutional Layers

- Spatial convolutions using filter kernels. Example (Padding: 0. Stride: 1):

$$\underbrace{\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}}_{\text{Data (3,3)}} * \underbrace{\begin{bmatrix} w_{11} = 1 & w_{12} = 2 \\ w_{21} = 3 & w_{22} = 4 \end{bmatrix}}_{\text{Kernel (2,2)}} = \underbrace{\begin{bmatrix} 23 & 33 \\ 53 & 63 \end{bmatrix}}_{\text{Feature map}}$$

Convolutional Neural Networks (CNN)

Pooling Layers

- Spatial down-sampling operations. Example (Padding: 0. Stride: 1. Kernel size: (2,2)):

$$\underbrace{\begin{bmatrix} 23 & 33 \\ 53 & 63 \end{bmatrix}}_{\text{Feature map}} \Rightarrow \begin{cases} 63 & \text{if Type : MAX} \\ 43 & \text{if Type : AVG} \end{cases}$$

Fully Connected Layers

- Similar structure to the Multilayer Perceptron (MLP).

Hand Detection using CNN

Problem Approach

Window Search

- Exhaustive:
 - Sliding
 - Fixed size
- Selective [2, 3]:
 - Segmentation
 - Non-fixed size

Window Classification

- Two classes.
- Insensitive to resampling.
- Non-maximum suppression.

Hand Detection using CNN

Data: Training & Testing [4]



(a) Camera #1.

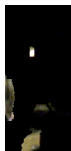


(b) Camera #2.

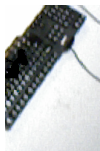


(c) Camera #3.

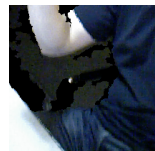
Figure: Positive classification samples. Frame #1.



(a) Camera #1.



(b) Camera #2.



(c) Camera #3.

Figure: Negative classification samples. Frame #1.

Hand Detection using CNN

Data: Validation [5]



(a) Sample #30.

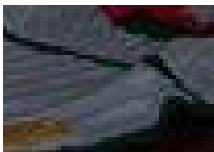


(b) Sample #31.



(c) Sample #47.

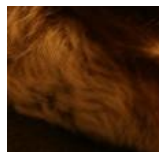
Figure: Positive classification samples.



(a) Sample #30.



(b) Sample #31.



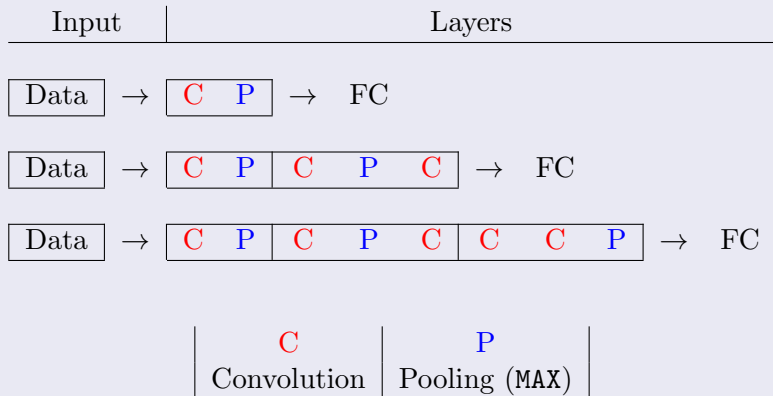
(c) Sample #47.

Figure: Negative classification samples.

Hand Detection using CNN

Network Architectures

Convolutional Layers [6]



- ReLU activation functions.

Hand Detection using CNN

Network Architectures

Fully Connected Layers [6]

Input		Layers
-------	--	--------

CL	→	n_{nf} 2
----	---	-------------------

CL	→	n_{nf} n_{nf} 2
----	---	-----------------------------------

n_{nf} Fully Connected

$$n_{\text{nf}} = 2^n \quad \text{for } n = 7, \dots, 12$$

Results

Classification Accuracy: Test Set

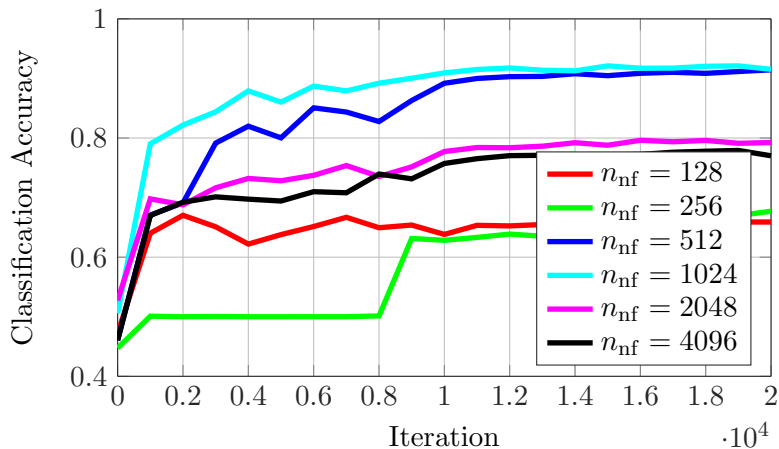


Figure: One convolution. Two fully connected layers.

Results

Classification Accuracy: Test Set

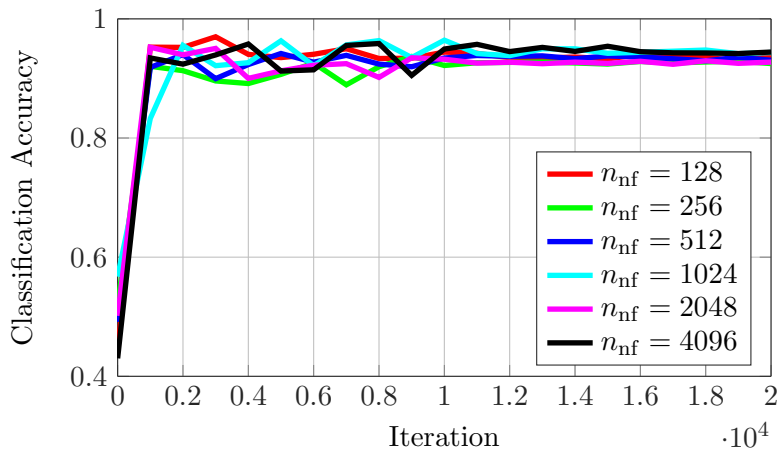


Figure: Three convolutions. Two fully connected layers.

Results

Classification Accuracy: Test Set

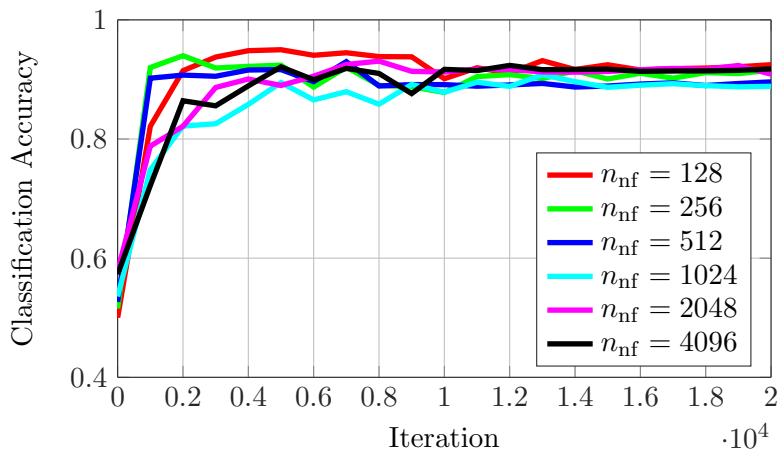


Figure: One convolution. Three fully connected layers.

Results

Classification Accuracy: Validation Set

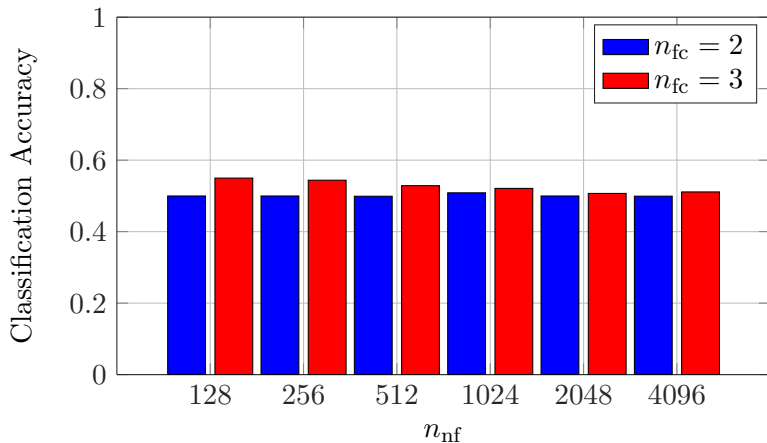


Figure: One convolution.

Results

Classification Accuracy: Validation Set

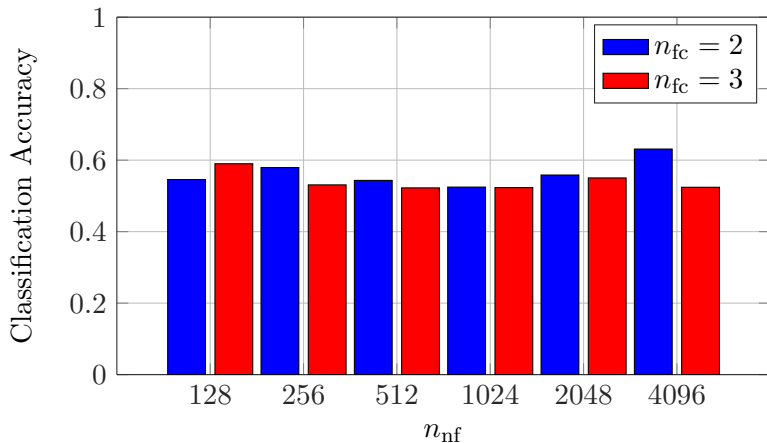


Figure: Three convolutions.

Results

Classification Accuracy: Validation Set

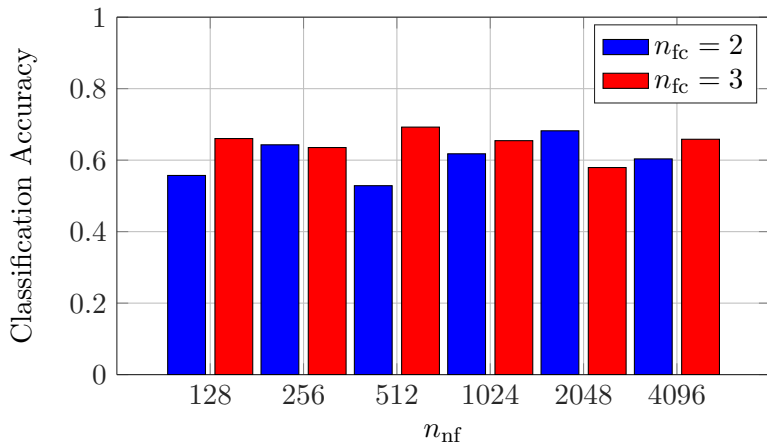


Figure: Five convolutions.

Results

Detection Test: Shallow Network

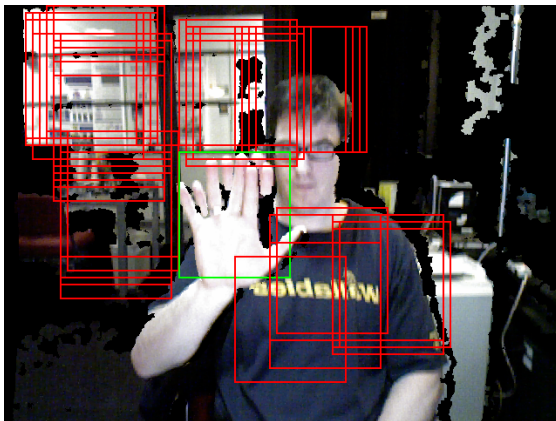


Figure: Camera #1. Frame #2000. Precision: 2.17 %.

Results

Detection Test: Shallow Network

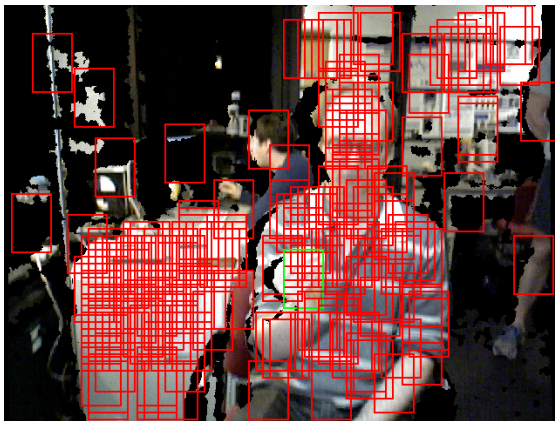


Figure: Camera #1. Frame #4000. Precision: 0.51 %.

Results

Detection Test: Deep Network

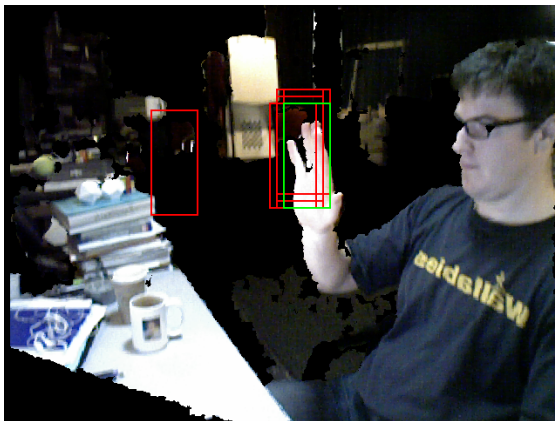


Figure: Camera #3. Frame #2000. Precision: 14.3 %.

Results

Detection Test: Deep Network

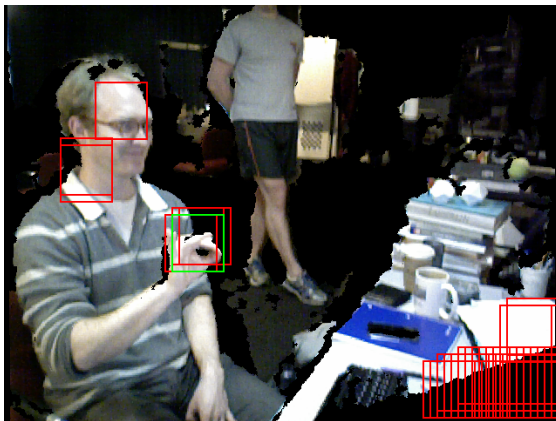


Figure: Camera #3. Frame #6000. Precision: 2.4 %.

Hand Pose Estimation using CNN

Problem Approach

Pose Estimation

- Hand dominated image.
- Compute pose or appearance description.
- Wide variety of pose descriptions.
- **Key-point locations**

Image to Pose Regression

- 36 key-points in (u, v, d) -space.
- No spatial resampling.

Hand Pose Estimation using CNN

Key-points

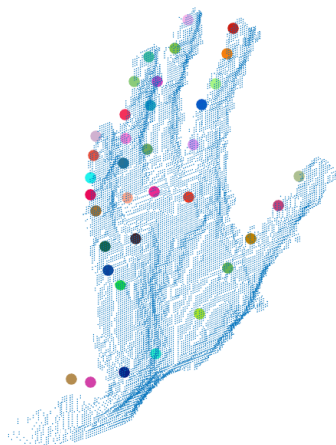


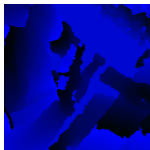
Figure: (x, y, z) -space rendering of (u, v, d) key-points overlaid PrimeSense depth data.

Hand Pose Estimation using CNN

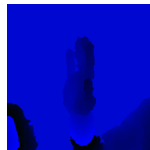
Data: Training & Testing [4]



(a) Camera #1.

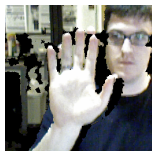


(b) Camera #2.

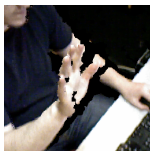


(c) Camera #3.

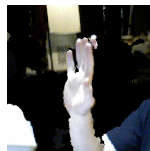
Figure: Depth data. Frame #1.



(a) Camera #1.



(b) Camera #2.



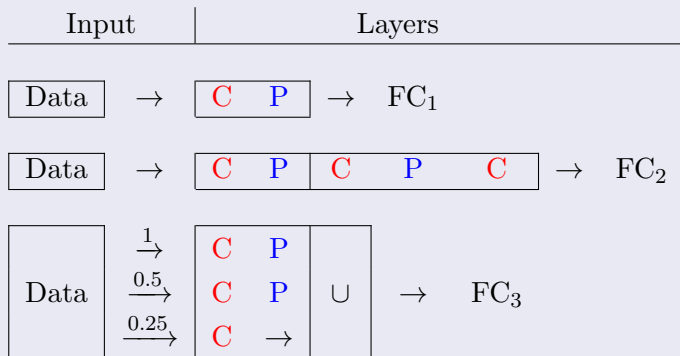
(c) Camera #3.

Figure: RGB data. Frame #1.

Hand Pose Estimation using CNN

Network Architectures

Convolutional Layers [7]



Hand Pose Estimation using CNN

Network Architectures

Fully Connected Layers [7]

Input		Layers	
CL_1	\rightarrow	n_{nf}	108
CL_2	\rightarrow	n_{nf} n_{nf}	108
CL_3	\rightarrow	n_{nf}	108

$$n_{\text{nf}} = 1024$$

Results

Euclidean Loss: Testing Set

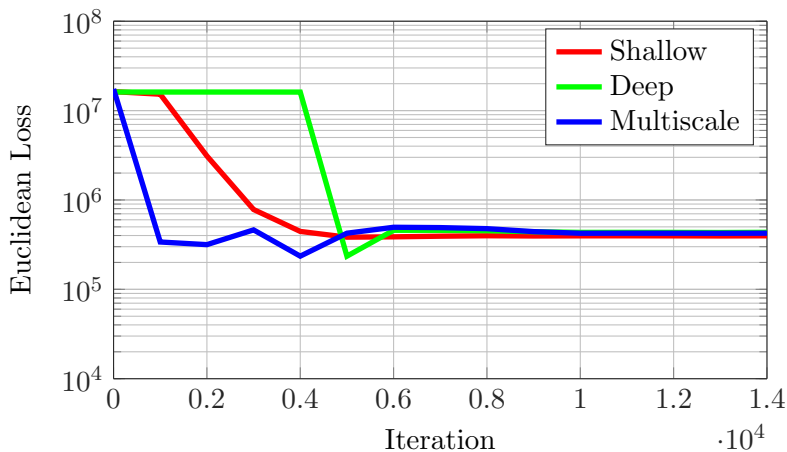


Figure: Depth Data.

Results

Euclidean Loss: Testing Set

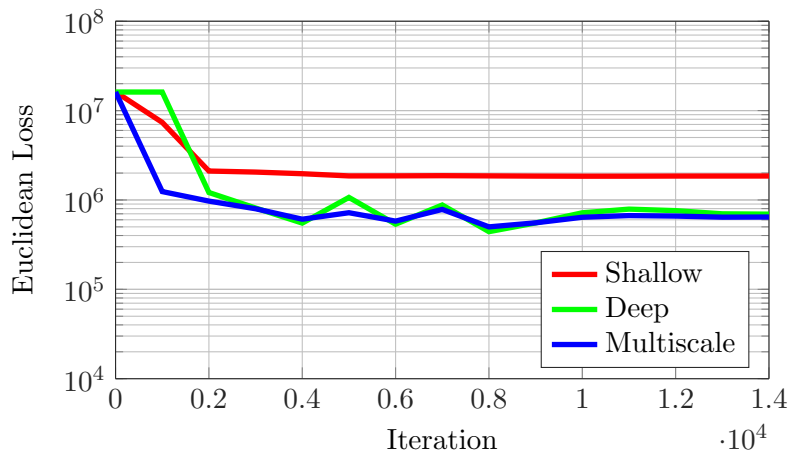


Figure: RGB Data.

Results

Deep Network Prediction

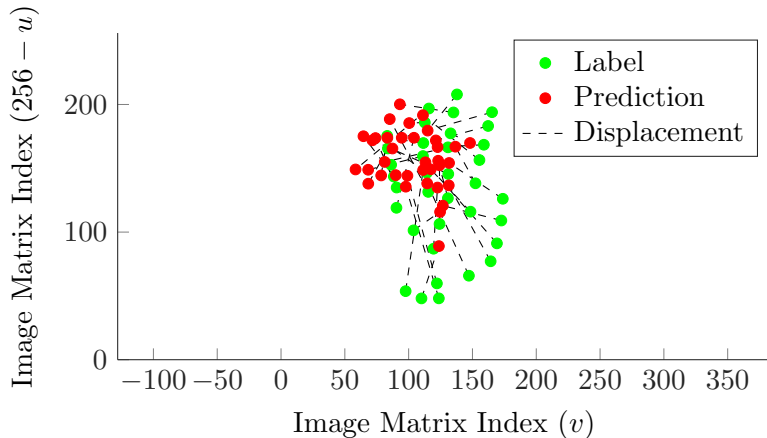


Figure: (u, v) -prediction given depth data.

Results

Deep Network Prediction

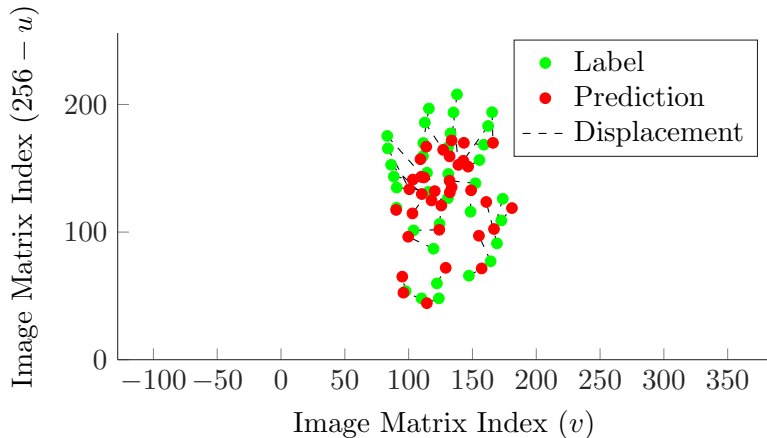


Figure: (u, v) -prediction given RGB data.

Results

Deep Network Prediction

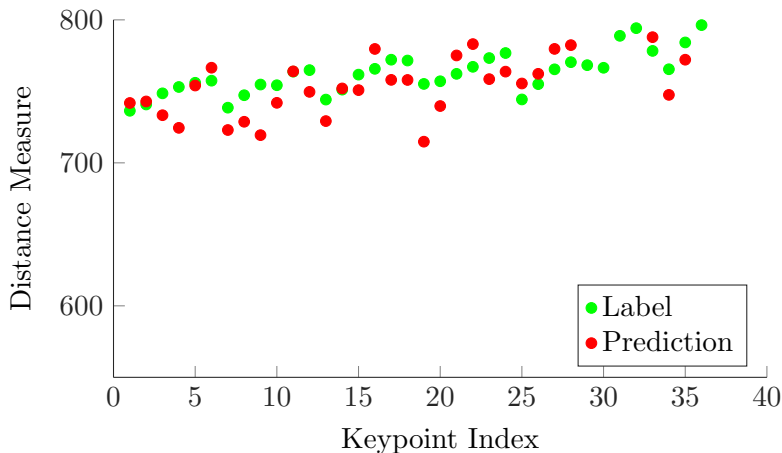


Figure: d -prediction given RGB data.

Conclusions

Hand Detection using CNN

- CNN considered suitable classifier.
- High diversity requirements on negative class.
- Larger dataset → reduce the number of false positives.
- Shallow network → color as feature descriptor.
- Deep networks → more abstract feature descriptors.

Conclusions (cont.)

Hand Pose Estimation using CNN

- Shallow network \rightarrow poor result on color data.
- Color \mapsto key-point locations more non-linear.
- Requires deeper networks.
- Color data offers smallest key-point displacement.
- Color data can be used to infer depth.

References & Further Reading I

- [1] Deep Learning. [Online]. Available:
http://en.wikipedia.org/wiki/Deep_learning
- [2] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, “Selective search for object recognition,” *International Journal of Computer Vision*, vol. 104, no. 2, pp. 154–171, 2013. [Online]. Available:
<https://ivi.fnwi.uva.nl/isis/publications/2013/UijlingsIJCV2013>
- [3] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, “Caffe: Convolutional architecture for fast feature embedding,” *arXiv preprint arXiv:1408.5093*, 2014.

References & Further Reading II

- [4] J. Tompson, M. Stein, Y. Lecun, and K. Perlin, “Real-time continuous pose recovery of human hands using convolutional networks,” *ACM Transactions on Graphics*, vol. 33, August 2014.
- [5] A. Mittal, A. Zisserman, and P. H. S. Torr, “Hand detection using multiple proposals,” in *British Machine Vision Conference*, 2011.
- [6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

References & Further Reading III

- [7] M. Oberweger, P. Wohlhart, and V. Lepetit, “Hands deep in deep learning for hand pose estimation,” *arXiv preprint arXiv:1502.06807*, 2015.
- [8] S. Marsland, *Machine Learning: An Algorithmic Perspective*. CRC Press, 2011. [Online]. Available: <http://books.google.se/books?id=n66O8a4SWGEC>
- [9] NVIDIA DIGITS - Interactive Deep Learning GPU Training System. [Online]. Available: <https://developer.nvidia.com/digits>

References & Further Reading IV

- [10] Y. LeCun. Convolutional Neural Networks: Machine Learning for Computer Perception. GPU Technology Conference On-Demand Webinar. [Online]. Available: [http://on-demand-gtc.gputechconf.com/gtcnew/on-demand-gtc.php?searchByKeyword=yann&searchItems=&sessionTopic=&sessionEvent=&sessionYear=2014&sessionFormat=&submit=&select=+](http://on-demand-gtc.gputechconf.com/gtcnew/on-demand-gtc.php?searchByKeyword=yann&searchItems=&sessionTopic=&sessionEvent=&sessionYear=2014&sessionFormat=&submit=&select=)
- [11] Artificial Neural Networks and Other Learning Systems (DD2432). [Online]. Available: <https://www.kth.se/student/kurser/kurs/DD2432?l=en>