Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

T A M S

Department of Informatics
Intelligent Robotics WS 2015/16

23.11.2015

Object Recognition
SIFT vs Convolutional Neural Networks

Josip Josifovski
4josifov@informatik.uni-hamburg.de

Outline

- **Object recognition**:
  - Definition, problem, human vision system and machine vision
- **Scale Invariant Feature Transform (SIFT)**
  - Algorithm details and example
- **Convolutional Neural Networks (CNNs)**
  - Algorithm details and example
- **Comparison of SIFT and CNN**
  - Biological plausibility, complexity, resources and applicability
- **Summary**

Object recognition - Definition

"The term **recognition** has been used to refer to many different visual capabilities, including **identification**, **categorization** and **discrimination**. Normally, when we speak of **recognizing an object** we mean that we have successfully categorized as an instance of a particular object class."

Liter, Jeffrey C., and Heinrich H. Bülthoff. "An introduction to object recognition."Zeitschrift für Naturforschung C 53.7-8 (1998): 610-621.

Identification – equality on a physical level

Categorization – assigning an object to some category, as humans do

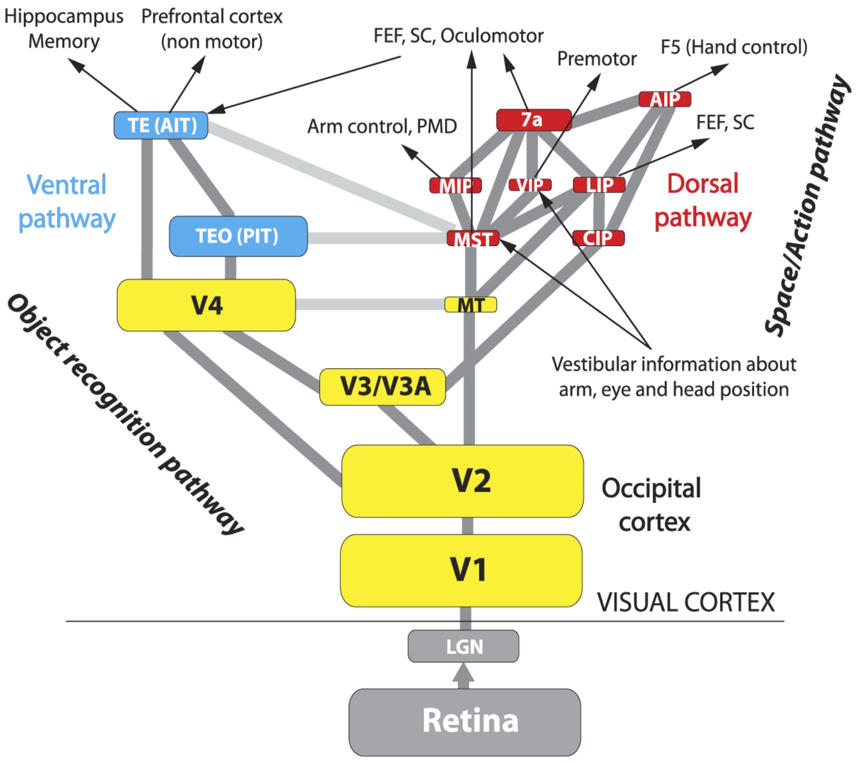
Discrimination – classification , assigning an object to one class

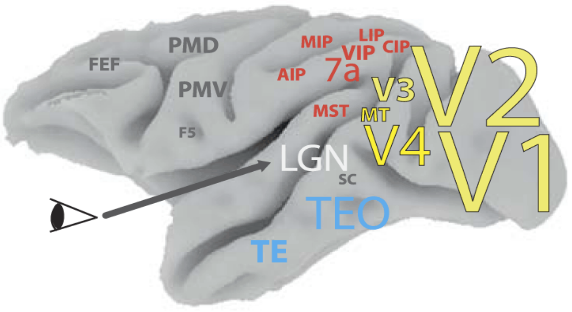
Object recognition – Problem

http://www.kyb.tuebingen.mpg.de/typo3temp/pics/915b4f5fb5.jpg

How humans do it?

- **Easy task** for human
- **Two pathways** for processing of visual input in the brain:
  - **Ventral** pathway
  - **Dorsal** pathway
- **Hierarchical processing** in the cortex:
  - Increasing receptive fields
  - Increasing complexity of details

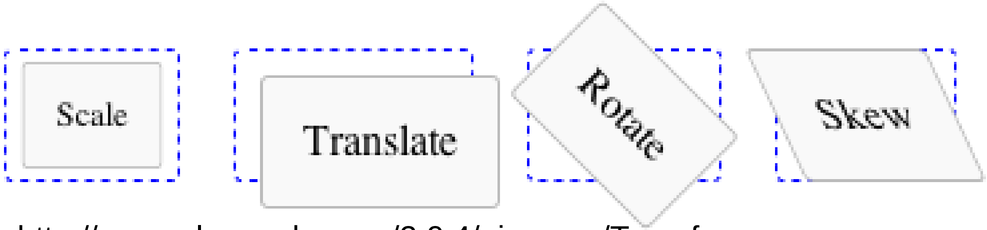




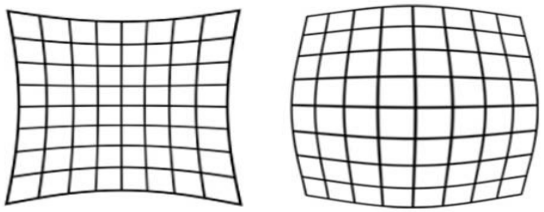
Kruger, Norbert, et al. "Deep hierarchies in the primate visual cortex: What can we learn for computer vision?." Pattern Analysis and Machine Intelligence, IEEE Transactions on 35.8 (2013): 1847-1871.

How machines do it?

- **Hard task** for machine
- Different transformations, distortions, scene conditions, viewing angles



http://manual.qooxdoo.org/2.0.4/_images/Transform.png



http://starizona.com/acb/basics/optics/distortion.jpg

- Most often recognition is done by **extracting local features** of object and trying to **match** them with features of unknown object

Scale Invariant Feature Transform

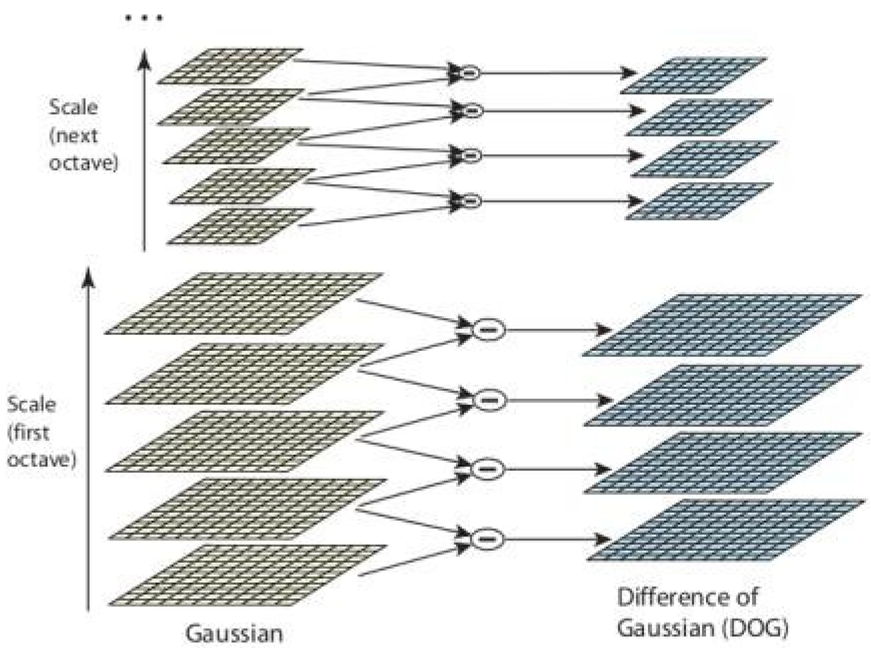
- Published by David G. Lowe in 1999
- Invariant to **scaling**, **rotation** and **translation**
- Partially invariant to **illumination changes** or **affine or 3D projection**
- Transforms an image into a large collection of local feature vectors (**local descriptors** called **SIFT keys**)
- Patented – University of British Columbia

Lowe, David G. "Object recognition from local scale-invariant features."Computer vision, 1999. The proceedings of the seventh IEEE international conference on. Vol. 2. Ieee, 1999.

SIFT steps

1) Scale-space extrema detection

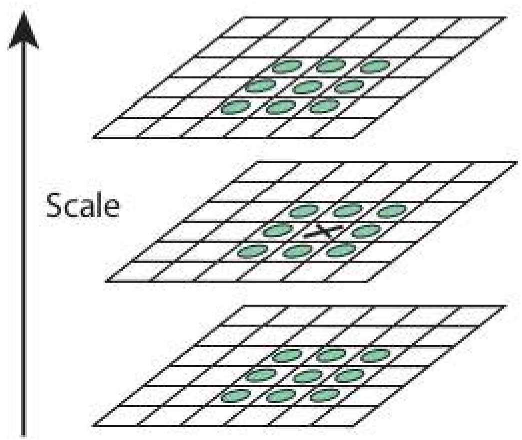
- Convolving image with Gaussian kernel repeatedly to get more and more blurred version of the image
- Calculating the difference image (DoG) as approximation to Laplacian of Gaussian (LoG)



http://docs.opencv.org/master/sift_dog.jpg

2) Key-point localization

- Finding the extrema (maxima or minima at each level of the pyramid)
- Comparing the extrema to layers above or below to check if it is stable



http://docs.opencv.org/master/sift_local_extrema.jpg

SIFT steps (cont)

3) Orientation assignment

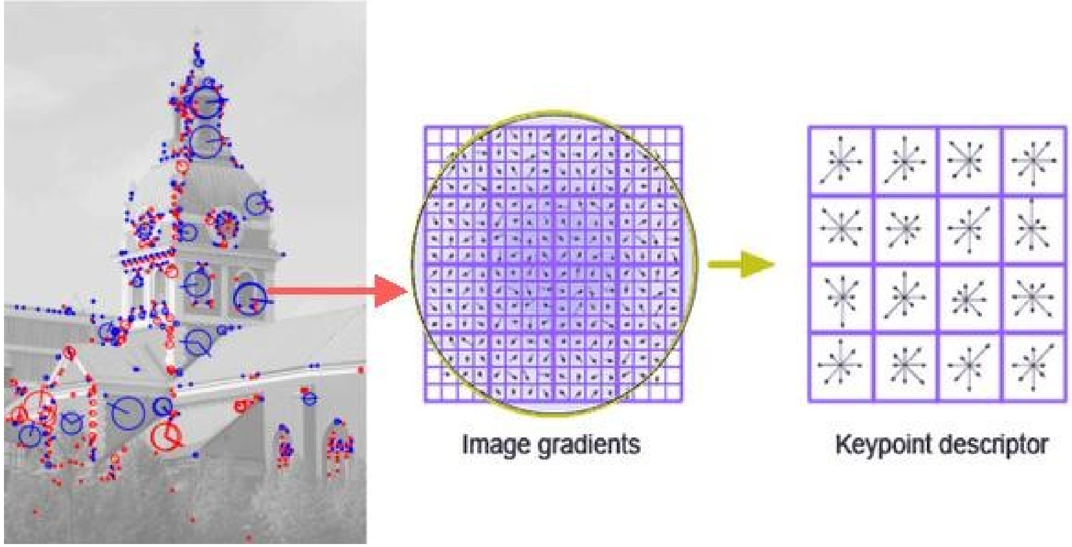
- Calculation of gradient magnitude and orientation at each pixel of the smoothed images in the pyramid
- Determining each key-point's orientation by calculating orientation histogram of its neighborhood

4) Description generation

- Consider an 8-pixel radius (16x16) around a key-point in the pyramid level at which the key is detected
- Calculate an 8-bin orientation histogram for each 4x4 region. The descriptor is the 128-dimensional vector containing the histogram values of the 16 regions.

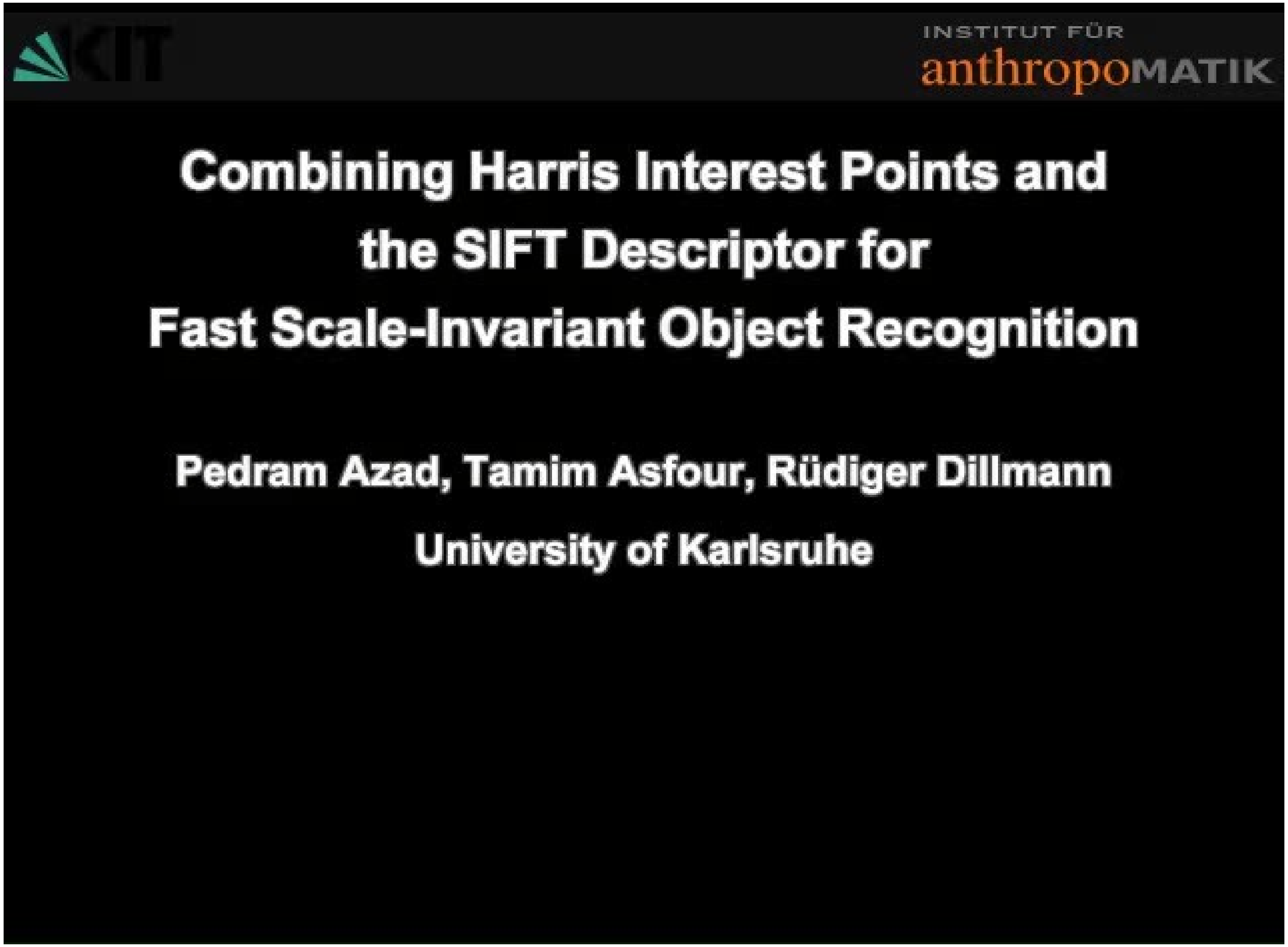
5) Indexing and matching

- Creating a hash table (dictionary) with descriptors of sample images
- Descriptors extracted from a new image are matched to the ones from the dictionary to recognize objects



http://www.codeproject.com/KB/recipes/619039/SIFT.JPG

SIFT - Example



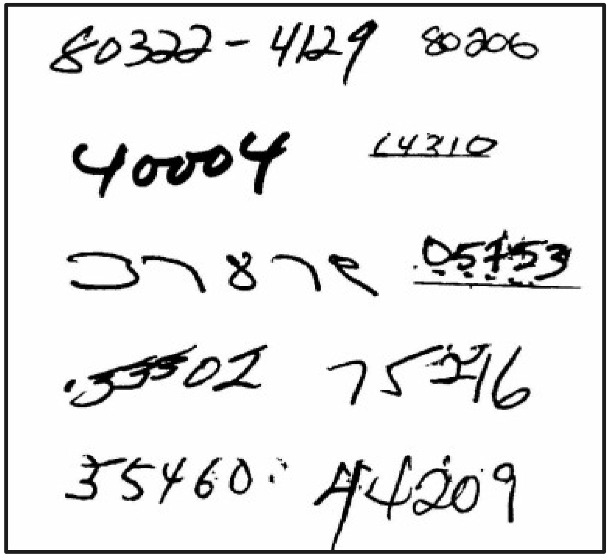
https://www.youtube.com/watch?v=3dY4uvSwiwE

Convolutional Neural Network (CNN)

- Follows the principles of visual processing in the brain
- Basic idea introduced by **Fukushima** in the 1980s
- Improved by **Jan LeCunn**, most popular model **LeNet**
- Convolutional neural networks

  have recently become **very popular**

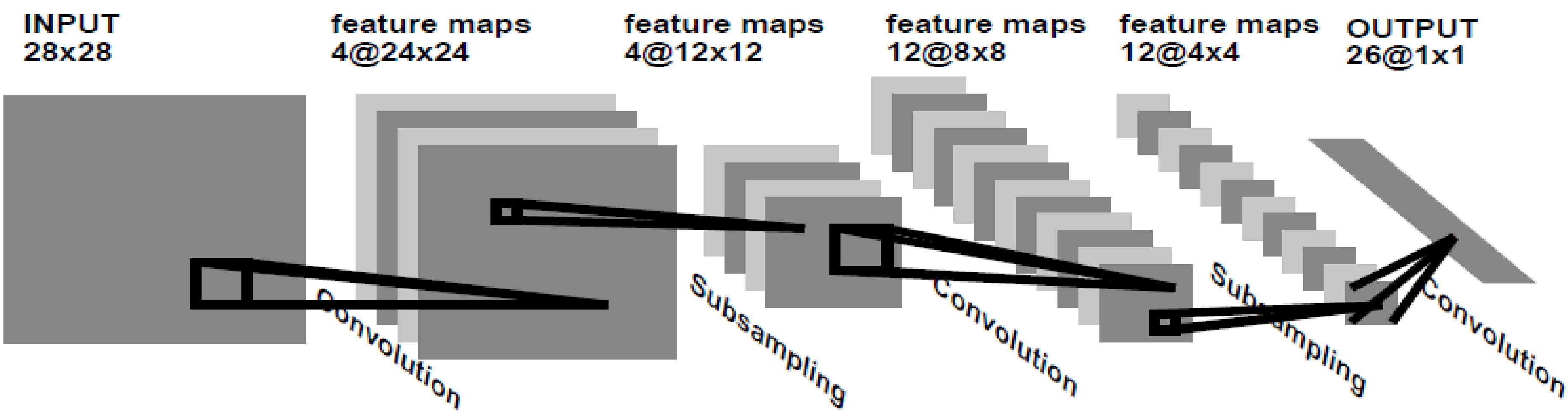
  in image and video processing



Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. Neural computation, 1(4):541-551, 1989.

CNN – the architecture

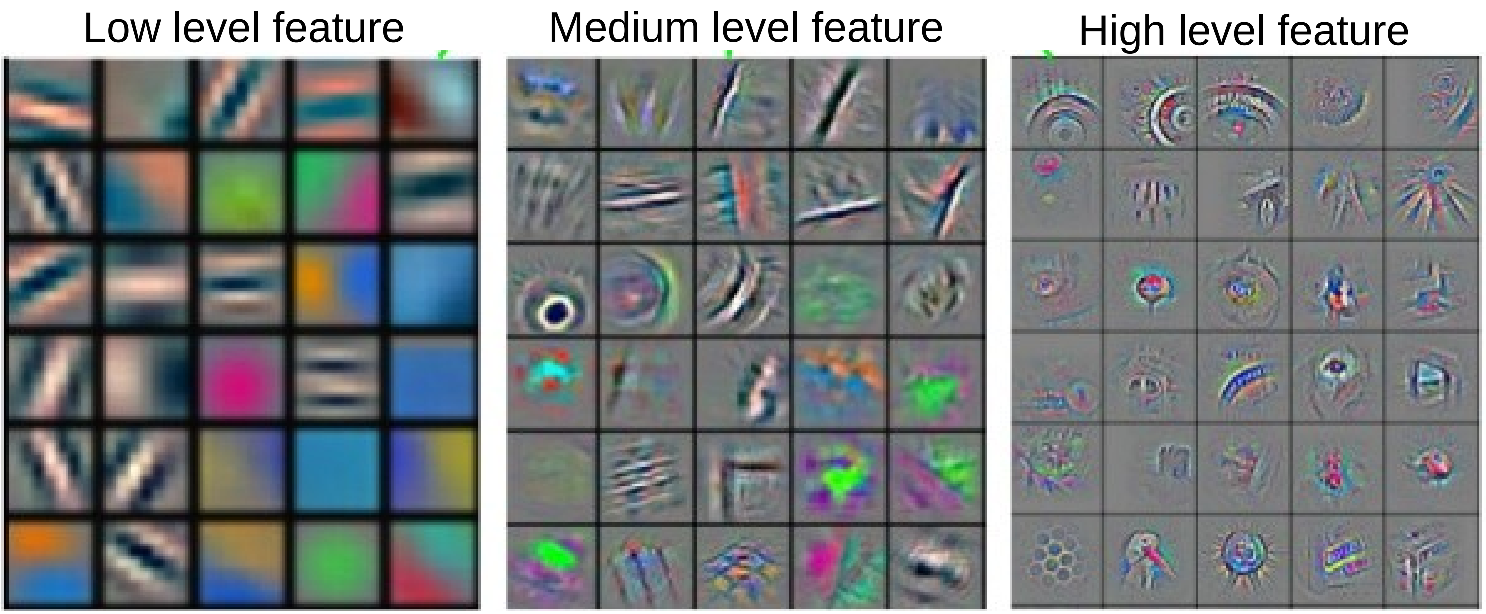
- **Basic principles:**
  - local receptive fields
  - weight sharing
  - subsampling
- **Layer types:**
  - input layer
  - convolutional layer
  - subsampling layer
  - output layer
- **Training:**
  - Backpropagation
  - Adaptive weights



Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. Neural computation, 1(4):541-551, 1989.

CNN – features and feature maps

- Different feature extractors (filters) emerge at different layers during the training of the network
  - Low layer features: lines, contrast, color
  - Medium layer features: corners or other edge/color conjunctions, textures
  - High layer features: more complex, class specific

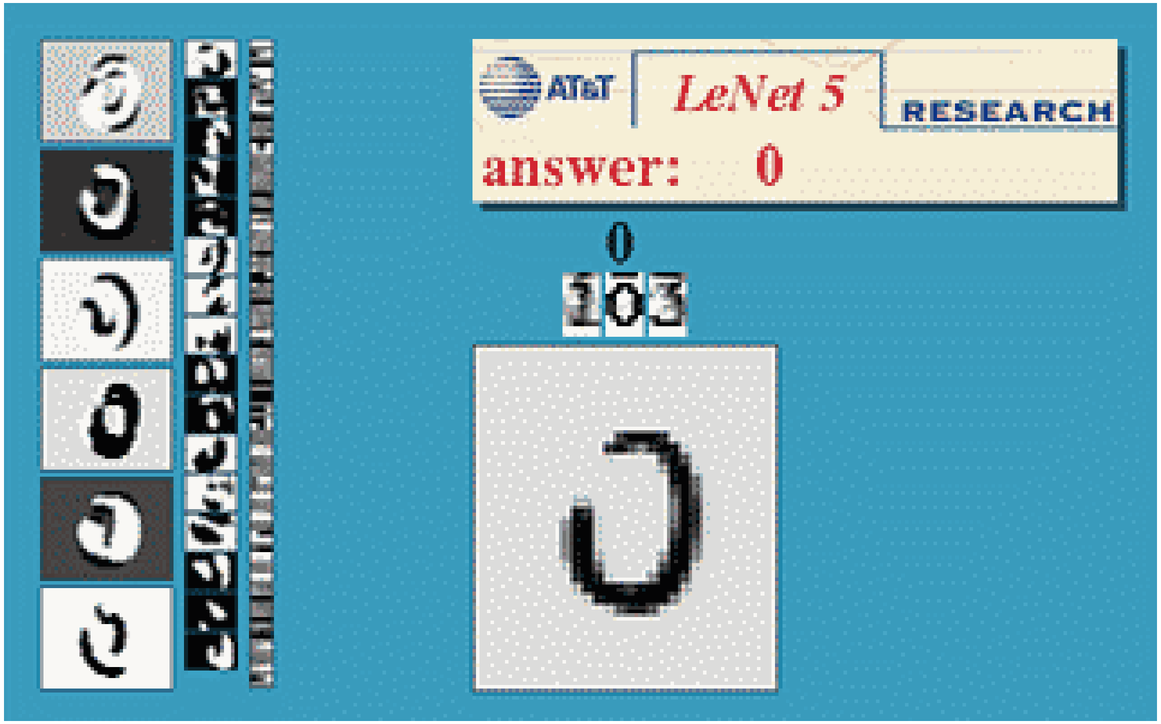


Zeiler, Matthew D., and Rob Fergus. "Visualizing and understanding convolutional networks." Computer Vision–ECCV 2014. Springer International Publishing, 2014. 818-833.

CNN – Example

1) LeNet 5

http://yann.lecun.com/exdb/lenet/index.html



2) ImageNet 2014: Interface for comparing human performance with the winner GoogLeNet

http://cs.stanford.edu/people/karpathy/ilsvrc/

Comparison of SIFT and CNN

Biological plausability: Since the most sophisticated vision system is the human one, the intuition is to understand it and apply its elements in computer vision

SIFT

- **Neurons** in the inferior temporal cortex that respond to complex, scale invariant features
- The feature extraction and learning process of SIFT is **very different** than the processing in the human brain

CNN

- It is a neural network model, it has been inspired by the way **brain** works.
- The way of feature extraction and generalization from simple to complex is much **more similar** to the processing in the human visual system

Comparison of SIFT and CNN (cont.)

Complexity and demand for resources: Design complexity, processing power and memory demands, training set, speed of output

SIFT	CNN
Simpler deign and less parameters to set compared to CNN	Needs **experience** to make design decisions
Less processing power needed, memory needed for storing features for each image	**High demand for processing** during the training phase, memory needed to store the weights of the network
Smaller training set	The **bigger the training set** the better
Fast	**Slower** than SIFT

Comparison of SIFT and CNN (cont.)

Applicability: Range of problems and scenarios in which SIFT and CNN can be applied.

SIFT

- More relevant for **identification** tasks
- SIFT and SIFT like descriptors are used in **vide range** of vision tasks
- Can be used for **real time** scenarios

CNN

- More relevant for **classification** and **categorization** tasks, has very good generalization abilities
- Currently **very popular** model for image and video tasks

CNN and SIFT - Pros & Cons

	Good	Bad
SIFT	• Identification tasks • Simple to implement • Fast	• Poor generalization • Not robust to non-linear transformations
CNN	• Classification tasks • Strongly bio-inspired • Very good generalization	• Lots of processing power • Big training datasets • Parameters to set

Questions?

Thank you for the attention

Literature

1) Liter, Jeffrey C., and Heinrich H. Bülthoff. "An introduction to object recognition."Zeitschrift für Naturforschung C 53.7-8 (1998): 610-621.

2) Kruger, Norbert, et al. "Deep hierarchies in the primate visual cortex: What can we learn for computer vision?." Pattern Analysis and Machine Intelligence, IEEE Transactions on 35.8 (2013): 1847-1871.

3) Lowe, David G. "Object recognition from local scale-invariant features."Computer vision, 1999. The proceedings of the seventh IEEE international conference on. Vol. 2. Ieee, 1999.

4) Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. Neural computation, 1(4):541-551, 1989.

5) Fischer, Philipp, Alexey Dosovitskiy, and Thomas Brox. "Descriptor matching with convolutional neural networks: a comparison to sift." arXiv preprint arXiv:1405.5769 (2014).

6) Zeiler, Matthew D., and Rob Fergus. "Visualizing and understanding convolutional networks." Computer Vision–ECCV 2014. Springer International Publishing, 2014. 818-833.