Boston University
CAS CS 640: AI

1st & ½ Lecture on Computer Vision

by Margrit Betke

October 12 & 17, 2023

Learning Objectives for this Lecture

Computer Science

- Understand formats of images used as inputs to AI models: greyscale, color, medical scans
- Understand differences and similarities between pre-2012 "traditional computer vision" and post-2012 neural-network-based computer vision & see examples
- Understand why convolution is powerful
- Understand how tools from estimation theory can be used to measure recognizability of objects in images
- Learn about breakthrough dataset ImageNet
- Learn about early CNNs used in computer vision

What is an image?

Computer Science

- Images are fields of colored dots
- Each dot is called a pixel =picture cell
- Standard test image with detail, shading, texture, sharp & blurry regions:

Lena Soderberg '72

(controversy!)

Slide credit: Diane Theriault

Color Models

Computer Science

- Images can be gray scale, color, or color with an alpha (transparency) channel
- Most common color representation is RGB (Red, Green, Blue). This is the representation used to put pixels on the screen

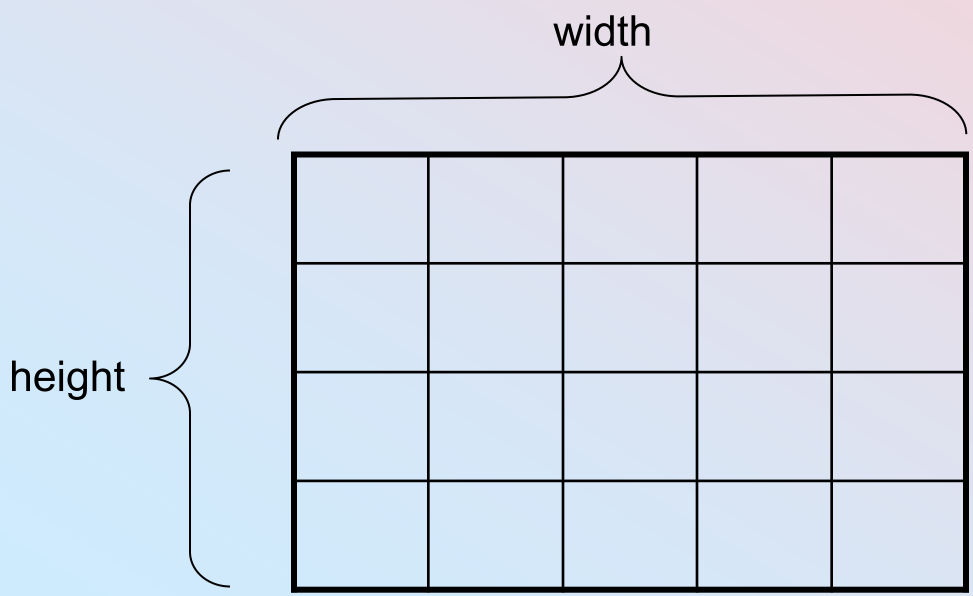
- Other models include CMYK (used for print) and YUV (often used for input from cameras, compression, and transmission)

Slide credit: Diane Theriault

What is an image?

Computer Science

- Images are 2 dimensional arrays of data, with an associated width, height, and color depth.
- Images typically use one byte per color channel per pixel.
- Gray images have 1 color channel. RGB images have 3 color channels. RGBA images have 4 color channels.



Slide credit: Diane Theriault

Digital Image File Formats

Computer Science

Image:

Header
Table = Map = Raster of Pixel Values

Size of table, color, compression scheme

Gray-scale images: generally 1 byte per pixel

Color images: 3 numbers (each 1 byte) per pixel

Medical images, e.g., CT, MRI:
typically 2 bytes per voxel

Example: PGM Image

Computer Science

Image file

Image ??

P2
3 3 255

0	255	0
220	0	20
0	130	0

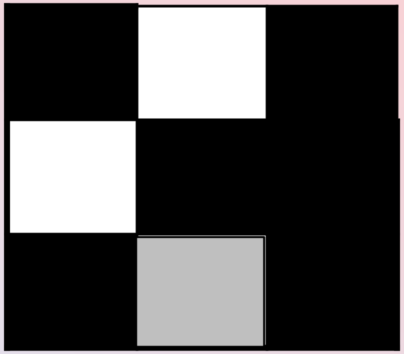
Example: PGM Image

Image file

```
P2
3 3 255

0 255 0
220 0 20
0 130 0
```

Image



Light: Electromagnetic Waves

Wavelength λ

Visible Range

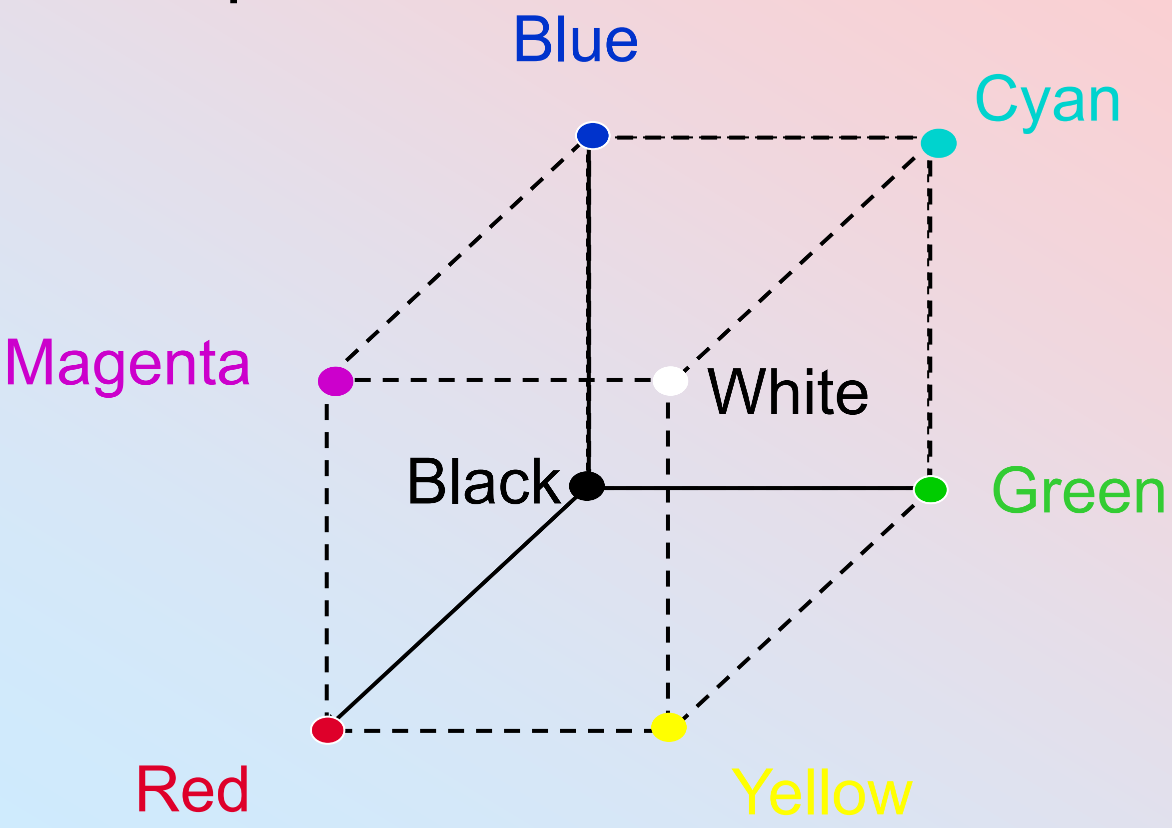
X-rays Violet Blue Yellow Green Red Radio

λ

50 nm 400 450 500 550 600 650 700 nm mm - km

RGB Color Space

Additive Space



Example: PPM Image

Computer Science

Image file

P3
3 3 255

0	0	0	255	0	0	0	0	0
0	255	0	0	0	0	255	255	0
0	0	0	0	0	255	0	0	0

Image ??

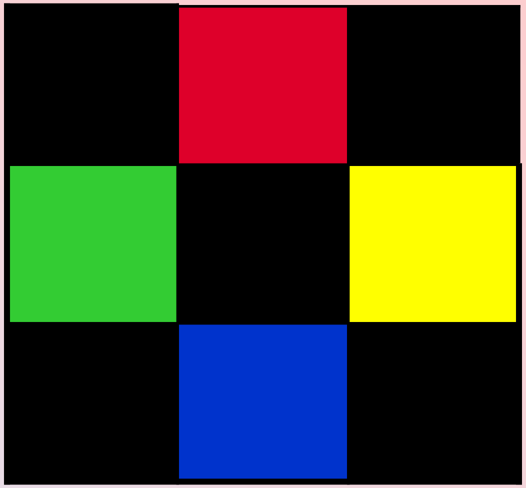
Example: PPM Image

Image file

```
P3
3   3   255

0   0   0   255  0   0   0    0    0
0  255  0    0   0   0  255  255   0
0   0   0    0   0  255  0    0    0
```

Image



How do I get at the data?

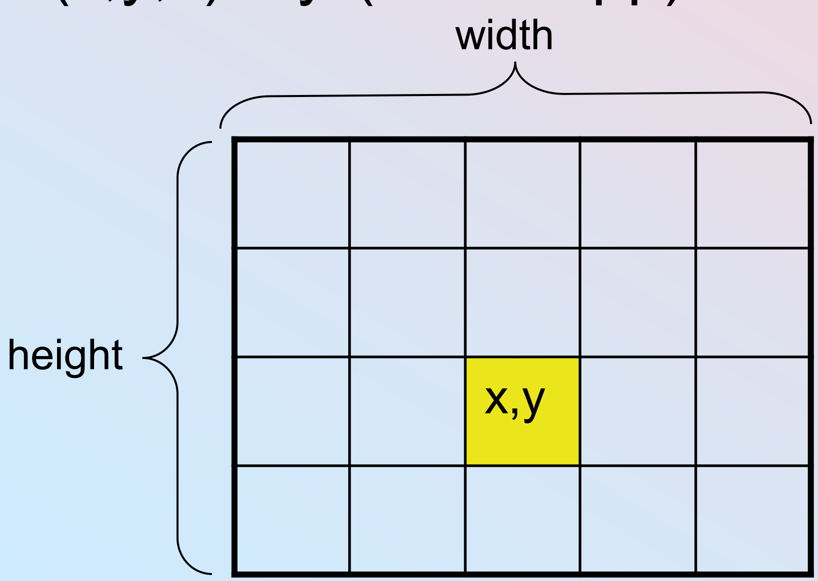
Computer Science

- Some image-handling APIs have nice interfaces, but speed can be a problem.
- You will probably have to handle the bytes of data directly at some point

Slide credit: Diane Theriault

How do I get at the data?

- X = desired row
- Y = desired column
- C = color channel (red, green, blue, …).
- Bpp = Bytes per pixel (color channels)
- Image data is normally stored in row major order
- Note that there may be multiple values associated with each x,y pixel
- Data(x,y,c) = y*(width*Bpp) + x*Bpp + c



Slide credit: Diane Theriault

Example of a “Traditional” Computer Vision Algorithm: Color to Gray Scale Conversion

- Pre-NN-revolution Computer Vision: Algorithms

- Example of such a pre-2012 algorithm:

  Converting from color to gray scale,
  a very common operation

Color-to-Grayscale Conversion

Computer Science

- “Quick and dirty” conversion: Grab the Green Channel
- Average R, G, B: (R+G+B)/3
- Max(R, G, B)
- Weigh them: 0.3*R + 0.6*G + 0.1*B

Slide credit: Diane Theriault

Image File Formats

Computer Science

- PPM / PGM is the simplest file format ever, but not supported by Photoshop or MS Image Viewer. Uncompressed.
- BMP: Microsoft's uncompressed image format
- GIF: Images are compressed using run-length encoding, and reducing the number of colors used. Licensed, not open
- JPEG: Images are compressed by throwing away high frequency information

Slide credit: Diane Theriault

Tools of the Trade

- OpenCV is a widely used, open-source computer vision library maintained by Intel
- Provides libraries for image I/O, movie I/O and camera capture
- Industrial strength computer vision and image processing implementations
- Quick and dirty GUI toolkit

Slide credit: Diane Theriault

Tools of the Trade

Computer Science

- Irfanview is a freely available image viewer and possibly one of the most useful programs ever.

Common Gotcha's

Computer Science

- Sometimes the mapping from a weird looking image to the actual error is not obvious

Slide credit: Diane Theriault

Common Gotcha's
Color Order

Computer Science

- RGB vs. BGR

Slide credit: Diane Theriault

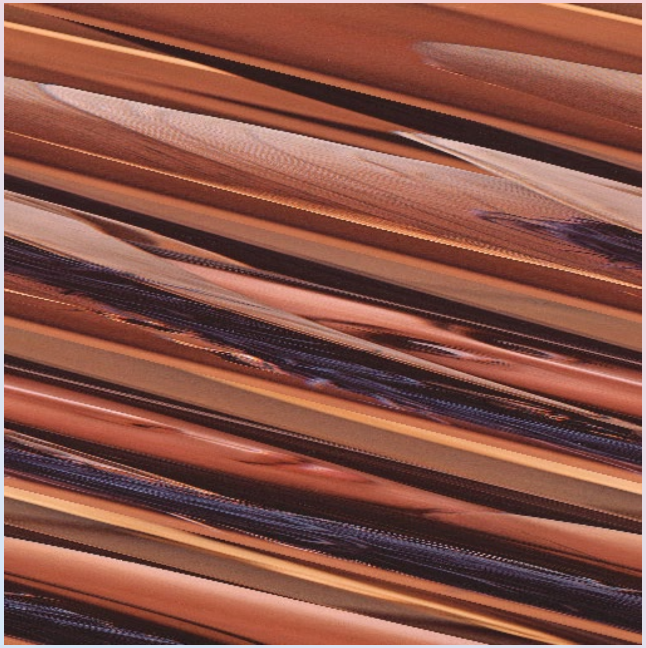
Common Gotcha's
Wrong Width

Computer Science

- Incorrect width can result in an image with strong diagonal structure

Actual width: 512

This image width: 508

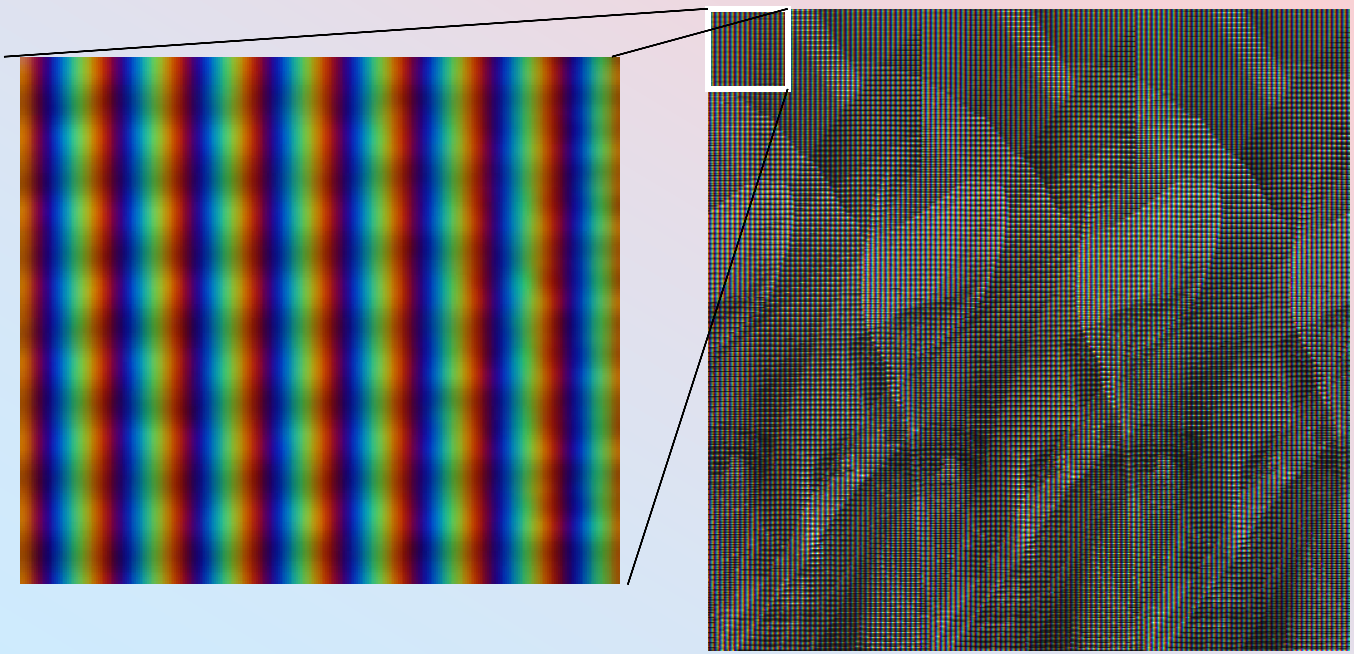


Slide credit: Diane Theriault

Common Gotcha's
Wrong Color Depth

Computer Science

- Mismatched color depth can result in an image with a rainbow effect



Slide credit: Diane Theriault

Common Gotcha's Windows line endings

Computer Science

- On Windows, it is critically important to open image files in binary mode.
- Otherwise,windows helpfully strips out any bytes with value '\r' (20).

Slide credit: Diane Theriault

Today's Computer Vision: Mostly Neural Networks

Computer Science

- Deep neural networks
- Convolutional neural networks
- Transformers
- Diffusion models

\+ traditional computer vision algorithms,
representations, geometry, and tricks

- In CS 640: Both traditional & NN Computer Vision

1D Discrete Convolution

1D Convolution:
Time signal *f* and shifted time signal *g* are multiplied and added:

$$(f * g)[n] \stackrel{\text{def}}{=} \sum_{m=-\infty}^{\infty} f[m]\, g[n-m]$$
$$= \sum_{m=-\infty}^{\infty} f[n-m]\, g[m].$$

2D generalization:
f = input image, g = template image
(or CNN function)

2D Convolution Example

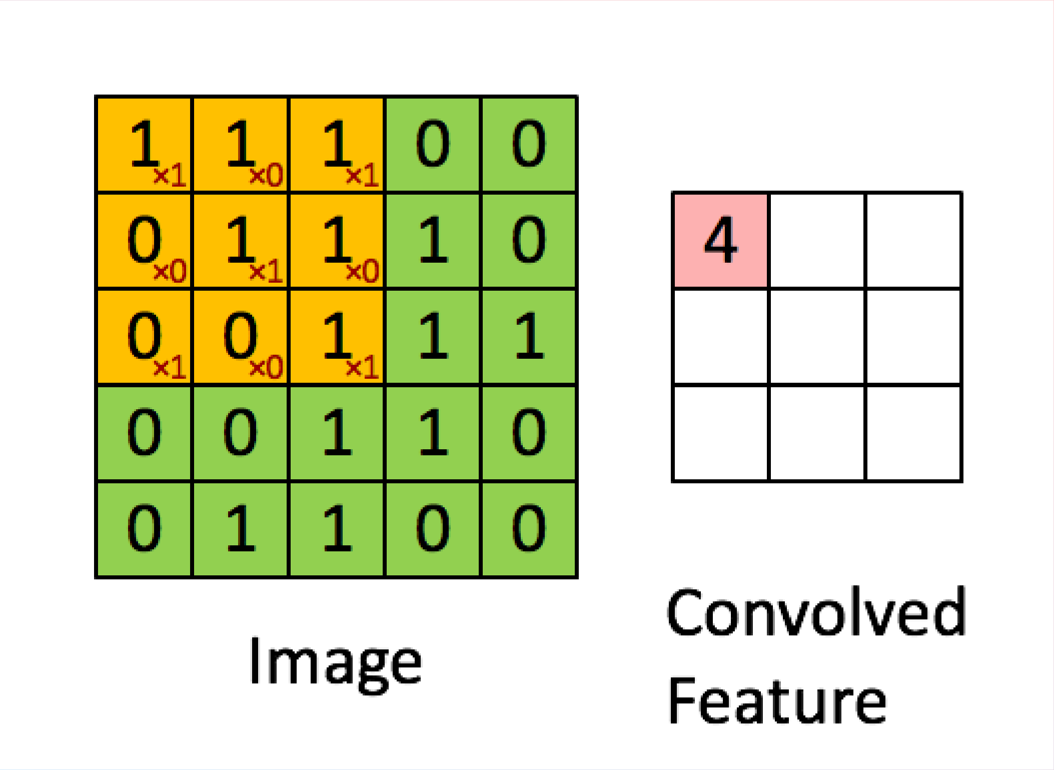
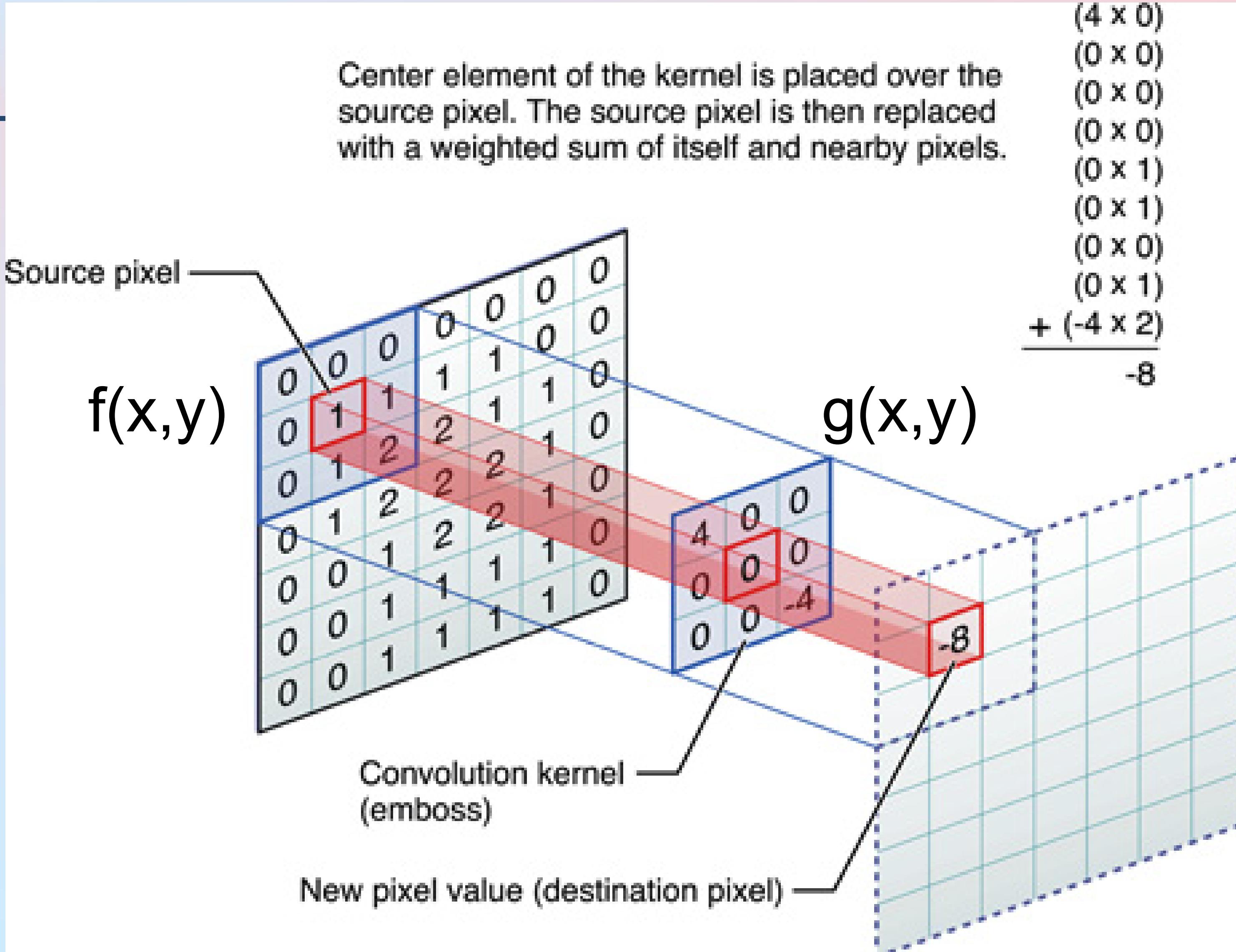


Image Credit: Nvidia



f*g

Image Credit: Madhushree Basavarajaiah

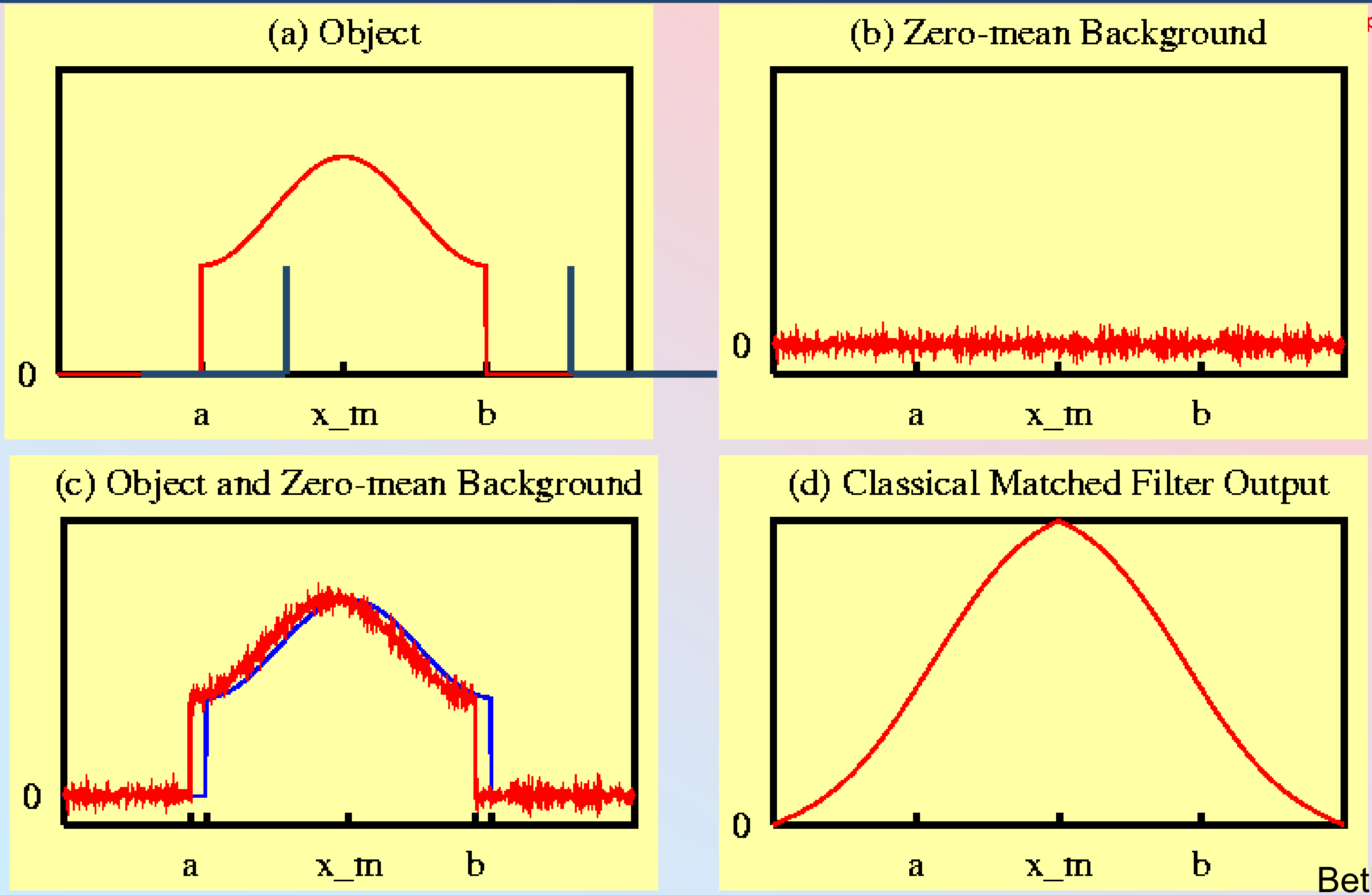
Why is Convolution Powerful?

Signal Processing:

Convolution is used to define a "matched filter" for locating "targets" in time signals

Optimal algorithm if noise is Gaussian.

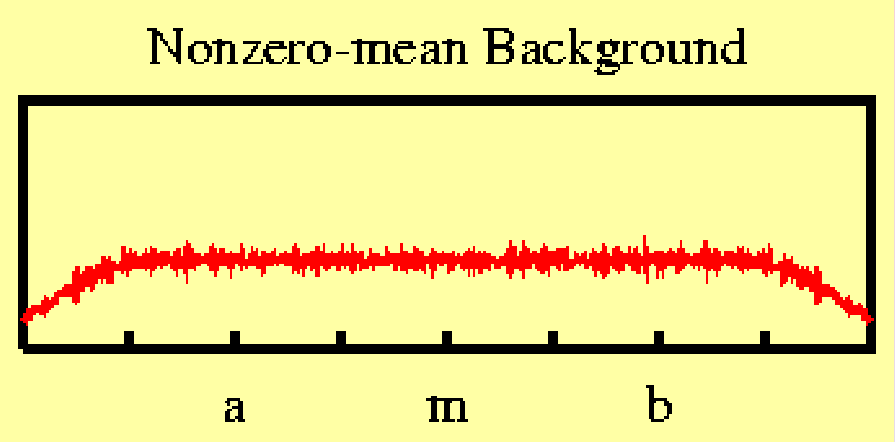
1D Position Estimation: Σ object*background

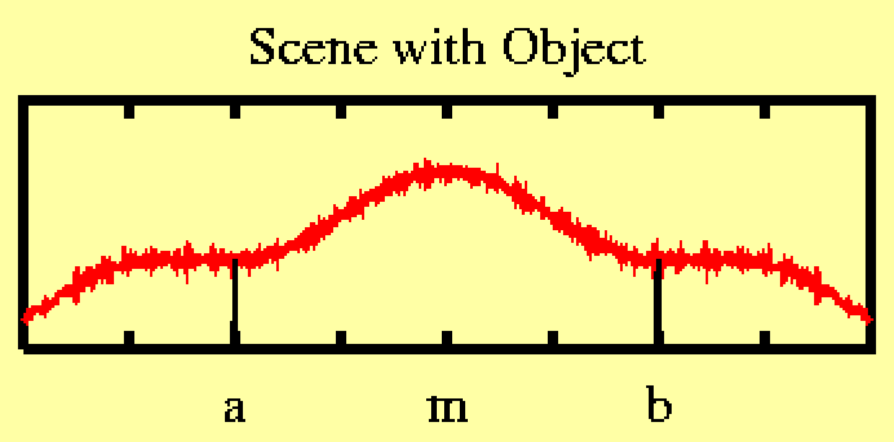


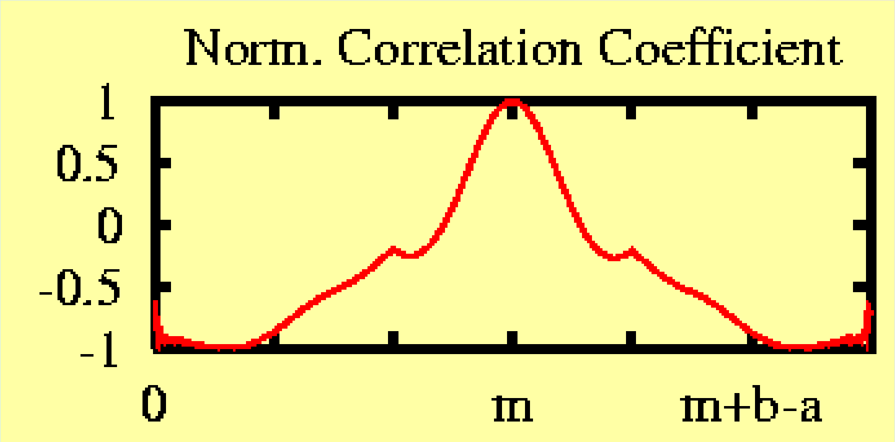
Betke, Makris, IJCV 2001

Another 1D convolution example:

Computer Science





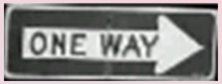


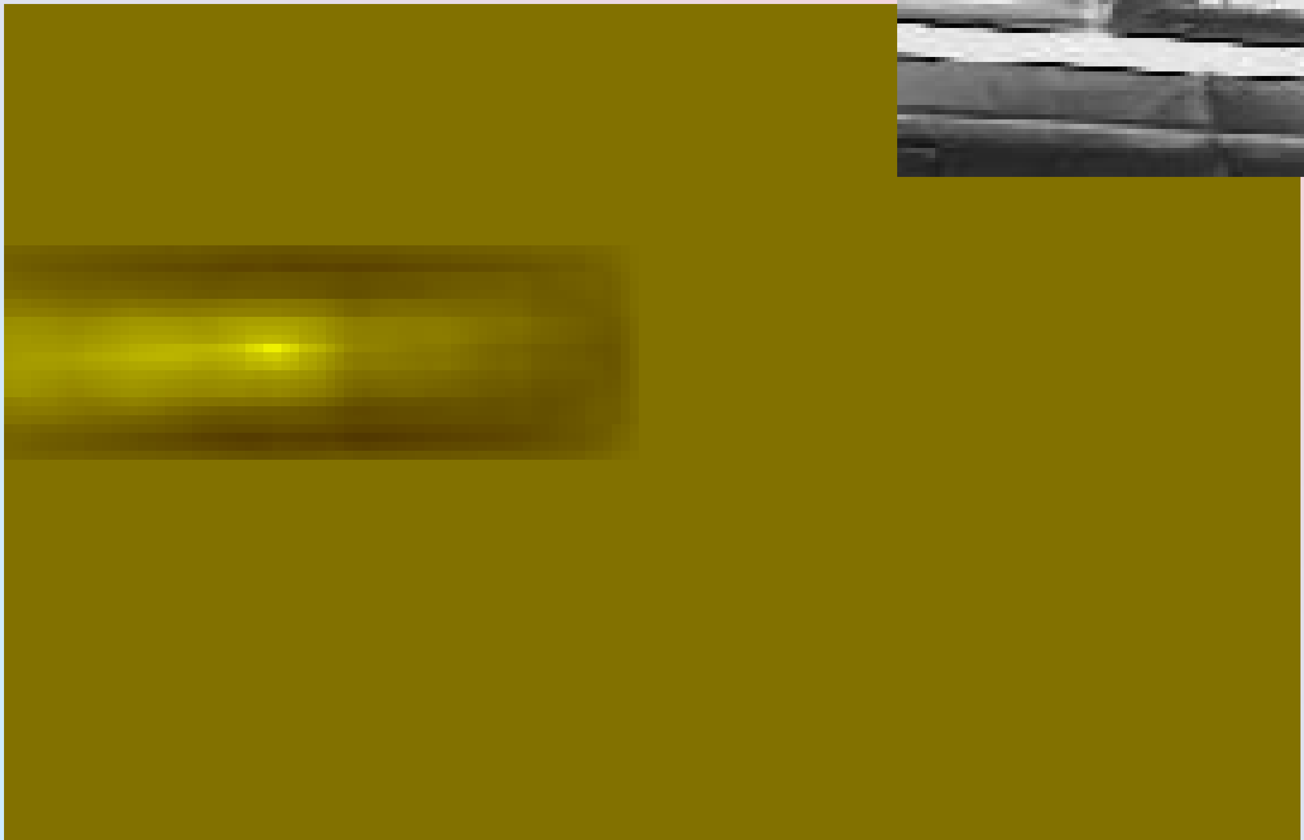
= convolution/std-devs

Betke, Makris, IJCV 2001

2D Position Estimation

Convolution of one-way sign with itself

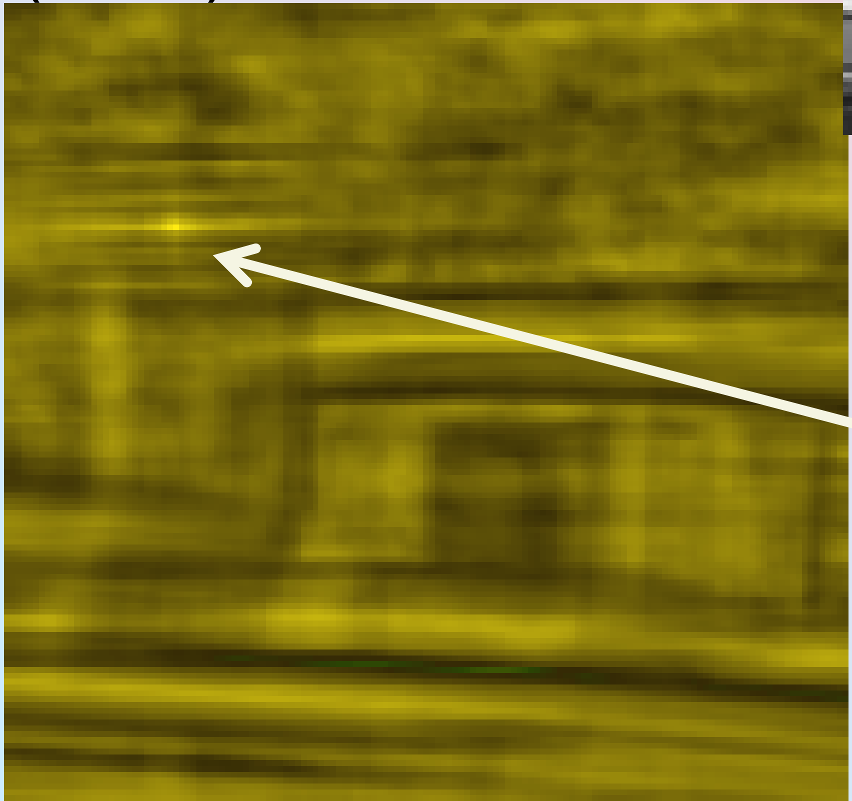




Betke, Makris,
IJCV 2001

2 D Position Estimation

Convolution of one-way sign with scene (NCC)

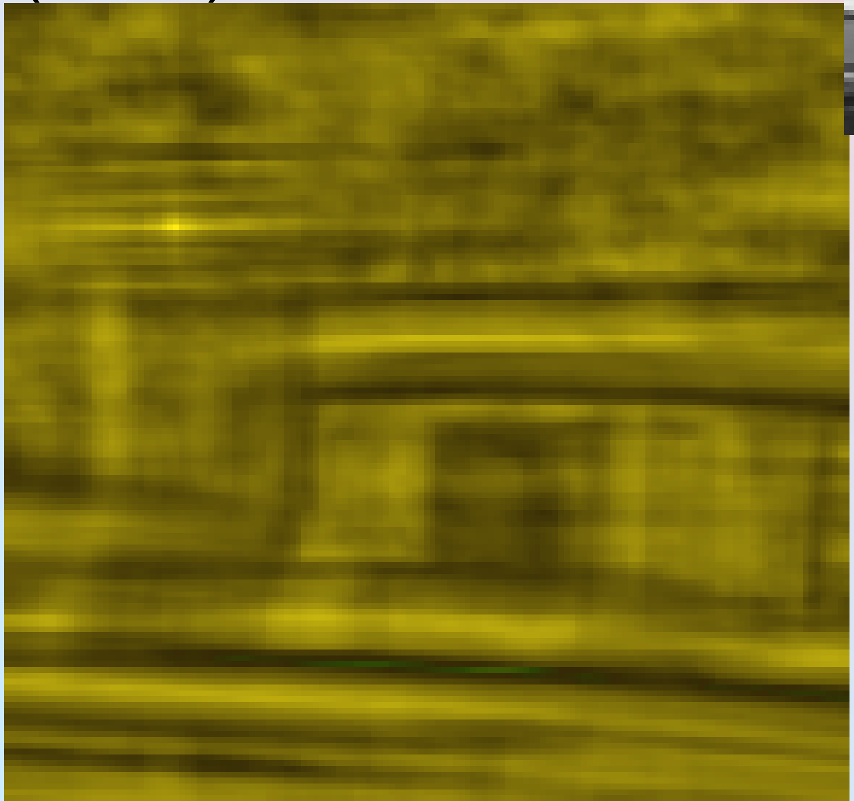


Peak in performance surface (= negative loss fct) at correct location

Betke, Makris, IJCV 2001

2 D Position Estimation

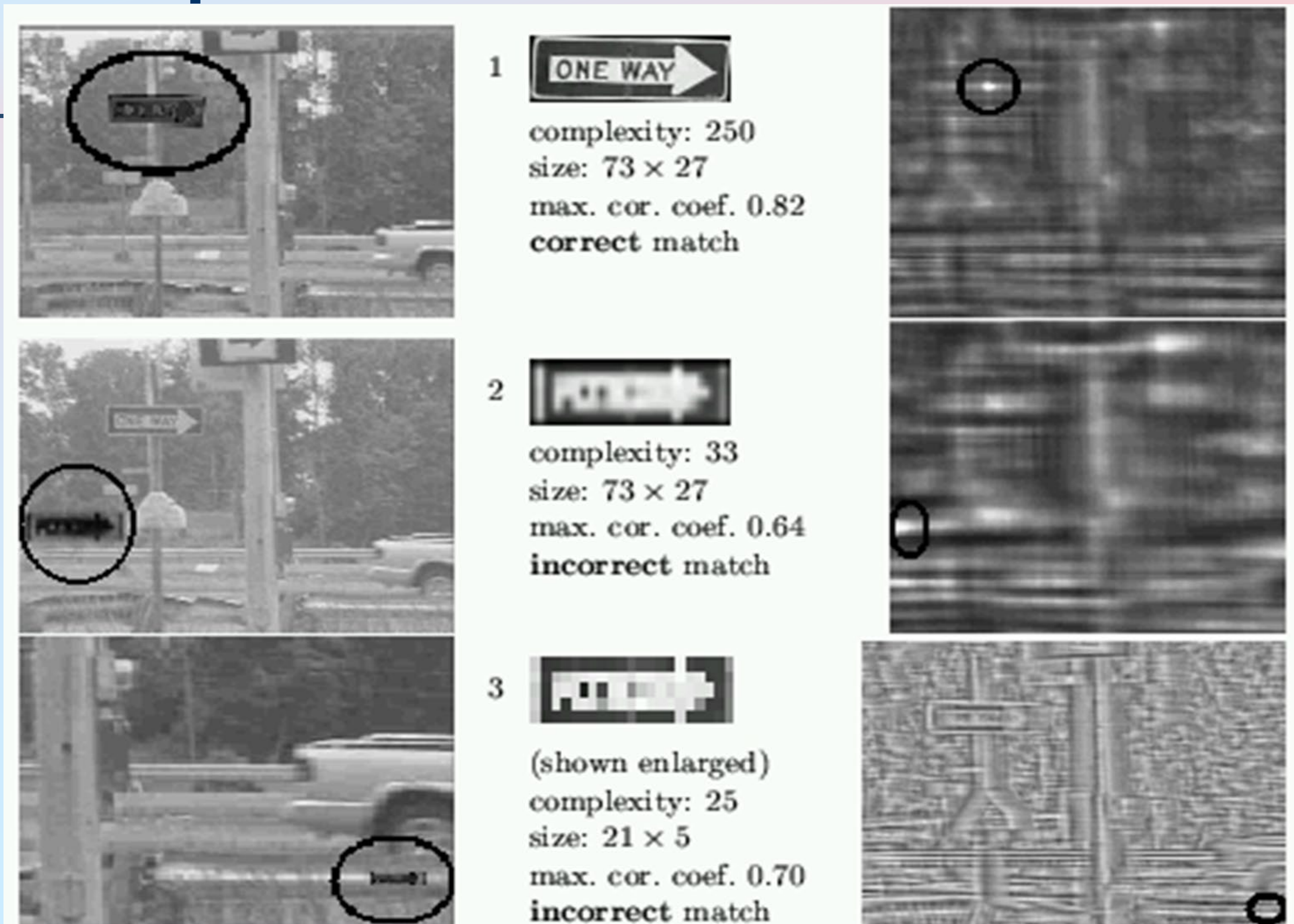
Convolution of one-way sign with scene (NCC)



This performance surface is computed for correct size of one-way sign

Different surfaces for different sizes of object

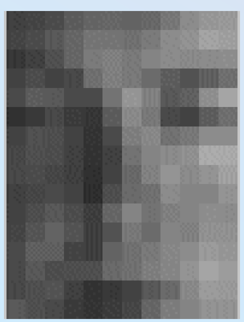
Sample Performance Surfaces



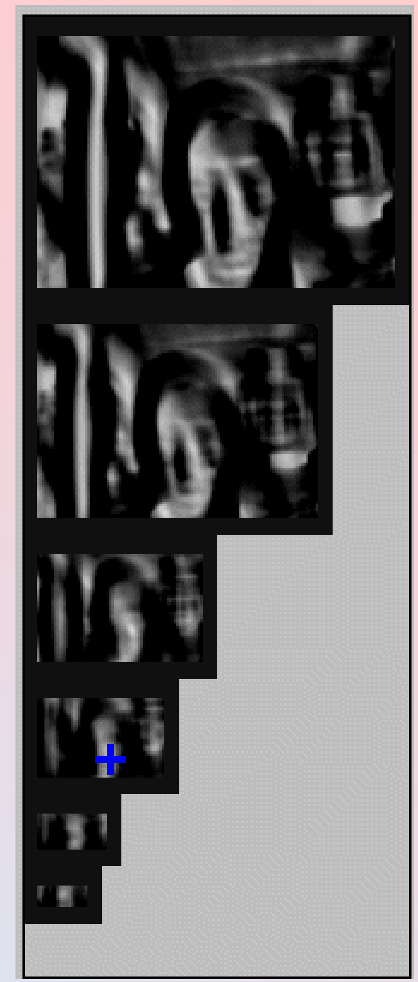
Multi-Resolution Matching

Normalized correlation coefficient over multi-resolution search space:

$$r = \frac{1/n \quad \Sigma_i (s_i - mean(s)) (m_i - mean(m))}{(\sigma_s \sigma_m)}$$



← Template matched over all resolutions →



BOSTON UNIVERSITY
Computer Science

Finding the Face and its Movement by Locating the Best Match of a Face Template

(a) Input

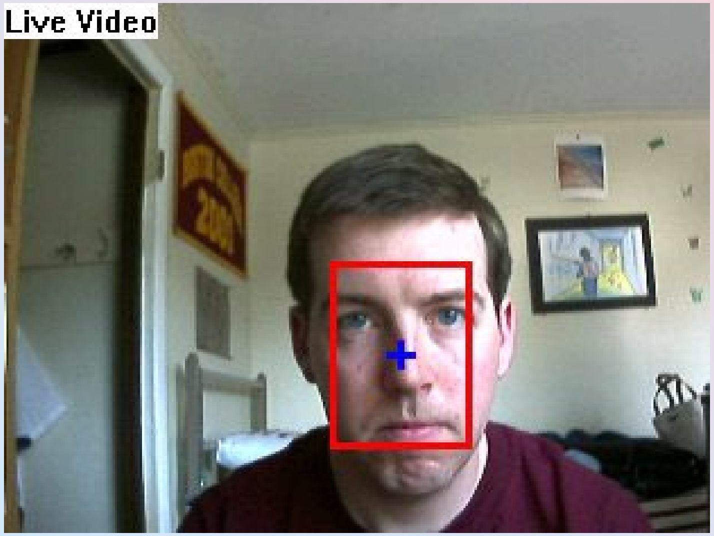
You can apply template matching to a small version of your input image and use that search result to start searching for a match in the 2nd smallest images. Repeat until the original size is processed.

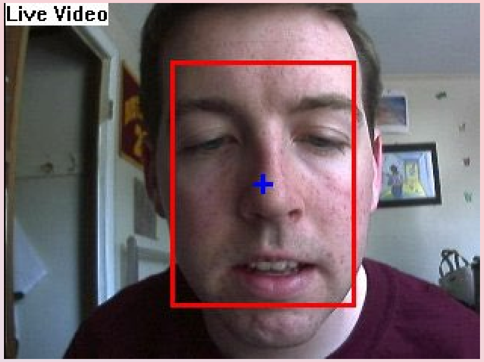
(d) Correlation

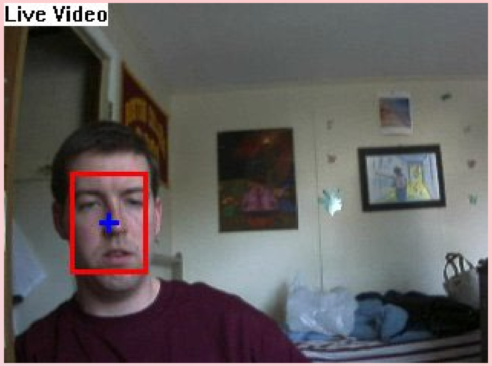
Face Detection

Computer Science

Data Variability







Large Face

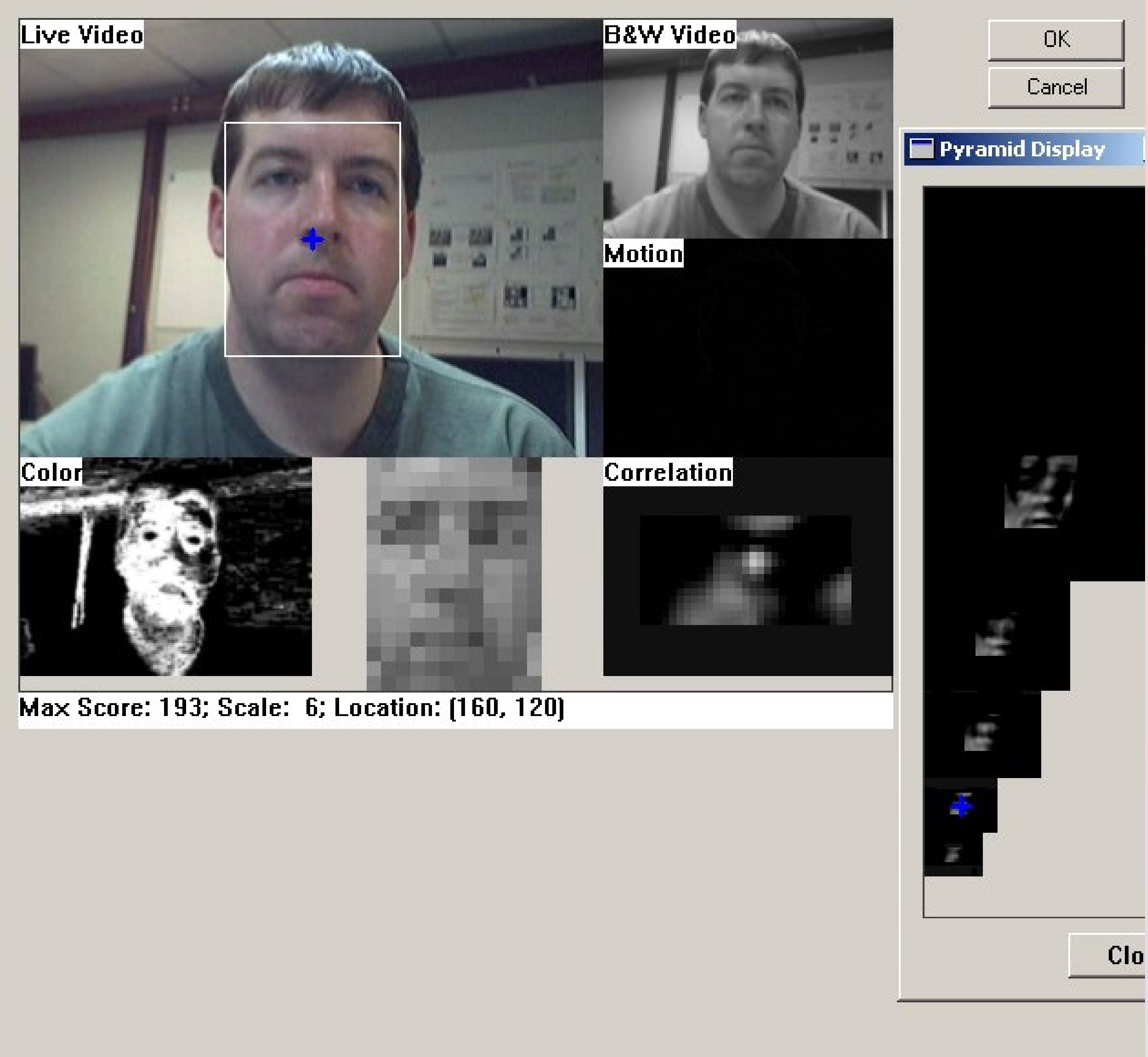
Small Face

Shadows

Cluttered background

Face Detection Interface

Computer Science



Live Video
B&W Video
Motion
Color
Correlation
Max Score: 193; Scale: 6; Location: (160, 120)
OK
Cancel
Pyramid Display
Clo

Object Recognition = Parameter Estimation

Computer Science

Affine parameterization $\mathbf{x}' = A\mathbf{x} + b$ => estimate $\mathbf{a}$

Likelihood function

$$P(\mathbf{I}|\mathbf{a}) = \frac{1}{(2\pi\sigma^2)^{\frac{NM}{2}}} \exp\left(-\frac{1}{2\sigma^2}\sum_{k=1}^{MN}(I_k - m_k(\mathbf{a}))^2\right)$$

CR lower bound

$$\mathrm{E}[(\hat{\mathbf{a}} - \mathbf{a})(\hat{\mathbf{a}} - \mathbf{a})^T] \geq \mathbf{J}^{-1}$$

Betke, Makris, IJCV 2001

Fisher Information

a_4 = s

$$J_{ij} = \frac{1}{\sigma^2} \sum_x \sum_y \left(\frac{\partial m(x,y,\mathbf{a})}{\partial a_i} \frac{\partial m(x,y,\mathbf{a})}{\partial a_j} \right)$$

a_1 = x

a_2 = y

$a_3 = \theta$

Object Coherence

Computer Science

CRLB: $$\mathrm{E}[(\hat{a}_i - a_i)^2] \geq [\mathbf{J}^{-1}]_{ii} = \frac{\sigma^2}{E}\,\ell_i^2$$

Energy: $$E = \sum_{(x,y)\in O} |m(x,y;\mathbf{a})|^2$$

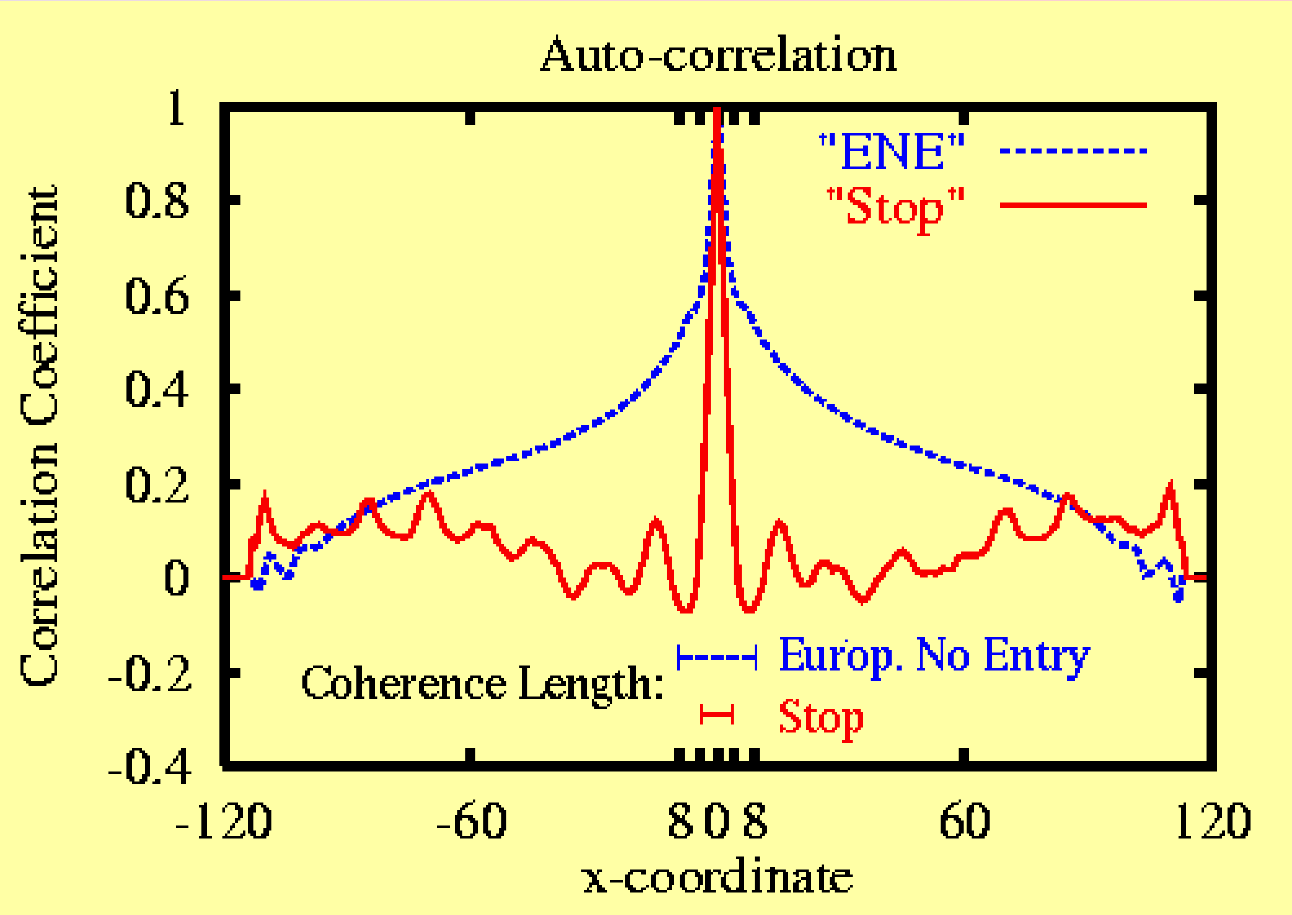
Coherence scale and volume:

$$\ell_i = \left([\mathbf{J}^{-1}]_{ii}\,\frac{E}{\sigma^2}\right)^{\frac{1}{2}}$$

$$V = \left(\frac{E}{\sigma^2}\right)^{\frac{n_a}{2}} |\mathbf{J}|^{-\frac{1}{2}}$$

Coherence Length Scale

Computer Science



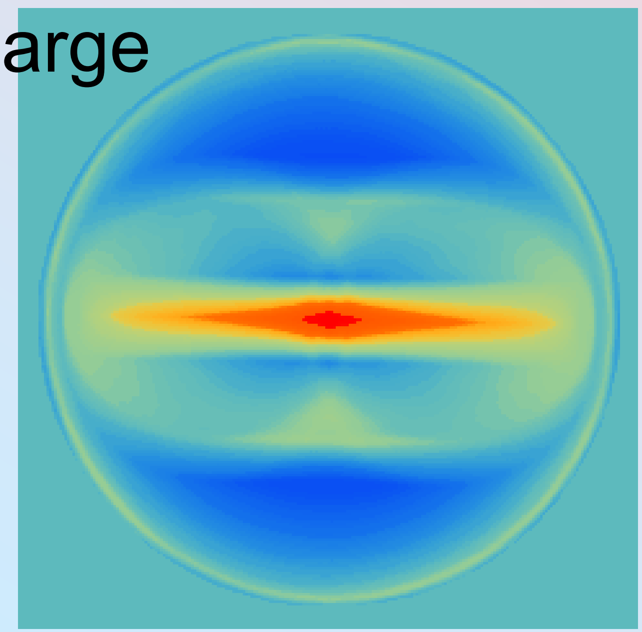
Since coherence length of Stop sign < No-Entry Sign, resolving location (x-coordinate) of Stop sign is easier

Coherence Area

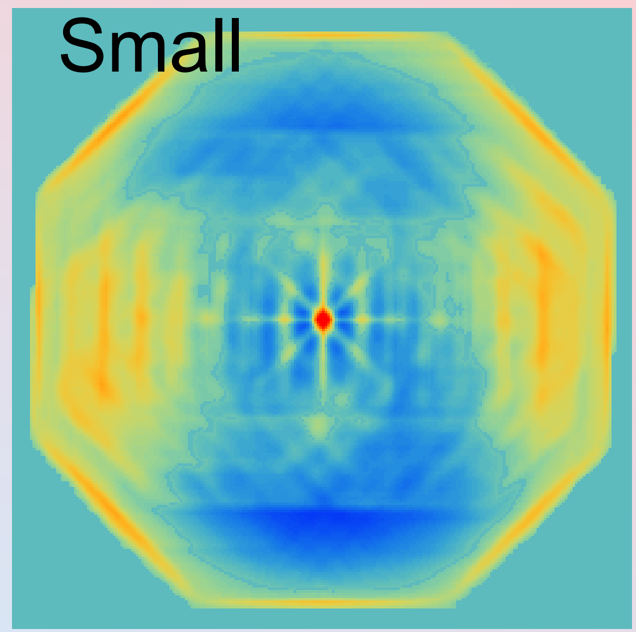
Computer Science

Betke, Makris, IJCV 2001

Large



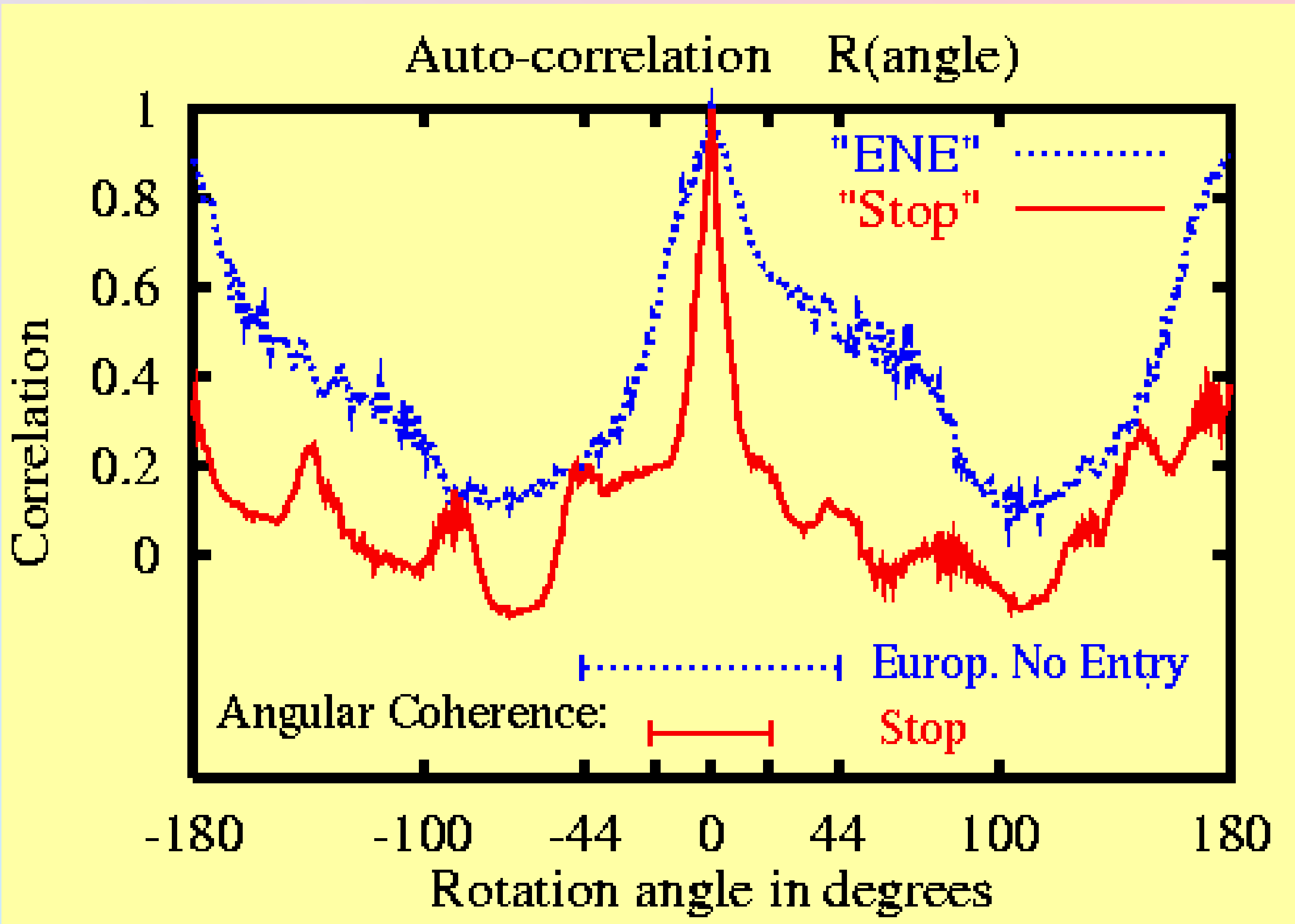
Small

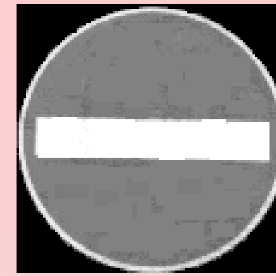


Resolving (x,y) location is easier for Stop sign

Angular Coherence Scale

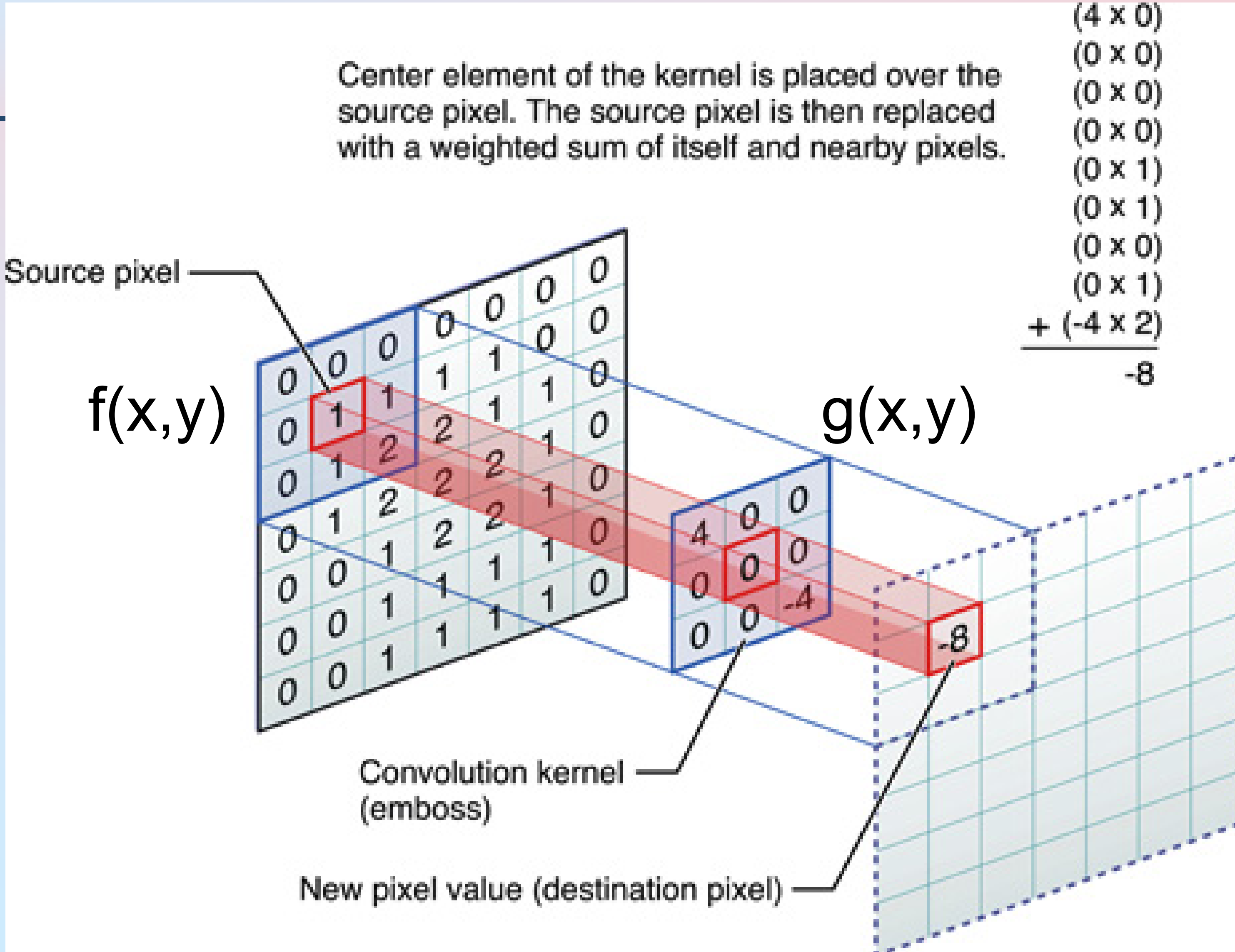
Computer Science





Peaks at ~45, 90, … degrees

Betke, Makris, IJCV 2001



f*g

Computer Science

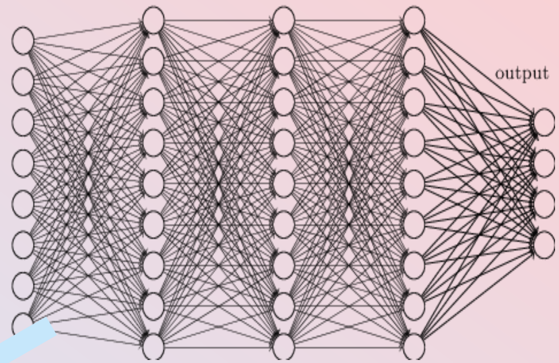
Image Credit: Madhushree Basavarajaiah

Conclusions on Coherence

- Using the Fisher Information matrix, we can compute the coherence scales of objects
- Coherence scales define the recognizability of object parameters
- Intuitively, coherence areas = “cells” = “interconnected parts” =“degrees of freedom”
- Coherence scales can be visualized with autocorrelations, i.e., “object convolution with itself”
- Neural nets compute many convolutions and memorize coherence scales of objects

Back to Neural Nets & their Success in Solving Computer Vision Problems

Large labeled datasets



Deep neural networks

GPU technology

Slide credit: Dinesh Jayaraman

Convolutional Neural Networks (CNN, ConvNet, DCN)

- CNN = a multi-layer neural network with
  - **Local** connectivity:
    - Neurons in a layer are only connected to a small region of the layer before it
  - **Share** weight parameters across spatial positions:
    - Learning shift-invariant filter

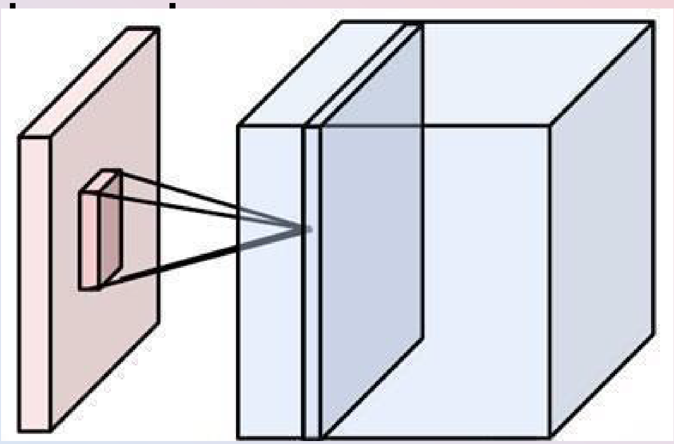
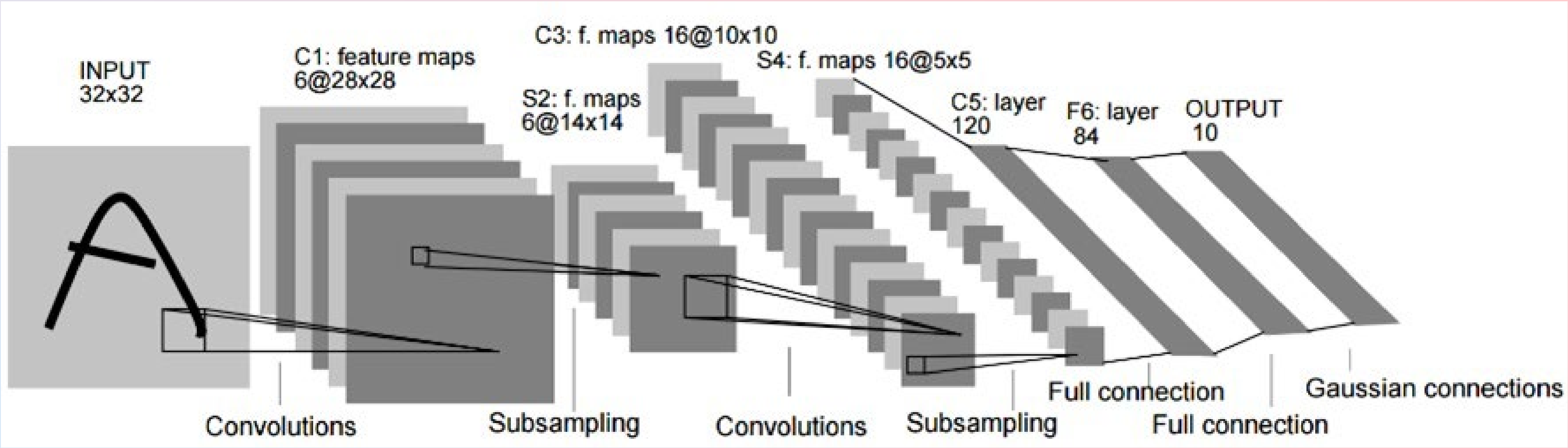


Image credit: A. Karpathy

LeNet [LeCun et al.]



1990: Zipcode recognition

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

LeNet-1 from 1993

Gradient-based learning applied to document recognition [LeCun, Bottou, Bengio, Haffner 1998]

LeCun Interview, Oct. 5, 2023

Computer Science

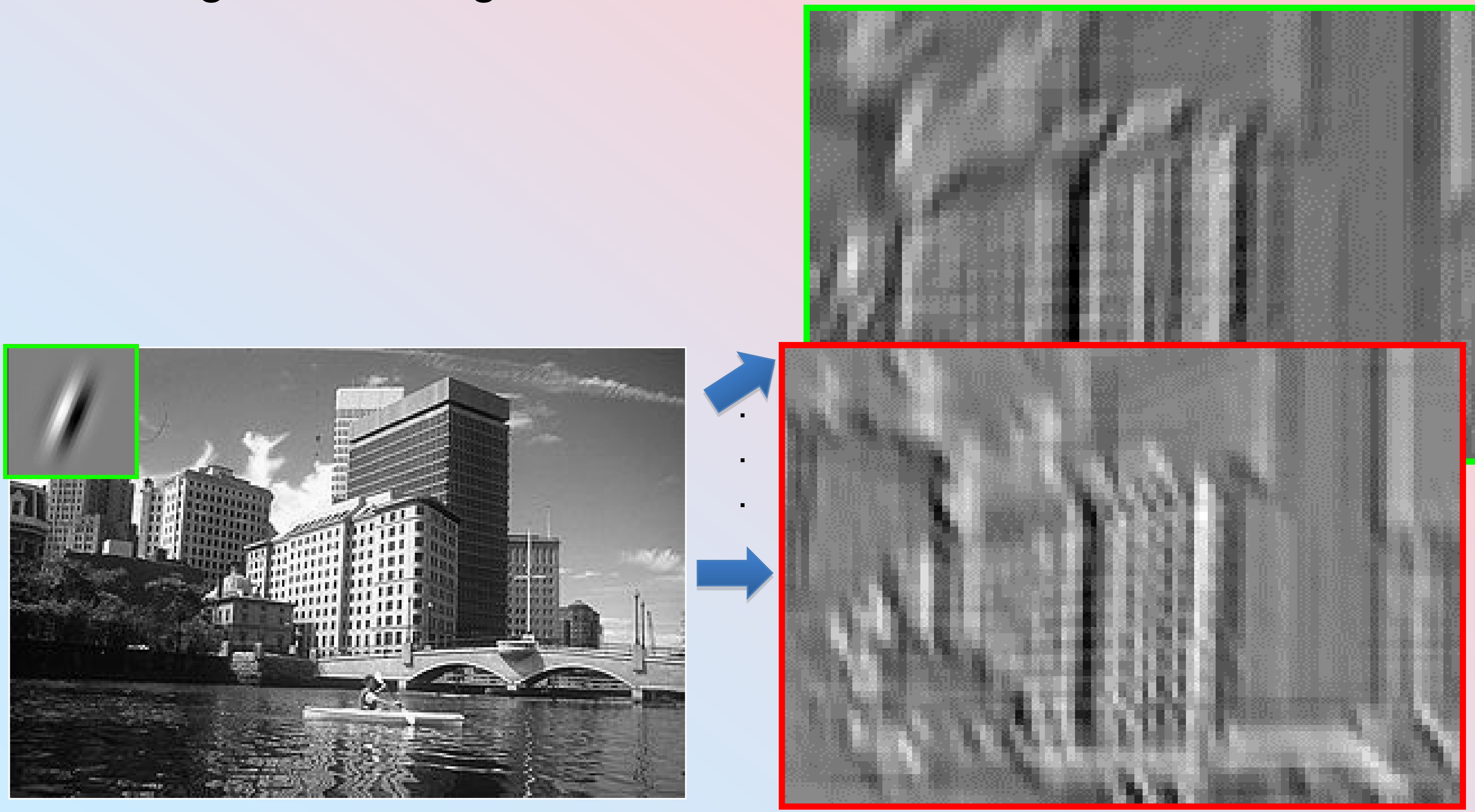
- https://www.rsipvision.com/ICCV2023-Thursday/

Yann LeCun

- VP and Chief AI Scientist, Facebook
- Silver Professor of Computer Science, Data Science, Neural Science, and Electrical and Computer Engineering, New York University.
- ACM Turing Award Laureate,
- Member, National Academy of Engineering

Another example of 2D Convolution

- Weighted moving sum

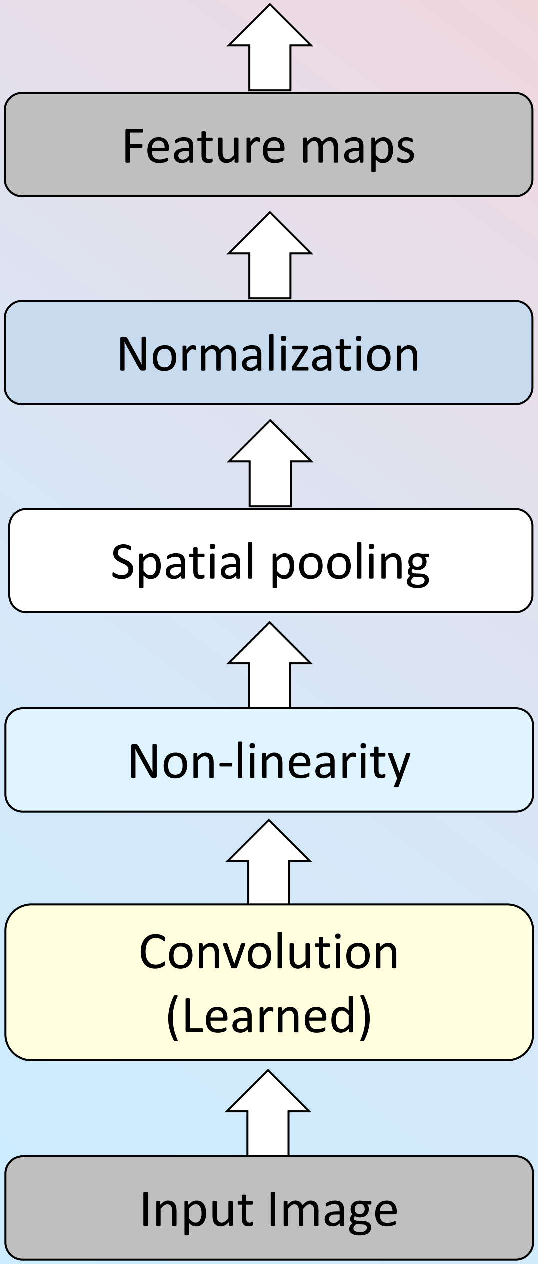


Input

Feature Activation Map

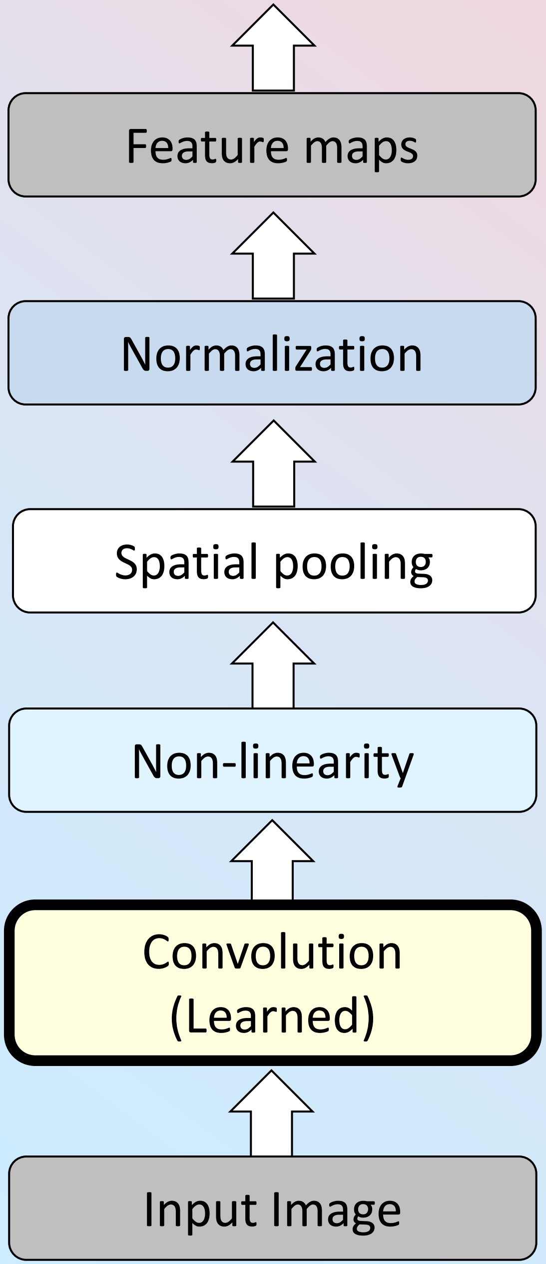
slide credit: S. Lazebnik

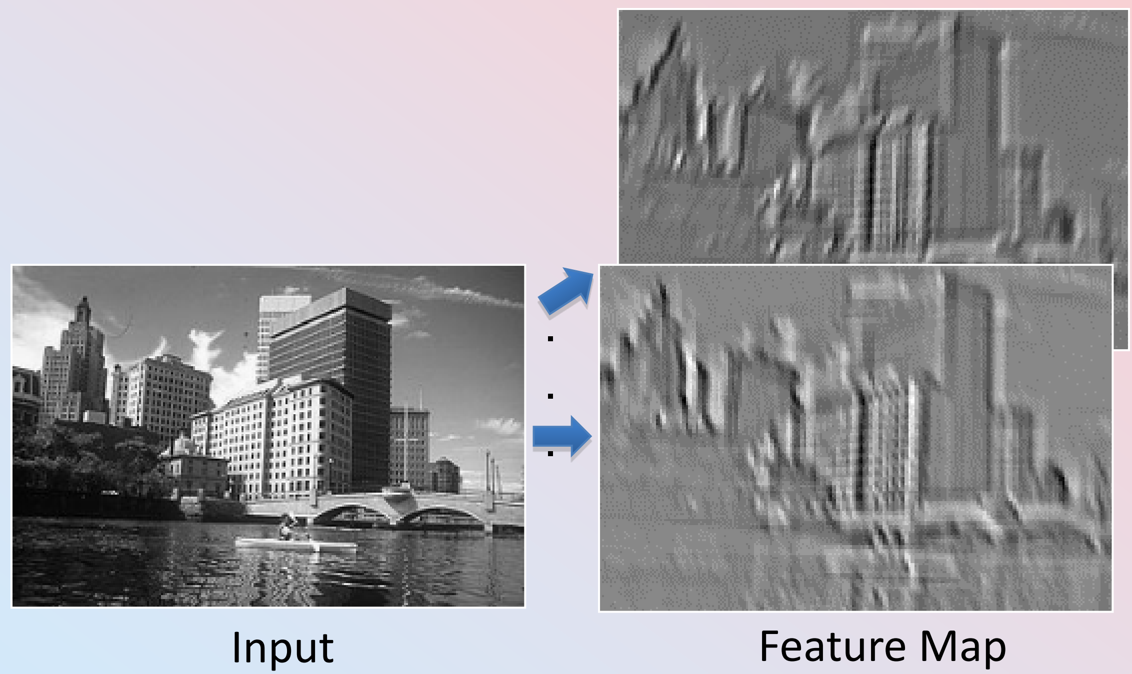
Convolutional Neural Networks



slide credit: S. Lazebnik

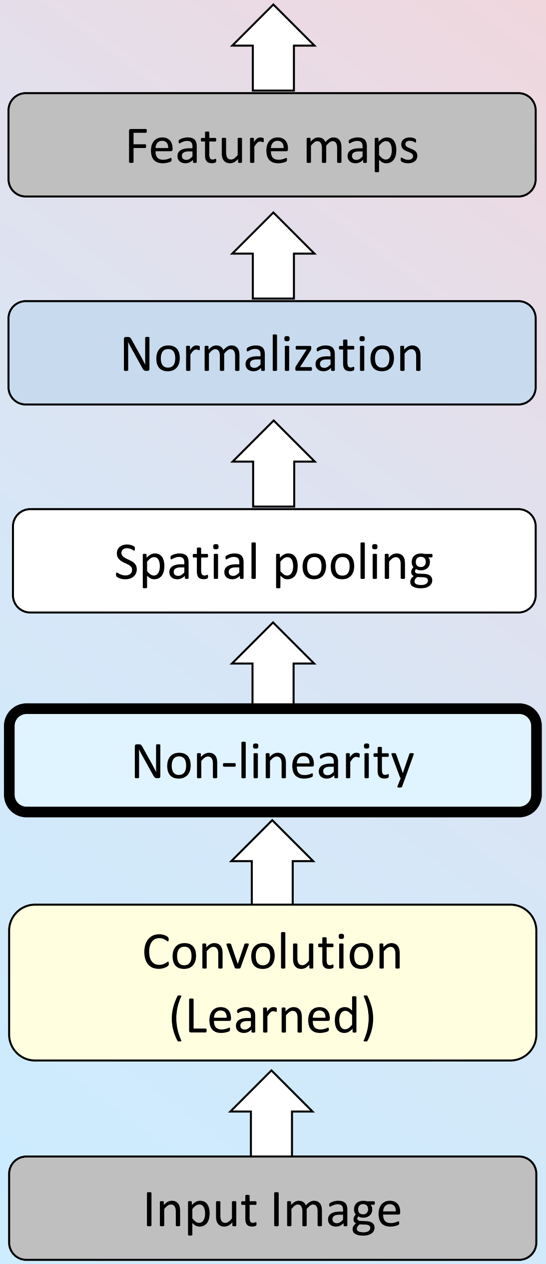
Convolutional Neural Networks



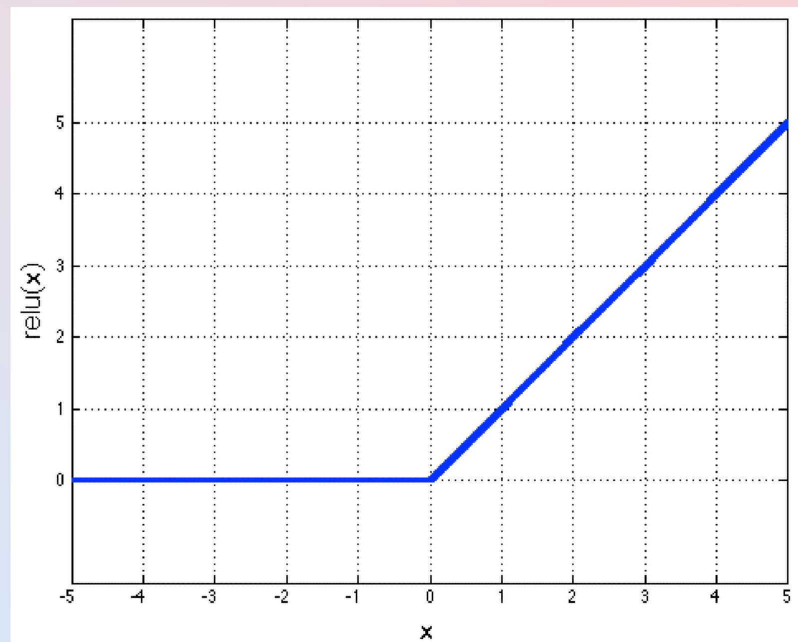


slide credit: S. Lazebnik

Convolutional Neural Networks

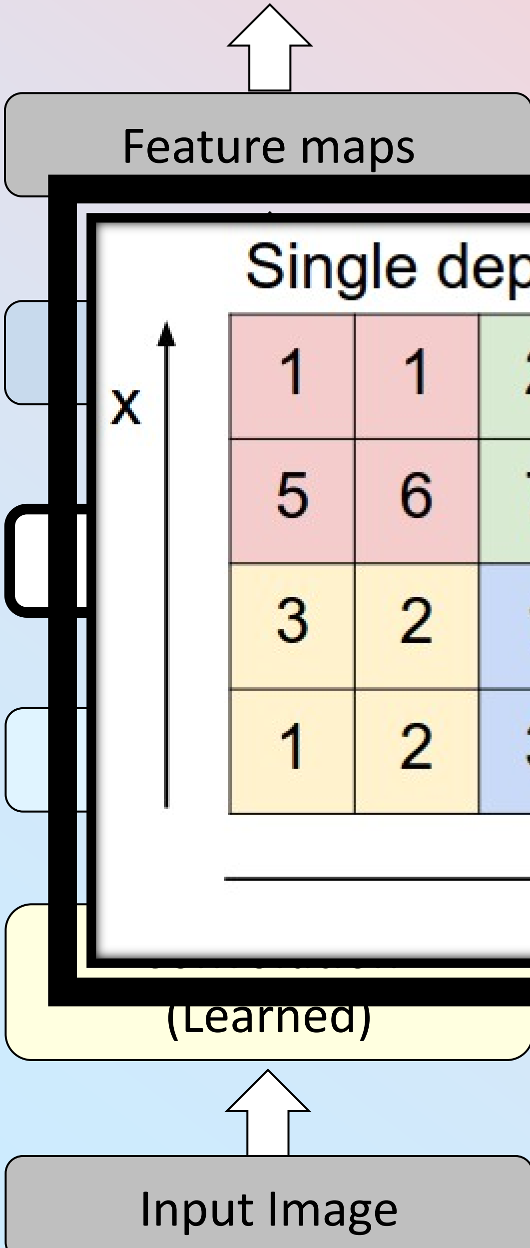


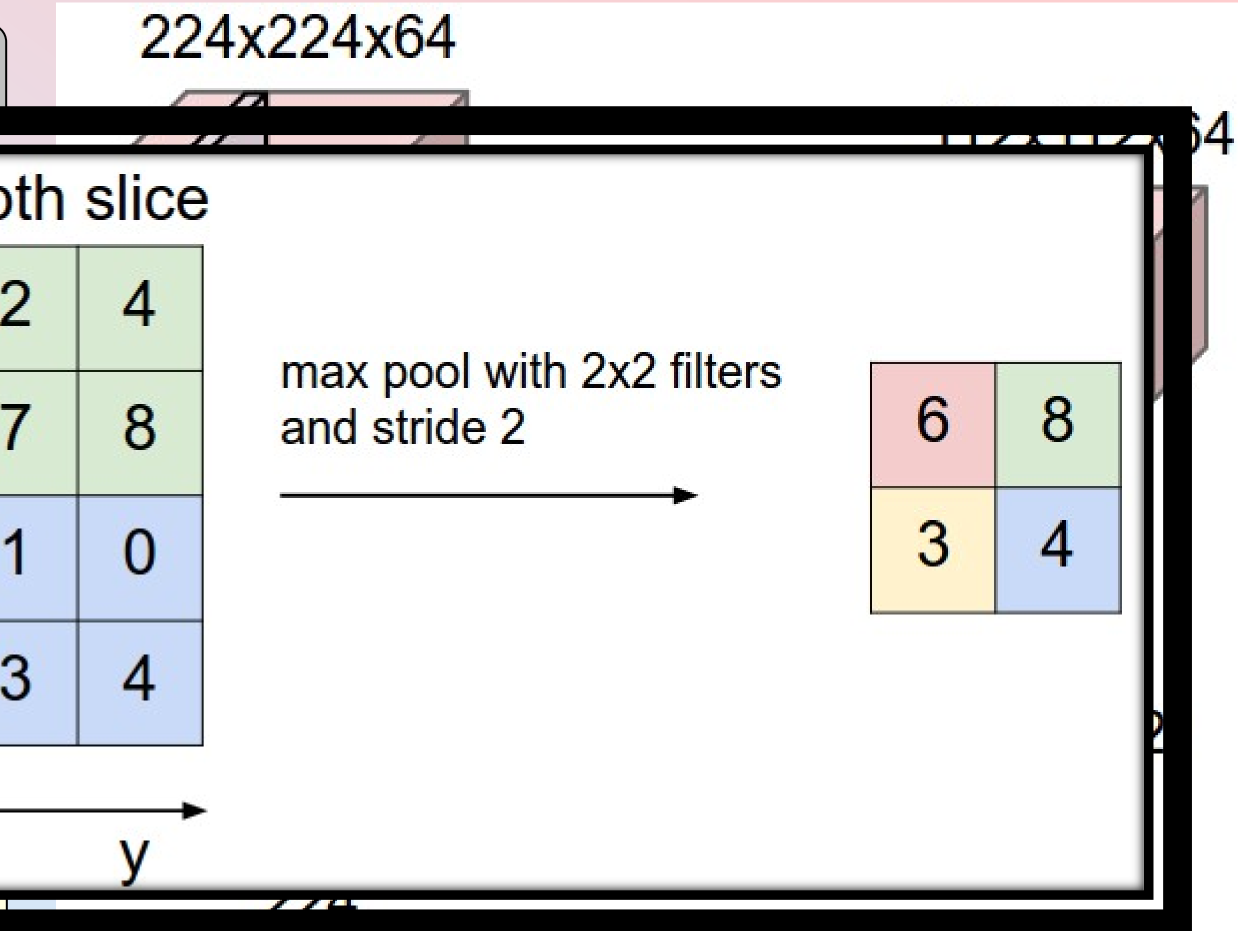
Rectified Linear Unit (ReLU)



slide credit: S. Lazebnik

Convolutional Neural Networks

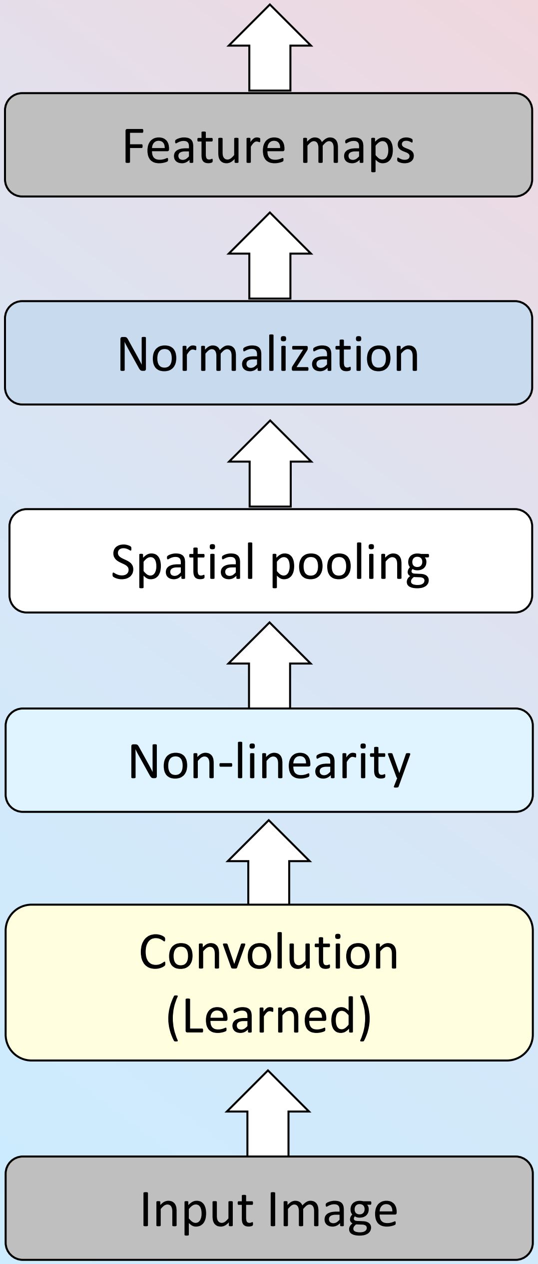




Provide *translation invariance*

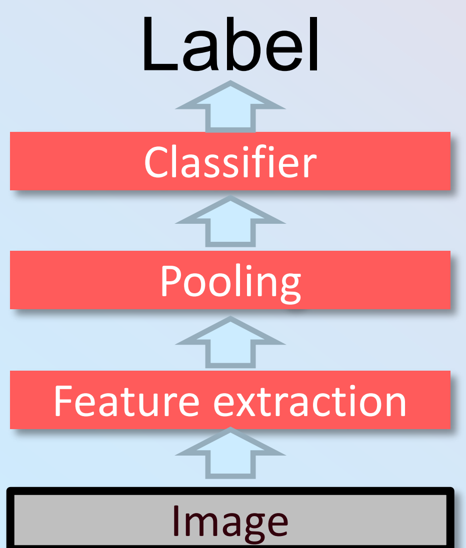
slide credit: S. Lazebnik

Convolutional Neural Networks

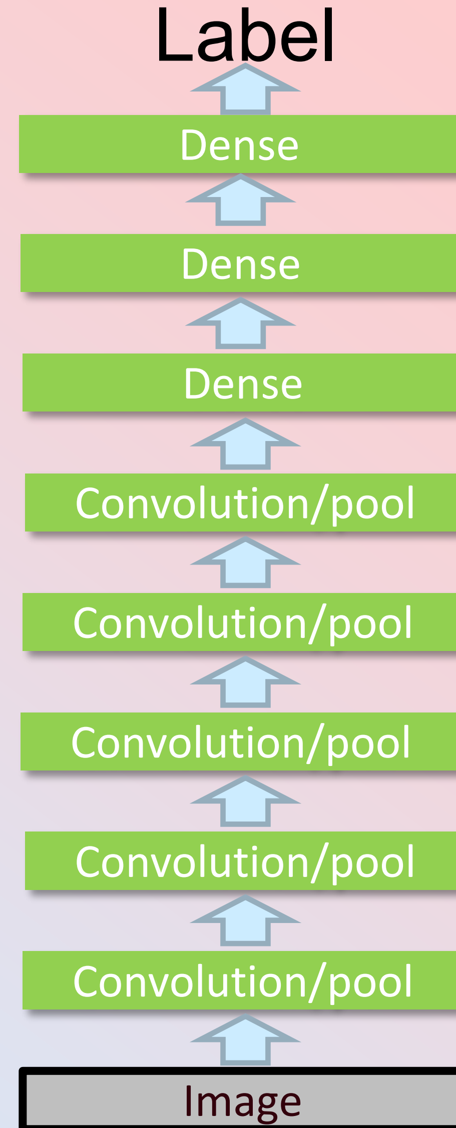


slide credit: S. Lazebnik

Traditional versus NN-based Computer Vision: Engineered versus Learned Features



Convolutional filters are trained in a supervised manner by back-propagating classification error



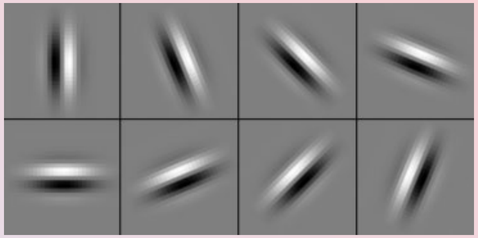
Jia-Bin Huang and Derek Hoiem, UIUC

SIFT Descriptor

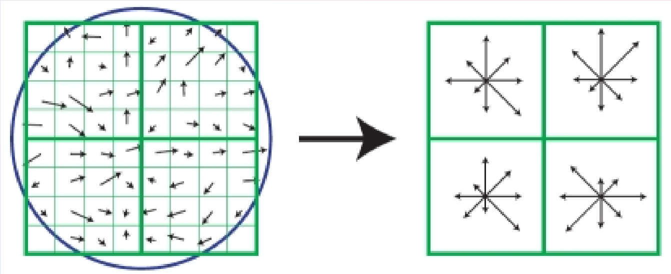
Lowe [IJCV 2004]

Image Pixels

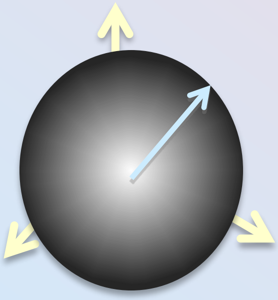
Apply oriented filters



Spatial pool (Sum)



Normalize to unit length

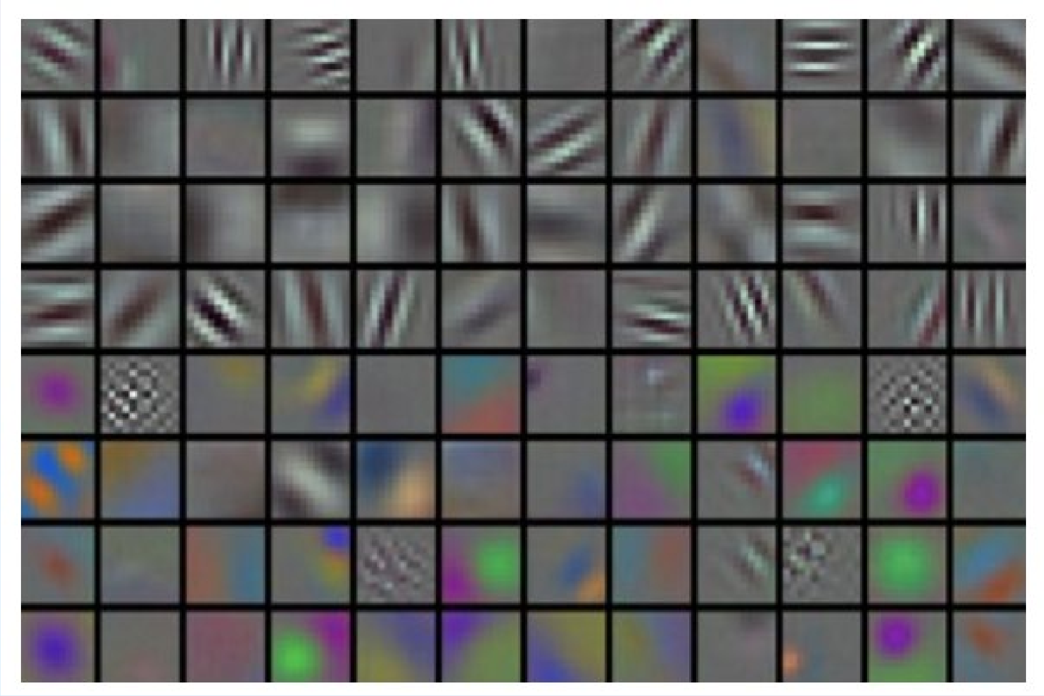


Feature Vector

slide credit: R. Fergus

Visualizing what was learned

- What do the learned filters look like?



Typical first layer filters

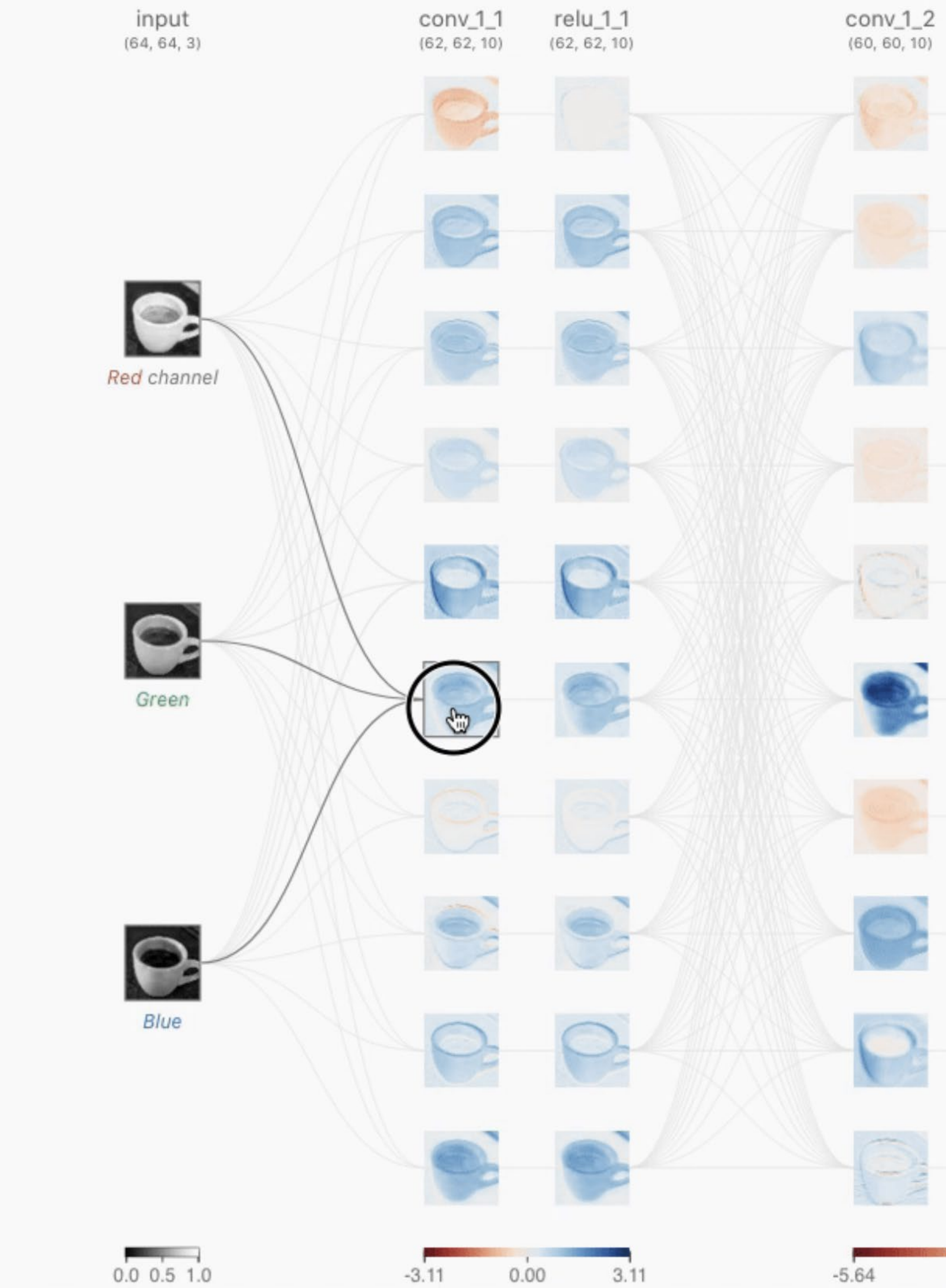
Image Credit: Kristen Grauman

The CNN Explainer

Thanks to CS640 classmate Mao Mao, we have a link to the *CNN Explainer*:

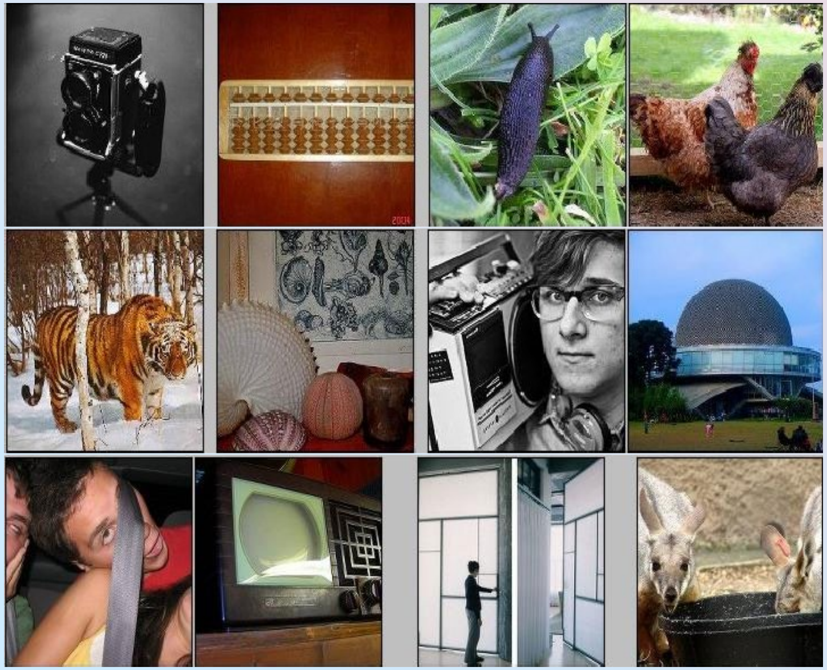
https://poloclub.github.io/cnn-explainer/

by Jay Wang, Robert Turko, Omar Shaikh, Haekyu Park, Nilaksh Das, Fred Hohman, Minsuk Kahng, and Polo Chau, a result of a research collaboration between Georgia Tech and Oregon State University



ImageNet –
The Data Set that Mattered and Still Matters!

[Deng et al. CVPR 2009]



- 14 million labeled images
- 20 thousand object classes

- Images collected from the Internet

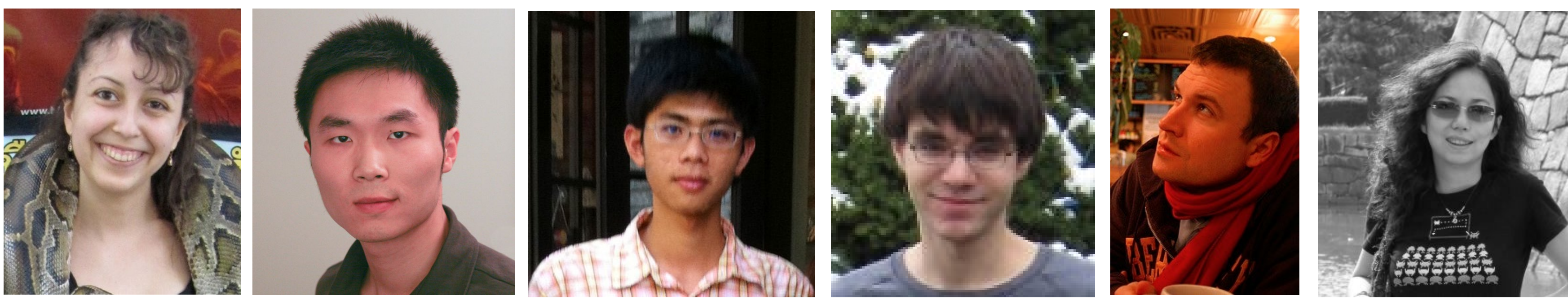
- Human labels obtained by crowdsourcing with Amazon Turk

- Still very important in 2023 because it is used for pretraining of “backbone neural nets”

Analysis of Large Scale Visual Recognition

Adapted for BU CS 440/640 by M. Betke

Fei-Fei Li and Olga Russakovsky



Olga Russakovsky, Jia Deng, Zhiheng Huang, Alex Berg, Li Fei-Fei
Detecting avocados to zucchinis: what have we done, and where are we going?
ICCV 2013 **http://image-net.org/challenges/LSVRC/2012/analysis**

Backpack

Flute

Strawberry

Traffic light

Backpack

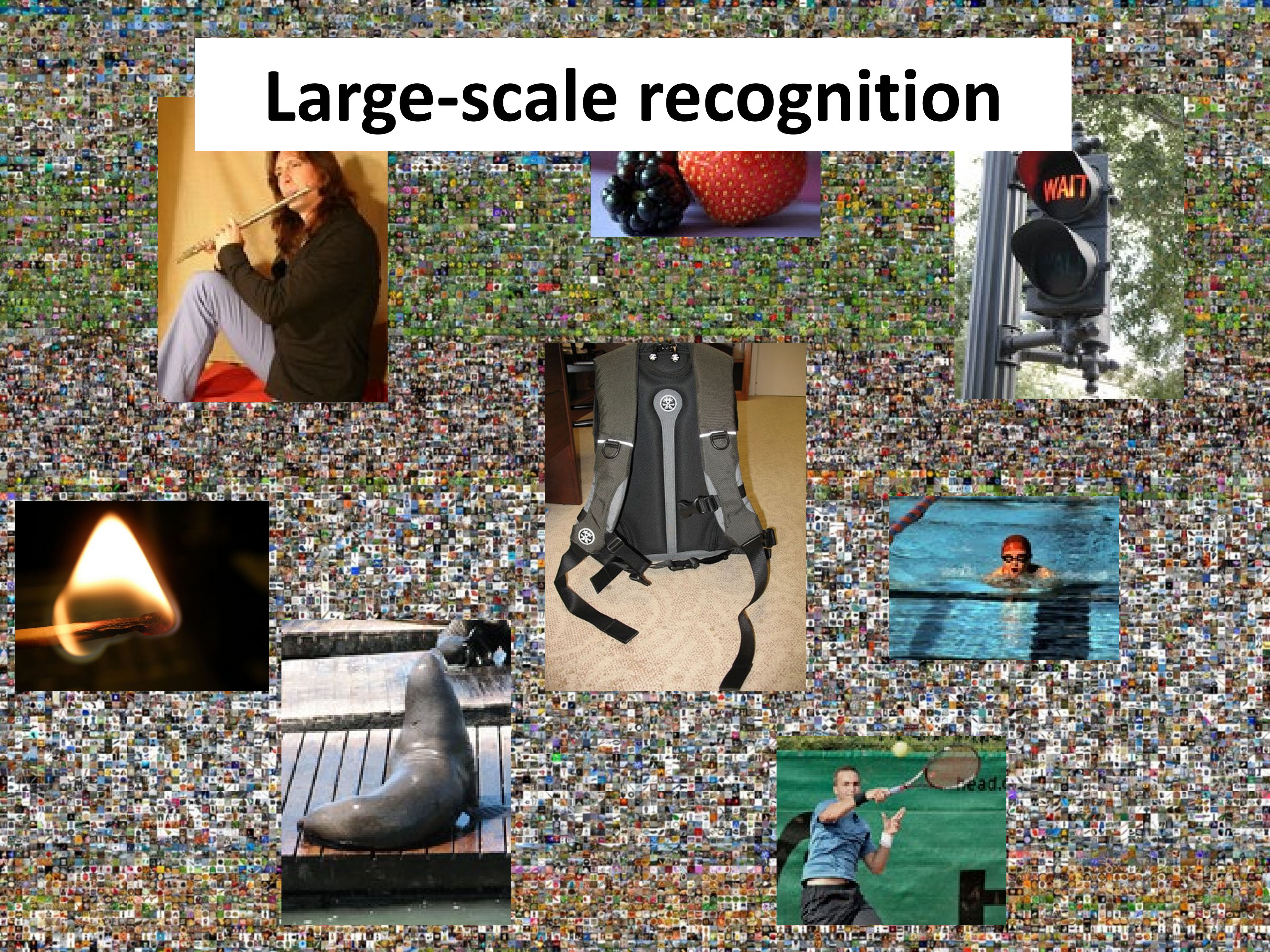
Matchstick

Bathing cap

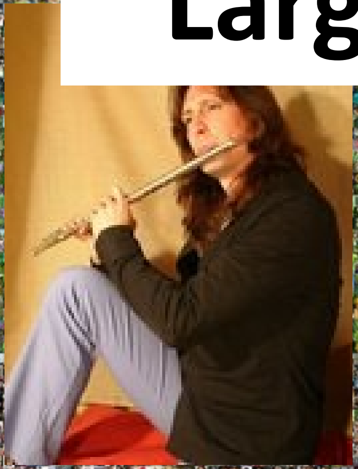
Sea lion

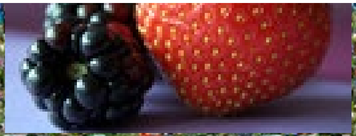
Racket

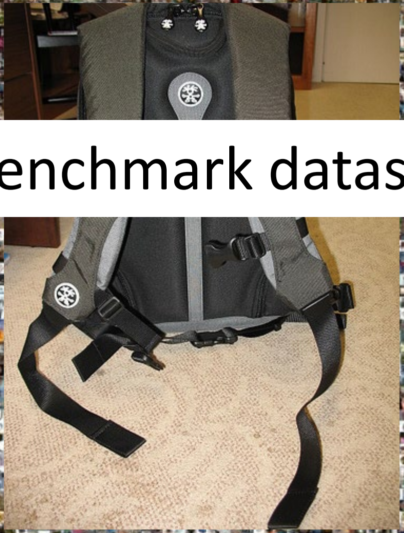
Large-scale recognition



Large-scale recognition





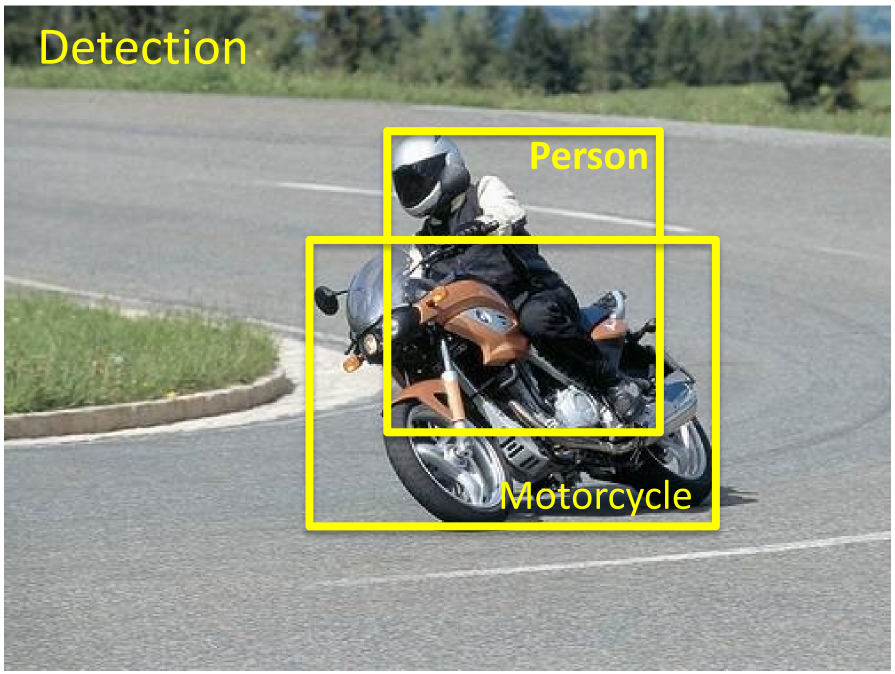


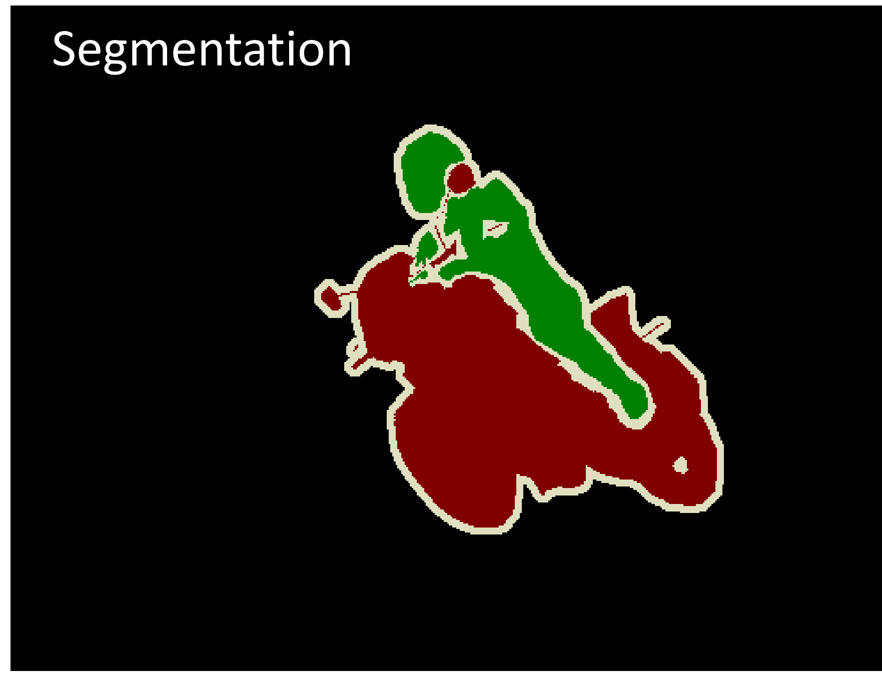
Need benchmark datasets

PASCAL VOC 2005-2012

20 object classes **22,591 images**

Classification: person, motorcycle





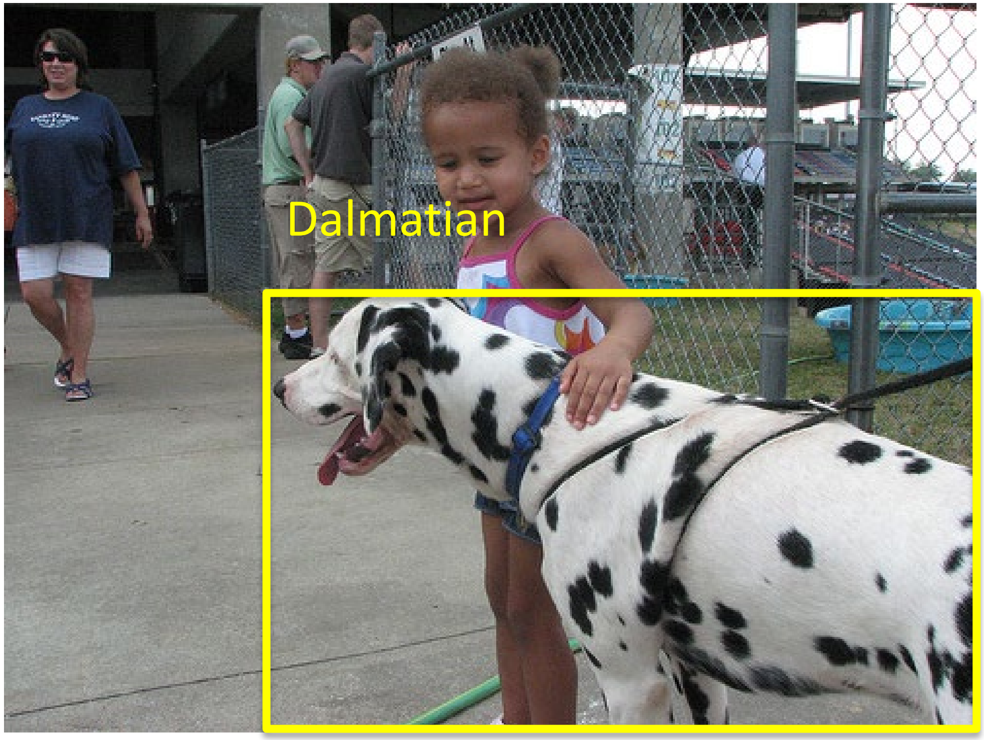
Action: riding bicycle

Everingham, Van Gool, Williams, Winn and Zisserman.
The PASCAL Visual Object Classes (VOC) Challenge. IJCV 2010.

IMAGENET Large Scale Visual Recognition Challenge (ILSVRC) 2010-2012

~~20 object classes 22,591 images~~

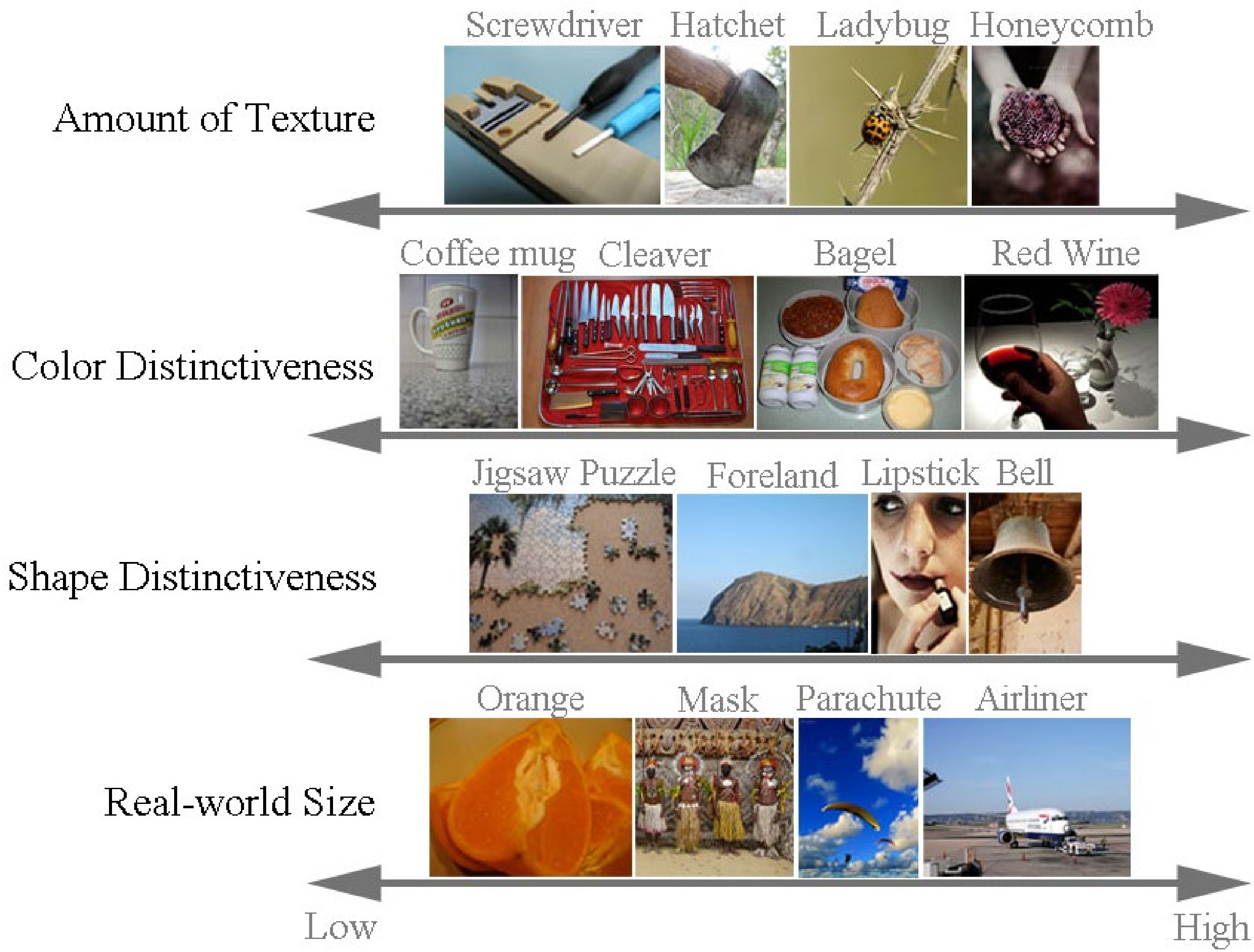
1000 object classes 1,431,167 images



http://image-net.org/challenges/LSVRC/{2010,2011,2012}

Variety of object classes in ILSVRC

Variety of object classes in ILSVRC



ILSVRC Task 1: Classification

Steel drum

ILSVRC Task 1: Classification

Allowed system output: 5 predictions per image

Goal: Get 1 of the 5 predictions correct

Steel drum

Output:		**Output:**	
Scale		Scale	
T-shirt		T-shirt	
<u>Steel drum</u>	✓	Giant panda	✗
Drumstick		Drumstick	
Mud turtle		Mud turtle	

Indicator Function:

1[System output correct on this image] = 1 = 0

ILSVRC Task 1: Classification

Steel drum

Output:
Scale
T-shirt
<u>Steel drum</u>
Drumstick
Mud turtle

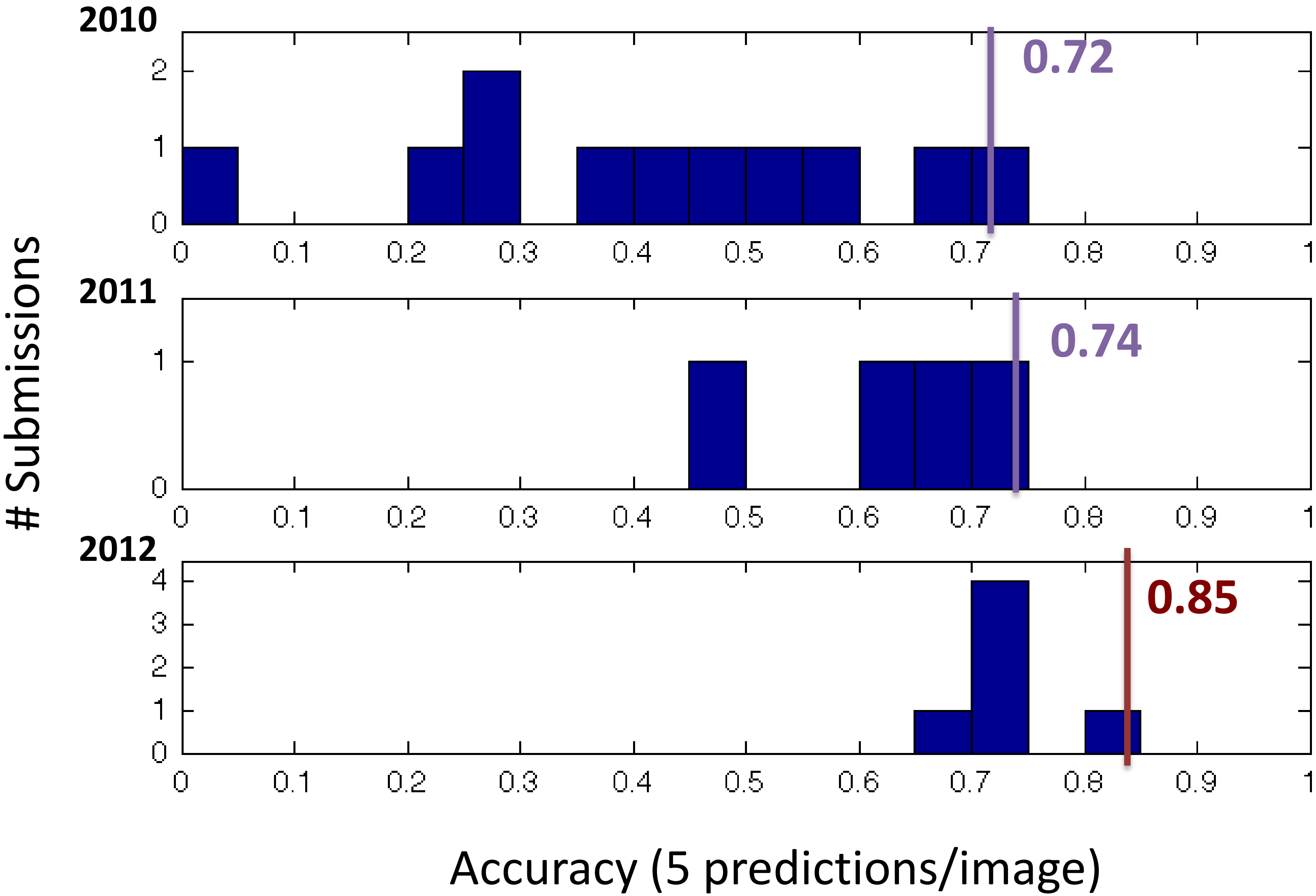
✓

Output:
Scale
T-shirt
Giant panda
Drumstick
Mud turtle

✗

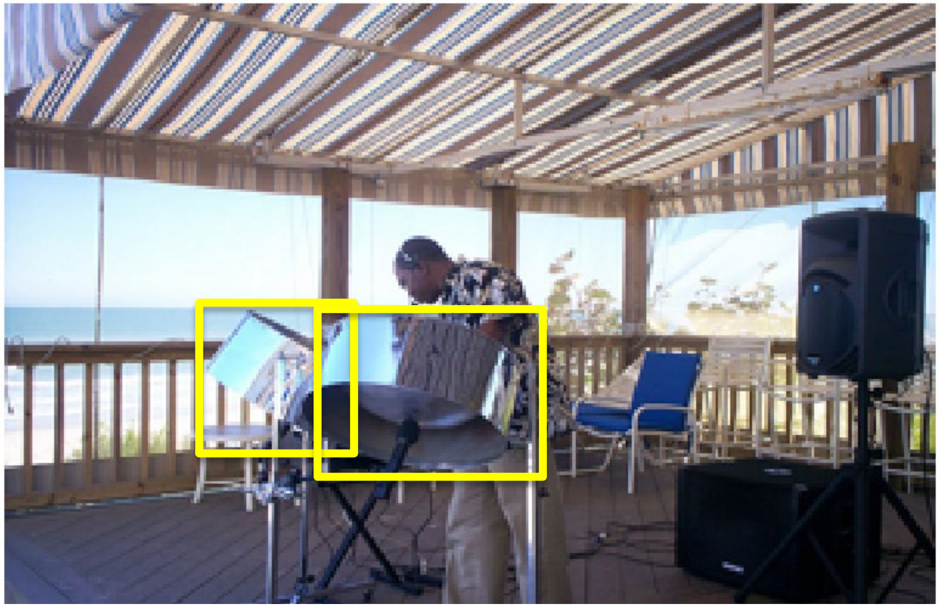
$$\text{Accuracy} = \frac{1}{100{,}000} \sum_{\substack{100{,}000 \\ \text{images}}} 1[\text{correct on image } i]$$

ILSVRC Task 1: Classification



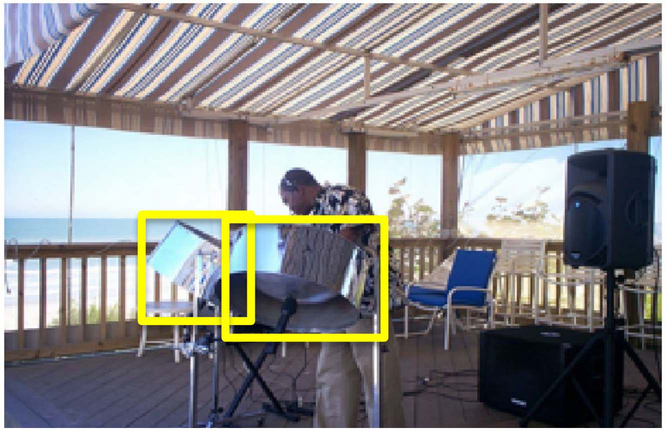
ILSVRC Task 2: Classification + Localization

Steel drum

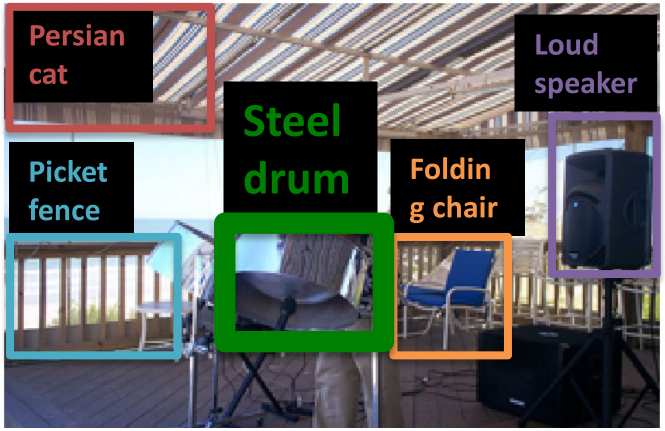


ILSVRC Task 2: Classification + Localization

Steel drum

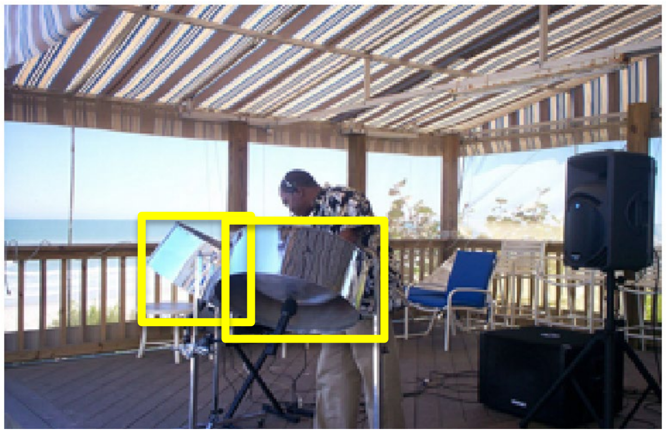


Output

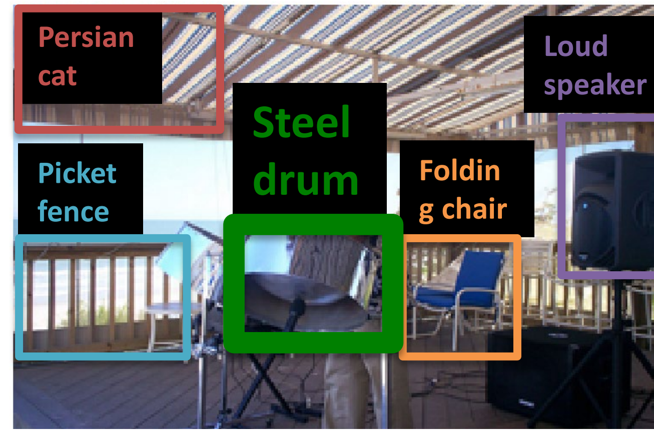


ILSVRC Task 2: Classification + Localization

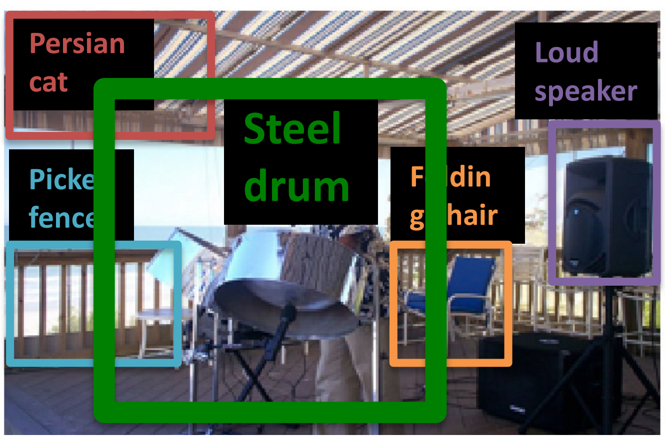
Steel drum



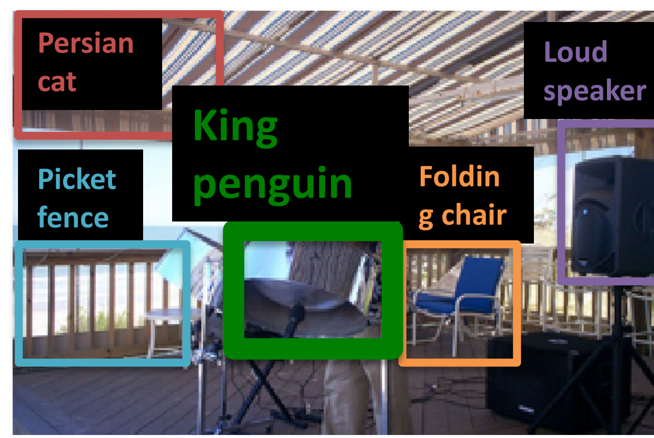
Output



Output (bad localization)

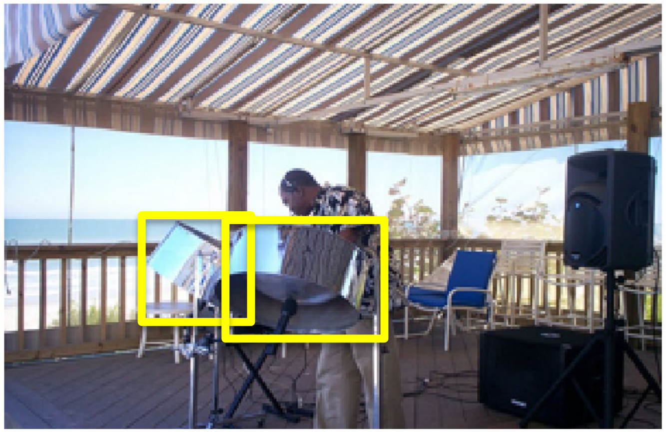


Output (bad classification)

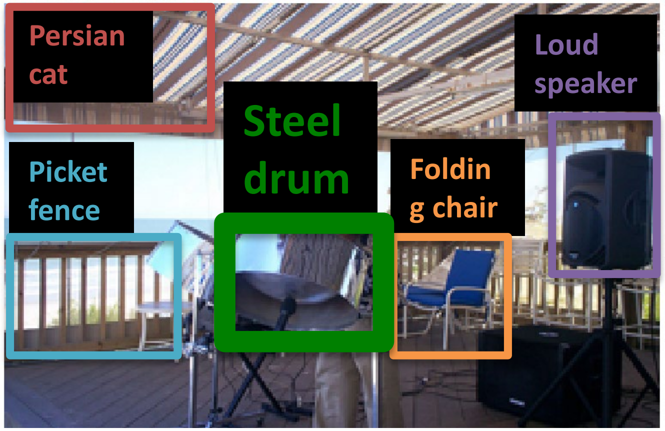


ILSVRC Task 2: Classification + Localization

Steel drum

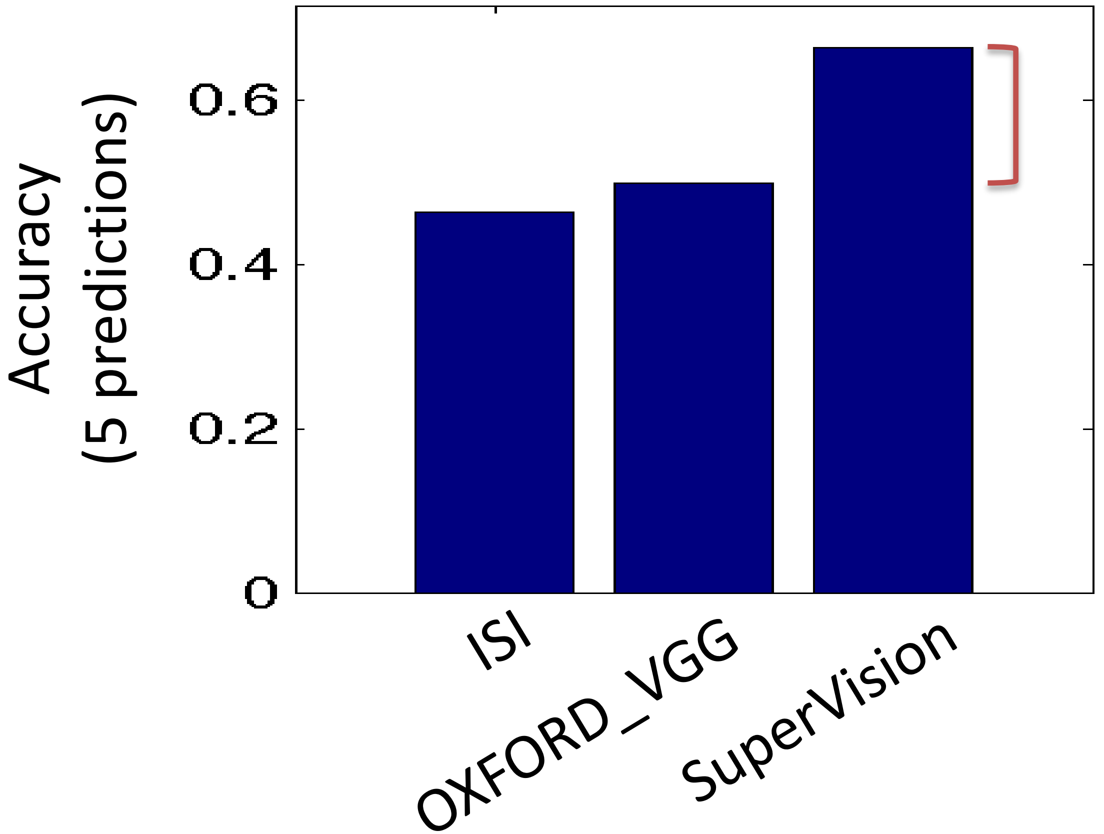


Output



$$\text{Accuracy} = \frac{1}{100{,}000} \sum_{\substack{100{,}000 \\ \text{images}}} 1[\text{correct on image i}]$$

ILSVRC Task 2: Classification + Localization



ISI=Uni. Tokyo Team

VGG=Uni. Oxford Team

SuperVision =
University of Toronto Team
Led by
Geoffrey Hinton,
Turing Award Winner

What happens under the hood?

Preliminaries:

- ILSVRC-500 (2012) dataset
- Leading algorithms

What happens under the hood on classification+localization?

- A closer look at small objects
- A closer look at textured objects

Olga Russakovsky, Jia Deng, Zhiheng Huang, Alex Berg, Li Fei-Fei
Detecting avocados to zucchinis: what have we done, and where are we going?
ICCV 2013 http://image-net.org/challenges/LSVRC/2012/analysis

ILSVRC (2012)

1000 object classes

Easy to localize

Hard to localize

ILSVRC-500 (2012)

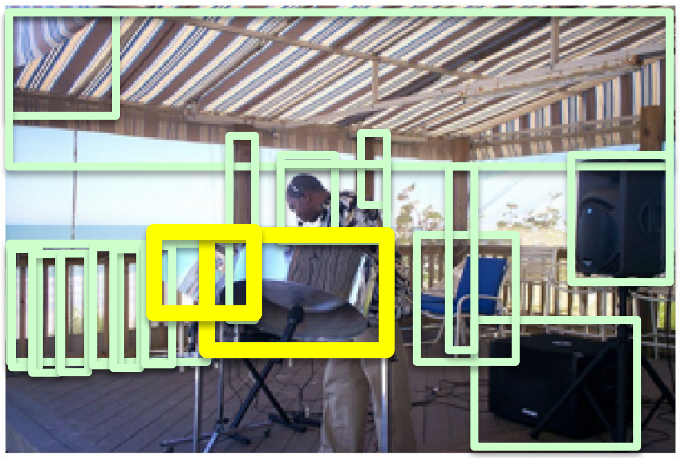
ILSVRC-500 (2012)

Object scale (fraction of image area occupied by target object)

ILSVRC-500 (2012)	500 object categories	25.3%
PASCAL VOC (2012)	20 object categories	25.2%

Level of clutter

Steel drum



- Generate candidate object regions using method of

  Selective Search for Object Detection
  vanDeSande et al. ICCV 2011
- Filter out regions inside object
- Count regions

ILSVRC-500 (2012)	500 object categories	128 ± 35
PASCAL VOC (2012)	20 object categories	130 ± 29

SuperVision = AlexNet

Alex Krizhevsky, Ilya Sutskever, Geoffrey Hinton (Krizhevsky NIPS12)

Image classification: Deep convolutional neural networks

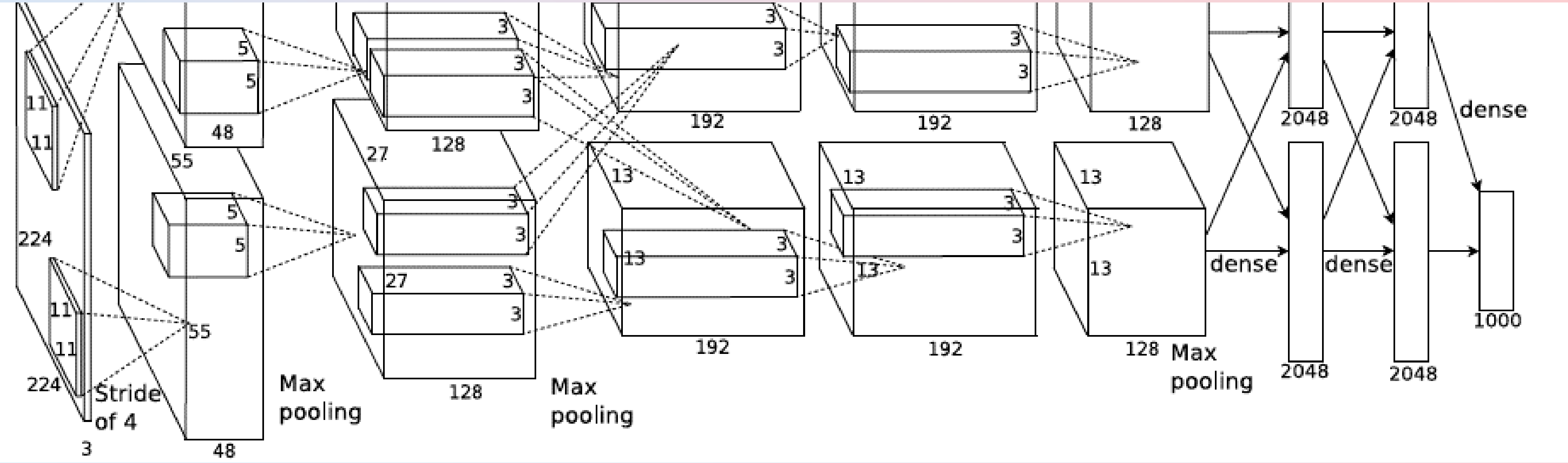
- 7 hidden “weight” layers, 650K neurons, 60M parameters, 630M connections
- Rectified Linear Units, max pooling, dropout trick
- Randomly extracted 224x224 patches for more data
- Trained with Stochastic Gradient Descent on two GPUs for a week, fully supervised (50x speed-up over CPU)

Localization: Regression on (x,y,w,h)

http://www.cs.toronto.edu/~fritz/absps/imagenet.pdf

AlexNet

- Similar to the model proposed by LeCun in 1998 but:
  - Larger model (7 hidden layers, 650,000 units, 60,000,000 params)
  - More data (10^6 vs. 10^3 images)



A. Krizhevsky, I. Sutskever, and G. Hinton,
ImageNet Classification with Deep Convolutional Neural Networks, NIPS 2012

Details of the Oxford VGG

This is **not** the neural net VGG but uses traditional computer vision techniques!

Karen Simonyan, Yusuf Aytar, Andrea Vedaldi, Andrew Zisserman

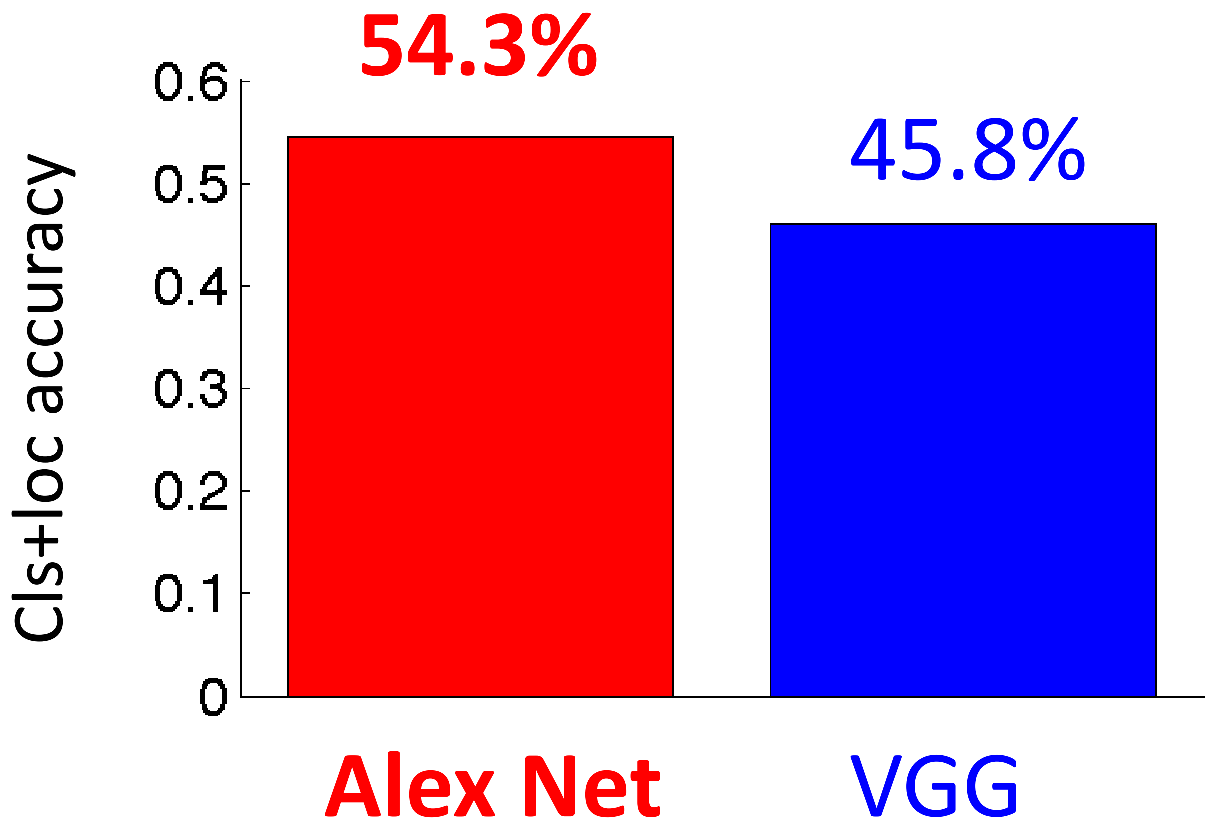
Image classification: Fisher vector + linear SVM (Sanchez CVPR11)

- Root-SIFT (Arandjelovic CVPR12), color statistics, augmentation with patch location (x,y) (Sanchez PRL12)
- Fisher vectors: 1024 Gaussians, 135K dimensions
- No SPM, product quantization to compress
- Semi-supervised learning to find additional bounding boxes
- 1000 one-vs-rest SVM trained with Pegasos SGD
  - 135M parameters!

Localization: Deformable part-based models (Felzenszwalb PAMI10), without parts (root-only)

http://image-net.org/challenges/LSVRC/2012/oxford_vgg.pdf

Results on ILSVRC-500



Preliminaries:

- ILSVRC-500 (2012) dataset – similar to PASCAL
- Leading algorithms: Alex Net and VGG

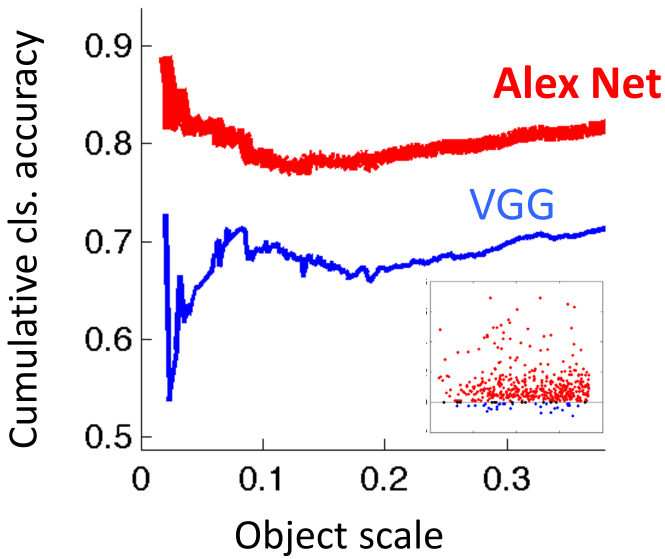
What happens under the hood on classification+localization?

- Alex Net always great at classification, but VGG does better than Alex Net localizing small objects
- A closer look at textured objects

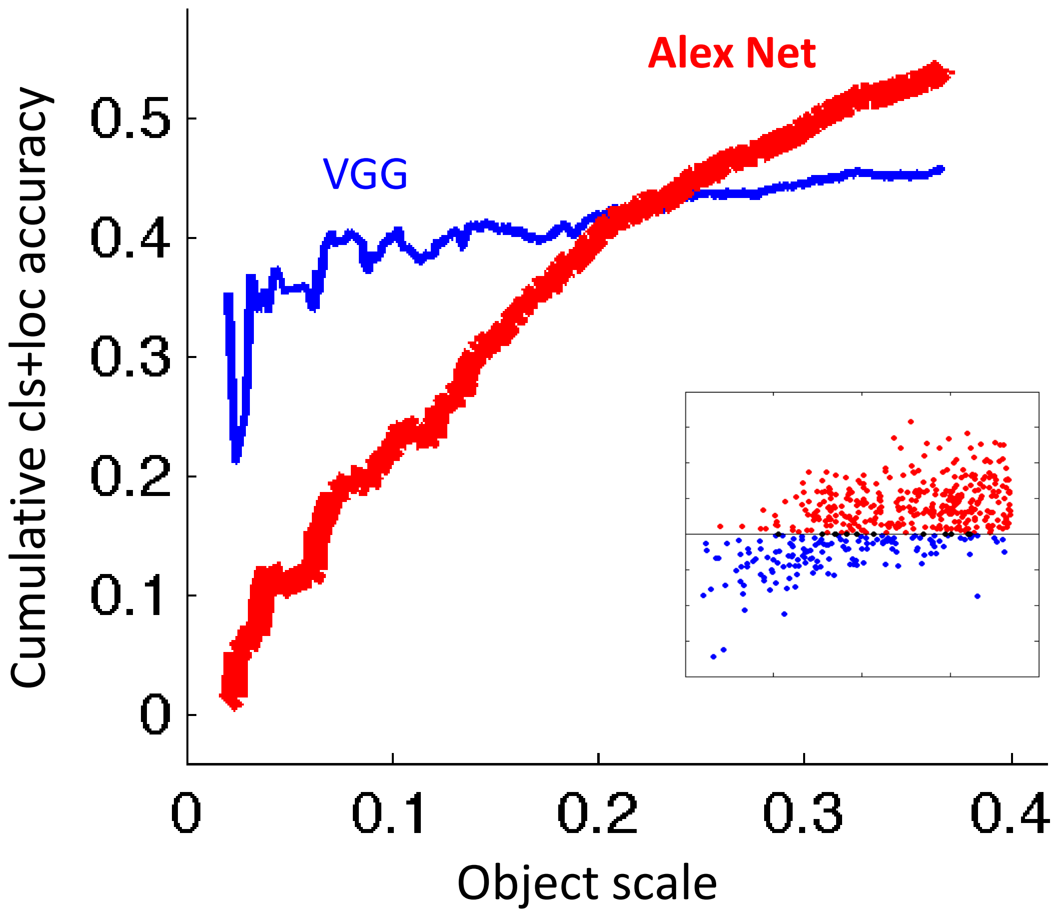
Olga Russakovsky, Jia Deng, Zhiheng Huang, Alex Berg, Li Fei-Fei
Detecting avocados to zucchinis: what have we done, and where are we going?
ICCV 2013 http://image-net.org/challenges/LSVRC/2012/analysis

Cumulative accuracy across scales

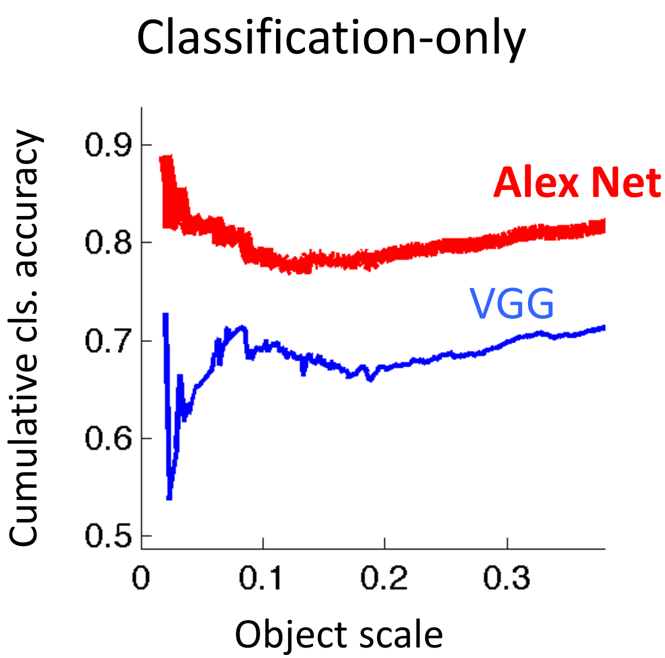
Classification-only

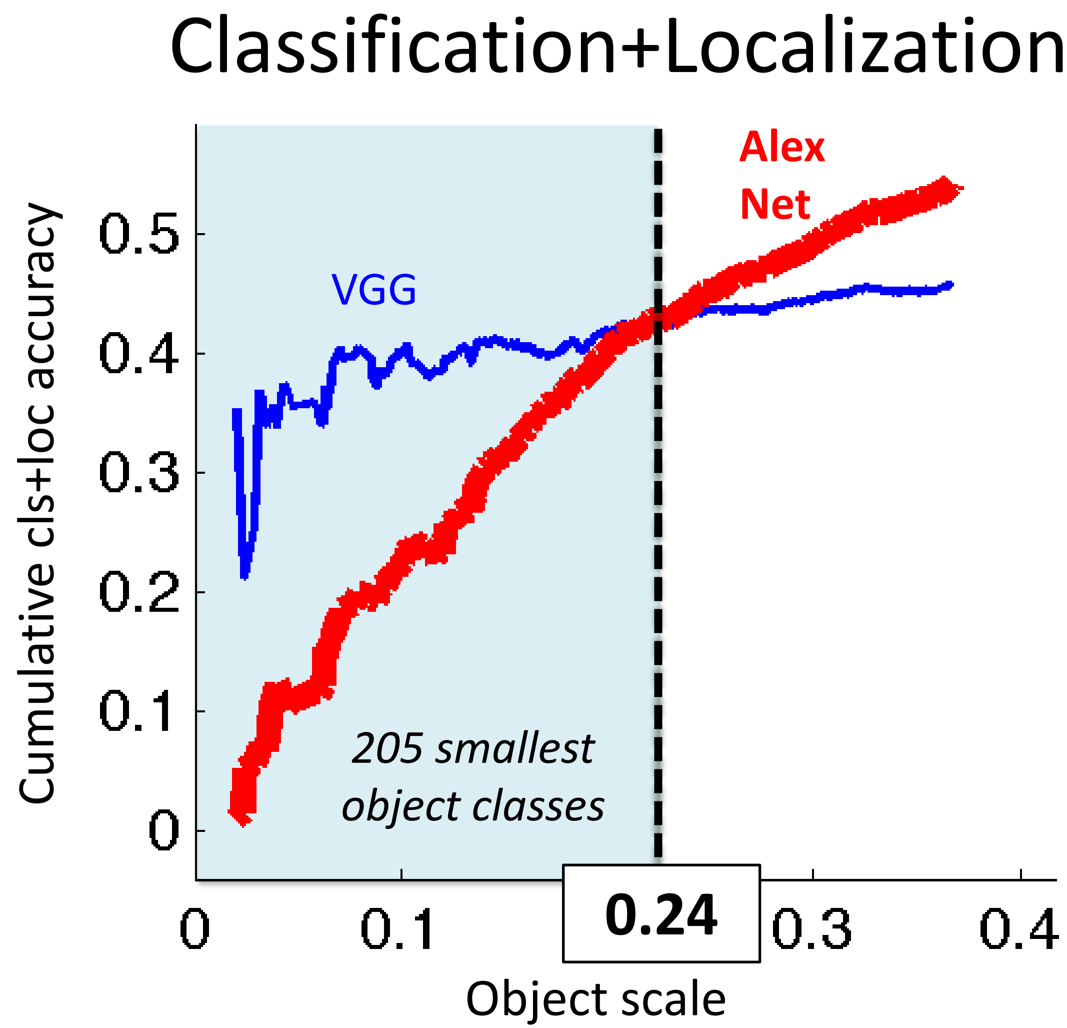


Classification+Localization



Cumulative accuracy across scales





Textured objects (ILSVRC-500)

Textured objects (ILSVRC-500)

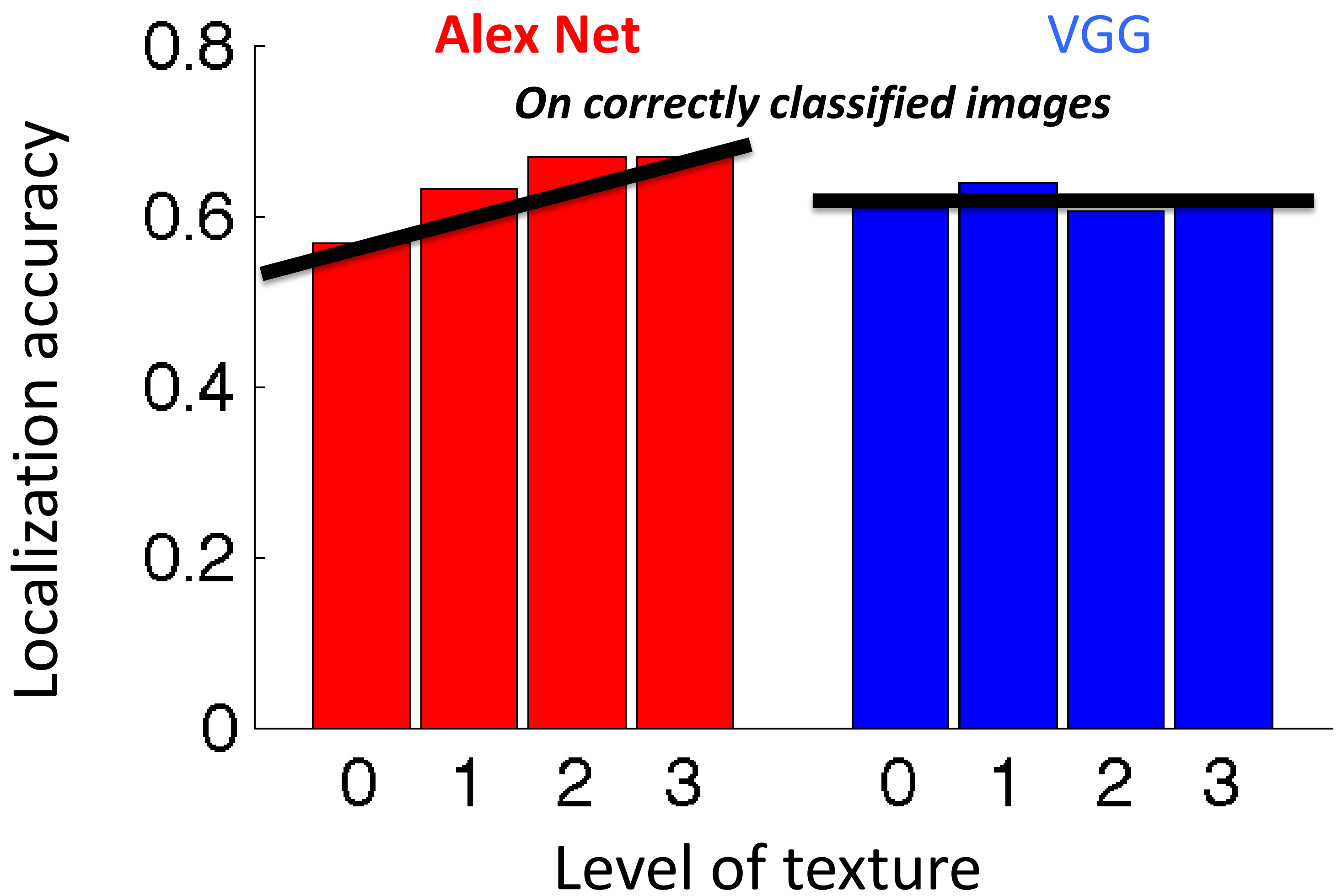
	No texture	Low texture	Medium texture	High texture
# classes	116	189	143	52
Object scale	20.8%	23.7%	23.5%	25.0%

Textured objects (416 classes)

	No texture	Low texture	Medium texture	High texture
# classes	116	~~189~~ 149	~~143~~ 115	~~52~~ 35
Object scale	20.8%	~~23.7%~~ 20.8%	~~23.5%~~ 20.8%	~~25.0%~~ 20.8%

Localizing textured objects

(416 classes, same average object scale at each level of texture)

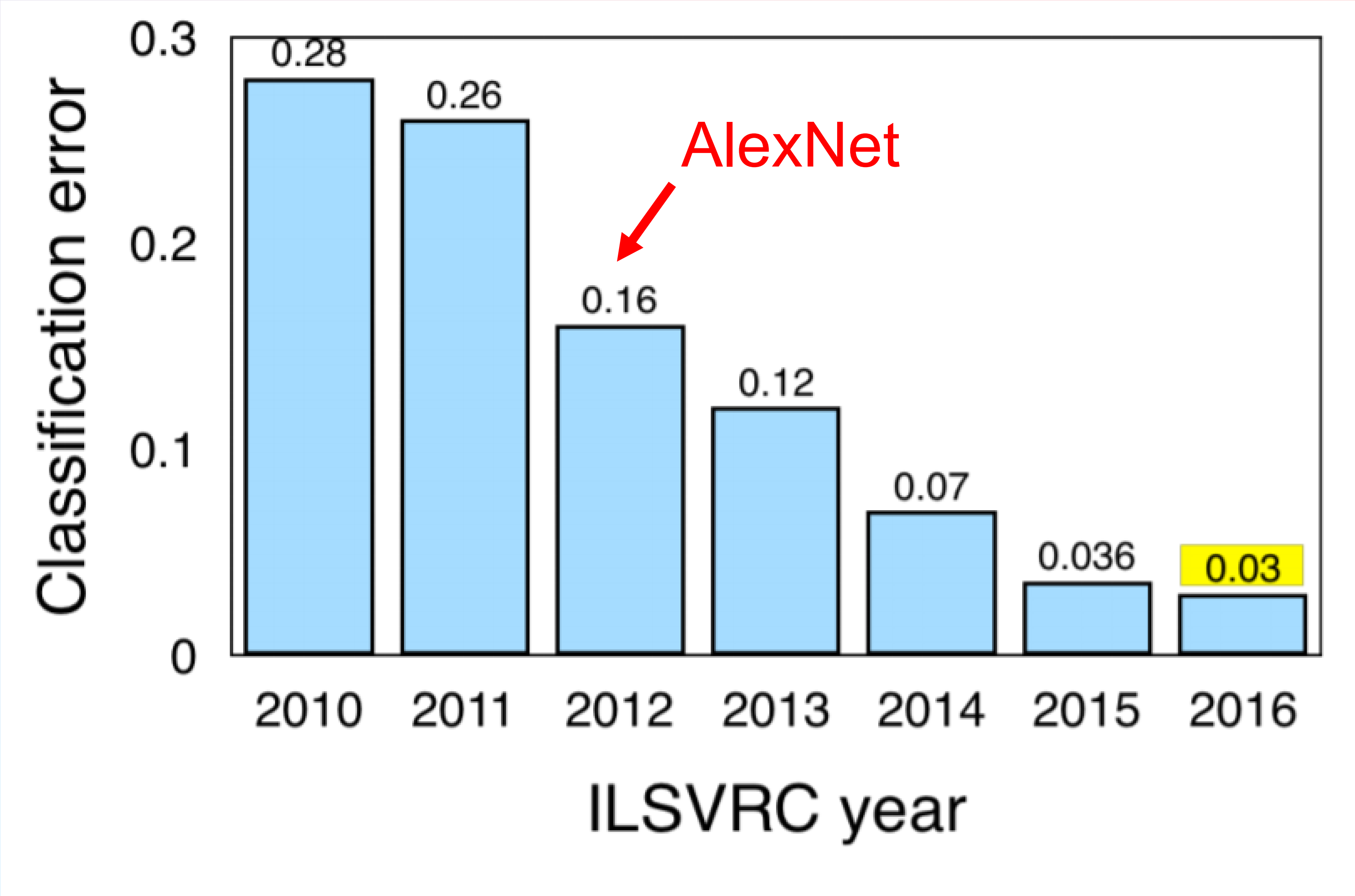


Conclusions on analysis of classification+localization results

- Alex Net always great at classification, but VGG does better than Alex Net localizing small objects
- Textured objects: VGG broadly successful. Alex Net better at higher textures, worse at smaller.

Olga Russakovsky, Jia Deng, Zhiheng Huang, Alex Berg, Li Fei-Fei
Detecting avocados to zucchinis: what have we done, and where are we going?
ICCV 2013 http://image-net.org/challenges/LSVRC/2012/analysis

ImageNet Classification Challenge



http://image-net.org/challenges/talks/2016/ILSVRC2016_10_09_clsloc.pdf

Recap of NN-based Computer Vision

- Neural networks
  - View of neural networks as learning hierarchy of features
- Convolutional neural networks
  - Architecture of network accounts for image structure
  - “End-to-end” recognition from pixels
  - Together with large labeled datasets and lots of computation → major success on benchmark ImageNet, i.e., object classification and localization

Learning Objectives for this Lecture

Computer Science

- Understand formats of images used as inputs to AI models: greyscale, color, medical scans
- Understand differences and similarities between pre-2012 “traditional computer vision” and post-2012 neural-network-based computer vision & see examples
- Understand why convolution is powerful
- Understand how tools from estimation theory can be used to measure recognizability of objects in images
- Learn about breakthrough dataset ImageNet
- Learn about early CNNs used in computer vision