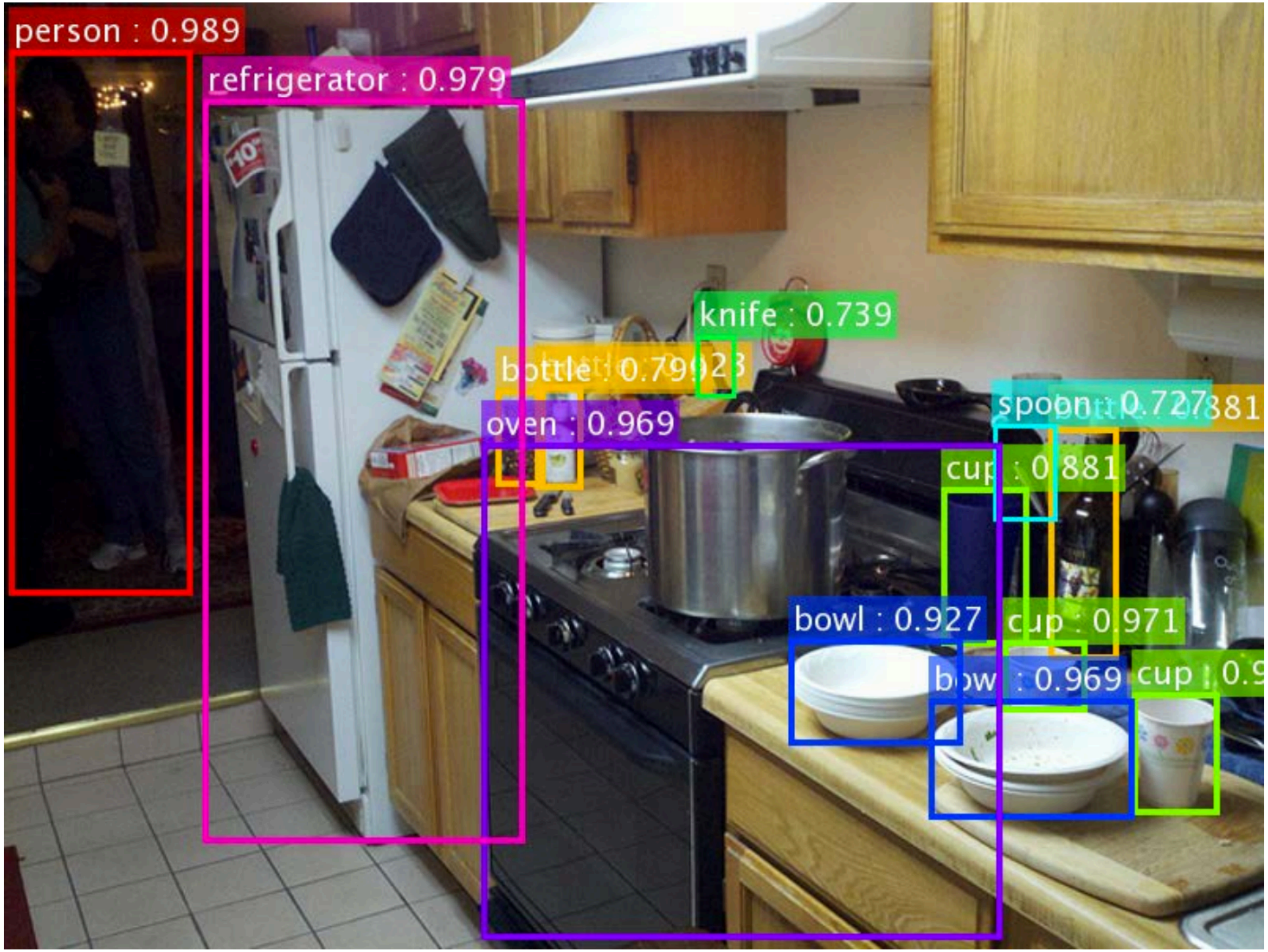
Introduction to Computer Vision:

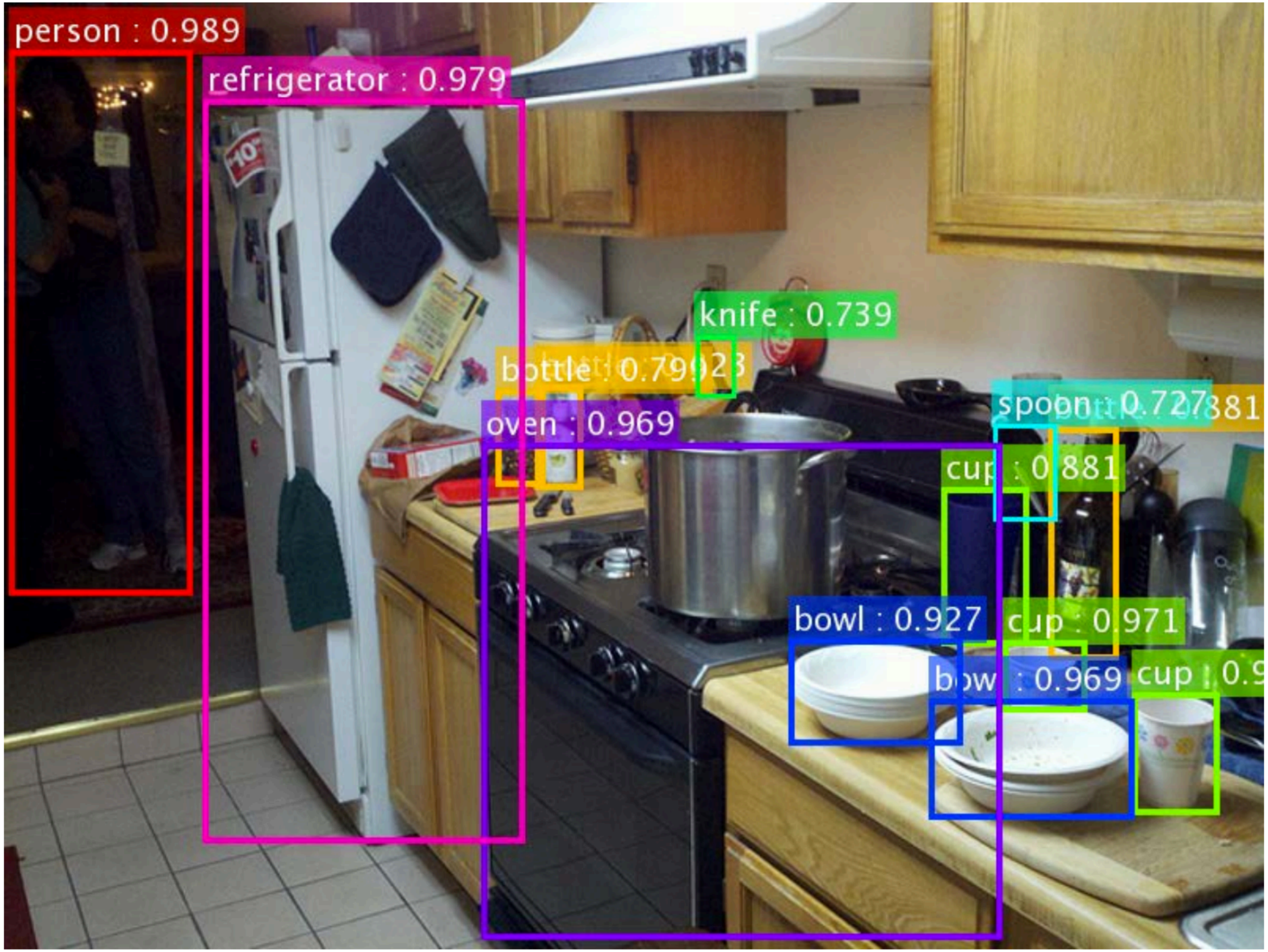
Object Recognition

Fereshteh Sadeghi

fsadeghi@cs.washington.edu

Lots of slides from Larry Zitnick and Alyosha Efros

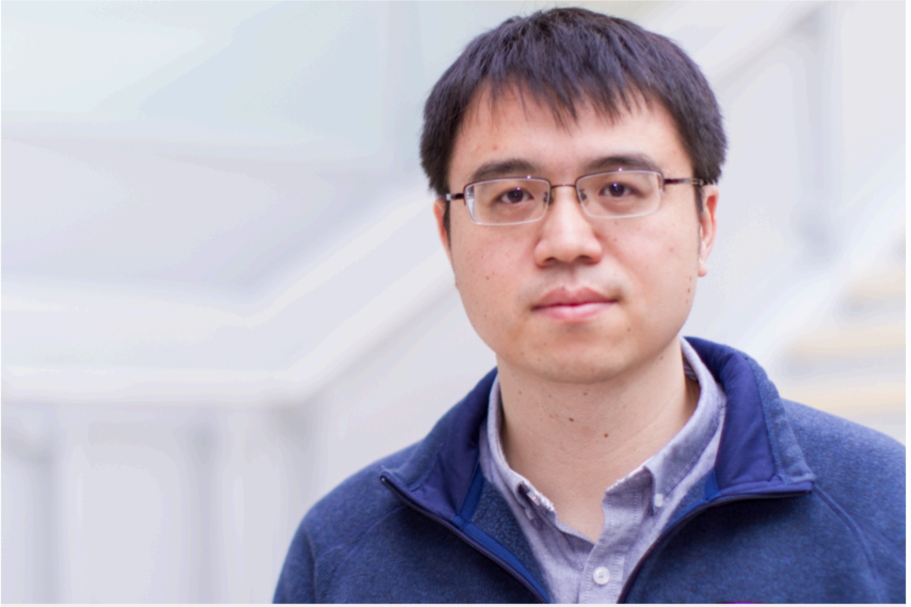

person : 0.989
refrigerator : 0.979
knife : 0.739
oven : 0.969
spoon : 0.727
cup : 0.881
bowl : 0.927
cup : 0.971
bowl : 0.969



*the original image is from the COCO dataset

Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. "Deep Residual Learning for Image Recognition". arXiv 2015.
Shaoqing Ren, Kaiming He, Ross Girshick, & Jian Sun. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks". NIPS 2015.

Microsoft researchers win ImageNet computer vision challenge



Jian Sun, a principal research manager at Microsoft Research, led the image understanding project. Photo: Craig Tuschhoff/Microsoft.

Posted December 10, 2015 By **Allison Linn**

Microsoft researchers on Thursday announced a major advance in technology designed to identify the objects in a photograph or video, showcasing a system whose accuracy meets and sometimes exceeds human-level performance.

Microsoft's new approach to recognizing images also took first place in several major categories of image recognition challenges Thursday, beating out many other competitors from academic, corporate and research institutions in the ImageNet and Microsoft Common Objects in Context challenges.

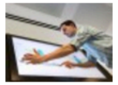
Featured Posts


Microsoft releases CNTK, its open source deep learning toolkit, on GitHub


Molecular biology meets computer science tools in new system for CRISPR


Microsoft researchers win ImageNet computer vision challenge

Popular Posts

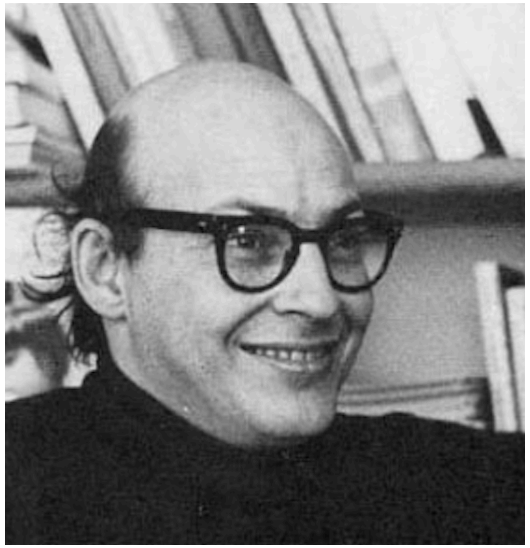

How Microsoft and Novartis created Assess MS


Microsoft releases CNTK, its open source deep learning toolkit, on GitHub


Molecular biology meets computer science tools in new system for CRISPR

1966

"Connect a television camera to a computer and get the machine to describe what it sees."



Marvin Minsky
Turing award, 1969

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
PROJECT MAC

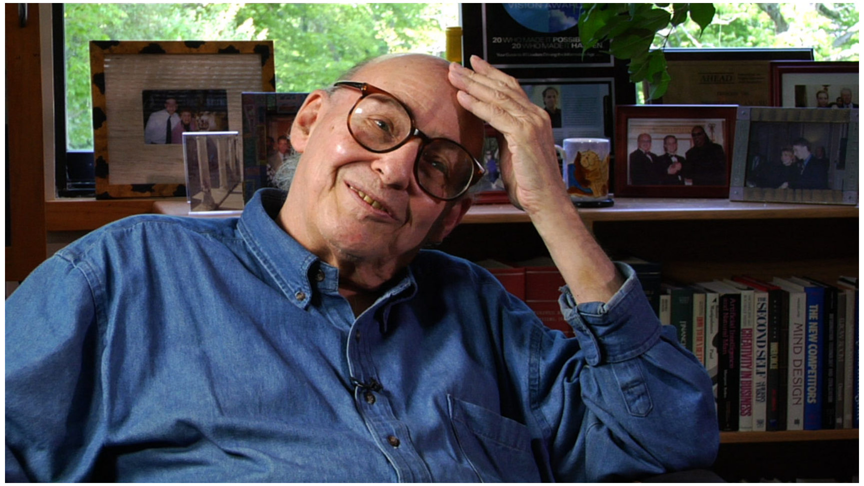
Artificial Intelligence Group July 7, 1966
Vision Memo. No. 100.

THE SUMMER VISION PROJECT

Seymour Papert

The summer vision project is an attempt to use our summer workers effectively in the construction of a significant part of a visual system. The particular task was chosen partly because it can be segmented into sub-problems which will allow individuals to work independently and yet participate in the construction of a system complex enough to be a real landmark in the development of "pattern recognition".

How hard is computer vision?



Marvin Minsky
Turing award, 1969

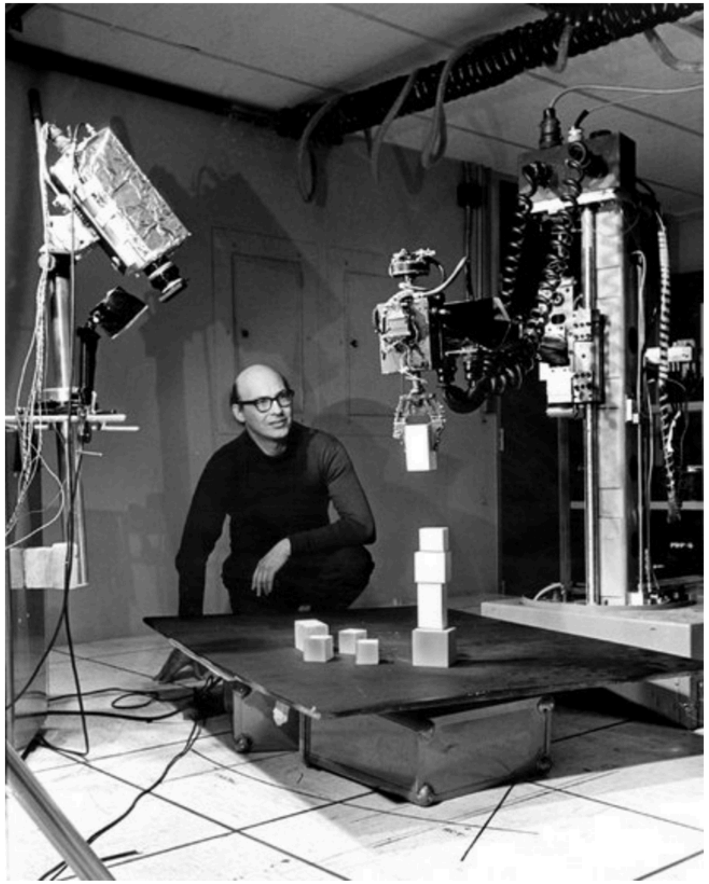
Gerald Sussman

"You'll notice that Sussman never worked in vision again"
-Berthold Horn

The New York Times

Marvin Minsky, Pioneer in Artificial Intelligence, Dies at 88

By GLENN RIFKIN JAN. 25, 2016



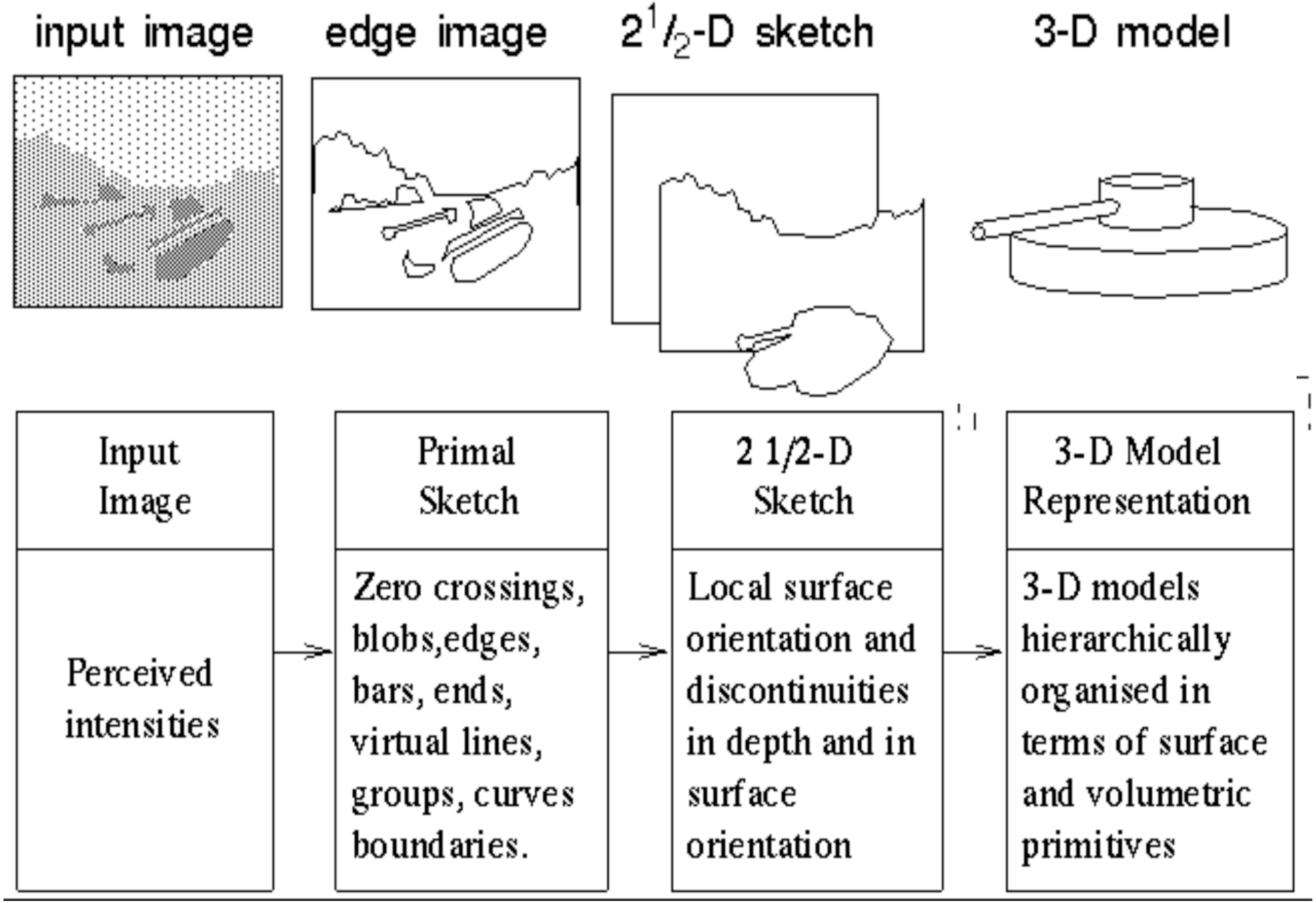
Marvin Minsky in a lab at M.I.T. in 1968. M.I.T.

Marvin Minsky, who combined a scientist's thirst for knowledge with a philosopher's quest for truth as a pioneering explorer of artificial intelligence, work that helped inspire the creation of the personal computer and the Internet, died on Sunday night in Boston. He was 88.

His family said the cause was a cerebral hemorrhage.

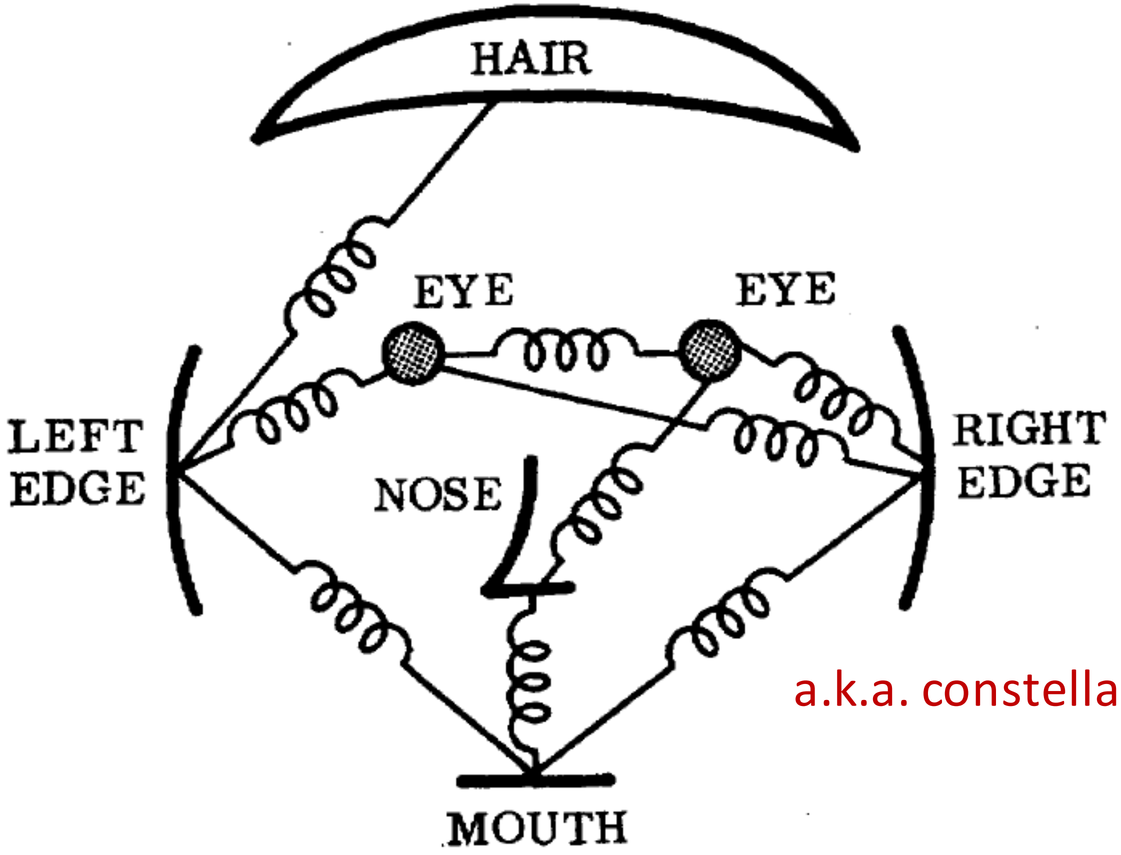
Well before the advent of the microprocessor and the supercomputer, Professor Minsky, a revered computer science educator at M.I.T., laid the foundation for the field of artificial intelligence by demonstrating the possibilities of imparting common-sense reasoning to computers.

"Marvin was one of the very few people in computing whose visions and perspectives liberated the computer



Stages of Visual Representation, David Marr, 1970

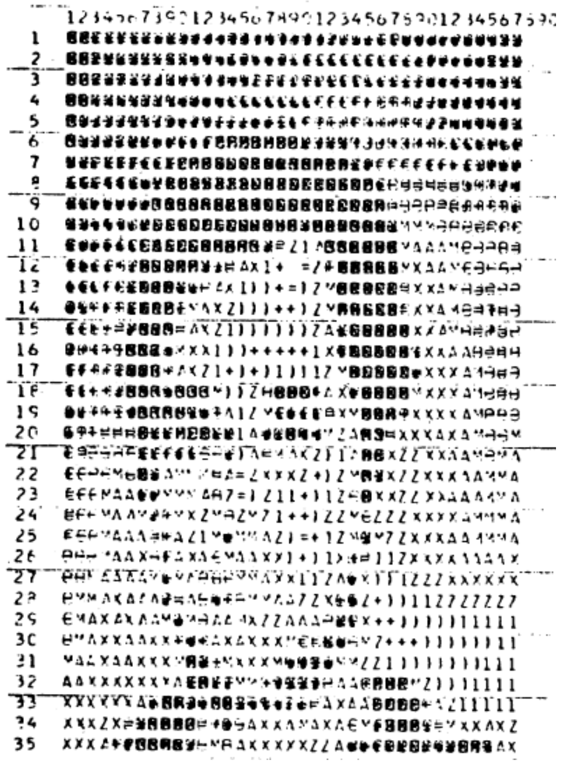
1973

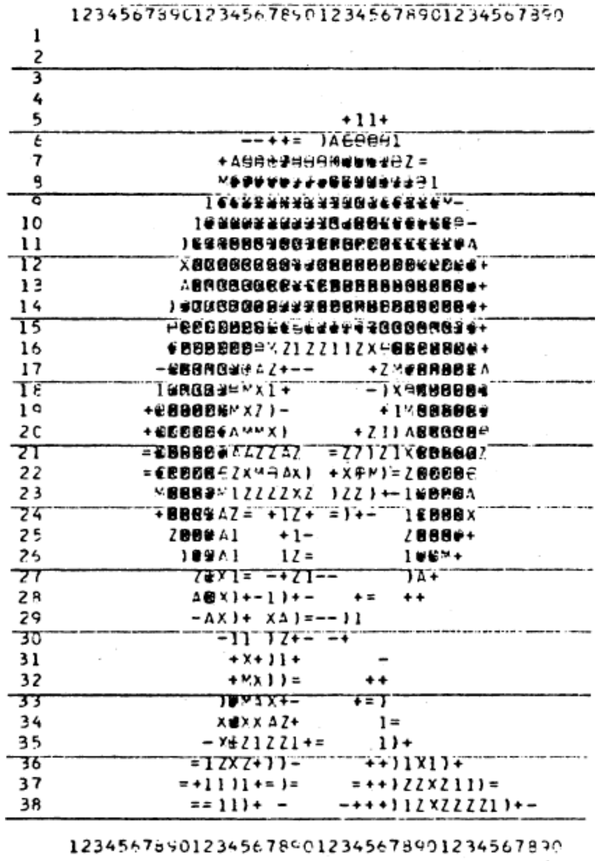


a.k.a. constellation model

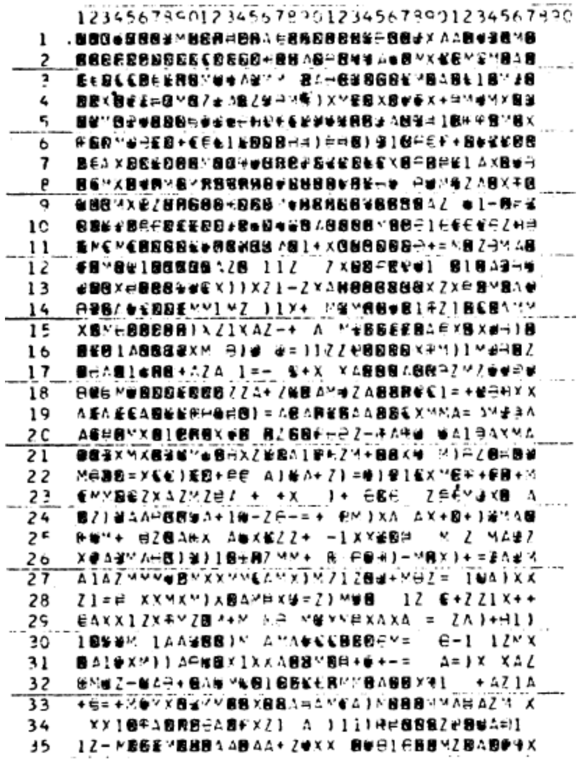
The representation and matching of pictorial structures, Fischler and Elschlager, 1973

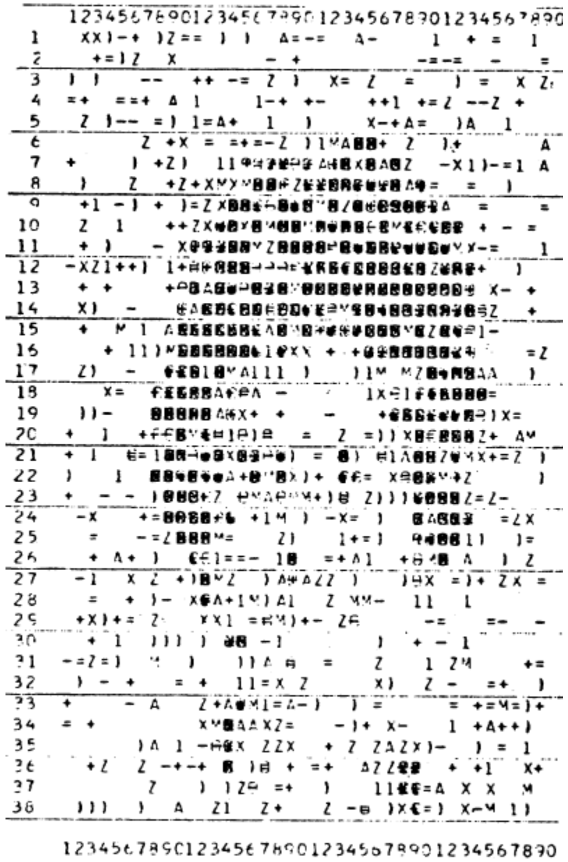
1973





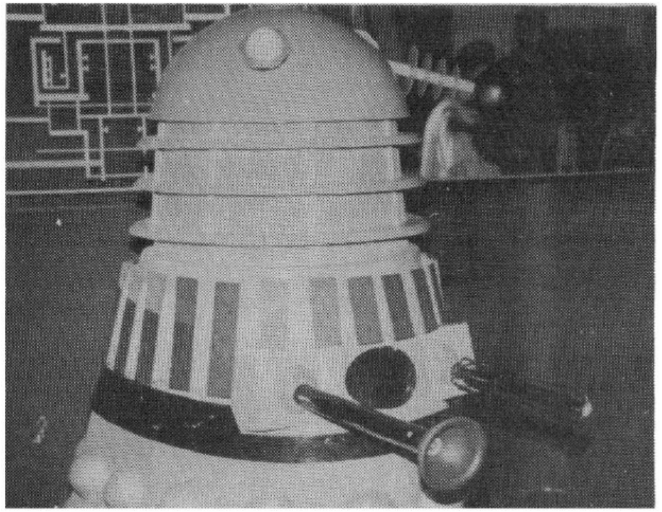
1973

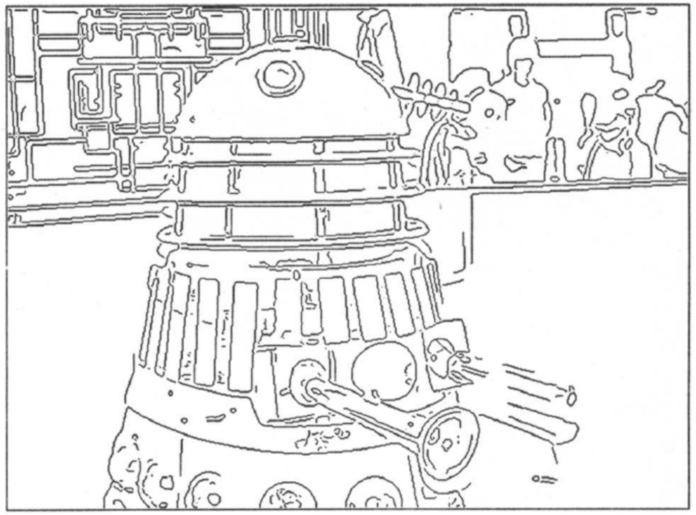




1980's

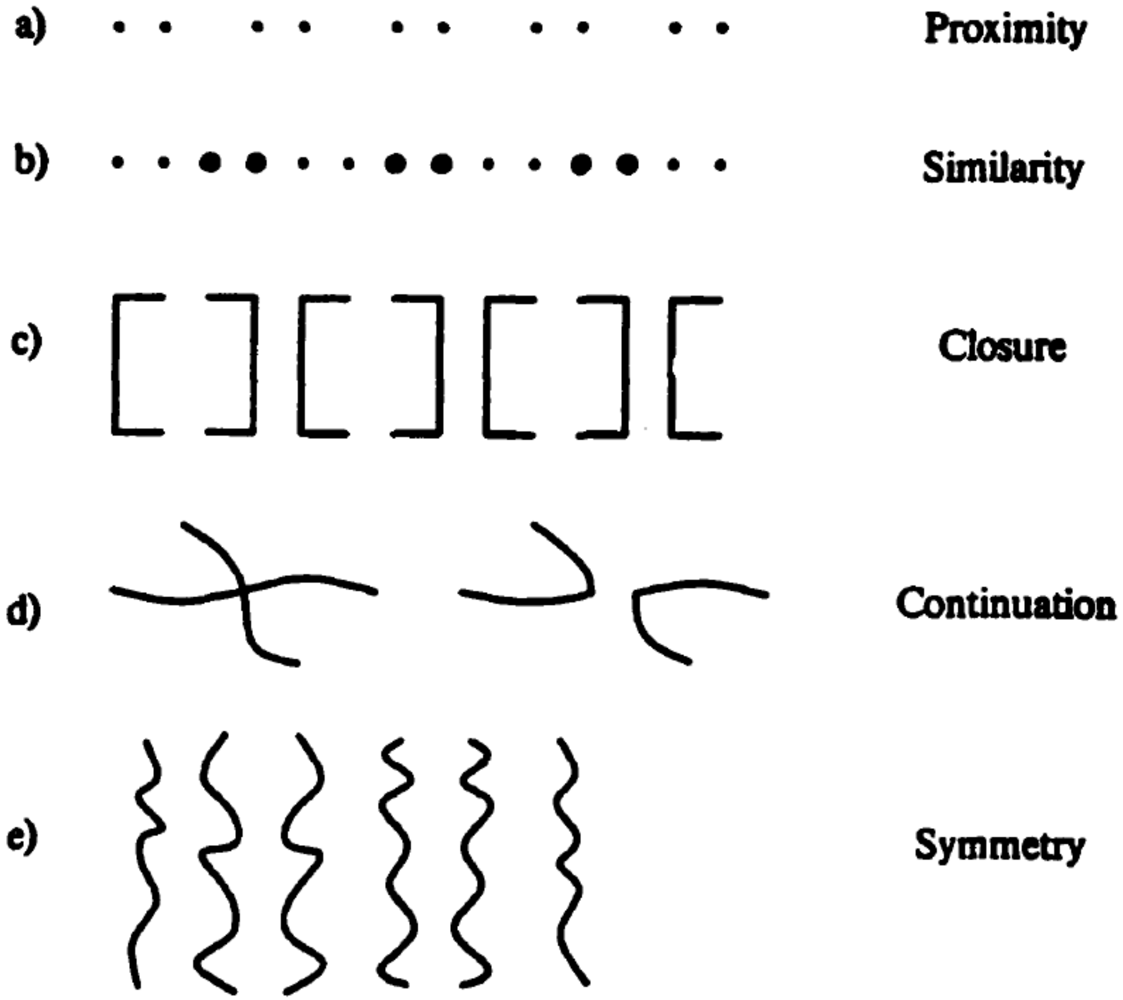
AI winter… …back to basics





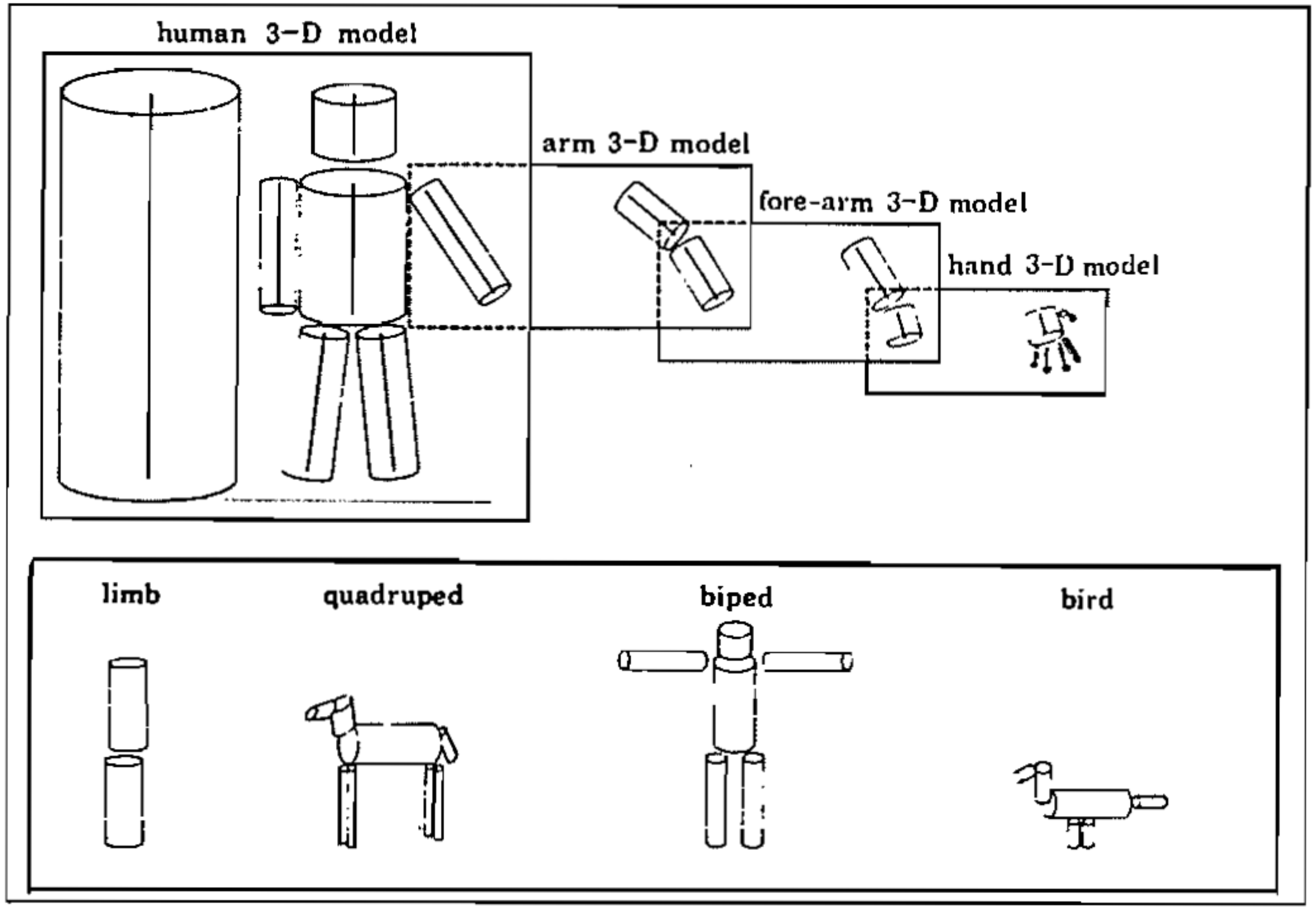
A Computational Approach to Edge Detection, Canny 1986

1984

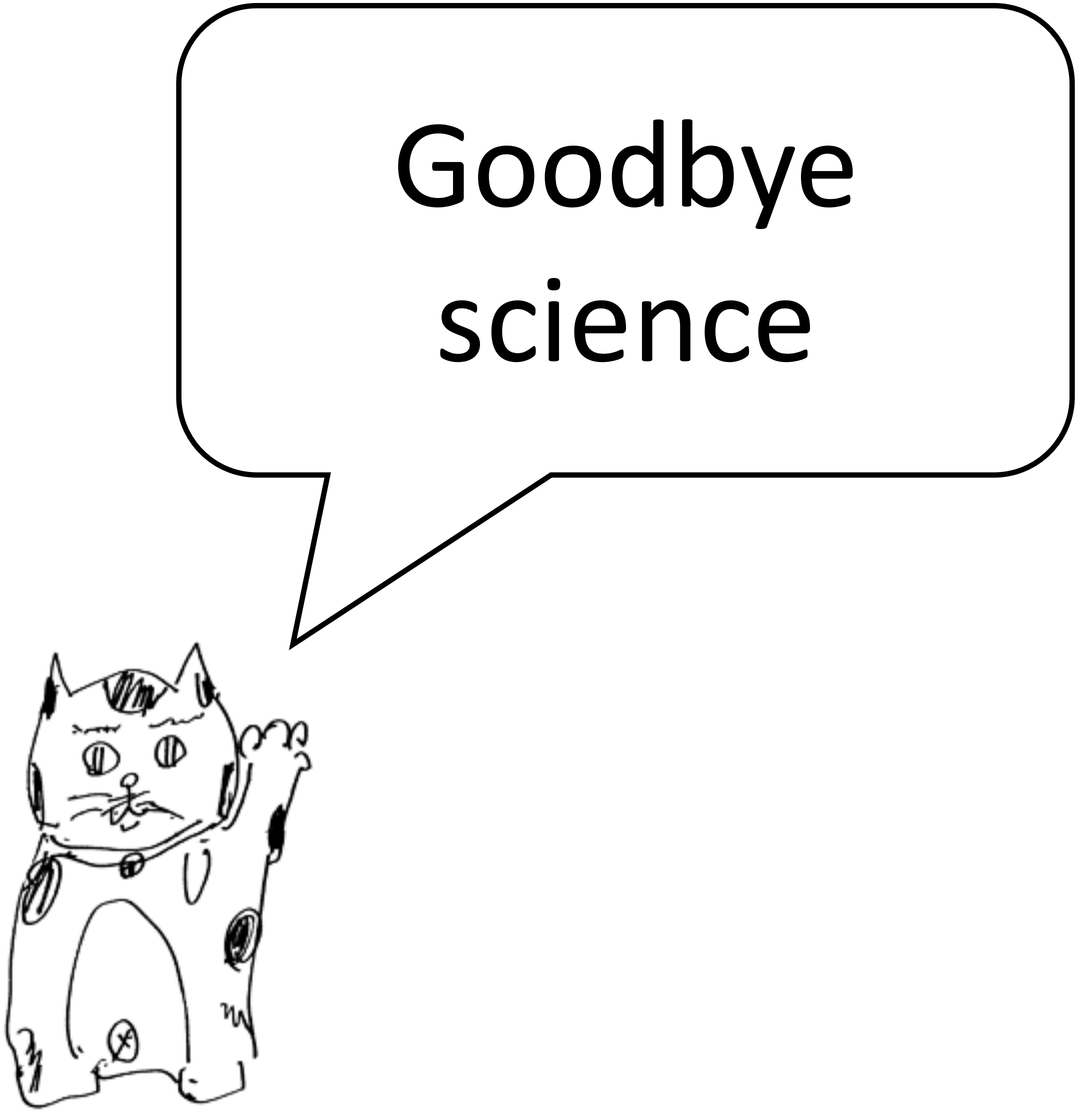


Perceptual Organization and Visual Recognition,
David Lowe, 1984

1986

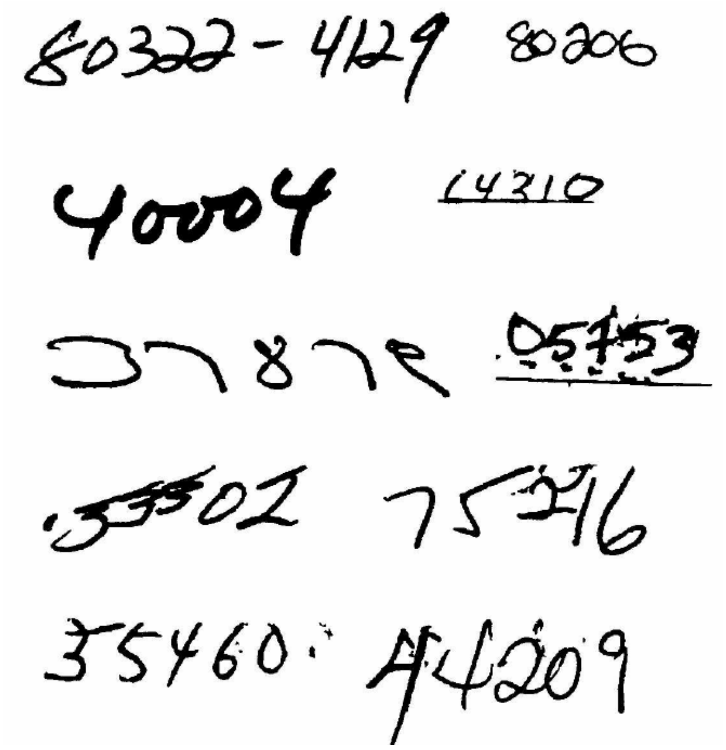


Perceptual organization and the representation of natural form,
Alex Pentland, 1986


Goodbye science

1989

MNIST

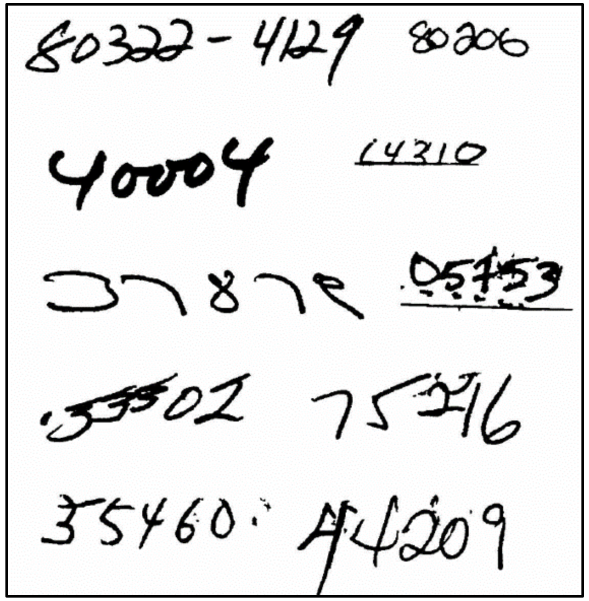


Zip codes

Backpropagation applied to handwritten zip code recognition,
Lecun et al., 1989

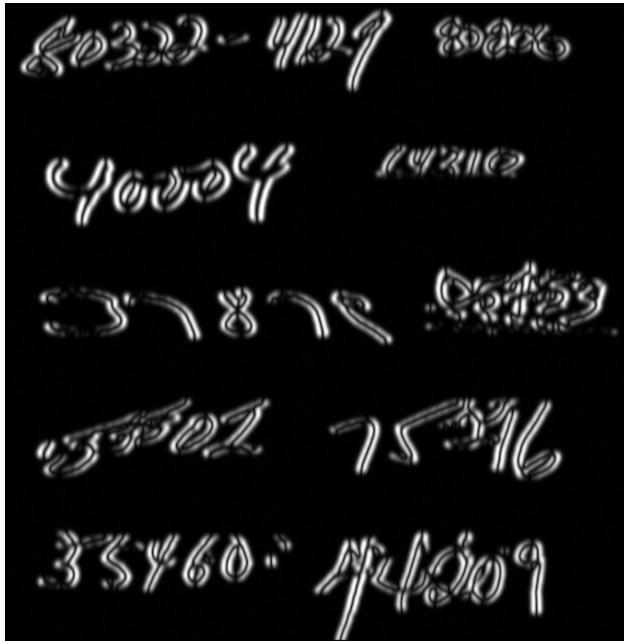
Filters

Input



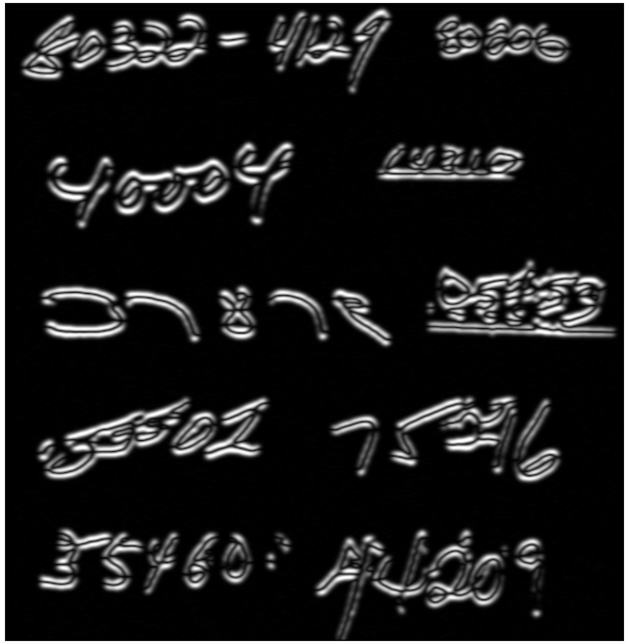
-1	0	+1
-2	0	+2
-1	0	+1

x filter

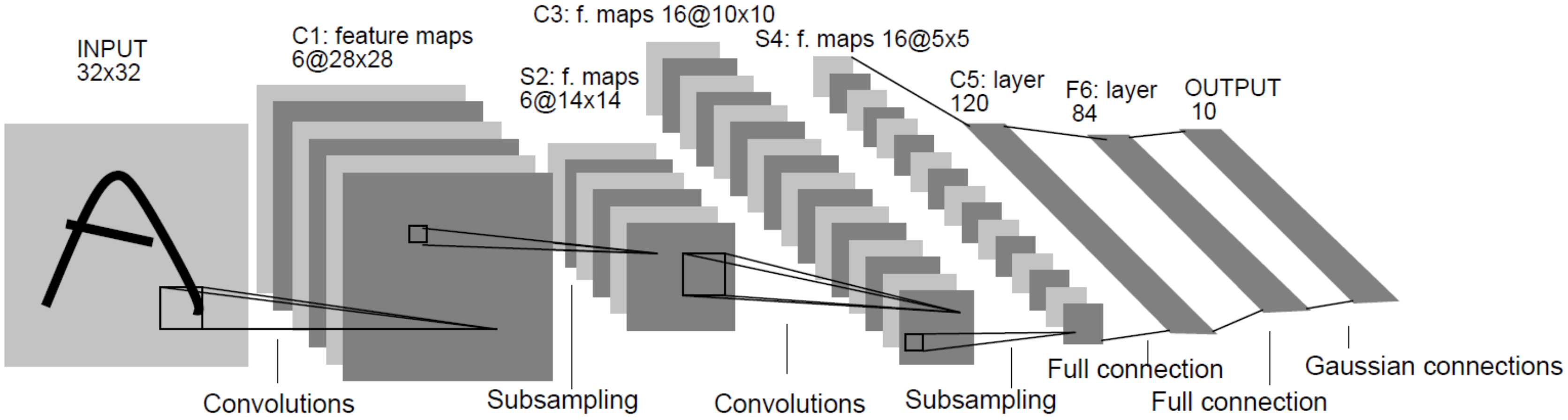


+1	+2	+1
0	0	0
-1	-2	-1

y filter

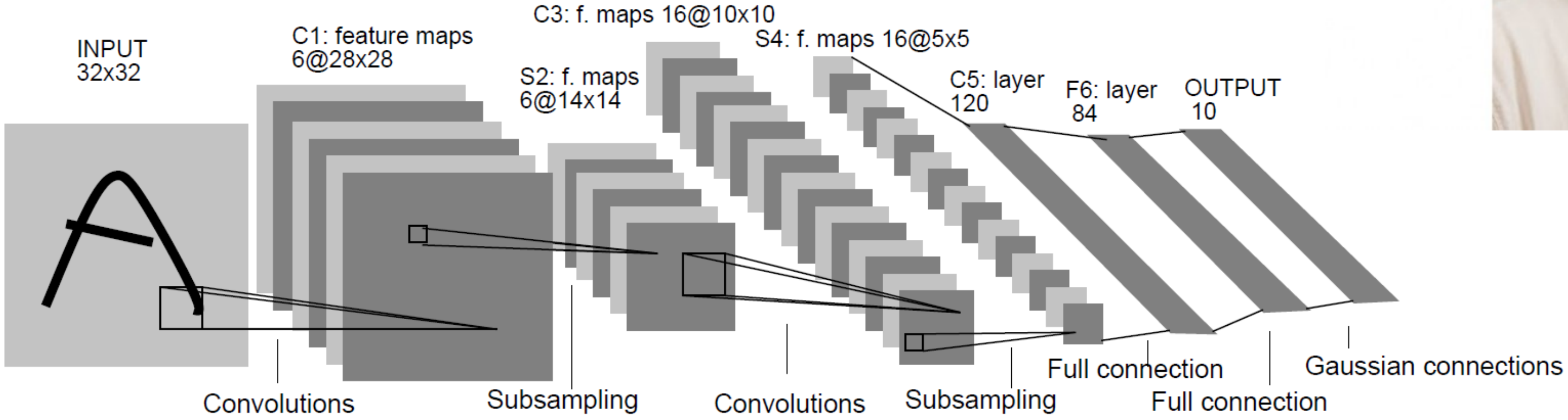


1989



Backpropagation applied to handwritten zip code recognition,
Lecun et al., 1989

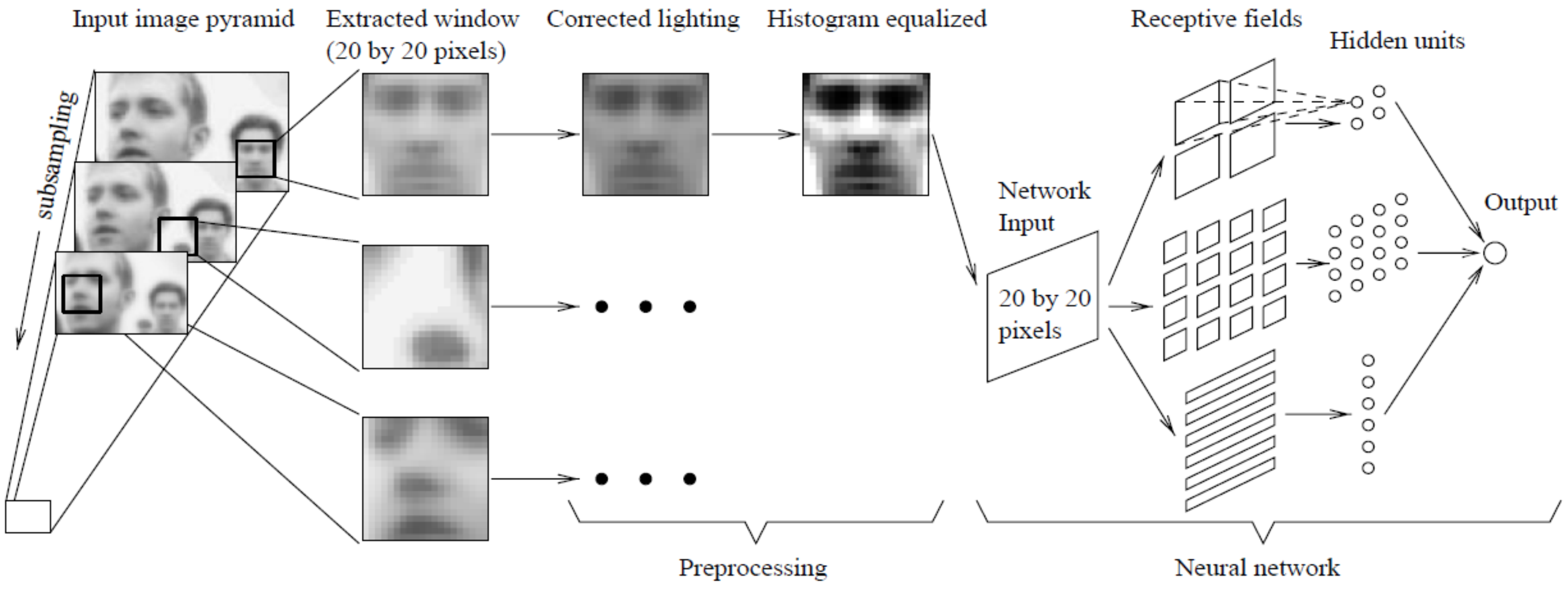
1989



Backpropagation applied to handwritten zip code recognition,
Lecun et al., 1989

1998

Faces

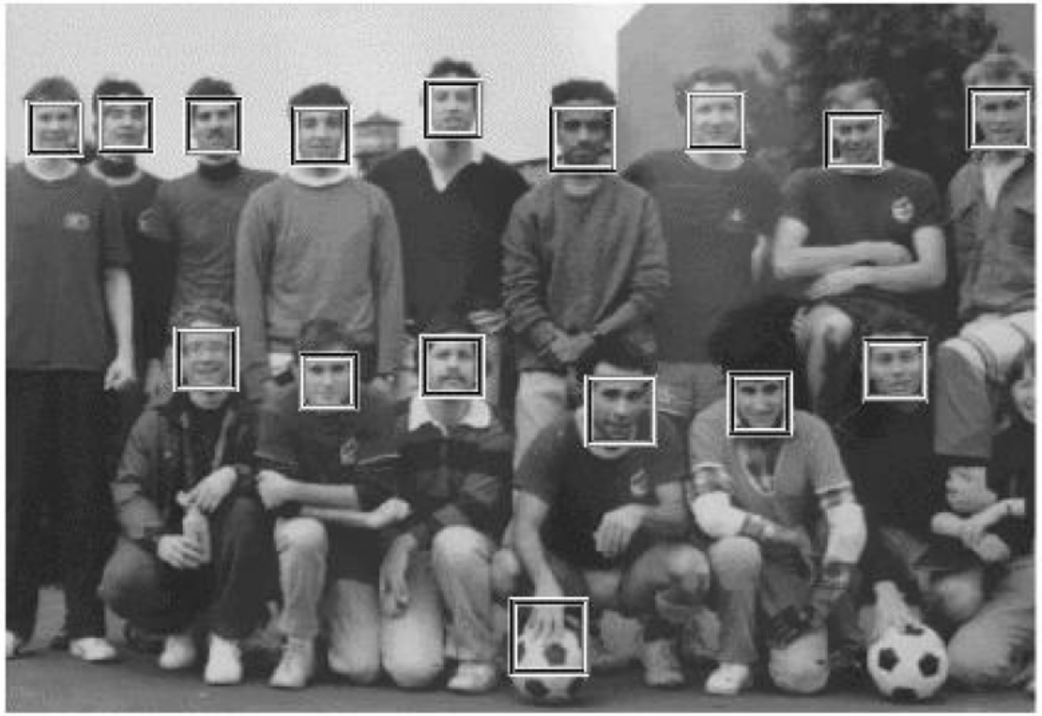


Neural Network-Based Face Detection, Rowley at al., PAMI 1998

2001

Sliding window in real time!

Boosting + Cascade = Speed



Rapid Object Detection using a Boosted Cascade of Simple Features,
Viola and Jones, CVPR 2001

Why did it work?

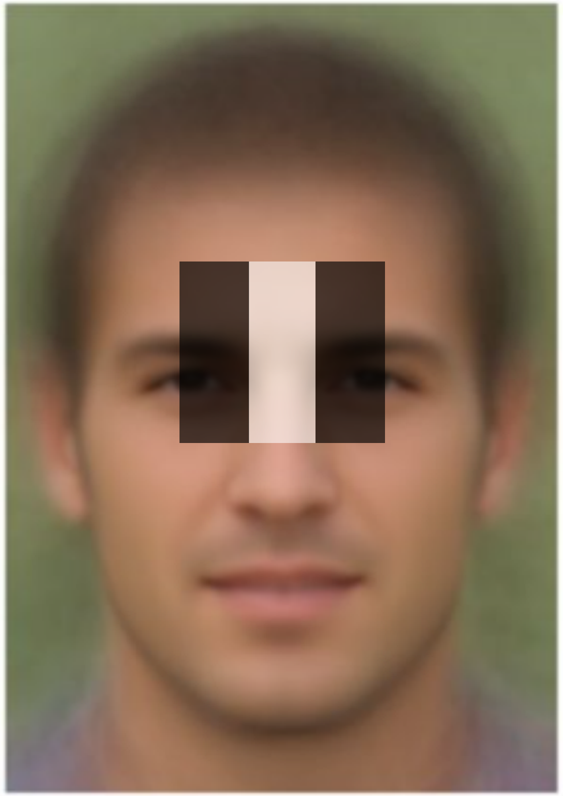
- Simple features (Haar wavelets)

□ - ■ = h

Integral images + Haar wavelets = fast

Face Detection, Viola & Jones, 2001

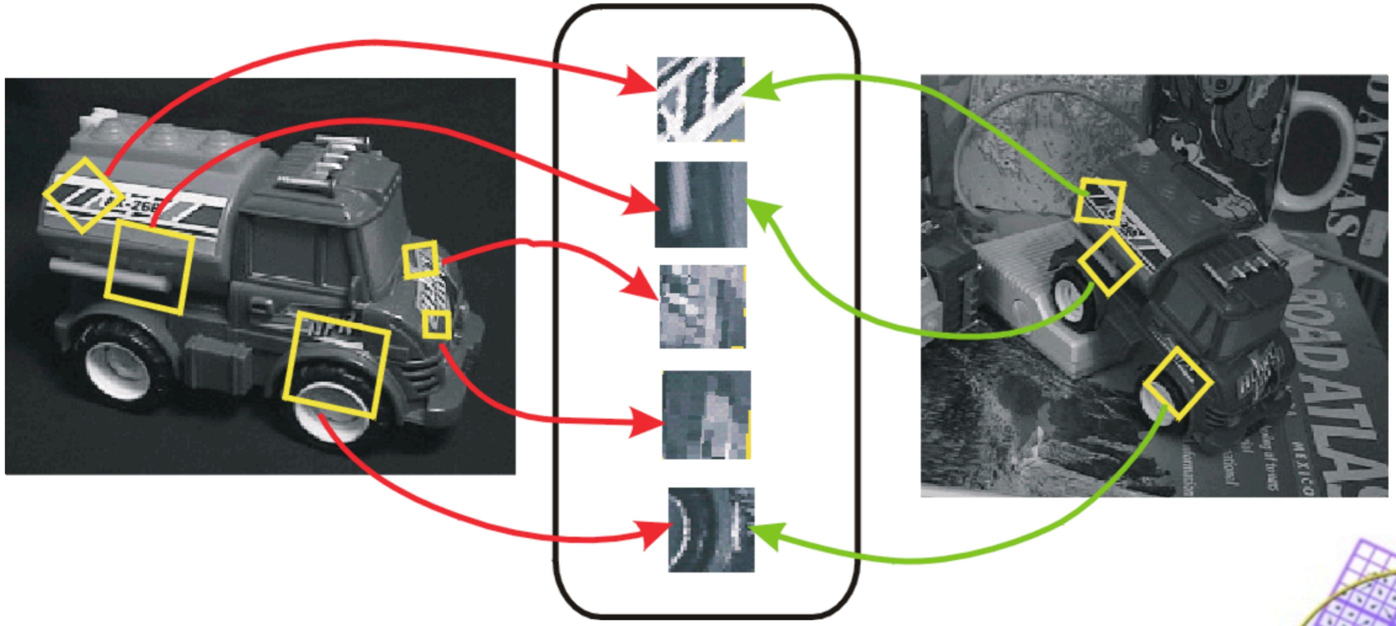
Why did it work?

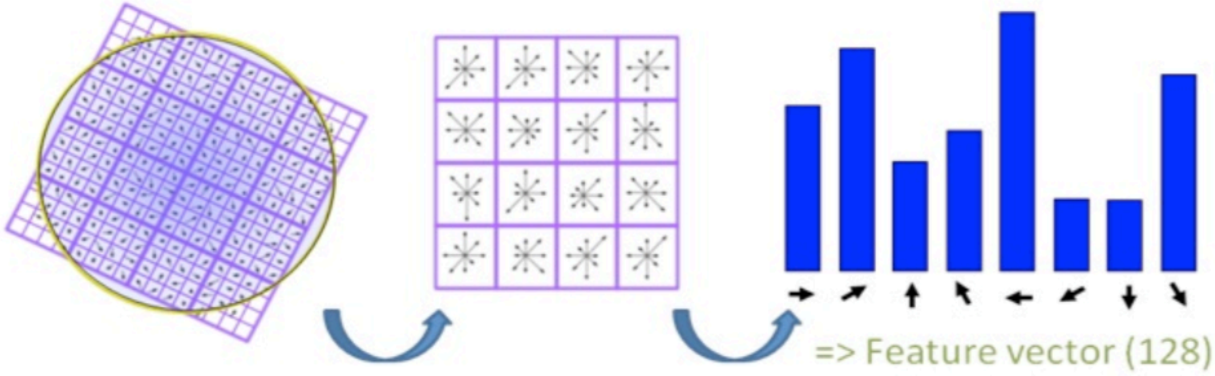


Why did it fail?

1999*

SIFT (Scale Invariant Feature Transform)

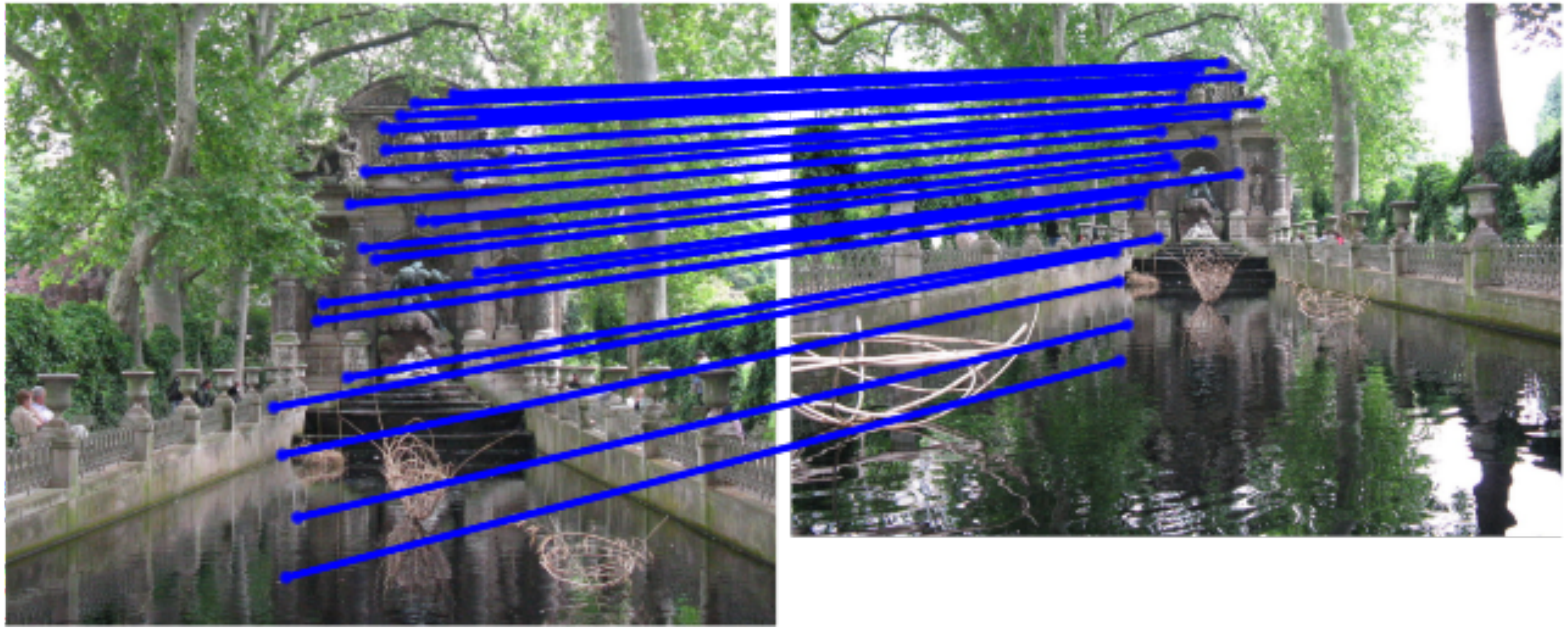




No more sliding windows (interest points)
Better features (use more computation)

Object Recognition from Local Scale-Invariant Features, Lowe, ICCV 1999.

SIFT Matching



[SIFT: Lowe, 2004]

What worked

Panorama stitching

Recognizing panoramas, Brown and Lowe, *ICCV* 2003

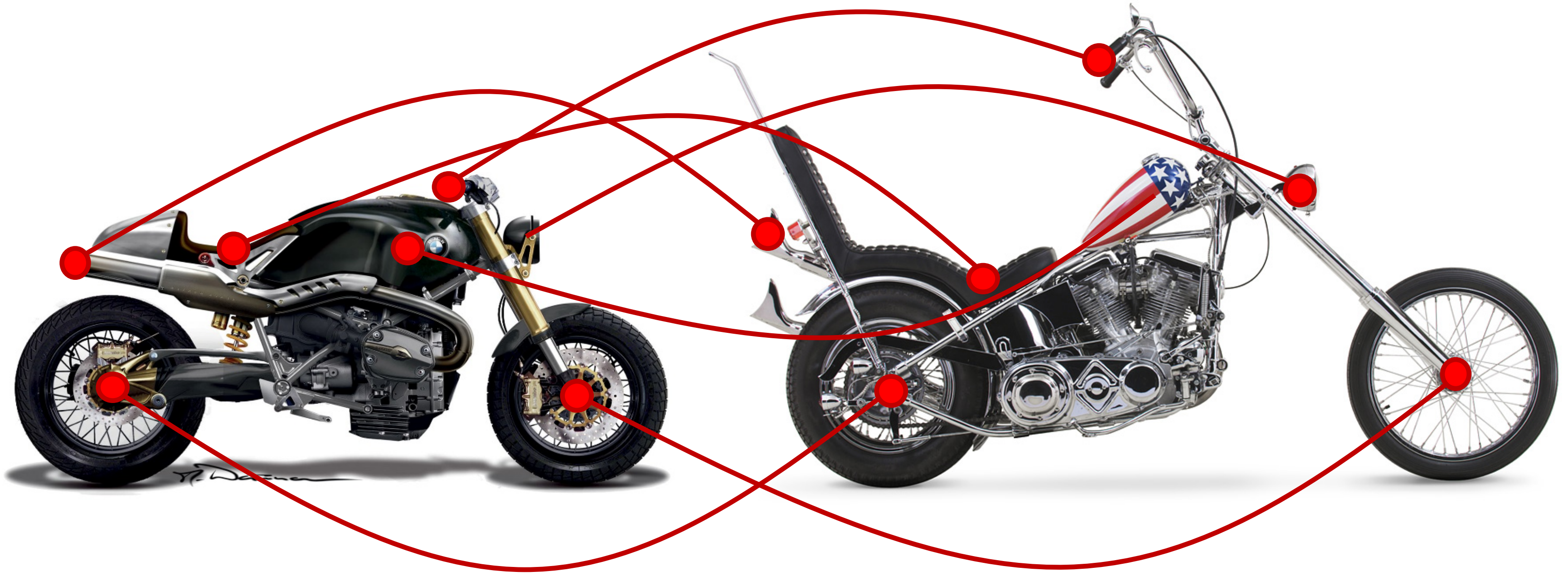
SIFT Matching

[SIFT: Lowe, 2004]

Interest points

2003

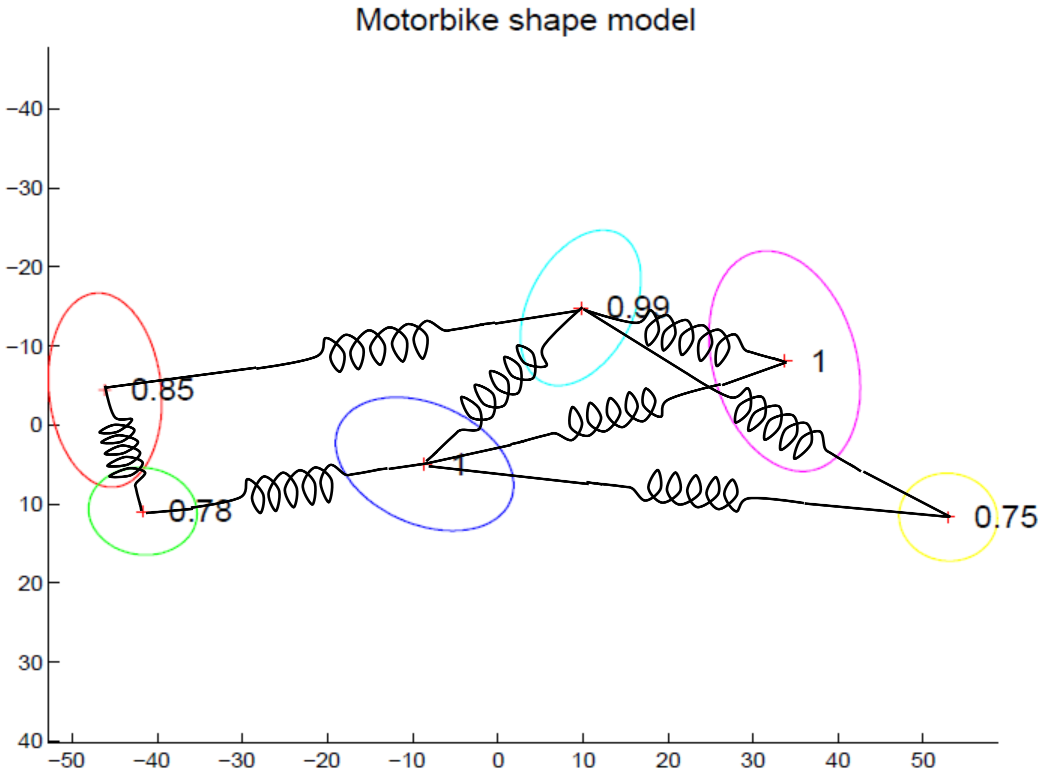
Constellation model (redux)



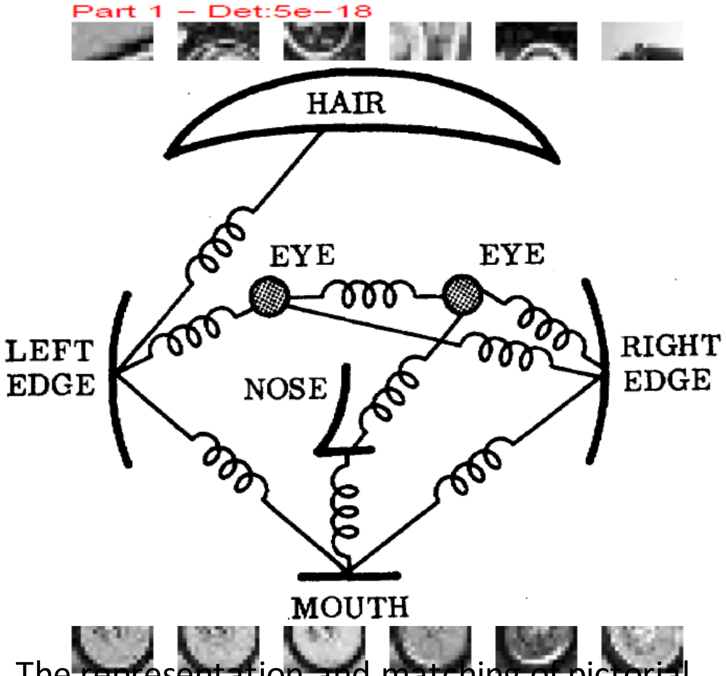
Object Class Recognition by Unsupervised Scale-Invariant Learning, Fergus et al., *CVPR* 2003.

2003

Constellation model (redux)

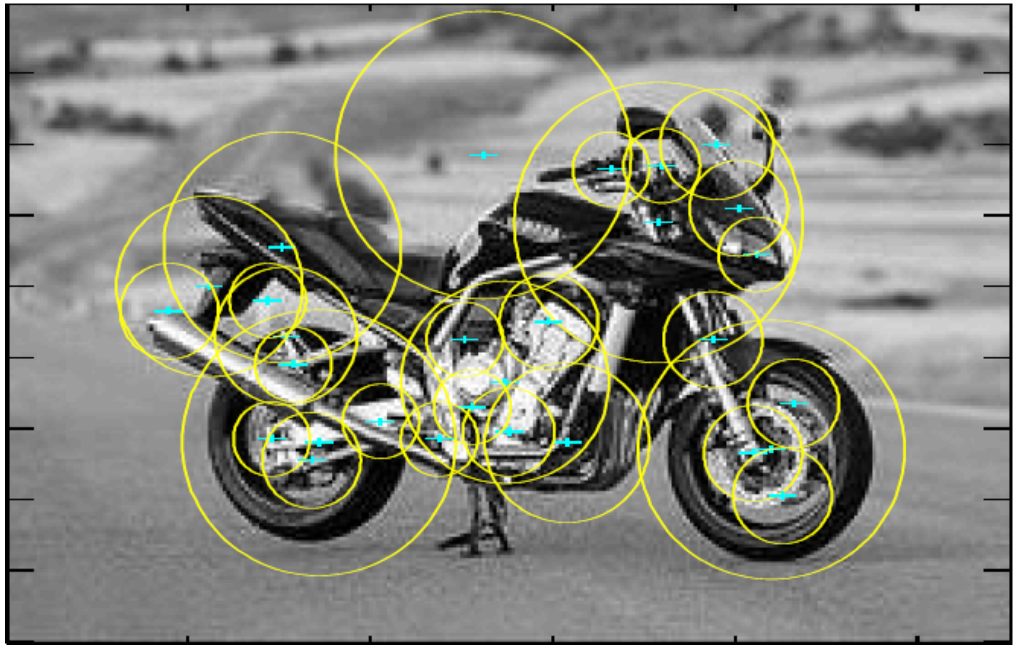


Joint Gaussian density



The representation and matching of pictorial structures, Fischler and Elschlager, 1973

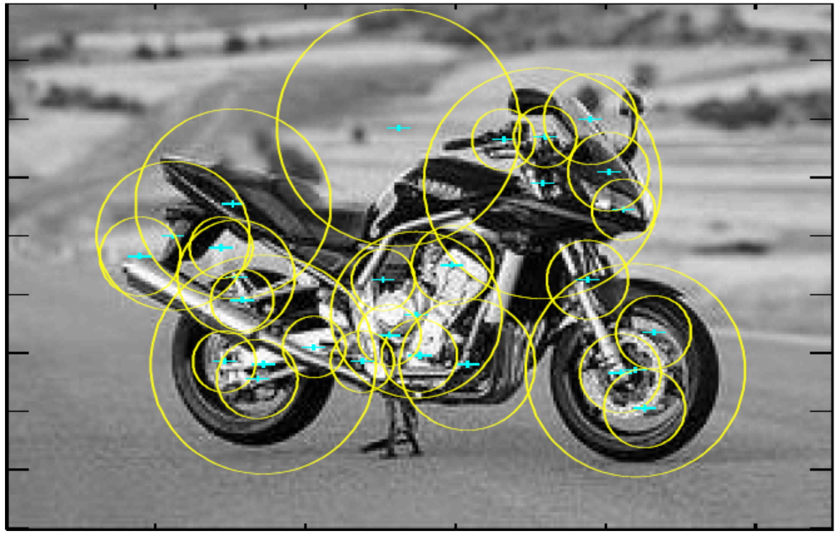
Interest points used to find parts:

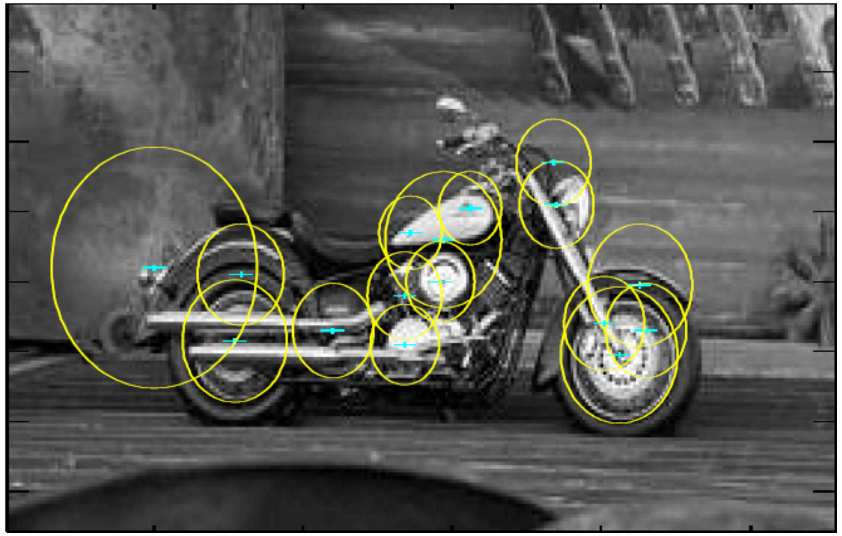


Smaller number of candidate parts allows for more complex spatial models.

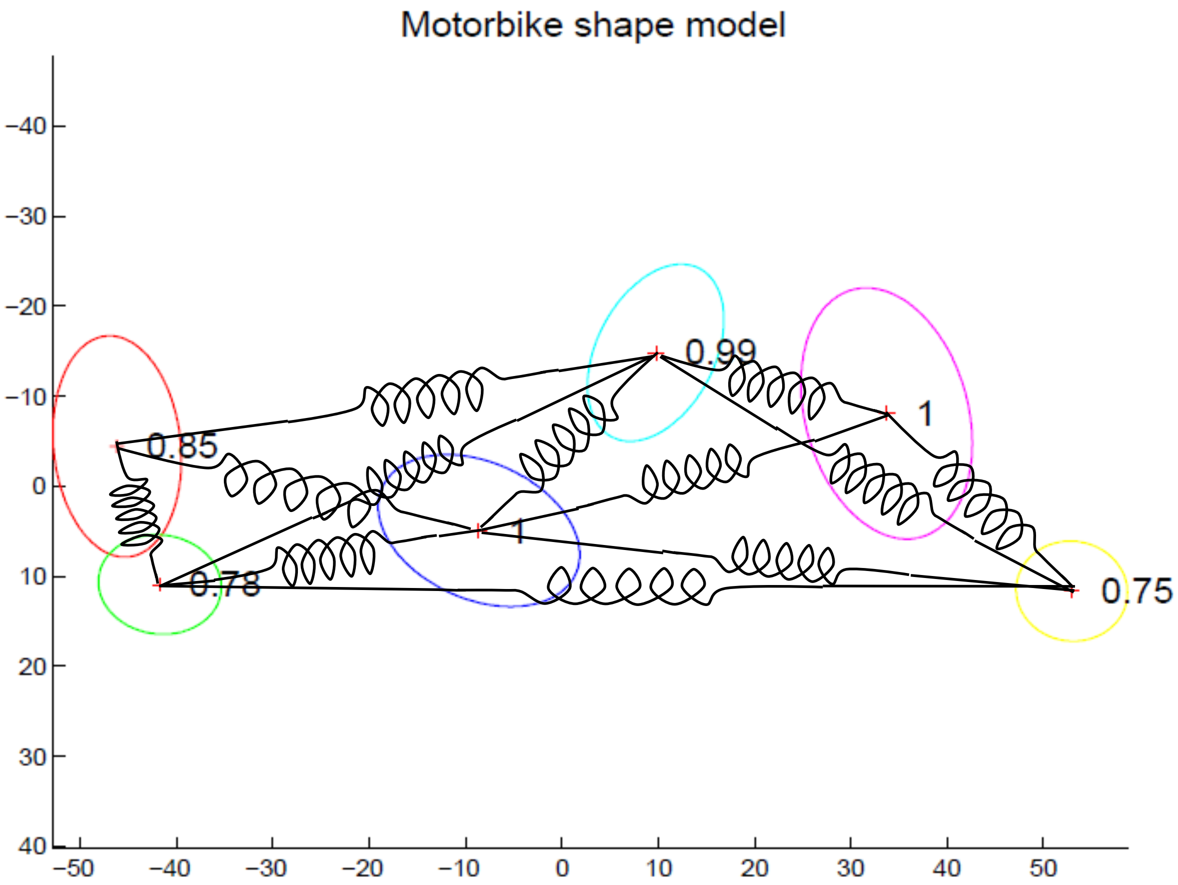
Why it fails

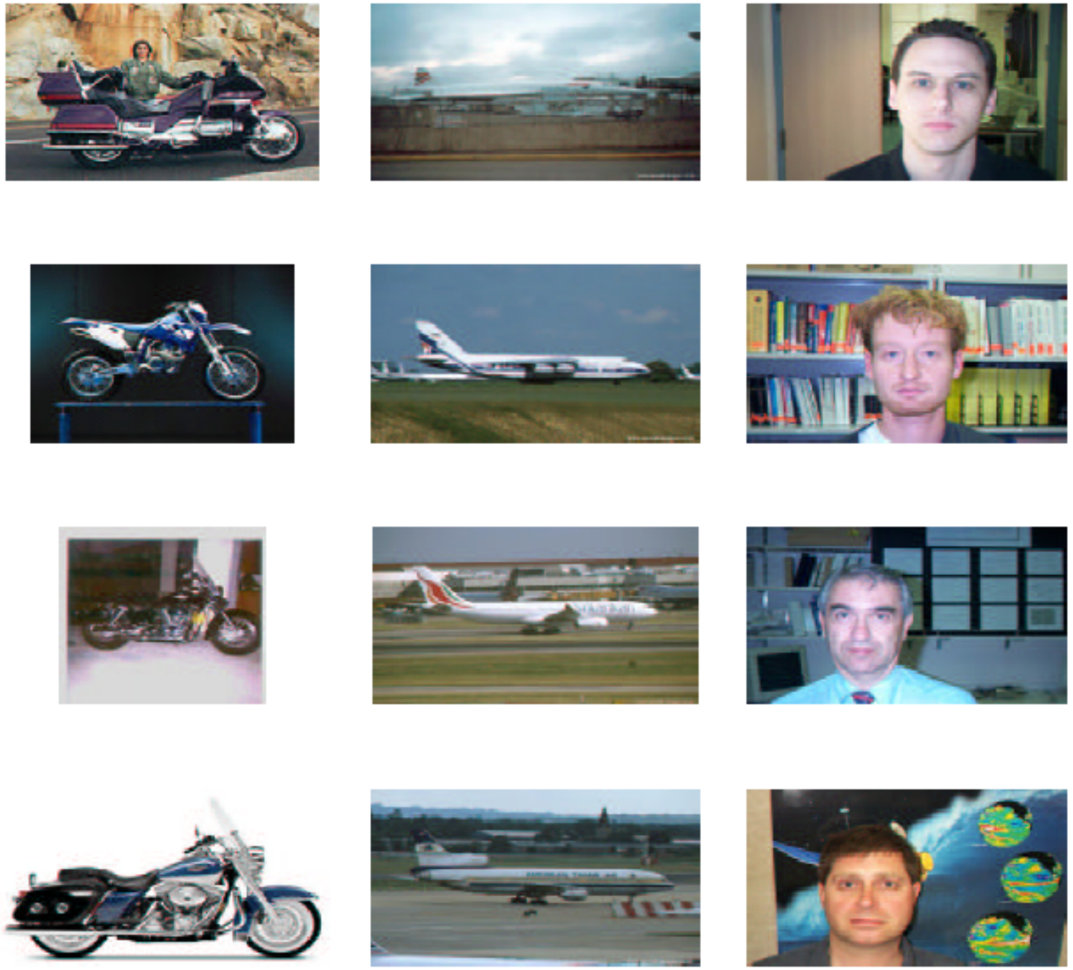
Interest points don't work for category recognition





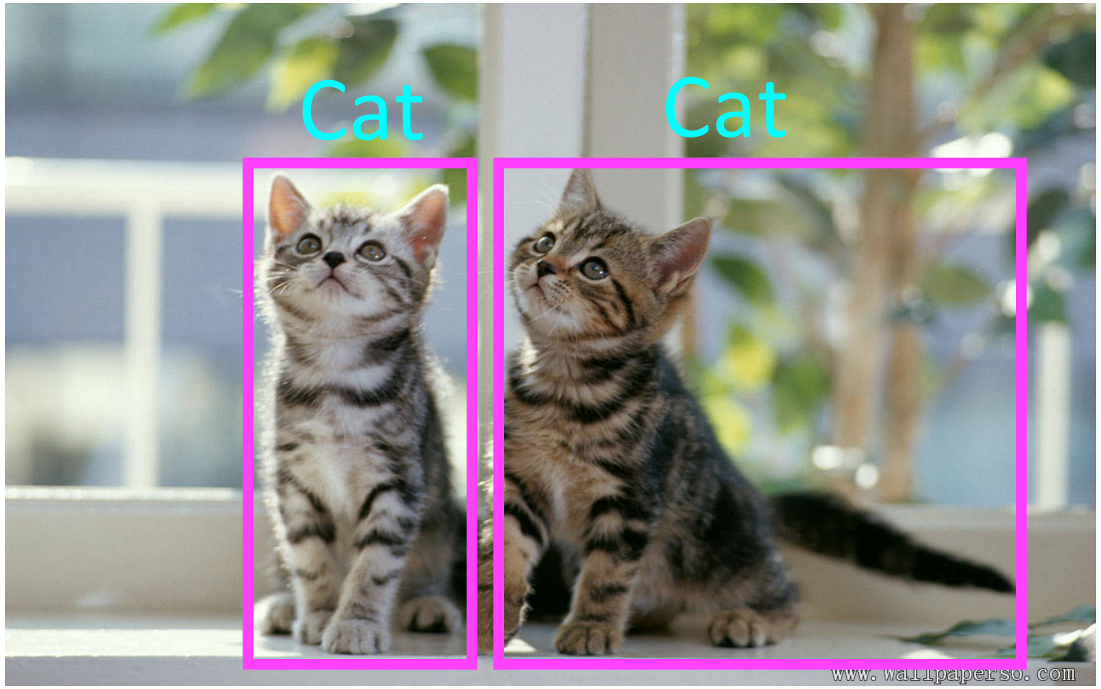
Too many springs...



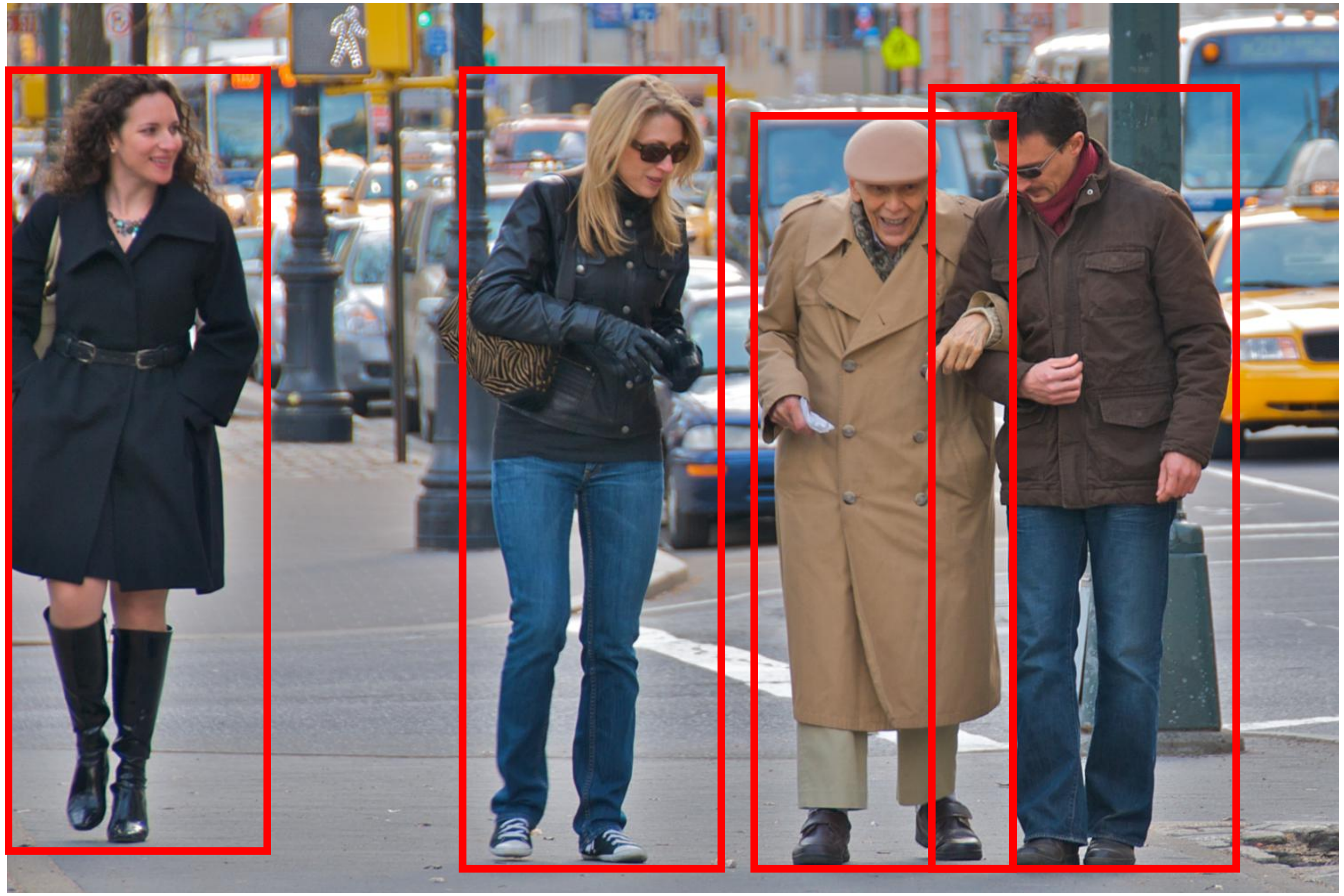


Cat?

Classification Vs. Detection



2005 HOG (histograms of oriented gradients)



Histograms of oriented gradients for human detection,
Dalal and Triggs, CVPR 2005.

Pedestrians

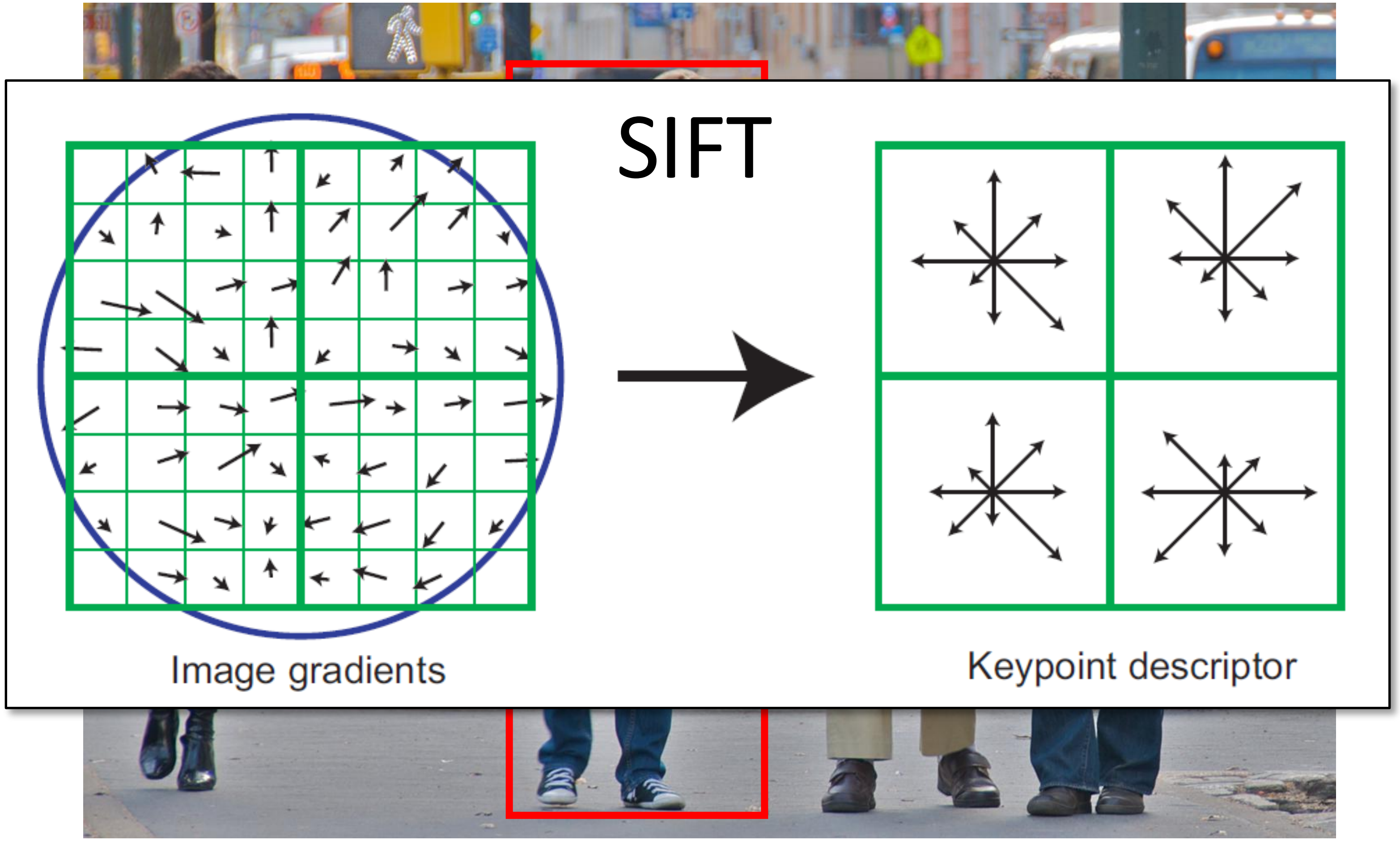
- Defined by their contours
- Cluttered backgrounds
- Significant variance in texture

Interest points won’t work…

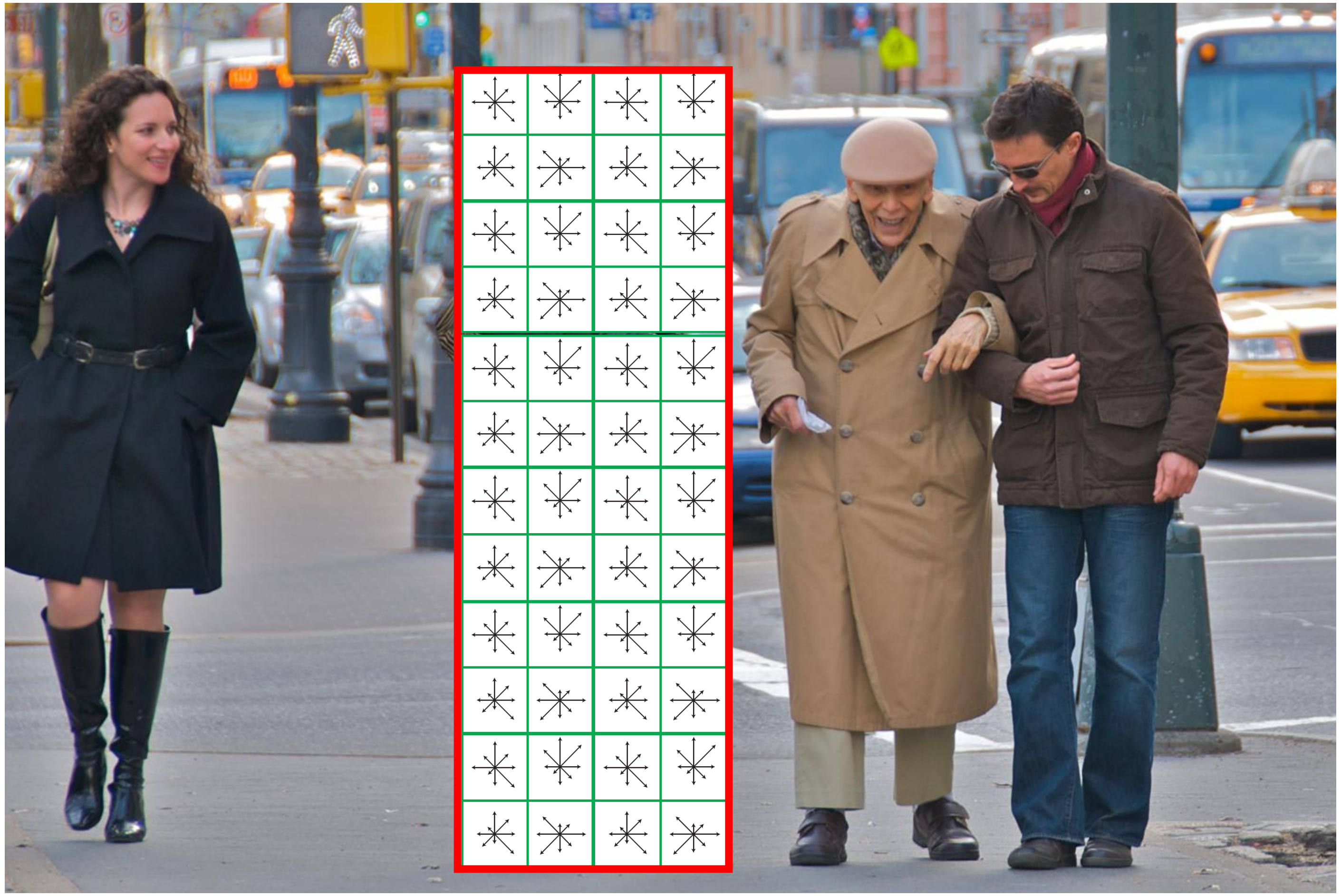
…back to sliding window.

2005 HOG (histograms of oriented gradients)

2005 HOG (histograms of oriented gradients)



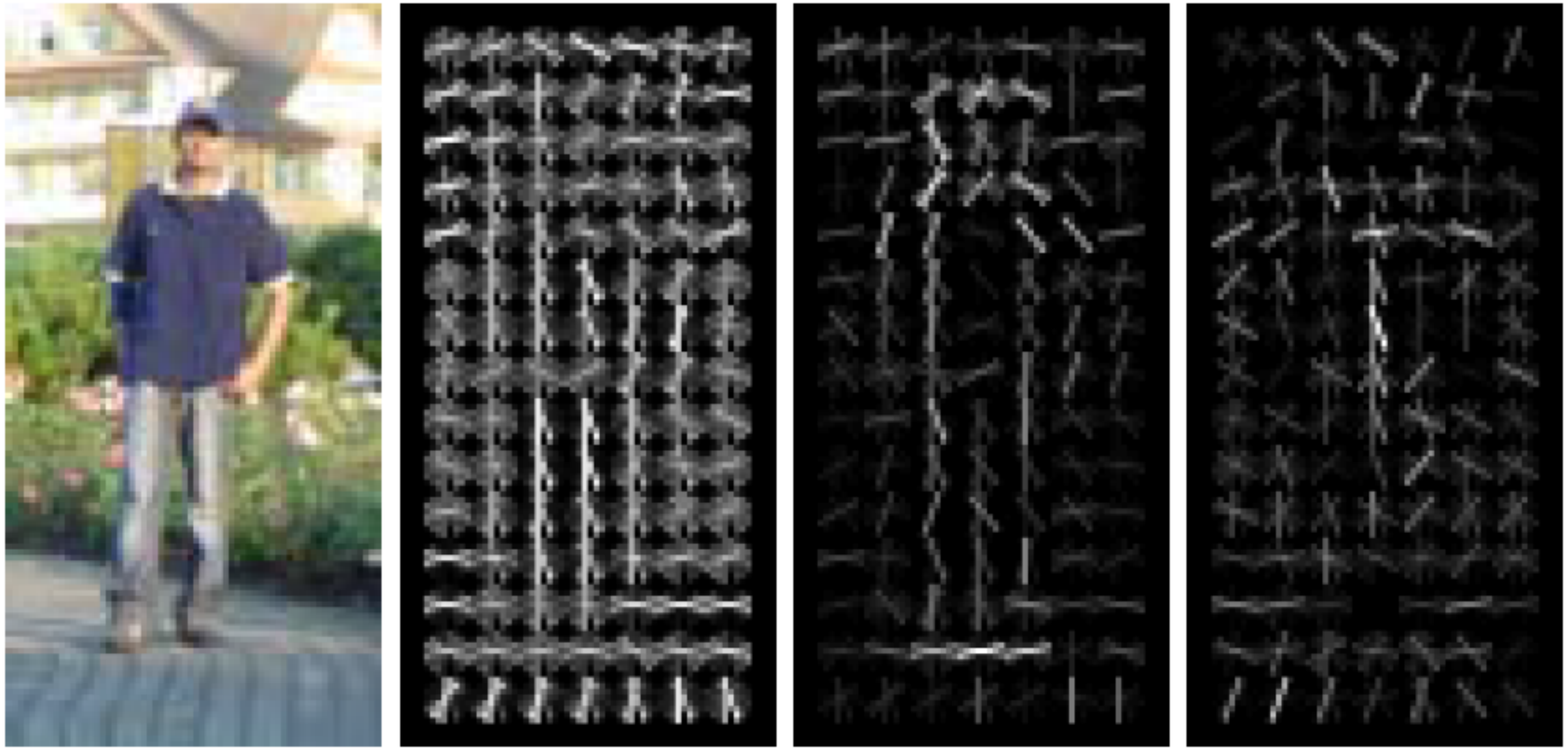
2005 HOG (histograms of oriented gradients)



Histograms of oriented gradients for human detection,
Dalal and Triggs, *CVPR* 2005.

2005 HOG (histograms of oriented gradients)

Presence > Magnitude



✓ Normalization by a local window

Why it worked

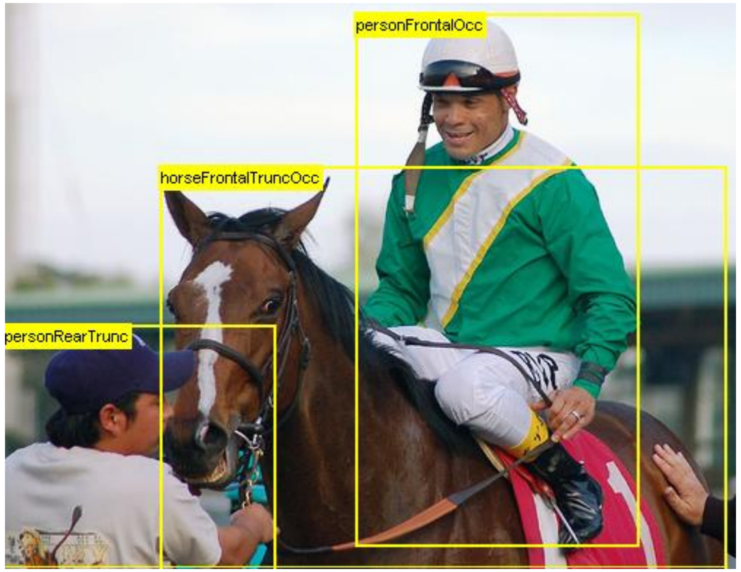
We can finally detect object boundaries in a reliable manner!

Hard negative mining

Computers are fast enough.

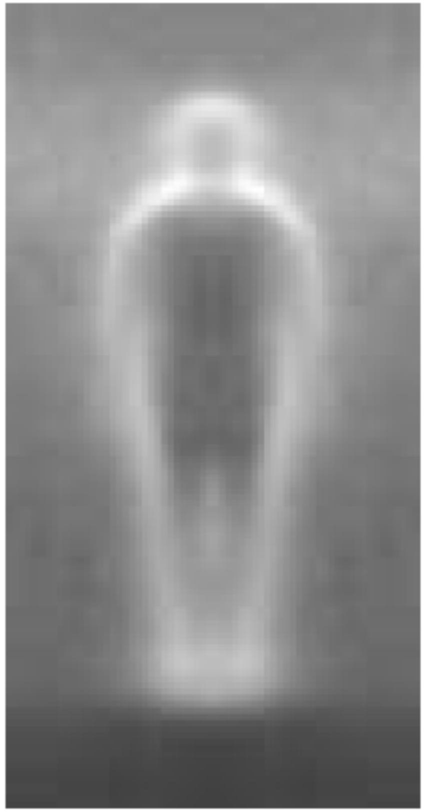
2007 PASCAL VOC

20 classes

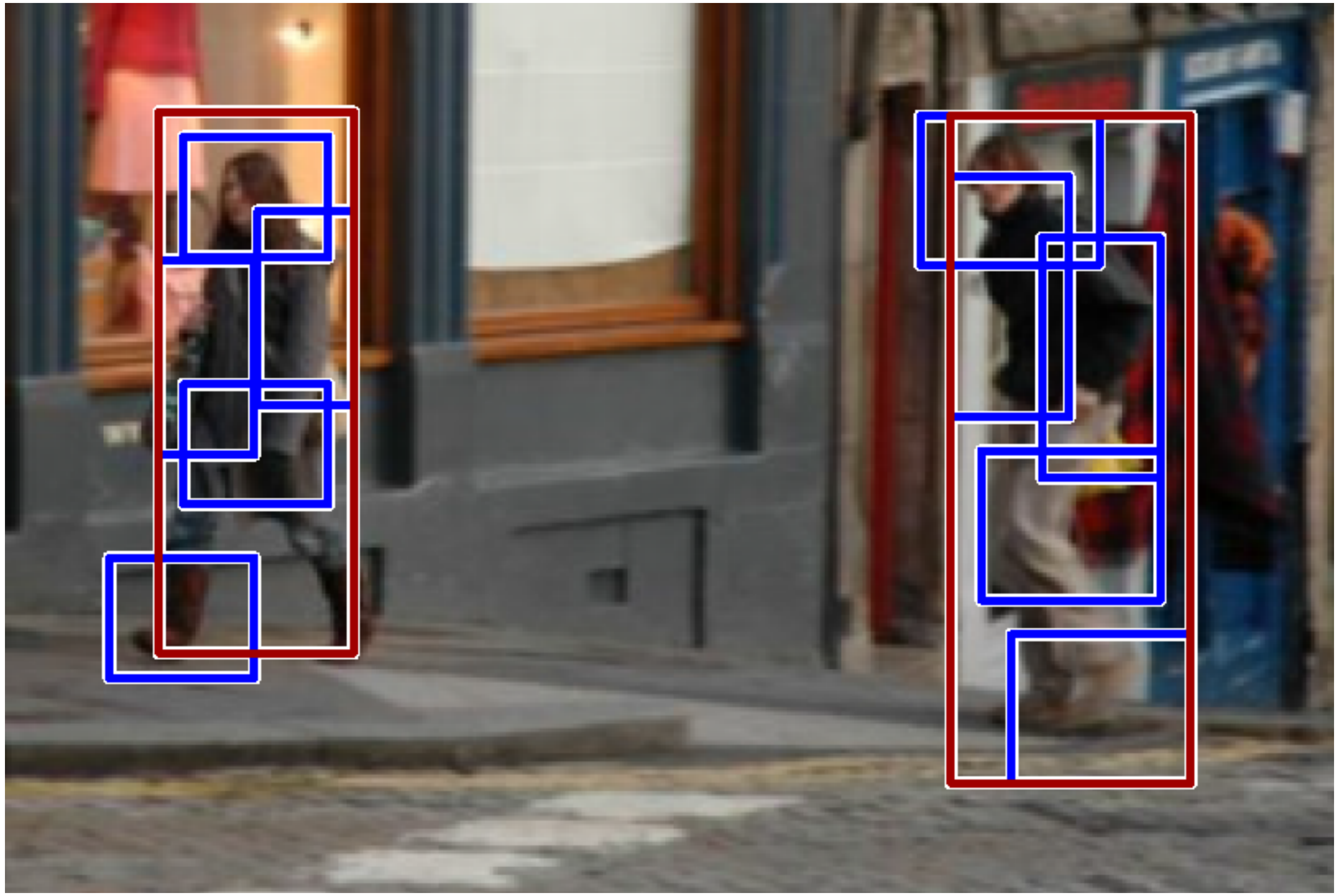


The PASCAL Visual Object Classes (VOC) Challenge, Everingham, Van Gool, Williams, Winn and Zisserman, *IJCV*, 2010

Why it failed

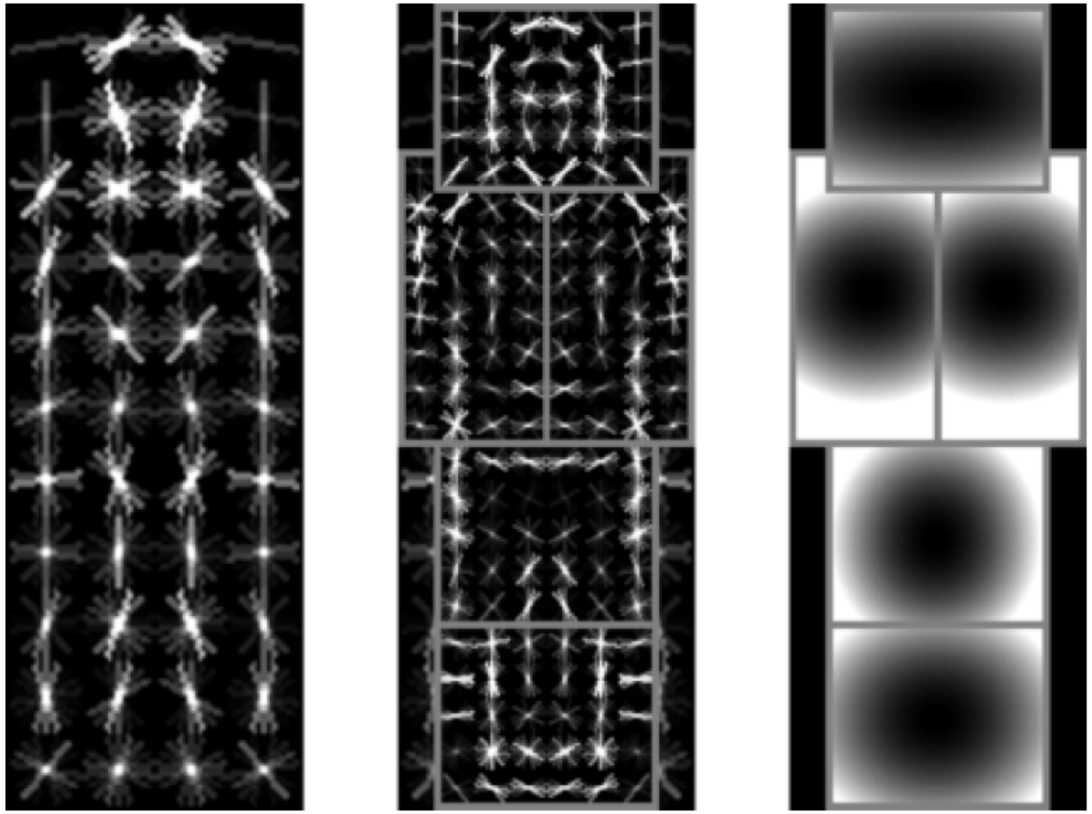


2008 DPM (Deformable parts model)



Object Detection with Discriminatively Trained Part Based Model,
Felzenszwalb, Girshick, McAllester and Ramanan, *PAMI*, 2010

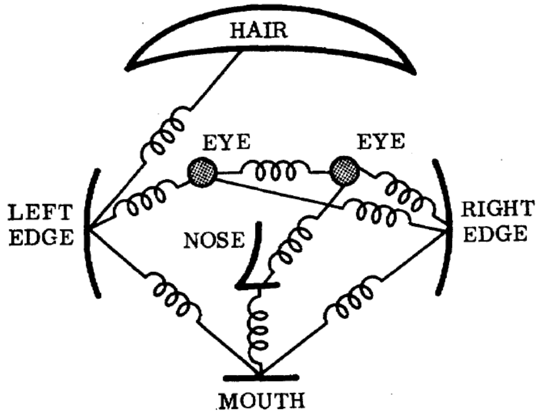
2008 DPM (Deformable parts model)

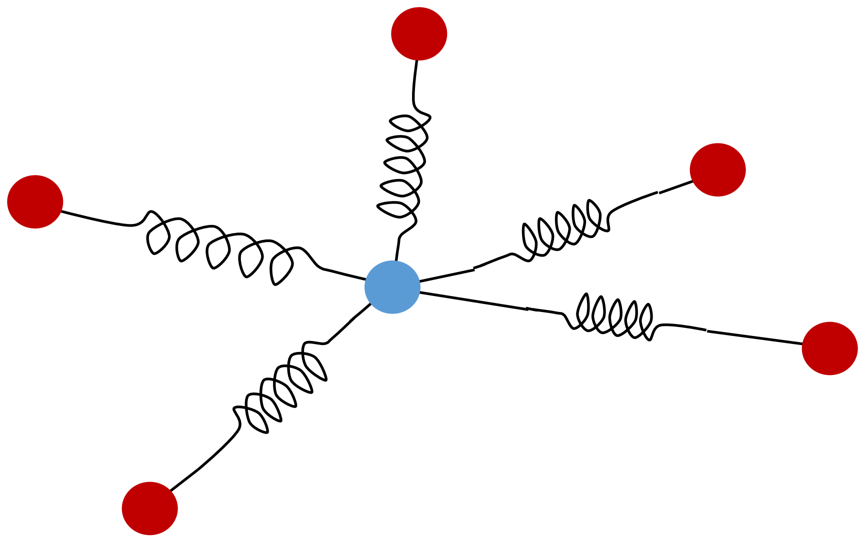


Object Detection with Discriminatively Trained Part Based Model,
Felzenszwalb, Girshick, McAllester and Ramanan, *PAMI*, 2010

Star-structure

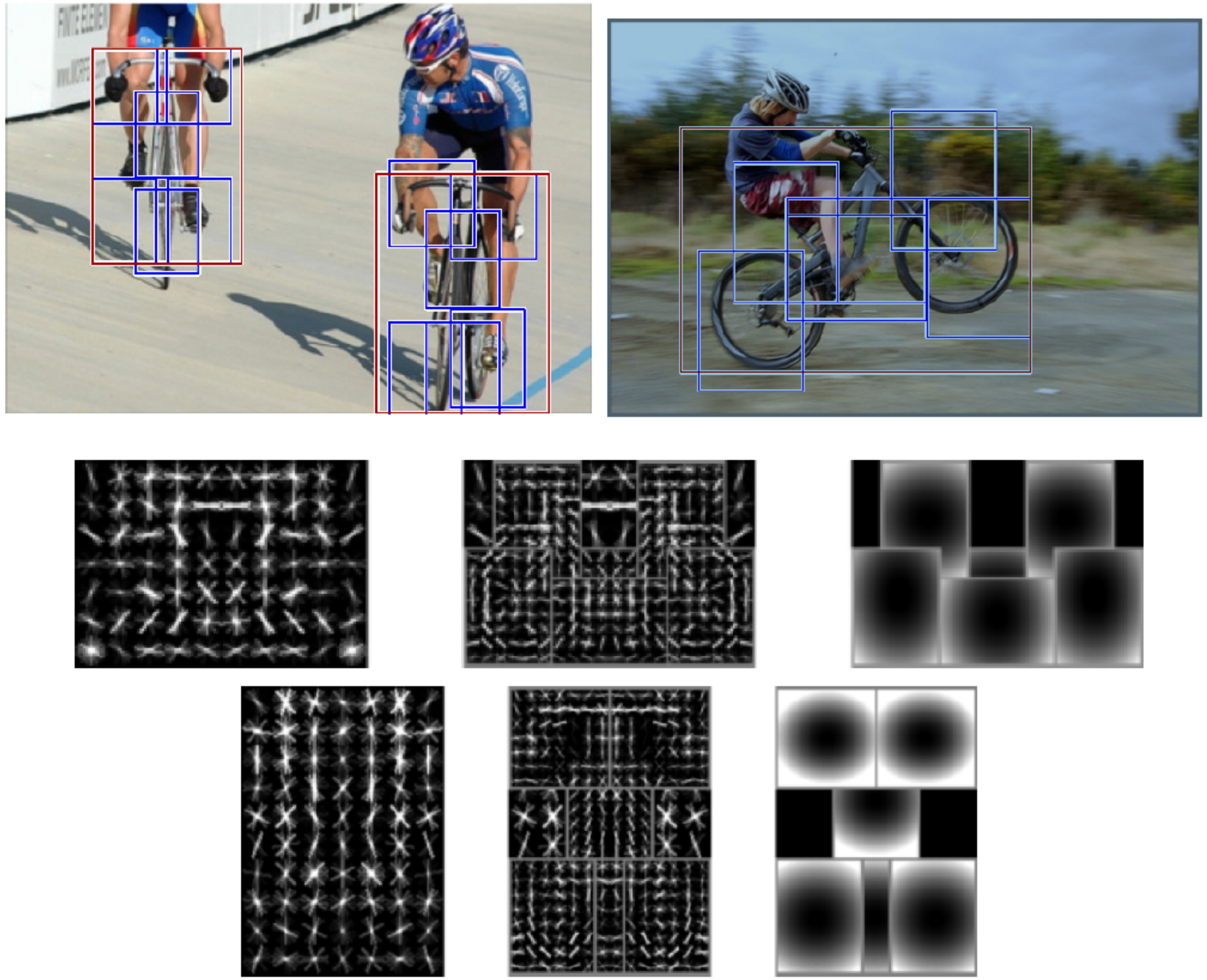
- Computationally efficient (distance transform)





Distance transforms of sampled functions, Felzenszwalb and Huttenlocher, Cornell University CIS, Tech. Rep. 2004.

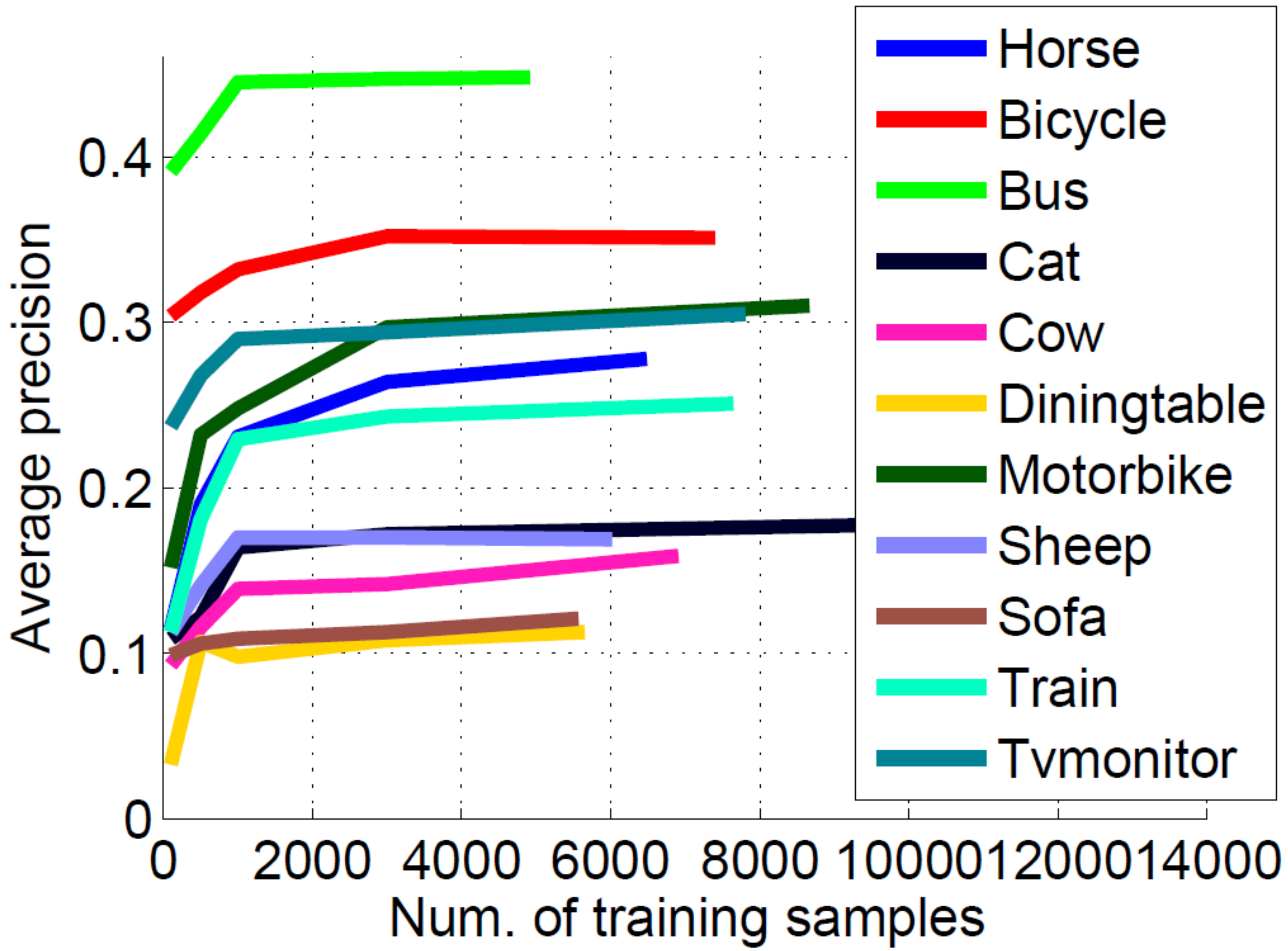
Multiple components



Why it worked

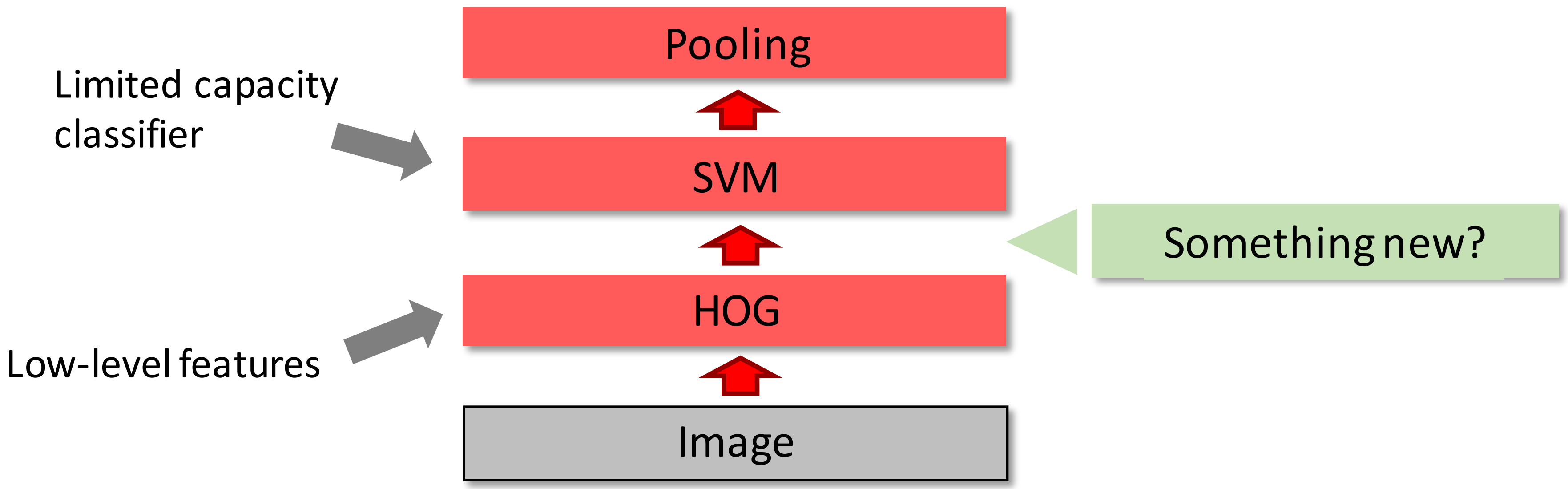
- Multiple components
- Deformable parts?
- Hard negative mining
- Good balance

"How important are 'Deformable Parts' in the Deformable Parts Model?“,
Divvala, Efros, and Hebert, *Parts and Attributes Workshop, ECCV,* 2012



Do We Need More Training Data or Better Models for Object Detection?
Zhu, Vondrick, Ramanan, Fowlkes, *BMVC* 2012.

DPM


Limited capacity classifier
Low-level features
Pooling
SVM
HOG
Image
Something new?

Problems with Visual Categories

- A lot of categories are functional

Char

- World is too varied

- Categories are 3D, but images are 2D

car

IMAGENET

www.image-net.org

22K categories and **14M** images

- Animals
  - Bird
  - Fish
  - Mammal
  - Invertebrate
- Plants
  - Tree
  - Flower
- Food
- Materials
- Structures
- Artifact
  - Tools
  - Appliances
  - Structures
- Person
- Scenes
  - Indoor
  - Geological Formations
- Sport Activities

Deng, Dong, Socher, Li, Li, & Fei-Fei, 2009

2009 ImageNet

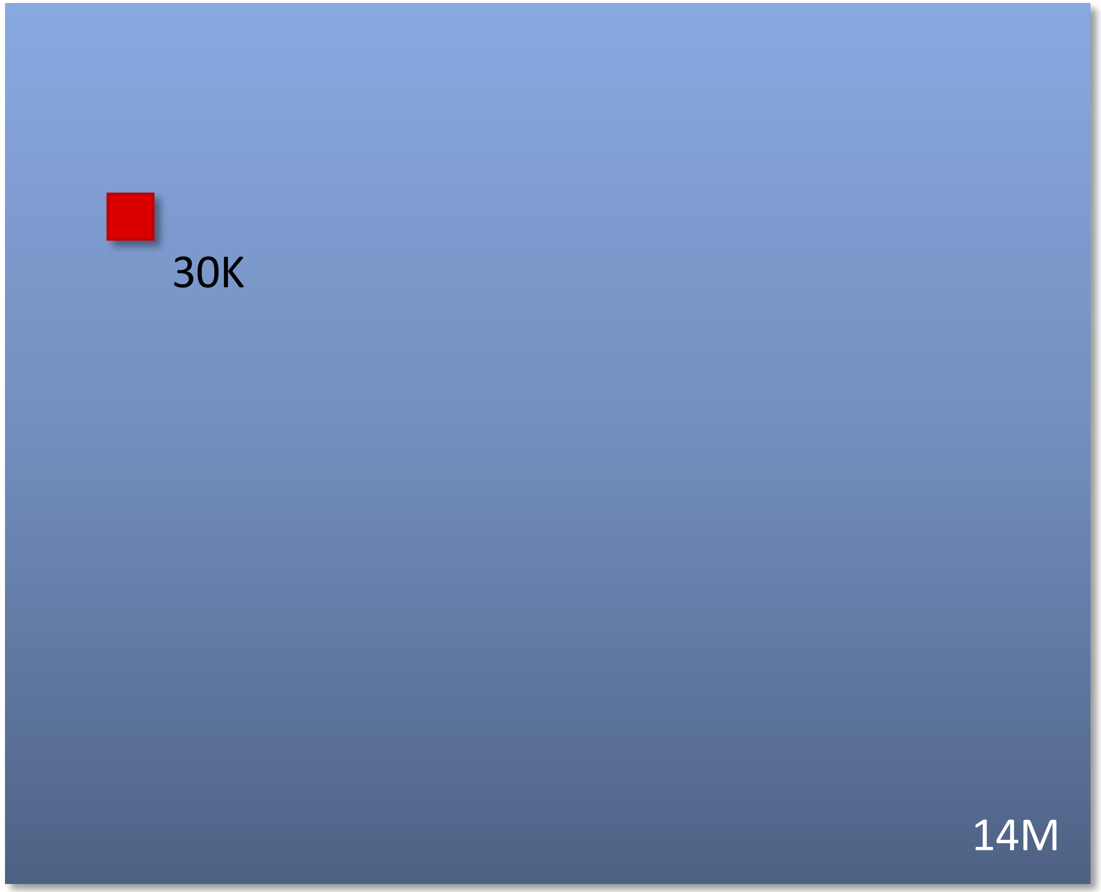
22K categories, 14M images

ImageNet: A Large-Scale Hierarchical Image Database,
Deng, Dong, Socher, Li, Li and Fei-Fei, *CVPR,* 2009

Images

2009

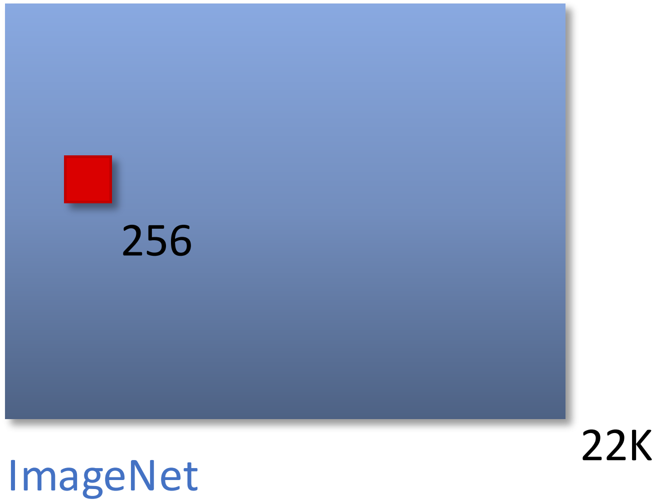
2012



ImageNet

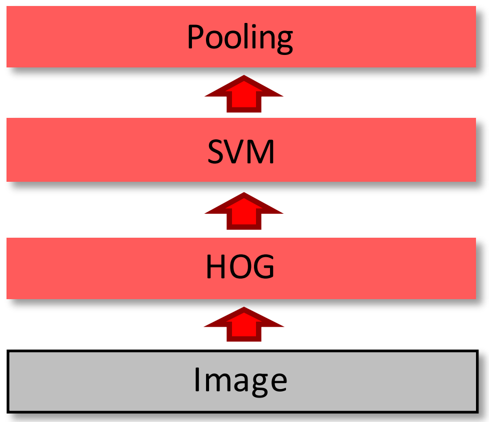
Categories

2009 2012

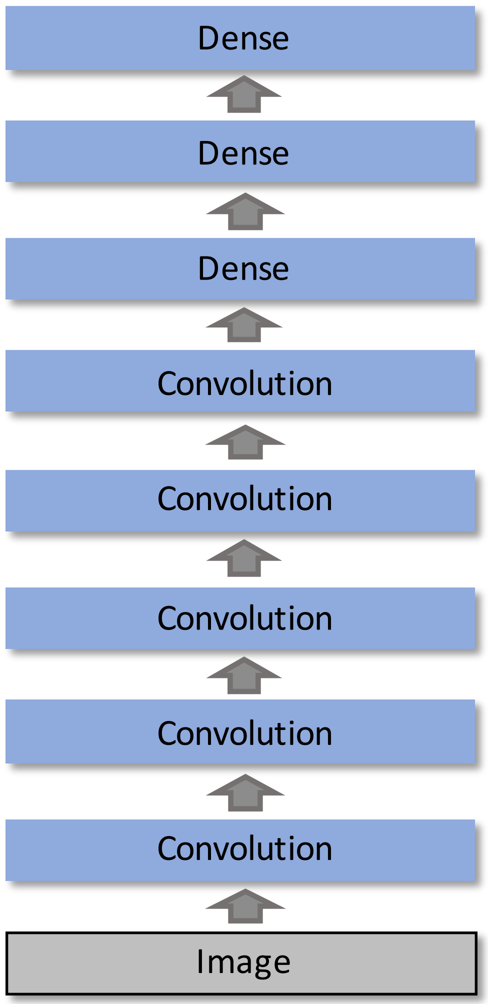


Algorithms

2009



2012



The New York Times

Seeking a Better Way to Find Web Images

By JOHN MARKOFF NOV. 19, 2012

STANFORD, Calif. — You may think you can find almost anything on the Internet.

But even as images and video rapidly come to dominate the Web, search engines can ordinarily find a given image only if the text entered by a searcher matches the text with which it was labeled. And the labels can be unreliable, unhelpful ("fuzzy" instead of "rabbit") or simply nonexistent.

To eliminate those limits, scientists will need to create a new generation of visual search technologies — or else, as the Stanford computer scientist Fei-Fei Li recently put it, the Web will be in danger of "going dark."

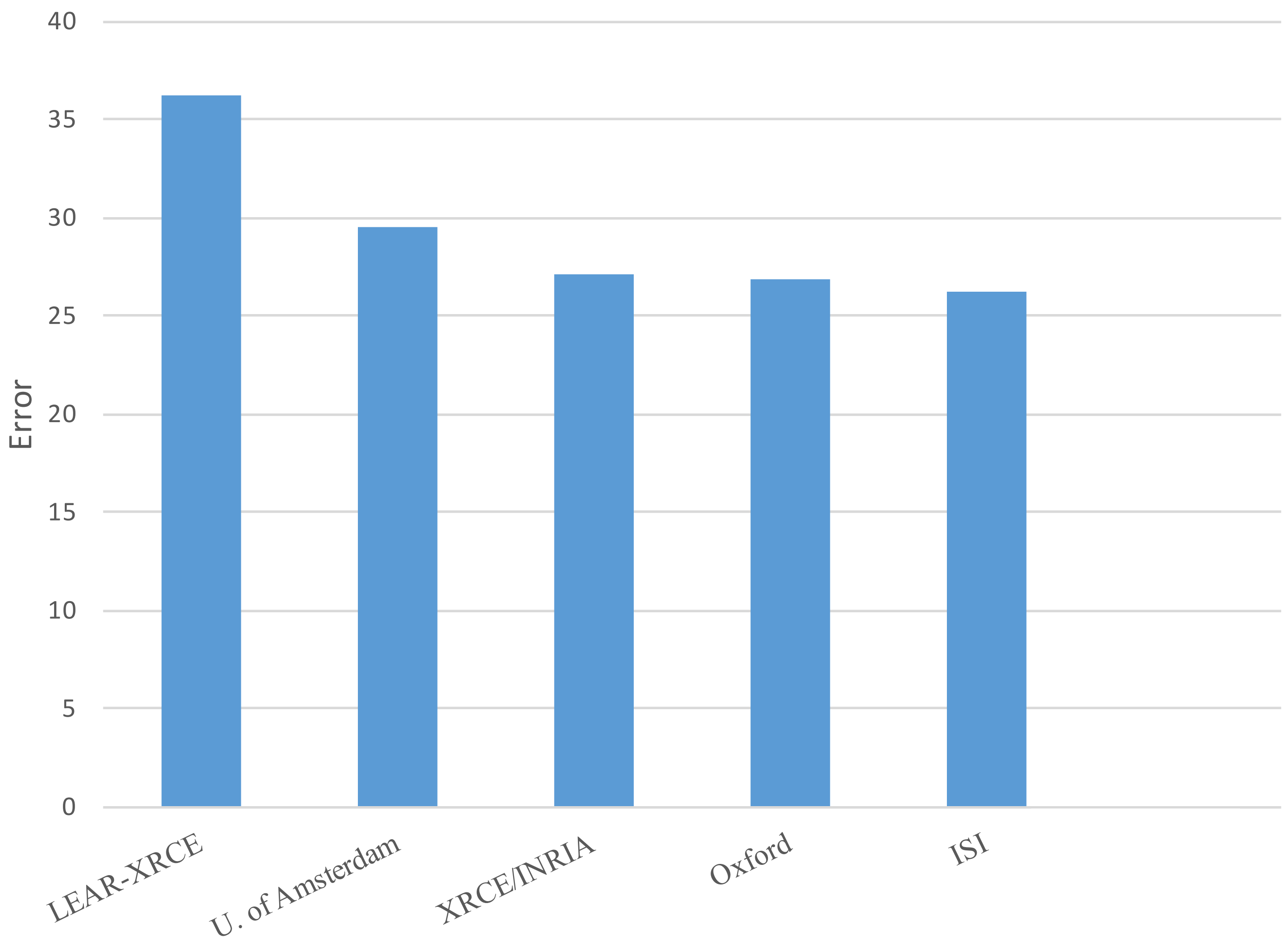
Now, along with computer scientists from Princeton, Dr. Li, 36, has built the world's largest visual database in an effort to mimic the human vision system. With more than 14 million labeled objects, from obsidian to orangutans to ocelots, the database has because a vital resource for computer vision researchers.

The labels were created by humans. But now machines can learn from the vast database to recognize similar, unlabeled objects, making possible a striking increase in recognition accuracy.

This summer, for example, two Google computer scientists, Andrew Y. Ng and Jeff Dean, tested the new system, known as ImageNet, on a huge collection of labeled photos.

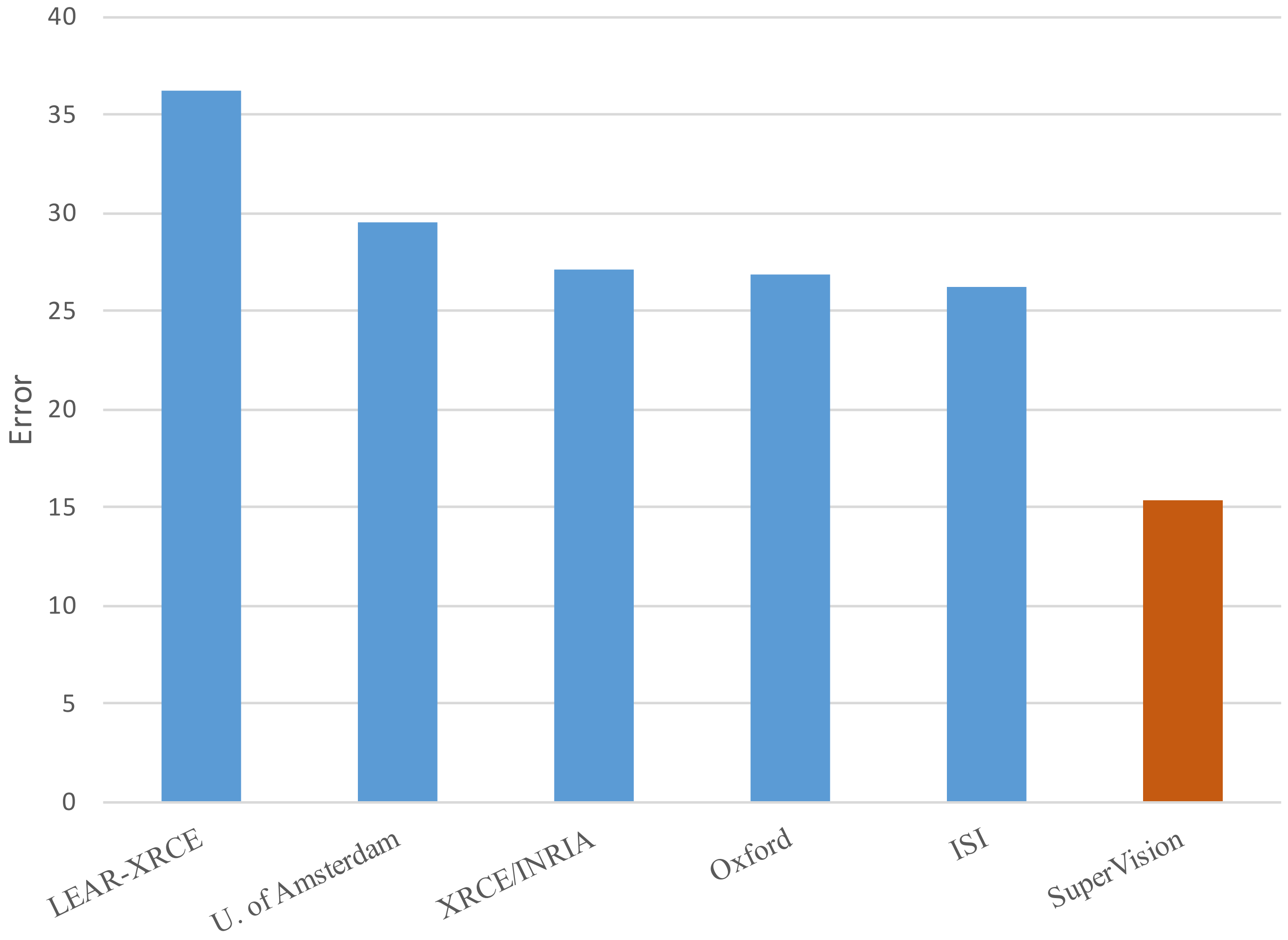
The system performed almost twice as well as previous "neural network" algorithms — software models that seek to replicate human brain functions.

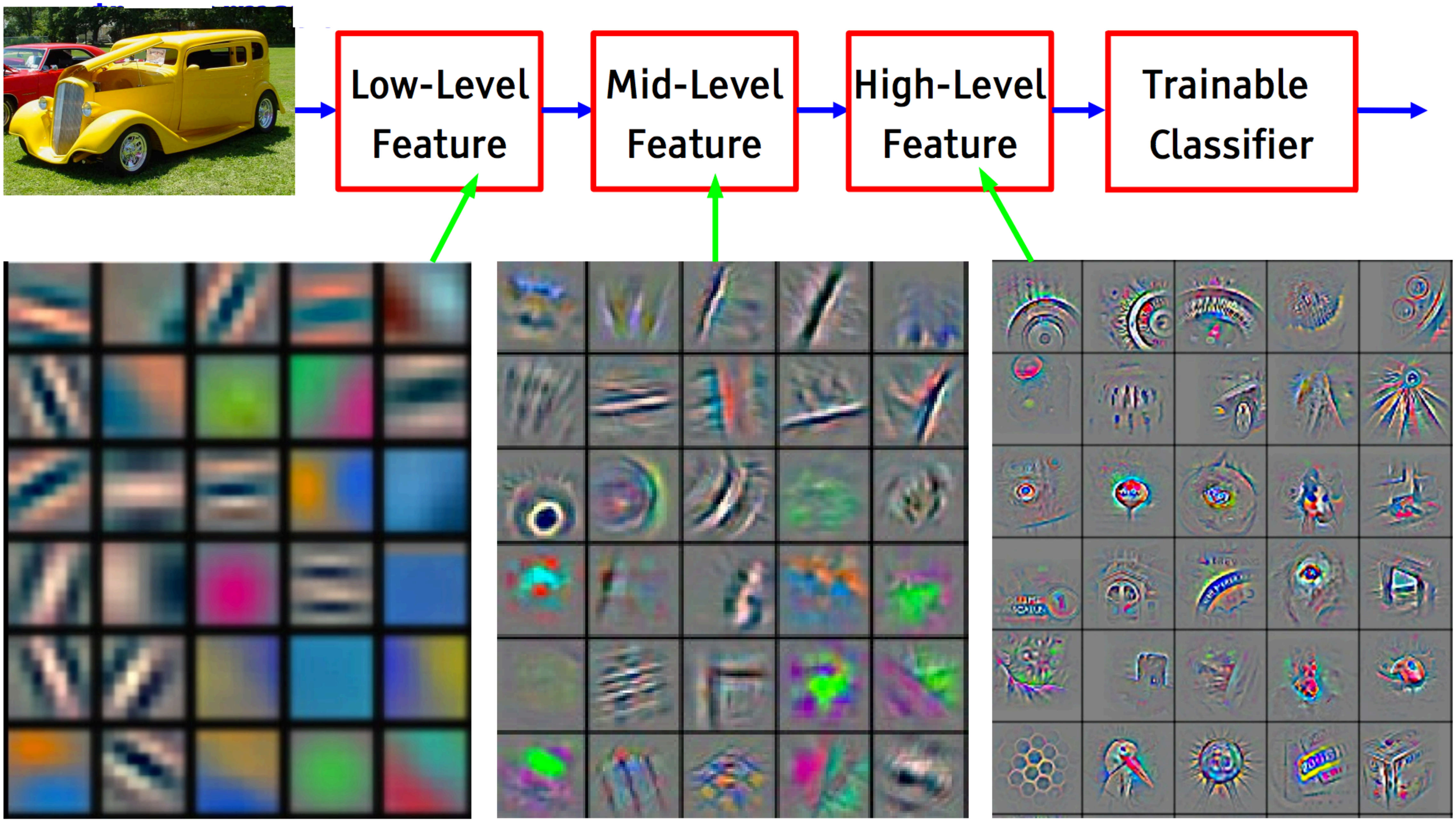
2012 ImageNet 1K
(Fall 2012)



2012 ImageNet 1K

(Fall 2012)



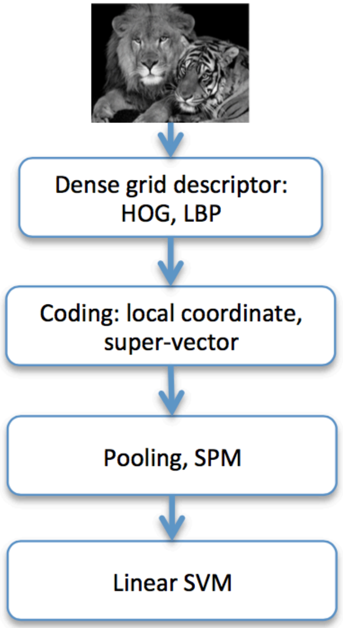


Feature visualization of convolutional net trained on ImageNet from [Zeiler & Fergus 2013]

IMAGENET Large Scale Visual Recognition Challenge

Year 2010

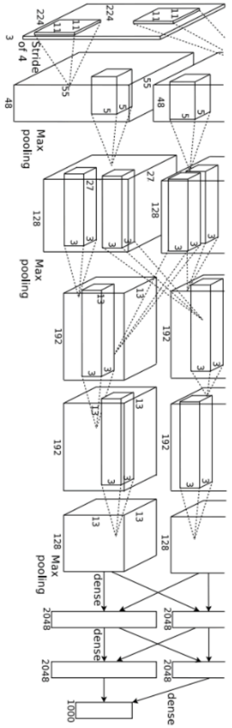
NEC-UIUC



[Lin CVPR 2011]

Year 2012

SuperVision



[Krizhevsky NIPS 2012]

Year 2014

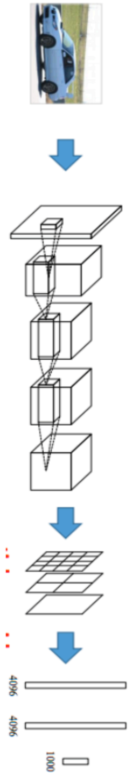
GoogLeNet

[Szegedy arxiv 2014]

VGG

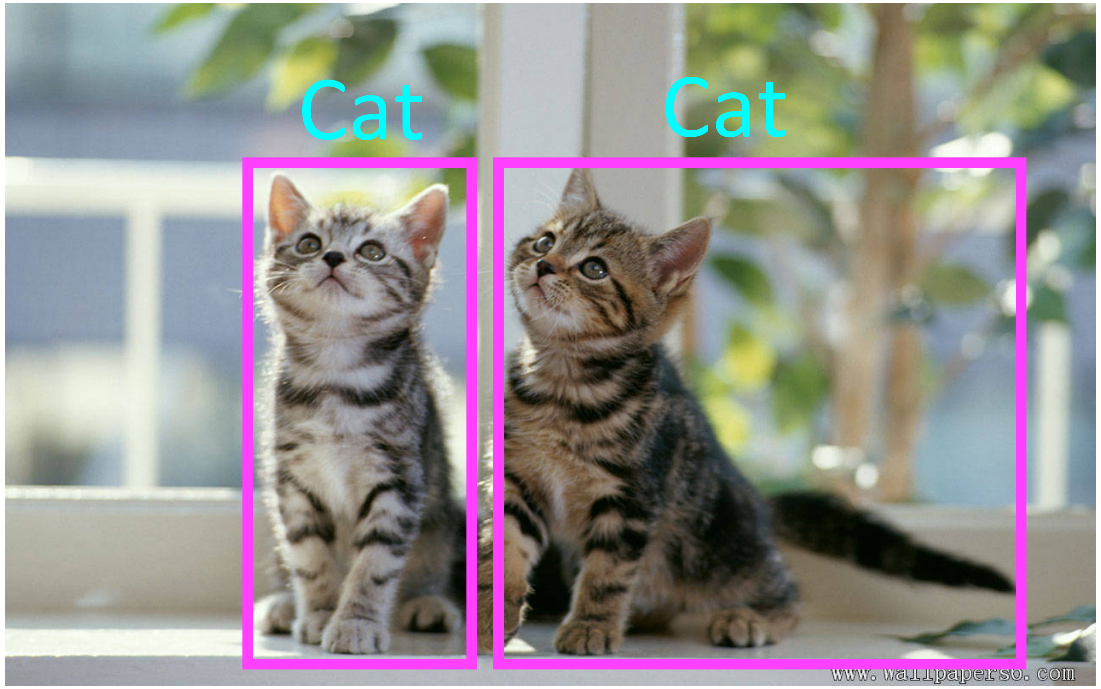
[Simonyan arxiv 2014]

MSRA



[He arxiv 2014]

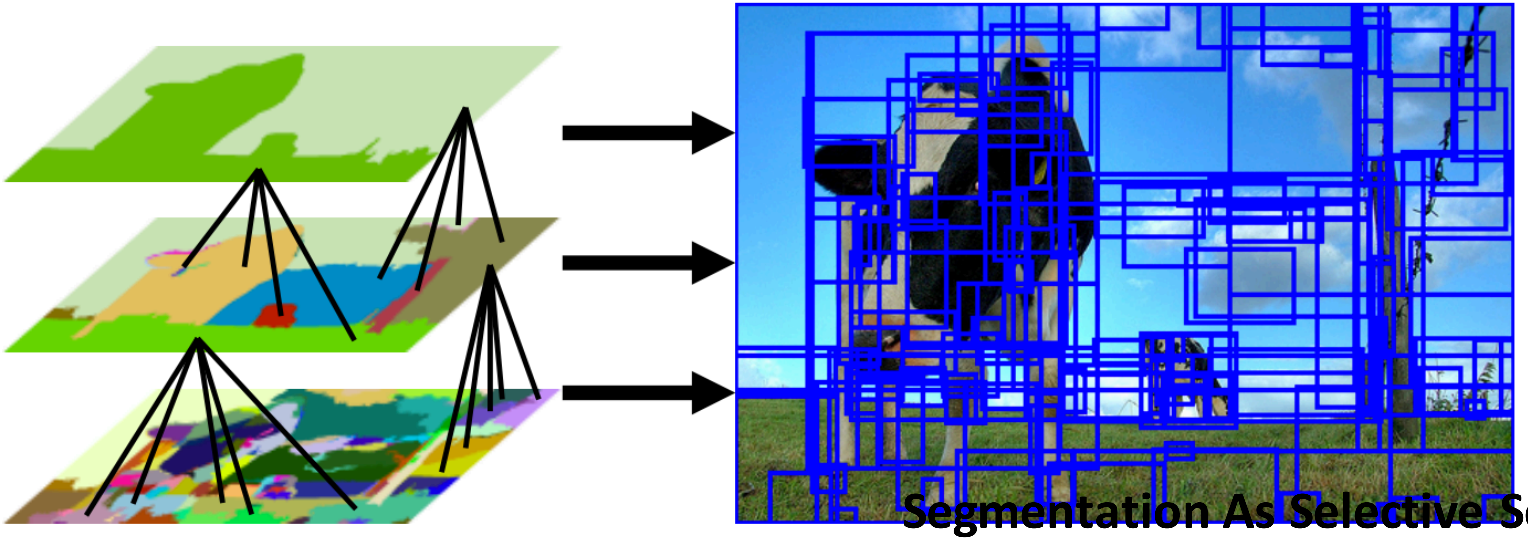
Classification Vs. Detection



Object Proposals

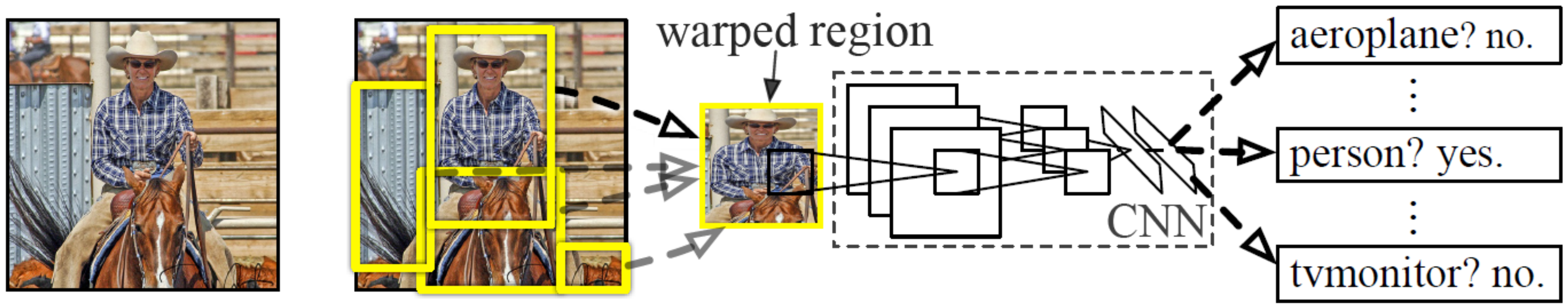
Ground truth

Object hypotheses



Segmentation As Selective Search for Object Recognition,
van de Sande, Uijlings, Gevers, Smeulders, ICCV 2011

Object detection



1. Input image

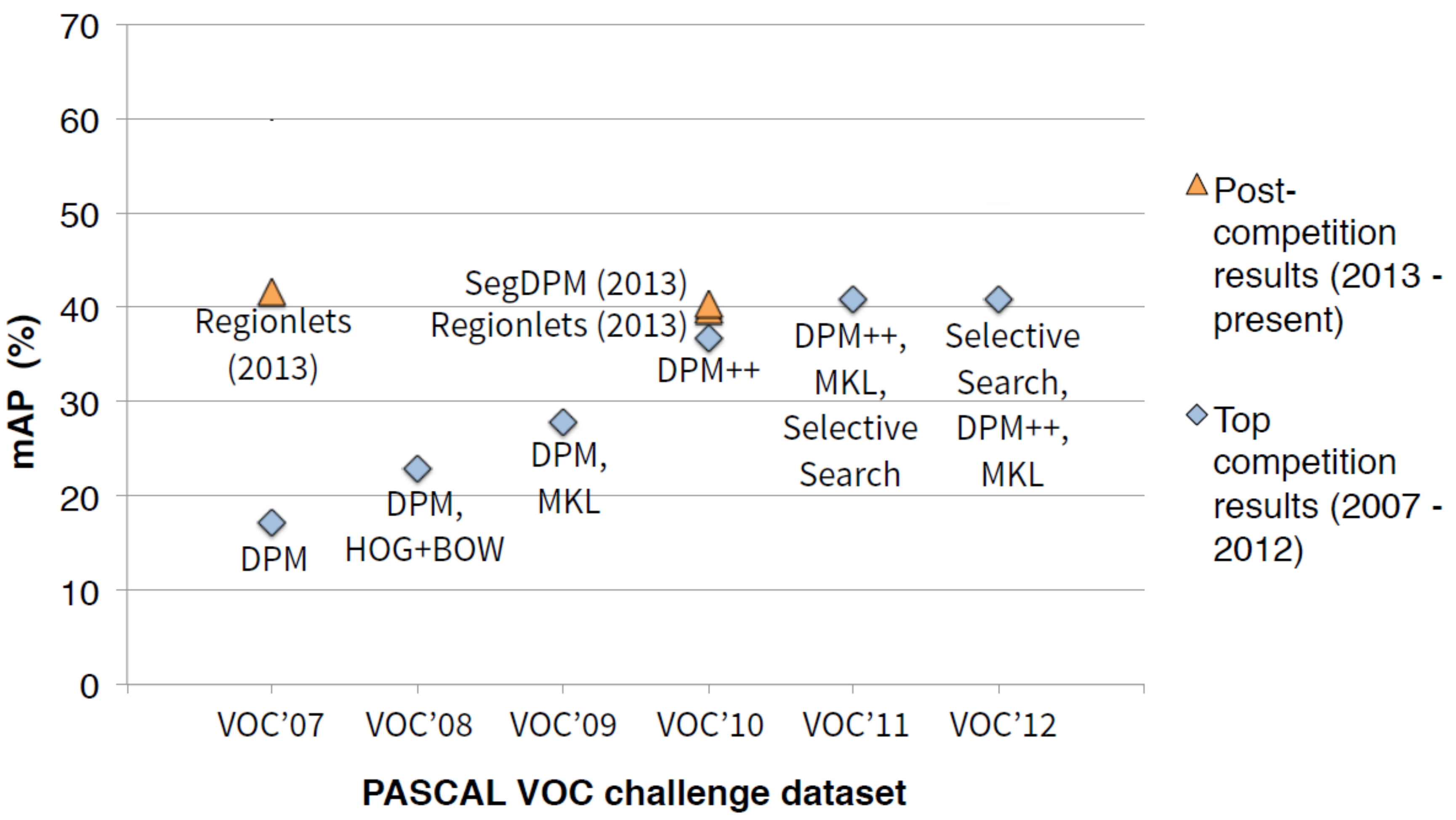
2. Extract region proposals (~2k)

3. Compute CNN features

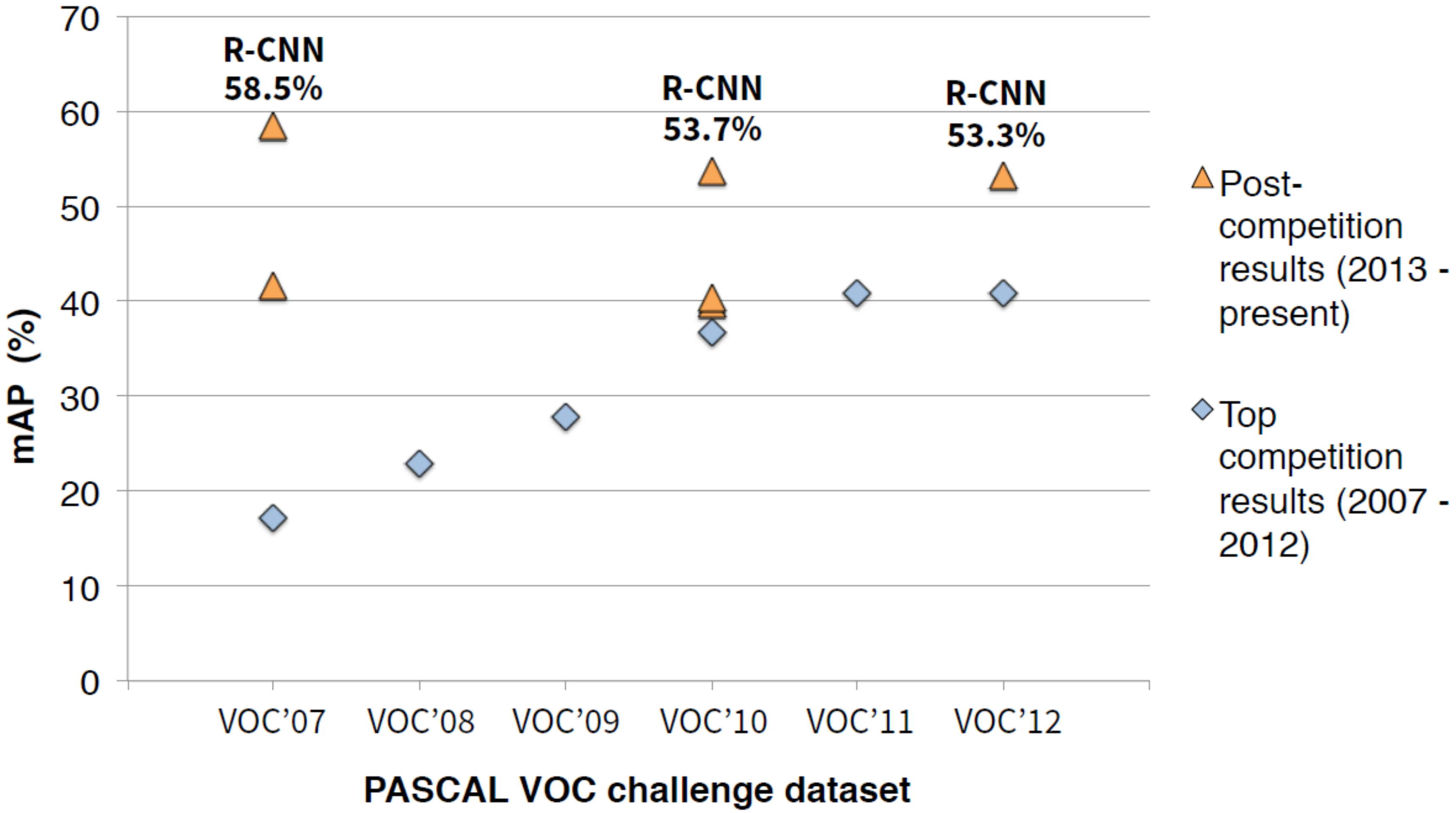
4. Classify regions

Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation, Girshick, Donahue, Darrell, Malik, *CVPR* 2014.

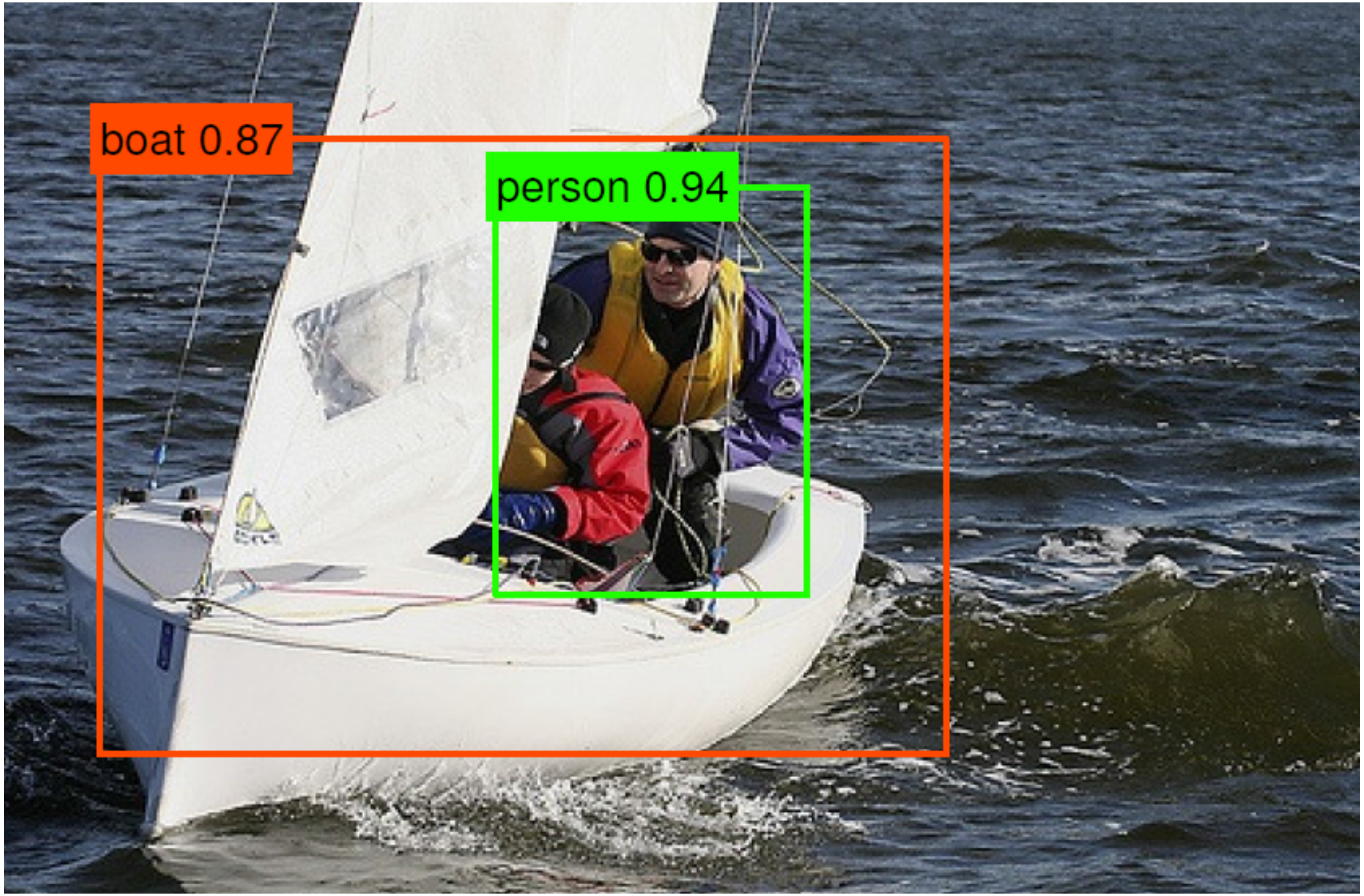
Online classification demo:
http://decaf.berkeleyvision.org/



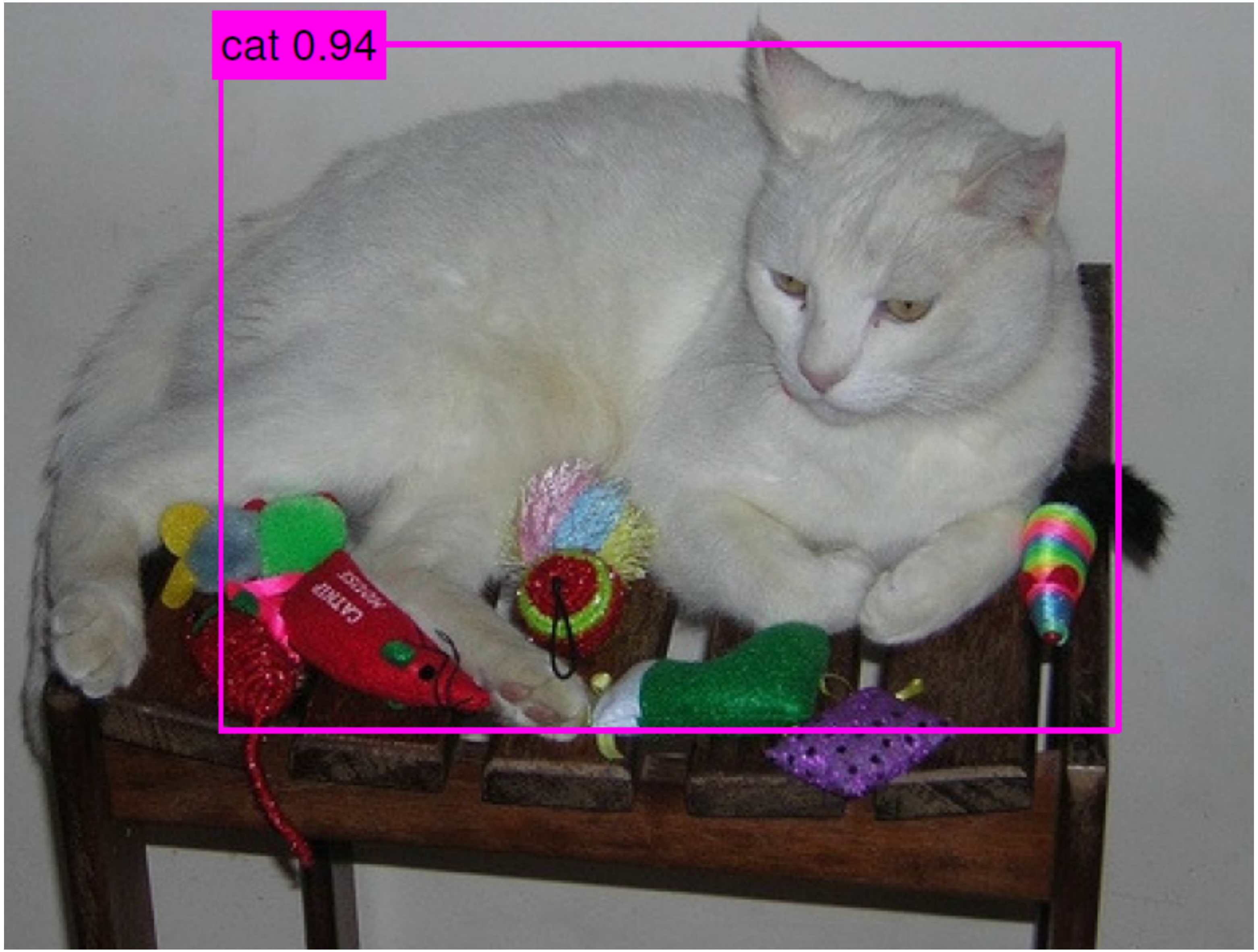
70
60
50
40
30
20
10
0
mAP (%)
Regionlets (2013)
DPM
DPM, HOG+BOW
DPM, MKL
SegDPM (2013)
Regionlets (2013)
DPM++
DPM++, MKL, Selective Search
Selective Search, DPM++, MKL
VOC'07
VOC'08
VOC'09
VOC'10
VOC'11
VOC'12
PASCAL VOC challenge dataset
Post-competition results (2013 - present)
Top competition results (2007 - 2012)



R-CNN
58.5%
R-CNN
53.7%
R-CNN
53.3%
70
60
50
40
30
20
10
0
mAP (%)
VOC'07
VOC'08
VOC'09
VOC'10
VOC'11
VOC'12
PASCAL VOC challenge dataset
Post-competition results (2013 - present)
Top competition results (2007 - 2012)


boat 0.87
person 0.94

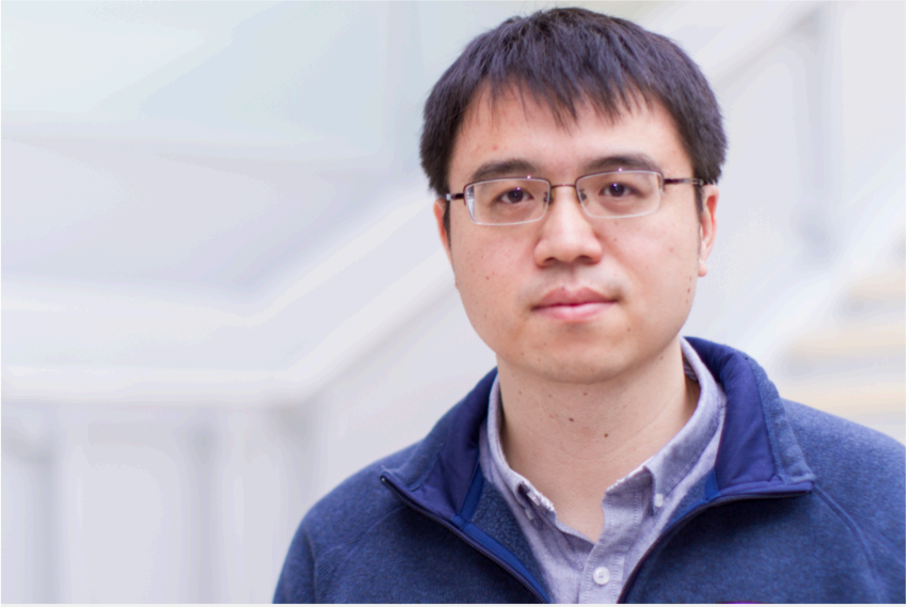

bottle 0.86
bottle 0.98
bottle 0.82


cat 0.94


cat 0.95


cat 0.82
http://phyllischan.blogspot.com

Microsoft researchers win ImageNet computer vision challenge



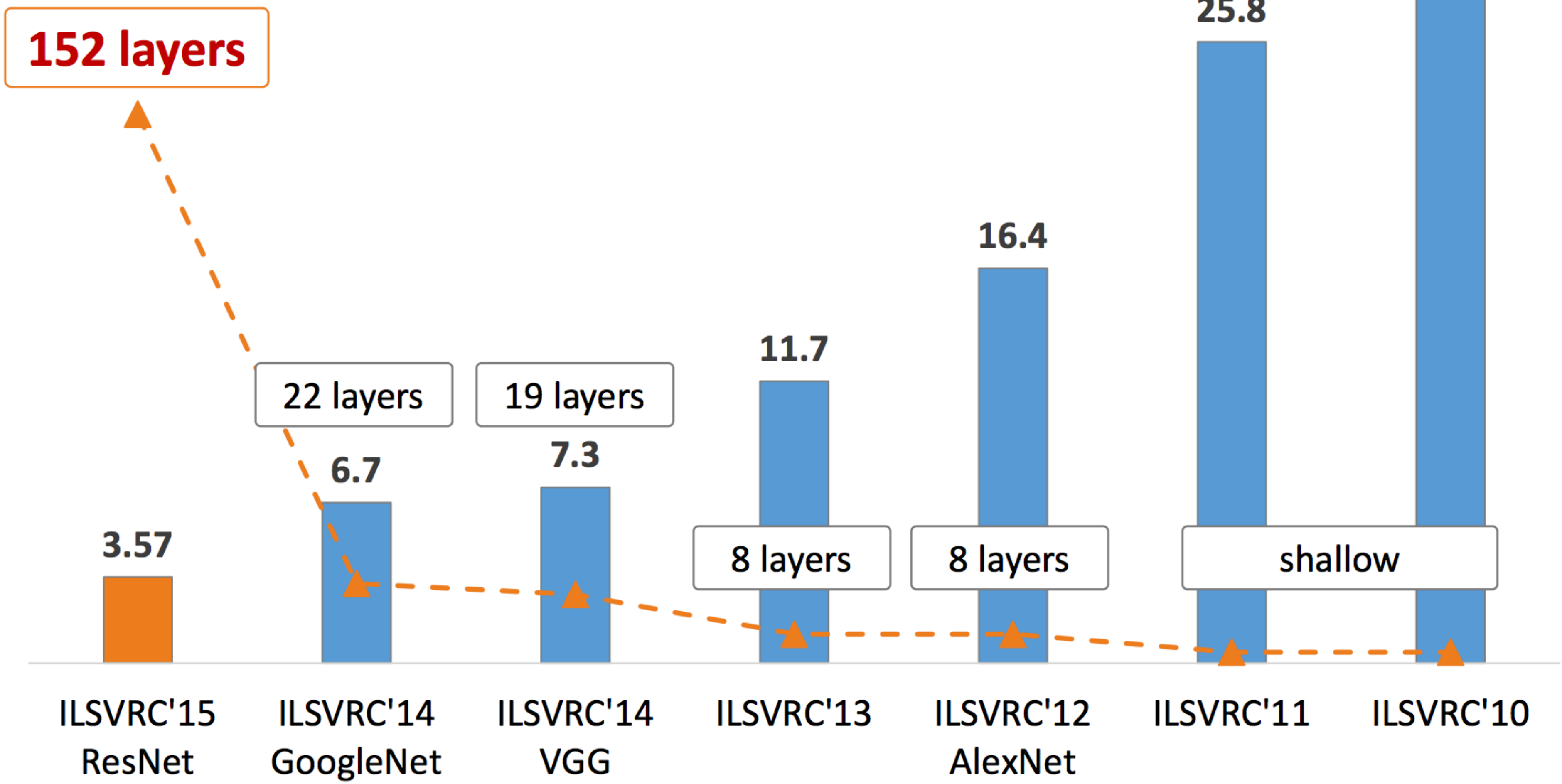
Jian Sun, a principal research manager at Microsoft Research, led the image understanding project. Photo: Craig Tuschhoff/Microsoft.

Posted December 10, 2015 By **Allison Linn**

Microsoft researchers on Thursday announced a major advance in technology designed to identify the objects in a photograph or video, showcasing a system whose accuracy meets and sometimes exceeds human-level performance.

Microsoft's new approach to recognizing images also took first place in several major categories of image recognition challenges Thursday, beating out many other competitors from academic, corporate and research institutions in the ImageNet and Microsoft Common Objects in Context challenges.

Revolution of Depth



ImageNet Classification top-5 error (%)

Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. "Deep Residual Learning for Image Recognition". arXiv 2015.

Microsoft Research

Revolution of Depth

AlexNet, 8 layers (ILSVRC 2012)

11x11 conv, 96, /4, pool/2
↓
5x5 conv, 256, pool/2
↓
3x3 conv, 384
↓
3x3 conv, 384
↓
3x3 conv, 256, pool/2
↓
fc, 4096
↓
fc, 4096
↓
fc, 1000

VGG, 19 layers (ILSVRC 2014)

3x3 conv, 64
↓
3x3 conv, 64, pool/2
↓
3x3 conv, 128
↓
3x3 conv, 128, pool/2
↓
3x3 conv, 256
↓
3x3 conv, 256
↓
3x3 conv, 256
↓
3x3 conv, 256, pool/2
↓
3x3 conv, 512
↓
3x3 conv, 512
↓
3x3 conv, 512
↓
3x3 conv, 512, pool/2
↓
3x3 conv, 512
↓
3x3 conv, 512
↓
3x3 conv, 512
↓
3x3 conv, 512, pool/2
↓
fc, 4096
↓
fc, 4096
↓
fc, 1000

GoogleNet, 22 layers (ILSVRC 2014)

ICCV15 International Conference on Computer Vision

Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. “Deep Residual Learning for Image Recognition”. arXiv 2015.

Microsoft Research

Revolution of Depth

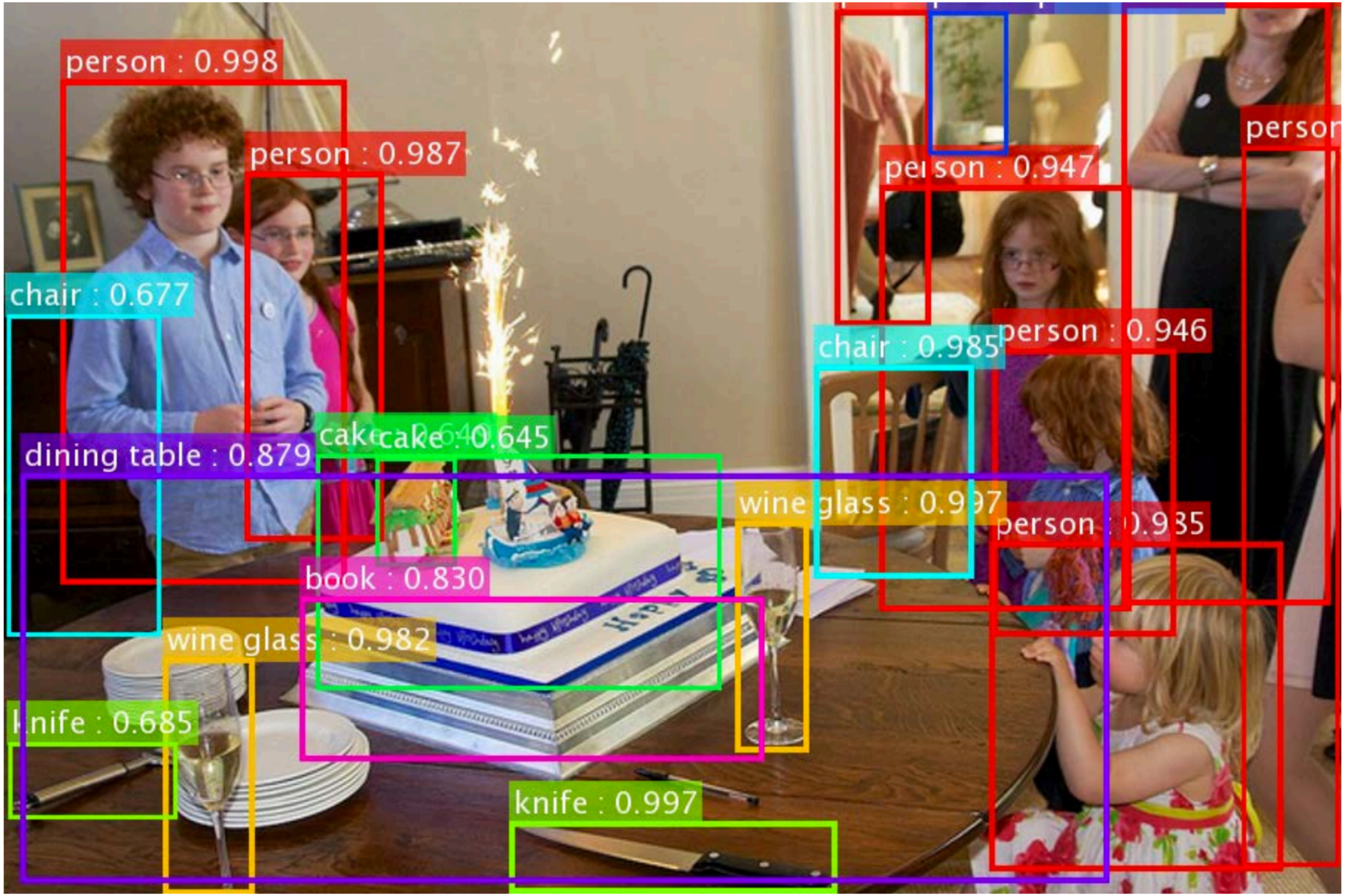
AlexNet, 8 layers
(ILSVRC 2012)

VGG, 19 layers
(ILSVRC 2014)

ResNet, 152 layers
(ILSVRC 2015)

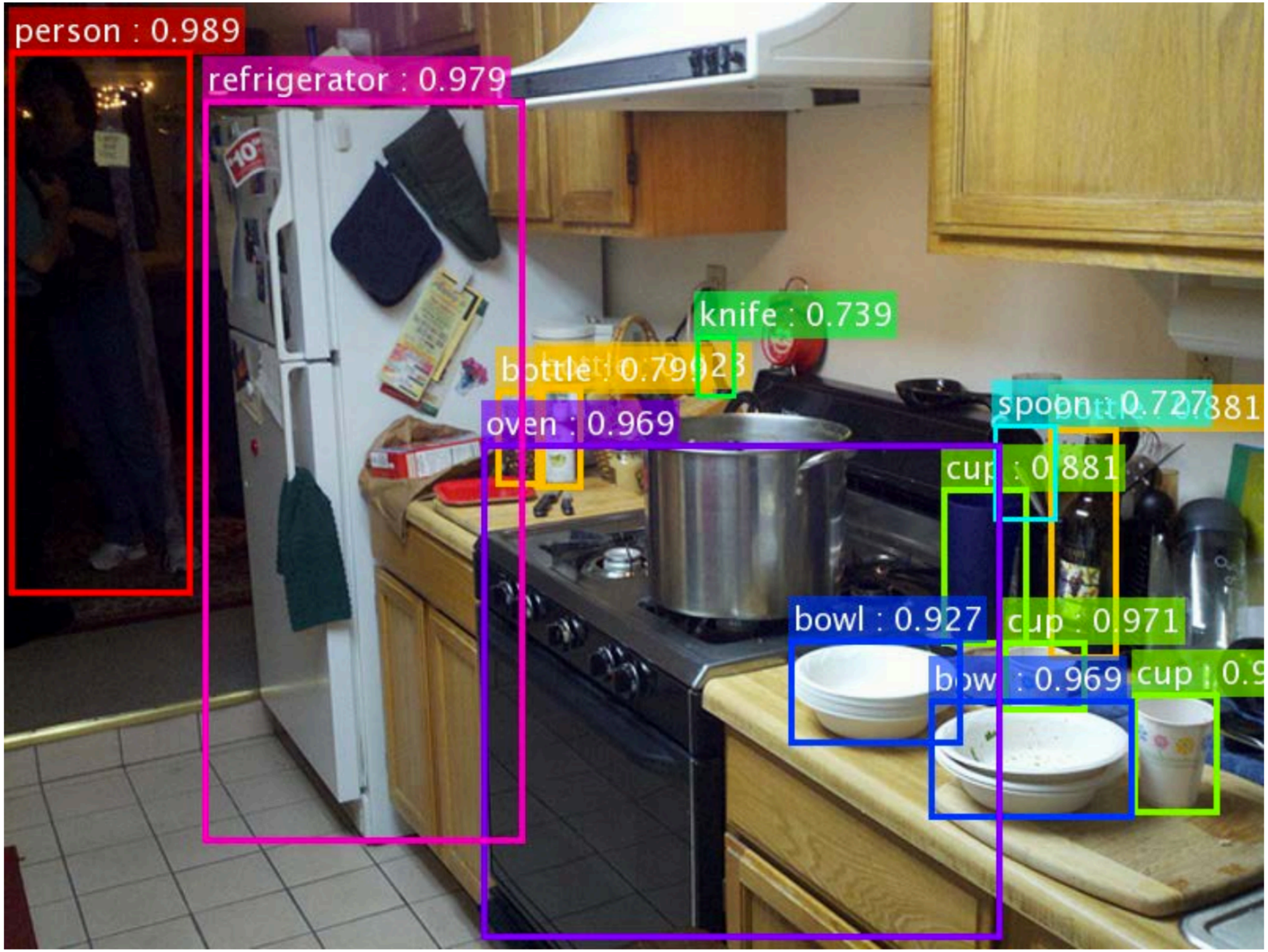
ICCV15 International Conference on Computer Vision

Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. “Deep Residual Learning for Image Recognition”. arXiv 2015.



*the original image is from the COCO dataset

Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. “Deep Residual Learning for Image Recognition”. arXiv 2015.
Shaoqing Ren, Kaiming He, Ross Girshick, & Jian Sun. “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”. NIPS 2015.

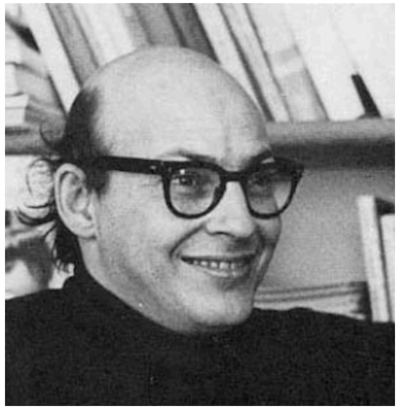


*the original image is from the COCO dataset

Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. “Deep Residual Learning for Image Recognition”. arXiv 2015.
Shaoqing Ren, Kaiming He, Ross Girshick, & Jian Sun. “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”. NIPS 2015.

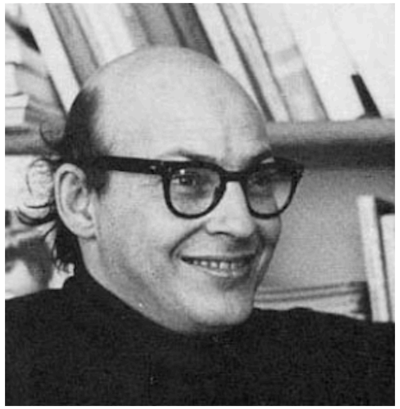
Going beyond categorization…

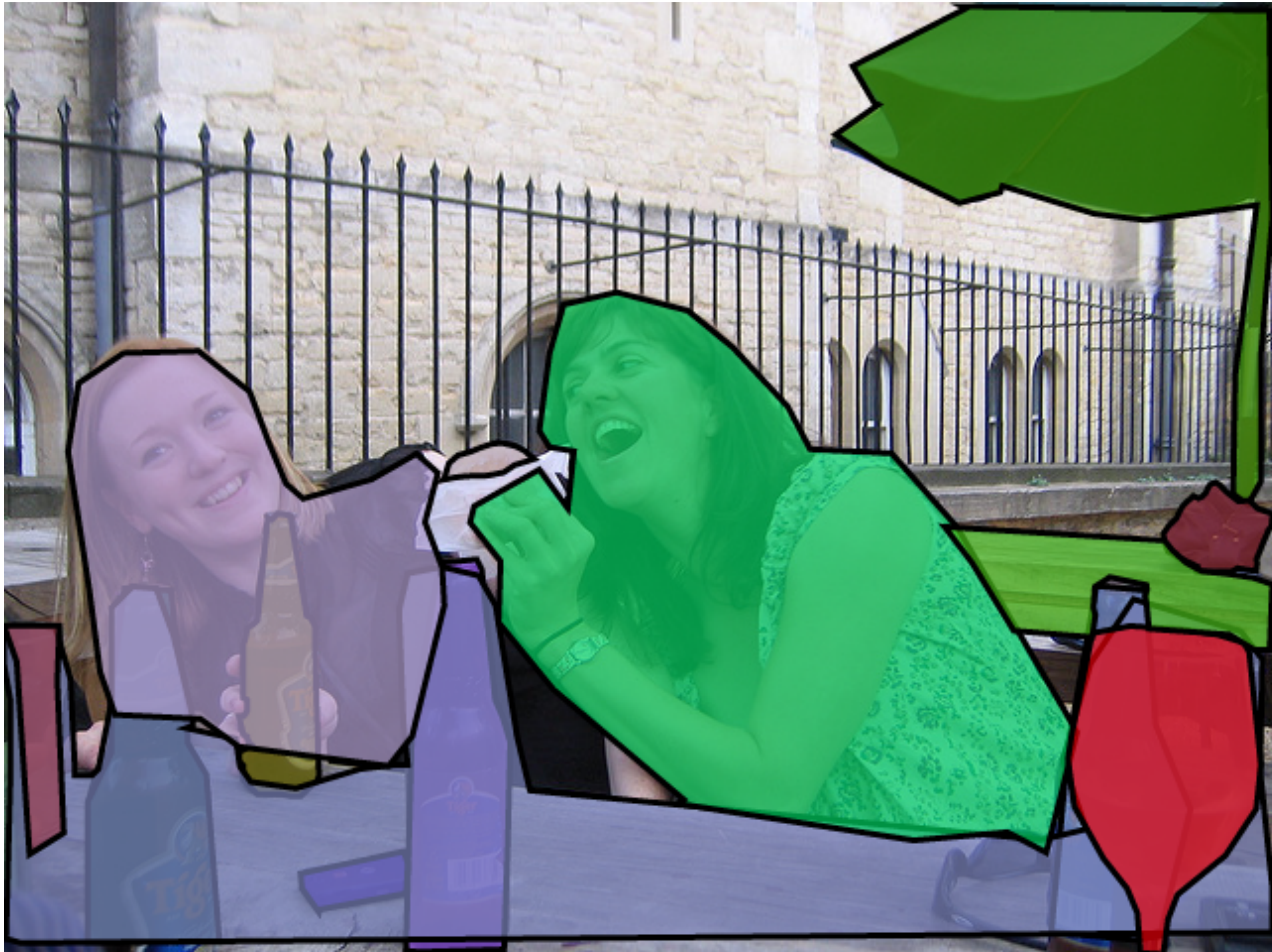
"Connect a television camera to a computer and **get the machine to describe what it sees.**"



Going beyond categorization…

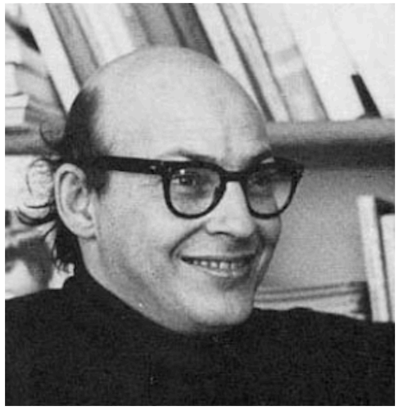
“Connect a television camera to a computer and **get the machine to describe what it sees.**”





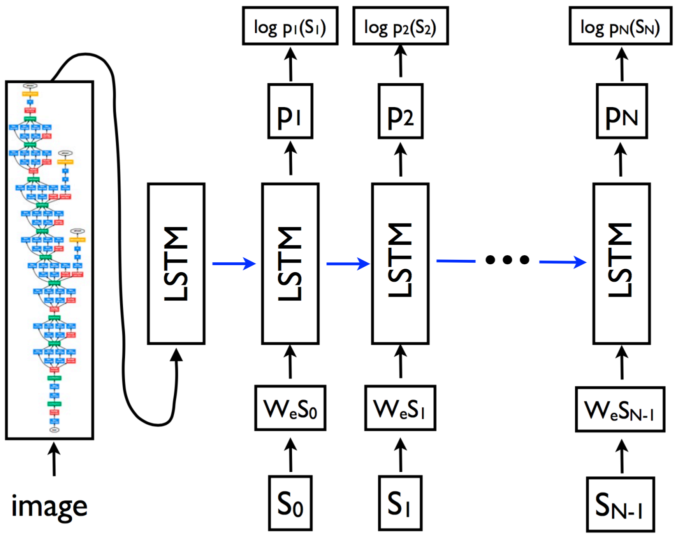
Going beyond categorization…

“Connect a television camera to a computer and **get the machine to describe what it sees.**”



two girls sitting at a table smiling and eating and drinking.
a woman is eating a doughnut and drinking beer.
there are two woman drinking beers and eating food
a woman leaning into another woman as she holds a sandwich towards her.
two ladies are enjoying beer and treats at the table.

Going beyond categorization…



The New York Times

Researchers Announce Advance in Image-Recognition Software

By JOHN MARKOFF NOV. 17, 2014

MOUNTAIN VIEW, Calif. — Two groups of scientists, working independently, have created artificial intelligence software capable of recognizing and describing the content of photographs and videos with far greater accuracy than ever before, sometimes even mimicking human levels of understanding.

Until now, so-called computer vision has largely been limited to recognizing individual objects. The new software, described on Monday by researchers at Google and at Stanford University, teaches itself to identify entire scenes: a group of young men playing Frisbee, for example, or a herd of elephants marching on a grassy plain.

The software then writes a caption in English describing the picture. Compared with human observations, the researchers found, the computer-written descriptions are surprisingly accurate.

The advances may make it possible to better catalog and search for the billions of images and hours of video available online, which are often poorly described and archived. At the moment, search engines like Google rely largely on written language accompanying an image or video to ascertain what it contains.

Captioned by Human and by Google's Experimental Program

Human: "A group of men playing Frisbee in the park."
Computer model: "A group of young people playing a game of Frisbee."

1 of 6

…the “giraffe-tree” problem ☹

a giraffe next to a tree

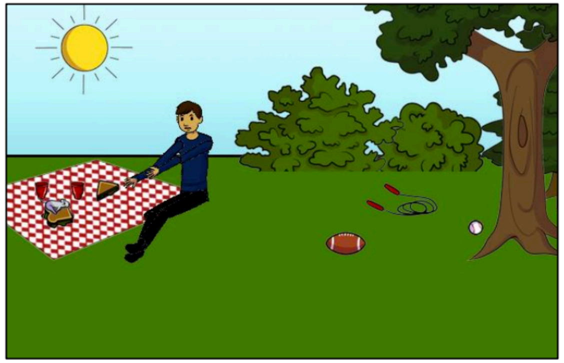
VQA: Visual Question Answering

www.visualqa.org

Stanislaw Antol*, Aishwarya Agrawal*, Jiasen Lu, Margaret Mitchell,
Dhruv Batra, C. Lawrence Zitnick, Devi Parikh

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?



Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?

What is Visual Question Answering (VQA)?

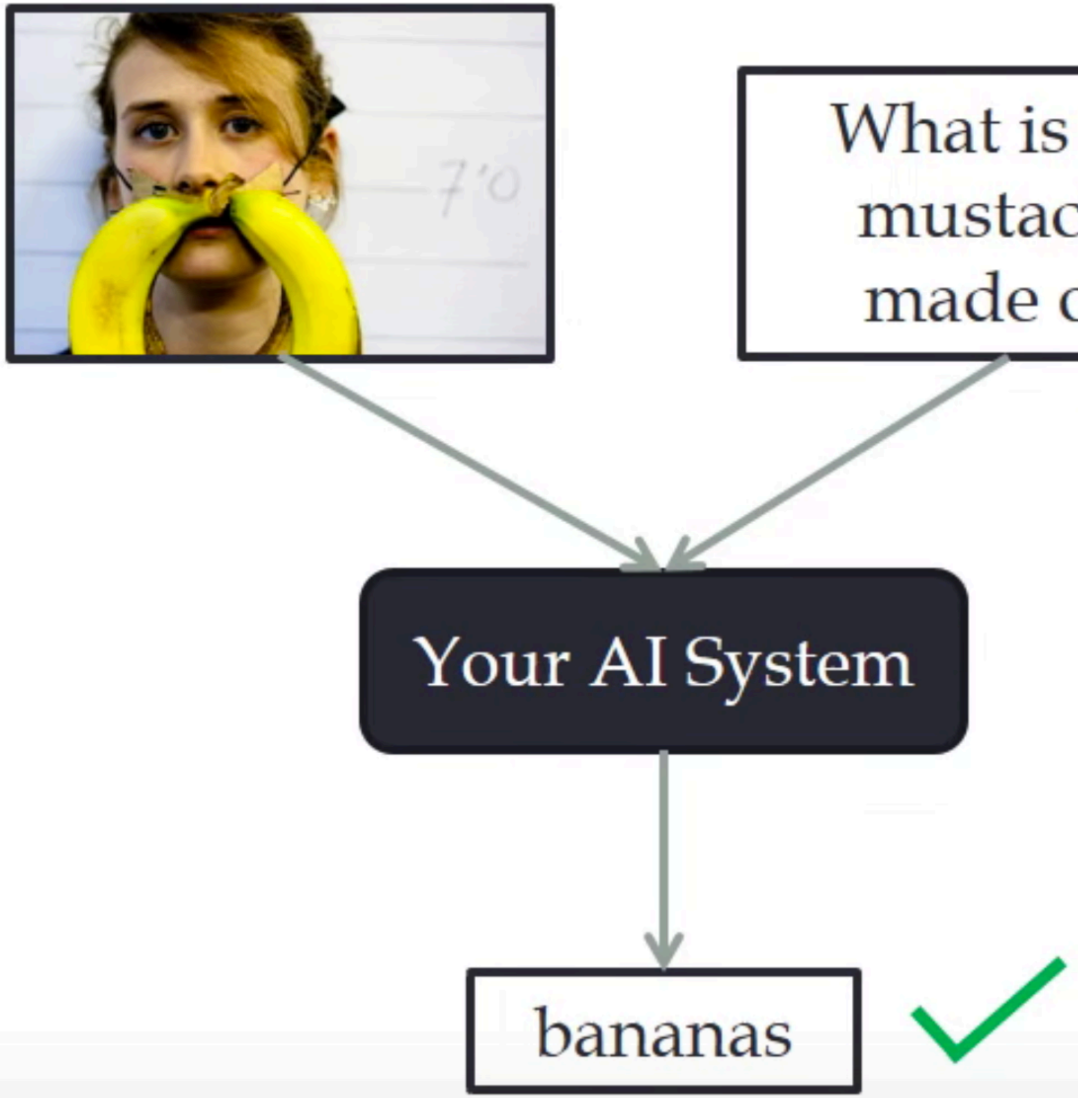
Ask any question about this image

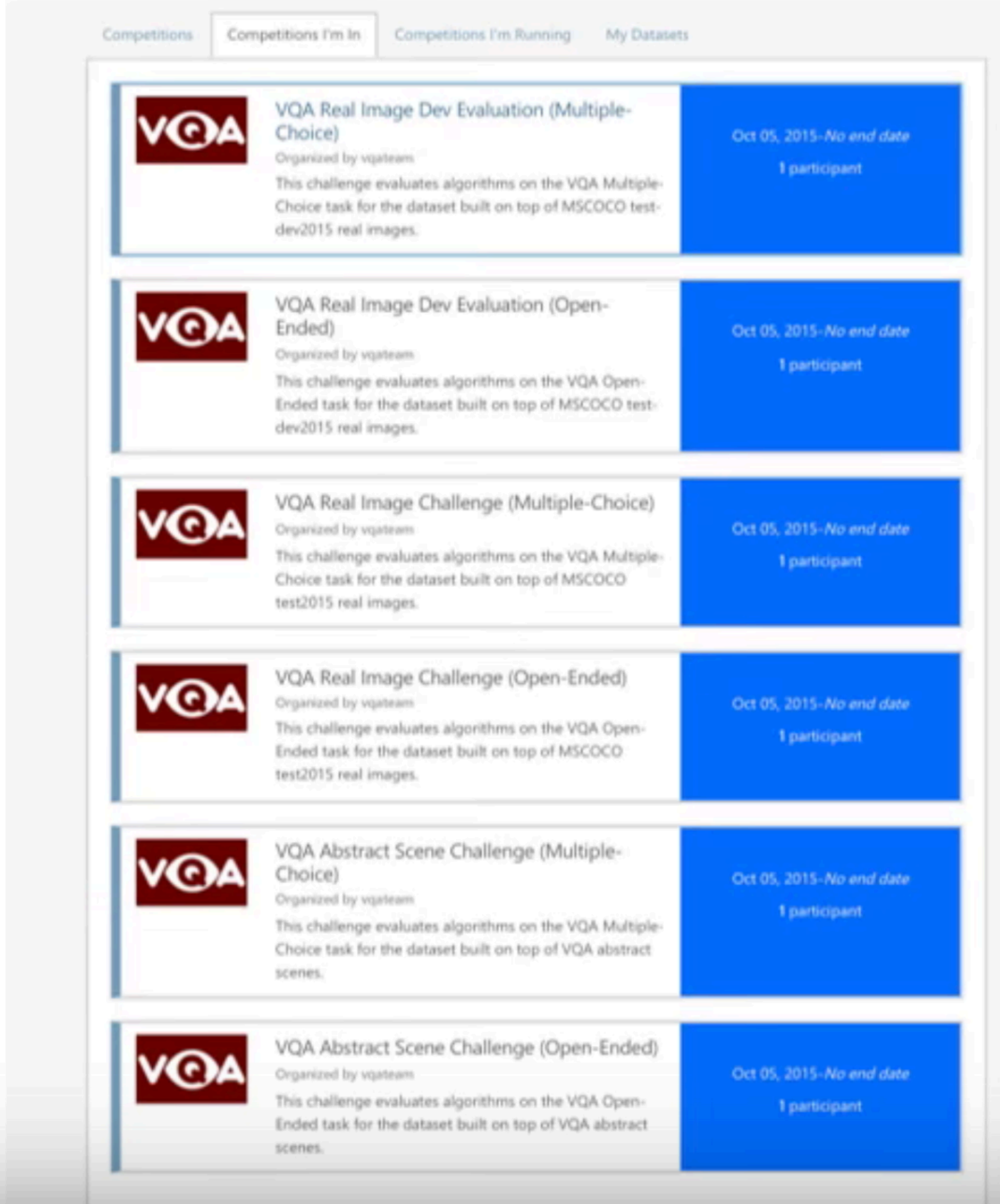
Is it daytime?

Answer

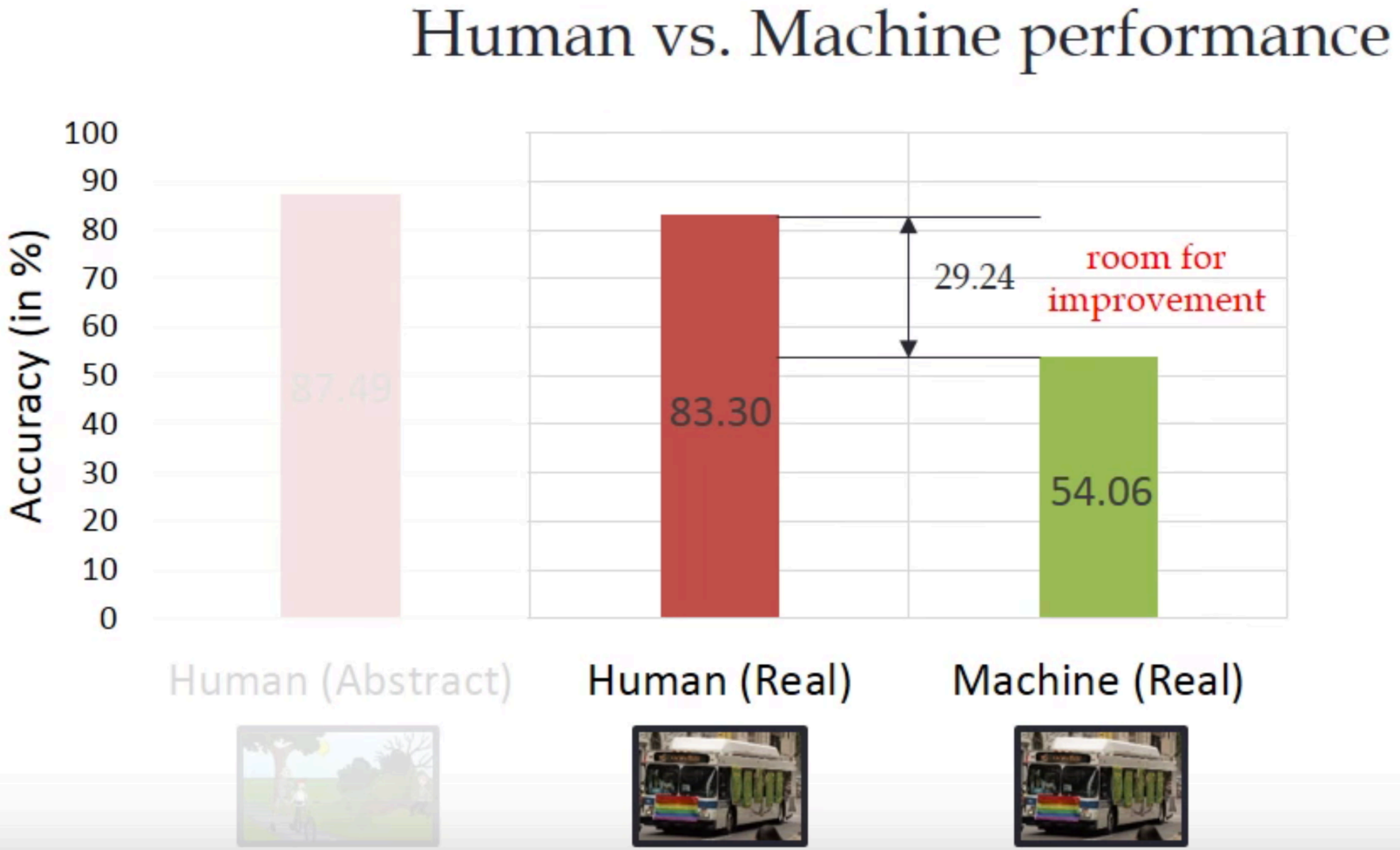
Answer	Confidence
yes	0.5721
no	0.4035
maybe	0.0017
night	0.0008
day	0.0008

VQA Challenges on www.codalab.org





Open-Ended Task Accuracies



Big Visual Data

flickr

6 billion images

3.5 trillion photographs

imgur — the simple image sharer

1 billion images served daily

YouTube

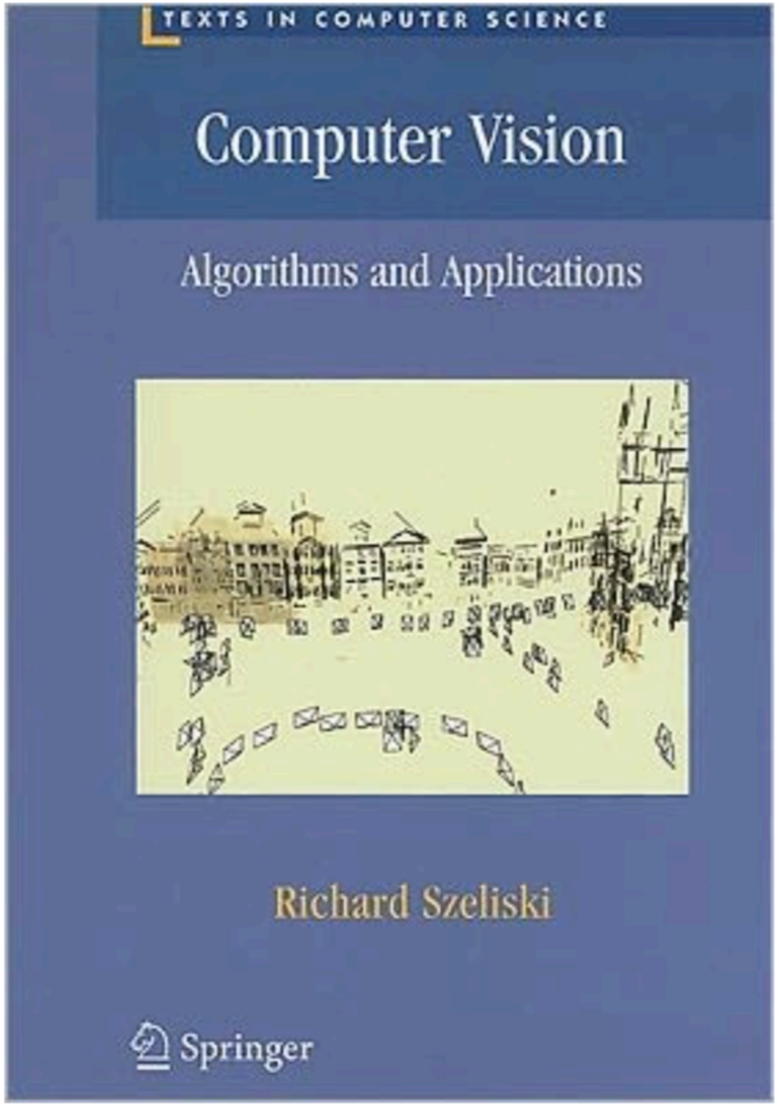
100 hours uploaded per minute

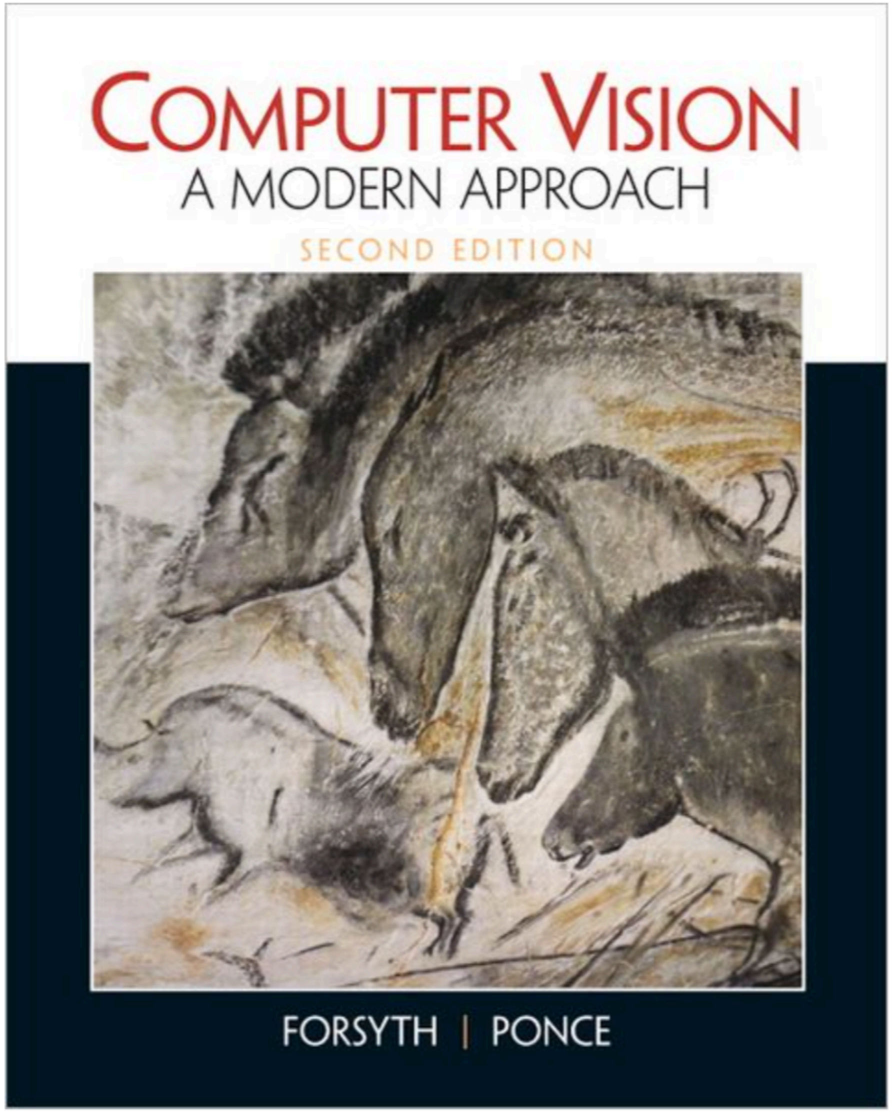
facebook

70 billion images


Too Big for Humans
Digital Dark Matter
[Perona 2010]

Books


TEXTS IN COMPUTER SCIENCE
Computer Vision
Algorithms and Applications
Richard Szeliski
Springer


COMPUTER VISION
A MODERN APPROACH
SECOND EDITION
FORSYTH | PONCE