9th Russian Summer School in Information Retrieval (RuSSIR)
August 24-28, 2015
St Petersburg, Russia

Visual object recognition and localization

Part 1:

Introduction to visual recognition

Ivan Laptev

ivan.laptev@inria.fr

INRIA, WILLOW, ENS/INRIA/CNRS UMR 8548
Laboratoire d'Informatique, Ecole Normale Supérieure, Paris, France

Includes slides from Steven M. Seitz and A. Torralba

What is Computer Vision?

What is Computer Vision?

Computer vision grand challenge: Dynamic scene understanding

Objects:
cars, glasses, people, etc…

Actions:
drinking, running, door exit, car enter, etc…

constraints

Scene categories:
indoors, outdoors, street scene, etc…

Geometry:
Street, wall, field, stair, etc…

Why visual recognition?

Data

- Image and video data is exploding (cheap cameras, easy storage and sharing)

BBC Motion Gallery

~2.5 Billion new images / month

ina

TV-channels recorded since 60's

flickr

~5K image uploads every min.

YouTube
Broadcast Yourself

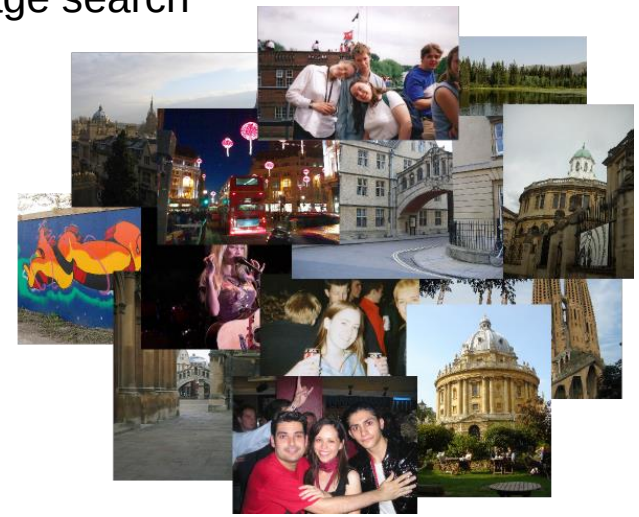
>34K hours of video upload every day

~30M surveillance cameras in US
=> ~700K video hours/day

And even more with future wearable devices

Applications

Very large-scale image search



Person identification

Driver-less cars

Robotics

Applications

First appearance of
N. Sarkozy on TV

Sociology research:
Influence of character
smoking in movies

Education: How do I
make a pizza?

Motion capture and animation

Predicting crowd behavior
Counting people

Where is my cat?

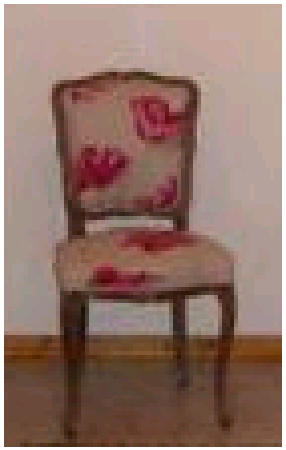
Source: http://www.youtube.com/watch?v=eYdUZdan5i8

Current solution: train a *person-throws-cat-into-trash-bin* classifier

What are the challenges?

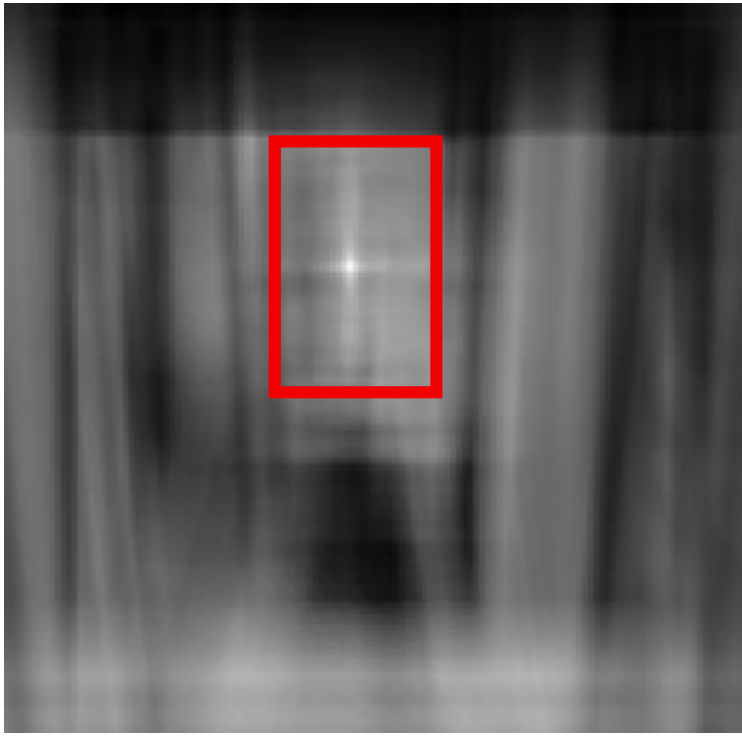
Object recognition
Is it really so hard?

This is a chair

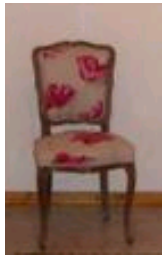


Find the chair in this image

Output of normalized correlation

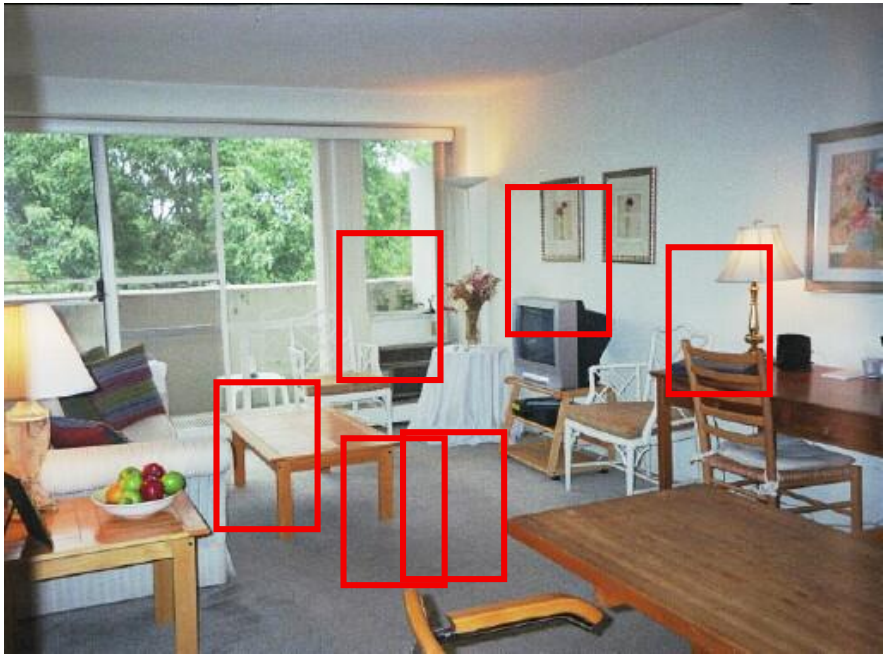


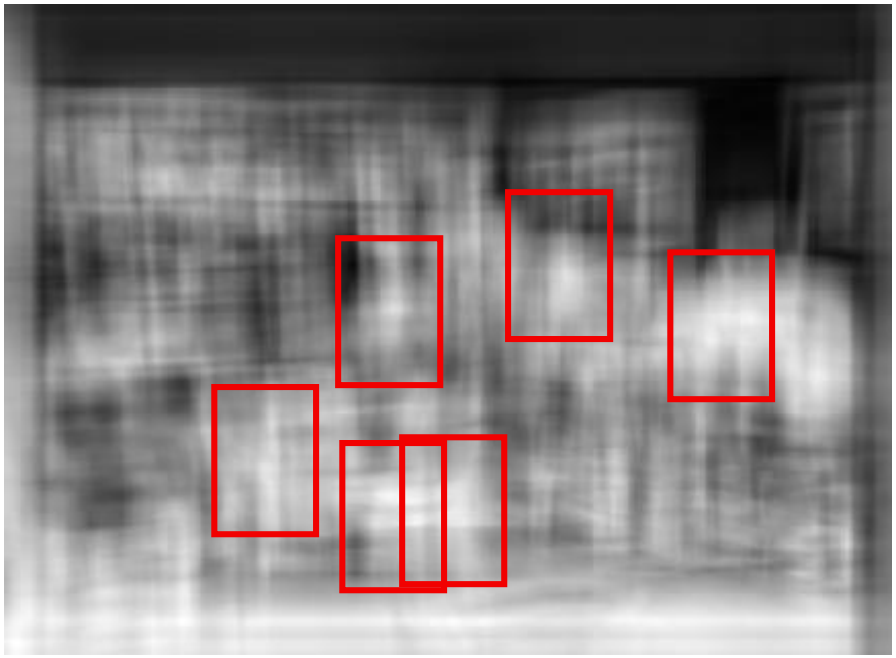
Slide credit: A. Torralba



Object recognition
Is it really so hard?

Find the chair in this image





Pretty much garbage
Simple template matching is not going to make it

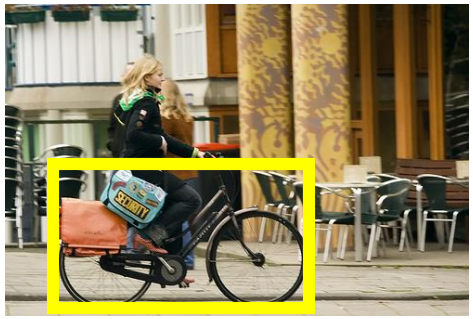
Slide credit: A. Torralba

And it can get a lot harder

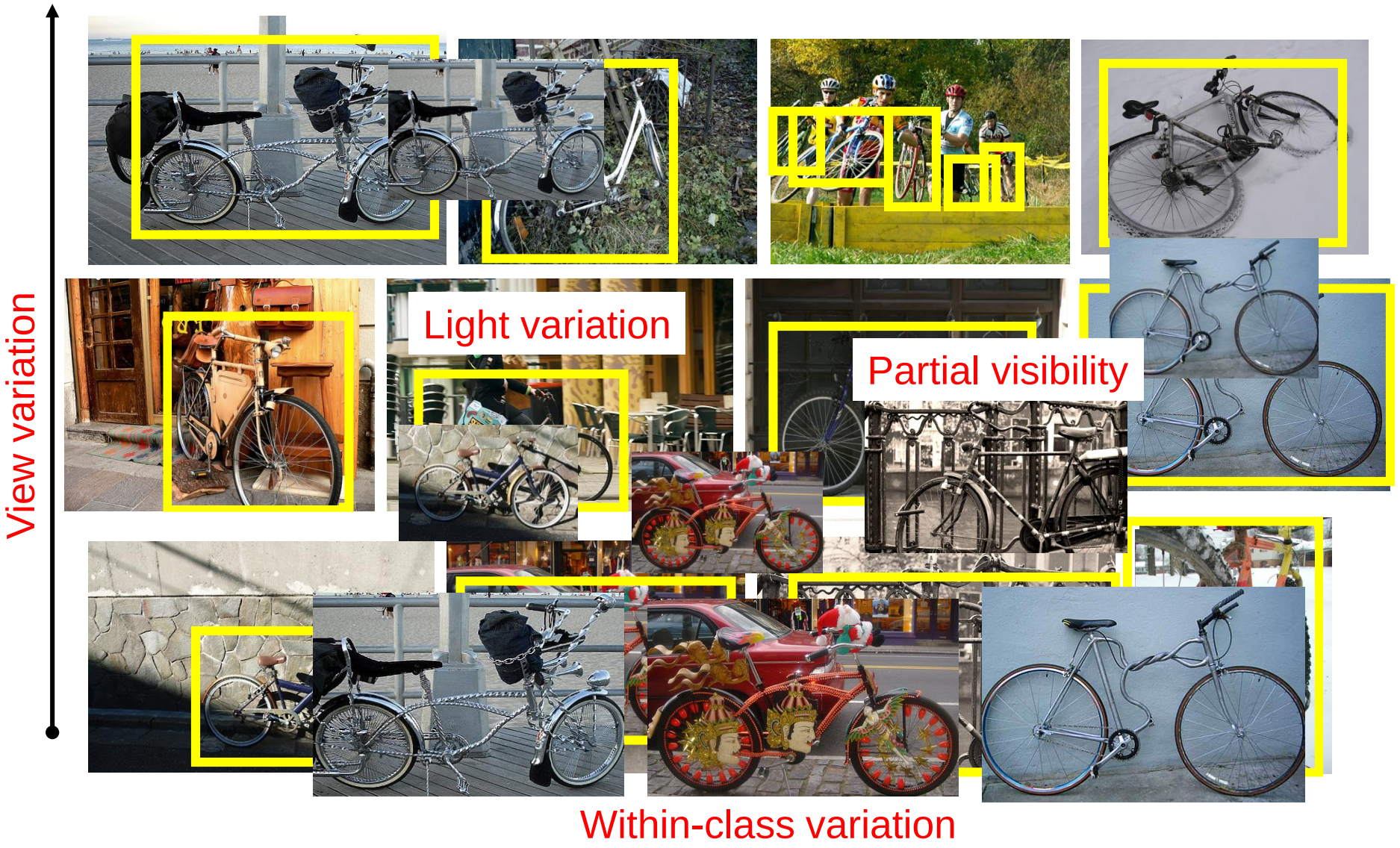
Brady, M. J., & Kersten, D. (2003). Bootstrapped learning of novel objects. J Vis, 3(6), 413-422

Slide credit: A. Torralba

Visual variability



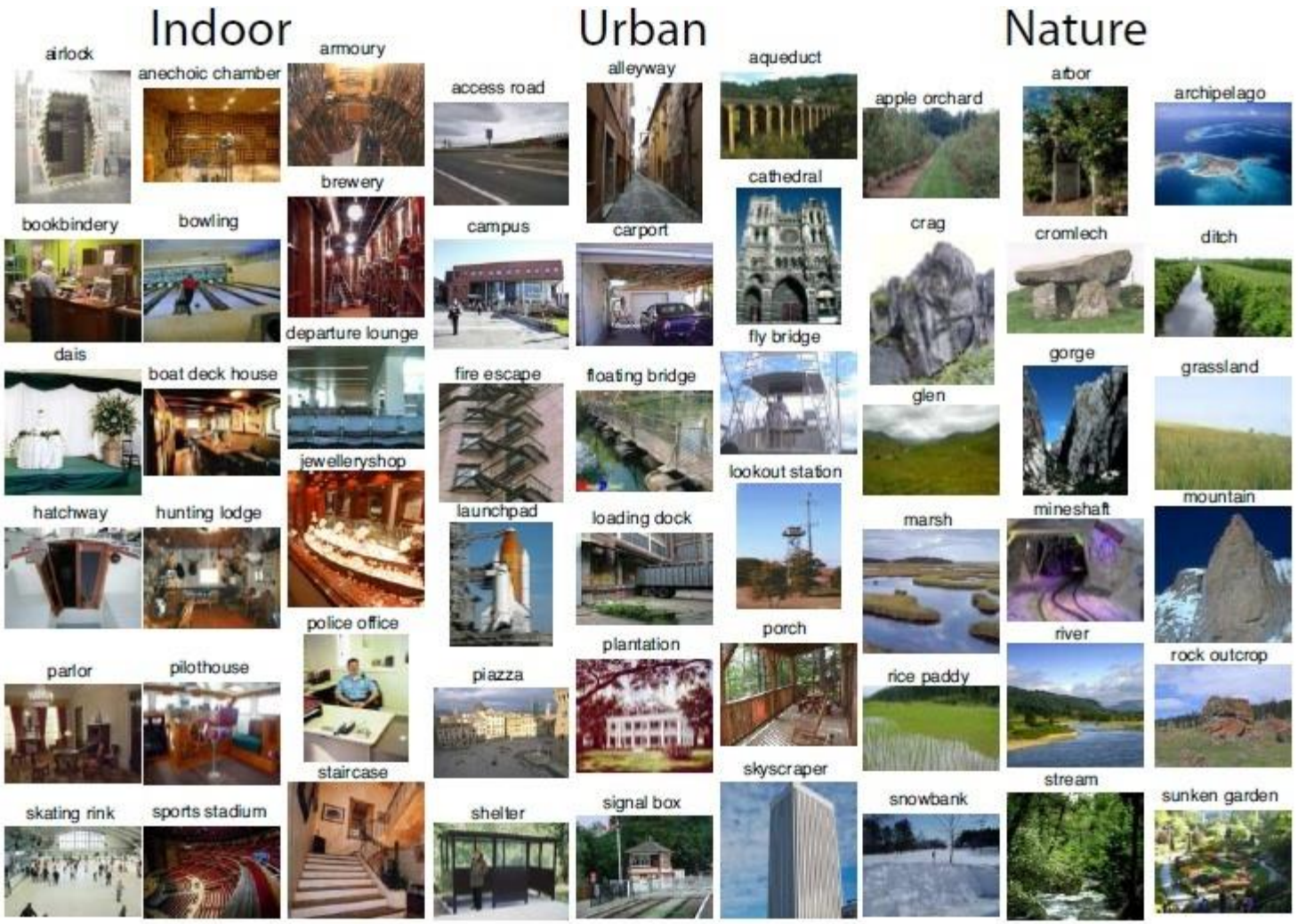
Visual variability



Many *object* categories

ImageNet: >22.000 object categories

Many *scene* categories



SUN dataset: ~900 scene categories

Many *action* categories

Interacting w. computer

Taking a photograph

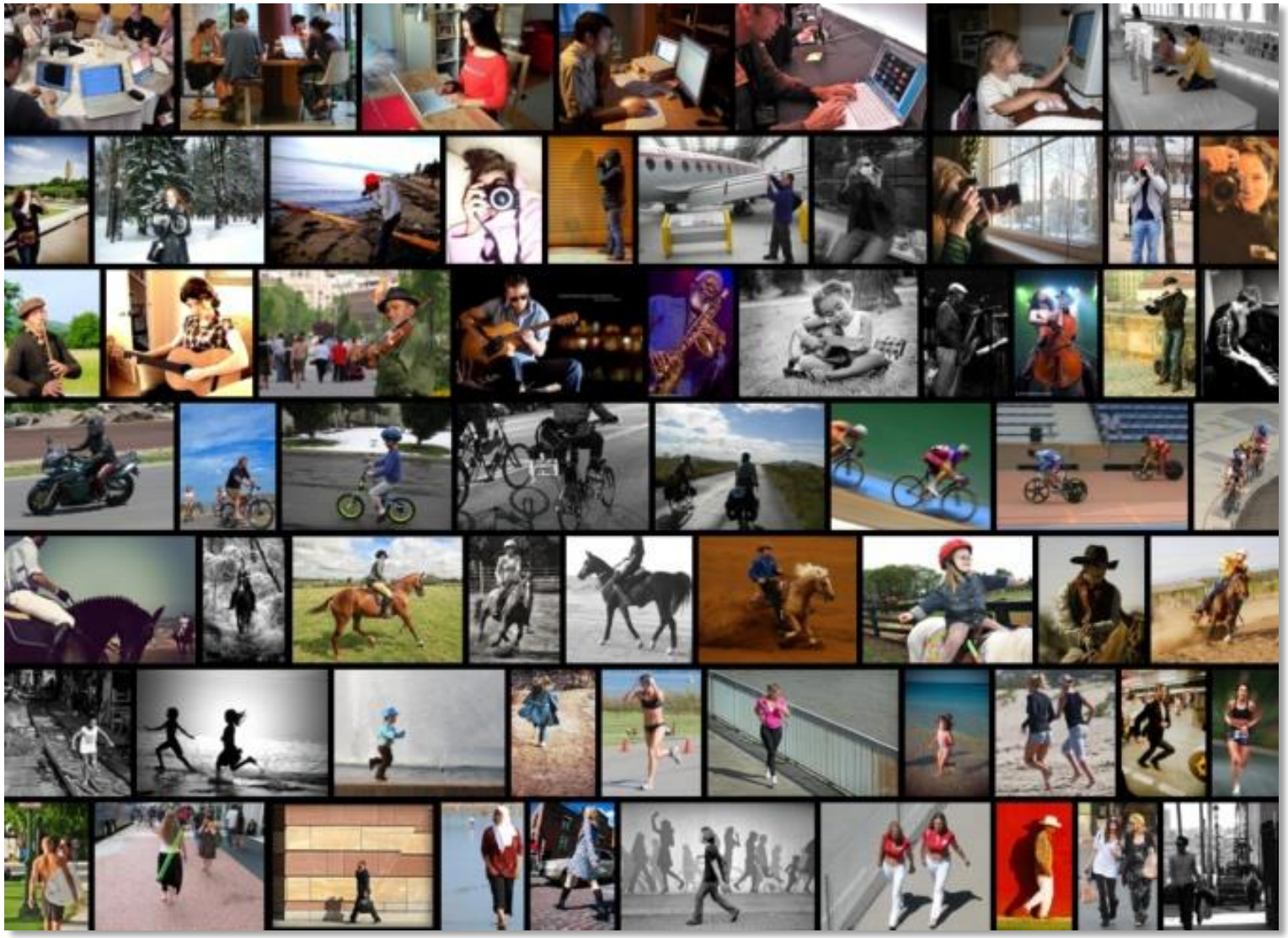
Playing music

Riding bike

Riding horse

Running

Walking



Subtle differences

Same or different person?

Same

Different

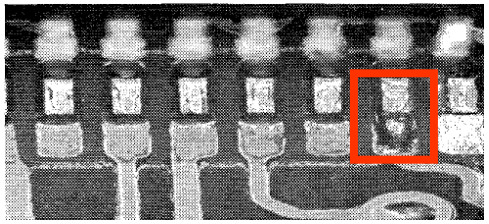
What is the recent progress?

Recent progress

Industry

Research

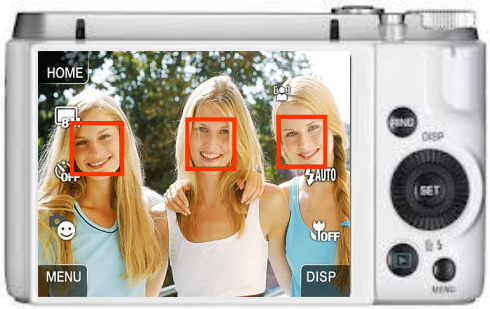
1990s:



Automated quality inspection
(controlled lighting, scale,…)

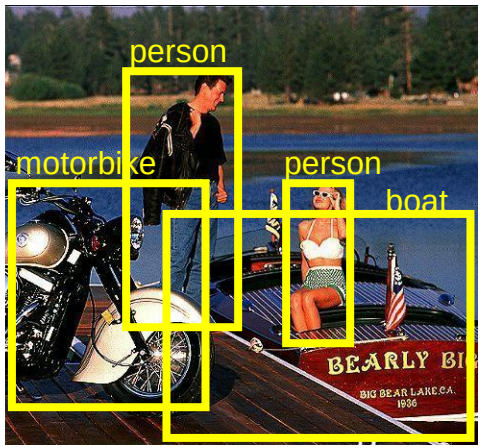
Recognition at the level of a few
toy objects (COIL 20 dataset)

Now:



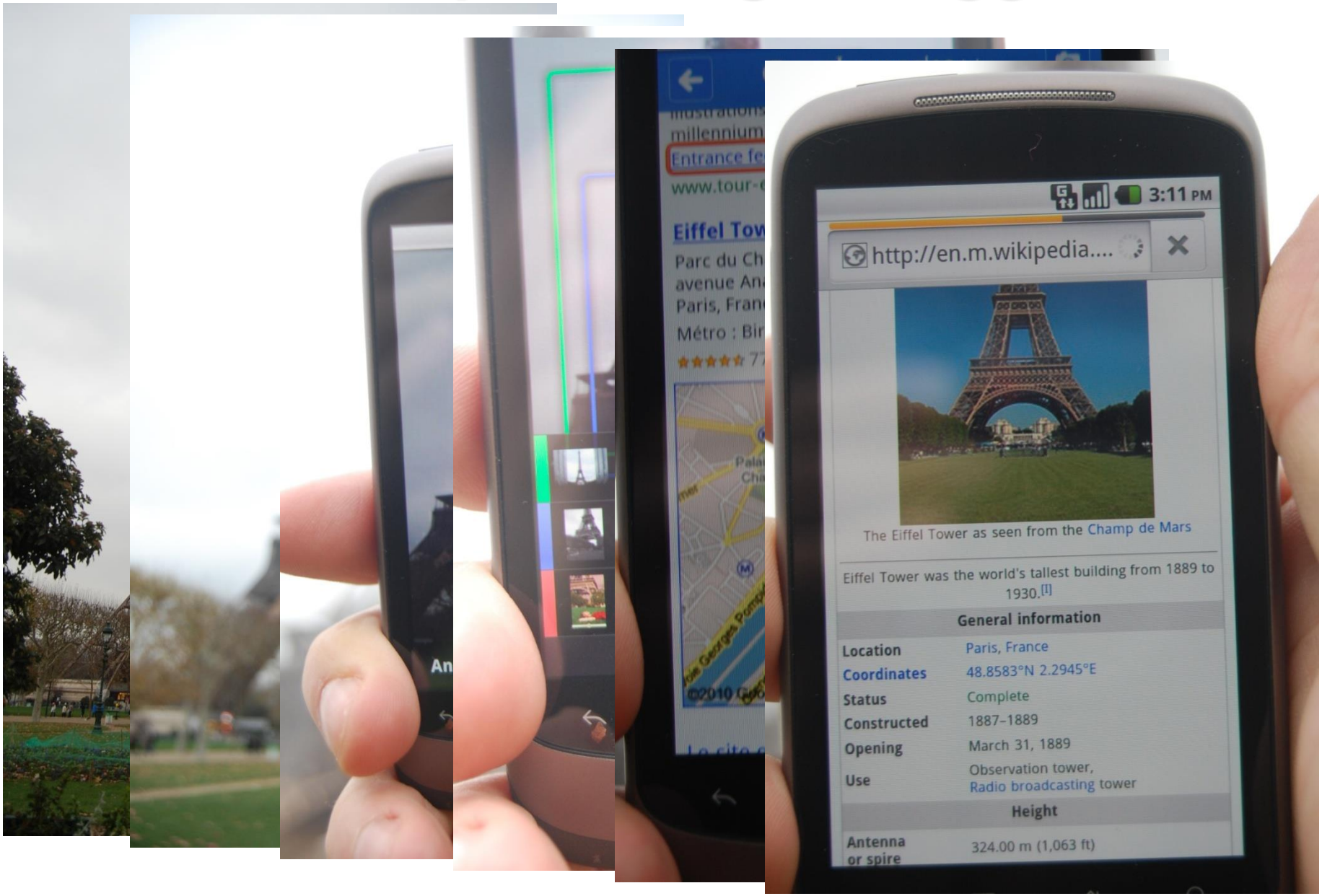
Facial expression
recognition

Gesture recognition

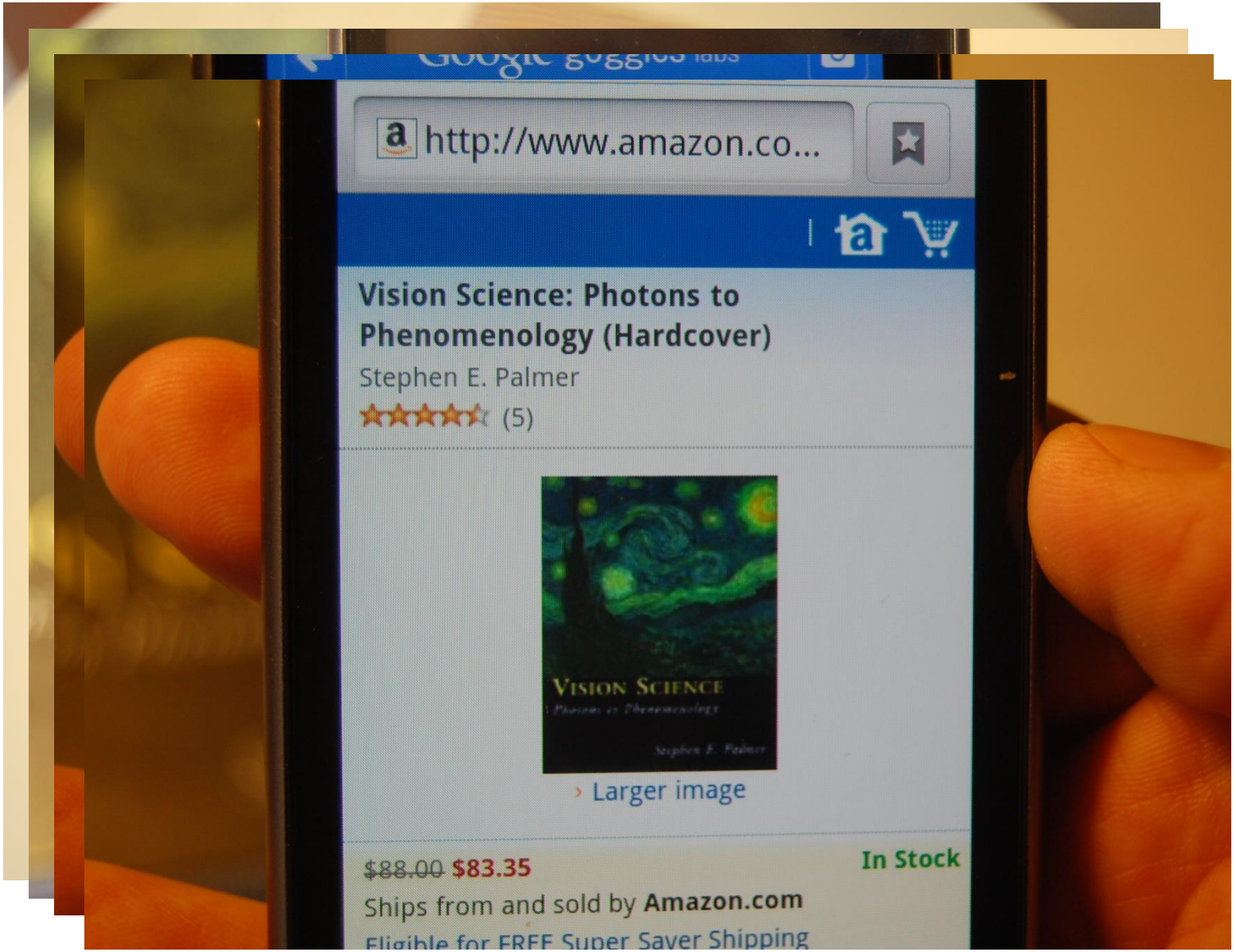


Object and action recognition in natural
images and video

Example: Google Goggles

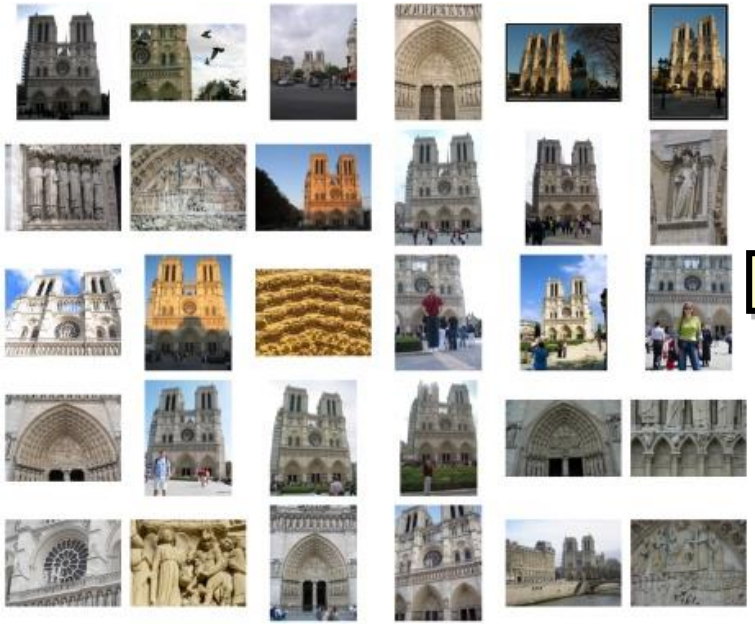


What is the recent progress?



Example: Photo Tourism

[Snavely, Seitz and Szeliski ’06]



Input photographs

Scene reconstruction

Relative camera positions and orientations
Point cloud
Sparse correspondence

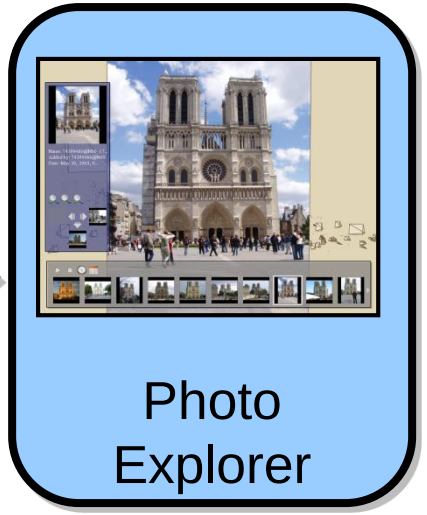


Photo Explorer

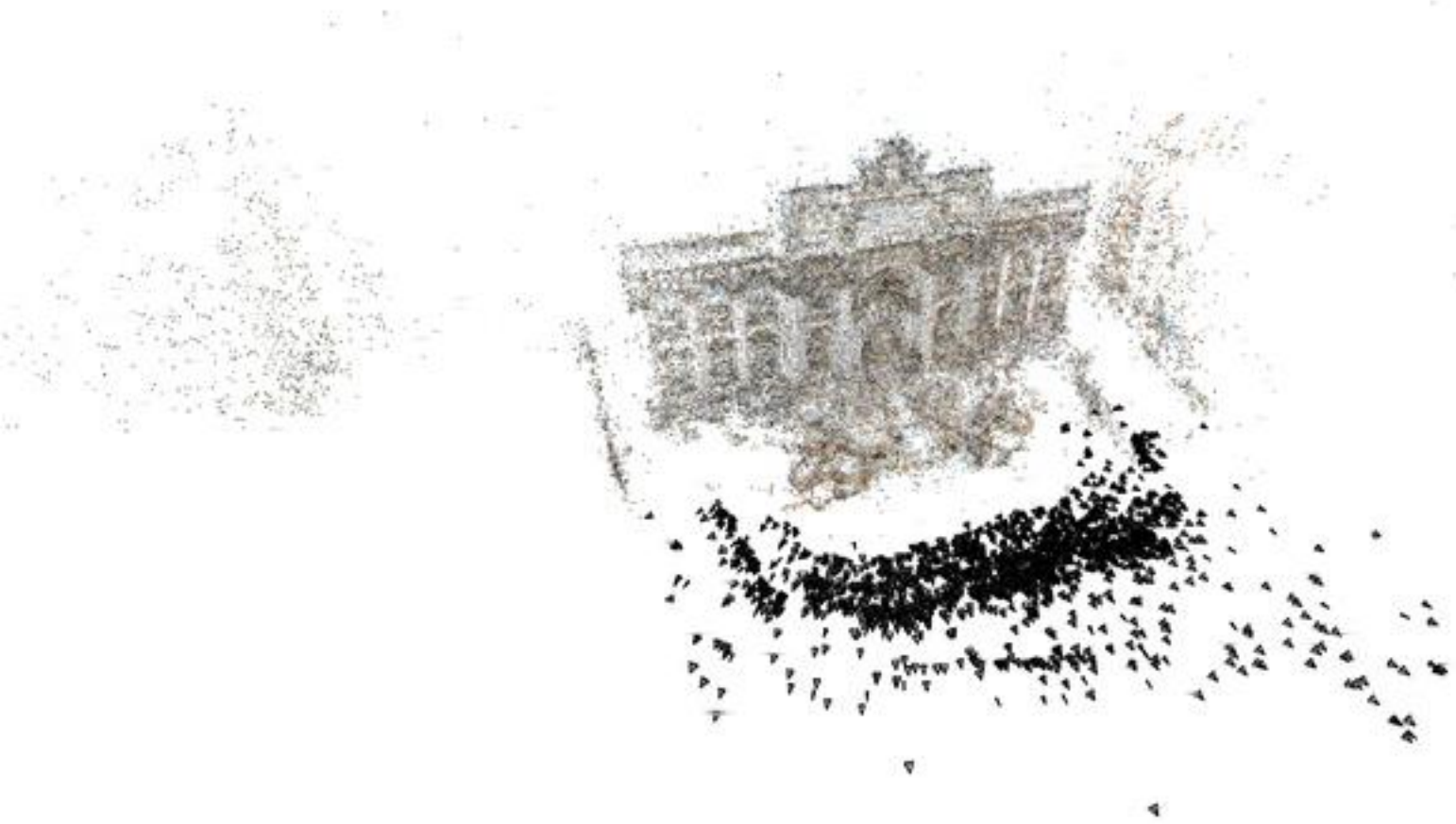
15,464

37,383

76,389

Reconstruction and photo explorer

[Agarwal, Snavely, Simon, Seitz and Szeliski ’09]



•150,000 images from Flickr.com associated with the tags "Rome" or "Roma"

• Matching and reconstruction: 21 hours on a cluster with 496 compute cores

Reconstruction and photo explorer

[Agarwal, Snavely, Simon, Seitz and Szeliski ’09]

•150,000 images from Flickr.com associated with the tags "Rome" or "Roma"

• Matching and reconstruction: 21 hours on a cluster with 496 compute cores

Reconstruction and photo explorer

[Agarwal, Snavely, Simon, Seitz and Szeliski ’09]

•250,000 Venice images from Flickr.com

• Matching and reconstruction: 27 hours on a cluster with 496 compute cores

Reconstruction and photo explorer

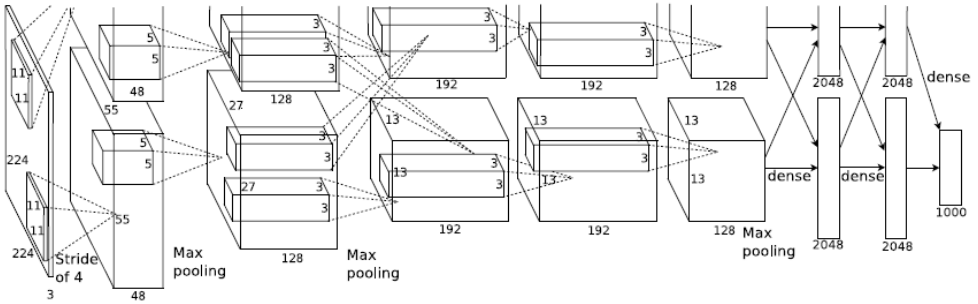
[Agarwal, Snavely, Simon, Seitz and Szeliski ’09]

•250,000 Venice images from Flickr.com

• Matching and reconstruction: 27 hours on a cluster with 496 compute cores

Convolutional Neural Networks

- Success in character recognition [LeCun'88].
- Limited performance on natural images until 2012.



- [Krizhevsky et al. 2012]: break-through in ImageNet object classification.

ILSVRC'12: 1.2M images, 1K classes

2012

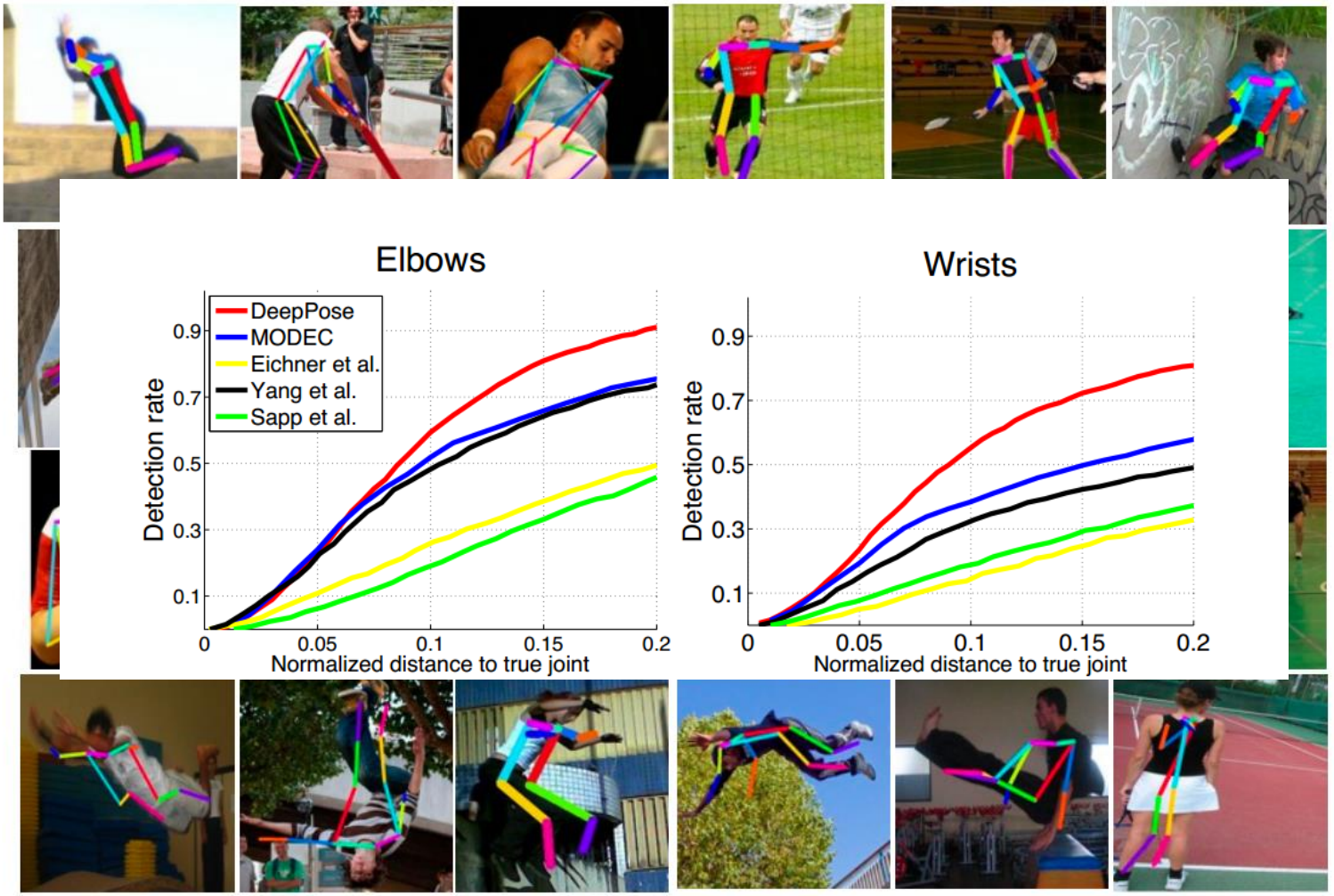
Method:	Top 5 error:
SIFT + FVs [7]	*26.2%*
1 CNN	—
5 CNNs	**16.4%**
1 CNN*	—
7 CNNs*	**15.3%**

2014-2015

GoogLeNet:	6.6%
VGG:	6.8%
BAIDU	5.3%
Human	*5.1%*

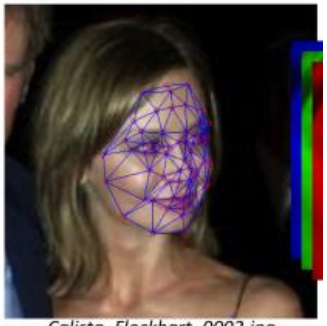
Other recent successes of ConvNets: Human pose estimation

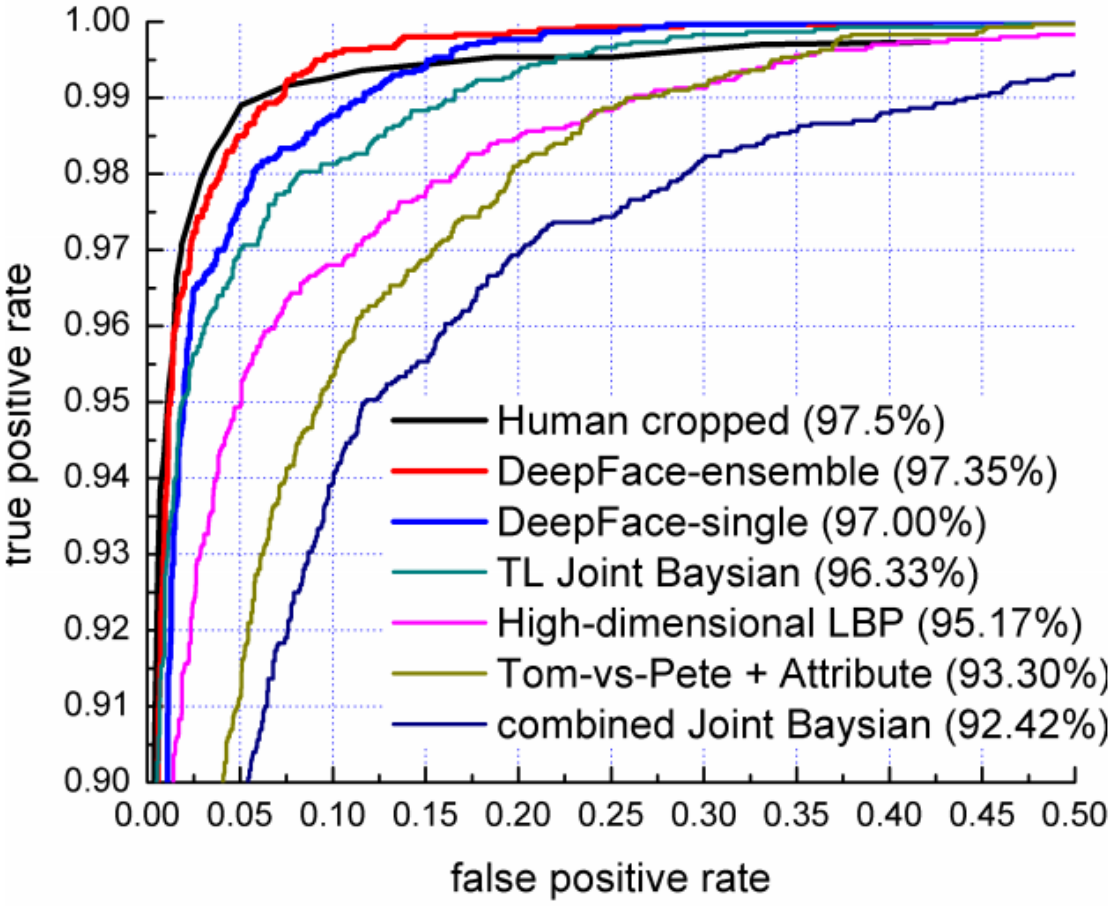
DeepPose: Human Pose Estimation via Deep Neural Networks
A. Toshev and C. Szegedy, CVPR 2014



Other recent successes of ConvNets: Face recognition

DeepFace: Closing the Gap to Human-Level Performance in Face Verification
Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, Lior Wolf, CVPR 2014





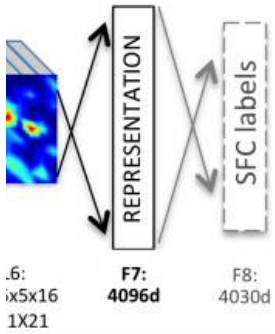
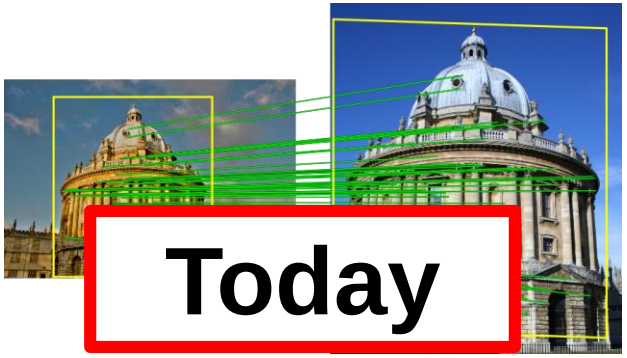


Figure 3. The ROC curves on the *LFW* dataset. Best viewed in color.

Summary of the course

Problems:

1. Instance-level recognition



Today

2. Category-level recognition

Object classes

Tomorrow

as well as scene and action classes

Methods:

Local features,
Bag-of-Features (BOF) representations

Geometric verification,
Indexing

Machine learning:
Support Vector Machines (SVM)
Convolutional Neural Nets (CNN)

Practical sessions:

If you are not fluent in Matlab, go through this tutorial:

http://www.di.ens.fr/willow/teaching/recvis12/matlab-tut.zip