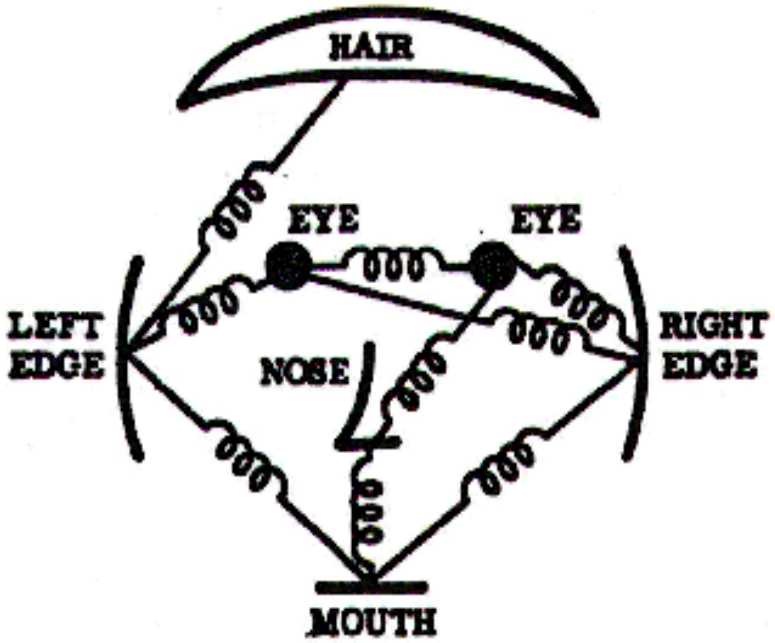
CS6670: Computer Vision

Noah Snavely

Lecture 17: Parts-based models and context

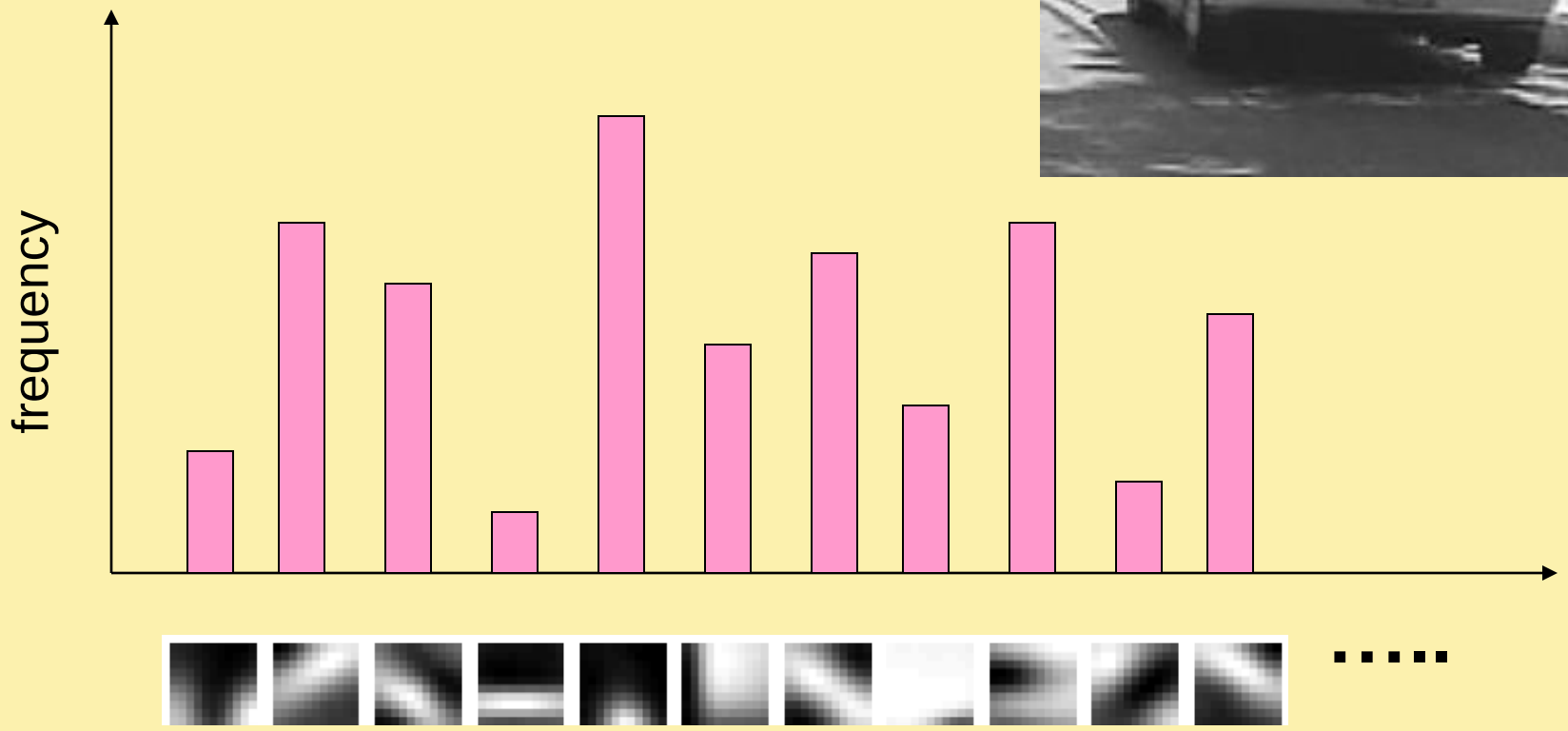


Announcements

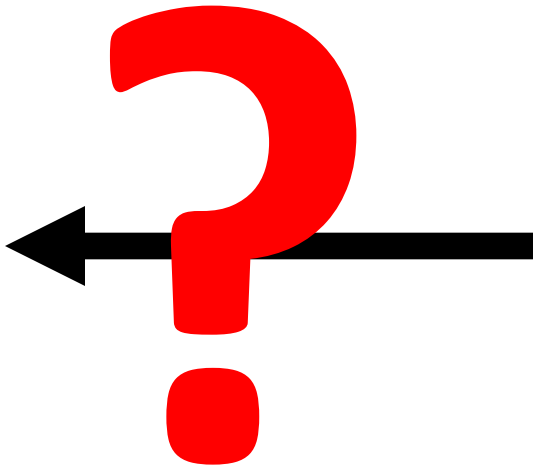
- Project 3: Eigenfaces
  - due Wednesday, November 11 at 11:59pm
  - solo project

Object → Bag of ‘words’

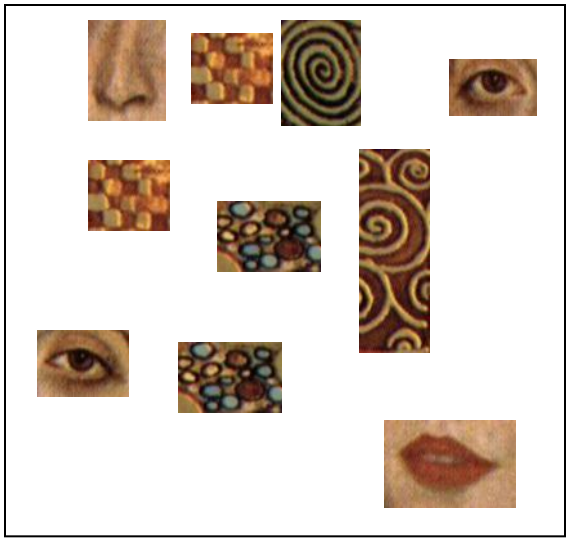
3. Image representation

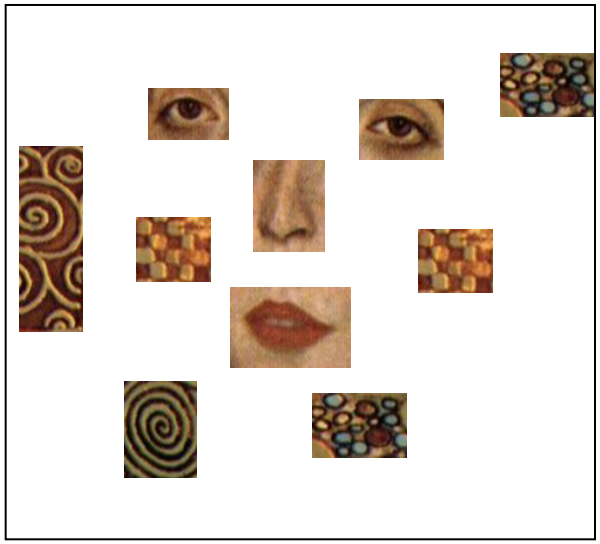


What about spatial info?



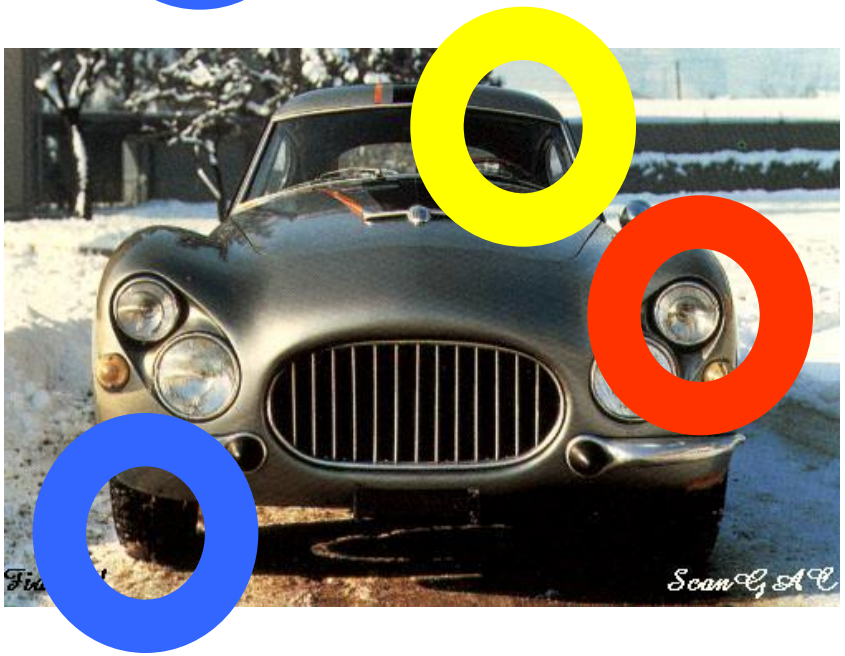
Problem with bag-of-words



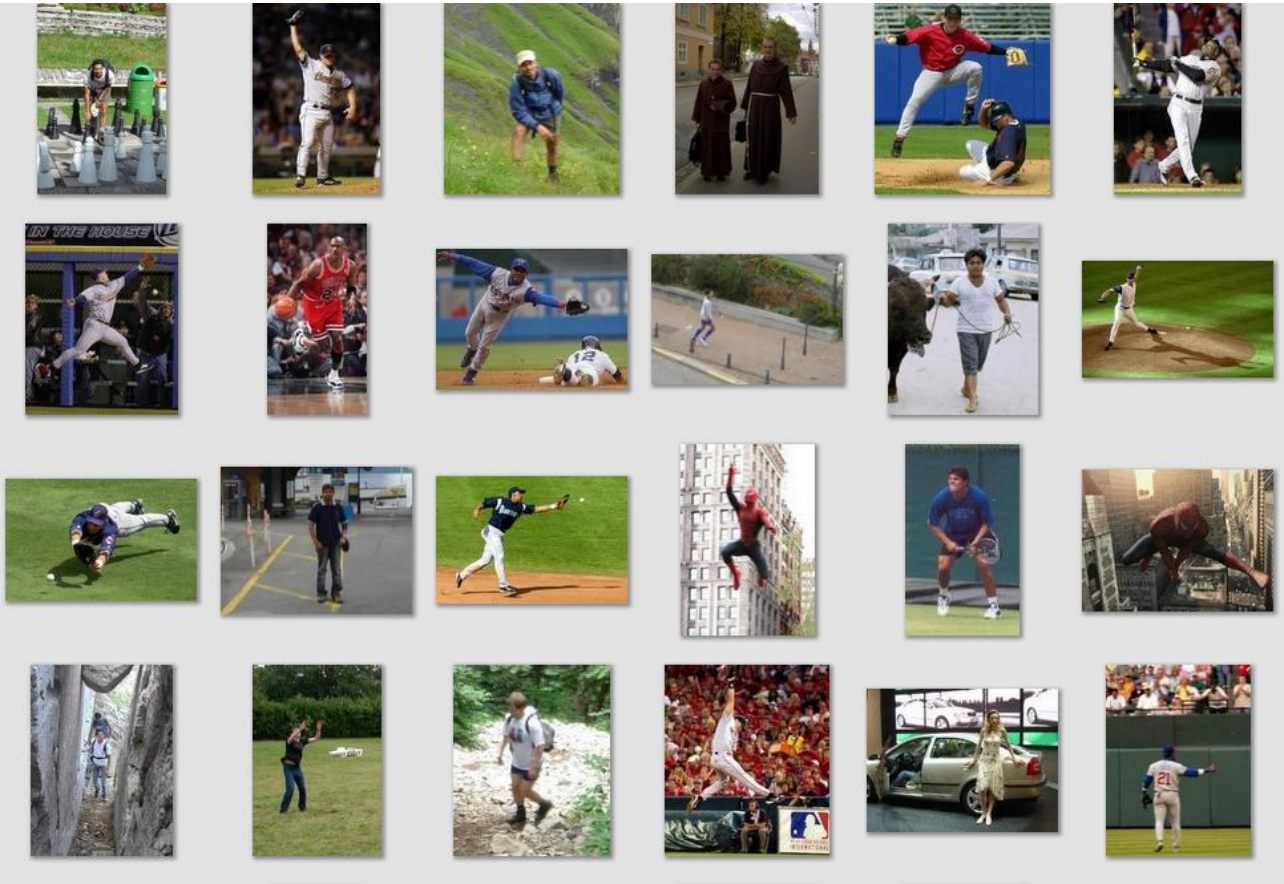


- All have equal probability for bag-of-words methods
- Location information is important

Model: Parts and Structure

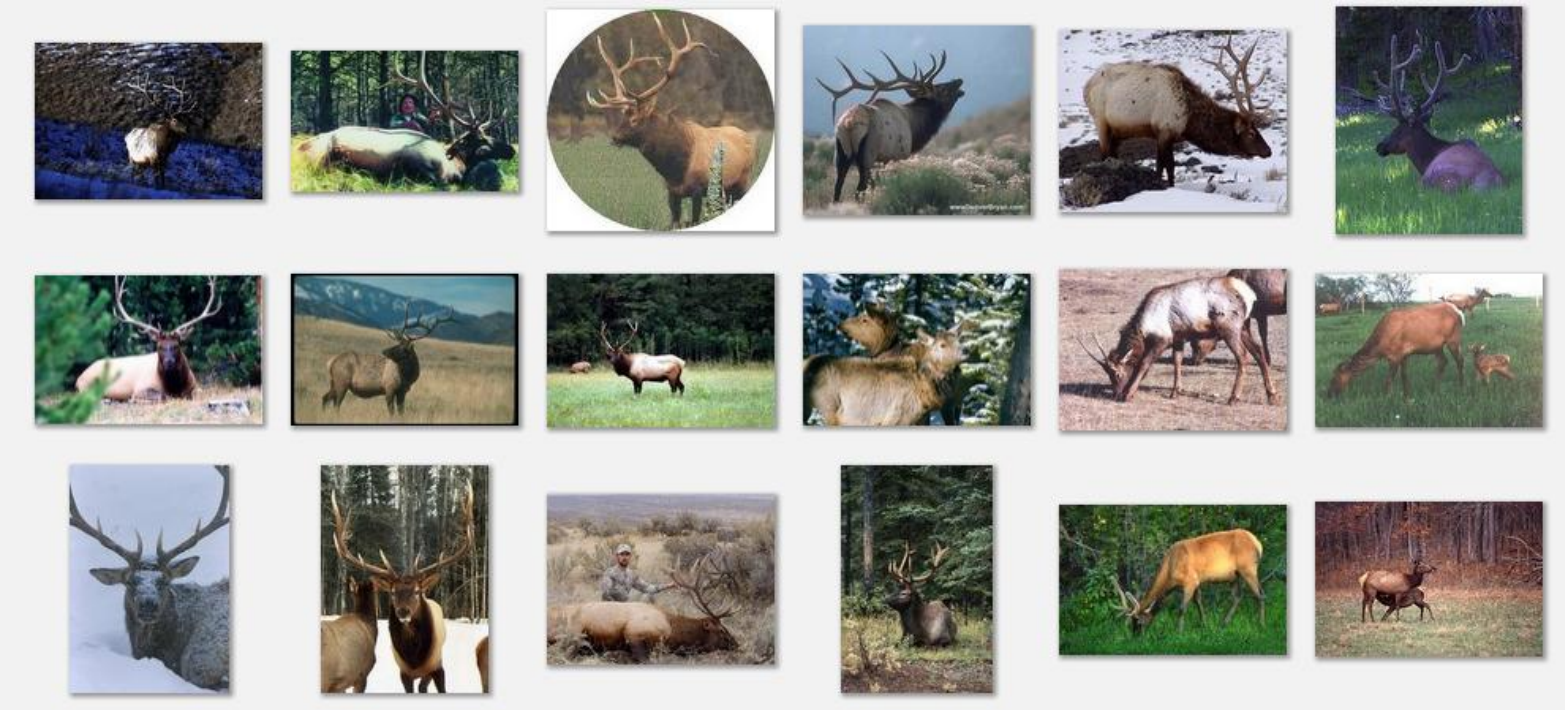


Deformable objects



Images from D. Ramanan's dataset

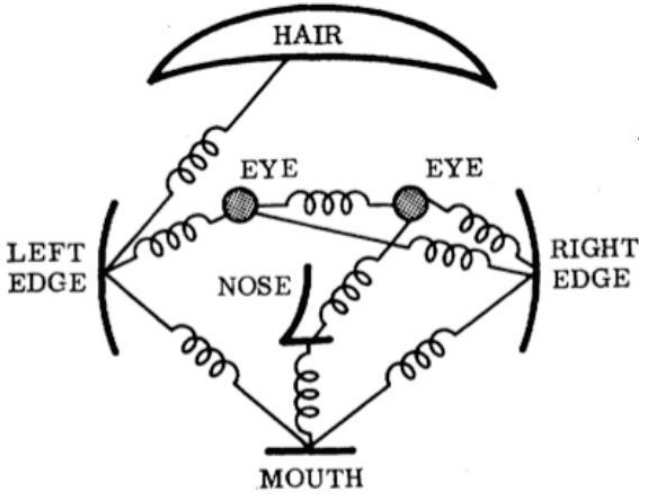
Deformable objects



Images from Caltech-256

Part-based representation

- Objects are decomposed into parts and spatial relations among parts



Fischler and Elschlager ‘73

Pictorial structures

- Two components:
  - Appearance model
    - How much does a given window look like a given part?
  - Spatial model
    - How well do the parts match the expected shape?

Formal Definition of Model

- Set of parts $V=\{v_1, \ldots, v_n\}$

Pictorial Structure

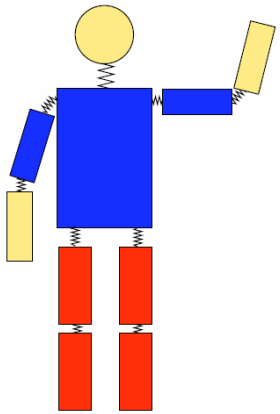
- Matching = Local part evidence + Global constraint

$$L^* = \arg\min_L \left(\sum_{i=1}^{n} m_i(l_i) + \sum_{(v_i,v_j)\in E} d_{ij}(l_i, l_j) \right)$$

- $m_i(l_i)$: matching cost for part I
- $d_{ij}(l_i,l_j)$: deformable cost for connected pairs of parts
- (v_i,v_j): connection between part i and j

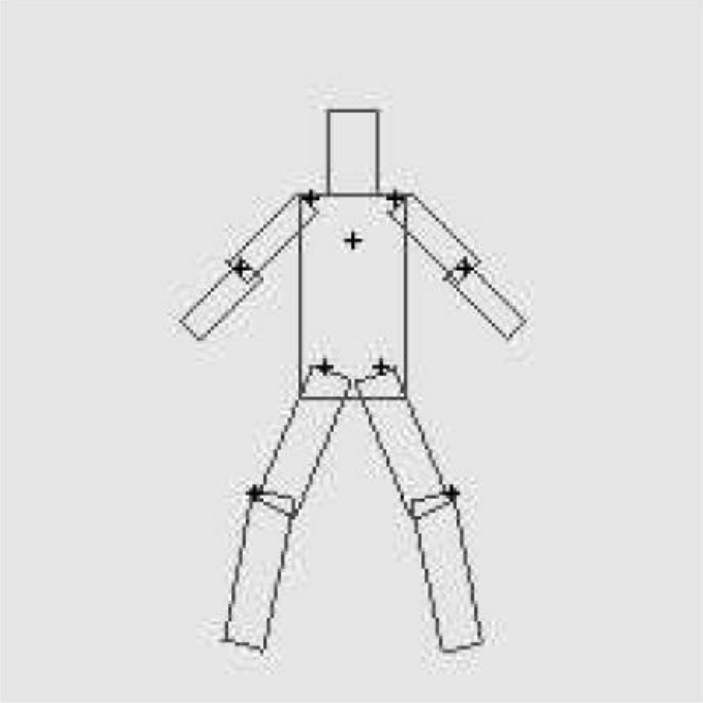
Flexible Template Algorithms

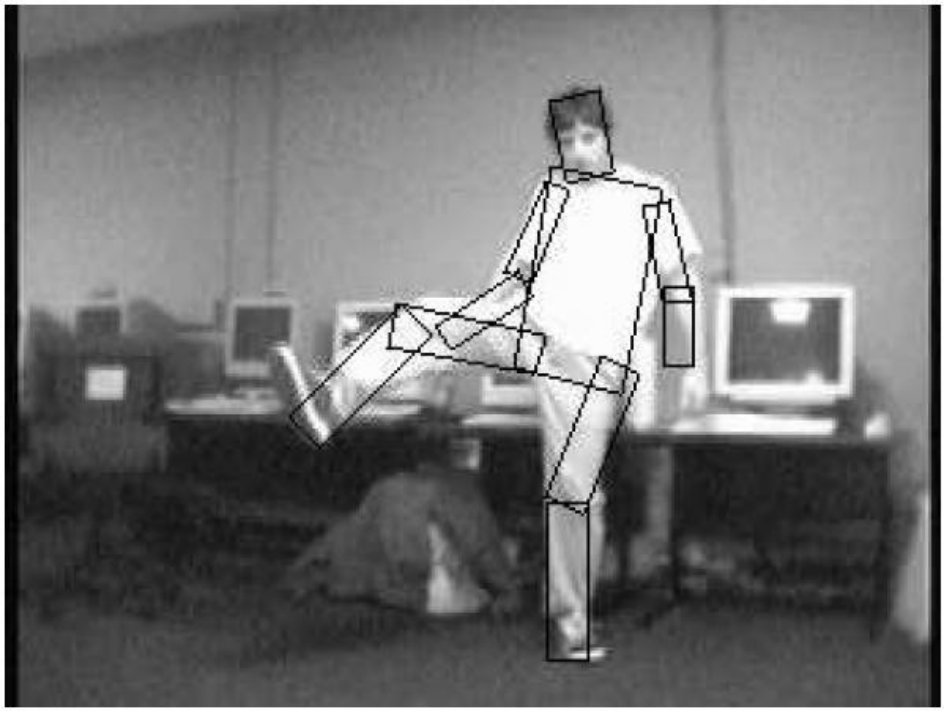
- Difficulty depends on structure of graph
  - Which parts connected and form of constraint



Part-based representation

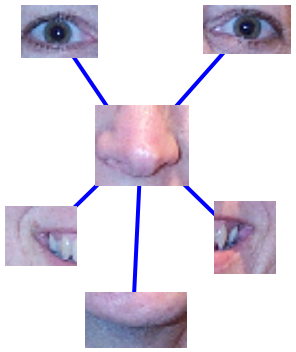
- Tree model ➔ Efficient inference by dynamic programming

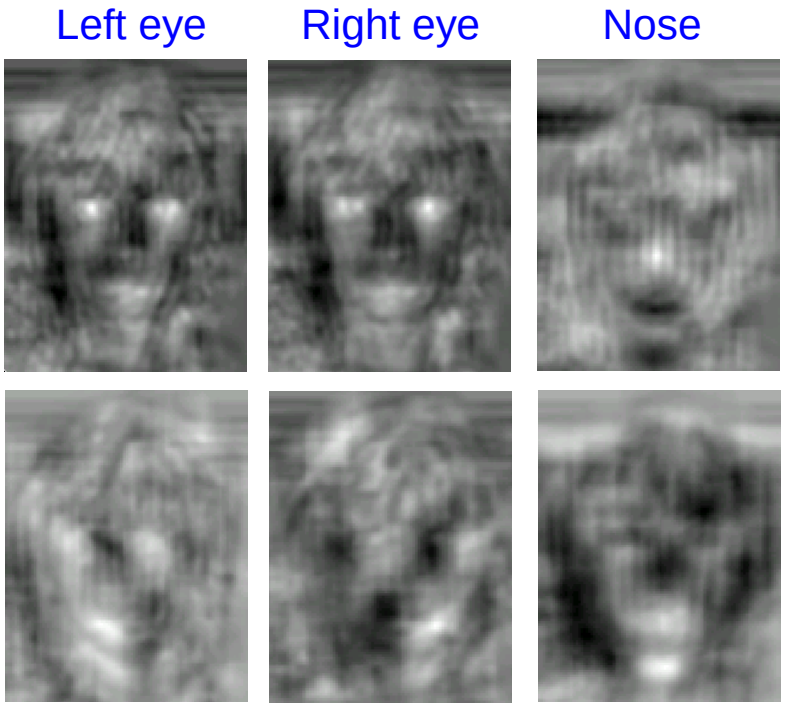




Appearance model

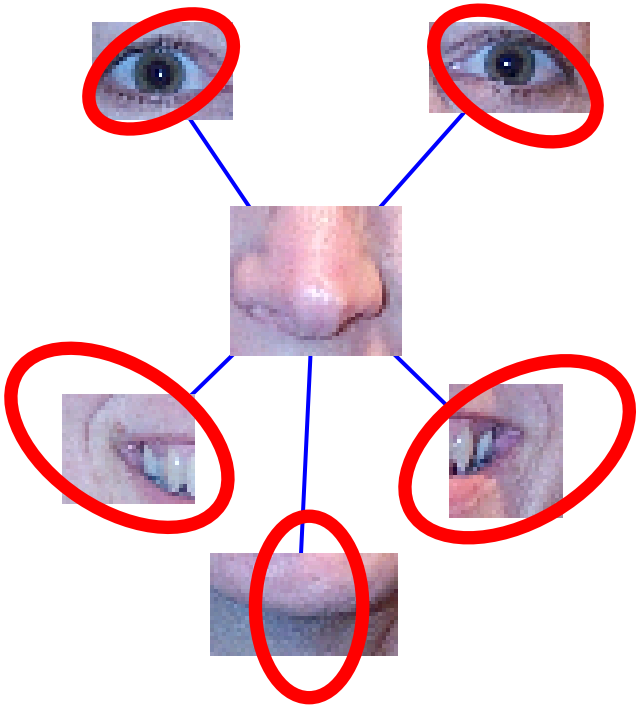
- Each part has an associated appearance model
  - E.g., a reference patch, gradient histogram, etc.





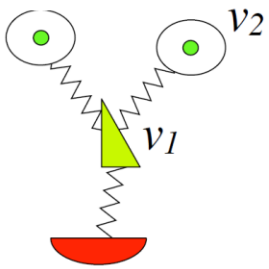
Spatial model

- Each edge represents a spring with a certain relative offset, covariance



Matching on tree structure

$$E(L) = \sum_{i=1}^{n} m_i(l_i) + \sum_{(v_i,v_j) \in E} d_{ij}(l_i,l_j)$$

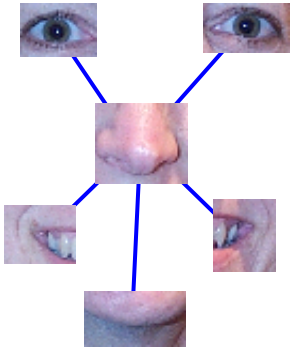


- For each l_1, find best l_2:

$$\text{Best}_2(l_1) = \min_{l_2} \left[m_2(l_2) + d_{12}(l_1,l_2) \right]$$

- Remove v_2, and repeat with smaller tree, until only a single part
- Complexity: $O(nk^2)$: n parts, k locations per part

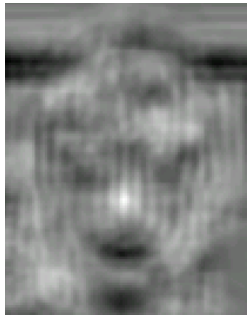
Putting it all together



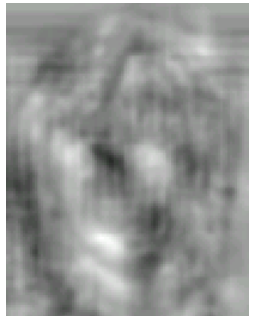
Left eye

Right eye

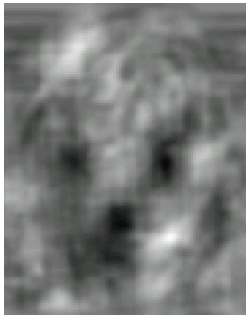
Nose



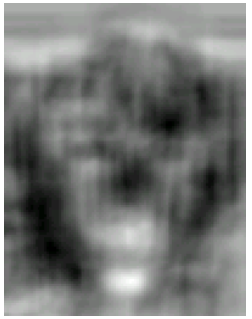
Left mouth



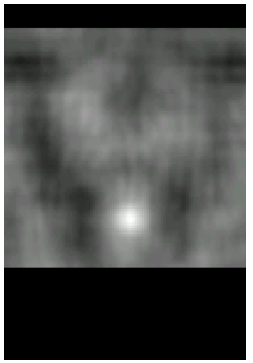
Right mouth

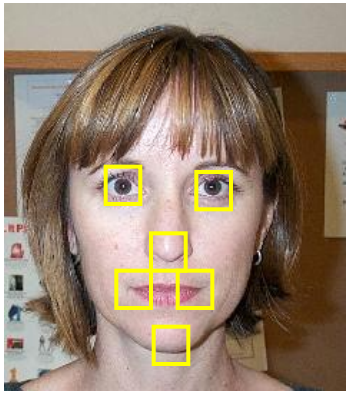


Chin

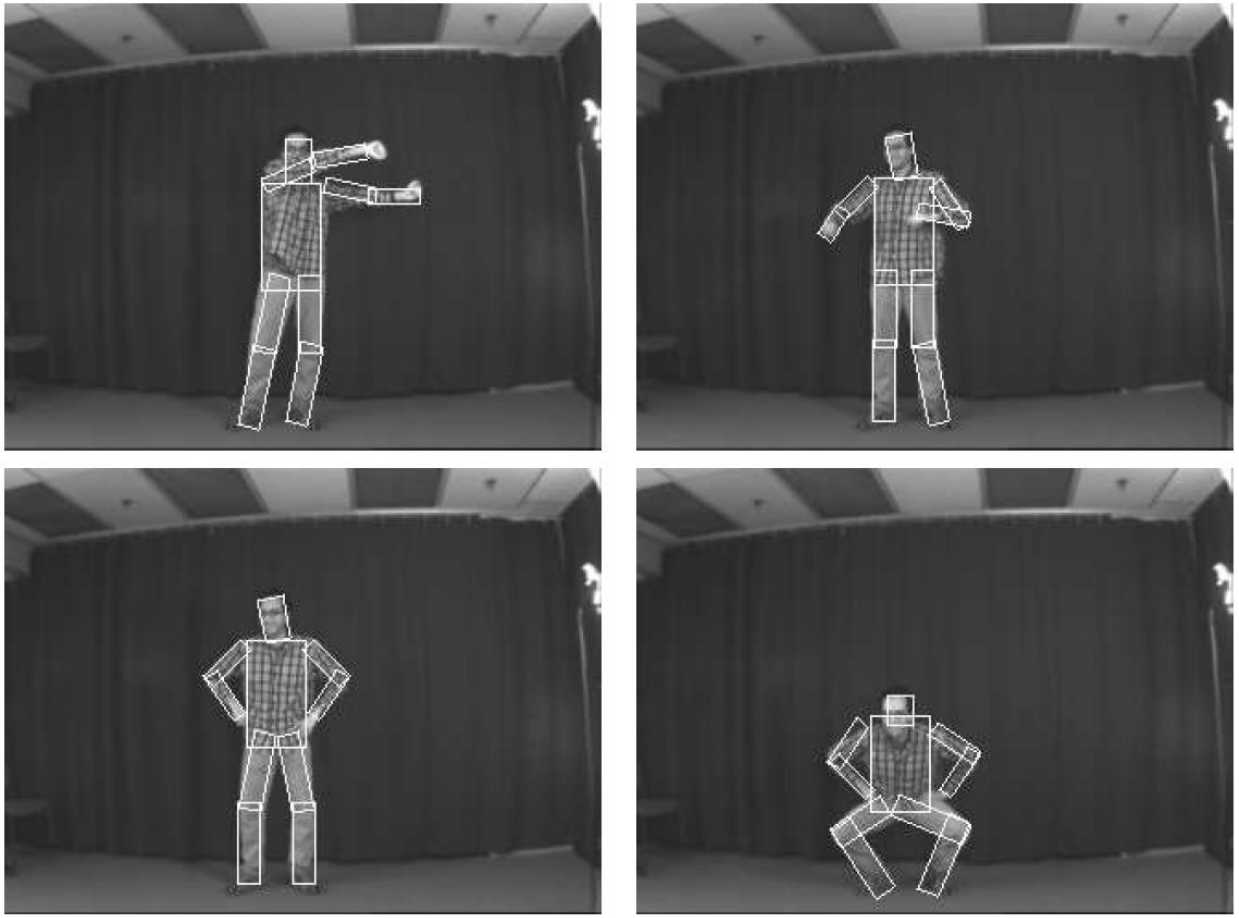


Marginal on Nose

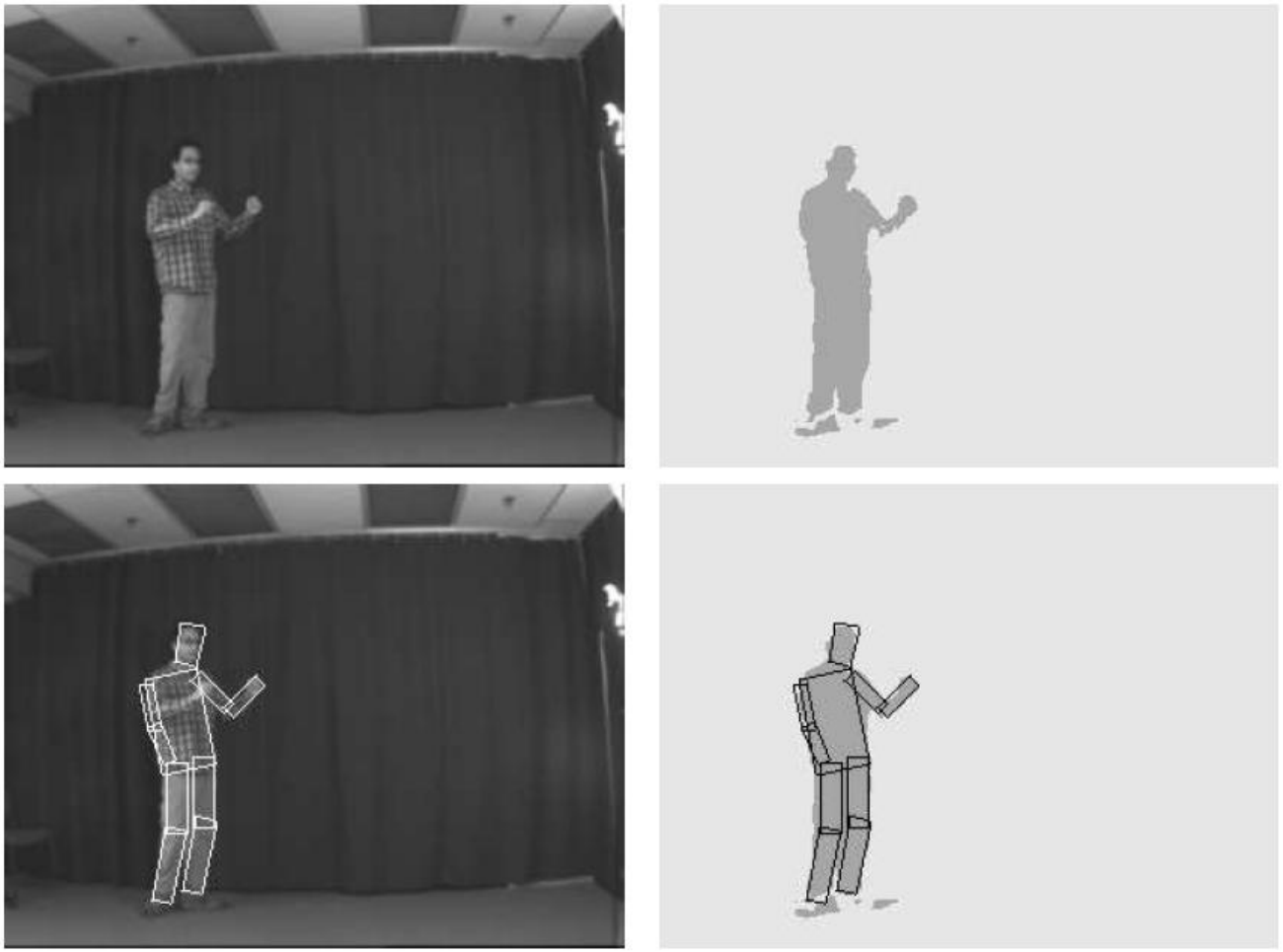




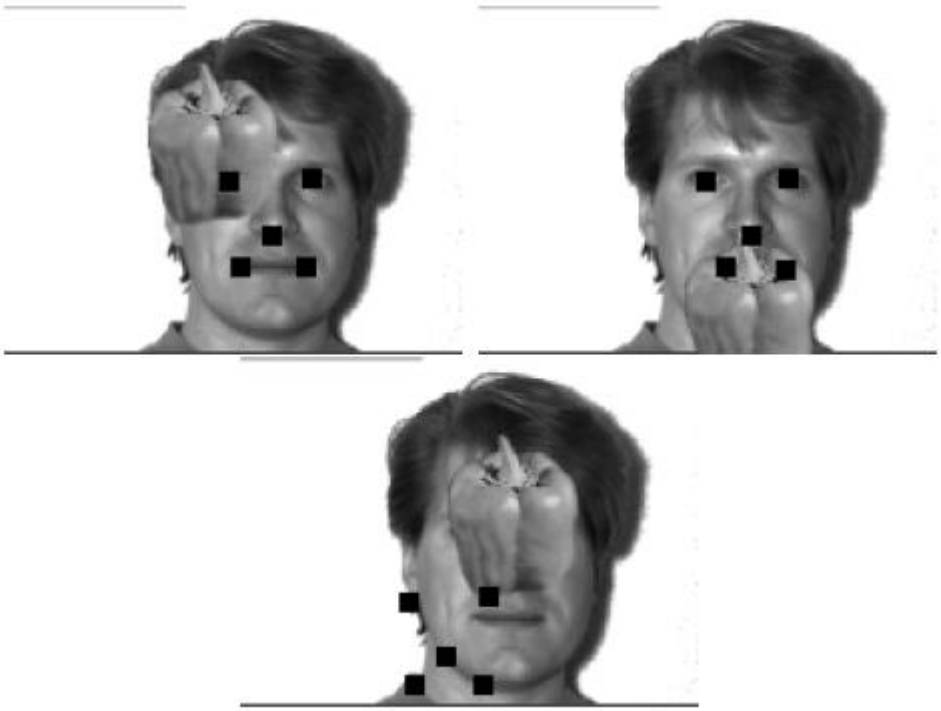
Sample result on matching human



Sample result on matching human



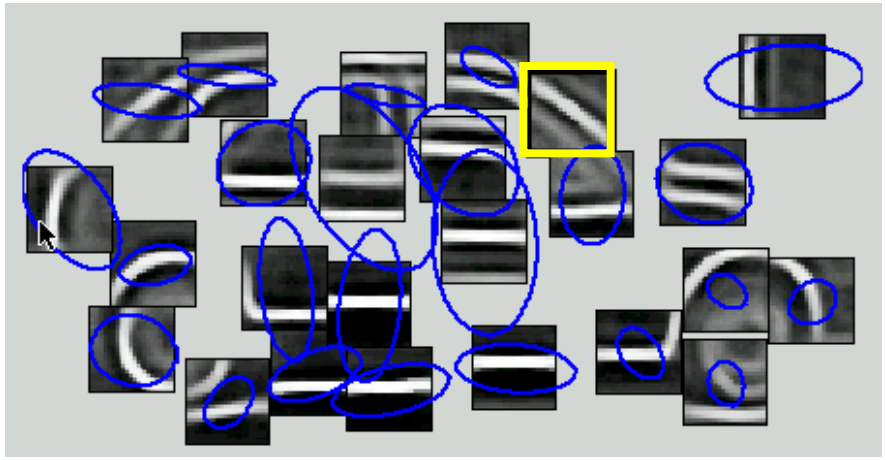
Matching results

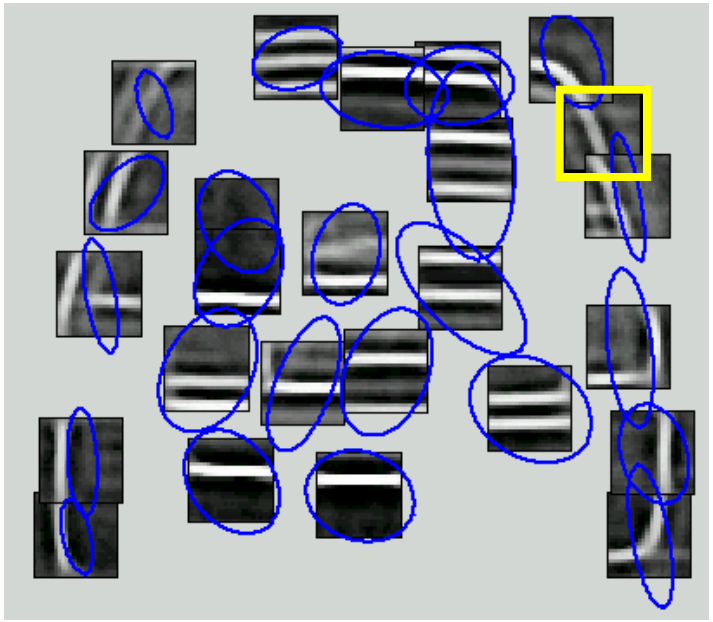


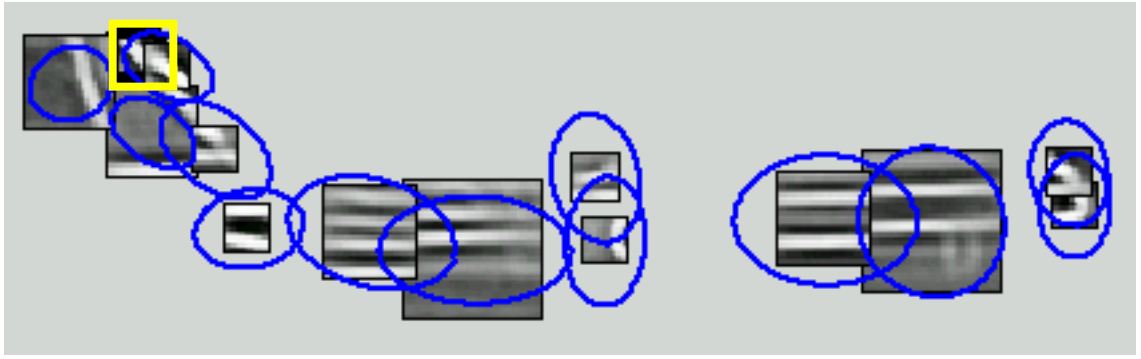
Learning the model parameters

- Easiest approach: supervised learning
  - Someone chooses the number and meaning of the parts, labels them in a bunch of training examples
  - Use this to learn the appearance and spatial models
- A lot of work has been done on unsupervised learning of these models

Some learned object models







Part-based representation

- K-fans model (D.Crandall, et.all, 2005)

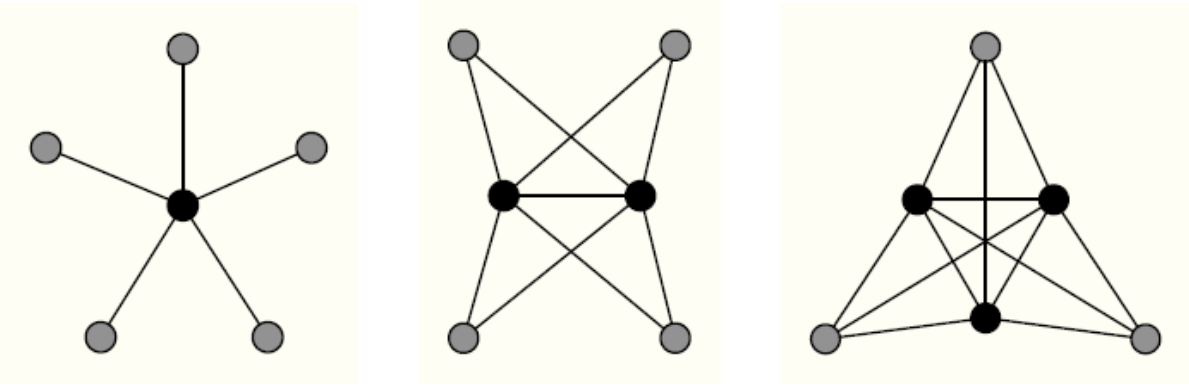
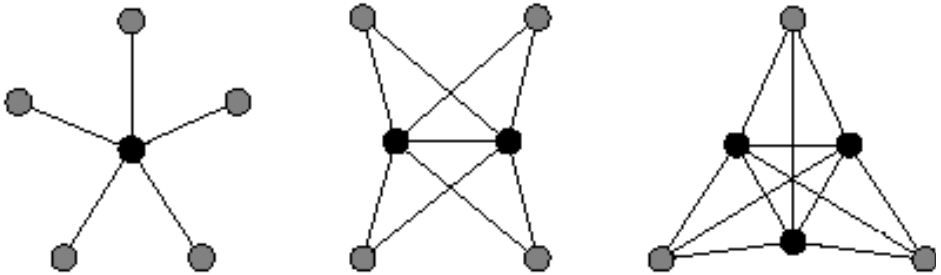
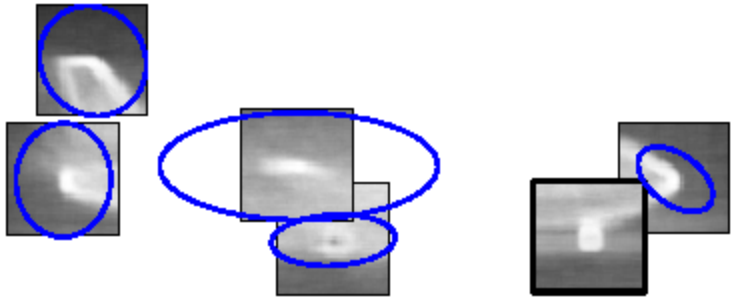


Figure 1. Some k-fans on 6 nodes. The reference nodes are shown in black while the regular nodes are shown in gray.

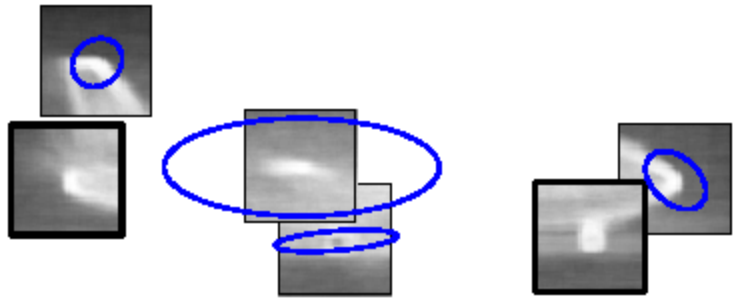
How much does shape help?

- Crandall, Felzenszwalb, Huttenlocher CVPR'05
- Shape variance decreases with increasing model complexity
- Do get some benefit from shape

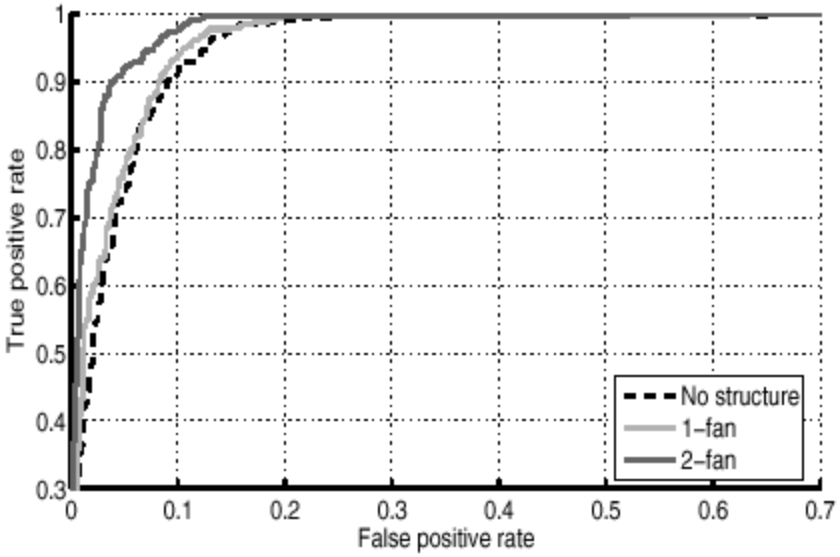




(a) Airplane, 1-fan



(b) Airplane, 2-fan



3-minute break

Context: thinking outside the (bounding) box

© Oliva & Torralba

Slides courtesy Alyosha Efros

Eye of the Beholder

Claude Monet
Gare St.Lazare Paris, 1877

Eye of the Beholder

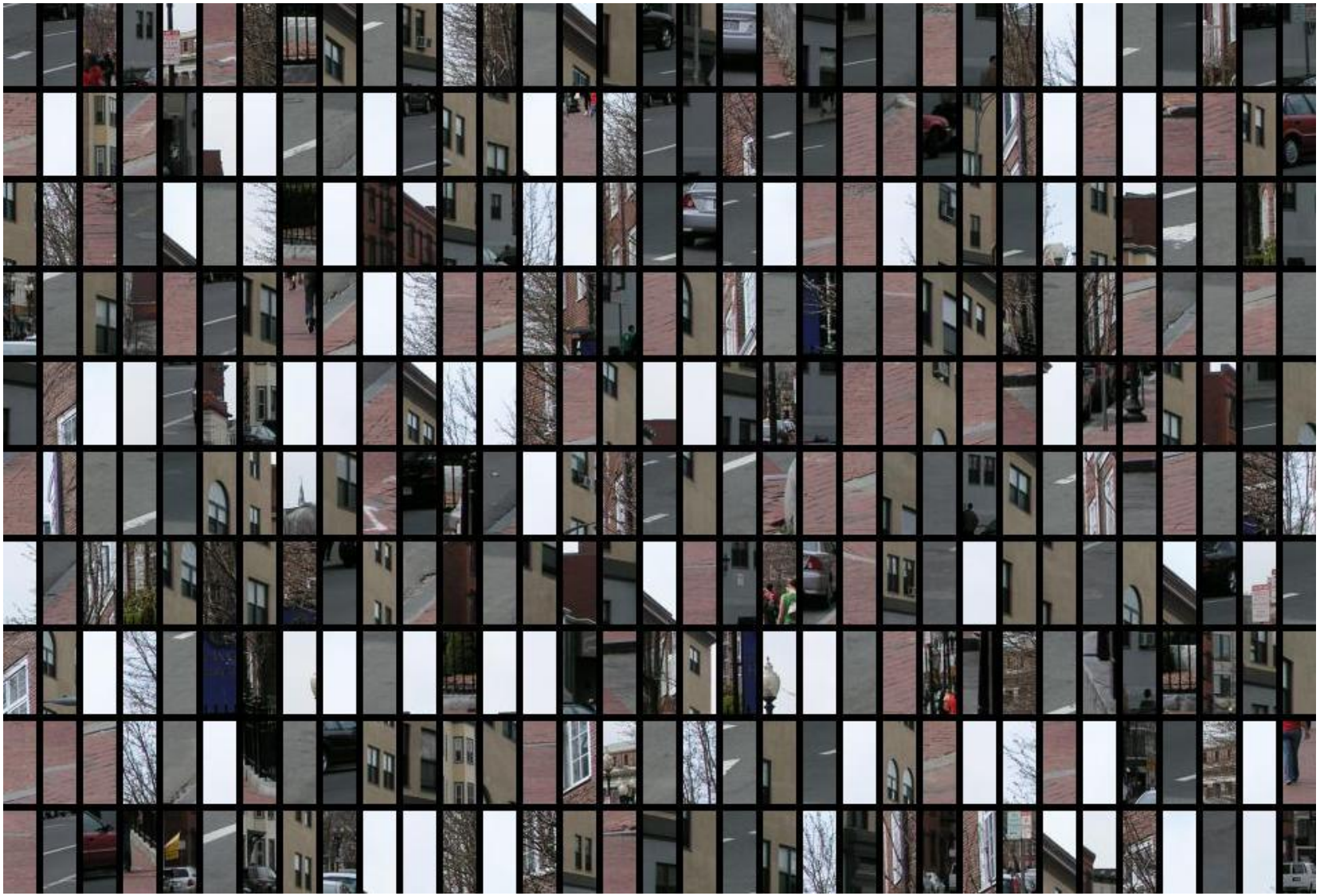
where did it go?

Seeing less than you think…

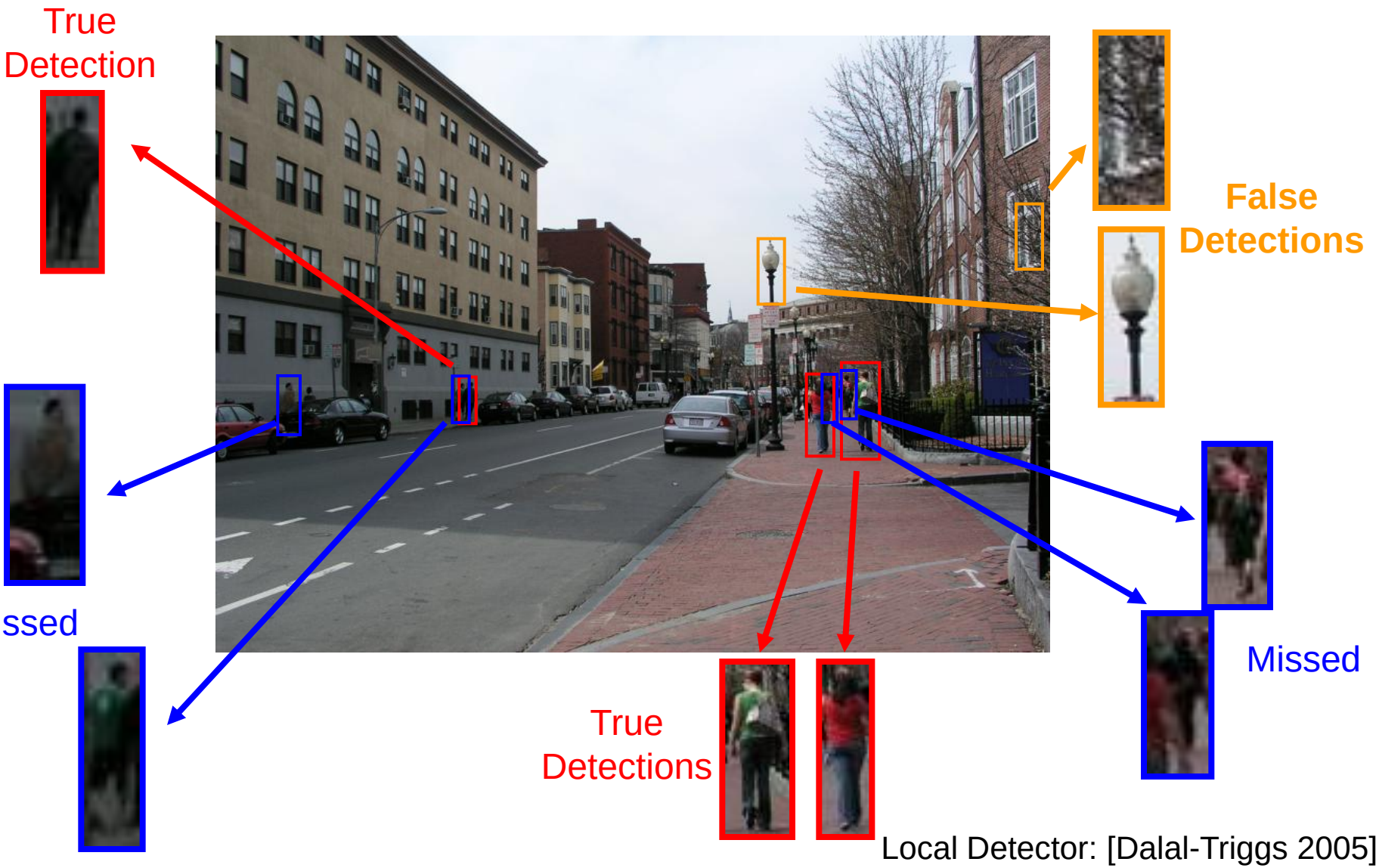
Seeing less than you think…

“The Miserable Life of a Person Detector”

What the Detector Sees



What the Detector Does



Local Detector: [Dalal-Triggs 2005]

with hundreds of categories…

If we have 1000 categories (detectors), and each detector produces 1 FP every 10 images, we will have 100 false alarms per image… pretty much garbage…

Slide by Antonio Torralba

Context to the rescue!

We know there is a keyboard present in this scene even if we cannot see it clearly.

We know there is no keyboard present in this scene

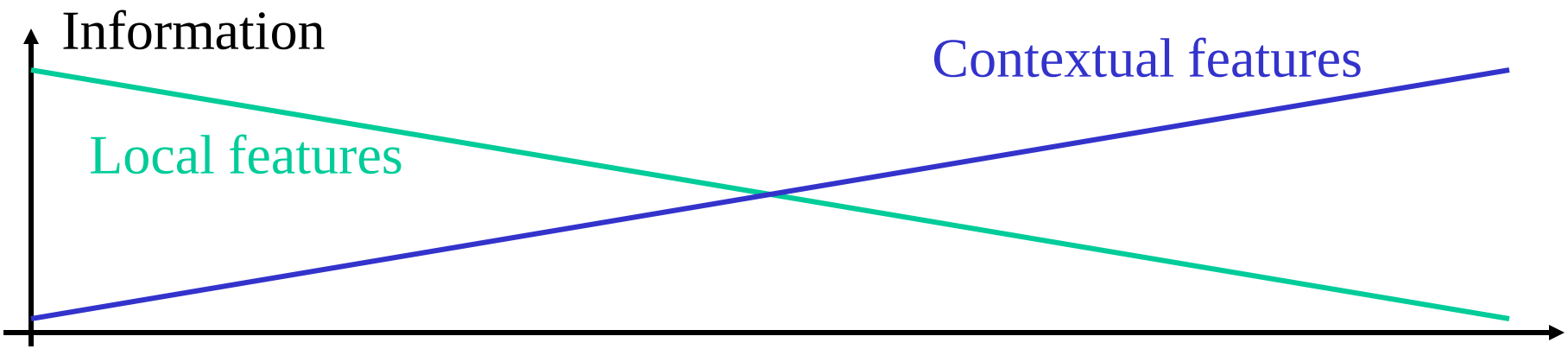
… even if there is one indeed.

Slide by Antonio Torralba

When is context helpful?





Slide by Antonio Torralba

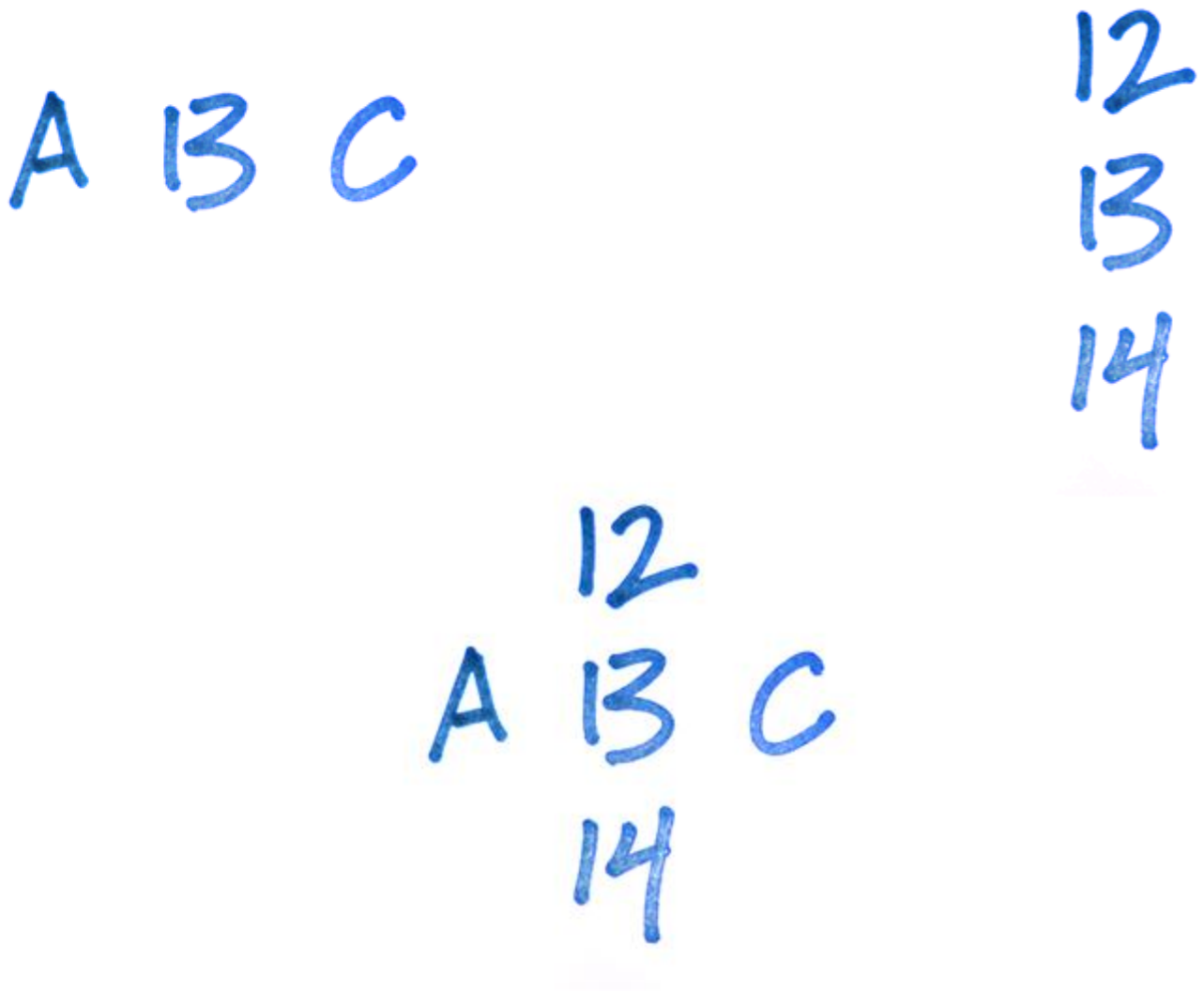
Is it just for small / blurry things?

Slide by Antonio Torralba

Is it just for small / blurry things?

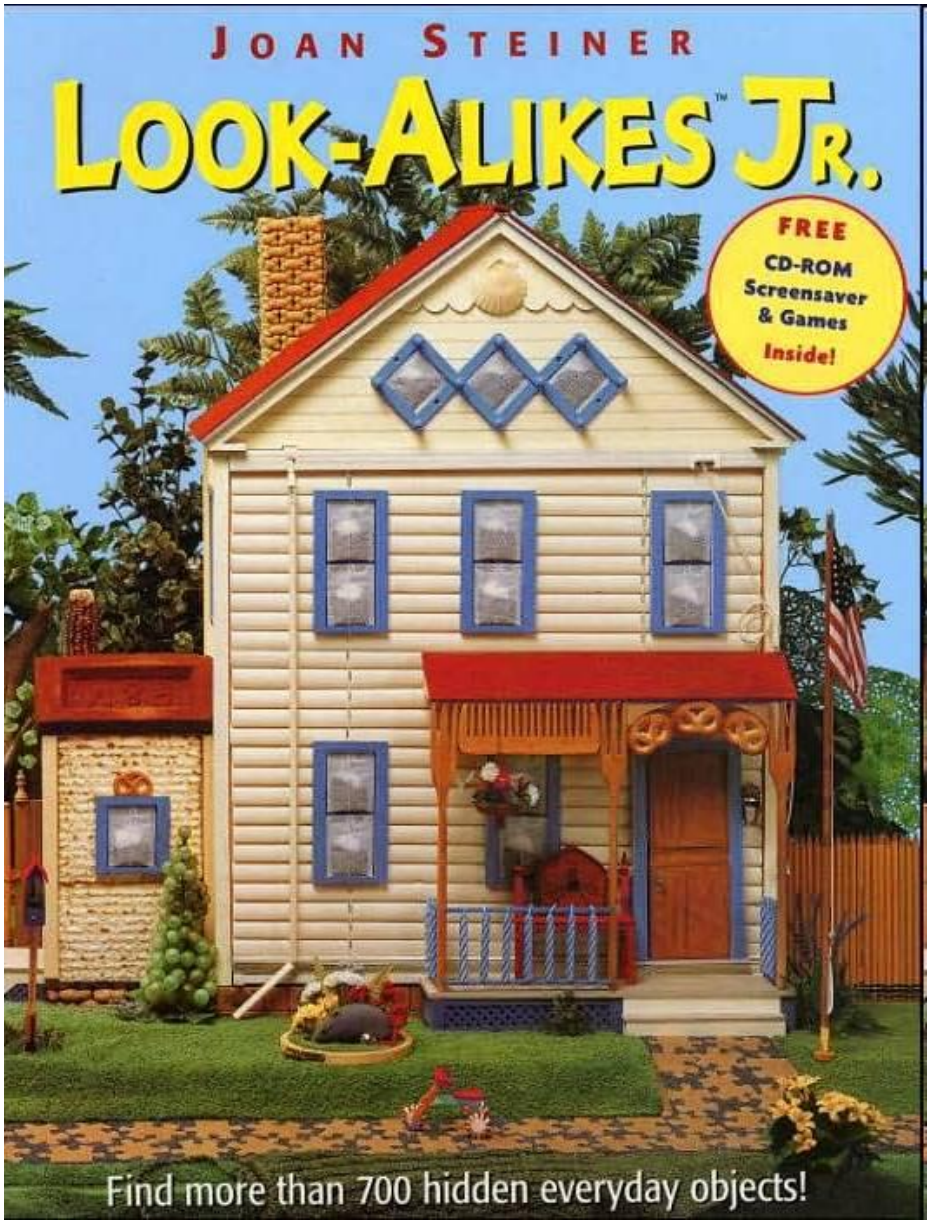
Slide by Antonio Torralba

Is it just for small / blurry things?



Slide by Antonio Torralba

Context is hard to fight!



Thanks to Paul Viola for showing me these

more “Look-alikes”

Don’t even need to see the object

Slide by Antonio Torralba

Don’t even need to see the object

Chance ~ 1/30000

Slide by Antonio Torralba

But object can say a lot about the scene

The influence of an object extends beyond its physical boundaries

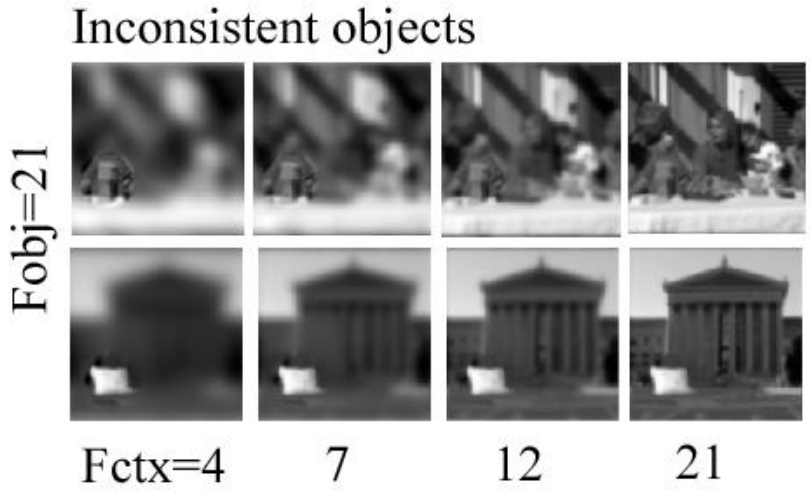
TRENDS in Cognitive Sciences

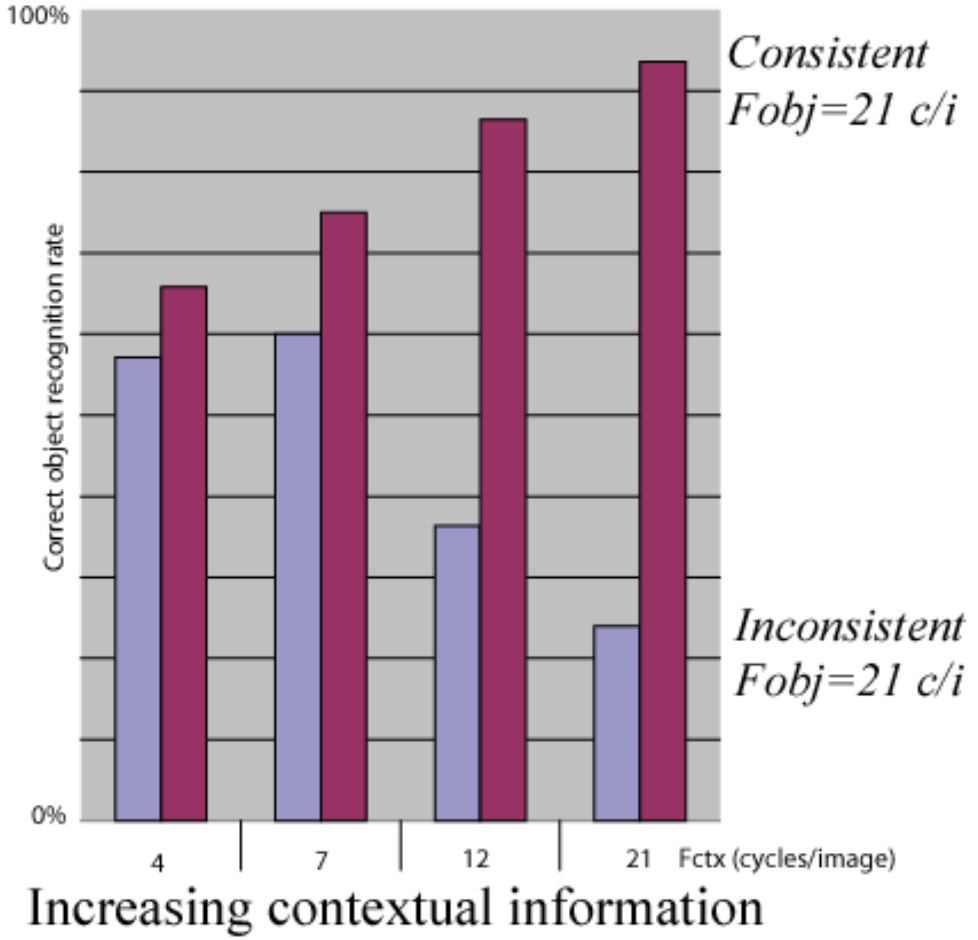
Object priming

Increasing contextual information

Torralba, Sinha, Oliva, VSS 2001

Object priming



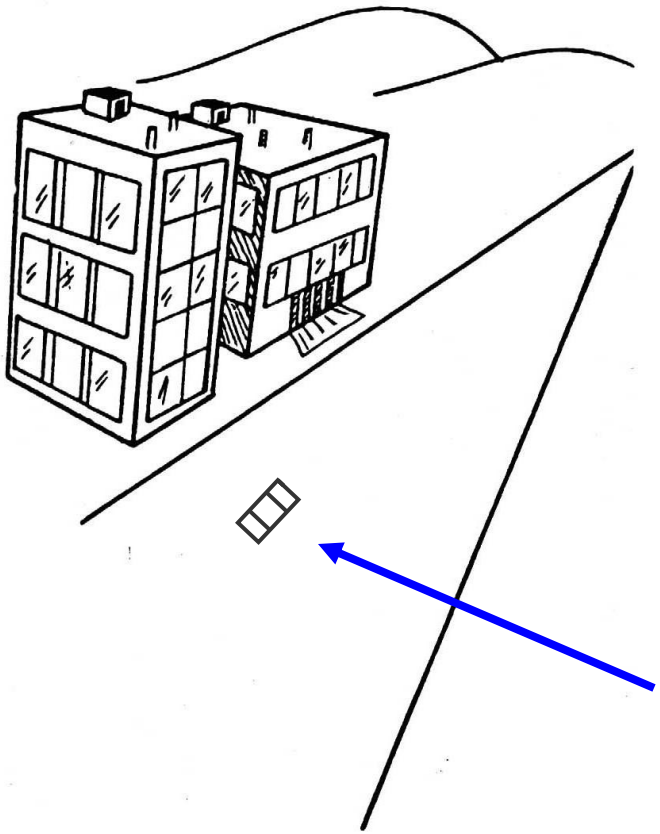


Torralba, Sinha, Oliva, VSS 2001

Why context is important?

- Typical answer: to “guess” small / blurry objects based on a prior
  - most current vision systems

So, you think it's so simple?

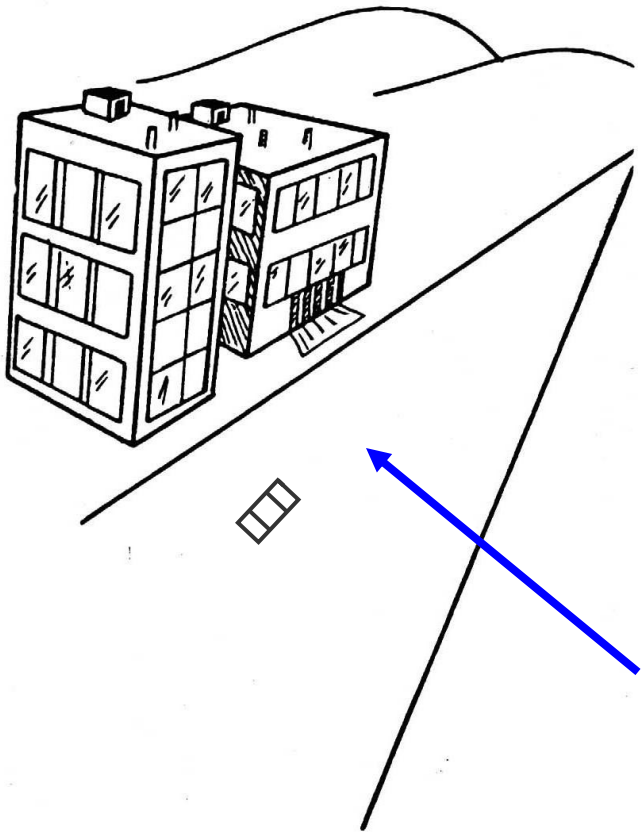


What is this?

slide by Takeo Kanade

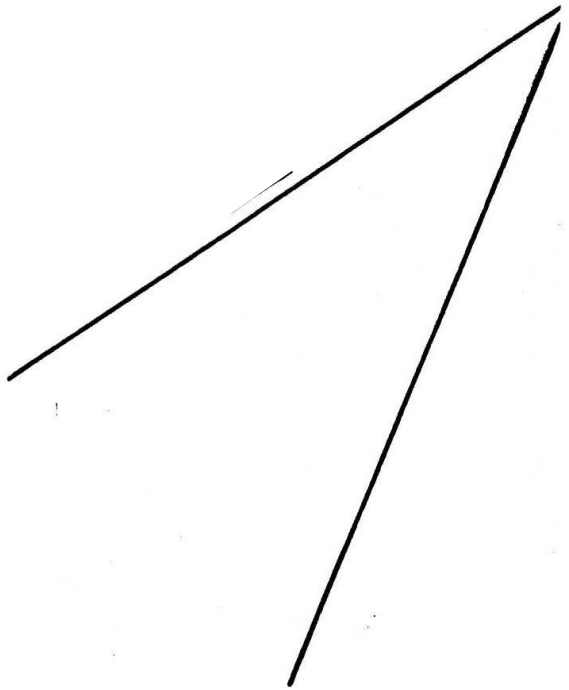
Why is this a car?

…because it's on the road!

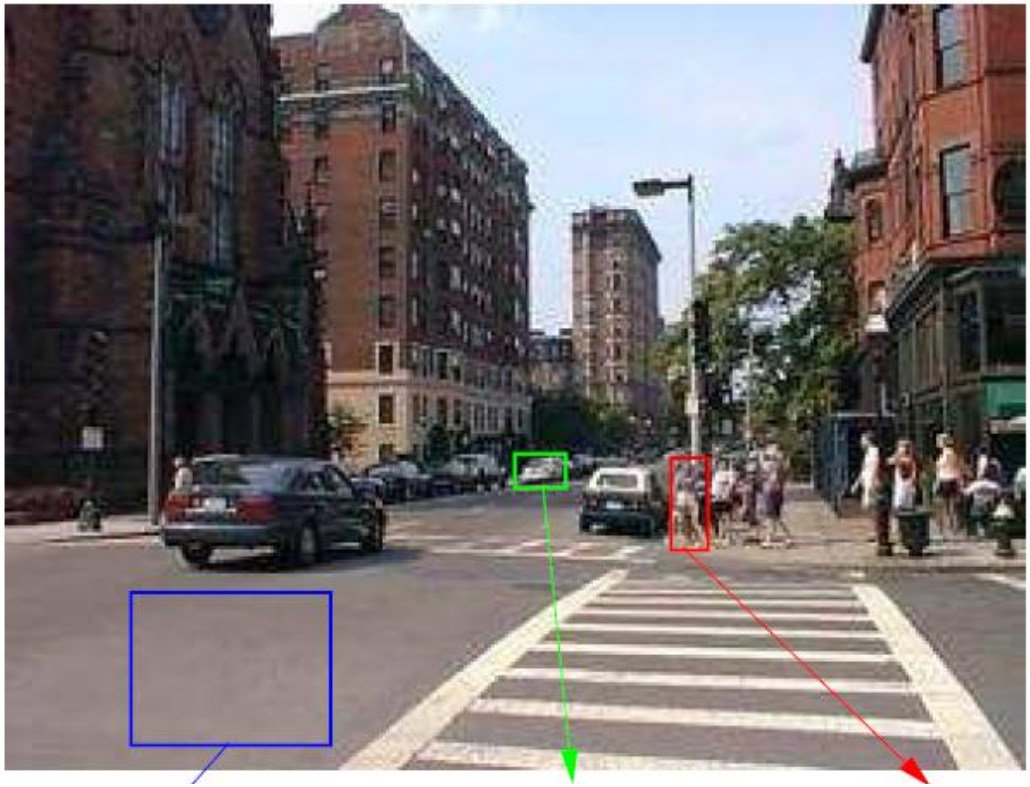


Why is this road?

Why is this a road?



Same problem in real scenes



Why context is important?

- Typical answer: to “guess” small / blurry objects based on a prior
  - Most current vision systems

- Deeper answer: to make sense of the visual world
  - much work yet to be done!

Why context is important?

- To resolve ambiguity
  - Even high-res objects can be ambiguous
    - e.g. there are more people than faces in the world!
  - There are 30,000+ objects, but only a few dozen can happen within a given scene

- To notice “unusual” things
  - Prevents mental overload
- To infer function of unknown object

Sources of Context

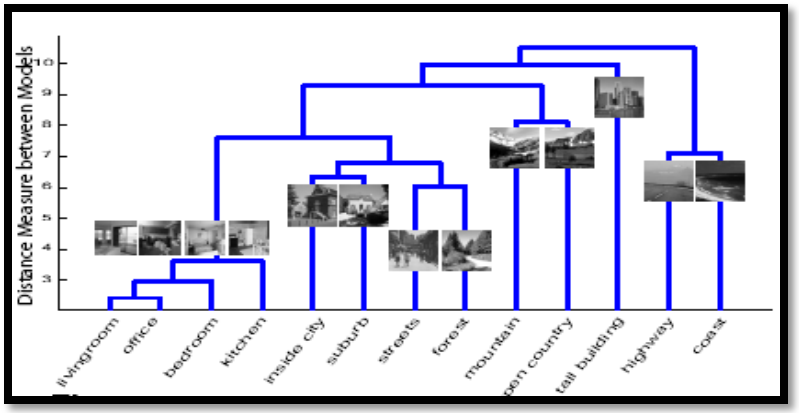
Local Pixel Context	window surround, image neighborhoods, object boundary/shape
Scene Gist Context	global image statistics
Geometric Context	3D scene layout, support surface, surface orientations, occlusions, contact points, etc.
Semantic Context	event/activity depicted, scene category, objects present in the scene, keywords
Photogrammetric Context	camera height, orientation, focal length, lens distortion, radiometric response function
Illumination Context	sun direction, sky color, cloud cover, shadow contrast, etc
Weather Context	current/recent precipitation, wind speed/direction, temperature, the season, etc.
Geographic Context	GPS location, terrain type, land use category, elevation, population density, etc.
Temporal Context	nearby frames (if video), temporally proximal images, videos of similar scenes
Cultural Context	photographer bias, dataset selection bias, visual clichés, etc

Table 1. Taxonomy of various sources of contextual information.

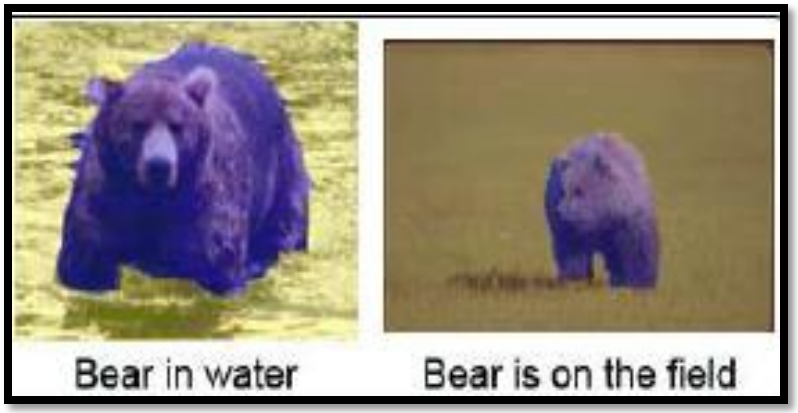
from Divvala et al, CVPR 2009

Semantic Context

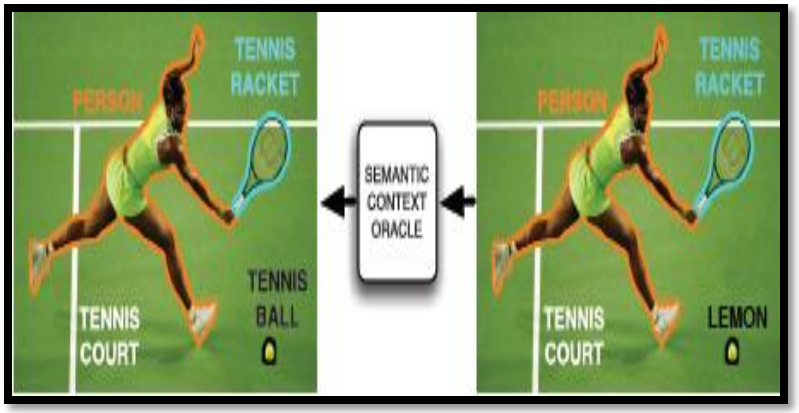
Berg et al. 2004



Fei-Fei and Perona 2005

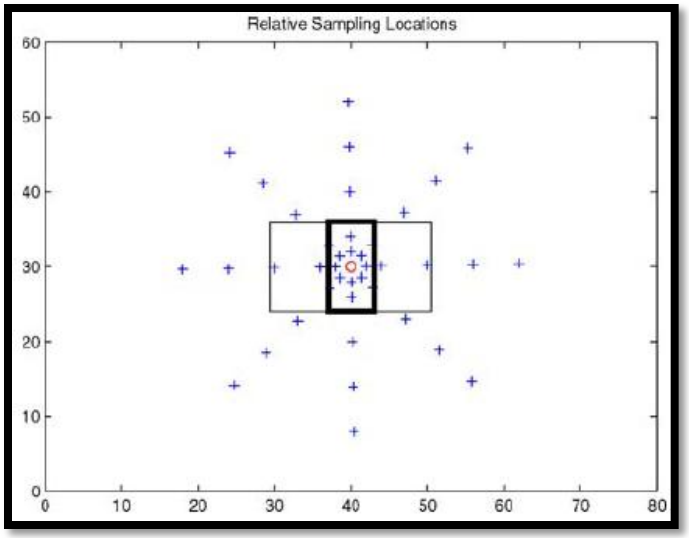


Gupta & Davis 2008

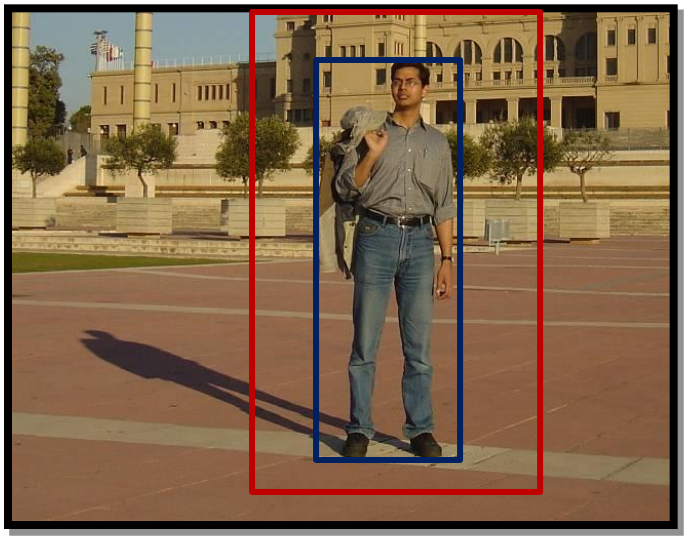


Rabinovich et al. 2007

Local Pixel Context



Wolf & Bileschi 2006



Dalal & Triggs 2005

Scene Gist Context

Global (low-dimensional) scene statistics

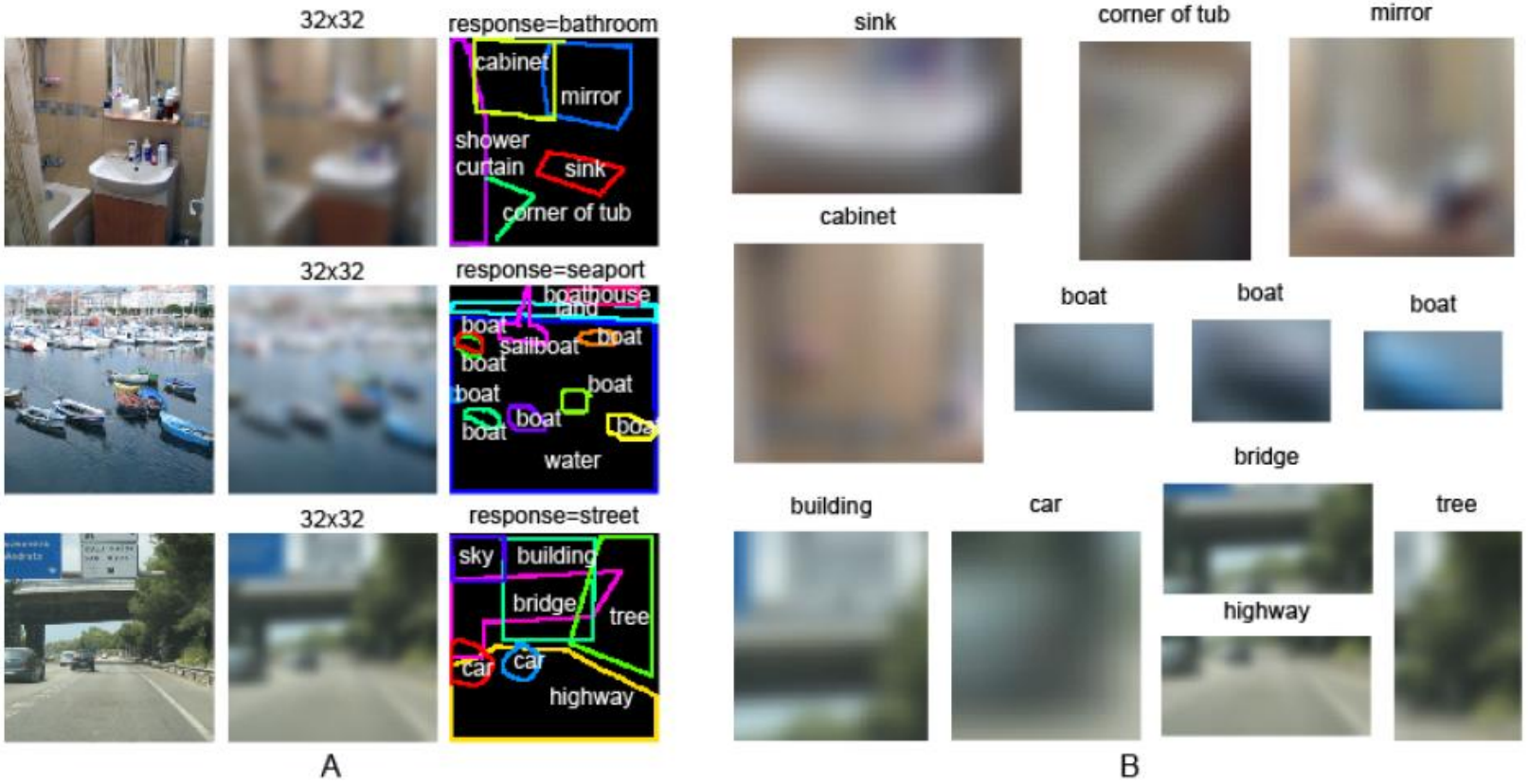
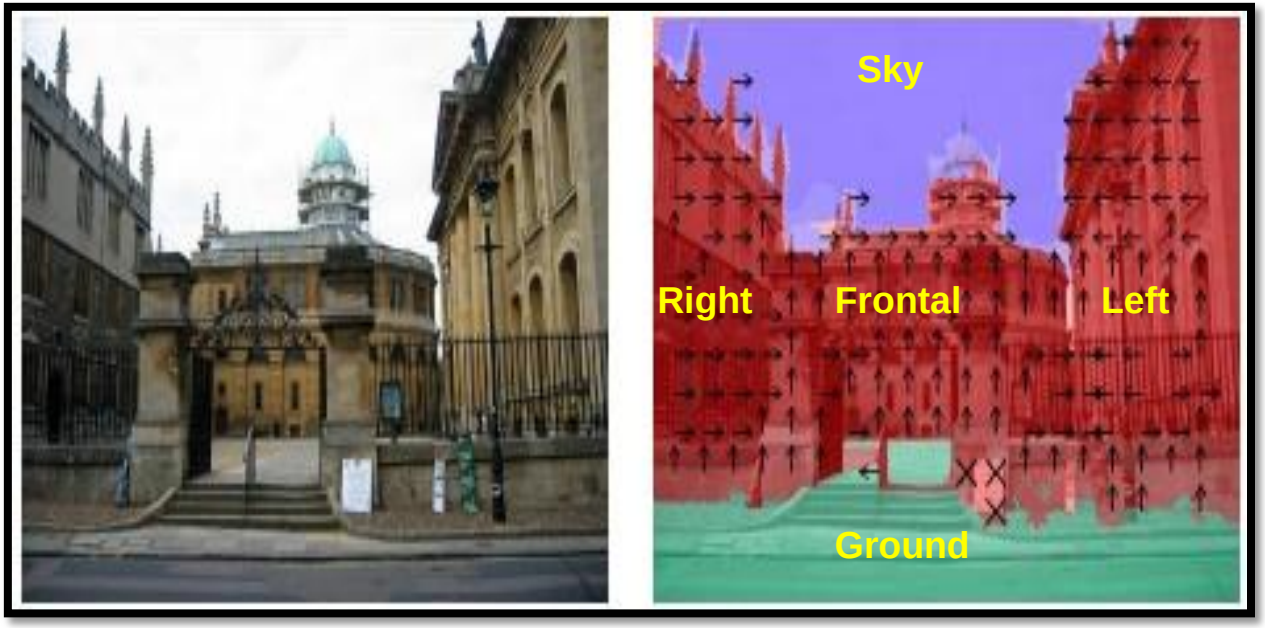


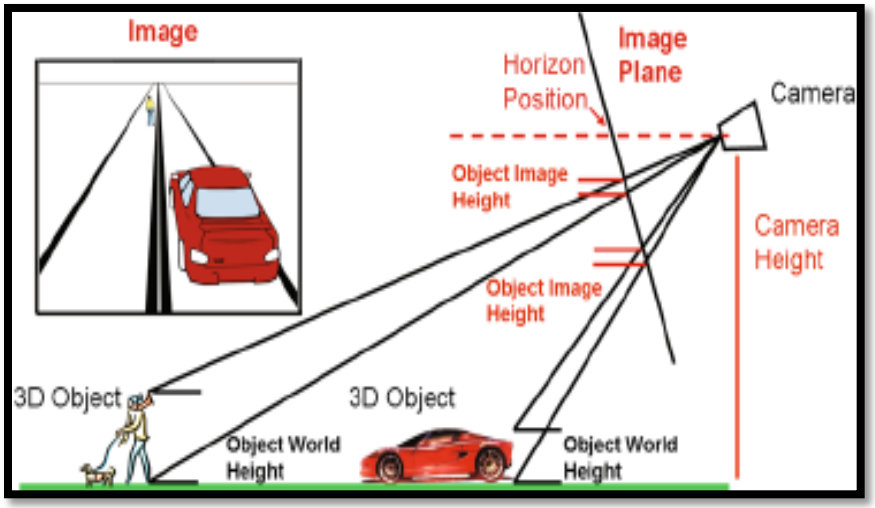
Figure 8: Images at a resolution of 32x32 pixels and the segmentations provided by the participants. Figure B shows some of the recognized objects cropped. Many of those objects become unrecognizable once they are extracted from the context.

Geometric Context

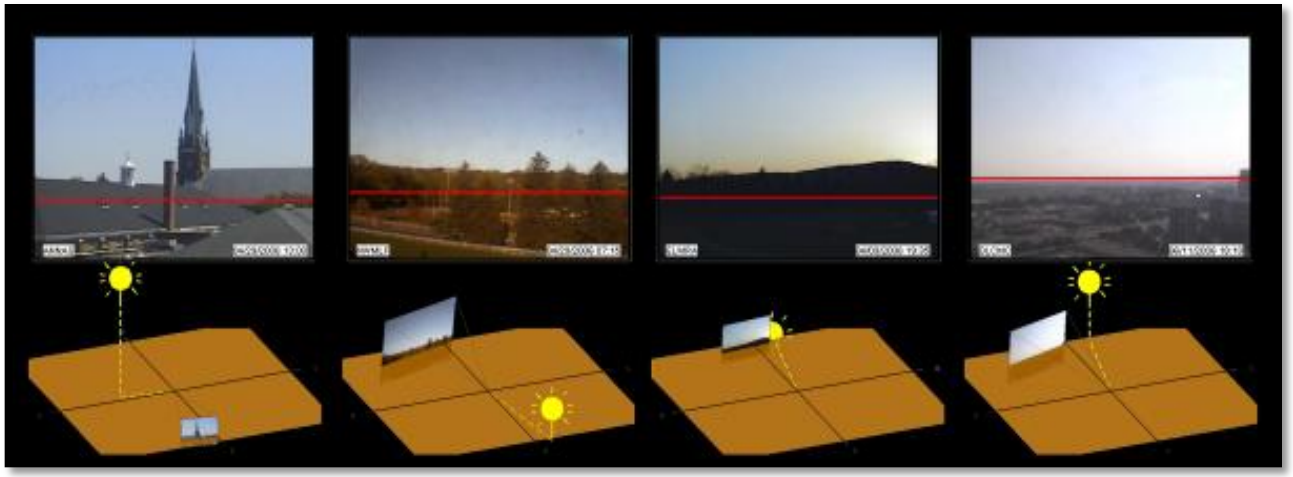


Geometric Context [Hoiem *et al.* ’2005]

Photogrammetric Context



Hoiem et al 2006

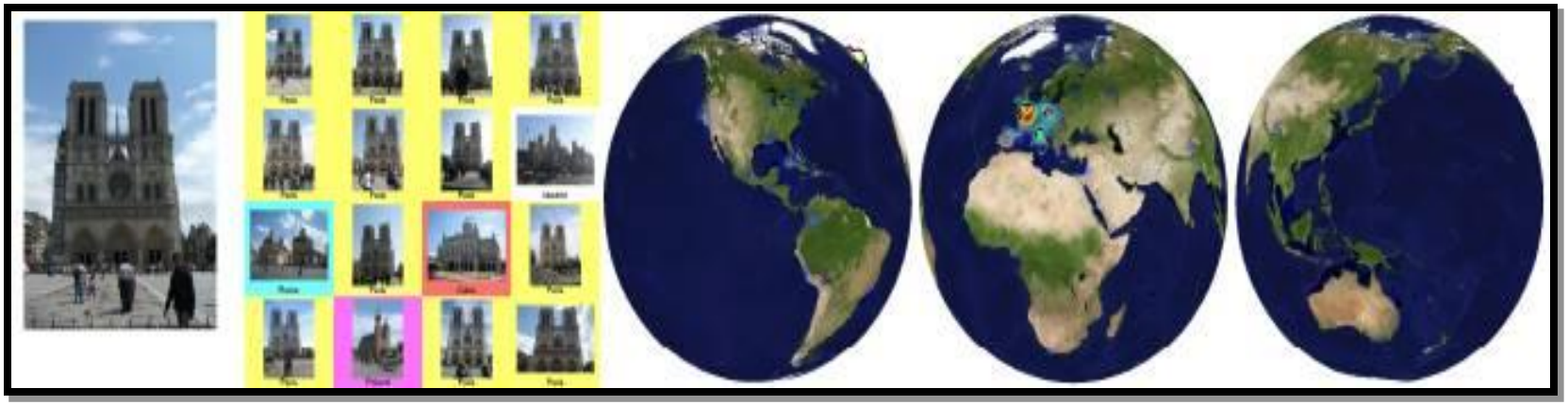


Lalonde et al. 2008

Illumination and Weather Context

Illumination context (Lalonde et al)

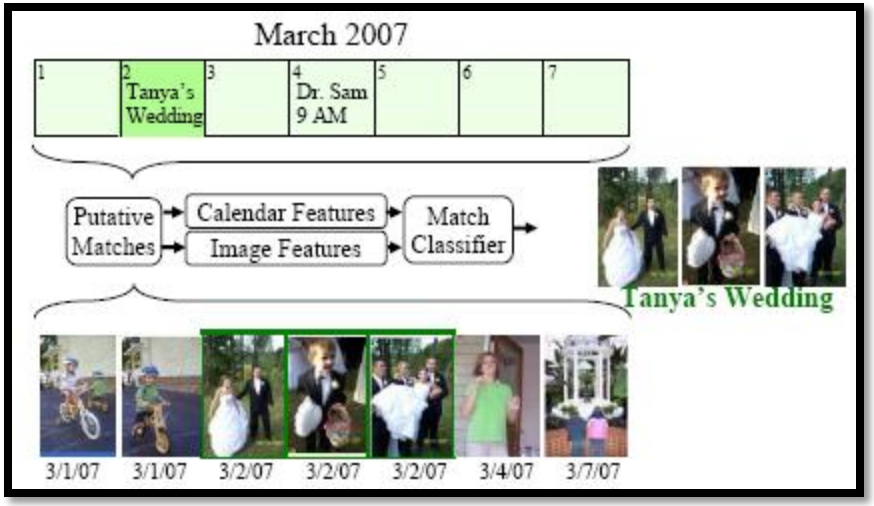
Geographic Context



Hays & Efros 2008

Temporal Context

Liu et al. 2008



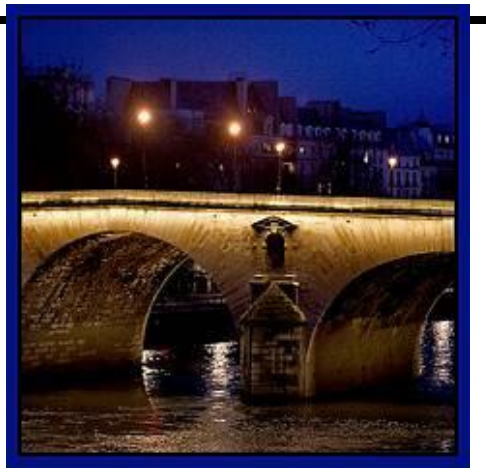
Gallagher et al 2008

Cultural Context

Photographer Bias

Society Bias

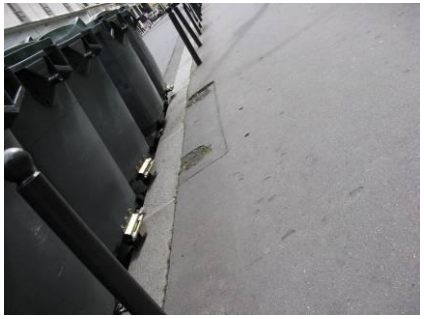
Flickr Paris



“uniformly sampled” Paris



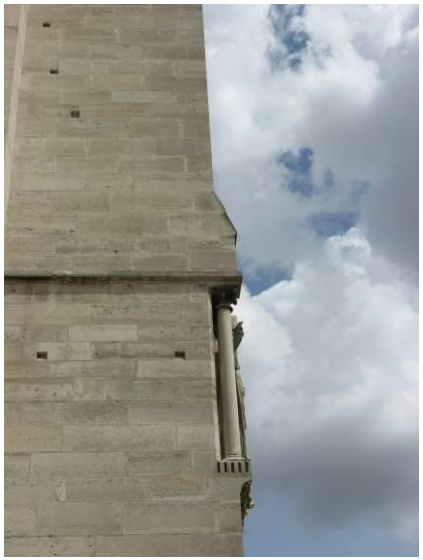






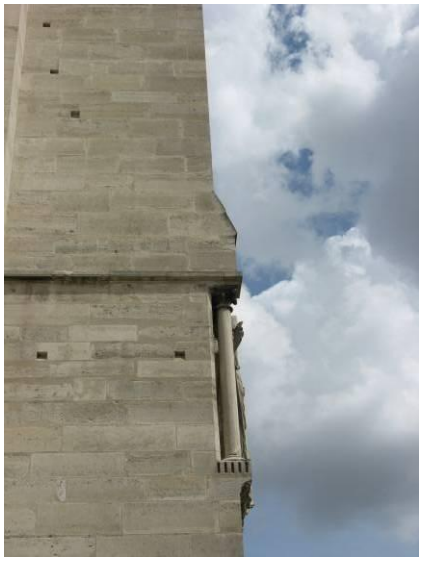



…or Notre Dame



Why Photographers are Biased?

People want their pictures to be recognizable and/or interesting



vs.

Why Photographers are Biased?

People follow photographic conventions

vs.

Simon & Seitz 2008

“100 Special Moments” by Jason Salavon

Little Leaguer

Kids with Santa

The Graduate

Newlyweds