Project presentations

Dec 6 and Dec 11

Dec 13 reports due

Assigned days in the website.
Sent us email if there is a mistake in the title/group members or a time conflict.
Presentations

4 Min + 1 min questions

• Send us presentation. We will run all presentations from the same computer.
How to give a talk

http://www.cs.berkeley.edu/~messer/Bad_talk.html

http://www-psych.stanford.edu/~lera/talk.html
First, some bad news

The more you work on a talk, the better it gets: if you work on it for 1 day, the talk you give will be better than if you had only worked on it for 1 hour. If you work on it for 2 days, it will be better still. 7 days, better yet…
All talks are important

There are no unimportant talks.
There are no big or small audiences.

Prepare each talk with the same enthusiasm.
How to give a talk

Delivering:
Look at the audience! Try not to talk to your laptop or to the screen. Instead, look at the other humans in the room.
You have to believe in what you present, be confident... even if it only lasts for the time of your presentation.
Do not be afraid to acknowledge limitations of whatever you are presenting. Limitations are good. They leave job for the people to come. Trying to hide the problems in your work will make the preparation of the talk a lot harder and your self confidence will be hurt.
Let the audience see your personality

• They want to see you enjoy yourself.
• They want to see what you love about the work.
• People really respond to the human parts of a talk. Those parts help the audience with their difficult task of listening to an hour-long talk on a technical subject. What was easy, what was fun, what was hard about the work?
• Don’t be afraid to be yourself and to be quirky.
The different kinds of talks you’ll have to give as a researcher

• 2-5 minute talks
• 20 -30 minute conference presentations
• 30-60 minute colloquia
How to give a talk

**Talk organization:** here there are as many theories as there are talks. Here there are some extreme advices:

1. Go into details / only big picture
2. Go in depth on a single topic / cover as many things as you can
3. Be serious (never make jokes, maybe only one) / be funny (it is just another form of theater)

Corollary: ask people for advice, but at the end, if will be just you and the audience. Chose what fits best your style.

What everybody agree on is that you have to practice in advance (the less your experience, the more you have to practice). Do it with an audience or without, but practice.

The best advice I got came from Yair Weiss while preparing my job talk:

“just give a good talk”
How to give the project class talk

Initial conditions:
• I started with a great idea
• It did not work
• The day before the presentation I found 40 papers that already did this work
• Then I also realized that the idea was not so great

How do I present?
• Just give a good talk
Sources on writing technical papers

24. The opening paragraph should be your best paragraph, and its first sentence should be your best sentence. If a paper starts badly, the reader will wince and be resigned to a difficult job of fighting with your prose. Conversely, if the beginning flows smoothly, the reader will be hooked and won’t notice occasional lapses in the later parts. Probably the worst way to start is with a sentence of the form “An $x$ is $y$.” For example,

   Bad: An important method for internal sorting is quicksort.
   Good: Quicksort is an important method for internal sorting, because ...
   Bad: A commonly used data structure is the priority queue.
   Good: Priority queues are significant components of the data structures needed for many different applications.
Knuth on equations

13. Many readers will skim over formulas on their first reading of your exposition. Therefore, your sentences should flow smoothly when all but the simplest formulas are replaced by “blah” or some other grunting noise.
The paper impact curve

- Lots of impact
- Nothing

Paper quality:
- So-so
- Ok
- Pretty good
- Creative, original and good.
Lecture 22
Scene understanding
Beyond single classes

- Multiclass
- Multiview
- Datasets
Beyond single classes

- Multiclass
- Multiview
- Datasets
Shared features

• Is learning the object class 1000 easier than learning the first?

• Can we transfer knowledge from one object to another?

• Are the shared properties interesting by themselves?
Reusable Parts

Krempp, Geman, & Amit “Sequential Learning of Reusable Parts for Object Detection”. TR 2002

Goal: Look for a vocabulary of edges that reduces the number of features.

Examples of reused parts
Additive models and boosting

• Independent binary classifiers:

- Screen detector
- Car detector
- Face detector

• Binary classifiers that share features:

- Screen detector
- Car detector
- Face detector
Generalization as a function of object similarities

12 unrelated object classes

12 viewpoints

Area under ROC

K = 2.1

Area under ROC

K = 4.8

Number of training samples per class

Number of training samples per class

Torralba, Murphy, Freeman. CVPR 2004. PAMI 2007
Beyond single classes

- Multiclass
- **Multiview**
- Datasets
Class experiment
Class experiment

**Experiment 1:** draw a horse (the entire body, not just the head) in a white piece of paper.

Do not look at your neighbor! You already know how a horse looks like... no need to cheat.
Experiment 2: draw a horse (the entire body, not just the head) but this time chose a viewpoint as weird as possible.
3D object categorization

Despite we can categorize all three pictures as being views of a horse, the three pictures do not look as being equally typical views of horses. And they do not seem to be recognizable with the same easiness.
Canonical Perspective

Experiment (Palmer, Rosch & Chase 81): participants are shown views of an object and are asked to rate “how much each one looked like the objects they depict” (scale; 1=very much like, 7=very unlike)

In a recognition task, reaction time correlated with the ratings.

Canonical views are recognized faster at the entry level.

From Vision Science, Palmer
Clocks are preferred as purely frontal
Object representations

**Explicit 3D models**: use volumetric representation. Have an explicit model of the 3D geometry of the object.

Appealing but hard to get it to work…
Object representations

**Implicit 3D models**: matching the input 2D view to view-specific representations.

(b) For cars, classifiers are trained on 8 viewpoints

Not very appealing but somewhat easy to get it to work…
Beyond single classes

• Multiclass
• Multiview
• Datasets
The PASCAL Visual Object Classes

In 2007, the twenty object classes that have been selected are:

*Person*: person

*Animal*: bird, cat, cow, dog, horse, sheep

*Vehicle*: aeroplane, bicycle, boat, bus, car, motorbike, train

*Indoor*: bottle, chair, dining table, potted plant, sofa, tv/monitor

M. Everingham, Luc van Gool, C. Williams, J. Winn, A. Zisserman 2007
Caltech 101 and 256

Fei-Fei, Fergus, Perona, 2004

Griffin, Holub, Perona, 2007
LabelMe

Tool went online July 1st, 2005
530,000 object annotations collected
Labelme.csail.mit.edu
80,000,000 images

75,000 non-abstract nouns from WordNet

7 Online image search engines

And after 1 year downloading images

Google: 80 million images

A. Torralba, R. Fergus, W.T. Freeman. PAMI 2008
• An ontology of images based on WordNet
• ImageNet currently has
  – 13,000+ categories of visual concepts
  – 10 million human-cleaned images (~700im/categ)
  – 1/3+ is released online @ www.image-net.org

Deng, Dong, Socher, Li & Fei-Fei, CVPR 2009
Dataset biases

Torralba, Efros. Unbiased Look at Dataset Bias. CVPR 2011
The texture

The object

The scene
The detector challenge

By looking at the output of a detector on a random set of images, can you guess which object is it trying to detect?
What object is the detector trying to detect?

By looking at the output of a detector on a random set of images, can you guess which object is it trying to detect?
What object is the detector trying to detect?

By looking at the output of a detector on a random set of images, can you guess which object is it trying to detect?
Microwave

Top 8 out of 4317 images

P. Felzenszwalb, D. McAllester, and D. Ramanan. CVPR, 2008
Microwave & refrigerator
What object is hidden behind the red box?
Objects in context

Carbonetto, de Freitas & Barnard (2004)

Sudderth, Torralba, Wilsky, Freeman (2005)

Kumar, Hebert (2005)


Toralba, Sinha (2001)

Torralba Murphy Freeman (2004)

Rabinovich et al (2007)

Hoiem, Efros, Hebert (2005)

Heitz and Koller (2008)

Desai, Ramanan, and Fowlkes (2009)
Increasing the context strength

- **4x4**
- **8x8**
- **16x16**
- **32x32**
- **64x64**
Scenes rule over objects

3D percept is driven by the scene, which imposes its ruling to the objects
Mary Potter (1976)

Mary Potter (1975, 1976) demonstrated that during a rapid sequential visual presentation (100 msec per image), a novel picture is instantly **understood** and observers seem to comprehend a lot of visual information.
Instructions: 9 photographs will be shown for half a second each. Your task is to memorize these pictures.
Which of the following pictures have you seen?

If you have seen the image clap your hands once

If you have not seen the image do nothing
Have you seen this picture?
Have you seen this picture?
NO
Have you seen this picture?
Have you seen this picture?
Have you seen this picture?
Yes
Have you seen this picture?
You have seen these pictures

You were tested with these pictures
The gist of the scene

In a glance, we remember the meaning of an image and its global layout but some objects and details are forgotten
Scene Categorization

Oliva and Torralba, 2001

Coast  Forest  Highway  Inside City  Mountain  Open Country  Street  Tall Building

Fei Fei and Perona, 2005

Bedroom  Kitchen  Living Room  Office  Suburb

Lazebnik, Schmid, and Ponce, 2006

Industrial  Store

15 Scene Database
Which are the important elements?

Different content (i.e. objects), different spatial layout
Which are the important elements?

- Cabinets
- Ceiling
- Windows
- Seats

Similar objects, and similar spatial layout

Different lighting, different materials, different “stuff”
What can be an alternative to objects?
Scene emergent features

"Recognition via features that are not those of individual objects but “emerge” as objects are brought into relation to each other to form a scene.” – Biederman 81

From “on the semantics of a glance at a scene”, Biederman, 1981
Examples of scene emergent features

Suggestive edges and junctions

Simple geometric forms

Biederman, 1981

Biederman, 1981

Brunet & Potter, 1969

Oliva & Torralba, 2001

Blobs

Textures ~ Sketch
Ensemble statistics

Ariely, 2001, Seeing sets: Representation by statistical properties
Chong, Treisman, 2003, Representation of statistical properties

Conclusion: observers had more accurate representation of the mean than of the individual members of the set.
Global image descriptors
Global image descriptors

Bag of words

Non localized textons

Spatially organized textures

Sivic et al., ICCV 2005
Fei-Fei and Perona, CVPR 2005

M. Gorkani, R. Picard, ICPR 1994
A. Oliva, A. Torralba, IJCV 2001

Walker, Malik. Vision Research 2004

S. Lazebnik, et al, CVPR 2006

Gist descriptor

Oliva and Torralba, 2001

- Apply oriented Gabor filters over different scales
- Average filter energy in each bin

Similar to SIFT (Lowe 1999) applied to the entire image

Gist descriptor
Gist descriptor

\[ V = \{\text{energy at each orientation and scale}\} = 6 \times 4 \text{ dimensions} \]

Oliva, Torralba. IJCV 2001
Example visual gists

Global features (I) ~ global features (I')

Oliva & Torralba (2001)
Global features

“The viewer is presented with a ‘potential image’, that is, a complex multiplicity of possible images, none of which ever finally resolves”.

Rob Pepperell
Textons

Filter bank

Vector of filter responses at each pixel

Kmeans over a set of vectors on a collection of images

Malik, Belongie, Shi, Leung, 1999
Textons

Walker, Malik, 2004

Malik, Belongie, Shi, Leung, 1999

best match

\( \chi^2 = 5.67 \)

\# occurrences in image

universal textons

\( \chi^2 = 4.17 \times 10^3 \)

\# occurrences in image

universal textons

Walker, Malik, 2004
Histogram Intersection

\[ \mathcal{I}(H(X), H(Y)) = \sum_{j=1}^{r} \min(H(X)_j, H(Y)_j) \]

Adapted from Kristen Grauman
A Support Vector Machine (SVM) learns a classifier with the form:

\[ H(x) = \sum_{m=1}^{M} a_m y_m k(x, x_m) \]

Where \( \{x_m, y_m\} \), for \( m = 1 \ldots M \), are the training data with \( x_m \) being the input feature vector and \( y_m = +1,-1 \) the class label. \( k(x, x_m) \) is the kernel and it can be any symmetric function satisfying the Mercer Theorem.

The classification is obtained by thresholding the value of \( H(x) \).

There is a large number of possible kernels, each yielding a different family of decision boundaries:

- Linear kernel: \( k(x, x_m) = x^T x_m \)
- Radial basis function: \( k(x, x_m) = \exp(-|x - x_m|^2/\sigma^2) \).
- Histogram intersection: \( k(x,x_m) = \sum_i (\min(x(i), \ x_m(i))) \)
Bag of words

Bag of words model

Spatially organized textures

65 17 23 36

20 0 0 0 0 11 1 0 2 14 0 3 3

3 0 12 4 0 0 4 16 3 6 0 11
Bag of words & spatial pyramid matching


S. Lazebnik, et al, CVPR 2006
Learning Scene Categorization

Forest path
Vs.
all

Living - room
Vs.
all
The 15-scenes benchmark

Oliva & Torralba, 2001
Fei Fei & Perona, 2005
Lazebnik, et al 2006
Scene recognition

100 training samples per class

SVM classifier in both cases

- store
- livingroom
- kitchen
- industrial
- bedroom
- office
- tall building
- street
- open country
- mountain
- inside city
- highway
- forest
- coast
- suburb

Percent correct

Recognition rate

Number training samples per class

Human performance

- all [88.1]
- phot [81.2]
- hog2x2 [81.0]
- texton histogram [77.8]
- ssim [77.2]
- gist [74.7]

- sparse SIFT histograms [56.6]
- geometric classification map [55.0]
- straight line histograms [50.9]
SUN Dataset Project

We want:
• Large variety of scene categories (we want them all)
• Lots of objects categories
• Multi-object scenes

1. We take all scene words from a dictionary
2. We download images and clean the categories
3. We segment all the images

Krista Ehinger    Jianxiong Xiao

Xiao, Hays, Ehinger, Oliva, Torralba; CVPR 2010
397 Well-sampled Categories
Performance with 400 categories

Xiao, Hays, Ehinger, Oliva, Torralba; maybe 2010
Training images

Abbey

Airplane cabin

Airport terminal

Alley

Amphitheater

Xiao, Hays, Ehinger, Oliva, Torralba; maybe 2010
<table>
<thead>
<tr>
<th>Category</th>
<th>Training images</th>
<th>Correct classifications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abbey</td>
<td><img src="image1" alt="Abbey Images" /></td>
<td><img src="image2" alt="Abbey Correct Images" /></td>
</tr>
<tr>
<td>Airplane cabin</td>
<td><img src="image3" alt="Airplane Images" /></td>
<td><img src="image4" alt="Airplane Correct Images" /></td>
</tr>
<tr>
<td>Airport terminal</td>
<td><img src="image5" alt="Airport Images" /></td>
<td><img src="image6" alt="Airport Correct Images" /></td>
</tr>
<tr>
<td>Alley</td>
<td><img src="image7" alt="Alley Images" /></td>
<td><img src="image8" alt="Alley Correct Images" /></td>
</tr>
<tr>
<td>Amphitheater</td>
<td><img src="image9" alt="Amphitheater Images" /></td>
<td><img src="image10" alt="Amphitheater Correct Images" /></td>
</tr>
<tr>
<td>Location</td>
<td>Training images</td>
<td>Correct classifications</td>
</tr>
<tr>
<td>------------------</td>
<td>-----------------</td>
<td>-------------------------</td>
</tr>
<tr>
<td>Abbey</td>
<td><img src="image1" alt="Abbey images" /></td>
<td><img src="image2" alt="Correct Abbey classifications" /></td>
</tr>
<tr>
<td>Airplane cabin</td>
<td><img src="image4" alt="Airplane cabin images" /></td>
<td><img src="image5" alt="Correct classifications" /></td>
</tr>
<tr>
<td>Airport terminal</td>
<td><img src="image7" alt="Airport terminal images" /></td>
<td><img src="image8" alt="Correct classifications" /></td>
</tr>
<tr>
<td>Alley</td>
<td><img src="image10" alt="Alley images" /></td>
<td><img src="image11" alt="Correct classifications" /></td>
</tr>
<tr>
<td>Amphitheater</td>
<td><img src="image13" alt="Amphitheater images" /></td>
<td><img src="image14" alt="Correct classifications" /></td>
</tr>
</tbody>
</table>

Xiao, Hays, Ehinger, Oliva, Torralba; maybe 2010
Categories or a continuous space?

Check poster by Malisiewicz, Efros
Categories or a continuous space?

From the city to the mountains in 10 steps
Objects in context
Is local information enough?
Is local information even enough?
Is local information even enough?

Local features

Contextual features

Distance

Information

Local features

Contextual features
The system does not care about the scene, but we do...

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.
The multiple personalities of a blob
The multiple personalities of a blob
A B C
Look-Alikes by Joan Steiner
Look-Alikes by Joan Steiner
The importance of context

- Cognitive psychology
  - Palmer 1975
  - Biederman 1981
  - ...

- Computer vision
  - Noton and Stark (1971)
  - Hanson and Riseman (1978)
  - Barrow & Tenenbaum (1978)
  - Ohta, kanade, Skai (1978)
  - Haralick (1983)
  - Strat and Fischler (1991)
  - Bobick and Pinhanez (1995)
  - Campbell et al (1997)
Objects and Scenes

Biederman’s violations (1981):

1. **Support** (e.g., a floating fire hydrant). The object does not appear to be resting on a surface.
2. **Interposition** (e.g., the background appearing through the hydrant). The objects undergoing this violation appear to be transparent or passing through another object.
3. **Probability** (e.g., the hydrant in a kitchen). The object is unlikely to appear in the scene.
4. **Position** (e.g., the fire hydrant on top of a mailbox in a street scene). The object is likely to occur in that scene, but it is unlikely to be in that particular position.
5. **Size** (e.g., the fire hydrant appearing larger than a building). The object appears to be too large or too small relative to the other objects in the scene.
# CONDOR system

Strat and Fischler (1991)

<table>
<thead>
<tr>
<th>Class</th>
<th>Context elements</th>
<th>Operator</th>
</tr>
</thead>
<tbody>
<tr>
<td>SKY</td>
<td>ALWAYS</td>
<td>ABOVE-HORIZON</td>
</tr>
<tr>
<td>SKY</td>
<td>SKY-IS-CLEAR $\land$ TIME-IS-DAY</td>
<td>BRIGHT</td>
</tr>
<tr>
<td>SKY</td>
<td>SKY-IS-CLEAR $\land$ TIME-IS-DAY $\land$ RGB-IS-AVAILABLE</td>
<td>BLUE</td>
</tr>
<tr>
<td>SKY</td>
<td>SKY-IS-OVERCAST $\land$ TIME-IS-DAY</td>
<td>BRIGHT</td>
</tr>
<tr>
<td>SKY</td>
<td>SKY-IS-OVERCAST $\land$ TIME-IS-DAY $\land$</td>
<td>UNTEXTURED</td>
</tr>
<tr>
<td>SKY</td>
<td>RGB-IS-AVAILABLE</td>
<td>WHITE</td>
</tr>
<tr>
<td>SKY</td>
<td>SPARSE-RANGE-IS-AVAILABLE</td>
<td>SPARSE-RANGE-IS-UNDEFINED</td>
</tr>
<tr>
<td>SKY</td>
<td>CAMERA-IS-HORIZONTAL</td>
<td>NEAR-TOP</td>
</tr>
<tr>
<td>SKY</td>
<td>CAMERA-IS-HORIZONTAL $\land$ CLIQUE-CONTAINS(complete-sky)</td>
<td>ABOVE-SKYLINE</td>
</tr>
<tr>
<td>SKY</td>
<td>CLIQUE-CONTAINS(sky)</td>
<td>SIMILAR-INTENSITY</td>
</tr>
<tr>
<td>SKY</td>
<td>CLIQUE-CONTAINS(sky)</td>
<td>SIMILAR-TEXTURE</td>
</tr>
<tr>
<td>SKY</td>
<td>RGB-IS-AVAILABLE $\land$ CLIQUE-CONTAINS(sky)</td>
<td>SIMILAR-COLOR</td>
</tr>
<tr>
<td>GROUND</td>
<td>CAMERA-IS-HORIZONTAL</td>
<td>HORIZONTALLY-STRIATED</td>
</tr>
<tr>
<td>GROUND</td>
<td>CAMERA-IS-HORIZONTAL</td>
<td>NEAR-BOTTOM</td>
</tr>
<tr>
<td>GROUND</td>
<td>SPARSE-RANGE-IS-AVAILABLE</td>
<td>SPARSE-RANGES-FORM-HORIZONT/</td>
</tr>
<tr>
<td>GROUND</td>
<td>DENSE-RANGE-IS-AVAILABLE</td>
<td>DENSE-RANGES-FORM-HORIZONTA</td>
</tr>
<tr>
<td>GROUND</td>
<td>CAMERA-IS-HORIZONTAL $\land$ CLIQUE-CONTAINS(complete-ground)</td>
<td>BELOW-SKYLINE</td>
</tr>
<tr>
<td>GROUND</td>
<td>CAMERA-IS-HORIZONTAL $\land$ CLIQUE-CONTAINS(geometric-horizon) $\land$ CLIQUE-CONTAINS(skyline)</td>
<td>BELOW-GEOMETRIC-HORIZON</td>
</tr>
<tr>
<td>GROUND</td>
<td>TIME-IS-DAY</td>
<td>DARK</td>
</tr>
</tbody>
</table>

- Guzman (*SEE*), 1968
- Noton and Stark 1971
- Hansen & Riseman (*VISIONS*), 1978
- Barrow & Tenenbaum 1978
- Brooks (*ACRONYM*), 1979
- Marr, 1982
- Ohta & Kanade, 1978
- Yakimovskiy & Feldman, 1973
An Age of Scene Understanding

- Guzman (SEE), 1968
- Noton and Stark 1971
- Hansen & Riseman (VISIONS), 1978
- Barrow & Tenenbaum 1978
- Brooks (ACRONYM), 1979
- Marr, 1982
- Ohta & Kanade, 1978
- Yakimovsky & Feldman, 1973

[Ohta & Kanade 1978]
\[ p(O \mid I) \propto p(I \mid O) \, p(O) \]

Diagram:
- **Objects**
- **Image**
- **Object model**
- **Context model**
\[ p(O \mid I) \alpha p(I \mid O) \, p(O) \]

- Object model
- Context model
- Full joint
- Scene model
- Approx. joint
\[ p(O \mid I) \propto p(I \mid O) \cdot p(O) \]

Object model

Context model

Full joint

Scene model

Approx. joint
\[ p(\text{O} \mid \text{I}) \propto p(\text{I} \mid \text{O}) \ p(\text{O}) \]

Object model

Context model

Full joint

Scene model

Approx. joint

\[ p(\text{O}) = \sum_s \prod_i p(\text{O}_i \mid S=s) \ p(S=s) \]
p(O | l) \propto p(l|O) \ p(O)

- Object model
- Context model
- Scene model
- Approx. joint
- Full joint
Context models

Independent model

Objects are correlated via the scene

Dependencies among objects
Context models

Independent model

Objects are correlated via the scene

Dependencies among objects
Global precedence

Forest Before Trees: The Precedence of Global Features in Visual Perception
Navon (1977)
Global and local representations

Urban street scene

building

car

sidewalk
Global and local representations

Image index: Summary statistics, configuration of textures

Urban street scene

- building
- car
- sidewalk

Histogram

Urban street scene
An integrated model of Scenes, Objects, and Parts

- Multiclass and pose invariant object detection,
Context-based vision system for place and object recognition

We use 17 annotated sequences for training

- Hidden states = location (63 values)
- Observations = $v^G_t$ (80 dimensions)
- Transition matrix encodes topology of environment
- Observation model is a mixture of Gaussians centered on prototypes (100 views per place)

Torralba, Murphy, Freeman and Rubin. ICCV 2003
Our mobile rig

Torralba, Murphy, Freeman, Rubin. 2003
Place recognition demo

Input image (120x160)

$t=930$, truth = 400-fl6-visionArea1

Shows the category and the identity of The place when the system is confident.
Runs at 4 fps on Matlab.
Identification and categorization of known places

Specific location

Location category

Indoor/outdoor

Thistle corridor
Theresa office
200 side street
Draper street
200 out street
400 Short street
Draper plaza
400 plaza
400 Back street
Jason corridor
elevator 200/7
office 200/936
Vision Area 2
Vision Area 1
kitchen floor 6
elevator 200/6
corridor 6c
corridor 6b
corridor 6a
office 400/628
office 400/627
office 400/625
office 400/611
office 400/610
elevator 400/1
elevator 400/1

kitchen
lobby
open space
corridor
office
plaza
street

outdoor
indoor

$P(Q_t \mid v^G_{t:t})$

$P(C_t \mid v^G_{t:t})$
An integrated model of Scenes, Objects, and Parts

Murphy, Torralba, Freeman; NIPS 2003. Torralba, Murphy, Freeman, CACM 2010.
Application of object detection for image retrieval

Results using the keyboard detector alone
Application of object detection for image retrieval

Results using the keyboard detector alone

Results using both the keyboard detector and the global scene features
Object retrieval: scene features vs. detector

Results using the keyboard detector alone

Results using both the detector and the global scene features

Murphy, Torralba, Freeman; NIPS 2003. Torralba, Murphy, Freeman, CACM 2010.
Localizing the object
An integrated model of Scenes, Objects, and Parts

[Diagram showing relationships between Scene, Scene gist features, N_car, Z_car, and S]
Predicting object location

Training set (cars)

\[
\{g^1, z^1\} \\
\{g^2, z^2\} \\
\{g^3, z^3\} \\
\vdots
\]

\[
Z|g = \sum (A_n g + b_n) w_n(g)
\]
Predicting location

Torralba & Sinha, 2001; Murphy, Torralba, Freeman, 2003; Hoeim, Efros, Hebert. 2006
An integrated model of Scenes, Objects, and Parts

We train a multiview car detector.

\[
p(d \mid F=1) = N(d \mid \mu_1, \sigma_1)
\]
\[
p(d \mid F=0) = N(d \mid \mu_0, \sigma_0)
\]
An integrated model of Scenes, Objects, and Parts
Two tasks
A car out of context ...
A car out of context ...
3d Scene Context

Hoiem, Efros, Hebert ICCV 2005
Figure 6. Stages of the recognition system: (a) initial detections before and (b) after applying ground plane constraints, (c) temporal integration on reconstructed map, (d) estimated 3D car locations, rendered back into the original image.
Context models

Independent model

Objects are correlated via the scene

Dependencies among objects
1) Generate candidate objects (run a detector, or segmentation)

M possible object labels
N regions

Label: \( c_k = [1\ldots M] \) with \( k = [1\ldots N] \)
Scores: \( s_k = \) vector length M

2) For each candidate, get a list of possible interpretations with their probabilities

\[ p(c_k = m \mid s_k) \]

3) Goal: to assign labels \( c_k \) to each candidate so that they are in contextual agreement. We want to optimize the joint probability of all the labels:

\[ p(c_1 = m_1, \ldots, c_N = m_N \mid s_1, \ldots, s_N) \]

**Goal:** to assign labels $c_k$ to each candidate so that they are in contextual agreement.

M possible object labels
N regions

Label: $c_k = [1 \ldots M]$ with $k = [1 \ldots N]$
Scores: $s_k =$ vector length $M$

We want to optimize the joint probability of all the labels:

$$p(c_1 = m_1, \ldots, c_N = m_N \mid s_1, \ldots, s_N)$$

**Solution 1:** Assume objects are independent:

$$p(c_1=m_1,\ldots,c_N=m_N|s_1,\ldots,s_N) = \prod_{i=1}^{N} p(c_i=m_i|s_i)$$

**Problem:** it does not makes use of the correlation between objects in the world. This is fine if the detectors are perfect.
**Goal:** to assign labels $c_k$ to each candidate so that they are in contextual agreement.

$M$ possible object labels  
$N$ regions  

Label: $c_k = [1...M]$ with $k = [1...N]$  
Scores: $s_k =$ vector length $M$

We want to optimize the joint probability of all the labels:

$$p(c_1 = m_1, \ldots, c_N = m_N \mid s_1, \ldots, s_N)$$

**Solution 2:** Assume objects are fully dependent:

$$p(c_1 = m_1, \ldots, c_N = m_N \mid s_1, \ldots, s_N) = \frac{p(s_1, \ldots, s_N \mid c_1 = m_1, \ldots, c_N = m_N) \ p(c_1 = m_1, \ldots, c_N = m_N)}{Z(s_1, \ldots, s_N)}$$

$$= \frac{\prod_{i=1}^{N} p(s_i \mid c_i = m_i) \ p(c_1 = m_1, \ldots, c_N = m_N)}{Z(s_1, \ldots, s_N)}$$

$$Z(s_1, \ldots, s_N) = \sum_{\text{All } [c_1, \ldots, c_N]} \prod_{i=1}^{N} p(s_i \mid c_i = m_i) \ p(c_1 = m_1, \ldots, c_N = m_N)$$

**Problem:** learning $p(c_1 = m_1, \ldots, c_N = m_N)$ will need a lot of data. Recognition can be slow.
Goal: to assign labels $c_k$ to each candidate so that they are in contextual agreement.

$M$ possible object labels
$N$ regions

Label: $c_k = [1...M]$ with $k = [1...N]$
Scores: $s_k =$ vector length $M$

We want to optimize the joint probability of all the labels:

$$p(c_1 = m_1, \ldots, c_N = m_N \mid s_1, \ldots, s_N)$$

Solution 3: Approximated model of dependencies:

$$p(c_1 = m_1, \ldots, c_N = m_N \mid s_1, \ldots, s_N) = \frac{\Pi p(s_i \mid c_i = m_i) p(c_1 = m_1, \ldots, c_N = m_N)}{Z(s_1, \ldots, s_N)}$$

$$p(c_1 = m_1, \ldots, c_N = m_N) = \exp(\Sigma \phi(c_i = m_i, c_j = m_j))$$

$\phi(c_i = m_i, c_j = m_j) =$ co-ocurrence matrix on training set (count how many times two objects appear together).

Problem: learning $p(c_1 = m_1, \ldots, c_N = m_N)$ will be easier, but recognition may still be slow.
\( \Phi(c_i=m_i, c_j=m_j) \) = co-occurrence matrix on training set (count how many times two objects appear together).
Objects in context

Carbonetto, de Freitas & Barnard (2004)


Kumar, Hebert (2005)

Heitz and Koller (2008)

Sudderth, Torralba, Wilsky, Freeman (2005)

Rabinovich et al (2007)

Torralba, Sinha (2001)

Torralba Murphy Freeman (2004)

Hoiem, Efros, Hebert (2005)

Desai, Ramanan, and Fowlkes (2009)
Object-Object Relationships

- Fink & Perona (NIPS 03)

Use output of boosting from other objects at previous iterations as input into boosting for this iteration

**Figure 5:** A-E. Emerging features of eyes, mouths and faces (presented on windows of raw images for legibility). The windows' scale is defined by the detected object size and by the map mode (local or contextual). C. faces are detected using face detection maps $H^{Face}$, exploiting the fact that faces tend to be horizontally aligned.
Pixel labeling using MRFs

Enforce consistency between neighboring labels, and between labels and pixels

\[ P(L, x) = P(L)P(x|L) = \frac{1}{Z} \prod_i \prod_{j \in N_i} \psi_{ij}(L_i, L_j) \left[ \prod_i P(x_i|L_i) \right] \]

Carbonetto, de Freitas & Barnard, ECCV’04
Beyond nearest-neighbor grids

• Most MRF/CRF models assume nearest-neighbor graph topology
• This cannot capture long-distance correlations
Dynamically structured trees

- Each node pick its parents  
  (Storkey & Williams, PAMI’03)

- 2D SCFGs  
  (Pollak, Siskind, Harper & Bouman ICASSP’03)
Object-Object Relationships

Use latent variables to induce long distance correlations between labels in a Conditional Random Field (CRF)

He, Zemel & Carreira-Perpinan (04)
Object-Object Relationships

[Kumar Hebert 2005]
3d Scene Context

Image | Support | Vertical | Sky

V-Left | V-Center | V-Right | V-Porous | V-Solid

Object Surface? | Support?

[Hoiem, Efros, Hebert ICCV 2005]
In this work, there is not labeling for stuff. Instead, they look for clusters of textures and model how each cluster correlates with the target object.
What, where and who? Classifying events by scene and object recognition

event: Rowing

scene: Lake

Athlete

Rowing boat

Water

Tree

L-J Li & L. Fei-Fei, ICCV 2007
Grammars

- Guzman (SEE), 1968
- Noton and Stark 1971
- Hansen & Riseman (VISIONS), 1978
- Barrow & Tenenbaum 1978
- Brooks (ACRONYM), 1979
- Marr, 1982
- Yakimovsky & Feldman, 1973

[Ohta & Kanade 1978]
Grammars for objects and scenes

Who needs context anyway?
We can recognize objects even out of context

Banksy