CS294-43: Recognition in Context

Prof. Trevor Darrell

Spring 2009

April 14th, 2009
Last Lecture – **Kernel Combination, Segmentation, and Structured Output**

- C. Pantofaru, C. Schmid, and M. Hebert, "Object recognition by integrating multiple image segmentations," CVPR 2008,
- Chunhui Gu, Joseph J. Lim, Pablo Arbelaez, Jitendra Malik, Recognition using Regions, CVPR 2009, to appear
Today – Image Context


Today – Image Context


- D. Hoiem, A. A. Efros, and M. Hebert, "Putting objects in perspective," in Computer Vision and Pattern Recognition, 2006 [Robert Carroll]


Why is detection hard?

Plus, we want to do this for ~ 1000 objects

10,000 patches/object/image

1,000,000 images/day
Standard approach to scene analysis

1) Object representation based on intrinsic features:

2) Detection strategy:

3) The scene representation

Slide credit: A. Torralba
Is local information enough?

Slide credit: A. Torralba
With hundreds of categories

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

Slide credit: A. Torralba
Is local information even enough?
Is local information even enough?

Information Contextual features

Local features

Distance

Slide credit: A. Torralba
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.

---

Slide credit: A. Torralba
The multiple personalities of a blob

Slide credit: A. Torralba
The multiple personalities of a blob

Slide credit: A. Torralba
ABC
Look-Alikes by Joan Steiner
Look-Alikes by Joan Steiner
Look-Alikes by Joan Steiner
The context challenge

How far can you go without using an object detector?
What are the hidden objects?

Slide credit: A. Torralba
What are the hidden objects?

Slide credit: A. Torralba
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)

<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 ▵ TIME-IS-DAY</td>
<td>BRIGHT</td>
</tr>
<tr>
<td>SKY</td>
<td>SKY-IS-CLEAR ▵ TIME-IS-DAY</td>
<td>UNTEXTURED</td>
</tr>
<tr>
<td>SKY</td>
<td>SKY-IS-CLEAR ▵ TIME-IS-DAY ▷ RGB-IS-AVAILABLE</td>
<td>BLUE</td>
</tr>
<tr>
<td>SKY</td>
<td>SKY-IS-OVERCAST ▵ TIME-IS-DAY</td>
<td>BRIGHT</td>
</tr>
<tr>
<td>SKY</td>
<td>SKY-IS-OVERCAST ▵ TIME-IS-DAY</td>
<td>UNTEXTURED</td>
</tr>
<tr>
<td>SKY</td>
<td>SKY-IS-OVERCAST ▵ TIME-IS-DAY ▷ 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 ▷ CLIQUE-CONTAINS(complete-sky)</td>
<td>NEAR-TOP</td>
</tr>
<tr>
<td>SKY</td>
<td>CLIQUE-CONTAINS(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>RGB-IS-AVAILABLE ▷ CLIQUE-CONTAINS(sky)</td>
<td>SIMILAR-TEXTURE</td>
</tr>
<tr>
<td>GROUND</td>
<td>CAMERA-IS-HORIZONTAL ▷ CLIQUE-CONTAINS(complete-ground)</td>
<td>SIMILAR-COLOR</td>
</tr>
<tr>
<td>GROUND</td>
<td>SPARSE-RANGE-IS-AVAILABLE</td>
<td>HORIZONTALLY-STRIATED</td>
</tr>
<tr>
<td>GROUND</td>
<td>DENSE-RANGE-IS-AVAILABLE</td>
<td>NEAR-BOTTOM</td>
</tr>
<tr>
<td>GROUND</td>
<td>DENSE-RANGES-FORM-HORIZONTA</td>
<td>SPARSE-RANGES-FORM-HORIZONTA</td>
</tr>
<tr>
<td>GROUND</td>
<td>BELOW-SKYLINE</td>
<td>DENSE-RANGES-FORM-HORIZONTA</td>
</tr>
<tr>
<td>GROUND</td>
<td>CAMERA-IS-HORIZONTAL ▷ CLIQUE-CONTAINS(geometric-horizon) ▷ 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>

Slide credit: A. Torralba
Today – Image Context


• D. Hoiem, A. A. Efros, and M. Hebert, "Putting objects in perspective," in Computer Vision and Pattern Recognition, 2006 [Robert Carroll]


• G. Heitz and D. Koller, "Learning spatial context: Using stuff to find things," in ECCV 2008, pp. 30-43. [Brain Kazian]


• Y. Li and R. Nevatia, "Key object driven multi-category object recognition, localization and tracking using spatio-temporal context," in ECCV 2008
Multiclass object detection and context modeling

Antonio Torralba

In collaboration with
Kevin P. Murphy and William T. Freeman
Object representations

Inside the object
(intrinsic features)

Object size

Global
Parts
Pixels

Agarwal & Roth, (02), Moghaddam, Pentland (97), Turk, Pentland (91), Vidal-Naquet, Ullman, (03)
Heisele, et al, (01), Agarwal & Roth, (02), Kremp, Geman, Amit (02), Dorko, Schmid, (03)
Fergus, Perona, Zisserman (03), Fei Fei, Fergus, Perona, (03), Schneiderman, Kanade (00), Lowe (99)
Etc.
1) Search space is HUGE

“Like finding needles in a haystack”

For each object:
- Need to search over locations and scales
- Error prone (classifier must have very low false positive rate)
- Slow (many patches to examine)

10,000 patches/object/image

1,000,000 images/day
2) Local features are not even sufficient.
Symptoms of local features only

Some false alarms occur in image regions in which it is impossible for the target to be present given the context.
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

Human vision: Biederman, Bar & Ullman, Palmer, …
What is context

• Scenes

• Other objects

• Properties of objects and scenes (pose, style, etc.)
Why is context important?

- Changes the interpretation of an object (or its function)

- Context defines what an unexpected event is
Why is context important?

• Reduces the search space

• Context features can be shared among many objects across locations and scales: more efficient than local features.
Object representations

Outside the object (contextual features)  |  Inside the object (intrinsic features)

Global context  |  Local context  |  Global appearance  |  Parts  |  Pixels

Krupa & Shiele, (03), Fink & Perona (03)
Carbonetto, Freitas, Barnard (03), Kumar, Hebert, (03)
He, Zemel, Carreira-Perpinan (04), Moore, Essa, Monson, Hayes (99)
Strat & Fischler (91), Murphy, Torralba & Freeman (03)

Agarwal & Roth, (02), Moghaddam, Pentland (97), Turk, Pentland (91), Vidal-Naquet, Ullman, (03)
Heisele, et al, (01), Agarwal & Roth, (02), Kremp, Geman, Amit (02), Dorko, Schmid, (03)
Fergus, Perona, Zisserman (03), Fei Fei, Fergus, Perona, (03), Schneiderman, Kanade (00), Lowe (99)
Etc.
Previous work on context

- Strat & Fischler (91)

Context defined using hand-written rules about relationships between objects

<table>
<thead>
<tr>
<th>#</th>
<th>Class</th>
<th>Context elements</th>
<th>Operator</th>
</tr>
</thead>
<tbody>
<tr>
<td>41</td>
<td>SKY</td>
<td>ALWAYS</td>
<td>ABOVE-HORIZON</td>
</tr>
<tr>
<td>42</td>
<td>SKY</td>
<td>SKY-IS-CLEAR ∧ TIME-IS-DAY</td>
<td>BRIGHT</td>
</tr>
<tr>
<td>43</td>
<td>SKY</td>
<td>SKY-IS-CLEAR ∧ TIME-IS-DAY</td>
<td>UNTEXTURED</td>
</tr>
<tr>
<td>44</td>
<td>SKY</td>
<td>SKY-IS-CLEAR ∧ TIME-IS-DAY ∧ RGB-IS-AVAILABLE</td>
<td>BLUE</td>
</tr>
<tr>
<td>45</td>
<td>SKY</td>
<td>SKY-IS-OVERCAST ∧ TIME-IS-DAY</td>
<td>BRIGHT</td>
</tr>
<tr>
<td>46</td>
<td>SKY</td>
<td>SKY-IS-OVERCAST ∧ TIME-IS-DAY ∧ RGB-IS-AVAILABLE</td>
<td>UNTEXTURED</td>
</tr>
<tr>
<td>47</td>
<td>SKY</td>
<td>SKY-IS-OVERCAST ∧ TIME-IS-DAY ∧ RGB-IS-AVAILABLE</td>
<td>WHITE</td>
</tr>
<tr>
<td>48</td>
<td>SKY</td>
<td>SPARSE-RANGE-IS-AVAILABLE</td>
<td>SPARSE-RANGE-IS-UNDEFINED</td>
</tr>
<tr>
<td>49</td>
<td>SKY</td>
<td>CAMERA-IS-HORIZONTAL</td>
<td>NEAR-TOP</td>
</tr>
<tr>
<td>50</td>
<td>SKY</td>
<td>CAMERA-IS-HORIZONTAL ∧</td>
<td>ABOVE-SKYLINE</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CLIQUE-CONTAINS(complete-sky)</td>
<td></td>
</tr>
<tr>
<td>51</td>
<td>SKY</td>
<td>CLIQUE-CONTAINS(sky)</td>
<td>SIMILAR-INTENSITY</td>
</tr>
<tr>
<td>52</td>
<td>SKY</td>
<td>CLIQUE-CONTAINS(sky)</td>
<td>SIMILAR-TEXTURE</td>
</tr>
<tr>
<td>53</td>
<td>SKY</td>
<td>RGB-IS-AVAILABLE ∧ CLIQUE-CONTAINS(sky)</td>
<td>SIMILAR-COLOR</td>
</tr>
<tr>
<td>54</td>
<td>GROUND</td>
<td>CAMERA-IS-HORIZONTAL</td>
<td>HORIZONTALLY-STRIATED</td>
</tr>
<tr>
<td>55</td>
<td>GROUND</td>
<td>CAMERA-IS-HORIZONTAL</td>
<td>NEAR-BOTTOM</td>
</tr>
<tr>
<td>56</td>
<td>GROUND</td>
<td>SPARSE-RANGE-IS-AVAILABLE</td>
<td>SPARSE-RANGES-FORM-HORIZONTAL-SURFACE</td>
</tr>
<tr>
<td>57</td>
<td>GROUND</td>
<td>DENSE-RANGE-IS-AVAILABLE</td>
<td>DENSE-RANGES-FORM-HORIZONTAL-SURFACE</td>
</tr>
<tr>
<td>58</td>
<td>GROUND</td>
<td>CAMERA-IS-HORIZONTAL ∧</td>
<td>BELOW-SKYLINE</td>
</tr>
<tr>
<td>59</td>
<td>GROUND</td>
<td>CLIQUE-CONTAINS(complete-ground)</td>
<td>BELOW-GEOMETRIC-HORIZON</td>
</tr>
<tr>
<td>60</td>
<td>GROUND</td>
<td>CAMERA-IS-HORIZONTAL ∧</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>CLIQUE-CONTAINS(geometric-horizon) ∧</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>CLIQUE-CONTAINS(skyline)</td>
<td></td>
</tr>
<tr>
<td>61</td>
<td>GROUND</td>
<td>TIME-IS-DAY</td>
<td>DARK</td>
</tr>
<tr>
<td>62</td>
<td>FOLIAGE</td>
<td>ALWAYS</td>
<td>HIGHLY-TEXTURED</td>
</tr>
<tr>
<td>63</td>
<td>FOLIAGE</td>
<td>ALWAYS</td>
<td>HIGH-VEGETATIVE-TRANSPARENCY</td>
</tr>
<tr>
<td>64</td>
<td>FOLIAGE</td>
<td>CAMERA-IS-HORIZONTAL</td>
<td>NEAR-TOP</td>
</tr>
<tr>
<td>65</td>
<td>FOLIAGE</td>
<td>RGB-IS-AVAILABLE</td>
<td>GREEN</td>
</tr>
<tr>
<td>66</td>
<td>RAISED-OBJECT</td>
<td>SPARSE-RANGE-IS-AVAILABLE</td>
<td>SPARSE-HEIGHT-ABOVE-GROUND</td>
</tr>
<tr>
<td>67</td>
<td>RAISED-OBJECT</td>
<td>DENSE-RANGE-IS-AVAILABLE</td>
<td>DENSE-HEIGHT-ABOVE-GROUND</td>
</tr>
<tr>
<td>68</td>
<td>RAISED-OBJECT</td>
<td>CAMERA-IS-HORIZONTAL ∧</td>
<td>ABOVE-SKYLINE</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CLIQUE-CONTAINS(complete-sky)</td>
<td></td>
</tr>
</tbody>
</table>

Table 5: Type II Context Sets: Candidate Evaluation
Previous work on context

• Fink & Perona (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 $H_{\text{Face}}$, exploiting the fact that faces tend to be horizontally aligned.](image)
Previous work on context

- Murphy, Torralba & Freeman (03)

Use global context to predict objects but there is no modeling of spatial relationships between objects.
Previous work on context

- Carbonetto, de Freitas & Barnard (04)
- Enforce spatial consistency between labels using MRF
Graphical models for image labeling

Nearest neighbor grid

Densely connected graphs with low informative connections

Want to model long-range correlations between labels
Previous work on context

- He, Zemel & Carreira-Perpinan (04)
  Use latent variables to induce long distance correlations between labels in a Conditional Random Field (CRF)
Outline of this talk

• Use global image features (as well as local features) in boosting to help object detection

• Learn structure of dense CRF (with long range connections) using boosting, to exploit spatial correlations
Image database

- ~2500 hand labeled images with segmentations
- ~30 objects and stuff
- Indoor and outdoor
- Sets of images are separated by locations and camera (digital/webcam)
- No graduate students or low-income-student-class exploited for labeling.
Which objects are important?

Average percentage of pixels occupied by each object.
Object representation

• **Discrete/bounded/rigid**

  Screen, car, pedestrian, bottle, …

• **Extended/unbounded/deformable**

  Building, sky, road, shelves, desk, …

We will use region labeling as a representation.
Learning local features
(intrinsic object features)

We maximize the probability of the true labels using Boosting.
Object local features

(Borenstein & Ullman, ECCV 02)

Convolve with oriented filter

Normalized correlation with an object patch

Threshold

Convolve with segmentation fragment

Patches from 5x5 to 30x30 pixels.
Results with local features
Results with local features

Screen
Results with local features

Car
Global context: location priming
How far can we go without object detectors?

Context features that represent the scene instead of other objects.
The global features can provide:
- Object presence
- Location priming
- Scale priming
Object global features

First we create a dictionary of scene features and object locations:

Only the vertical position of the object is well constrained by the global features.
Object global features

How to compute the global features

Downsample (8x8, 16x16, 32x32)

Downsample (10x10)
Car detection with global features

Features selected by boosting:

![Car features selected by boosting](image)
Combining global and local

Global and local
Only local

ROC for same total number of features (100 boosting rounds):
Clustering of objects with local and global feature sharing

Objects are similar if they share local features and they appear in the same contexts.
Outline of this talk

• Use global image features (as well as local features) in boosting to help object detection

• Learn structure of dense CRF (with long range connections) using boosting, to exploit spatial correlations
Adding correlations between objects

We need to learn

• The structure of the graph
• The pairwise potentials
Learning in CRFs

• Parameters
  – Lafferty, McCallum, Pereira (ICML 2001)
    • Find global optimum using gradient methods plus exact inference (forwards-backwards) in a chain
  – Kumar & Herbert, NIPS 2003
    • Use pseudo-likelihood in 2D CRF
  – Carbonetto, de Freitas & Barnard (04)
    • Use approximate inference (loopy BP) and pseudo-likelihood on 2D MRF

• Structure
  – He, Zemel & Carreira-Perpinan (CVPR 04)
    • Use contrastive divergence
  – Torralba, Murphy, Freeman (NIPS 04)
    • Use boosting
Sequentially learning the structure
Sequentially learning the structure

At each iteration of boosting

• We pick a weak learner applied to the image (local or global features)

• We pick a weak learner applied to a subset of the label-beliefs at the previous iteration. These subsets are chosen from a dictionary of labeled graph fragments from the training set.
Car detection
Car detection

From intrinsic features

A car out of context is less of a car

From contextual features
Screen/keyboard/mouse
Cascade

Viola & Jones (2001)

Set to zero the beliefs of nodes with low probability of containing the target.

Perform message passing only on undecided nodes.

The detection of the screen reduces the search space for the mouse detector.
Cascade

[Diagram showing the process of cascade detection with images and labels at different time steps (t=1, t=2, t=4, t=20, t=40).]
Cascade

[Graphs showing size of search space and detection rate with varying rounds for Screen, Keyboard, Mouse, and b(car) at different time points: t=1, t=2, t=4, t=20, t=40]
Future work

- Learn relationships between more objects (things get interesting beyond the 10 objects bar)
- Integrate segmentation and multiscale detection
- Add scenes/places
Today – Image Context


• D. Hoiem, A. A. Efros, and M. Hebert, "Putting objects in perspective," in Computer Vision and Pattern Recognition, 2006 [Robert Carroll]


• G. Heitz and D. Koller, "Learning spatial context: Using stuff to find things," in ECCV 2008, pp. 30-43. [Brain Kazian]


• Y. Li and R. Nevatia, "Key object driven multi-category object recognition, localization and tracking using spatio-temporal context," in ECCV 2008
Today – Image Context


• D. Hoiem, A. A. Efros, and M. Hebert, "Putting objects in perspective," in Computer Vision and Pattern Recognition, 2006 [Robert Carroll]


• G. Heitz and D. Koller, "Learning spatial context: Using stuff to find things," in ECCV 2008, pp. 30-43. [Brain Kazian]


• Y. Li and R. Nevatia, "Key object driven multi-category object recognition, localization and tracking using spatio-temporal context," in ECCV 2008
A picture tells us many stories

Fei-Fei et al. JoV 2007
A picture tells us many stories
This was a picture with some dark sploches in it. Yeah. . .that's about it. (Subject: KM)

I think I saw two people on a field. (Subject: RW)

Outdoor scene. There were some kind of animals, maybe dogs or horses, in the middle of the picture. It looked like they were running in the middle of a grassy field. (Subject: IV)

two people, whose profile was toward me. looked like they were on a field of some sort and engaged in some sort of sport (their attire suggested soccer, but it looked like there was too much contact for that). (Subject: AI)

Fei-Fei et al. JoV 2007
What, where and who? Classifying events by scene and object recognition

event: Rowing

scene: Lake

L-J Li & L. Fei-Fei, ICCV 2007
Related vision works toward holistic interpretation of images

- Murphy, Torralba & Freeman (‘Seeing Forest and Trees’, 2003)
- Jin & Geman (‘Hierarchical Image Model’, 2006)
- Hoiem, Efros & Herbert (‘Objects in Perspective’, 2006)
- …
event

scene pathway

object pathway

\[ L.-J. \text{ Li} & L. \text{ Fei-Fei ICCV 2007} \]
Compute SIFT descriptor

[Lowe’99]

Codewords representation

L.-J. Li & L. Fei-Fei ICCV 2007
G:
Geometry features

Hoiem et al. Siggraph 2006
O = 'horse'

object pathway

L.-J. Li & L. Fei-Fei ICCV 2007
Example: Polo

L.-J. Li & L. Fei-Fei ICCV 2007
Example: Sailing

L.-J. Li & L. Fei-Fei ICCV 2007
scene pathway

"Polo Field"

L.-J. Li & L. Fei-Fei ICCV 2007
Recognition in an Unknown Image

L.-J. Li & L. Fei-Fei ICCV 2007
Labeling objects in an unknown image
-- objects (who)

\[ p(A_n, G_n | O_k) = \sum_z P(A_n, G_n | z) P(z | O_k) \]

L.-J. Li & L. Fei-Fei ICCV 2007
Labeling objects in an unknown image

-- scene (where)

\[
p(I | S, \rho, \theta) = \int p(\omega | \rho, S) \left( \prod_{m=1}^{M} \sum_{t_m} p(t_m | \omega) : p(X_m | t_m, \theta) \right) d\omega
\]

L.-J. Li & L. Fei-Fei ICCV 2007
Labeling objects in an unknown image -- events (what)

\[ p(I|E) \propto P(I|S)P(S|E) \prod_{n=1}^{N} \sum_{Q} P(A_n, G_n|Q)P(Q|E) \]

L.-J. Li & L. Fei-Fei ICCV 2007
The 3W stories

what

event: Polo

who

Sky
Tree
Horse
Grass

where

scene: Polo Field

event: Bocce

scene: Bocce court

L.-J. Li & L. Fei-Fei ICCV 2007
event: Rowing

Tree
Athlete
Rowing boat
Water

scene: Lake

event: Sailing

Sky
Sailing boat
Water

scene: Lake

event: Snowboarding

Sky
Athlete
Snowfield

scene: Snow mountain
Quantitative result

Average Perf. = 73.4%

L.-J. Li & L. Fei-Fei ICCV 2007
Today – Image Context


• D. Hoiem, A. A. Efros, and M. Hebert, "Putting objects in perspective," in Computer Vision and Pattern Recognition, 2006 [Robert Carroll]


• G. Heitz and D. Koller, "Learning spatial context: Using stuff to find things," in ECCV 2008, pp. 30-43. [Brain Kazian]


• Y. Li and R. Nevatia, "Key object driven multi-category object recognition, localization and tracking using spatio-temporal context," in ECCV 2008
Today – Image Context


• D. Hoiem, A. A. Efros, and M. Hebert, "Putting objects in perspective," in Computer Vision and Pattern Recognition, 2006 [Robert Carroll]


• G. Heitz and D. Koller, "Learning spatial context: Using stuff to find things," in ECCV 2008, pp. 30-43. [Brain Kazian]


• Y. Li and R. Nevatia, "Key object driven multi-category object recognition, localization and tracking using spatio-temporal context," in ECCV 2008
Peripheral-Foveal Vision for Real-time Object Recognition

Stephen Gould, Benjamin Sapp, Morgan Quigley, Andrew Y. Ng
Overview

- Human object recognition in a 3d environment is far superior to that of any robotic vision system.

- One reason (out of many) for this is that humans use a **fovea** to fixate on, or near an object, thus obtaining a very high resolution image of the object and rendering it easy to recognize.

- We present a novel method for identifying and tracking objects using a two camera system.

- **Our method** is motivated by biological vision systems:
  - uses a learned "**attentive** interest map" on a low resolution view to direct a high resolution "**fovea**."
  - objects that are recognized in the **fovea** can then be tracked using **peripheral vision**.

http://ai.stanford.edu/~sgould/vision/nips06poster.pdf
STAIR robot

http://ai.stanford.edu/~sgould/vision/nips06poster.pdf
Visual attention model

- Our system uses two separate cameras:
  - fixed wide-angle camera for peripheral vision
  - controllable pan-tilt-zoom (PTZ) camera for foveal vision.
- The PTZ camera can focus on any region of the scene to obtain a high-resolution image for object recognition.
- Previously identified objects are tracked using peripheral vision.
• The attention system periodically decides between the following actions:

  • **Confirmation** of a tracked object by fixating the fovea over the predicted location of the object;

  • **Search** for unidentified objects by moving the fovea to some new part of the scene.

• Estimating the reduction in entropy, $H$, by taking each action (**Confirmation** or **Search**), we take the action which maximizes the expected reduction in entropy.
Interest modeling

• Our interest model allows us to choose which foveal region to examine next by rapidly identifying pixels which have a high probability of containing objects that we can classify.

• We define a pixel to be interesting if it is part of an unknown, yet classifiable object.

  • A consequence of this definition is that our model automatically encodes the biological phenomena of saliency and inhibition of return [Itti and Koch, 2001].

  • Interestingness of every pixel in the peripheral view is modeled using a dynamic Bayesian network (DBN) whose parameters are learned from training videos.
Object recognition and tracking

- A Kalman filter tracks the location and velocity of identified objects in the 2d image plane.

- Use subset of biologically inspired C1 features [Serre et. al, 2004] and learn a boosted decision tree classifier for each object.
Experimental results

- Our method compared to three naive approaches:
  (i) fixing the foveal gaze to the center of view,
  (ii) linearly scanning over the scene from top-left to bottom-right, and,
  (iii) randomly moving the fovea around the scene.

<table>
<thead>
<tr>
<th>Fovea control</th>
<th>Recall</th>
<th>Precision</th>
<th>$F_1$-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed at center</td>
<td>9.49%</td>
<td>97.4%</td>
<td>17.3%</td>
</tr>
<tr>
<td>Linear scanning</td>
<td>13.6%</td>
<td>100.0%</td>
<td>24.0%</td>
</tr>
<tr>
<td>Random scanning</td>
<td>27.7%</td>
<td>84.1%</td>
<td>41.6%</td>
</tr>
<tr>
<td><strong>Our method</strong></td>
<td>62.2%</td>
<td>83.9%</td>
<td><strong>71.5%</strong></td>
</tr>
</tbody>
</table>

- Videos demonstrating our results are available at http://ai.stanford.edu/~sgould/vision/
Today – Image Context


- D. Hoiem, A. A. Efros, and M. Hebert, "Putting objects in perspective," in Computer Vision and Pattern Recognition, 2006 [Robert Carroll]


Key Object Driven Multi-Category Object Recognition, Localization and Tracking Using Spatio-Temporal Context

Yuan Li and Ram Nevatia
Institute for Robotic and Intelligent Systems
University of Southern California
The Idea

- **Objective**: recognizing and tracking multiple categories of objects that are involved in interaction with humans in video.

- **Motivation**: the significance of recognizing objects involved in interaction with humans lies not only in the static image domain but also in video understanding and analysis.

- **Difficulty**: varying appearance of objects, lack of image detail, multi-object co-occurrence, motion make it difficult for purely appearance-based approach.

- **Approach**: incorporate spatial and temporal contextual information to aid object recognition and localization.
Approach Overview

Image feature extraction and key object detection

Inference with spatial-temporal context

Spatial relationship

Temporal relationship

Result

- Human (head-shoulder)
- Table
- Whiteboard
- Computer
- Projector
- Paper
Approach Overview

- **Spatial relationships** between different object categories are utilized so that co-inference enhances accuracy;
- **Temporal context** is utilized to accumulate object evidence and to track objects continuously;
- **Robust key object detection**: boosting + edgelet features; use key objects to reduce inference space for other objects.
- **Observation model for other objects**: interest point detection + SIFT + image region features.
- **Modeled by a dynamic MRF**:
  - Node: object state and observation in one frame;
  - Intra-frame edge: spatial relationship;
  - Inter-frame edge: temporal relationship;
  - Inference done by nonparametric belief propagation.
Fig. 2. the MRF defined in our problem (left) and an ideal graph structure for one input frame (right). Section 5 explains how to build such a graph.
Fig. 4. Use of augmenting nodes to update graph structure. Augmenting nodes for each category are shown as one (dotted circle). For weighted samples, red indicates the highest possible weight, while blue indicates the lowest.
Fig. 3. An example of finding paper based on appearance. (a) Input image; (b) SIFT features (green: feature with positive weight in the classifier, red: feature with negative weight); (c) Segmentation; (d) observation likelihood $p(paper|r_i)$ for each region $r_i$ (yellow: high likelihood).
Experiment setting

• CHIL meeting video corpus
  – Test on 16 videos from 3 sites (IBM, AIT and UPC), 3 camera views for each site, 400 frames / seq.
Experiment setting

• Compare three methods with different levels of context:
  – No context, *i.e.* object observation model is directly applied to each frame;
  – Spatial context only, *i.e.* a MRF without the temporal edges is applied in a frame-by-frame manner;
  – Spatio-temporal context, *i.e.* the full model with both spatial and temporal edges applied to the sequence.

(a) Observation likelihood of different categories without context

(b) Key object detection

(c) Frame-based inference (spatial-only)

(d) Inference using the complete model (spatio-temporal)
Results

- Detection and segmentation

- Recognition confusion matrix

![Results](image)
Results

- Human (head-shoulder)
- Table
- Whiteboard
- Computer
- Projector
- Paper
- Cup
- Chair

(a) IBM Site
(b) IBM Site
(c) IBM Site
(d) Zoomed view
(e) AIT Site
(f) AIT Site
(g) AIT Site
(h) Zoomed view
Results
Today – **Image Context**

- D. Hoiem, A. A. Efros, and M. Hebert, "Putting objects in perspective," in Computer Vision and Pattern Recognition, 2006
Who needs context anyway?
We can recognize objects even out of context

BARELY LEGAL

Banksy
Slide credit: A. Torralba
Next Class – **Shared Structures (Features, Parts)**


