Understanding Visual Scenes

Antonio Torralba

Computer Science and Artificial Intelligence Laboratory (CSAIL)
Department of Electrical Engineering and Computer Science
A VIEW OF A PARK ON A NICE SPRING DAY
Do not feed the ducks sign

PEOPLE WALKING IN THE PARK

DUCKS LOOKING FOR FOOD

PERSON FEEDING DUCKS IN THE PARK
PEOPLE UNDER THE SHADOW OF THE TREES

DUCKS ON TOP OF THE GRASS
Why do we care about recognition?

Perception of function: We can perceive the 3D shape, texture, material properties, without knowing about objects.

But, the concept of category encapsulates also information about what can we do with those objects.

“We therefore include the perception of function as a proper –indeed, crucial- subject for vision science”, from Vision Science, chapter 9, Palmer.
The perception of function

• Direct perception (affordances): Gibson

Flat surface
Horizontal
Knee-high
...

Sittable
upon

• Mediated perception (Categorization)

Flat surface
Horizontal
Knee-high
...

Chair

Sittable
upon

Chair
Chair
Chair?
Direct perception

Some aspects of an object function can be perceived directly

- Functional form: Some forms clearly indicate to a function ("sittable-upon", container, cutting device, …)

It does not seem easy to sit-upon this…
Scenes, as objects, also have affordances.
The function of the scene
Direct perception

Some aspects of an object function can be perceived directly

• Observer relativity: Function is observer dependent

From http://lastchancerescueflint.org
Limitations of Direct Perception

Objects of similar structure might have very different functions.

Not all functions seem to be available from direct visual information only.
Limitations of Direct Perception
Object detection and recognition

Short overview of current approaches
Object recognition
Is it really so hard?

This is a chair

Find the chair in this image

Output of normalized correlation
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
So, let’s make the problem simpler: Block world

Fig. 1. A system for recognizing 3-d polyhedral scenes. a) L.G. Roberts. b) A blocks world scene. c) Detected edges using a 2x2 gradient operator. d) A 3-d polyhedral description of the scene, formed automatically from the single image. e) The 3-d scene displayed with a viewpoint different from the original image to demonstrate its accuracy and completeness. (b) - (e) are taken from [64] with permission MIT Press.)
Binford and generalized cylinders

Recognition by components

Fig. 3. The representation of objects by assemblies of generalized cylinders. a) Thomas Binford. b) A range image of a doll. c) The resulting set of generalized cylinders. (b) and (c) are taken from Agin [1] with permission.


Introduced in computer vision by A. Pentland, 1986.

Irving Biederman

Families of recognition algorithms

Bag of words models
Csurka, Dance, Fan, Willamowski, and Bray 2004
Sivic, Russell, Freeman, Zisserman, ICCV 2005

Voting models
Viola and Jones, ICCV 2001
Heisele, Poggio, et. al., NIPS 01
Schneiderman, Kanade 2004
Vidal-Naquet, Ullman 2003

Shape matching
Deformable models
Berg, Berg, Malik, 2005
Cootes, Edwards, Taylor, 2001

Constellation models
Fischler and Eischlager, 1973
Burl, Leung, and Perona, 1995
Weber, Welling, and Perona, 2000
Fergus, Perona, & Zisserman, CVPR 2003

Rigid template models
Sirovich and Kirby 1987
Turk, Pentland, 1991
Dalal & Triggs, 2006
The face age

Feret dataset, 1996 DARPA

- Graded Learning for Object Detection - Fleuret, Geman (1999)
- Robust Real-time Object Detection - Viola, Jones (2001)
- …
Face detection

[Face priority AE] When a bright part of the face is too bright
Haar-like filters and cascades

Viola and Jones, ICCV 2001

The average intensity in the block is computed with four sums independently of the block size.

Also Fleuret and Geman, 2001
Generic objects:
Edge based descriptors

Gavrila, Philomin, ICCV 1999

Papageorgiou & Poggio (2000)

J. Shotton, A. Blake, R. Cipolla. PAMI 2008.
Histograms of oriented gradients

SIFT, D. Lowe, ICCV 1999

Shape context
Belongie, Malik, Puzicha, NIPS 2000

Count the number of points inside each bin, e.g.:

Count = 4
:
Count = 10

Compact representation of distribution of points relative to each point
Histograms of oriented gradients

Dalal & Trigs, 2006

Orientation Voting

Overlapping Blocks

Local Normalization

Input Image  Gradient Image  Output

Not a person

person
Adding parts

Felzenszwalb, McAllester, Ramanan. 2008.
Adding parts

Felzenszwalb, McAllester, Ramanan. 2008.
Evaluation of performance

Before plotting and ROC or precision-recall curves…

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 detector trying to detect?

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 detector trying to detect?

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?
1. chair, 2. table, 3. road, 4. road, 5. table, 6. car, 7. keyboard.
Some symptoms of standard approaches
Scenes rule over objects

3D percept is driven by the scene, which imposes its ruling to the objects
Scene recognition
The gist of the scene
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.
Demo: Rapid image understanding
By Aude Oliva

Instructions: 9 photographs will be shown for half a second each. Your task is to memorize these pictures.
Memory Test

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?
NO
Have you seen this picture?
NO
Have you seen this picture?
NO
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
From objects to scenes

SceneType 2 {street, office, …}

Object localization

Local features

Image

Riesenhuber & Poggio (99); Vidal-Naquet & Ullman (03); Serre & Poggio, (05); 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)
What makes scenes different?

Different objects, different spatial layout
What makes scenes different?

Different objects, similar spatial layout
What makes scenes different?

Similar objects, different spatial layout
What makes scenes different?

Similar objects, different spatial layout
What makes scenes different? 

Similar objects, and similar spatial layout

Different lighting, different materials, very specific object categories
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
- Blobs
- Textures
Ensemble statistics

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

Set | Test

Conclusion: observers had more accurate representation of the mean than of the individual members of the set.
From scenes to objects

SceneType 2 {street, office, …}

Object localization

Local features

Image

Object localization

Scene emergent

Ensemble statistics

Global features
How far can we go without objects?

SceneType 2 \{street, office, \ldots\}

Local features

Image

- Scene emergent
- Ensemble statistics
- Global features
Global image descriptors
Global image descriptors

Bag of words

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

Non localized textons

Walker, Malik. Vision Research 2004

Spatially organized textures

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

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

Textons

Filter bank  \rightarrow  K-means (100 clusters)

Malik, Belongie, Shi, Leung, 1999

Walker, Malik, 2004

\( \chi^2 = 5.87 \)

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

label = bedroom

label = beach
Bag of words & spatial pyramid matching


S. Lazebnik, et al, CVPR 2006
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]
- phow [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]
Large Scale Scene Recognition

~1,000 categories

>130,000 images

>12,000 fully annotated images

Xiao, Hays, Ehinger, Oliva, Torralba; maybe 2010
Performance with 400 categories

Xiao, Hays, Ehinger, Oliva, Torralba; maybe 2010
Xiao, Hays, Ehinger, Oliva, Torralba; maybe 2010
<table>
<thead>
<tr>
<th></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="Correct classifications" /></td>
</tr>
<tr>
<td>Airplane cabin</td>
<td><img src="image3" alt="Airplane cabin images" /></td>
<td><img src="image4" alt="Correct classifications" /></td>
</tr>
<tr>
<td>Airport terminal</td>
<td><img src="image5" alt="Airport terminal images" /></td>
<td><img src="image6" alt="Correct classifications" /></td>
</tr>
<tr>
<td>Alley</td>
<td><img src="image7" alt="Alley images" /></td>
<td><img src="image8" alt="Correct classifications" /></td>
</tr>
<tr>
<td>Amphitheater</td>
<td><img src="image9" alt="Amphitheater images" /></td>
<td><img src="image10" alt="Correct classifications" /></td>
</tr>
</tbody>
</table>

Xiao, Hays, Ehinger, Oliva, Torralba; maybe 2010
<table>
<thead>
<tr>
<th>Location</th>
<th>Training images</th>
<th>Correct classifications</th>
<th>Miss-classifications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abbey</td>
<td>![Abbey Images]</td>
<td>![Correct Abbey]</td>
<td>![Miss Abbey]</td>
</tr>
<tr>
<td>Airplane cabin</td>
<td>![Airplane Images]</td>
<td>![Correct Airplane]</td>
<td>![Miss Airplane]</td>
</tr>
<tr>
<td>Airport terminal</td>
<td>![Airport Images]</td>
<td>![Correct Airport]</td>
<td>![Miss Airport]</td>
</tr>
<tr>
<td>Alley</td>
<td>![Alley Images]</td>
<td>![Correct Alley]</td>
<td>![Miss Alley]</td>
</tr>
<tr>
<td>Amphitheater</td>
<td>![Amphitheater Images]</td>
<td>![Correct Amphitheater]</td>
<td>![Miss Amphitheater]</td>
</tr>
</tbody>
</table>

Xiao, Hays, Ehinger, Oliva, Torralba; maybe 2010
Example of three different scenes

RIVER

BEACH

VILLAGE

Xiao, Hays, Ehinger, Oliva, Torralba; maybe 2010
But they are all part of the same picture.
But they are all part of the same picture
Scene detection

beach detections

harbor detections

village detections

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
Spatial envelope: a continuous space of scenes

Oliva & Torralba, 2001
Spatial envelope: a continuous space of scenes

Oliva & Torralba, 2001
Context for object recognition
Who needs context anyway?
We can recognize objects even out of context

Banksy
Even in high resolution, we can not shut down contextual processing and it is hard to recognize the true identities of the elements that compose this scene.
Why is context important?

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

• Context defines what an unexpected event is
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.
Global precedence

Forest Before Trees: The Precedence of Global Features in Visual Perception
Navon (1977)
Scene recognition without object recognition

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

Murphy, Torralba, Freeman; NIPS 2003. Torralba, Murphy, Freeman, CACM 2010.
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.
The layered structure of scenes

Assuming a human observer standing on the ground

In a display with multiple targets present, the location of one target constraints the ‘y’ coordinate of the remaining targets, but not the ‘x’ coordinate.

Torralba, Oliva, Castelhano, Henderson. 2006
Context driven object detection

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

We train a multiview car detector.

\[
p(d \mid F=1) = \mathcal{N}(d \mid \mu_1, \sigma_1) \\
p(d \mid F=0) = \mathcal{N}(d \mid \mu_0, \sigma_0)
\]

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

Murphy, Torralba, Freeman; NIPS 2003. Torralba, Murphy, Freeman, CACM 2010.
Two tasks