Scenes and objects

Ivan Laptev and Josef Sivic
http://www.di.ens.fr/~josef
INRIA, WILLOW, ENS/INRIA/CNRS UMR 8548
Laboratoire d’Informatique, Ecole Normale Supérieure, Paris

With slides from: A. Torralba, L. Fei Fei, D. Hoiem and R. Fergus
Announcements

• Final project presentations next week!
  http://www.di.ens.fr/willow/teaching/recvis10/final_project/
  – Send us the **project title** and **names** of people in the group asap!
  – Schedule of the presentations will be emailed this week.

• **Final project report deadline extended to January 5**th.

• If you have any suggestions or comments on the course, please fill-in the feed-back form.
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 3 hours, the talk you give will be better than if you had only worked on it for 2 hours. If you work on it for 5 hours, it will be better still. 7 hours, 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.
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
Sources on writing technical papers

• Notes on technical writing, Don Knuth, 1989.
Today: Scenes and objects

1. Scenes as textures (without modeling objects and their relations)

2. Detecting single objects in context; geometric context.

3. Recognizing multiple objects in an image.

4. Recognizing unseen objects.
What is a scene?
A VIEW OF A PARK ON A NICE SPRING DAY
Do not feed the ducks sign
PEOPLE UNDER THE SHADOW OF THE TREES

DUCKS ON TOP OF THE GRASS
Scene views vs. objects

“By scene we mean a place in which a human can act within, or a place to which a human being could navigate. Scenes are a lot more than just a combination of objects (just as objects are more than the combinations of their parts). Like objects, scenes are associated with specific functions and behaviors, such as eating in a restaurant, drinking in a pub, reading in a library, and sleeping in a bedroom.” – A. Torralba
Scene views vs. objects

A photograph of a fire hydrant

A photograph of a street
Part I: Scenes as textures

(No explicit modeling of objects and their relations)
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
Global scene representations

Bag of words

SpaKally organized textures

Non localized textons

Spatial structure is important in order to provide context for object localization

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
Bag of words for scenes

Bag of words model

Spatially organized textures
Scene categorization

Can we use this representation to categorize scenes?
The 15-scenes benchmark

Oliva & Torralba, 2001
Fei Fei & Perona, 2005
Lazebnik, et al 2006
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: \( k(x,x_m) = \sum_i (\min(x(i), x_m(i))) \)
Scene recognition

100 training samples per class
SVM classifier in all cases

Pixels: Gaussian kernel
Gist: Gaussian kernel
Bag of words: Histogram intersection
Pyr: Pyramid matching kernel
Large Scale Scene Recognition

> 400 categories

> 140,000 images

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

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

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

From the city to the mountains in 10 steps
Exploiting regularities in real-world scenes
Scenes are unique
But not all scenes are so original
But not all scenes are so original
Find similar scenes by matching image descriptors
Find similar scenes by matching image descriptors

Query image

GIST

Top matches
Nearest neighbors classification

- Given a new test sample, assign the label of the nearest neighbor
K-Nearest neighbors classification

Find the K closest points to the test sample
Use labels of the K neighbors to vote

If query lands here, the 5 NN consist of 3 negatives and 2 positives, so we classify it as negative.
instead of using objects labels, the web provides other kinds of metadata associate to large collections of images

Figure 2. The distribution of photos in our database. Photo locations are cyan. Density is overlaid with the jet colormap (log scale).

20 million geotagged and geographic text-labeled images

Hays & Efros. CVPR 2008
Figure 5. *Geolocation performance across features*. Percentage of test cases geolocated to within 200km for each feature. We compare geolocation by 1-NN vs. largest mean-shift mode.
Image completion

Instead, generate proposals using millions of images

Hays, Efros, 2007

Input

16 nearest neighbors (gist+color matching)

output

Hays, Efros, 2007
Scene matching with camera transformations
Scene matching with camera view transformations: Translation

1. Move camera
2. View from the virtual camera
3. Find a match to fill the missing pixels
4. Locally align images
5. Find a seam
6. Blend in the gradient domain
Scene matching with camera view transformations: Camera rotation

1. Rotate camera
2. View from the virtual camera
3. Find a match to fill-in the missing pixels
4. Stitched rotation
5. Display on a cylinder
Scene matching with camera view transformations: Forward motion

1. Move camera
2. View from the virtual camera
3. Find a match to replace pixels
Tour from a single image

Navigate the virtual space using intuitive motion controls
Basic camera motions

Camera translation
Basic camera motions

Forward motion
Basic camera motions

Camera rotation
Exploring famous sites
If images are from the same place...

Google Street View
(controlled image capture)

PhotoToursim/PhotoSynth
[Snively et al., 2006]
(register images based on multi-view geometry)
Dense correspondence between different scenes

Ce Liu, Jenny Yuen, A. Torralba, J. Sivic, B. Freeman
Matching frames / views

The two images are taken from the same scene with different time and/or perspective.
Matching scenes
Two images taken from the same scene category, but different instances
• Contain different objects with different scales, perspectives and spatial location
Image representation

Image gradients

Keypoint descriptor
Matching dense SIFT descriptor

RGB images

SIFT images
p ... position on the grid
s(p) ... SIFT descriptor at position p
w ... displacement vector with components w=(u,v).
p  ... position on the grid
s(p)  ... SIFT descriptor at position p
w  ... displacement vector with components w=(u,v).
The objective function of SIFT flow

• The energy function is similar to that of optical flow

\[ E(w) = \sum_p \left\| s_1(p) - s_2(p + w) \right\|_1 + \frac{1}{\sigma^2} \sum_p \left( u^2(p) + v^2(p) \right) + \sum_{(p,q) \in \varepsilon} \min(\alpha |u(p) - u(q)|, d) + \min(\alpha |v(p) - v(q)|, d) \]

• \( p, q \): grid coordinate, \( w \): displacement vector, \( u, v \): \( x \)- and \( y \)-component, \( s_1, s_2 \): SIFT descriptor

• Decoupled smoothness; truncated L1 norm
Same scene instance matching
Matching different scenes
Matching: objects
Scene matching
Scene matching
Failures

• The nearest neighbors may not contain similar scenes or object categories (SIFT flow tries to match image structures anyway)
With good image correspondence and a lot of data...

The space of world images

• Labels
• Motion
• Depth
• ...

Hays, Efros, Siggraph 2006
Russell, Liu, Torralba, Fergus, Freeman. NIPS 2007
Predicting events

Predicting events

Query

Query

Retrieved video

Motion synthesis results

Still image

Video of the best match

Motion synthesis results
Retrieved video

Synthesized video

Query

Synthesized video

Discussion

• Regularities in scene appearance can be used for a number of applications (label transfer - recognition, scene completion, gps location prediction, event prediction...)

• Performance depends on the quality of the matches, i.e. is the particular scene represented in the database?
  
  • Increase database size [Torralba, PAMI 2008].
  
  • Combine multiple database images [Russell et al. 2009]

However, some “atypical” scenes might still not be represented well.
Today: Scenes and objects

1. Scenes as textures (without modeling objects and their relations)

2. Detecting single objects in context; geometric context.

3. Recognizing multiple objects in an image.

4. Recognizing unseen objects.
Part II: Scene as a context for single object classes
Who needs context anyway?
We can recognize objects even out of context

Banksy
Why is context important?

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

• Context defines what an unexpected event is
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.
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)
What is the context for a single object category?
The influence of an object extends beyond its physical boundaries
Global and local representations

Image index: Summary statistics, configuration of textures

Urban street scene

Global and local representations
An integrated model of Scenes, Objects, and Parts

Scene gist features

\[ P(N_{\text{car}} \mid S = \text{street}) \]

\[ P(N_{\text{car}} \mid S = \text{park}) \]
Context driven object detection

Scene

Scene gist features

\[ P(N_{\text{car}} | S = \text{street}) \]
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

\[ P(F, S | x, d, g) \propto p(F | S) p(S | g) \prod_{i} p(x_i | g) \prod_{i} N(x_i; \mu_b, \sigma_b^2) \prod_{i} N(d_i; \mu_{tp}, \sigma_{tp}^2) \prod_{i} N(d_i; \mu_{tn}, \sigma_{tn}^2) \]
A car out of context ...
See also…

H. Harzallah, F. Jurie and C. Schmid,
*Combining efficient object localization and image classification*, ICCV 2009

V. Delaitre, I. Laptev and J. Sivic
*Action recognition in still images...*, BMVC 2010
We are wired for 3D

~6cm
We can not shut down 3D perception

(c) 2006 Walt Anthony
Scenes rule over objects

3D percept is driven by the scene, which imposes its ruling to the objects
3D from pixel values


A. Saxena, M. Sun, A. Y. Ng. "Learning 3-D Scene Structure from a Single Still Image"
In ICCV workshop on 3D Representation for Recognition (3dRR-07), 2007.
Confidences from Boosted Decision Trees

\[ P(label \mid \text{good segment, data}) \]

[Collins et al. 2002]
Surface Estimation

[Hoiem, Efros, Hebert ICCV 2005]
Object Support
3D Scene Context

Hoiem, Efros, Hebert ICCV 2005
3D scene context
Object Size ↔ Camera Viewpoint
Object Size ↔ Camera Viewpoint

Input Image

Loose Viewpoint Estimate
Object Size ↔ Camera Viewpoint

Object Position/Sizes ↔ Viewpoint
Object Size ↔ Camera Viewpoint
Object Size $\leftrightarrow$ Camera Viewpoint
Object Size ↔ Camera Viewpoint
How surfaces and viewpoint help detection

Image

\[ P(\text{object}) \]

\[ P(\text{surfaces}) \]

\[ P(\text{viewpoint}) \]

\[ P(\text{object} \mid \text{surfaces}) \]

\[ P(\text{object} \mid \text{viewpoint}) \]
How surfaces and viewpoint help detection

Image

P(surfaces)

P(viewpoint)

P(object)

P(object | surfaces, viewpoint)
Qualitative Results

Car: TP / FP  Ped: TP / FP

Initial: 2 TP / 3 FP

Final: 7 TP / 4 FP

Local Detector from [Murphy-Torralba-Freeman 2003]
3D City Modeling using Cognitive Loops

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.
Single view metrology
Criminisi, et al. 1999

Need to recover:
• Ground plane
• Reference height
• Horizon line
• Where objects contact the ground
Announcements

• Final project presentations next week!
  http://www.di.ens.fr/willow/teaching/recvis10/final_project/
  – Send us the **project title** and **names** of people in the group asap!
  – Schedule of the presentations will be emailed this week.

• **Final project report deadline extended to January 5**th.

• If you have any suggestions or comments on the course, please fill-in the feed-back form.