Computer Vision II –
Scene Understanding

Michael Yang
Roadmap (4 lectures)

• Object Detection (26.06)

• Image Categorization (03.07)

• Convolutional neural network (10.07)

• Scene Understanding (17.07)

• Poster, Q&A (24.07)
Slides credits

- Bernt Schiele
- Li Fei-Fei
- Rob Fergus
- Kirsten Grauman
- Derek Hoiem
- Antonio Torralba
- James Hays
- Jianxiong Xiao
- Stefan Roth
- Andreas Geiger
- Jamie Shotton
- Antonio Criminisi
- Carsten Rother
Roadmap (last lecture)

• Shallow vs. deep architectures

• Convolutional neural network (CNN)

• Training CNN

• CNN for X
“Shallow” vs. “deep” architectures

Traditional recognition: “Shallow” architecture

Image/Video Pixels \arrow{->} Hand-designed feature extraction \arrow{->} Trainable classifier \arrow{->} Object Class

Deep learning: “Deep” architecture

Image/Video Pixels \arrow{->} Layer 1 \arrow{->} \ldots \arrow{->} Layer N \arrow{->} Simple classifier \arrow{->} Object Class
Neural Net Events

- Founded by Warren McCulloch and Walter Pitts in 1943
- Back propagation by Rumelhart and Hinton in 1986
- Criticism by Minsky in his book "Perceptron" in 1969
- Deep belief networks by Hinton in 1998
- ImageNet classification over millions of images in 2012
- Convolutional neural networks
- Google CAT in 2012
- ImageNet classification over millions of images in 2012
- Convolutional neural networks

ImageNet classification over millions of images
Convolutional Neural Network (CNN/Convnet)

- Neural network with specialized connectivity structure
- Stack multiple stages of feature extractors
- Higher stages compute more global, more invariant features
- Classification layer at the end

Convolutional Neural Network (CNN/Convnet)

• Feed-forward feature extraction:
  1. Convolve input with learned filters
  2. Non-linearity
  3. Spatial pooling
  4. Normalization

• Supervised training of convolutional filters by back-propagating classification error
ImageNet Challenge 2012

- Similar framework to LeCun’98 but:
  - Bigger model (7 hidden layers, 650,000 units, 60,000,000 params)
  - More data ($10^6$ vs. $10^3$ images)
  - GPU implementation (50x speedup over CPU)
    - Trained on two GPUs for a week
  - Better regularization for training (DropOut)

Training CNN

• Backpropagation + stochastic gradient descent
  • *Neural Networks: Tricks of the Trade*

• Dropout

• Data augmentation

• Initialization
  • Transfer learning
Transfer Learning

- Improvement of learning in a **new** task through the *transfer of knowledge* from a **related** task that has already been learned.

- Weight initialization for CNN

---

Learning and Transferring Mid-Level Image Representations using Convolutional Neural Networks [Oquab et al. CVPR 2014]
Beyond classification

- Detection
- Segmentation
- Regression
- Pose estimation
- Matching patches
- Synthesis

and many more...
Tools

- Caffe
- cuda-convnet2
- Torch
- MatConvNet
- Pylearn2
Roadmap (this lecture)

• Defining the Problem

• Context

• Spatial Layout

• 3D Scene Understanding
Scene Understanding

• **What is goal of scene understanding:**
  • Build machine that can see like humans to automatically interpret the content of the images

• **Comparing with traditional vision problem:**
  • Study on larger scale
  • Human vision related tasks
Larger Scale

More image information. Context information.

focal length = 35 mm
Human vision related task

More similar as the way that human understand the image
Infer more useful information from image
How DO human learn?

• Bayesian Rules:

\[ P(A | B) = P(B | A) \times P(A) / P(B) \]

• In practice: Infer abstract knowledge based on observation

\[ P(W | I) = P(I | W) \times P(W) / P(I) \]

µ \( P(I | W) \times P(W) \)

Posterior probability

Likelihood: The probability of getting I given model W

Prior: The probability of W w/o seeing any observation
How DO human learn?

• To teach human baby what is “horse”: show 3 pictures and let them learn by themselves.

• They can be very successful to learn the correct concept.

• But all the following concepts can explain the images:
  • “horse” = all horse
  • “horse” = all horse but not Clydesdales
  • “horse” = all animal

/ =

“horse”
Roadmap (this lecture)

• Defining the Problem

• Context

• Spatial Layout

• 3D Scene Understanding
Objects usually are surrounded by a scene that can provide context in the form of nearby objects, surfaces, scene category, geometry, etc.
Contextual Reasoning

• Definition: Making a decision based on more than *local* image evidence.
Context provides clues for function

• What is this?
Context provides clues for function

• What is this?

• Now can you tell?
Context provides clues for function

- once more how amazing is the visual system
Context provides clues for function

• once more how amazing is the visual system
Is local information enough?
Is local information enough?

- Local features
- Contextual features

Distance

Information

Local features vs. Contextual features with respect to distance.
Context in Recognition

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.
Context in Recognition
Context in Recognition

Look-Alikes by Joan Steiner
Context in Recognition

• Pictures shown for 150 ms
• Objects in appropriate context were detected more accurately than objects in an inappropriate context
• Scene consistency affects object detection

Biederman 1982
Why is context important?

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

• Context defines what an unexpected event is
There are many types of context

- **Local pixels**
  - window, surround, image neighborhood, object boundary/shape, global image statistics

- **2D Scene Gist**
  - global image statistics

- **3D Geometric**
  - 3D scene layout, support surface, surface orientations, occlusions, contact points, etc.

- **Semantic**
  - event/activity depicted, scene category, objects present in the scene and their spatial extents, keywords

- **Photogrammetric**
  - camera height orientation, focal length, lens distortion, radiometric, response function

- **Illumination**
  - sun direction, sky color, cloud cover, shadow contrast, etc.

- **Geographic**
  - GPS location, terrain type, land use category, elevation, population density, etc.

- **Temporal**
  - nearby frames of video, photos taken at similar times, videos of similar scenes, time of capture

- **Cultural**
  - photographer bias, dataset selection bias, visual cliches, etc.

from Divvala et al. CVPR 2009
Roadmap (this lecture)

• Defining the Problem

• Context

• Spatial Layout

• 3D Scene Understanding
Spatial layout is especially important

1. Context for recognition
Spatial layout is especially important

1. Context for recognition
Spatial layout is especially important

1. Context for recognition
2. Scene understanding
Spatial layout is especially important

1. Context for recognition
2. Scene understanding
3. Many direct applications
   a) Assisted driving
   b) Robot navigation/interaction
   c) 2D to 3D conversion for 3D TV
   d) Object insertion
Spatial Layout: 2D vs. 3D
Context in Image Space

[Torralba Murphy Freeman 2004]

[Kumar Hebert 2005]

[He Zemel Cerreia-Perpiñán 2004]
But object relations are in 3D...
How to represent scene space?
Wide variety of possible representations

Scene-Level Geometric Description

a) Gist, Spatial Envelope

b) Stages
Wide variety of possible representations

Retinotopic Maps

(c) Geometric Context

d) Depth Maps
Wide variety of possible representations

Highly Structured 3D Models

- e) Ground Plane
- f) Ground Plane with Billboards
- g) Ground Plane with Walls
- h) Blocks World
- i) 3D Box Model
Key Trade-offs

• Level of detail: rough “gist”, or detailed point cloud?
  • Precision vs. accuracy
  • Difficulty of inference

• Abstraction: depth at each pixel, or ground planes and walls?
  • What is it for: e.g., metric reconstruction vs. navigation
Low detail, Low abstraction

Holistic Scene Space: “Gist”

Oliva & Torralba 2001

Torralba & Oliva 2002
High detail, Low abstraction

Depth Map

Saxena, Chung & Ng 2005, 2007
Medium detail, High abstraction

Room as a Box

[Hedau Hoiem Forsyth 2009]
The challenge
Our World is Structured

Abstract World

Our World

Image Credit (left): F. Cunin and M.J. Sailor, UCSD
Learn the Structure of the World

Training Images
Infer the most likely interpretation

Unlikely

Likely
Geometry estimation as recognition

Region → Features
Color
Texture
Perspective
Position → Surface Geometry Classifier

Vertical, Planar

Training Data
Surface Layout Algorithm

Input Image → Segmentation → Features (Perspective, Color, Texture, Position) → Surface Labels

Training Data

[Trained Region Classifier]

[Hoiem Efros Hebert 2007]
Surface Layout Algorithm

Input Image → Multiple Segmentations → Features (Perspective, Color, Texture, Position) → Confidence-Weighted Predictions → Final Surface Labels

Training Data → Trained Region Classifier → [Hoiem Efros Hebert 2007]
Results

Input Image  Ground Truth  Result
Failures: Reflections, Rare Viewpoint

Input Image  |  Ground Truth  |  Result
**Average Accuracy**

Main Class: 88%

Subclasses: 61%

<table>
<thead>
<tr>
<th>Main Class</th>
<th>Support</th>
<th>Vertical</th>
<th>Sky</th>
</tr>
</thead>
<tbody>
<tr>
<td>Support</td>
<td>0.84</td>
<td>0.15</td>
<td>0.00</td>
</tr>
<tr>
<td>Vertical</td>
<td>0.09</td>
<td>0.90</td>
<td>0.02</td>
</tr>
<tr>
<td>Sky</td>
<td>0.00</td>
<td>0.10</td>
<td>0.90</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Vertical Subclass</th>
<th>Left</th>
<th>Center</th>
<th>Right</th>
<th>Porous</th>
<th>Solid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Left</td>
<td>0.37</td>
<td>0.32</td>
<td>0.08</td>
<td>0.09</td>
<td>0.13</td>
</tr>
<tr>
<td>Center</td>
<td>0.05</td>
<td>0.56</td>
<td>0.12</td>
<td>0.16</td>
<td>0.12</td>
</tr>
<tr>
<td>Right</td>
<td>0.02</td>
<td>0.28</td>
<td>0.47</td>
<td>0.13</td>
<td>0.10</td>
</tr>
<tr>
<td>Porous</td>
<td>0.01</td>
<td>0.07</td>
<td>0.03</td>
<td>0.84</td>
<td>0.06</td>
</tr>
<tr>
<td>Solid</td>
<td>0.04</td>
<td>0.20</td>
<td>0.04</td>
<td>0.17</td>
<td>0.55</td>
</tr>
</tbody>
</table>
Automatic Photo Popup

Labeled Image → Fit Ground-Vertical Boundary with Line Segments → Form Segments into Polylines → Cut and Fold

Final Pop-up Model

[Hoiem Efros Hebert 2005]
Mini-conclusions

- Can learn to predict surface geometry from a single image
- Very rough models, much room for improvement
Things to remember

• Objects should be interpreted in the context of the surrounding scene
  • Many types of context to consider

• Spatial layout is an important part of scene interpretation, but many open problems
  • How to represent space?
  • How to learn and infer spatial models?

• Consider trade-offs of detail vs. accuracy and abstraction vs. quantification
Roadmap (this lecture)

• Defining the Problem

• Context

• Spatial Layout

• 3D Scene Understanding
10 Minutes break

Evaluation
Complete Scene Understanding

Involves

- Localization of all instances of foreground objects ("things")
- Localization of all background classes ("stuff")
- Pixel-wise segmentation
- 3D reconstruction
- Pose detection
- Action recognition
- Event recognition
- .....
Semantic Scene Understanding

We're interested in whole scene understanding. Given an image, detect every *thing* in it.

**Thing**: An object with a specific size and shape.

*Adelson, Forsyth et al. 96*

*Slides credit: Ľubor Ladický*
Semantic Scene Understanding

We're interested in whole scene understanding
Given an image, label all the *stuff*

**Stuff**: Material defined by a homogeneous or repetitive pattern, with no specific spatial extent / shape.

*Adelson, Forsyth et al. 96*
Combining Object Detectors and CRFs

Why not combine?

– State of the art sliding window object detection

– State of the art segmentation techniques
Algorithms for Object Localization

Sliding window detectors

- HOG descriptor (Dalal & Triggs CVPR05)
- Based on histograms of features (Vedaldi et al. ICCV09)
- Part-based models (Felzenszwalb et al. CVPR09)
Sliding window detectors

- Sliding window + Segmentation
  - OBJCUT (Kumar et al. 05)
  - Updating colour model (GrabCut - Rother et al. 04)
Sliding window detectors

Sliding window detectors not good for “stuff”

Sky is irregular shape not suited to the sliding window approach
Algorithms for Object-class Segmentation

Pairwise CRF over pixels

Input image

Final segmentation

CRF construction

Training of Potentials

\[ E(x) = \sum_{i \in V} \psi_i(x_i) + \sum_{i \in V, j \in N_i} \psi_{ij}(x_i, x_j) \]

Shotton et al. ECCV06
Algorithms for Object-class Segmentation

Pairwise CRF over Super-pixels / Segments

Input image

Unsupervised segmentation

Training of potentials

MAP

$$E(x) = \sum_{i \in V} \psi_i(x_i) + \sum_{i \in V, j \in N_i} \psi_{ij}(x_i, x_j)$$

Final segmentation

Batra et al. CVPR08, Yang et al. CVPR07, Zitnick et al. CVPR08, Rabinovich et al. ICCV07, Boix et al. CVPR10
Algorithms for Object-class Segmentation

Associative Hierarchical CRF

Input image

Multiple segmentations or hierarchies

MAP

CRF construction

Final segmentation

Ladický et al. ICCV09, Russell et al. UAI10
CRF Formulation with Detectors

- CRF formulation altered with a potential for each detection

\[ E(x) = E_{pix}(x) + \sum_{d \in \mathcal{D}} \psi_d(x_d, H_d, l_d) \]

- AH-CRF energy without detectors
- Set of pixels of d-th detection
- Classifier response
- Detected label

CRF graph over pixels
CRF Formulation with Detectors

- Joint CRF formulation should contain
  - Possibility to reject detection hypothesis
  - Recover the status of the detection (0 / 1)
- Thus, potential is a minimum over indicator variable \( y_d \in \{0, 1\} \)

\[
\psi_d(x_d, H_d, l_d) = \min_{y_d} \phi_d(y_d, x_d, H_d, l_d)
\]
Results on CamVid dataset

<table>
<thead>
<tr>
<th></th>
<th>Building</th>
<th>Tree</th>
<th>Sky</th>
<th>Car</th>
<th>Sign-Symbol</th>
<th>Road</th>
<th>Pedestrian</th>
<th>Fence</th>
<th>Column-Pole</th>
<th>Sidewalk</th>
<th>Bicyclist</th>
<th>Global</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brostow et al.</td>
<td>46.2</td>
<td>61.9</td>
<td>89.7</td>
<td>68.6</td>
<td>42.9</td>
<td>89.5</td>
<td>53.6</td>
<td>46.6</td>
<td>0.7</td>
<td>60.5</td>
<td>22.5</td>
<td>69.1</td>
<td>53.0</td>
</tr>
<tr>
<td>Sturgess et al.</td>
<td>84.5</td>
<td>72.6</td>
<td>97.5</td>
<td>72.7</td>
<td>34.1</td>
<td>95.3</td>
<td>34.2</td>
<td>45.7</td>
<td>8.1</td>
<td>77.6</td>
<td>28.5</td>
<td>83.8</td>
<td>59.2</td>
</tr>
<tr>
<td>Our method</td>
<td>81.5</td>
<td>76.6</td>
<td>96.2</td>
<td>78.7</td>
<td>40.2</td>
<td>93.9</td>
<td>43.0</td>
<td>47.6</td>
<td>14.3</td>
<td>81.5</td>
<td>33.9</td>
<td>83.8</td>
<td>62.5</td>
</tr>
</tbody>
</table>

*Brostow et al. ECCV08, Sturgess et al. BMVC09*
Results on CamVid dataset

Result without detections

Set of detections

Final Result

Also provides number of object instances (using $y_d$'s)
### Results on VOC2009 dataset

<table>
<thead>
<tr>
<th></th>
<th>Background</th>
<th>Aeroplane</th>
<th>Bicycle</th>
<th>Bird</th>
<th>Boat</th>
<th>Bottle</th>
<th>Bus</th>
<th>Car</th>
<th>Cat</th>
<th>Chair</th>
<th>Cow</th>
<th>Dining table</th>
<th>Dog</th>
<th>Horse</th>
<th>Motorbike</th>
<th>Person</th>
<th>Potted plant</th>
<th>Sheep</th>
<th>Sofa</th>
<th>Train</th>
<th>TV/monitor</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>BONN_SVM-SEGm</td>
<td>83.9</td>
<td>64.3</td>
<td>21.8</td>
<td>21.7</td>
<td>32.0</td>
<td>40.2</td>
<td>57.3</td>
<td>49.4</td>
<td>38.8</td>
<td>5.2</td>
<td>28.5</td>
<td>22.0</td>
<td>19.6</td>
<td>33.6</td>
<td>45.5</td>
<td>33.6</td>
<td>27.3</td>
<td>40.4</td>
<td>18.1</td>
<td>33.6</td>
<td>46.1</td>
<td>36.3</td>
</tr>
<tr>
<td>CVC_HOCRf</td>
<td>80.2</td>
<td>67.1</td>
<td>26.6</td>
<td>30.3</td>
<td>31.6</td>
<td>30.0</td>
<td>44.5</td>
<td>41.6</td>
<td>25.2</td>
<td>5.9</td>
<td>27.8</td>
<td>11.0</td>
<td>23.1</td>
<td>40.5</td>
<td>53.2</td>
<td>32.0</td>
<td>22.2</td>
<td>37.4</td>
<td>23.6</td>
<td>40.3</td>
<td>30.2</td>
<td>34.5</td>
</tr>
<tr>
<td>UOCTTL_SVM-MDPM</td>
<td>78.9</td>
<td>35.3</td>
<td>22.5</td>
<td>19.1</td>
<td>23.5</td>
<td>36.2</td>
<td>41.2</td>
<td>50.1</td>
<td>11.7</td>
<td>8.9</td>
<td>28.5</td>
<td>1.4</td>
<td>5.9</td>
<td>24.0</td>
<td>35.3</td>
<td>33.4</td>
<td>35.1</td>
<td>27.7</td>
<td>14.2</td>
<td>34.1</td>
<td>41.8</td>
<td>29.0</td>
</tr>
<tr>
<td>NECUIUC_CLS-DTCT</td>
<td>81.8</td>
<td>41.9</td>
<td>23.1</td>
<td>22.4</td>
<td>22.0</td>
<td>27.8</td>
<td>43.2</td>
<td>51.8</td>
<td>25.9</td>
<td>4.5</td>
<td>18.5</td>
<td>18.0</td>
<td>23.5</td>
<td>26.9</td>
<td>36.6</td>
<td>34.8</td>
<td>8.8</td>
<td>28.3</td>
<td>14.0</td>
<td>35.5</td>
<td>34.7</td>
<td>29.7</td>
</tr>
<tr>
<td>LEAR_SEGDET</td>
<td>79.1</td>
<td>44.6</td>
<td>15.5</td>
<td>20.5</td>
<td>13.3</td>
<td>28.8</td>
<td>29.3</td>
<td>35.8</td>
<td>25.4</td>
<td>4.4</td>
<td>20.3</td>
<td>1.3</td>
<td>16.4</td>
<td>28.2</td>
<td>30.0</td>
<td>24.5</td>
<td>12.2</td>
<td>31.5</td>
<td>18.3</td>
<td>28.8</td>
<td>31.9</td>
<td>25.7</td>
</tr>
<tr>
<td>BROOKESMSRC_AHCRF</td>
<td>79.6</td>
<td>48.3</td>
<td>6.7</td>
<td>19.1</td>
<td>10.0</td>
<td>16.6</td>
<td>32.7</td>
<td>38.1</td>
<td>25.3</td>
<td>5.5</td>
<td>9.4</td>
<td>25.1</td>
<td>13.3</td>
<td>12.3</td>
<td>35.5</td>
<td>20.7</td>
<td>13.4</td>
<td>17.1</td>
<td>18.4</td>
<td>37.5</td>
<td>36.4</td>
<td>24.8</td>
</tr>
<tr>
<td>Our method</td>
<td>81.2</td>
<td>46.1</td>
<td>15.4</td>
<td>24.6</td>
<td>20.9</td>
<td>36.9</td>
<td>50.0</td>
<td>43.9</td>
<td>28.4</td>
<td>11.5</td>
<td>18.2</td>
<td>14.7</td>
<td>25.4</td>
<td>25.1</td>
<td>37.7</td>
<td>34.1</td>
<td>27.7</td>
<td>29.6</td>
<td>18.4</td>
<td>43.8</td>
<td>40.8</td>
<td>32.1</td>
</tr>
</tbody>
</table>
3D Traffic Scene Understanding

KITTI (video)

3D Traffic Scene Understanding from Movable Platforms

Andreas Geiger
3D Traffic Scene Understanding

• Goal: Infer from short video sequences (moving observer)
  • Topology and geometry of the scene
  • Semantic information (traffic situation)

• Probabilistic generative model of 3D urban scenes
Topology and Geometry Model

Topology Model \((\kappa)\)  

Geometry Model \((c, w, r, \alpha)\)

Road Layout \(\mathcal{R} = \{\kappa, c, w, r, \alpha\}\)
Image Evidence $E = \{T; V; S; F; O\}$
Probabilistic Graphical Model

\[ p(\mathcal{E}, \mathcal{R}) = p(\mathcal{R}) \prod_i p(t_i | \mathcal{R}) \prod_i p(v_i | \mathcal{R}) \prod_i p(s_i | \mathcal{R}) \prod_i p(\rho_i | \mathcal{R}) \prod_i p(f_i | \mathcal{R}) \]
Vehicle Tracklets

- Object detection [Felzenszwalb et al. 2010]
- Associate objects over time (tracking by detection)

- Projection to 3D object tracklet \( t = \{d_1, \ldots, d\} \)
  (\( d \) captures the object location and orientation)
Probabilistic Graphical Model

Vanishing Points
Probabilistic Graphical Model

Semantic Labels
Probabilistic Graphical Model

Occupancy, Scene Flow
Inference

Denote

- $\mathcal{E}$ the image evidence
- $\mathcal{R}$ the road layout
- $\mathcal{C}$ the location of cars in the scene

Given $\mathcal{E}$, inference of $\mathcal{R}$ and $\mathcal{C}$ is solved in two steps:

- Infer road layout $\mathcal{R}$ while marginalizing $\mathcal{C}$
  \[ \hat{\mathcal{R}} = \arg\max_{\mathcal{R}} p(\mathcal{R}|\mathcal{E}) \quad \text{(Metropolis-Hastings)} \]

- Infer car locations $\mathcal{C}$ using MAP road layout $\mathcal{R}$
  \[ \hat{\mathcal{C}} = \arg\max_{\mathcal{C}} p(\mathcal{C}|\mathcal{E}, \mathcal{R}) \quad \text{(Dynamic programming)} \]
Experimental Results

Experiments
• 113 sequences 5-30 seconds (9438 frames)
• Best results when combining all feature cues
• Most important: Occupancy grid, tracklets, 3D scene flow
• Less important: Semantic labels, vanishing points

Metrics
• Topology Accuracy: 92.0%
• Location Error: 3.0 m
• Street Orientation Error: 3.0
• Tracklet-to-Lane Accuracy: 82.0%
• Vehicle Orientation Error: 14.0
Experimental Results
3D Scene Understanding

• Defining the Problem

• Context

• Spatial Layout

• 3D Scene Understanding