Computer Vision: Looking Back to Look Forward

Svetlana Lazebnik
IRIM Short Course
Spring 2020
Outline of lectures

Tues. 1/28: Where are we going, where have we been?
Wed. 1/29: Stepping back: What is the nature of visual perception?
**Thurs. 1/30: History of ideas in recognition – Part I**

Tues. 2/4: History of ideas in recognition – Part II – **TSRB banquet hall**
Wed. 2/5: Emerging trends – or, what should we be working on?
Thurs. 2/6: Ethical issues for computer vision researchers
History of Recognition: Outline

Object *instance* recognition

Object *category* recognition

Texture recognition

Scene recognition
History of Recognition: Outline

- **Object instance recognition**
  - 3D geometric
    - Blocks world
    - Alignment
    - Aspect graphs
    - Invariants
    - Primitive-based approaches: generalized cylinders, geons
  - Appearance-based
    - Eigenfaces
    - Appearance manifolds
    - Color indexing
  - Keypoint-based
    - Schmid & Mohr
    - Lowe
    - Video Google

- **Texture recognition**
  - Textons
  - Filter banks
  - Histograms

- **Object category recognition**
  - Pictorial structures
  - Early statistical: Late 90s face detectors
  - Constellation models
  - Bags of features
  - HOG, DPMs

- **Scene recognition**
  - Early knowledge-based
  - GIST
  - Bags of features, pyramids
History of Recognition: Outline

Object instance recognition

Texture recognition

Object category recognition

Scene recognition
Recognizing known, rigid 3D objects

- Recognition as inverse graphics
- Ignore image intensities, focus on edges or point features
- Main problem is alignment, or estimation of object pose that would give rise to projected image

J. Mundy, Object recognition in the geometric era: A retrospective, 2006
Model-based methods: Alignment

Perkins (1978)  
Grimson & Lozano-Perez (1984)  
Lowe (1985)  

Ayache & Faugeras (1986)  
Huttenlocher & Ullman (1987)
View-based methods: Aspect graphs

“Model-free” methods: Invariants

Examples of model features used to compute invariants

J. Mundy and A. Zisserman (eds.), *Geometric Invariance in Computer Vision*, 1992
3D shape primitives: Generalized cylinders

- Binford (1971), Agin & Binford (1973)
- Brooks (1981)
- Zerroug & Nevatia (1994)
- Zisserman et al. (1995)
Marr’s 3D object representation

Figure 5-3. This diagram illustrates the organization of shape information in a 3-D model description. Each box corresponds to a 3-D model, with its model axis on the left side of the box and the arrangement of its component axes on the right. In addition, some component axes have 3-D models associated with them, as indicated by the way the boxes overlap. The relative arrangement of each model’s component axes, however, is shown improperly, since it should be in an object-centered system rather than the viewer-centered projection used here (a more correct 3-D model is given by the table shown in Figure 5-5c). The important characteristics of this type of organization are: (1) Each 3-D model is a self-contained unit of shape information and has a limited complexity; (2) information appears in shape contexts appropriate for recognition (the disposition of a finger is most stable when specified relative to the hand that contains it); and (3) the representation can be manipulated flexibly. This approach limits the representation’s scope, however, since it is only useful for shapes that have well-defined 3-D model decompositions.


Marr & Nishihara (1978)
Psychological theory: Recognition by components

Primitives (*geons*)

- Cube: Straight Edge, Straight Axis, Constant
- Wedge: Straight Edge, Straight Axis, Expanded
- Pyramid: Straight Edge, Straight Axis, Expanded
- Cylinder: Curved Edge, Straight Axis, Constant
- Barrel: Curved Edge, Straight Axis, Exp. & Cont.
- Arch: Straight Edge, Curved Axis, Constant
- Cone: Curved Edge, Straight Axis, Expanded
- Expanded Cylinder: Curved Edge, Straight Axis, Expanded
- Handle: Curved Edge, Curved Axis, Constant
- Expanded Handle: Curved Edge, Curved Axis, Expanded

Objects

- Watering can
- Iron
- Lamp
- Sailboat

Biederman (1987)

Geometric recognition: Discussion

• Summary
  • Emphasis on 3D modeling was sign of “right” thinking
  • People liked the math, combinatorics involved in reasoning about 3D shape, invariants, and aspect graphs
  • For rigid objects, some robustness to occlusion, viewpoint change was achieved
  • Practical applications in industrial inspection, robotics, target recognition
  • Recognition by components conceptually appealing, gave rise to supporting theories in cognitive science

• What went wrong?
  • There was no reliable way to establish matches between model and image features
  • Numerical pose estimation techniques were not mature enough?
  • Nobody knew how to or compute part decompositions of general categories or model intra-class variations of 3D shape
Today: Revival of 3D primitives?

“Here we do not wish to reprise the classic debates on the value of volumetric primitives – while they were oversold in the 70s and 80s, they suffer from complete neglect now, and we hope that this demonstration of feasibility of learning how to assemble an object from volumetric primitives will reignite interest.”

Appearance-based instance recognition

- Maybe grayscale (or color) images are not as scary as they seem?

![COIL-100 Dataset](image-url)
Color histograms

M. Swain and D. Ballard, Color Indexing, IJCV 1991
Eigenfaces


<table>
<thead>
<tr>
<th>Condition</th>
<th>Correct/Unknown Recognition Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lighting</td>
<td>Orientation</td>
</tr>
<tr>
<td>Forced classification</td>
<td>96/0</td>
</tr>
<tr>
<td>Forced 100% accuracy</td>
<td>100/19</td>
</tr>
<tr>
<td>Forced 20% unknown rate</td>
<td>100/20</td>
</tr>
</tbody>
</table>
Appearance manifolds

H. Murase and S. Nayar, [Visual learning and recognition of 3-d objects from appearance](https://doi.org/10.1007/BF00887090), IJCV 1995

J. Mundy et al., [An Experimental Comparison of Appearance and Geometric Model Based Recognition](https://doi.org/10.1007/978-1-4612-5800-1), 1996
Keypoints + local appearance

Keypoints: corners (Harris & Stephens, 1988)
Descriptors: Local jets (Koenderink & Van Doorn, 1987)

Scale-invariant feature transform (SIFT)

D. Lowe, *Object recognition from local scale-invariant features*, ICCV 1999

D. Lowe, *Distinctive image features from scale-invariant keypoints*, IJCV 2004
Covariant keypoints + invariant descriptors
Large-scale image search

• Borrow ideas from text indexing and retrieval!
  • Visual words, inverted indices, stop words, query expansion...

Sivic & Zisserman (2003), Nister & Stewenius (2006), Philbin et al. (2007), etc.

Figure source:
K. Grauman and B. Leibe
Large-scale image registration, reconstruction

San Marco Square: 13,699 images, 4,515,157 points

Snavely et al. (2006), Agarwal et al. (2009), Frahm et al. (2012), etc.

Sattler et al. (2011)
Appearance-based methods: Discussion

• Summary
  • In the early- to mid-90’s, people realized that you could use images to represent themselves
    • This was supported by viewpoint-centric theories of human recognition – see, e.g., Edelman & Weinshall (1990), Tarr & Pinker (1991)
  • Initial appearance-based models (eigenfaces, appearance manifolds) were global and lacking invariance
  • In the late 90’s, methods based on local keypoint invariants achieved a good combination of discriminability, robustness and invariance

• What went wrong?
  • Nothing really, except that keypoint detectors and descriptors (and subsequent indexing pipelines) had to be extensively hand-engineered
History of Recognition: Outline

Object *instance* recognition

Object *category* recognition

Texture recognition

Scene recognition
Texture recognition

Statistics of textons
- Julesz (1962, 1981)

Autocorrelation, co-occurrence functions
- Haralick (1979)

Filter banks, pyramids

K-means clustering, texton vocabularies
- Leung & Malik (1999)

Blobs
- Voorhees & Poggio (1987)

Blobs + textons
- Lazebnik et al. (2003)
Texture recognition: Discussion

• Summary
  • Texture recognition was the only sub-field in which statistical methods held sway from the beginning
  • Dominant model is local statistics or histograms of local appearance
    • For a long time, this was assumed to be appropriate only for textures
  
• At this point, texture is mainly interesting as part of scenes
History of Recognition: Outline

Object instance recognition

Texture recognition

Object category recognition

Scene recognition

MISSION: ACCOMPLISHED
Category recognition: Deformable templates
Deformable templates

Deformable templates

Yuille (1991)

Cootes & Taylor (1995)
Pictorial structures revived


P. Felzenszwalb and D. Huttenlocher, *Pictorial structures for object recognition*, IJCV 2005
Finding people was really hard

**Fig. 6.** Typical control images wrongly classified as containing naked people. These images contain people or skin-colored material (animal skin, wood, bread, off-white walls) and structures which the geometric grouper mistakes for spines or girdles. The grouper is frequently confused by groups of parallel edges, as in the industrial image.

**Fleck, Forsyth & Bregler (1996)**

**Figure 17.** A negative image for which a human assembly was found. The assembly indeed looks like a configuration of a person. A better segment finder would not produce these segments and thus a person would not be detected. The white regions in the image are the pixels that have been masked out because they could not belong to a person due to their color.

**Ioffe & Forsyth (2001)**
Not so today...

F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero and M. Black, 
Keep it SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image, ECCV 2016

A. Kanazawa, J. Zhang, P. Felsen, J. Malik, 
Learning 3D Human Dynamics from Video, CVPR 2019
Constellation models

Deformable templates: Discussion

• Summary
  • Idea: represent a category as a group of meaningful parts (or primitives) that can deform in some constrained way
  • Optimize an energy function (possibly derived from a probabilistic formalism) to find the “best” part layout in the image
  • Such models are also known as part-based and are conceptually similar to primitive-based geometric models of the 70’s and 80’s

• What went wrong?
  • Until the mid-2000s, there were no reliable part (or primitive) detectors
Features and classifiers

- **Appearance manifolds + neural network**
  - Sung & Poggio (1994)

- **Support vector machines**
  - Osuna, Freund, Girosi (1997)

- **Neural network**
  - Rowley, Baluja, Kanade (1998)

- **Statistics of feature responses, probabilistic classifier**

\[
\prod_{j=1}^{n_{\text{macro}}} \prod_{i=1}^{n_{\text{sub}}} \frac{P(q_j|i, \text{object})P(pos_i|q_j, \text{object})}{P(q_j|i, \text{object})} \geq \lambda = \frac{P(\text{object})}{P(\text{object})}
\]

  - Schneiderman & Kanade (1998)

- **Rectangle features, boosting**

  - Viola & Jones (2001)
Datasets

Caltech-4

Instances

Alignment

Aspect graphs

Generalized cylinders

1970

1980

1990

2000

Invariants

Keypoints

Appearance models

Appearance models

SIFT

Video Google

Textons

Filter banks

Texton histograms

Bags of features

Misc. statistics

Deformable templates

Textons

Filter banks

Texton histograms

Bags of features

Probabilistic models

Constellation models

Pictorial structures

Object categories

Generalized cylinders

Aspect graphs

Alignment

Misc. statistics

Deformable templates

Textons

Filter banks

Texton histograms

Bags of features

Probabilistic models

Constellation models

Pictorial structures

Datasets

Caltech-4
Datasets

Caltech-4

Facets

Instances

Alignment

Aspect graphs

Generalized cylinders

Textures

Misc. statistics

Textons

Filter banks

Deformable templates

Object categories

Probabilistic models

Constellation models

Features & classifiers

Appearance models

Key points

Appearance invariants

Alignment invariants

Appearance appearance

Alignment appearance

Generative models

Probabilistic models

Constellation models

Pictorial structures

Features & classifiers

Texton histograms

Bags of features

Texton histograms

SIFT

Video Google

1970

1980

1990

2000

1970

1980

1990

2000

Alignment

Aspect graphs

Generalized cylinders

Textures

Misc. statistics

Textons

Filter banks

Deformable templates

Object categories

Probabilistic models

Constellation models

Pictorial structures

Features & classifiers
Early 2000’s: Keypoints for everything

Bags of features

Csurka et al. (2004), Willamowski et al. (2005), Grauman & Darrell (2005), Sivic et al. (2003, 2005)
Implicit shape models

B. Leibe, A. Leonardis, and B. Schiele, *Combined Object Categorization and Segmentation with an Implicit Shape Model*, ECCV Workshop on Statistical Learning in Computer Vision, 2004
Spatial pyramids

Lazebnik, Schmid & Ponce (2006)
Caltech-101 dataset

http://www.vision.caltech.edu/Image_Datasets/Caltech101/

L. Fei-Fei, R. Fergus, and P. Perona, Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories, CVPR 2004 Workshop on Generative-Model Based Vision
Detection: Histograms of oriented gradients

N. Dalal and B. Triggs, Histograms of Oriented Gradients for Human Detection, CVPR 2005
PASCAL VOC dataset

• PASCAL VOC Challenge (2005-2012):
  http://pascallin.ecs.soton.ac.uk/challenges/VOC/

Challenge classes:

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

Dataset size (by 2012): 11.5K training/validation images, 27K bounding boxes, 7K segmentations
Discriminative deformable part models

Root filter  Part filters  Deformation weights

P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan, Object Detection with Discriminatively Trained Part-Based Models, PAMI 2009
Discriminative deformable part models

P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan, Object Detection with Discriminatively Trained Part-Based Models, PAMI 2009
Progress on PASCAL detection

Before CNNs

After CNNs
Developments after 2010: Selective search for detection

ROI features: color SIFT, codebook of size 4K, spatial pyramid with four levels = 360K dimensions

J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders, Selective Search for Object Recognition, IJCV 2013
Developments after 2010: COCO dataset

What is COCO?

COCO is a large-scale object detection, segmentation, and captioning dataset. COCO has several features:

- Object segmentation
- Recognition in context
- Superpixel stuff segmentation
- 330K images (>200K labeled)
- 1.5 million object instances
- 80 object categories
- 91 stuff categories
- 5 captions per image
- 250,000 people with keypoints

http://cocodataset.org/#home
Developments after 2010: Instance-level detection and segmentation

K. He, G. Gkioxari, P. Dollar, and R. Girshick, Mask R-CNN, ICCV 2017 (Best Paper Award)
Category-level recognition: Discussion

• Summary
  • In deep neural networks, deformable templates have converged with features and classifiers, winning the day
  • We now have clean end-to-end architectures for simultaneously handling instance detection, classification, and segmentation

• What went wrong?
  • We are saturating ImageNet and COCO
  • We are no closer to understanding what a “category” is – existing label taxonomies are as unsatisfying as ever
  • Existing datasets and models assume that every category is common and known
  • We’re still not sure whether our category models need parts or explicit 3D information (more on this tomorrow)
Category-level recognition: What’s next?

• 3D representations

A. Kanazawa, S. Tulsiani, A. Efros, J. Malik, Learning Category-Specific Mesh Reconstruction from Image Collections, ECCV 2018
History of Recognition: Outline

Object instance recognition

Object category recognition

Texture recognition

Scene recognition

MISSION: ACCOMPLISHED
Early scene understanding

- Approach has everything: color, super-pixels, bottom-up segmentation, top-down parsing, inter- and intra-region reasoning, Bayesian formulation!

Y. Yakimovskv and J. Feldman, A semantics-based decision theory region analyzer, IJCAI 1973
Early scene understanding

A. Hanson and E. Riseman, VISIONS: A computer system for interpreting scenes, Computer Vision Systems, 1978
Early scene understanding

What went wrong?

Slide credit:
A. Efros

Ohta & Kanade (1978)
What went wrong?

Slide credit: A. Efros
Structured scene representations revisited

Appearance-based scene representation: GIST

Appearance-based scene representation: GIST

• Matching of scenes based on GIST works surprisingly well – *given a large enough dataset*

Semantic segmentation

MSRC Dataset (2006)
TextonBoost

\[
\log P(c|x, \theta) = \sum_i \psi_i(c_i, x; \theta_{\psi}) + \pi(c_i, x_i; \theta_{\pi}) + \lambda(c_i, i; \theta_{\lambda}) + \sum_{(i,j) \in \mathcal{E}} \phi(c_i, c_j, g_{ij}(x); \theta_{\phi})
\]

J. Shotton, J. Winn, C. Rother, and A. Criminisi,
*TextonBoost: Joint Appearance, Shape And Context Modeling For Multi-class Object Recognition And Segmentation*, ECCV 2006
Scene flow

C. Liu, J. Yuen, and A. Torralba, *Nonparametric Scene Parsing via Label Transfer*, PAMI 2011
Semantic segmentation with superpixels

Superpixel features

<table>
<thead>
<tr>
<th>Feature Type</th>
<th>Description</th>
<th>Dimensions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shape</td>
<td>Mask of superpixel shape over its bounding box (8 x 8)</td>
<td>64</td>
</tr>
<tr>
<td></td>
<td>Bounding box width/height relative to image width/height</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Superpixel area relative to the area of the image</td>
<td>1</td>
</tr>
<tr>
<td>Location</td>
<td>Mask of superpixel shape over the image</td>
<td>64</td>
</tr>
<tr>
<td></td>
<td>Top height of bounding box relative to image height</td>
<td>1</td>
</tr>
<tr>
<td>Texture/SIFT</td>
<td>Texton histogram, dilated texton histogram</td>
<td>100 x 2</td>
</tr>
<tr>
<td></td>
<td>SIFT histogram, dilated SIFT histogram</td>
<td>100 x 2</td>
</tr>
<tr>
<td></td>
<td>Left/right/top/bottom boundary SIFT histogram</td>
<td>100 x 4</td>
</tr>
<tr>
<td>Color</td>
<td>RGB color mean and std. dev.</td>
<td>3 x 2</td>
</tr>
<tr>
<td></td>
<td>Color histogram (RGB, 11 bins per channel), dilated hist.</td>
<td>33 x 2</td>
</tr>
<tr>
<td>Appearance</td>
<td>Color thumbnail (8 x 8)</td>
<td>192</td>
</tr>
<tr>
<td></td>
<td>Masked color thumbnail</td>
<td>192</td>
</tr>
<tr>
<td></td>
<td>Grayscale gist over superpixel bounding box</td>
<td>320</td>
</tr>
</tbody>
</table>

[After Malisiewicz and Efros, 2008]

Semantic segmentation with superpixels and detectors

J. Tighe and S. Lazebnik, Finding Things: Image Parsing with Regions and Per-Exemplar Detectors, CVPR 2013
Semantic segmentation today: Mask R-CNN

K. He, G. Gkioxari, P. Dollar, and R. Girshick, Mask R-CNN, ICCV 2017 (Best Paper Award)
Scene understanding: What’s next?

- We have challenging standardized benchmarks, though we’re beginning to saturate current datasets (COCO)
- Need to push more on 3D and situated scene understanding


M.-F. Chang et al., *Argoverse: 3D Tracking and Forecasting with Rich Maps*, CVPR 2019
History of Recognition: Summary

• **Instance recognition**
  • 3D geometric
    • Blocks world
    • Alignment
    • Aspect graphs
  • Invariants
  • Primitive-based approaches: generalized cylinders, geons

• **Appearance-based**
  • Eigenfaces
  • Appearance manifolds
  • Color indexing

• **Keypoint-based**
  • Schmid & Mohr
  • Lowe
  • Video Google

• **Texture recognition**
  • Textons
  • Filter banks
  • Histograms

• **Object category recognition**
  • Pictorial structures
  • Early statistical: Late 90s face detectors
  • Constellation models
  • Bags of features
  • HOG, DPMs

• **Scene recognition**
  • Early knowledge-based
  • GIST
  • Bags of features, pyramids
CNN vs. hand-crafted pipelines
Convolution as feature extraction
Normalization and spatial pooling

Image pixels

Apply oriented filters

Take max filter response (L-inf normalization)

Spatial pool (Sum), L2 normalization

SIFT features

Filter with Visual Words

Take max VW response (L-inf normalization)

Multi-scale spatial pool (Sum)

Image descriptors

Source: R. Fergus
Deformable part models as CNNs

Deformable part models as CNNs

R. Girshick, F. Iandola, T. Darrell, and J. Malik, Deformable Part Models are Convolutional Neural Networks, CVPR 2015
In general, what were the “big ideas”?

• Optimization – not procedural reasoning
• Rich descriptors of pixel values – not simple point and line features
• Data and learning – not rules
• Discriminative classifiers – not just probabilistic models
• Sliding window operations and spatial pooling
• Deformable templates
Where did we go wrong?

• Computer vision had several periods of “spinning its wheels”
  • We’ve always prioritized methods that could already do interesting things over potentially more promising methods that could not yet deliver
  • We’ve undervalued simple methods, data, and learning
  • When nothing worked, we distracted ourselves with fancy math
  • We’ve had some problems with bandwagon jumping and intellectual snobbery
  • On a few occasions, we unaccountably ignored methods that later proved to be “game changers” (RANSAC, SIFT)

• But it’s not clear whether any of it mattered in the end...
Lana’s MVP’s (Most Valuable Papers)

2. Pictorial structures – Fischler & Elschlager, 1973
3. RANSAC – Fischler & Bolles, 1981
4. Edge detection – Canny, 1986
5. Corner detection – Harris & Stephens, 1988
7. Graph cuts – Boykov et al., 2001
9. SIFT – Lowe, 2004
10. Deformable part models – Felzenszwalb et al., 2010
What did my historical survey omit?

• Image filtering
  • Wavelets, steerable filters, bilateral filtering...
  • Biologically inspired low-level representations (Olhausen & Field, etc.)

• Generation methods
  • E.g., texture generation (Heeger & Bergen, Efros & Leung, etc.)

• Image-based modeling and computational photography
• Dense reconstruction, stereo
• Graphical models craze
• Perceptual organization