Solved, Half-Solved and Unsolved Problems in Visual Recognition

Jitendra Malik
University of California at Berkeley

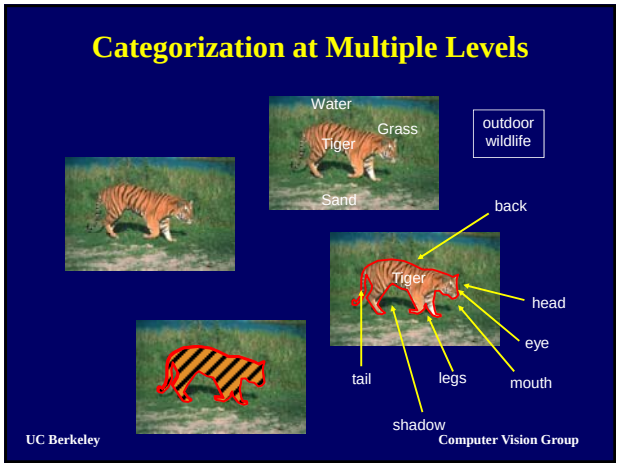
The more you look, the more you see!

Image shown to subjects	40ms	80ms	107ms	500ms
	"Possibly outdoor scene, maybe a farm. I could not tell for sure."	"There seem to be two people in the center of the scene."	" People playing rugby. Two persons in close contact, wrestling, on grass. Another man more distant. Goal in sight."	"Some kind of game or fight. Two groups of two men. One in the foreground was getting a fist in the face. Outdoors, because I see grass and maybe lines on the grass? That is why I think of a game, rough game though, more like rugby than football because they weren't in pads and helmets..."

Figure 2. Human subjects reporting on what he/she saw in an image shown for different presentation durations (PD=27, 40, 67, 80, 107, 500ms). From Fei-Fei and Perona [26].

PASCAL Visual Object Challenge

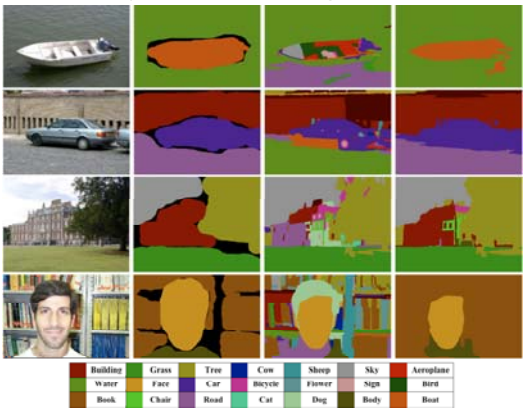
Categorization at Multiple Levels



Actually, we should want more…

Orig. Image Segmentation Orig. Image Segmentation

Complete Semantic Segmentation



The more you look, the more you see!

Image shown to subjects	40ms	80ms	107ms	500ms
	"Possibly outdoor scene, maybe a farm. I could not tell for sure."	"There seem to be two people in the center of the scene."	" People playing rugby. Two persons in close contact, wrestling, on grass. Another man more distant. Goal in sight."	"Some kind of game or fight. Two groups of two men. One in the foreground was getting a fist in the face. Outdoors, because I see grass and maybe lines on the grass? That is why I think of a game, rough game though, more like rugby than football because they weren't in pads and helmets..."

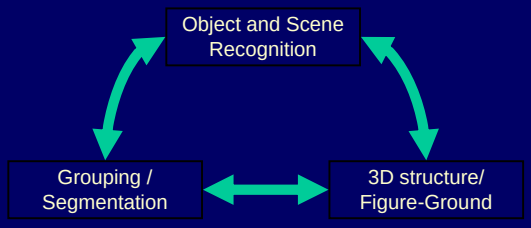
Figure 2. Human subjects reporting on what he/she saw in an image shown for different presentation durations (PD=27, 40, 67, 80, 107, 500ms). From Fei-Fei and Perona [26].

We need to identify

- Objects
- Agents
- Relationships among objects with objects, objects with agents, agents with agents …
- Events and Actions

University of California
Berkeley **Computer Vision Group**

The central problems of vision



UC Berkeley **Computer Vision Group**

A brief history of computer vision ..

Those who cannot remember the past are condemned to repeat it
-George Santayana

10

Fifty years of computer vision 1963-2013

- 1960s: Beginnings in artificial intelligence, image processing and pattern recognition
- 1970s: Foundational work on image formation: Horn, Koenderink, Longuet-Higgins …
- 1980s: Vision as applied mathematics: geometry, multi-scale analysis, probabilistic modeling, control theory, optimization
- 1990s: Geometric analysis largely completed, vision meets graphics, statistical learning approaches resurface
- 2000s: Significant advances in visual recognition, range of practical applications

UC Berkeley **Computer Vision Group**

Object recognition in computer vision

- Recognition as Pose Estimation
- Recognition as Description using Volumetric primitives
- Recognition as Pattern Classification
- Recognition as Deformable Matching

University of California
Berkeley **Computer Vision Group**

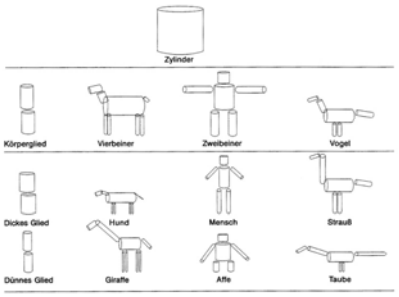
Recognition as Pose Estimation: Object as a set of points in 3D

- Roberts (1963) , Faugeras & Hebert (1983), Huttenlocher & Ullman (1987)
- Variants
  - Geometric Hashing : Lamdan & Wolfson (1988)
  - Pose Clustering : Stockman (1987), Olson (1994)
  - Linear Combination of Views: Basri & Ullman (1991)

Recognition as Fitting Volumetric Primitives: Object as a hierarchy of simple shapes

- Binford (1971) , Marr & Nishihara (1978), Biederman(1987)
- Discredited as an approach for recognition in general, it has retained appeal for analyzing images of people

The Stick Figure Ideal



Recognition as Statistical Pattern Classification: Object as a feature vector

- Optical Character Recognition studied as far back as the 1950s. Recent years focus on handwritten digit classification and face detection.
- Some examples:
  - Neural networks: Neocognitron (Fukushima, 1980, 1988) , Convolution Neural Networks (LeCun et al), C2 Features (Serre, Wolf & Poggio 2005)
  - Support Vector Machines (various)
  - Decision Trees (Amit, Geman, & Wilder, 1997)
  - Boosted Decision Trees (Viola & Jones, 2001)

Recognition as Pictorial Structure Matching: Object as a configuration of feature points

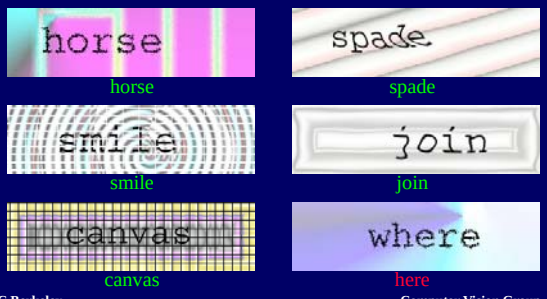
- Transformations to model shape variation-D'Arcy Wentworth Thompson (1910)
- Grenander (1970s and later)probabilistic models ontransformations
- Fischler and Elschlager (1973) - deformable matching of landmarks ,"point masses", in a configuration of "springs" to model deformable templates.
- Von derMalsburg-dynamic link architecture for neural modelling, elastic graph matching for face recognition (1993, 1997)
- Felzenszwalb and Huttenlocher (2000) - pictorial structures for aligning human bodies to stick figures using dynamic programming
- Belongie, Malik &Puzicha (2001) use"shape contexts" as point descriptors, and thin plate splines to model deformation.

Handwritten digit recognition (MNIST,USPS)

- LeCun's Convolutional Neural Networks variations (0.8%, 0.6% and 0.4% depending on different ways of virtually augmenting dataset)
- SVMs (DeCoste & Scholkopf : 0.6%)
- K-NN based Shape context/TPS matching (Belongie, Malik & Puzicha: 0.6%)
- On USPS comparison to humans: 2.5% (Bromley and Sackinger, 1991), cf. Zhang et al based on Simard's tangent distance; 2.59%

EZ-Gimpy Results (Mori & Malik, 2003)

- 171 of 192 images correctly identified: 92 %



Face Detection

Carnegie Mellon University

Multiscale sliding window

Paradigm introduced by Rowley, Baluja & Kanade 96 for face detection
Viola & Jones 01, Dalal & Triggs 05, Felzenszwalb, McAllester, Ramanan 08

Problems with the multi-scale scanning paradigm

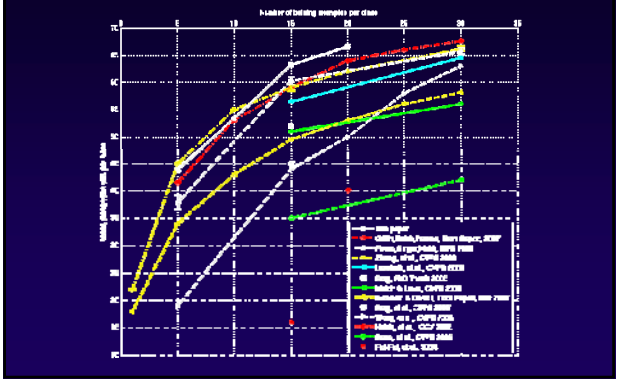
- Computational complexity
  - 10^6 windows, 10 scales, 10^4 categories
- Not natural for irregularly shaped objects
- Segmentation is delinked
- Context is delinked

Caltech-101 [Fei-Fei et al. 04]

- 102 classes, 31-300 images/class

Caltech 101 classification results

(even better by combining cues..)



Current Works on Caltech-101 low-level features

Image →

- SIFT (Lazebnik&Schmid&Ponce, Grauman&Darrell, Wang&Zhang&Feifei)
- "S1" features (Serre&Wolf&Poggio, Mutch&Lowe)
- Geometric Blur (Berg&Berg&Malik,Zhang&Berg&Maire&Malik,Frome&Singer&Malik)
- Other histogram of local edges (Ommer&Buhmann)

University of California **Berkeley** **Computer Vision Group**

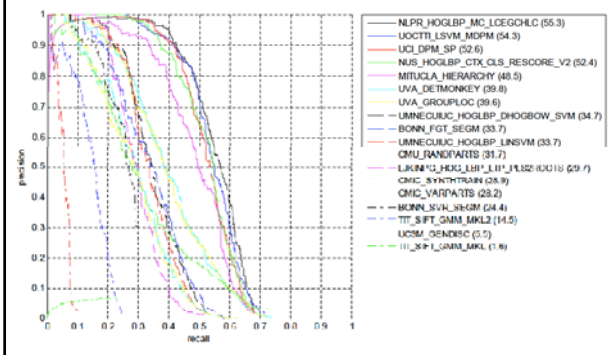
The PASCAL Visual Object Classes Challenge 2010 (VOC2010)

Part 2 – Detection Task

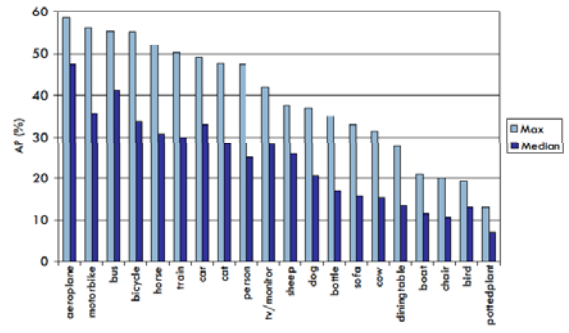
Mark Everingham
Luc Van Gool
Chris Williams
John Winn
Andrew Zisserman

PASCAL Visual Object Challenge

Precision/Recall - Bicycle

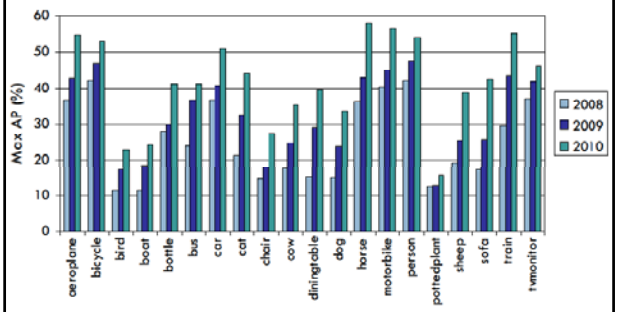


AP by Class



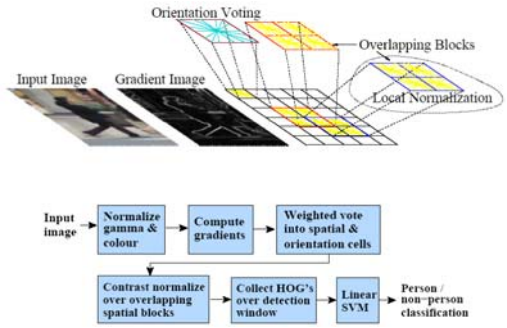
- Max AP: 58.4% (aeroplane) ... 13.0% (potted plant)

Progress 2008-2010



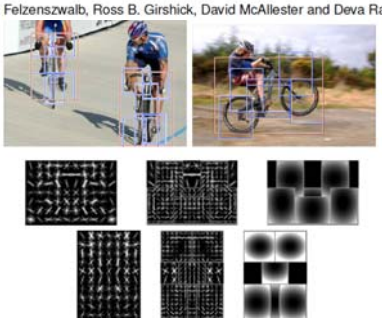
- Results on 2008 data improve for best 2009 and 2010 methods for all categories, by over 100% for some categories
  - Caveat: Better methods or more training data?

A good building block is a linear SVM trained on HOG features (Dalal&Triggs)



Object Detection with Discriminatively Trained Part Based Models

Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester and Deva Ramanan



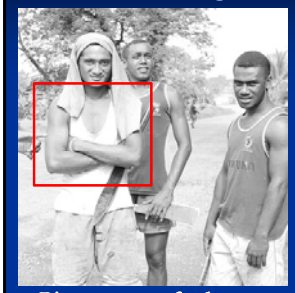
Examples of poselets

Patches are often far **visually**, but they are close **semantically**

(Bourdev& Malik, 09; Bourdev et al, 10)

How do we train a poselet for a given pose configuration?

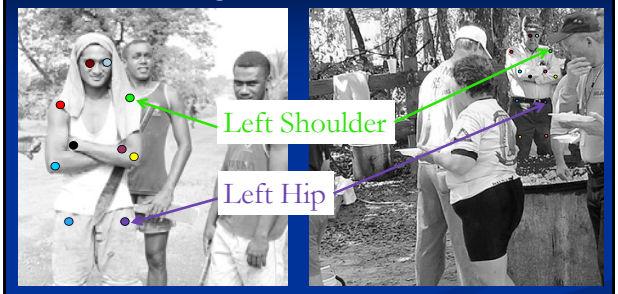
Finding Correspondences



Given part of a human pose

How do we find a similar pose configuration in the training set?

Finding Correspondences



We use keypoints to annotate the joints, eyes, nose, etc. of people

Finding Correspondences

Residual Error

Training poselet classifiers

Residual Error: 0.15 0.20 0.10 0.85 0.15 0.35

1. Given a seed patch
2. Find the closest patch for every other person
3. Sort them by residual error
4. Threshold them

Training poselet classifiers

1. Given a seed patch
2. Find the closest patch for every other person
3. Sort them by residual error
4. Threshold them
5. Use them as positive training examples for a classifier (HOG features, linear SVM)

How do we find poselets?

- Choose thousands of random windows, generate poselet candidates, train linear SVMs

- Select a small set of poselets that are:
  - Individually effective
  - Complementary

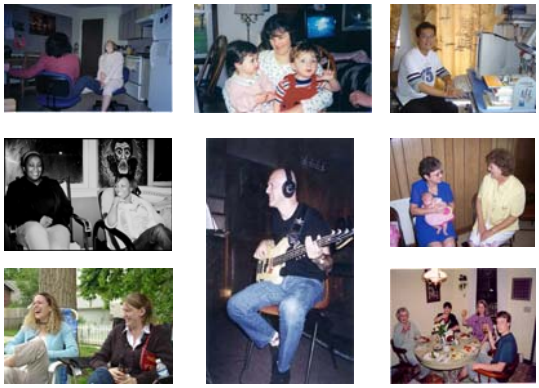
Segmenting people (Brox et al, CVPR 2011)

Actions in still images …

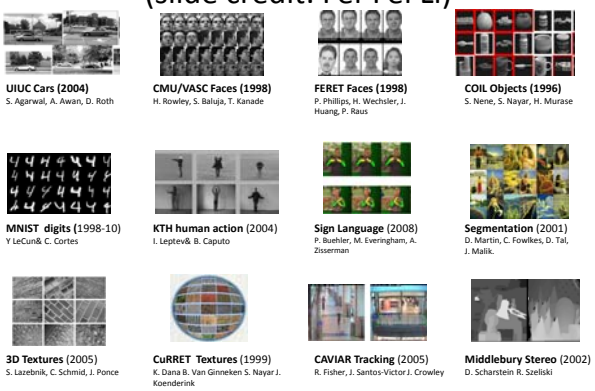
- have characteristic :
  - pose and appearance
  - interaction with objects and agents

Some discriminative poselets

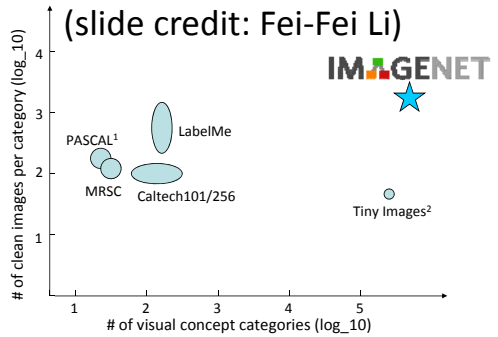
AP=0.16



Datasets and computer vision (slide credit: Fei-Fei Li)



Comparison among freedatasets (slide credit: Fei-Fei Li)



1. Excluding the Caltech101 datasets from PASCAL
2. No image in this dataset is human annotated. The # of clean images per category is a rough estimation

SUN Database: Large-scale Scene Recognition from Abbey to Zoo

Jianxiong Xiao James Hays[†] Krista A. Ehinger Aude Oliva Antonio Torralba

Figure 7. Selected SUN scene classification results using all features.

Examples of Actions

- Movement and posture change
  - run, walk, crawl, jump, hop, swim, skate, sit, stand, kneel, lie, dance (various), …
- Object manipulation
  - pick, carry, hold, lift, throw, catch, push, pull, write, type, touch, hit, press, stroke, shake, stir, turn, eat, drink, cut, stab, kick, point, drive, bike, insert, extract, juggle, play musical instrument (various)…
- Conversational gesture
  - point, …
- Sign Language

Key cues for action recognition

- "Morpho-kinetics" of action (shape and movement of the body)
- Identity of the object/s
- Activity context

- ACTION = MOVEMENT + GOAL

Recognition

Far field

- 3-pixel man
- Blob tracking

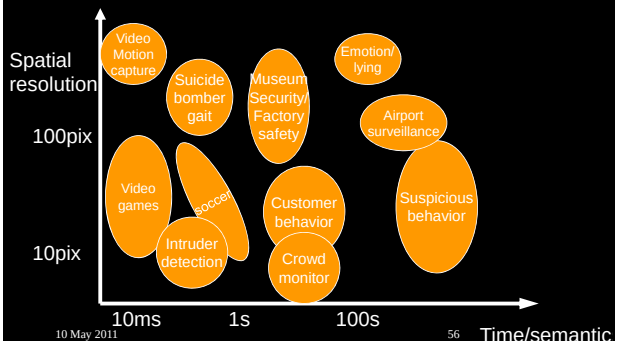
Near field

- 300-pixel man
- Limb shape

Medium-field Recognition

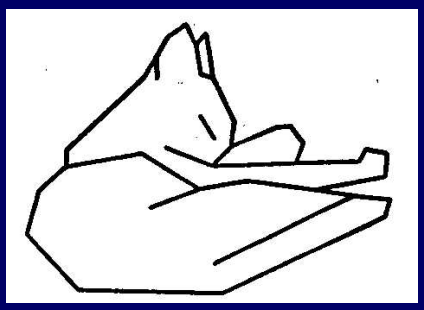
10 May 2011 The 30-Pixel Man 55

Taxonomy



10 May 2011 56

Attneave's Cat (1954)
Line drawings convey most of the information



UC Berkeley Computer Vision Group

The more you look, the more you see!

Image shown to subjects	40ms	80ms	107ms	500ms
	"Possibly outdoor scene, maybe a farm. I could not tell for sure."	"There seem to be two people in the center of the scene."	" People playing rugby. Two persons in close contact, wrestling, on grass. Another man more distant. Goal in sight."	"Some kind of game or fight. Two groups of two men. One in the foreground was getting a fist in the face. Outdoors, because I see grass and maybe lines on the grass? That is why I think of a game, rough game though, more like rugby than football because they weren't in pads and helmets..."

Figure 2. Human subjects reporting on what he/she saw in an image shown for different presentation durations (PD=27, 40, 67, 80, 107, 500ms). From Fei-Fei and Perona [26].

So much remains to be done…

- Objects, Scenes, Events
- The semantic gap is to be confronted, not avoided!

UC Berkeley Computer Vision Group