Object Recognition

Computer Vision
CSE576, Spring 2008
Richard Szeliski

Recognition problems

What is it?

- Object and scene recognition

Who is it?

- Identity recognition

Where is it?

- Object detection

What are they doing?

- Activities

All of these are **classification** problems

- Choose one class from a list of possible candidates

What is recognition?

A different taxonomy from [Csurka *et al.* 2006]:

- Recognition
  - Where is *this* particular object?
- Categorization
  - What *kind* of object(s) is(are) present?
- Content-based image retrieval
  - Find me something that looks similar
- Detection
  - Locate *all* instances of a given class

Readings

- **Weakly Supervised Scale-Invariant Learning of Models for Visual Recognition**
  Fergus, R. , Perona, P. and Zisserman, A.
  International Journal of Computer Vision, Vol. 71(3), 273-303, March 2007

Sources

- Steve Seitz, CSE 455/576, previous quarters
- Fei-Fei, Fergus, Torralba, CVPR'2007 course
- Efros, CMU 16-721 Learning in Vision
- Freeman, MIT 6.869 Computer Vision: Learning
- Linda Shapiro, CSE 576, Spring 2007

Today's lecture

- Known object recognition [Lowe]
- Bag of keypoints [Csurka *etc.*]
- Location recognition [Schindler *et al.*]
- Deformable object/category recognition [Fergus *et al.*]
- Recognition by segmentation

Recognizing and Learning Object Categories: Year 2007

Li Fei-Fei, Princeton
Rob Fergus, MIT
Antonio Torralba, MIT

(see other slide deck)

Today's lecture

- Known object recognition [Lowe]
- Bag of keypoints [Csurka *etc.*]
- Location recognition [Schindler *et al.*]
- Deformable object/category recognition [Fergus *et al.*]
- Recognition by segmentation

Single object recognition



Single object recognition

- Lowe, et al. 1999, 2003
- Mahamud and Herbert, 2000
- Ferrari, Tuytelaars, and Van Gool, 2004
- Rothganger, Lazebnik, and Ponce, 2004
- Moreels and Perona, 2005
- …

Planar object recognition [Lowe]

- Use SIFT features
- Verify affine (or homography) geometric alignment

Planar object recognition [Lowe]

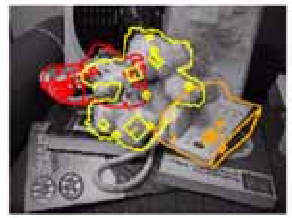
- Use SIFT features
- Verify affine (or homography) geometric alignment


3D object recognition [Lowe]

- Extract object outlines with background subtraction

3D object recognition [Lowe]

- Use 3 matches to recognize
- Use additional matches for verification
- Tolerant to occlusions



Feature-based recognition

How can we scale to millions of objects?

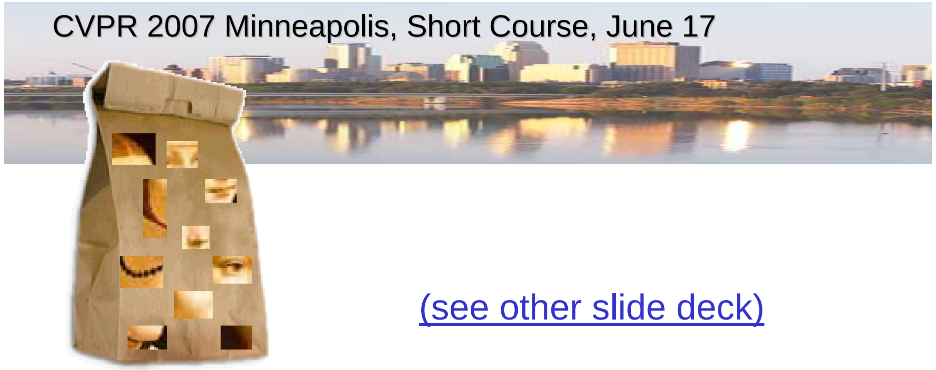
Comparison to *all* stored objects/features is infeasible.

Answer:

- quantize features into *words* [Csurka *et al.* 04]
- use information retrieval (inverted index)
- use *metric tree* for faster quantization (NN) [Nister & Stewenius 05]

Today's lecture

- Known object recognition [Lowe]
- Bag of keypoints [Csurka *etc.*]
- Location recognition [Schindler *et al.*]
- Deformable object/category recognition [Fergus *et al.*]
- Recognition by segmentation



(see other slide deck)

Part 1: Bag-of-words models

by Li Fei-Fei (Princeton)

Today's lecture

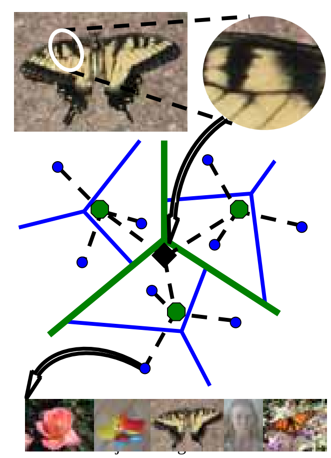
- Known object recognition [Lowe]
- Bag of keypoints [Csurka *etc.*]
- Location recognition [Schindler *et al.*]
- Deformable object/category recognition [Fergus *et al.*]
- Recognition by segmentation

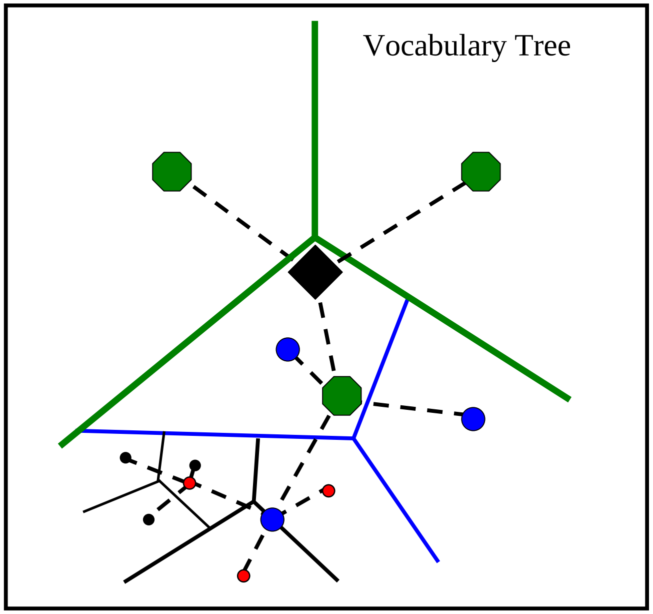
How to scale to 10^6s of images?

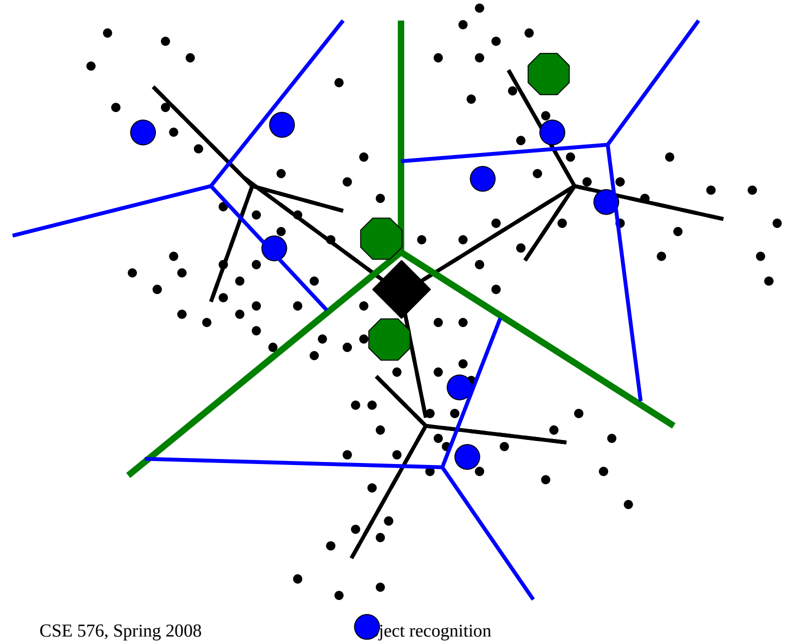
Make "word" generation even more efficient:
  "Vocabulary tree"

Scalable Recognition with a Vocabulary Tree

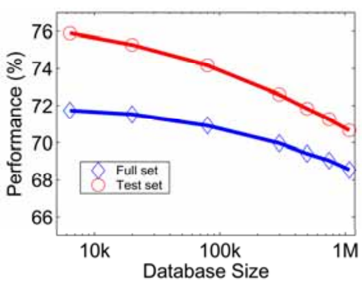
David Nistér, Henrik Stewénius

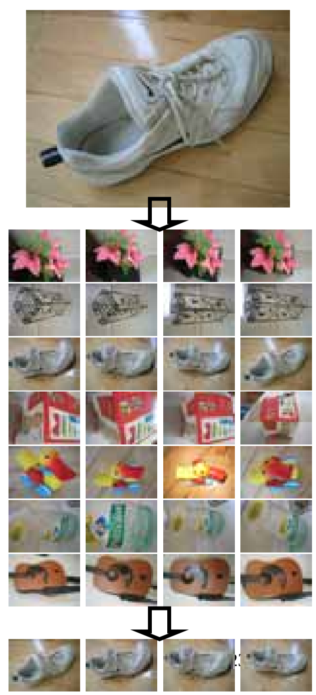




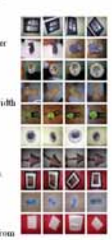


Performance





http://vis.uky.edu/~stewe/ukbench/



Recognition Benchmark Images

Henrik Stewénius and David Nistér

The set consists of 2604 groups of 4 images each for a total of 10416 images. All the images are 640x480.

If you use the dataset, please refer to:

- D. Nistér and H. Stewénius, Scalable Recognition with a Vocabulary Tree, CVPR 2006. PDF

Subsets

For users of subsets of the database please note that the difficulty is dependent on the chosen subset. Important factors are:

1. Difficulty of the objects themselves. CD-covers are much easier than flowers. See performance curve below.
2. Sharpness of the images. Many of the indoor images are somewhat blurry and this can affect some algorithms.
3. Similar or identical objects. All the pictures where taken by CS students/faculty/staff and thus keyboards and computer equipment are popular motives. So is computer vision literature.

Download

Please note BEFORE starting your download that the file is almost 2GB. Please save a local copy in order to save bandwidth at our server.

- Zipped File.

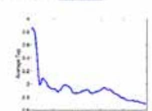
Performance

In the paper we give results either for a subset of 6376 images (all we had at that time) or a smaller subset of 1400 images. The smaller set was used when we did not have an efficient enough implementation in order to handle the larger set.

Performance Measures

- Our simplest measure of performance is to count how many of the 4 images which are top-4 when using a query image from that set of four images.

A matlab implementation which computes this measure: Download.



How our performance varies when taking subsets 0:n from the set. These results were run with settings optimized for speed.

Location Recognition

Can we apply this to recognizing your location from a cell-phone photo?

City-Scale Location Recognition

Grant Schindler, Matthew Brown, and Richard Szeliski
CVPR'2007

The Problem

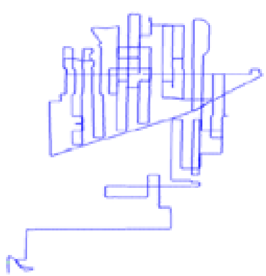
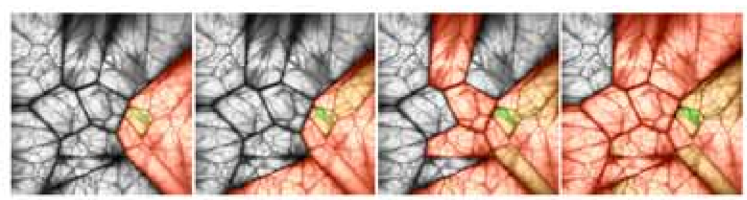


Figure 1. We perform location recognition on 20 km of urban streetside imagery, storing 100 million features in a vocabulary tree, the structure of which is determined by the features that are most informative about each location. Shown here is the path of our vehicle over 20 km of urban terrain.

Main idea

Find N-best matches in vocabulary tree



(a) N=1 (b) N=2 (c) N=5 (d) N=9

Figure 3. Greedy N-Best Paths Search. From left to right, we increase the number of nodes N whose children are considered at each level of the tree. Cells are colored from red to green according to the depth at which they are encountered in the tree, while gray cells are never searched. By considering more nodes in the tree, recognition performance is improved at a computational cost that varies with N.

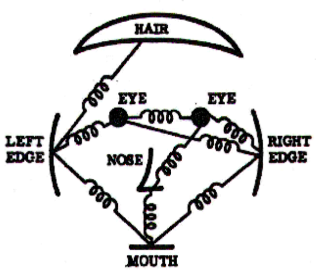
Other ideas

- Use only informative features (ignore trees…)
- Integrate matches with adjacent (streetside) neighbors

Figure 9. Typical examples of the 278 query images (left) and the corresponding top matches returned from the database (right) using a 1000^2 vocabulary tree with $N = 4$.

Today's lecture

- Known object recognition [Lowe]
- Bag of keypoints [Csurka *etc.*]
- Location recognition [Schindler *et al.*]
- Deformable object/category recognition [Fergus *et al.*]
- Recognition by segmentation



(see other slide deck)

Part 2: part-based models

by Rob Fergus (MIT)

Today's lecture

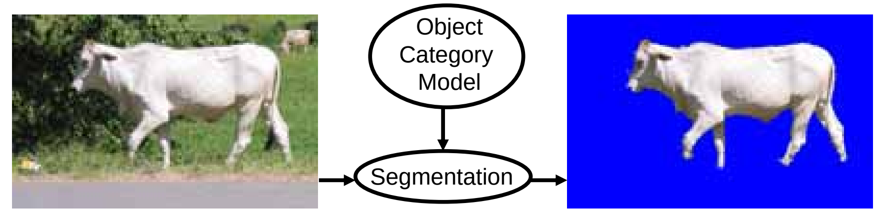
- Known object recognition [Lowe]
- Bag of keypoints [Csurka *etc.*]
- Location recognition [Schindler *et al.*]
- Deformable object/category recognition [Fergus *et al.*]
- Recognition by segmentation

Part 4: Combined segmentation and recognition

by Rob Fergus (MIT)

Aim

Given an image and object category, segment the object



Cow Image

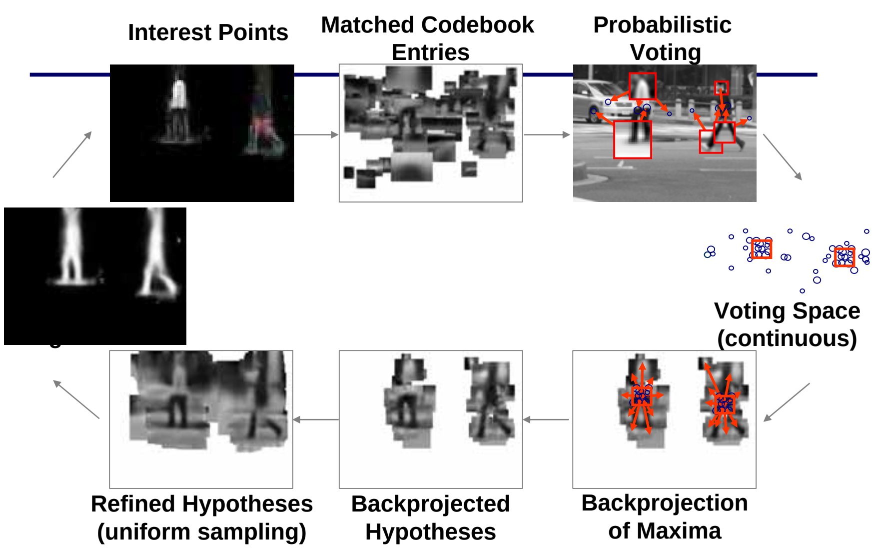
Segmented Cow

Segmentation should (ideally) be

- shaped like the object e.g. cow-like
- obtained efficiently in an unsupervised manner
- able to handle self-occlusion

Slide from Kumar '05

Implicit Shape Model - Liebe and Schiele, 2003



Other topics: context (scenes)

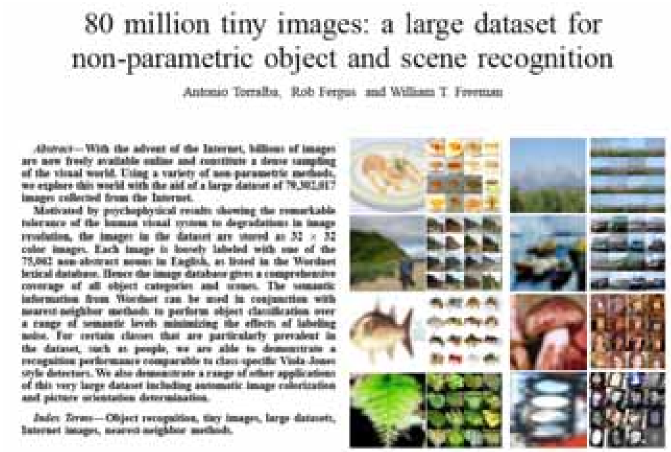

Contextual Priming for Object Detection 171

(a) (b)

Figure 2. In presence of image degradation (e.g. blur), object recognition is strongly influenced by contextual information. Recognition makes assumptions regarding object identities based on its size and location in the scene. In this picture subjects describe the scenes as (a) a car in

Antonio Torralba, **Contextual Priming for Object Detection**, *IJCV(53)*, No. 2, July 2003, pp. 169-191

New work: tiny images



80 million tiny images: a large dataset for non-parametric object and scene recognition

Antonio Torralba, Rob Fergus and William T. Freeman

Abstract—With the advent of the Internet, billions of images are now freely available online and constitute a dense sampling of the visual world. Using a variety of non-parametric methods, we explore this world with the aid of a large dataset of 79,302,017 images collected from the Internet.

Motivated by psychophysical results showing the remarkable tolerance of the human visual system to degradations in image resolution, the images in the dataset are stored as 32 × 32 color images. Each image is loosely labeled with one of the 75,062 non-abstract nouns in English, as listed in the Wordnet lexical database. Hence the image database gives a comprehensive coverage of all object categories and scenes. The semantic information from Wordnet can be used in conjunction with nearest-neighbor methods to perform object classification over a range of semantic levels minimizing the effects of labeling noise. For certain classes that are particularly prevalent in the dataset, such as people, we are able to demonstrate a recognition performance comparable to class-specific Viola-Jones style detectors. We also demonstrate a range of other applications of this very large dataset including automatic image colorization and picture orientation determination.

Index Terms—Object recognition, tiny images, large datasets, Internet images, nearest-neighbor methods.

(see other slide deck)

Datasets and object collections

Summary of object recognition

- Known object recognition [Lowe]
- Bag of keypoints [Csurka *etc.*]
- Location recognition [Schindler *et al.*]
- Deformable object/category recognition [Fergus *et al.*]
- Recognition by segmentation
- Context and scenes