**Course Outline**

- Image Processing Basics
- Segmentation & Grouping
- Object Recognition
- Object Categorization I
  - Sliding Window based Object Detection
- Local Features & Matching
  - Local Features - Detection and Description
  - Recognition with Local Features
  - Indexing & Visual Vocabularies
- Object Categorization II
  - Bag-of-Words Approaches & Part-based Approaches
- 3D Reconstruction
- Motion and Tracking

**Recap: Local Feature Matching Outline**

1. Find a set of distinctive keypoints
2. Define a region around each keypoint
3. Extract and normalize the region content
4. Compute a local descriptor from the normalized region
5. Match local descriptors

**Recap: Recognition with Local Features**

- Image content is transformed into local features that are invariant to translation, rotation, and scale
- Goal: Verify if they belong to a consistent configuration

**Recap: Object Recognition by Alignment**

- Assumption
  - Known object, rigid transformation compared to model image
  - If we can find evidence for such a transformation, we have recognized the object.
- You learned methods for
  - Fitting an affine transformation from 2-3 correspondences
  - Fitting a homography from 4 correspondences
    - Affine: solve a system \( A t = b \)
    - Homography: solve a system \( A h = 0 \)
- Correspondences may be noisy and may contain outliers
  - Use RANSAC for robust fitting

**Announcements**

- Happy new year everybody!
- Seminar registration period starts today
  - We will offer a seminar in the summer semester
  - “Current Topics in Computer Vision and Machine Learning”
  - Block seminar, presentations in 1st week of semester break
  - If you’re interested, you can register at [http://www.graphics.rwth-aachen.de/apse](http://www.graphics.rwth-aachen.de/apse)
  - Registration period: 07.01.2010 - 17.01.2010
Recap: Robust Estimation with RANSAC

RANSAC loop:
1. Randomly select a seed group of points on which to base transformation estimate (e.g., a group of matches)
2. Compute transformation from seed group
3. Find inliers to this transformation
4. If the number of inliers is sufficiently large, re-compute least-squares estimate of transformation on all of the inliers
   • Keep the transformation with the largest number of inliers

Recap: Generalized Hough Transform

• Suppose our features are scale- and rotation-invariant
  • Then a single feature match provides an alignment hypothesis (translation, scale, orientation).
  • Of course, a hypothesis from a single match is unreliable.
  • Solution: let each match vote for its hypothesis in a Hough space with very coarse bins.

Recap: Visual Words

• Quantize the feature space into “visual words”
• Perform matching only to those visual words.

Topics of This Lecture

• Part-Based Models for Object Categorization
  • Structure representations
  • Different connectivity structures
• Bag-of-Words Model
  • Representation
  • Use for image classification
• Implicit Shape Model
  • Generalized Hough Transform for object category detection
  • Example results and extensions
• Deformable Part-based Model
  • Multi-resolution models
• Top-Down Segmentation
Recognition of Object Categories

- We no longer have exact correspondences...
- On a local level, we can still detect similar parts.
- Represent objects by their parts → Bag-of-features
- How can we improve on this?
  - Encode structure

Part-Based Models

- Fischler & Elschlager 1973
- Model has two components
  - parts
  - structure (configuration of parts)

Different Connectivity Structures

- Bag of visual words
  - Courts et al. '04
  - Vezzosi et al. '06
- Constellation
  - Fergus et al. '03
  - Fel-Fel et al. '03
- Star shape
  - Leibe et al. '04, '05
  - Fergus et al. '05
- Tree
  - Felzenszwalb & Huttenlocher '04

Some Class-Specific Graphs

- Articulated motion
  - People
  - Animals
- Special parameterisations
  - Limb angles

Topics of This Lecture

- Part-Based Models for Object Categorization
  - Structure representations
  - Different connectivity structures
- Bag-of-Words Model
  - Representation
  - Use for image classification
- Implicit Shape Model
  - Generalized Hough Transform for object category detection
  - Example results and extensions
- Deformable Part-Based Model
  - Multi-resolution models
- Top-Down Segmentation

Analogy to Documents

- China is forecasting a trade surplus of $90bn (£51bn) to $100bn this year, a threefold increase on 2004's $32bn. The Commerce Ministry said the surplus would be created to export surpluses of 10% in line with domestic demand. Under normal circumstances, China's trade surplus in 2005 has made it clear that it will take its time and tread carefully before allowing the yuan to rise further in value.

- Of all the sensory impressions proceeding to the brain, the visual experiences are the dominant ones. Our perception of the world around us is based essentially on the messages we receive from our eyes. The eyes send a stream of images to the brain, which is able to recognize objects by the way they are reflected in the eye. In the brain, each cell of the visual cortex is responsible for a specific area of the visual field, or receptive field. The optic nerve transmits impulses from the retina to the brain, where they are interpreted and organized into meaningful patterns. The sensory cortex, which is responsible for processing these patterns, is located in the occipital lobe of the cerebral cortex.
Bags of Visual Words

- Summarize entire image based on its distribution (histogram) of word occurrences.
- Analogous to bag of words representation commonly used for documents.

Similarly, Bags-of-Textons for Texture Repr.

Comparing Bags of Words

- We build up histograms of word activations, so any histogram comparison measure can be used here.
- E.g. we can rank frames by normalized scalar product between their (possibly weighted) occurrence counts
  - Nearest neighbor search for similar images.

Learning/Recognition with BoW Histograms

- Bag of words representation makes it possible to describe the unordered point set with a single vector (of fixed dimension across image examples)
  - Provides easy way to use distribution of feature types with various learning algorithms requiring vector input.
Recap: Categorization with Bags-of-Words

- Compute the word activation histogram for each image.
- Let each such BoW histogram be a feature vector.
- Use images from each class to train a classifier (e.g., an SVM).

BoW for Object Categorization

- Works pretty well for image-level classification

Csurka et al. (2004), Wilmanski et al. (2005), Grauman & Darrell (2005), Sivic et al. (2003, 2005)

BoW for Object Categorization

- Good performance for pure classification (object present/absent)
  - Better than more elaborate part-based models with spatial constraints.
  - What could be possible reasons why?

Limitations of BoW Representations

- The bag of words removes spatial layout.
- This is both a strength and a weakness.
  - Why a strength?
  - Why a weakness?

Spatial Pyramid Representation

- Representation in-between orderless BoW and global appearance
Spatial Pyramid Representation

- Representation in-between orderless BoW and global appearance

Summary: Bag-of-Words

- Pros:
  - Flexible to geometry / deformations / viewpoint
  - Compact summary of image content
  - Provides vector representation for sets
  - Empirically good recognition results in practice

- Cons:
  - Basic model ignores geometry - must verify afterwards, or encode via features.
  - Background and foreground mixed when bag covers whole image
  - Interest points or sampling: no guarantee to capture object-level parts.
  - Optimal vocabulary formation remains unclear.

Implicit Shape Model (ISM)

- Basic ideas
  - Learn an appearance codebook
  - Learn a star-topology structural model
  - Features are considered independent given obj. center

- Algorithm: probabilistic Gen. Hough Transform
  - Exact correspondences $\Rightarrow$ Prob. match to object part
  - NN matching $\Rightarrow$ Soft matching
  - Feature location on obj. $\Rightarrow$ Part location distribution
  - Uniform votes $\Rightarrow$ Probabilistic vote weighting
  - Quantized Hough array $\Rightarrow$ Continuous Hough space

Topics of This Lecture

- Part-Based Models for Object Categorization
  - Structure representations
  - Different connectivity structures
- Bag-of-Words Model
  - Representation
  - Use for image classification
- Implicit Shape Model
  - Generalized Hough Transform for object category detection
  - Example results and extensions
  - Deformable Part-based Model
  - Multi-resolution models
  - Top-Down Segmentation
Implicit Shape Model: Basic Idea

- Objects are detected as consistent configurations of the observed parts (visual words).

Example: Results on Cows

Test image

Example: Results on Cows

- Matched patches

Example: Results on Cows

- Prob. Votes

Example: Results on Cows

- 1st hypothesis

Example: Results on Cows

- 2nd hypothesis

Example: Results on Cows

- 3rd hypothesis

Scale Invariant Voting

- Scale-invariant feature selection
  - Scale-invariant interest points
  - Rescale extracted patches
  - Match to constant-size codebook

- Generate scale votes
  - Scale as 3rd dimension in voting space
    - $x_{vote} = x_{img} - x_{ref} (s_{img}/s_{ref})$
    - $y_{vote} = y_{img} - y_{ref} (s_{img}/s_{ref})$
    - $s_{vote} = (s_{img}/s_{ref})$
  - Search for maxima in 3D voting space

- Search window
Continuous Generalized Hough Transform

- Binned accumulator array similar to standard Gen. Hough Transf.
- Quickly identify candidate maxima locations
- Refine locations by Mean-Shift search only around those points
- Avoid quantization effects by keeping exact vote locations.

Mean-shift interpretation as kernel prob. density estimation.

Scale Voting: Efficient Computation

\[ \hat{y}(u_i, x_j) = \frac{1}{V_b} \sum_k \sum_j \hat{y}(u_k, x_j) \cdot f_k(x_j, \hat{u}_i) \cdot K\left( \frac{x - x_j}{b} \right) \]

Detection Results

- Qualitative Performance
  - Recognizes different kinds of objects
  - Robust to clutter, occlusion, noise, low contrast

Detections Using Ground Plane Constraints

Battery of 5 ISM detectors for different car views

Extension: Rotation-Invariant Detection

- Polar instead of Cartesian voting scheme
- Benefits:
  - Recognize objects under image-plane rotations
  - Possibility to share parts between articulations.
- Caveats:
  - Rotation invariance should only be used when it’s really needed.
  - (Also increases false positive detections)

Sometimes, Rotation Invariance Is Needed...
Topics of This Lecture

- Part-Based Models for Object Categorization
  - Structure representations
  - Different connectivity structures
- Bag-of-Words Model
  - Representation
  - Use for image classification
- Implicit Shape Model
  - Generalized Hough Transform for object category detection
  - Example results and extensions
- Deformable Part-based Model
  - Multi-resolution models
- Top-Down Segmentation

Starting Point: HOG Sliding-Window Detector

Filter $F$

Score of $F$
at position $p$ is

$F \cdot \phi(p, H)$

$\phi(p, H) = \text{concatenation of HOG features from window specified by } p$

- Array of weights for features in window of HOG pyramid
- Score is dot product of filter and vector

Deformable Part-based Models

- Mixture of deformable part models (pictorial structures)
- Each component has global template + deformable parts
- Fully trained from bounding boxes alone

2-Component Bicycle Model

- Root filters coarse resolution
- Part filters finer resolution
- Deformation models

Object Hypothesis

- Multiscale model captures features at two resolutions

Score of a Hypothesis

$\text{score}(p_1, \ldots, p_r) = \sum_{i=1}^{r} F_i \cdot \phi(H, p_i)$

$\text{score}(z) = \beta \cdot \Psi(H, z)$

concatenation filters and deformation parameters

concatenation of HOG features and part displacement features
Recognition Model

\[ f_w(x) = w \cdot \Phi(x) \]

\[ f_w(x) = \max_z w \cdot \Phi(x,z) \]

- \( z \): vector of part offsets
- \( \Phi(x,z) \): vector of HOG features (from root filter & appropriate part sub-windows) and part offsets

Results: Persons

- Results (after non-maximum suppression)
  - \(-1s\) to search all scales

Results: Bicycles

Slide adapted from Trevor Darrell

False Positives

- Bicycles

Results: Cats

High-scoring true positives

High-scoring false positives
(not enough overlap)

You Can Try It At Home...

- Deformable part-based models have been very successful at several recent evaluations.

\[
\Rightarrow \quad \text{Currently, state-of-the-art approach in object detection}
\]

- Source code and models trained on PASCAL 2006, 2007, and 2008 data are available here:
  - [http://www.cs.uchicago.edu/~pff/latent](http://www.cs.uchicago.edu/~pff/latent)
Topics of This Lecture

- Part-Based Models for Object Categorization
  - Structure representations
  - Different connectivity structures
- Bag-of-Words Model
  - Representation
  - Use for image classification
- Implicit Shape Model
  - Generalized Hough Transform for object category detection
  - Example results and extensions
- Deformable Part-Based Model
  - Multi-resolution models
- Top-Down Segmentation

Figure-Ground Segregation

- What happens first - segmentation or recognition?
- Problem extensively studied in Psychophysics
- Experiments with ambiguous figure-ground stimuli
- Results:
  - Evidence that object recognition can and does operate before figure-ground organization
  - Interpreted as Gestalt cue familiarity.


Top-Down Segmentation: Basic Idea

- During initial voting
  - When we first observe a feature, we do not know its context.
  - Different figure-ground labels may be consistent with the appearance.
  - ⇒ Strategy: we cast votes for many locations...
- After voting
  - Voting groups features that are consistent with the same object.
  - We can now consider each feature conditioned on the selected object location hypothesis.
  - This allows us to backproject a local figure-ground label from selected votes.

Implicit Shape Model - Representation

- Learn appearance codebook
  - Extract local features at interest points
  - Agglomerative clustering ⇒ codebook
- Learn spatial distributions
  - Match codebook to training images
  - Record matching positions on object Training images (+reference segmentation)

Spatial occurrence distributions ⇒ local figure-ground labels

Implicit Shape Model - Segmentation

- Local Features
- Matched Codebook Entries
- Probabilistic Voting
- Backproject Maxima
- Backproject Meta-information
- 3D Voting Space (continuous)

Pixel Contributions

Segmentation

- Interpretation of $p(\text{figure})$ map
  - per-pixel confidence in object hypothesis
  - Use for hypothesis verification

Segmentation $p(\text{figure})$

$p(\text{ground})$

$p(\text{figure})$

$p(\text{ground})$
Perceptual and Sensory Augmented Computing

13

B. Leibe

Top-Down Segmentation: Motivation

- Secondary hypotheses (“mixtures of cars/cows/etc.”)
  - Desired property of algorithm! ⇒ robustness to occlusion
  - Standard solution: reject based on bounding box overlap
  - Problematic: may lead to missing detections!

Derivation: ISM Recognition

- Algorithm stages
  1. Voting
  2. Mean-shift search
  3. Backprojection

- Vote weights: contribution of a single feature $f$

  - Probability that object $o_i$ occurs at location $x$ given $(f,t)$
  \[
  p(o_i|x,f,t) = \sum_{C} \frac{p(C,f)}{p(o_i|C,f)}
  \]

  - How to measure those probabilities? Soft matching
  \[
  p(C,f) = \frac{1}{|C|} \quad \text{where} \quad |C| = |C|_{id(C)} \leq \theta
  \]

  \[
  p(o_i|C,f) = \frac{1}{\text{occurrence}(C_i)}
  \]

  - Activated codebook entries

  - Matched probability
  - Occurrence distribution

  \[
  \text{matching probability distribution}
  \]

Derivation: ISM Recognition

- Algorithm stages
  1. Voting
  2. Mean-shift search
  3. Backprojection

- Vote weights: contribution of a single feature $f$

  - Probability that object $o_i$ occurs at location $x$ given $(f,t)$
  \[
  p(o_i|x,f,t) = \sum_{C} \frac{p(C,f)}{p(o_i|C,f)}
  \]

  - Likelihood of the observed features given the object hypothesis
  \[
  p(f|x_o,s) = \frac{p(o_i|x,f,t)p(f|x,t)}{p(o_i,s)} = \sum p(o_i|x,C_i)p(C_i|f)p(f|x,t)
  \]

  - Likelihood of the observed features given the object hypothesis
  \[
  p(f|x_o,s) = \frac{p(o_i|x,f,t)p(f|x,t)}{p(o_i,s)} = \sum p(o_i|x,C_i)p(C_i|f)p(f|x,t)
  \]
Derivation: ISM Recognition

1. Algorithm stages
   1. Voting
   2. Mean-shift search
   3. Backprojection

2. Vote weights: contribution of a single feature $f$
$$p(f | i_{o,x}) = \frac{p(v_{o,x} | f) p(f | i_{o,x})}{p(i_{o,x})} \sum_{i} p(v_{o,x} | C_i) p(C_i | f) p(f | i_{o,x})$$

Derivation: ISM Top-down Segmentation

1. Algorithm stages
   1. Voting
   2. Mean-shift search
   3. Backprojection

2. Vote weights: contribution of a single feature $f$
$$p(f | i_{o,x}) = \frac{p(v_{o,x} | f) p(f | i_{o,x})}{p(i_{o,x})} \sum_{i} p(v_{o,x} | C_i) p(C_i | f) p(f | i_{o,x})$$

3. Figure-ground backprojection
$$p(p = figure | i_{o,x}, f | C_i) = p(p = figure, i_{o,x}, C_i, f) p(f | i_{o,x})$$
Top-Down Segmentation Algorithm

This may sound quite complicated, but it boils down to a very simple algorithm...

Example Results: Motorbikes

Example Results: Cows

Example Results: Chairs

Inferring Other Information: Part Labels
References and Further Reading

- Details about the ISM approach can be found in

- Details about the DPMs can be found in

- Try the ISM Linux binaries
  - http://www.vision.ee.ethz.ch/bleibe/code

- Try the Deformable Part-based Models
  - http://www.cs.uchicago.edu/~pff/latent

Some Remarks

- These were just two representative approaches.
- Several other part-based models in active use, e.g.
  - Tree-structured models
e. g. [Felzenszwalb & Huttenlocher ’05]

- Hierarchical representations
e. g. [Bouchard & Triggs ’04]

- Dense part layouts
e. g. [Winn & Shotton ’06]

Discussion: Implicit Shape Model

- Pros:
  - Works well for many different object categories
  - Both rigid and articulated objects
  - Flexible geometric model
  - Can recombine parts seen on different training examples
  - Learning from relatively few (50-100) training examples
  - Optimized for detection, good localization properties

- Cons:
  - Needs supervised training data
  - Object bounding boxes for detection
  - Reference segmentations for top-down segm.
  - Only weak geometric constraints
  - Result segmentations may contain superfluous body parts.
  - Purely representative model
  - No discriminative learning

You Can Try It At Home...

- Linux binaries available
  - Including datasets & several pre-trained detectors
  - http://www.vision.ee.ethz.ch/bleibe/code

Inferring Other Information: Depth Maps

“Depth from a single image”