Announcements

- Happy new year everybody!
- Exercise sheet 5 available
  - Interest points, matching, homographies
  - Exercise will take place next Wednesday
- Several more announcements to follow tomorrow...

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 ≥3 correspondences
  - Fitting a homography from ≥4 correspondences
- Correspondences may be noisy and may contain outliers
  - Use RANSAC for robust fitting
  - Randomly select a seed group of correspondences to estimate a transformation hypothesis.
  - Find inliers to the transformation.
  - Iterate until a solution with enough inliers has been found or the non-existence of such a solution can be inferred with high probability.
Recap: Generalized Hough Transform

- Suppose our features are scale- and rotation-invariant:
  - Then a single feature match provides an alignment hypothesis (translation, scale, orientation).

![Image](image1.png)

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.

![Image](image2.png)

Recap: Identification vs. Categorization

- Find this particular object
- Recognize ANY car
- Recognize ANY cow

![Image](image3.png)

Recap: Visual Words

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

![Image](image4.png)

Recap: Bag-of-Word Representations (BoW)

Object → Bag of “words”

![Image](image5.png)

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).

![Image](image6.png)
Recap: Advantage of BoW Histograms

- Bag of words representations make 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.

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
**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 (2D image fragments)
  - structure (configuration of parts)

**Different Connectivity Structures**

- O(N)
- O(N^2)
- O(N^3)
- O(N^4)
- O(N^5)

- Fergus et al. '03
- Leibe et al. '04, '06
- Crandall et al. '05
- Fergus et al. '05
- Felzenszwalb & Huttenlocher '05
- Carreira & Lowe '06

**Some Class-Specific Graphs**

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

**Spatial Models Considered Here**

- Fully connected shape model
- “Star” shape model

- e.g. Constellation Model
- Parts fully connected
- Recognition complexity: O(N)
- Method: Exhaustive search
- e.g. Implicit Shape Model (ISM)
- Parts mutually independent
- Recognition complexity: O(NP)
- Method: Gen. Hough Transform

**Topics of This Lecture**

- Part-Based Models for Object Categorization
  - Structure representations
  - Different connectivity structures
- Constellation Model
  - Probabilistic model
  - Example results
  - Weakly supervised learning
  - Implicit Shape Model
    - Basic approach
      - Connection with segmentation
    - Theoretical derivation
    - Example results
    - Extension
- Current Challenges
Probabilistic Model

\[ P(image \backslash object) = P(\text{appearance}, \text{shape} \backslash object) \]

\[ = \sum_h P(\text{appearance} \backslash h, \text{object}) p(\text{shape} \backslash h, \text{object}) p(h \backslash object) \]

Distribution over patch descriptors

High-dimensional appearance space

Probabilistic Model

\[ P(image \backslash object) = P(\text{appearance}, \text{shape} \backslash object) \]

\[ = \sum_h P(\text{appearance} \backslash h, \text{object}) p(\text{shape} \backslash h, \text{object}) p(h \backslash object) \]

Distribution over joint part positions

2D image space
**Probabilistic Model**

\[ P(\text{image} \mid \text{object}) = P(\text{appearance}, \text{shape} \mid \text{object}) = \sum P(\text{appearance} \mid \text{h}, \text{object}) P(\text{shape} \mid \text{h}, \text{object}) P(\text{h} \mid \text{object}) \]

- **Interpretation**: How likely is this configuration?
  - This term may be used to model occlusion effects.
  - E.g. that a certain object part is not visible in the test image.

- **Marginalization over all configurations h**
  - This means we consider all possible feature-part assignments.
  - If there are several such assignments, all will contribute to the probability score.

So, this feature could be part 2!

But also this feature.

We don't need to decide for one of them!

**Constellation Model: Practical Use**

- **Test likelihood ratio compared to clutter model**

**Example Results: Constellation Model**

- **Data from four categories**

  - **Faces**
  - **Motorbikes**
  - **Airplanes**
  - **Spotted cats**
Learning with Weak Supervision

- How can we learn these models in the presence of clutter?

   - Main ideas:
     - Use interest operator to detect local features (on both fg and bg).
     - If training objects have similar appearance, these regions will often be similar in different training examples.
     - Cluster features: large clusters used to select candidate fg parts.
     - Choose most informative parts while simultaneously estimating model parameters.
     - Iteratively try different combinations of a small number of parts and check model performance on validation set to evaluate quality.

Detect Features

- Use a scale invariant detector (like DoG in SIFT detection)
Learning with Weak Supervision

Which of the candidate parts define the class, in what configuration?

- Let’s assume
  - We know the number of parts that define the model (and can keep it small).
  - The object of interest is only consistent thing somewhere in each training image.

Discussion: Constellation Model

- Pros:
  - Works well for many different object categories
  - Can adapt well to categories where
    - Shape is more important
    - Appearance is more important
  - Everything is learned from training data
  - Weakly supervised training possible

- Cons:
  - Model contains many parameters that need to be estimated
  - Cost increases exponentially with increasing number of parameters
  - Fully connected model restricted to small number of parts.

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 ➔ Prob. match to object part
  - NN matching ➔ Soft matching
  - Feature location on obj. ➔ Part location distribution
  - Uniform votes ➔ Probabilistic vote weighting
  - Quantized Hough array ➔ Continuous Hough space

Topics of This Lecture

- Part-Based Models for Object Categorization
  - Structure representations
  - Different connectivity structures
- Constellation Model
  - Probabilistic model
  - Example results
  - Weakly supervised learning
- Implicit Shape Model
  - Basic approach
  - Connection with segmentation
  - Theoretical derivation
  - Example results
  - Extensions
- Current Challenges
Implicit Shape Model: Basic Idea

- Visual vocabulary is used to index votes for object position [a visual word = “part”].

Implicit Shape Model: Basic Idea

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

Implicit Shape Model - Representation

- Learn appearance codebook
  - Extract local features at interest points
  - Agglomerative clustering of codebook

- Learn spatial distributions
  - Match codebook to training images
  - Record matching positions on object

Spatial occurrence distributions

- Local figure-ground labels

Implicit Shape Model - Recognition

- Interest Points
- Matched Codebook Entries
- Probabilistic Voting

Backprojection of Maxima

Probabilistic vote weighting

3D voting space (continuous)

Example: Results on Cows
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
    \[ x_{vote} = x_{map} - x_{cen}(s_{map}/s_{cen}) \]
    \[ y_{vote} = y_{map} - y_{cen}(s_{map}/s_{cen}) \]
    \[ s_{vote} = (s_{map}/s_{cen}) \]
  - Search for maxima in 3D voting space

Scale Voting: Efficient Computation

- 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.

Detection Results

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

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: Motivation

- Secondary hypotheses ("mixtures of cars/cows/etc.")
  - Desired property of algorithm! \(\Rightarrow\) robustness to occlusion
  - Standard solution: reject based on bounding box overlap
  - Problematic \(\Rightarrow\) 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 \((f,t)\)
    \[ p(o_i|t) = \sum_{i} p(C|t) \]
- How to measure those probabilities?
  - \(p(C|t) = \frac{1}{|C|}\) where \(C = \{C_i | d(C_i, f) \leq \theta\}\)

Top-Down Segmentation: Motivation

- Secondary hypotheses ("mixtures of cars/cows/etc.")
  - Desired property of algorithm! \(\Rightarrow\) robustness to occlusion
  - Standard solution: reject based on bounding box overlap
  - Problematic \(\Rightarrow\) may lead to missing detections!
  - Use segmentations to resolve ambiguities instead.
- Basic idea: each observed pixel can only be explained by (at most) one detection.

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 \((f,t)\)
    \[ p(o_i|t) = \sum_{i} p(C|t) \]
- Matching occurrence 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 \((f,t)\)
    \[ p(o_i|t) = \sum_{i} p(C|t) \]
- Matching occurrence 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 \((f,t)\)
    \[ p(o_i|t) = \sum_{i} p(C|t) \]
- Matching occurrence distribution

- Location of object \(o_i\) at location \(x\)

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 \((f,t)\)
    \[ p(o_i|t) = \sum_{i} p(C|t) \]
- Matching occurrence distribution

- Location of object \(o_i\) at location \(x\)
Derivation: ISM Recognition

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

- Vote weights: contribution of a single feature $f$
  
  \[ p(f|\omega, x) = \frac{p(\omega, x|f)p(f|x)}{p(\omega, x)} = \frac{p(\omega, x|f)p(f|x)}{p(\omega, x)} \]

Derivation: ISM Top-Down Segmentation

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

- Vote weights: contribution of a single feature $f$
  
  \[ p(f|\omega, x) = \frac{p(\omega, x|f)p(f|x)}{p(\omega, x)} = \frac{p(\omega, x|f)p(f|x)}{p(\omega, x)} \]

- Figure-ground backprojection
  
  \[ p(p = \text{fig}|\omega, x, f, i) = \sum p(p = \text{fig}|\omega, x, C, i) \]

Marginalize over all codebook entries matched to $i$
Top-Down Segmentation Algorithm

Algorithm 5 The top-segmentation algorithm.

\[
\text{// Given: hypothesis } h \text{ and supporting votes } V_h.
\]
\[
\text{for all supporting votes } (x, y, v_{x'y'}) \in V_h, \text{ do}
\]
\[
\text{Let } m_{y_{new}} \text{ be the segmentation mask corresponding to } v_{x'y'}.
\]
\[
\text{Let } a \text{ be the size at which the interest region } I \text{ was sampled.}
\]
\[
\text{Rescale } m_{y_{new}} \text{ to } a.
\]
\[
\text{m}_{y_{new}} = (m_{y_{new}} - \frac{1}{2}a) \times a.
\]
\[
\text{for all } y \in \{0, a - 1\}, \text{ do}
\]
\[
\text{m}_{y_{new}}(a - m_{y_{new}} - y) = w \times m_{y_{new}}(a - m_{y_{new}} - y) + w \times (1 - m_{y_{new}}(a - m_{y_{new}} - y))
\]
\[
\text{end for}
\]
\[
\text{end for}
\]

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

Segmentation

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

Example Results: Motorbikes

- Training
  - 112 hand-segmented images

Example Results: Cows

- Results on novel sequences:
  - Single-frame recognition - No temporal continuity used!

Example Results: Chairs

- Office chairs
- Dining room chairs

Detections Using Ground Plane Constraints

- Battery of 5 ISM detectors for different car views
Inferring Other Information: Part Labels

- Training
- Test
- Output

[Thomas, Ferrari, Tuytelaars, Leibe, Van Gool, 3DRR'07; RSS'08]

Inferring Other Information: Depth Maps

- "Depth from a single image"

[Thomas, Ferrari, Tuytelaars, Leibe, Van Gool, 3DRR'07; RSS'08]

Extension: Estimating Articulation

- Try to fit silhouette to detected person

- Basic idea
  - Search for the silhouette that simultaneously optimizes the
  - Chamfer match to the distance-transformed edge image
  - Overlap with the top-down segmentation
  - Enforces global consistency
  - Caveat: introduces again reliance on global model

[B. Leibe]

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)

[Thomas, Ferrari, Tuytelaars, Leibe, Van Gool, 3DRR'07; RSS'08]

Sometimes, Rotation Invariance Is Needed...

[Figure from Mikolajczyk et al., CVPR'06]
**You Can Try It At Home...**

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

**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

**Topics of This Lecture**

- **Part-Based Models for Object Categorization**
  - Structure representations
  - Different connectivity structures
- **Constellation Model**
  - Probabilistic model
  - Example results
  - Weekly supervised learning
- **Implicit Shape Model**
  - Basic approach
  - Connection with segmentation
  - Theoretical derivation
  - Example results
  - Extensions

**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]

**Multi-Category Discrimination**

- Distinguish similar categories.
- Need to look at specific details!

**Multi-Aspect Recognition**

- Detectors for different viewpoints
  - How can this be improved?
Multi-Aspect Recognition

- Some approaches...

References and Further Reading

- Details about the Constellation Model can be found in
  - R. Fergus, A. Zisserman, P. Perona
  Object Class Recognition by Unsupervised Scale-Invariant Learning, in CVPR '03, 2003.

- Details about the ISM approach can be found in
  - B. Leibe, A. Leonardis, and B. Schiele,

- Try the Constellation Model demo

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