Perceptual and Sensory Augmented Computing

Computer Vision - Lecture 16
Part-based Models for Object Categorization

07.01.2014

Bastian Leibe
RWTH Aachen
http://www.vision.rwth-aachen.de
leibe@vision.rwth-aachen.de

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

Topics of This Lecture

- Recap: Specific Object Recognition with Local Features
- Part-Based Models for Object Categorization
  - Structure representations
  - Different connectivity structures
- Bag-of-Words Model
  - Use for image classification
- Implicit Shape Model
  - Generalized Hough Transform for object category detection
- Deformable Part-based Model
  - Discriminative part-based detection

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: Indexing features

Detect or sample features
List of positions, scales, orientations
Describe features
Associated list of d-dimensional descriptors

⇒ Shortlist of possibly matching images + feature correspondences

Recap: Fast Indexing with Vocabulary Trees

- Recognition

Geometric verification

Slide credits:
- Kristen Grauman
- David Lowe
- David Nister
- Bastian Leibe
Recap: Geometric Verification 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 $\geq 3$ correspondences
  - Fitting a homography from $\geq 4$ correspondences
    \[ A \cdot t = b \quad \text{Affine: solve a system} \]
    \[ A \cdot h = 0 \quad \text{Homography: solve a system} \]
- Correspondences may be noisy and may contain outliers
  - Need to use robust methods that can filter out outliers
  - Use RANSAC or the Generalized Hough Transform

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

- Articulated motion
  - People
  - Animals
- Special parameterisations
  - Limb angles
Topics of This Lecture
- Recap: Specific Object Recognition with Local Features
- Part-Based Models for Object Categorization
  - Structure representations
  - Different connectivity structures
- Bag-of-Words Model
  - Use for image classification
- Implicit Shape Model
- Generalized Hough Transform for object category detection
- Deformable Part-based Model
  - Discriminative part-based detection

Analogy to Documents
China is forecasting a trade surplus of $60bn (21bn) to $80bn this year, a threefold increase on 2004’s $22bn. The Commerce Ministry said the surplus would be created by the current account surpluses of $33bn and the services surplus of $18bn. China has been under an agreement ending its currency undervaluation, a condition of WTO accession. The US and major trading partners have pressed China to improve its currency regime. Zhou Xiaochuan, governor of the People’s Bank of China, has made it clear that it will take its time to allow the renminbi (yuan) to float. However, it will be allowed to trade freely. However, he has made it clear that it will take its time and tread carefully before allowing the yuan to rise further in value.

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.
- Main difference to text retrieval: visual words are not given a priori, but obtained through clustering (e.g., using k-means)

Similarly, Bags-of-Textons for Texture Repr.

Object → Bag of ‘words’

Source: ICCV 2005 short course, Li Fei-Fei

Source: ICCV 2005 short course, Li Fei-Fei

Source: ICCV 2005 short course, Li Fei-Fei
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.

\[
\text{sim}(d_i, q) = \frac{d_i \cdot q}{|d_i| \times |q|} = \frac{\sum_{w_{i,j} \in d_i} w_{i,j} \times w_{j,q}}{\sqrt{\sum_{w_{i,j} \in d_i} w_{i,j}^2} \times \sqrt{\sum_{w_{j,q} \in q} w_{j,q}^2}}
\]

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), Willamowski et al. (2005), Grauman & Darrell (2005), Sivic et al. (2003, 2005)

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?
BoW Representation: Spatial Information

- A bag of words is an orderless representation: throwing out spatial relationships between features
- Middle ground:
  - Visual “phrases” : frequently co-occurring words
  - Semi-local features : describe configuration, neighborhood
  - Let position be part of each feature
  - Count bags of words only within sub-grids of an image
  - After matching, verify spatial consistency (e.g., look at neighbors - are they the same too?)

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.

Topics of This Lecture

- Recap: Specific Object Recognition with Local Features
- Part-Based Models for Object Categorization
  - Structure representations
  - Different connectivity structures
- Bag-of-Words Model
  - Use for image classification
- Implicit Shape Model
  - Generalized Hough Transform for object category detection
- Deformable Part-based Model
  - Discriminative part-based detection
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

- **Features are considered independent given obj. center**
- **Objects are detected as consistent configurations of the observed parts (visual words).**

Implicit Shape Model: Basic Idea

- Visual vocabulary is used to index votes for object position [a visual word = “part”].
- Visual codeword with displacement vectors
- Test image
- Training image
- Implicit Shape Model - Representation
- Learn appearance codebook
  - Extract local features at interest points
  - Agglomerative clustering \(\rightarrow\) codebook
- Learn spatial distributions
  - Match codebook to training images
  - Record matching positions on object
- Spatial occurrence distributions

Implicit Shape Model - Recognition

- Interest Points
- Matched Codebook Entries
- Probabilistic Voting
- Interpretation (Codebook match)
- Object Position
- 1D Voting Space (continuous)
- Probabilistic vote weighting
- Backprojected Hypotheses
- Backprojection of Maxima
Example: Results on Cows

Detection Results

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

Detection Using Ground Plane Constraints

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)

Scale Invariant Voting

- Scale-invariant feature selection
  - Scale-invariant interest regions
  - Extract scale-invariant descriptors
  - Match to appearance codebook

- Generate scale votes
  - Scale as 3rd dimension in voting space
    \[ x_{scale} = x_{img} - x_{rec}(s_{img}/s_{rec}) \]
    \[ y_{scale} = y_{img} - y_{rec}(s_{img}/s_{rec}) \]
    \[ s_{scale} = (s_{img}/s_{rec}) \]
  - Search for maxima in 3D voting space

Sometimes, Rotation Invariance Is Needed...
**Starting Point: HOG Sliding-Window Detector**

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

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

\[
\text{Score of } F \text{ at position } p \text{ is } F \cdot \phi(p, H)
\]

**Deformable Part-based Models**

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

**Example Results: Motorbikes**

**Implicit Shape Model - Segmentation**

- Local Features
- Matched Codebook Entries
- Probabilistic Voting
- Backprojected Hypotheses
- Backprojection of Maxima

**Topics of This Lecture**

- Recap: Specific Object Recognition with Local Features
- Part-Based Models for Object Categorization
  - Structure representations
  - Different connectivity structures
- Bag-of-Words Model
  - Use for image classification
- Implicit Shape Model
  - Generalized Hough Transform for object category detection
- Deformable Part-based Model
  - Discriminative part-based detection

**You Can Try It At Home...**

- Linux source code & binaries available
  - Including datasets & several pre-trained detectors
  - [http://www.vision.rwth-aachen.de/software](http://www.vision.rwth-aachen.de/software)
2-Component Bicycle Model

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

Object Hypothesis

- Score of filter: dot product of filter with HOG features underneath it
- Score of object hypothesis is sum of filter scores minus deformation costs

Score of a Hypothesis

\[
score(p_0, \ldots, p_m) = \sum_{i=0}^m F_i \cdot \phi(H, p_i) - \sum_{i=1}^m d_i \cdot (dx^2, dy^2)
\]

\[
score(z) = \beta \cdot \Psi(H, z)
\]

- "data term"
- "spatial prior"
- concatenation filters and deformation parameters
- concatenation of HOG features and part displacement features

Recognition Model

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

\[
f_w(x) = \max_w \Phi(x, z)
\]

- \( z \): 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.
  - 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

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