Lecture 10
Recognition

Dr Christopher Brunner
AMME4710 – Computer Vision and Image Processing
Semester 2, 2013

By the end of this lecture you will be able to:

• Appreciate the importance of image recognition and its applications
• Describe the main approaches for category recognition
• Understand the approaches for context and scene understanding
Resources

• Readings
  – Szeliski, Computer Vision, Ch. 14
    http://szeliski.org/Book/

• Additional material (optional)

• Slides credits
  – L. Fei-Fei, R. Fergus, A. Torralba, S. Lazebnik, K. Grauman, T. Darrell
Example Applications

Mobile tourist guide
  • Self-localization
  • Object/building recognition
  • Photo/video augmentation

B. Leibe

[Quack, Leibe, Van Gool, CIVR’08]
Application: Large-Scale Retrieval

Query

Results from 5k Flickr images (demo available for 100k set)

[Philbin CVPR’07]
Web Demo: Movie Poster Recognition

50'000 movie posters indexed

Query-by-image from mobile phone available in Switzerland

Outline

• Category Recognition
  – Bag of Words
  – Part-based models

• Context and scene understanding
  – Large image collections
BAG OF WORDS
Example of object recognition

1000+ descriptors per frame

Slide credit: J. Sivic
Match regions between frames using SIFT descriptors and spatial consistency

Multiple regions overcome problem of partial occlusion

- Shape adapted regions
- Maximally stable regions

Slide credit: J. Sivic
Indexing local features

- When we see close points in feature space, we have similar descriptors, which indicates similar local content.
Visual search using local regions

Schmid and Mohr ’97 – 1k images
Sivic and Zisserman’03 – 5k images
Nister and Stewenius’06 – 50k images (1M)
Philbin et al.’07 – 100k images
Chum et al.’07 + Jegou and Schmid’07 – 1M images
Chum et al.’08 – 5M images

Index 1 billion (10^9) images
– 200 servers each indexing 5M images?

Slide credit: J. Sivic
Indexing local features: inverted file index

• For text documents, an efficient way to find all pages on which a word occurs is to use an index...

• We want to find all images in which a feature occurs.

• To use this idea, we’ll need to map our features to “visual words”.

Kristen Grauman
Object \rightarrow \text{Bag of ‘words’}
Bag-of-words models

- Orderless document representation: frequencies of words from a dictionary

Salton & McGill (1983)

1962-10-22: Soviet Missiles in Cuba

abandon achieving adversaries aggression agricultural appropriate armaments arms assessments atlantic ballistic berlin buildup burdens cargo college commitment communist constitution consumers cooperation crisis dangers declined defensive deficit depended disarmament divisions domination doubled economic education elimination emergence endangered equals europe expand exports fact false family forum freedom fulfill gromyko halt hazards hemisphere hospitals ideals independent industries inflation labor latin limiting minister missiles modernization neglect nuclear oas obligation observer offensive peril pledged predicted purchasing quarantine quote recession rejection republics retaliatory safeguard sites solution soviet space spur stability standby strength surveillance tax territory treaty undertakings unemployment weapons war warhead weapons

1941-12-08: Request for a Declaration of War

abandoning acknowledge aggression aggressors airplanes armaments armed army assault assembly authorizations bombing britain british cheerfully claiming constitution curtail december defeats defending delays democratic dictators disclose economic empire endanger facts false forgotten fortunes france freedom fulfilled fullness fundamental gangsters german germany god guam harbor hawaii hemisphere hitler hostilities immune improving indies innumerable invasion islands isole japanese labor metals midst midway navy nazis obligation offensive officially pacific partisanship patriotism pearl peril perpetrated perpetual philippine preservation privilege reject repaired resisting retain revealing rumors seas soldiers speaks speedy stamina strength sunday sunk supremacy tanks taxes treachery true tyranny undertaken victory war wartime washington
Bag-of-words models

- We analyzed 700 million words, phrases, and topic instances collected from the Facebook messages of 75,000 volunteers, who also took standard personality tests, and found striking variations in language with personality, gender, and age. In our open-vocabulary technique, the data itself drives a comprehensive exploration of language that distinguishes people, finding connections that are not captured with traditional closed-vocabulary word-category analyses.

Gender and the IGF: participation and language used.
www.diplomacy.edu
Visual words

- Map high-dimensional descriptors to tokens/words by quantizing the feature space.

- Quantize via clustering, let cluster centers be the prototype “words”.

- Determine which word to assign to each new image region by finding the closest cluster center.

Kristen Grauman
Bag-of-words steps

1. Extract features
2. Learn “visual vocabulary”
3. Quantize features using visual vocabulary
4. Represent images by frequencies of “visual words”
1. Feature detection and representation

- Detect patches
  - [Mikołajczyk and Schmid '02]
  - [Mata, Chum, Urban & Pajdla, '02]
  - [Sivic & Zisserman, '03]

- Local interest operator
  - or
- Regular grid

Compute descriptor
  - e.g. SIFT [Lowe'99]

Normalize patch
1. Feature extraction
2. Learning the visual vocabulary
2. Learning the visual vocabulary

Slide credit: Josef Sivic
2. Learning the visual vocabulary

Clustering

Visual vocabulary

Clustering

Slide credit: Josef Sivic
Uses of BoW representation

• Treat as feature vector for standard classifier
  – e.g. SVM

• Cluster BoW vectors over image collection
  – Discover visual themes

• Hierarchical models
  – Decompose scene/object
Clustering Visual Words

- Use models from ‘information retrieval’ literature
  - Probabilistic latent semantic analysis (pLSA)
  - Latent Dirichlet allocation (LDA)

\[
d = \text{image}, \quad w = \text{visual word}, \quad z = \text{topic (cluster)}
\]
Clustering and vector quantization

- Clustering is a common method for learning a visual vocabulary or codebook
  - Unsupervised learning process.
  - Each cluster center produced by k-means becomes a codevector.
  - Codebook can be learned on separate training set.
  - Provided the training set is sufficiently representative, the codebook will be “universal”.

- The codebook is used for quantizing features
  - A vector quantizer takes a feature vector and maps it to the index of the nearest codevector in a codebook.
  - Codebook = visual vocabulary.
  - Codewords = visual words.
Examples of codewords

Sivic et al. 2005
Example of codebook

Source: B. Leibe
Example: Video Google System

1. Collect all words within query region
2. Inverted file index to find relevant frames
3. Compare word counts
4. Spatial verification

Sivic & Zisserman, ICCV 2003
Visual vocabularies: Issues

• How to choose vocabulary size?
  – Too small: visual words not representative of all patches
  – Too large: quantization artifacts, overfitting

• Computational efficiency
  – Vocabulary trees
    (Nister & Stewenius, 2006)
Vocabulary Trees: hierarchical clustering for large vocabularies

Tree construction:
Vocabulary Tree

Training: Filling the tree

Slide credit: David Nister
Vocabulary Tree

Training: Filling the tree
Vocabulary Tree

Training: Filling the tree

[Nister & Stewenius, CVPR’06]

Slide credit: David Nister
Vocabulary Tree

Training: Filling the tree
Vocabulary Tree

Training: Filling the tree
Vocabulary Tree

Recognition

Verification on spatial layout

Slide credit: David Nister
Bag of words: Summary

+ flexible to geometry / deformations / viewpoint
+ compact summary of image content
+ provides vector representation for sets
+ very good results in practice

- background and foreground mixed when bag covers whole image
- optimal vocabulary formation remains unclear
- basic model ignores geometry – must verify afterwards, or encode via features

Kristen Grauman
PART-BASED MODELS
Problem with bag-of-words

- All have equal probability for bag-of-words methods
- Location information is important
- BoW + location still doesn’t give correspondence
Model: Parts and Structure
Representation

- Object as set of parts
  - Generative representation

- Model:
  - Appearance of part
  - Relative locations between parts

- Issues:
  - How to represent appearance
  - How to model location
  - How to handle occlusion/clutter

Figure from [Fischler & Elschlager 73]
History of Parts and Structure approaches

- Fischler & Elschlager 1973
- Yuille ‘91
- Brunelli & Poggio ‘93
- Lades, v.d. Malsburg et al. ‘93
- Cootes, Lanitis, Taylor et al. ‘95
- Amit & Geman ‘95, ‘99
- Felzenszwalb & Huttenlocher ’00, ’04
- Crandall & Huttenlocher ’05, ’06
- Leibe & Schiele ’03, ’04
- Many papers since 2000
Sparse representation

+ Computationally tractable (10^5 pixels → 10^1 -- 10^2 parts)
+ Generative representation of class
+ Avoid modeling global variability
+ Success in specific object recognition

- Throw away most image information
- Parts need to be distinctive to separate from other classes
The correspondence problem

- Model with $P$ parts
- Image with $N$ possible assignments for each part
- Consider mapping to be 1-1

$N^P$ combinations!!!
Different connectivity structures

\[
\begin{align*}
O(N^6) &\quad \text{Fergus et al. '03} \\
O(N^2) &\quad \text{Fei-Fei et al. '03} \\
O(N^3) &\quad \text{Crandall et al. '05} \\
O(N^2) &\quad \text{Crandall et al. '05} \\
O(N^2) &\quad \text{Felzenszwalb & Huttenlocher '00} \\
\end{align*}
\]

\begin{itemize}
\item[a)] Constellation [13]
\item[b)] Star shape [9, 14]
\item[c)] \(k\)-fan \((k = 2)\) [9]
\item[d)] Tree [12]
\item[e)] Bag of features [10, 21]
\item[f)] Hierarchy [4]
\item[g)] Sparse flexible model
\end{itemize}

from Sparse Flexible Models of Local Features
G. Carneiro and D. Lowe, ECCV 2006
Efficient methods

• Distance transforms

• Felzenszwalb and Huttenlocher ‘00 and ‘05

• $O(N^2P) \rightarrow O(NP)$ for tree structured models

• Removes need for region detectors
How much does shape help?

- Crandall, Felzenszwalb, Huttenlocher CVPR’05
- Shape variance increases with increasing model complexity
- Do get some benefit from shape
Appearance representation

- SIFT
- Decision trees
  [Lepetit and Fua CVPR 2005]
- PCA

Figure from Winn & Shotton, CVPR ‘06

- Image gradients
- Keypoint descriptor

Figure from Winn & Shotton, CVPR ‘06

[Lepetit and Fua CVPR 2005]
Learn Appearance

• Generative models of appearance
  – Can learn with little supervision
  – E.g. Fergus et al’ 03

• Discriminative training of part appearance model
  – SVM part detectors
  – Felzenszwalb, Mcallester, Ramanan, CVPR 2008
  – Much better performance
Felzenszwalb, Mcallester, Ramanan, CVPR 2008

• 2-scale model
  – Whole object
  – Parts

• HOG representation + SVM training to obtain robust part detectors

• Distance transforms allow examination of every location in the image
Hierarchical Representations

- Pixels $\rightarrow$ Pixel groupings $\rightarrow$ Parts $\rightarrow$ Object

- Multi-scale approach increases number of low-level features

- Amit and Geman ’98
- Ullman et al.
- Bouchard & Triggs ’05
- Zhu and Mumford
- Jin & Geman ‘06
- Zhu & Yuille ’07
- Fidler & Leonardis ‘07

Images from [Amit98]
Example: Implicit shape models

- Visual codebook is used to index votes for object position

B. Leibe, A. Leonardis, and B. Schiele, *Combined Object Categorization and Segmentation with an Implicit Shape Model*, ECCV Workshop on Statistical Learning in Computer Vision 2004
Implicit shape models

- Visual codebook is used to index votes for object position

B. Leibe, A. Leonardis, and B. Schiele, *Combined Object Categorization and Segmentation with an Implicit Shape Model*, ECCV Workshop on Statistical Learning in Computer Vision 2004
Implicit shape models: Training

1. Build codebook of patches around extracted interest points using clustering
Implicit shape models: Training

1. Build codebook of patches around extracted interest points using clustering
2. Map the patch around each interest point to closest codebook entry
Implicit shape models: Training

1. Build codebook of patches around extracted interest points using clustering
2. Map the patch around each interest point to closest codebook entry
3. For each codebook entry, store all positions it was found, relative to object center
Implicit shape models: Testing

1. Given test image, extract patches, match to codebook entry
2. Cast votes for possible positions of object center
3. Search for maxima in voting space
4. Extract weighted segmentation mask based on stored masks for the codebook occurrences
Detection Results

• Qualitative Performance
  - Recognizes different kinds of objects
  - Robust to clutter, occlusion, noise, low contrast
The Naïve Bayes model

- Generative model for a bag of features
- Assume that each feature $f_i$ is conditionally independent given the class $c$:

$$p(image \mid c) = p(f_1, \ldots, f_N \mid c) = \prod_{i=1}^{N} p(f_i \mid c)$$
Probabilistic constellation model

\[ P(image \mid object) = P(appearance, shape \mid object) \]

Part descriptors

Part locations

Candidate parts

Source: Lana Lazebnik
Probabilistic constellation model

\[ P(\text{image} \mid \text{object}) = P(\text{appearance}, \text{shape} \mid \text{object}) \]
**Probabilistic constellation model**

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

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

h: assignment of features to parts

Source: Lana Lazebnik
Probabilistic constellation model

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

\[ = \max_h P(\text{appearance} \mid h, \text{object}) p(\text{shape} \mid h, \text{object}) p(h \mid \text{object}) \]
Probabilistic constellation model

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

\[ = \max_h P(\text{appearance} \mid h, \text{object}) p(\text{shape} \mid h, \text{object}) p(h \mid \text{object}) \]
Results: Faces

Face shape model

Recognition results

Patch appearance model
Parts-based models: Summary

- Explicit notion of correspondence between image and model
- Efficient methods for large # parts and # positions in image
- With powerful part detectors, can get state-of-the-art performance
- Hierarchical models allow for more parts