CS 188: Artificial Intelligence
Spring 2011

Lecture 23: Computer Vision
4/20/2011

Pieter Abbeel – UC Berkley
Slides adapted from Trevor Darrell (and his sources)

Announcements

- On-going: contest (optional and FUN!)
- "Bonus" lectures:
  - Wednesday: Machine Learning for Computer Vision
  - Next Monday: Case Studies in Speech/Language and Robotics
  - Next Wednesday:
    - Course Wrap-Up
    - Pointers to courses and books for those who want to learn more AI
    - Contest!
- RRR Week Monday and Wednesday:
  - Review Sessions

Object Categorization

- How to recognize ANY car
- How to recognize ANY cow

Challenges: robustness

- Detection in Crowded Scenes
  - Learn object variability
  - Changes in appearance, scale, and articulation
  - Compensate for clutter, overlap, and occlusion

Recognition Challenges / Overview
Challenges: context and human experience

Context cues

Challenges: learning with minimal supervision

This is a pottpod

Find the pottpod

Rough evolution of focus in recognition research

Slide from Pietro Perona, 2004 Object Recognition workshop

Slide from Pietro Perona, 2004 Object Recognition workshop
Inputs/outputs/assumptions

- What is the goal?
  - Say yes/no as to whether an object present in image
  - Determine pose of an object, e.g., for robot to grasp
  - Categorize all objects
  - Forced choice from pool of categories
  - Bounding box on object
  - Full segmentation
  - Build a model of an object category

Today

- Scanning window paradigm
- HOG
- Boosted Face Detection
- BOW Indexing
- Scene Statistics

Scanning windows...

Detection via classification: Main idea

Basic component: a binary classifier

Car/non-car Classifier

No
Yes

car.

Detection via classification: Main idea

If object may be in a cluttered scene, slide a window around looking for it.

1. Obtain training data
2. Define features
3. Define classifier

Fleshing out this pipeline a bit more, we need to:

- Training examples
- Feature extraction
- Car/non-car Classifier
Detection via classification: Main idea

- Consider all subwindows in an image
  - Sample at multiple scales and positions (and orientations)
- Make a decision per window:
  - “Does this contain object category X or not?”

Feature extraction: Global appearance

- Simple holistic descriptions of image content
  - grayscale / color histogram
  - vector of pixel intensities

Eigenfaces: Global appearance description

An early appearance-based approach to face recognition

- Training images
- Mean
- Eigenvectors computed from covariance matrix
- Project new images to “face space”.
- Recognition via nearest neighbors in face space

Eigenfaces:

- Turk & Pentland, 1991

Gradient-based representations

- Consider edges, contours, and (oriented) intensity gradients

Derivatives and edges

- An edge is a place of rapid change in the image intensity function.
  - image intensity function (along horizontal scanline)
  - First derivative
  - Edges correspond to extrema of derivative

Feature extraction: Global appearance

- Pixel-based representations sensitive to small shifts
- Color or grayscale-based appearance description can be sensitive to illumination and intra-class appearance variation

Cartoon example: an albino koala
Partial derivatives of an image

\[ \frac{\partial f(x, y)}{\partial x} \]

\[ \frac{\partial f(x, y)}{\partial y} \]

Which shows changes with respect to \( x \)?

Assorted finite difference filters

Prewitt:

\[ M_x \]

\[ M_y \]

Sobel:

\[ M_x \]

\[ M_y \]

Roberts:

\[ M_x \]

\[ M_y \]

\[ \text{My} = \text{fspecial('sobel')}; \]
\[ \text{My} = \text{imfilter(double(im), My)}; \]
\[ \text{imagesc(outim)}; \]
\[ \text{colormap gray}; \]

Gradient-based representations

- Consider edges, contours, and (oriented) intensity gradients

- Summarize local distribution of gradients with histogram
  - Locally orderless: offers invariance to small shifts and rotations
  - Contrast-normalization: try to correct for variable illumination

Histograms of oriented gradients (HoG)

Map each grid cell in the input window to a histogram counting the gradients per orientation.

Code available:
http://pascal.inrialpes.fr/soft/olt/

K. Grauman, B. Leibe

Dalal & Triggs, CVPR 2005

Side credits: Dalal, Triggs, A. Boroumand
• Tested with
  – RGB
  – LAB
  – Grayscale
• Gamma Normalization and Compression
  – Square root
  – Log

• Histogram of gradient orientations
  -Orientation
  -Position
  – Weighted by magnitude

\[ E_1 - \text{norm: } x \rightarrow \sqrt{||x||^2 + \epsilon} \]
\[ E_2 - \text{max: } x \rightarrow \max\{\sqrt{||x||^2 + \epsilon} \} \]
Boosted Face Detection with Gradient Features

Gradient-based representations: Rectangular features

- Compute differences between sums of pixels in rectangles
- Captures contrast in adjacent spatial regions, efficient to compute
- Each feature parameterized by scale, position, type.

Viola & Jones, CVPR 2001
K. Grauman, B. Leibe
Boosting

- Build a strong classifier by combining number of “weak classifiers”, which need only be better than chance
- Sequential learning process: at each iteration, add a weak classifier
- Flexible to choice of weak learner
  - including fast simple classifiers that alone may be inaccurate
- We’ll look at Freund & Schapire’s AdaBoost algorithm
  - Easy to implement
  - Base learning algorithm for Viola-Jones face detector

AdaBoost: Intuition

Consider a 2-d feature space with positive and negative examples. Each weak classifier splits the training examples with at least 50% accuracy. Examples misclassified by a previous weak learner are given more emphasis at future rounds.

AdaBoost: Intuition

Final classifier is combination of the weak classifiers

AdaBoost Algorithm

Start with uniform weights on training examples

\[ \{x_1, \ldots, x_n\} \]

For T rounds

1. Initialize weights:
   \[ w_i = \frac{1}{n} \text{ for } i = 1, \ldots, n \]

2. For each feature, \( j \), train a classifier \( h_j \), which is restricted to using a single feature. The error is estimated with respect to \( w_i \):
   \[ e_j = \sum_{i} w_i I[y_i \neq h_j(x_i)] \]

3. Choose the classifier \( h_j \) with the lowest error \( e_j \)

4. Update the weights:
   \[ w_i^{t+1} = w_i^t e_j^{1-e_j} \]
   where \( e_j = 0 \) if example \( x_i \) is classified correctly, and \( e_j = 1 \) otherwise.

The final strong classifier is:

\[ h(x) = \arg \max_j \left[ \sum_{i} \log \left( \frac{1}{e_j} \right) \right] \]

where \( e_j = \log \frac{1}{1-e_j} \)

Freund & Schapire 1995

Example: Face detection

- Frontal faces are a good example of a class where global appearance models + a sliding window detection approach fit well:
  - Regular 2D structure
  - Center of face almost shaped like a "patch"/window
- Now we’ll take AdaBoost and see how the Viola-Jones face detector works
**Feature extraction**

“Rectangular” filters

Efficiently computable with integral image: any sum can be computed in constant time

Avoid scaling images ➔ scale features directly for same cost

**Visual Object Recognition Tutorial**

**Perceptual and Sensory Augmented Computing**

AdaBoost for feature+classifier selection

- Want to select the single rectangle feature and threshold that best separates positive (faces) and negative (non-faces) training examples, in terms of weighted error.

Resulting weak classifier:

$$ h(x) = \begin{cases} +1 & \text{if } f_i(x) > \theta_i \\ -1 & \text{otherwise} \end{cases} $$

For next round, reweight the examples according to errors, choose another filter/threshold combo.

**Visual Object Recognition Tutorial**

**Large library of filters**

Consider all possible filter parameters: position, scale, and type:

180,000+ possible features associated with each 24 x 24 window

Use AdaBoost both to select the informative features and to form the classifier

**Visual Object Recognition Tutorial**

**AdaBoost for Efficient Feature Selection**

- **Image Features ➔ Weak Classifiers**
- For each round of boosting:
  - Evaluate each rectangle feature on each example
  - Sort examples by filter values
  - Select best threshold for each filter (min error)
    - Sorted list can be quickly scanned for the optimal threshold
  - Select best filter/threshold combination
  - Weight on this feature is a simple function of error rate
  - Reweight examples

**AdaBoost Algorithm**

Start with uniform weights on training examples

For T rounds

- Evaluate weighted error for each feature, pick best.
- Re-weight the examples:
  - Incorrectly classified ➔ more weight
  - Correctly classified ➔ less weight

Final classifier is combination of the weak ones, weighted according to error they had.

Freund & Schapire 1995

**Visual Object Recognition Tutorial**

**4/24/2011**

- Even if the filters are fast to compute, each new image has a lot of possible windows to search.
- **How to make the detection more efficient?**
Cascading classifiers for detection

For efficiency, apply less accurate but faster classifiers first to immediately discard windows that clearly appear to be negative; e.g.,

- Filter for promising regions with an initial inexpensive classifier
- Build a chain of classifiers, choosing cheap ones with low false negative rates early in the chain

Viola-Jones Face Detector: Summary

- Train with 5K positives, 350M negatives
- Real-time detector using 38 layer cascade
- 6061 features in final layer
- [Implementation available in OpenCV: http://www.intel.com/technology/computing/opencv/]
Detecting profile faces?
Detecting profile faces requires training separate detector with profile examples.

Viola-Jones Face Detector: Results

Example application
Frontal faces detected and then tracked, character names inferred with alignment of script and subtitles.

Example application: faces in photos

Highlights
- Sliding window detection and global appearance descriptors:
  - Simple detection protocol to implement
  - Good feature choices critical
  - Past successes for certain classes

Limitations
- High computational complexity
  - For example: 250,000 locations x 30 orientations x 4 scales = 30,000,000 evaluations!
  - If training binary detectors independently, means cost increases linearly with number of classes
  - With so many windows, false positive rate better be low
Limitations (continued)

- Not all objects are “box” shaped

- Non-rigid, deformable objects not captured well with representations assuming a fixed 2d structure; or must assume fixed viewpoint

- Objects with less-regular textures not captured well with holistic appearance-based descriptions

Limitations (continued)

- If considering windows in isolation, context is lost

- In practice, often entails large, cropped training set (expensive)

- Requiring good match to a global appearance description can lead to sensitivity to partial occlusions

Context can constrain a sliding window search

Scaling up:
BOW Indexing
Outline of a large-scale retrieval strategy

1. Compute affine covariant regions in each frame independently
2. "Label" each region by a vector of descriptors based on its intensity
3. Finding corresponding regions is transformed to finding nearest neighbour vectors
4. Rank retrieved frames by number of corresponding regions
5. Verify retrieved frame based on spatial consistency

Example of object recognition

Match regions between frames using SIFT descriptors and spatial consistency

Multiple regions overcome problem of partial occlusion

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?

Beyond Nearest Neighbors...

Indexing local features using 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".

Object → Bag of ‘words’
**Analogy to documents**

Comparison of sensory, brain, visual, perception, eye, cell, optical nerve, image.Hubel, Wiesel. Hubel and Wiesel demonstrate that the image falling on the retina undergoes a step in the eye was projected. Through the messages 75x280 of sensory impressions proceeding to the brain, the visual experience are the final result. Our perception of the world is a result of sensory messages. The visual system is involved in visual processing: the retina, the visual cortex, and other brain centers. The cerebral cortex is the primary organ of sensory perception in the retina. The retina contains a specific detail in the pattern of the visual image.

China is forecasting a trade surplus of $90bn ($57bn to $123bn this year, a forecasted increase on 2004’s $32bn). The Commerce Ministry said the surplus would be created by a predicted 33% increase in exports to $275bn, compared with a 18% rise in imports to $183bn. The US wants the yuan to be allowed to trade more freely. However, Beijing has made it clear that it would not permit it to trade within a narrow band, but also not permitting it to trade when a recession, and the US wants the yuan to be allowed further in value.

---

**A clarification: definition of “BoW”**

**Looser definition**

- Independent features

**Stricter definition**

- Independent features
- Histogram representation

---

**Visual words: main idea**

- Extract some local features from a number of images ...
- e.g., SIFT descriptor space: each point is 128-dimensional

---

**Visual words: main idea**
Visual words: main idea

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

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

Visual words

Example: each group of patches belongs to the same visual word

Figure from Sivic & Zisserman, ICCV 2003

Slide credit: J. Sivic

Slide credit: D. Nister

K. Grauman, B. Leibe

Slide credit: D. Nister
Inverted file index for images comprised of visual words

- Score each image by the number of common visual words (tentative correspondences)
- But: does not take into account spatial layout of regions

Clustering / quantization methods

- k-means (typical choice), agglomerative clustering, mean-shift,
- Hierarchical clustering: allows faster insertion / word assignment while still allowing large vocabularies
  - Vocabulary tree [Nister & Stewenius, CVPR 2006]

“Bag of visual words”

Scene Interpretation based on “Global” Image Statistics

Spectral Signatures

Why are Fields, Beaches and Coasts less isotropic than other natural environments?

Scene Scale

“The point of view that any given observer adopts on a specific scene is constrained by the volume of the scene.”
How does the amount of clutter vary against scene scale in man-made environments? In natural environments?
Spatially Localized Statistics

- Windowed FFT

Examples (man-made)

Examples (Natural)

- Image statistics become non-stationary as scene scale increases.