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

- Illumination
- Object pose
- Clutter
- Occlusions
- Intra-class appearance
- Viewpoint

Challenges: robustness

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

K. Grauman, B. Leibe
Challenges: context and human experience

Challenges: learning with minimal supervision

Find the pottopod

Rough evolution of focus in recognition research
Inputs/outputs/assumptions

- What is the **goal**?
  - Say yes/no as to whether an object present in image
  - And/or:
  - 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

---

Detection via classification: Main idea

Basic component: a binary classifier

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

Fleshing out this pipeline a bit more, we need to:
1. Obtain training data
2. Define features
3. Define classifier

---

Scanning windows...
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
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.
Partial derivatives of an image

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

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

\[
\begin{array}{cc}
-1 & 1 \\
\end{array}
\]

or

\[
\begin{array}{cc}
1 & -1 \\
\end{array}
\]

Assorted finite difference filters

\[
Prewitt: \quad M_x = \begin{bmatrix}
1 & 0 & -1
\end{bmatrix} \quad ; \quad M_y = \begin{bmatrix}
-1 & 0 & 1
\end{bmatrix}
\]

\[
Sobel: \quad M_x = \begin{bmatrix}
1 & 2 & 1
\end{bmatrix} \quad ; \quad M_y = \begin{bmatrix}
-1 & -2 & -1
\end{bmatrix}
\]

\[
Roberts: \quad M_x = \begin{bmatrix}
1 & 0
\end{bmatrix} \quad ; \quad M_y = \begin{bmatrix}
0 & -1
\end{bmatrix}
\]

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

\[
\text{imagesc(outim)} \quad ; \quad \text{colormap gray}
\]

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

Gradient-based representations: Histograms of oriented gradients (HoG)

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

Dalal & Triggs, CVPR 2005

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

HOG
- Tested with
  - RGB
  - LAB
  - Grayscale
- Gamma Normalization and Compression
  - Square root
  - Log

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

- Sobel

\[
\begin{align*}
\text{centered} & : & 1 & 0 & 1 \\
\text{diagonal} & : & 0 & 1 & -1 \\
\text{uncentered} & : & 1 & 4 & 0 & 8 & -1 \\
\text{cubic-corrected} & : & 1 & 0 & 0 & 1 & -1 & 0 & 2 & 0 & 2 & -1 & 0 & 1
\end{align*}
\]
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 Algorithm**

Start with uniform weights on training examples

\[
\text{For } T \text{ rounds}
\]

- 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 classifiers

\[
\text{Final classifier is } \sum_{t=1}^{T} \alpha_t h_t(x)
\]

where \( \alpha_t = \log \frac{1}{Z_t} \)

**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

---

*Freund & Schapire 1995*
Feature extraction

"Rectangular" filters

- Feature output is difference between adjacent regions
- Efficiently computable with integral image: any sum can be computed in constant time
- Avoid scaling images ➔ scale features directly for same cost

Value at \((x, y)\) is sum of pixels above and to the left of \((x, y)\)

Integral Image

\[ D_{x,y} = \sum_{x' < x, y' < y} I(x', y') \]

\[ D_{x,y} = \sum_{x' < x} \sum_{y' < y} I(x', y') \]

\[ D_{x,y} = \sum_{y' < y} \sum_{x' < x} I(x', y') \]

\[ D_{x,y} = \sum_{x' < x} I(x', y) \]

\[ D_{x,y} = \sum_{y' < y} I(x, y') \]

\[ D_{x,y} = 0 \]

\[ D_{x,y} = D_{x-1,y} + I(x, y) \]

\[ D_{x,y} = D_{x-1,y} - D_{x-1,y-1} + I(x, y) \]

\[ D_{x,y} = D_{x-1,y-1} + I(x, y) \]

\[ D_{x,y} = D_{x-1,y-1} - D_{x-1,y} + I(x, y) \]

\[ D_{x,y} = D_{x-1,y} - D_{x-1,y-1} + I(x, y) \]

Viola & Jones, CVPR 2001

Large library of filters

- Considering 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

Viola & Jones, CVPR 2001

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.

Outputs of a possible rectangle feature on faces and non-faces.

Resulting weak classifier:

\[ h_i(x) = \begin{cases} +1 & \text{if } f_i(x) > 0, \\ -1 & \text{otherwise} \end{cases} \]

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

AdaBoost for Efficient Feature Selection

- Image Features ➔ Weak Classifiers
- For each round of boosting:
  - Evaluate each rectangle filter 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

Viola and Jones, Robust object detection using a boosted cascade of simple features, CVPR 2001
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/]

Viola-Jones Face Detector: Results

First two features selected

Viola-Jones Face Detector: Results

Viola-Jones Face Detector: Results

Viola-Jones Face Detector: Results
Detecting profile faces?

Detecting profile faces requires training separate detector with profile examples.

Example application

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

Example application: faces in photos

Everingham, M., Sivic, J. and Zisserman, A. “Hello! My name is... Buffy” - Automatic naming of characters in TV video, BMVC 2006.

http://www.robots.ox.ac.uk/~vgg/research/infacer/index.html

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

Limitations (continued)

• 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

Context can constrain a sliding window search

Limitations (continued)

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

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

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’
China is forecasting a trade surplus of $90bn ($51bn to $72bn this year, a forecasted increase on 2004's $32bn). The Commerce Ministry defines China's trade surplus as exports to the US of $275bn, compared with a predicted 30% jump in exports to $750bn, overtaking China's imports to the US by $660bn. The figures are likely to further demonstrate the Chinese government's deliberate undervalued yuan. Beijing wants the yuan to be allowed to trade among other currencies, but now, according to Hubel and Weisel, the system of nerve cells in the cortex is wired to be sensitive to the visual nature of things. The cerebral cortex is the part of the brain that is responsible for higher brain functions such as sensory, visual, and cognitive processing. It receives information from the retinal, which is the outermost layer of the eyeball and is responsible for processing visual data. The visual cortex, located in the occipital lobe of the brain, is responsible for processing visual information and sending it to the sensory cortex for further processing. The visual cortex is divided into several areas, each responsible for processing different aspects of visual information. The visual cortex is also connected to other areas of the brain, such as the premotor cortex, which is responsible for planning and controlling movements.
Visual words: main idea

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

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

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

Today

- Scanning window paradigm
- HOG
- Boosted Face Detection
- BOW Indexing
- GIST

Slide Credits

- As attributed...