CSE 152: Computer Vision
Manmohan Chandraker

Lecture 10: Recognition
Shape and Motion Estimation
Autonomous driving

Where is our car?
Structure from Motion
Visual SLAM

Where are other agents?
Object detection
3D localization

What is a safe path?
Behavior prediction
Path planning

Where are scene elements?
Semantic segmentation
Correspondence is one of the most fundamental problems in computer vision. Solving correspondence well goes a long way towards solving many problems. Examples include 3D reconstruction, object retrieval, action recognition.
**Scale Invariant Feature Transform**

Basic idea:

- Take 16x16 square window around detected feature
- Compute edge orientation (angle of the gradient - 90°) for each pixel
- Throw out weak edges (threshold gradient magnitude)
- Create histogram of surviving edge orientations

Adapted from slide by David Lowe
Modeling projection

- A matrix multiplication using homogeneous coordinates

\[
\Pi = K \begin{bmatrix}
1 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 1
\end{bmatrix}
\begin{bmatrix}
R & 0 & c \\
0 & 0 & c \\
0 & 0 & c \\
0 & 0 & 0 & 1
\end{bmatrix}
\]
Fundamental Matrix

\[ x_1 \leftrightarrow x_2 \]

\[ x_1^T F x_2 = 0 \]
RANSAC

1. Randomly choose $s$ samples (correspondences)
   – For fundamental matrix, what is the size of $s$?

2. Fit the model (fundamental matrix) to those samples
   – How would you fit the model to $s$ correspondences?

3. Count the number of inliers among all other correspondences
   – How can you determine which points are inliers?

4. Repeat $N$ times
   – How do you choose $N$?

5. Choose the model with the largest set of inliers
Triangulation to get 3D points

$R_1, t_1$

$R_2, t_2$
Structure from motion

\[ \prod_1 X_1 \sim p_{11} \]

minimize

\[ g(R, T, X) \]

non-linear least squares
Your basic stereo algorithm

For each epipolar line
  For each pixel in the left image
    • compare with every pixel on same epipolar line in right image
    • pick pixel with minimum match cost

Improvement: match *windows*
Stereo Matching Framework

- For every disparity, compute matching costs
  \[ E_0(x, y; d) = \rho(I_L(x' + d, y') - I_R(x', y')) \]

- Obtain disparity space image
  \[ E(x, y; d) = \sum_{(x',y') \in N(x,y)} E_0(x', y', d) \]

- Choose winning disparity at each pixel
  \[ d(x, y) = \arg \min_d E(x, y; d) \]

- Interpolate to sub-pixel accuracy.
Multiview stereo
Recognition

[Fei-Fei Li, Rob Fergus and Antonio Torralba]
Autonomous driving

Where is our car?
- Structure from Motion
- Visual SLAM

Where are other agents?
- Object detection
- 3D localization

What is a safe path?
- Behavior prediction
- Path planning

Where are scene elements?
- Semantic segmentation
So what does object recognition involve?
Verification: is that a bus?
Detection: are there cars?
Identification: is that a picture of Mao?
Object categorization

- sky
- building
- flag
- banner
- face
- street lamp
- wall
- bus
- cars
Scene and context categorization

• outdoor
• city
• traffic
• ...

[Image of a busy city street with traditional architecture and vehicles]
A Few Challenges in Computer Vision
Why is computer vision difficult?

- Viewpoint
- Lighting
- Scale
- Deformation
Why is computer vision difficult?

Intra-class variation

Background clutter

Motion (Source: S. Lazebnik)

Occlusion
Challenges: intra-class variation

Can you design an algorithm to describe a chair?
Object categorization: the statistical viewpoint

- Bayes rule:

\[
\frac{p(\text{zebra} \mid \text{image})}{p(\text{no zebra} \mid \text{image})} = \frac{p(\text{image} \mid \text{zebra})}{p(\text{image} \mid \text{no zebra})} \cdot \frac{p(\text{zebra})}{p(\text{no zebra})}
\]

- posterior ratio
- likelihood ratio
- prior ratio
Object categorization: the statistical viewpoint

\[ \frac{p(\text{zebra} \mid \text{image})}{p(\text{no zebra} \mid \text{image})} = \frac{p(\text{image} \mid \text{zebra})}{p(\text{image} \mid \text{no zebra})} \]

- Discriminative methods model posterior
- Generative methods model likelihood and prior
Discriminative

- Direct modeling of \[ \frac{p(\text{zebra} \mid \text{image})}{p(\text{no zebra} \mid \text{image})} \]
Generative

- Model $p(\text{image} \mid \text{zebra})$ and $p(\text{image} \mid \text{no zebra})$

<table>
<thead>
<tr>
<th>$p(\text{image} \mid \text{zebra})$</th>
<th>$p(\text{image} \mid \text{no zebra})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>Middle</td>
</tr>
<tr>
<td>High</td>
<td>Middle $\rightarrow$ Low</td>
</tr>
</tbody>
</table>
Three main issues

• Representation
  – How to represent an object category

• Learning
  – How to form the classifier, given training data

• Inference
  – How the classifier is to be used on novel data
Representation

– Generative or discriminative or hybrid
Representation

– Generative or discriminative or hybrid
– Appearance only or location and appearance
Representation

- Generative or discriminative or hybrid
- Appearance only or location and appearance
- Invariances
  - View point
  - Illumination
  - Occlusion
  - Scale
  - Deformation
  - Clutter
  - etc.
Representation

- Generative or discriminative or hybrid
- Appearance only or location and appearance
- Invariances
- Part-based or global
Representation

- Generative or discriminative or hybrid
- Appearance only or location and appearance
- Invariances
- Part-based or global
- Use set of features or each pixel in image
Learning

- Unclear how to model categories, so we learn what distinguishes them rather than manually specify the difference
  - Thus, the current interest in machine learning
Learning

- Unclear how to model categories, so we learn what distinguishes them rather than manually specify the difference
  - Thus, the current interest in machine learning
- Method of training: generative or discriminative
  - What are you maximizing? Likelihood (generative) or performances on train and validation set (discriminative)
Learning

- Unclear how to model categories, so we learn what distinguishes them rather than manually specify the difference
  - Thus, the current interest in machine learning
- Method of training: generative or discriminative
  - What are you maximizing? Likelihood (generative) or performances on train and validation set (discriminative)
- Level of supervision
  - Image label, bounding box, object boundaries, semantic parts

Contains a motorbike
Learning

- Unclear how to model categories, so we learn what distinguishes them rather than manually specify the difference
  - Thus, the current interest in machine learning
- Method of training: generative or discriminative
  - What are you maximizing? Likelihood (generative) or performances on train and validation set (discriminative)
- Level of supervision
  - Image label, bounding box, object boundaries, semantic parts
- Batch or incremental (on category and image level, user-feedback)
Learning

– Unclear how to model categories, so we learn what distinguishes them rather than manually specify the difference
  • Thus, the current interest in machine learning
– Method of training: generative or discriminative
  • What are you maximizing? Likelihood (generative) or performances on train and validation set (discriminative)
– Level of supervision
  • Image label, bounding box, object boundaries, semantic parts
– Batch or incremental (on category and image level, user-feedback )
– Training images:
  • Issue of overfitting
Learning

– Unclear how to model categories, so we learn what distinguishes them rather than manually specify the difference
  • Thus, the current interest in machine learning
– Method of training: generative or discriminative
  • What are you maximizing? Likelihood (generative) or performances on train and validation set (discriminative)
– Level of supervision
  • Image label, bounding box, object boundaries, semantic parts
– Batch or incremental (on category and image level, user-feedback)
– Training images:
  • Issue of overfitting
– Priors
Inference

– Scale or orientation range to search
– Speed
Instance Recognition

[David Nister, Kristen Grauman, Noah Snavely]
Recognizing or retrieving specific objects

Example 1: Place recognition for self-driving or robot navigation
Recognizing or retrieving specific objects

Example 2: Visual search in movies

Visually defined query

“Find this clock”

“Find this place”

“Groundhog Day” [Rammis, 1993]
Recognizing or retrieving specific objects

Example 3: Search photos for particular places

Find these landmarks ...in these images and 1M more
Why is it difficult?

Want to find the object despite possibly large changes in scale, viewpoint, lighting and partial occlusion.

- **Scale**
- **Viewpoint**
- **Lighting**
- **Occlusion**

Slide credit: J. Sivic
Object → Bag of ‘words’
Origin: Bag-of-words models

- Orderless document representation: frequencies of words from a dictionary  
  Salton & McGill (1983)

US Presidential Speeches Tag Cloud
http://chir.ag/phernalia/preztags/
Origin: Bag-of-words models

- Orderless document representation: frequencies of words from a dictionary
  Salton & McGill (1983)

US Presidential Speeches Tag Cloud
http://chir.ag/phernalia/preztags/
Origin: Bag-of-words models

- Orderless document representation: frequencies of words from a dictionary  
  Salton & McGill (1983)
Bags of features for object recognition

- Works quite well for image-level classification and for recognizing object *instances*
## Bags of features for object recognition

### Caltech6 dataset

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>airplanes</td>
<td>98.8</td>
<td>97.1</td>
<td>90.2</td>
</tr>
<tr>
<td>cars (rear)</td>
<td>98.3</td>
<td>98.6</td>
<td>90.3</td>
</tr>
<tr>
<td>cars (side)</td>
<td>95.0</td>
<td>87.3</td>
<td>88.5</td>
</tr>
<tr>
<td>faces</td>
<td>100</td>
<td>99.3</td>
<td>96.4</td>
</tr>
<tr>
<td>motorbikes</td>
<td>98.5</td>
<td>98.0</td>
<td>92.5</td>
</tr>
<tr>
<td>spotted cats</td>
<td>97.0</td>
<td>—</td>
<td>90.0</td>
</tr>
</tbody>
</table>
Recall: matching local features

To generate **candidate matches**, find patches that have the most similar appearance (e.g., lowest SSD)

Simplest approach: compare them all, take the closest (or closest k, or within a thresholded distance)
Matching two given views for depth

Search for a matching view for recognition

[Kristen Grauman]
Indexing local features

[Kristen Grauman]
Bag of features: outline

1. Extract features
Bag of features: outline

1. Extract features
2. Learn “visual vocabulary”
Bag of features: outline

1. Extract features
2. Learn “visual vocabulary”
3. Quantize features using visual vocabulary
Bag of features: outline

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

Regular grid
- Vogel & Schiele, 2003
- Fei-Fei & Perona, 2005
1. Feature extraction

Regular grid
- Vogel & Schiele, 2003
- Fei-Fei & Perona, 2005

Interest point detector
- Csurka et al. 2004
- Fei-Fei & Perona, 2005
- Sivic et al. 2005
1. Feature extraction

Regular grid
  • Vogel & Schiele, 2003
  • Fei-Fei & Perona, 2005

Interest point detector
  • Csurka et al. 2004
  • Fei-Fei & Perona, 2005
  • Sivic et al. 2005

Other methods
  • Random sampling (Vidal-Naquet & Ullman, 2002)
  • Segmentation-based patches (Barnard et al. 2003)
Indexing local features

• Each patch or region has a descriptor, which is a point in some high-dimensional feature space (for example, SIFT)
Indexing local features

- When we see close points in feature space, we have similar descriptors, which indicates similar local content.
Visual words: main idea

• Extract some local features from a number of images …

Example: For SIFT descriptor space, each point is 128-dimensional

D. Nister, CVPR 2006
Visual words: main idea
Visual words: main idea
Visual words: main idea
Each point is a local descriptor, e.g. SIFT vector.
2. Learning the visual vocabulary
2. Learning the visual vocabulary

Slide credit: Josef Sivic
2. Learning the visual vocabulary

Visual vocabulary

Clustering
K-means clustering

• Want to minimize sum of squared Euclidean distances between points $x_i$ and their nearest cluster centers $c_k$

$$D(X, C) = \sum_{\text{cluster } C_k} \sum_{\text{point } i \in C_k} (x_i - c_k)^2$$

Algorithm:

• Randomly initialize K cluster centers
• Iterate until convergence:
  • Assign each data point to the nearest center
  • Recompute each cluster center as the mean of all points assigned to it
From clustering to 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
  - Codevector = visual word
Example: each group of patches belongs to the same visual word.

Figure from Sivic & Zisserman, ICCV 2003
Visual words

Sivic et al. 2005
3. Image representation

![Image](image.png)

**frequency**

**codewords***

......
4. Image classification

- Given the bag-of-features representations of images from different classes, how do we learn a model for distinguishing them?
Comparing bags of words

- Rank frames by normalized scalar product between their (possibly weighted) occurrence counts—*nearest neighbor* search for similar images.

\[
sim(d_j, q) = \frac{\langle d_j, q \rangle}{\|d_j\| \|q\|}
\]

\[
= \frac{\sum_{i=1}^{V} d_j(i) \times q(i)}{\sqrt{\sum_{i=1}^{V} d_j(i)^2} \times \sqrt{\sum_{i=1}^{V} q(i)^2}}
\]

for vocabulary of \( V \) words

\([Kristen Grauman]\)
Image classification using BoW

- Treat as feature vector for a standard classifier
  - For example, support vector machine