Instance recognition

Thurs Oct 29

Last time

• Depth from stereo: main idea is to triangulate from corresponding image points.
• Epipolar geometry defined by two cameras
  – We’ve assumed known extrinsic parameters relating their poses
• Epipolar constraint limits where points from one view will be imaged in the other
  – Makes search for correspondences quicker
• To estimate depth
  – Limit search by epipolar constraint
  – Compute correspondences, incorporate matching preferences
Virtual viewpoint video


http://research.microsoft.com/IVM/VVV/
Review questions:
What stereo rig yielded these epipolar lines?

Figure from Hartley & Zisserman

Epipole has same coordinates in both images.
Points move along lines radiating from e: “Focus of expansion”

Review questions

- When solving for stereo, when is it necessary to break the soft disparity gradient constraint?
- What can cause a disparity value to be undefined?
- What parameters relating the two cameras in the stereo rig must be known (or inferred) to compute depth?
Today

- Instance recognition
  - Indexing local features efficiently
  - Spatial verification models

Recognizing or retrieving specific objects

Example I: Visual search in feature films

Visually defined query

“Find this clock”

“Find this place”

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

Example II: Search photos on the web for particular places

Find these landmarks ...in these images and 1M more

Slide credit: J. Sivic
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)

---

Multi-view matching

*Matching two given views for depth*  
*Search for a matching view for recognition*
Indexing local features

- Each patch / region has a descriptor, which is a point in some high-dimensional feature space (e.g., SIFT)
Indexing local features

• When we see close points in feature space, we have similar descriptors, which indicates similar local content.

Indexing local features

• With potentially thousands of features per image, and hundreds to millions of images to search, how to efficiently find those that are relevant to a new image?

• Possible solutions:
  - Inverted file
  - Nearest neighbor data structures
    • Kd-trees
    • Hashing
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”.

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.
Visual words: main idea

• Extract some local features from a number of images …

e.g., SIFT descriptor space: each point is 128-dimensional

Slide credit: D. Nister, CVPR 2006
Visual words: main idea

Visual words: main idea
Each point is a local descriptor, e.g. SIFT vector.
Visual words

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

Figure from Sivic & Zisserman, ICCV 2003

Visual words

• Also used for describing scenes and object categories for the sake of indexing or classification.

Sivic & Zisserman 2003; Csurka, Bray, Dance, & Fan 2004; many others.
Visual words and textons

- First explored for texture and material representations
- **Texton** = cluster center of filter responses over collection of images
- Describe textures and materials based on distribution of prototypical texture elements.

Leung & Malik 1999; Varma & Zisserman, 2002

Recall: Texture representation example

<table>
<thead>
<tr>
<th>Window</th>
<th>( \text{mean } \frac{d}{dx} \text{ value} )</th>
<th>( \text{mean } \frac{d}{dy} \text{ value} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Win. #1</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>Win. #2</td>
<td>18</td>
<td>7</td>
</tr>
<tr>
<td>Win. #9</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
</tbody>
</table>

Statistics to summarize patterns in small windows
Visual vocabulary formation

Issues:
- Sampling strategy: where to extract features?
- Clustering / quantization algorithm
- Unsupervised vs. supervised
- What corpus provides features (universal vocabulary?)
- Vocabulary size, number of words

Inverted file index

- Database images are loaded into the index mapping words to image numbers
Inverted file index

When will this give us a significant gain in efficiency?

New query image

- New query image is mapped to indices of database images that share a word.

<table>
<thead>
<tr>
<th>Word #</th>
<th>Image #</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1, 2</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>91</td>
<td>2</td>
</tr>
</tbody>
</table>

Instance recognition: remaining issues

- How to summarize the content of an entire image? And gauge overall similarity?
- How large should the vocabulary be? How to perform quantization efficiently?
- Is having the same set of visual words enough to identify the object/scene? How to verify spatial agreement?
- How to score the retrieval results?

Kristen Grauman
Of all the sensory impressions proceeding to the brain, the visual experiences are the dominant ones. Our perception of the world around us is based essentially on the messages that proceed from our eyes. For a long time it was thought that the retinal image was transmitted point by point to visual centers in the brain; the cerebral cortex was a movie screen, so to speak, upon which the image in the eye was projected. Through the discoveries of Hubel and Wiesel we now know that behind the origin of the visual perception in the brain there is a considerably more complicated course of events. By following the visual impulses along their path to the various cell layers of the optical cortex, Hubel and Wiesel have demonstrated that the message about the image falling on the retina undergoes a step-wise analysis in a system of nerve cells stored in columns. In this system each cell has its specific function and is responsible for a specific detail in the pattern of the retinal image.

China is forecasting a trade surplus of $90bn (£51bn) to $100bn this year, a threefold increase on 2004's $32bn. The Commerce Ministry said the surplus would be created by a predicted 30% jump in exports to $750bn, compared with an 18% rise in imports to $660bn. The figures are likely to further annoy the US, which has long argued that China's exports are unfairly helped by a deliberately undervalued yuan. Beijing agrees the surplus is too high, but says the yuan is only one factor. Bank of China governor Zhou Xiaochuan said the country also needed to do more to boost domestic demand so more goods stayed within the country. China increased the value of the yuan against the dollar by 2.1% in July and permitted it to trade within a narrow band, but the US wants the yuan to be allowed to trade freely. However, Beijing has made it clear it will take its time and tread carefully before allowing the yuan to rise further in value.

Object → Bag of ‘words’
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.
Comparing bags of words

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

\[
\text{sim}(d_j, q) = \frac{\langle d_j, q \rangle}{\|d_j\| \|q\|} = \frac{\sum_{i=1}^{V} d_j(i) \cdot q(i)}{\sqrt{\sum_{i=1}^{V} d_j(i)^2} \sqrt{\sum_{i=1}^{V} q(i)^2}}
\]

for vocabulary of \( V \) words

\( d_j \)  \( q \)

\[ \begin{bmatrix} 1 & 8 & 1 & 4 \end{bmatrix} \]

\[ \begin{bmatrix} 5 & 1 & 1 & 0 \end{bmatrix} \]

\( \text{tf-idf weighting} \)

- Term frequency – inverse document frequency
- Describe frame by frequency of each word within it, downweight words that appear often in the database
- (Standard weighting for text retrieval)

\[
t_i = \frac{n_{id}}{n_d} \log \frac{N}{n_i}
\]

Number of occurrences of word \( i \) in document \( d \)

Number of words in document \( d \)

Total number of documents in database

Number of documents word \( i \) occurs in, in whole database
Inverted file index and bags of words similarity

1. Extract words in query
2. Inverted file index to find relevant frames
3. Compare word counts

Kristen Grauman

Bags of words for content-based image retrieval

Visually defined query

“Find this clock”

“Find this place”

“Groundhog Day” [Rammis, 1993]

Slide from Andrew Zisserman
Sivic & Zisserman, ICCV 2003
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

• Demo online at:
  http://www.robots.ox.ac.uk/~vgg/research/vgoogle/index.html
Vocabulary Trees: hierarchical clustering for large vocabularies

- Tree construction:

Vocabulary Tree

- Training: Filling the tree
Vocabulary Tree

- Training: Filling the tree

[Nister & Stewenius, CVPR'06]

K. Grauman, B. Leibe

Slide credit: David Nister

Vocabulary Tree

- Training: Filling the tree

[Nister & Stewenius, CVPR'06]

K. Grauman, B. Leibe

Slide credit: David Nister
What is the computational advantage of the hierarchical representation bag of words, vs. a flat vocabulary?

Vocabulary size

Results for recognition task with 6347 images

Influence on performance, sparsity?  
Nister & Stewenius, CVPR 2006
Bags of words: pros and cons

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

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

Instance recognition: remaining issues

• How to summarize the content of an entire image? And gauge overall similarity?
• How large should the vocabulary be? How to perform quantization efficiently?
• Is having the same set of visual words enough to identify the object/scene? How to verify spatial agreement?
• How to score the retrieval results?
Which matches better?

Spatial Verification

Both image pairs have many visual words in common.
Spatial Verification

Only some of the matches are mutually consistent

Spatial Verification: two basic strategies

• RANSAC
  – Typically sort by BoW similarity as initial filter
  – Verify by checking support (inliers) for possible transformations
    • e.g., “success” if find a transformation with > N inlier correspondences

• Generalized Hough Transform
  – Let each matched feature cast a vote on location, scale, orientation of the model object
  – Verify parameters with enough votes
RANSAC verification

Recall: Fitting an affine transformation

Approximates viewpoint changes for roughly planar objects and roughly orthographic cameras.

\[
\begin{bmatrix}
    x'_i \\
    y'_i
\end{bmatrix} =
\begin{bmatrix}
    m_1 & m_2 \\
    m_3 & m_4
\end{bmatrix}
\begin{bmatrix}
    x_i \\
    y_i
\end{bmatrix} +
\begin{bmatrix}
    t_1 \\
    t_2
\end{bmatrix}
\begin{bmatrix}
    \cdots \\
    \cdots \\
    \cdots \\
    \cdots \\
    \cdots \\
    \cdots \\
    \cdots
\end{bmatrix}
\begin{bmatrix}
    0 & 0 & 0 & 0 & 1 & 0 \\
    0 & 0 & x_i & y_i & 0 & 1 \\
    \cdots \\
    \cdots \\
    \cdots
\end{bmatrix}
\begin{bmatrix}
    x'_i \\
    y'_i
\end{bmatrix} =
\begin{bmatrix}
    m_1 \\
    m_2 \\
    m_3 \\
    m_4 \\
    t_1 \\
    t_2
\end{bmatrix}
\begin{bmatrix}
    \cdots \\
    \cdots \\
    \cdots \\
    \cdots \\
    \cdots
\end{bmatrix}
RANSAC verification

Spatial Verification: two basic strategies

• RANSAC
  – Typically sort by BoW similarity as initial filter
  – Verify by checking support (inliers) for possible transformations
    • e.g., “success” if find a transformation with > N inlier correspondences

• Generalized Hough Transform
  – Let each matched feature cast a vote on location, scale, orientation of the model object
  – Verify parameters with enough votes
Voting: Generalized Hough Transform

- If we use scale, rotation, and translation invariant local features, then each feature match gives an alignment hypothesis (for scale, translation, and orientation of model in image).

Voting: Generalized Hough Transform

- A hypothesis generated by a single match may be unreliable,
- So let each match vote for a hypothesis in Hough space
Gen Hough Transform details (Lowe’s system)

- **Training phase:** For each model feature, record 2D location, scale, and orientation of model (relative to normalized feature frame)
- **Test phase:** Let each match between a test SIFT feature and a model feature vote in a 4D Hough space
  - Use broad bin sizes of 30 degrees for orientation, a factor of 2 for scale, and 0.25 times image size for location
  - Vote for two closest bins in each dimension
- Find all bins with at least three votes and perform geometric verification
  - Estimate least squares affine transformation
  - Search for additional features that agree with the alignment


Example result

Background subtract for model boundaries

Objects recognized

Recognition in spite of occlusion

[Lowe]
Recall: difficulties of voting

- Noise/clutter can lead to as many votes as true target
- Bin size for the accumulator array must be chosen carefully
- In practice, good idea to make broad bins and spread votes to nearby bins, since verification stage can prune bad vote peaks.

Example Applications

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

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

<table>
<thead>
<tr>
<th>Query</th>
<th>Results from 5k Flickr images (demo available for 100k set)</th>
</tr>
</thead>
</table>

[Philbin CVPR'07]

### Web Demo: Movie Poster Recognition

- **50,000 movie posters indexed**
- **Query-by-image from mobile phone available in Switzerland**

Instance recognition: remaining issues

• How to summarize the content of an entire image? And gauge overall similarity?

• How large should the vocabulary be? How to perform quantization efficiently?

• Is having the same set of visual words enough to identify the object/scene? How to verify spatial agreement?

• How to score the retrieval results?

Scoring retrieval quality

Query

Database size: 10 images
Relevant (total): 5 images

Results (ordered):

\[ \text{precision} = \frac{\text{relevant}}{\text{returned}} \]
\[ \text{recall} = \frac{\text{relevant}}{\text{total relevant}} \]
Recognition via alignment

**Pros:**

- Effective when we are able to find reliable features within clutter
- Great results for matching specific instances

**Cons:**

- Scaling with number of models
- Spatial verification as post-processing – not seamless, expensive for large-scale problems
- Not suited for category recognition.

---

China is forecasting a trade surplus of $90bn (£51bn) to $100bn this year, a threefold increase on 2004’s $32bn. The Commerce Ministry said the surplus would be created by a predicted 30% jump in exports to $750bn, compared with a 18% rise in imports to $660bn. The figures are likely to further annoy the US, which has long argued that China’s exports are unfairly helped by a deliberately undervalued yuan. Beijing agrees the surplus is too high, but says the yuan is only one factor. Bank of China governor Zhou Xiaochuan said the country also needed to do more to boost domestic demand so more goods stayed within the country. China increased the value of the yuan against the dollar by 2.1% in July and permitted it to trade within a narrow band, but the US wants the yuan to be allowed to rise freely. However, Beijing has made it clear that it will take its time and tread carefully before allowing the yuan to rise further in value.
Query expansion

Query: **golf green**

Results:

- How can the grass on the **greens** at a **golf** course be so perfect?
- For example, a skilled **golfer** expects to reach the **green** on a par-four hole in...
- Manufactures and sells synthetic **golf putting greens** and mats.

Irrelevant result can cause a ‘topic drift’:

Query Expansion Step by Step

Query Image

Retrieved image

Originally not retrieved

Slide credit: Ondrej Chum
Query Expansion Step by Step

Query Expansion Results

Slide credit: Ondrej Chum
Summary

• **Matching local invariant features**
  – Useful not only to provide matches for multi-view geometry, but also to find objects and scenes.

• **Bag of words** representation: quantize feature space to make discrete set of visual words
  – Summarize image by distribution of words
  – Index individual words

• **Inverted index**: pre-compute index to enable faster search at query time

• **Recognition of instances via alignment**: matching local features followed by spatial verification
  – Robust fitting: RANSAC, GHT

Kristen Grauman