Announcements

• PS2 due next Monday 11:59 am

Recap: Features and filters

Transforming and describing images; textures, edges
Recap: Grouping & fitting

Clustering, segmentation, fitting; what parts belong together?

Recognition and learning

Matching local features
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).

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?

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”.

### Database images

### Query image

### Descriptor’s feature space
Text retrieval vs. image search

- What makes the problems similar, different?

- Text words are discrete "tokens", whereas local image descriptors are high-dimensional, real-valued feature points.
- Need to quantize the visual features into discrete visual words.

Visual words: main idea

- Extract some local features from a number of images …

 e.g., SIFT descriptor space: each point is 128-dimensional
Each point is a local descriptor, e.g. SIFT vector.
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.

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

Windows with primarily horizontal edges
Windows with small gradient in both directions
Windows with primarily vertical edges
Both

<table>
<thead>
<tr>
<th>Dimension 1 (mean d/dx value)</th>
<th>Dimension 2 (mean d/dy value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Win. #1 4</td>
<td>10</td>
</tr>
<tr>
<td>Win. #2 18</td>
<td>7</td>
</tr>
<tr>
<td>Win. #9 20</td>
<td>20</td>
</tr>
</tbody>
</table>

Issues:
- Sampling strategy: where in image to extract features?
- Clustering / quantization algorithm?
- What corpus/dataset provides features (universal vocabulary)?
- Vocabulary size (number of words)?
Inverted file index

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

<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>9</td>
<td>3</td>
</tr>
<tr>
<td>91</td>
<td>2</td>
</tr>
</tbody>
</table>

Kristen Grauman

When will this give us a significant gain in efficiency?

- 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>9</td>
<td>3</td>
</tr>
<tr>
<td>91</td>
<td>2</td>
</tr>
</tbody>
</table>

Kristen Grauman

- If a local image region is a visual word, how can we summarize an image (the document)?
Analogy to documents

- All the sensory impressions reaching the brain, the visual experiences are the dominant ones, for perception is a visual one, and it is based mostly on visual memories.

- Through a sensory input, nerve impulses are transmitted to the visual area in the brain, which undergoes a step-by-step analysis in a stepwise split-and-split column. This column has a specific function and is responsible for detailed visual patterns in the retina image.

- China is forecasting a trade surplus of $118bn to $120bn this year, a threshold increase of 20% to 22%. The International Monetary Fund predicted a 5% jump in output, which is less than 1% rise in the US. China has the world’s biggest and fastest growing economy, but it’s more unknown that value of the yuan’s appreciation.

- Analyzed and permitted by coolness, but the US wants $90bn a year to be allowed to trade freely. However, Beijing has made it clear that it will take its time and tread carefully allowing the yuan to rise further in value.

- Bag 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 occurrence counts—nearest neighbor search for similar images.

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

for vocabulary of \( V \) words

---

Bags of words for content-based image retrieval

Visually defined query

"Groundhog Day" [Ramnis, 1993]

Find this clock

Find this place

---

Example

retrieved shots
Scoring retrieval quality

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

Results (ordered):

- Precision = Relevant / Returned
- Recall = Relevant / Total relevant

Vocabulary Trees: hierarchical clustering for large vocabularies

- Tree construction:

[Vocabulary Trees diagram]

[Nister & Stewenius, CVPR'06]

Slide credit: Ondrej Chum

What is the computational advantage of the hierarchical representation vs. a flat vocabulary?
Vocabulary Tree

• Training: Filling the tree

[Slide credit: David Nister]

[Nister & Stewenius, CVPR’06]
Vocabulary Tree

- Training: Filling the tree

Slide credit: David Nister

(Nister & Stewenius, CVPR'06)
Bags of words: pros and cons

+ flexible to geometry / deformations / viewpoint
+ compact summary of image content
+ provides vector representation for sets
+ decent 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

Summary So Far

• Matching local invariant features: useful to provide matches to find objects and scenes.
• Bag of words representation: quantize feature space to make discrete set of visual words
• Inverted index: pre-compute index to enable faster search at query time

Instance recognition

• Motivation – visual search
• Visual words
  • quantization, index, bags of words
• Spatial verification
  • affine; RANSAC, Hough
• Other text retrieval tools
  • tf-idf, query expansion
• Example applications
Instance recognition: remaining issues

- How to summarize the content of an entire image? And gauge overall similarity?
- How to score the retrieval results?
- 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?

Vocabulary size

Results for recognition task with 6347 images

Nister & Stewenius, CVPR 2006
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
Recall: Fitting an affine transformation

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

\[
\begin{bmatrix}
  x'_1 \\ y'_1 \\
  \vdots
\end{bmatrix}
= \begin{bmatrix}
  m_1 & m_2 & x_i \\ m_3 & m_4 & y_i \\
  \vdots & \vdots & \vdots
\end{bmatrix}
+ \begin{bmatrix}
  t_1 \\ t_2 \\
  \vdots
\end{bmatrix}
\]

\[
\begin{bmatrix}
  m_1 & m_2 & x_i & 0 & 0 & 1 & 0 \\
  m_3 & m_4 & y_i & 0 & 0 & 1 & 0 \\
  \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots
\end{bmatrix}
\begin{bmatrix}
  x'_1 \\ y'_1 \\
  \vdots
\end{bmatrix}
= \begin{bmatrix}
  \cdots \\
  \cdots
\end{bmatrix}
\]
**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/vgg/index.html](http://www.robots.ox.ac.uk/~vgg/research/vgg/index.html)

**Example Applications**

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

[Quack, Leibe, Van Gool, CVR'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 (acquired by Qualcomm)

What else can we borrow from text retrieval?

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)

Kristen Grauman
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

Query image

Spatial verification

New query

New results

Chum, Philbin, Sivic, Isard, Zisserman: Total Recall ..., ICCV 2007

Recognition via local-feature based alignment

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

Cons:
- Spatial verification as post-processing – not seamless, expensive for large-scale problems
- Not suited for generic category recognition
Summary

• Matching local invariant features
  – Useful 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

Questions?

See you Thursday!