6.819 / 6.869: Advances in Computer Vision

Image Retrieval:
Retrieval: Information, images, objects, large-scale

Website:
http://6.869.csail.mit.edu/fa15/

Instructor: Yusuf Aytar

Lecture TR 9:30AM – 11:00AM
(Room 34-101)
What is Image Retrieval?

User

Query

Text
Fall in Boston

Image

Speech

Retrieval Results
Applications

Art Retrieval

Medical Image Retrieval

Product Image Retrieval (Reviews, other prices etc.)

Fashion Image Retrieval
Overview

Information Retrieval
Bag of Words, TF-IDF, Cosine Similarity, Inverted Index

Object Instance Retrieval
Bag of Visual Words, Video Google, Object Instance Retrieval

Fast Object Detection/Retrieval
Fast detection, Part representations, Generalization from exemplar

Large Scale Image Search
KD-trees, Locality Sensitive Hashing, Semantic Hashing, Compact Codes
Information Retrieval
The last duel
After quarrelling over a bank loan, two men took part in the last fatal duel staged on Scottish soil. BBC News's James Landale retraces the steps of his ancestor, who made that final challenge.

West Bank water row
Palestinians have accused Israel of diverting water away from their towns in order to keep Jewish settlements in the occupied territories fully supplied. Israel denies the charge saying Palestinian farmers are to blame for using illegal connections to irrigate their fields.

A widely used document representation method
## Term Frequency (TF)

### Uncovering the trenches
I know quite a lot about these people, through documentary evidence relating to them and contemporary accounts of their times. I can identify the ship my ancestor served on in the Seven Years' War, his father's house on Holy Island and probably the...

```plaintext
Bank : 0
Loan : 1
Water : 0
Farmer : 0
```

### Morning sickness
Just a few of the opening bars are enough to transport many unsuspecting souls back to the school assembly halls of their childhood...

```plaintext
Bank : 0
Loan : 3
Water : 1
Farmer : 1
```

### Normalization

<table>
<thead>
<tr>
<th>Documents</th>
<th>Doc-1</th>
<th>Doc-2</th>
<th>Doc-3</th>
<th>Doc-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bank</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Loan</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Water</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Farmer</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Documents</th>
<th>Doc-1</th>
<th>Doc-2</th>
<th>Doc-3</th>
<th>Doc-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bank</td>
<td>0.5</td>
<td>0.33</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Loan</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>0.6</td>
</tr>
<tr>
<td>Water</td>
<td>0</td>
<td>0.66</td>
<td>1</td>
<td>0.2</td>
</tr>
<tr>
<td>Farmer</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.2</td>
</tr>
</tbody>
</table>
Inverse Document Frequency (IDF)

IDF of $i^{th}$ word: $idf_i = \log \left( \frac{n}{df_i} \right)$

The last duel
After quarrelling over a bank loan, two men took part in the last fatal duel staged on Scottish soil. BBC News's James Landale retraces the steps of his ancestor, who made that final challenge.

$\vec{v}(d_1) = tf \times idf$
Cosine Similarity

Query: **fall in Boston**

Cosine Similarity Score $= \v(q.)^T \v(d_1)$
Inverted Index

Allows quick lookup of document ids with a particular word
Bag of Words &
Object Instance Retrieval
Feature Detectors

Roles of the detector:
- provide invariance to transformations
- reduce the number of descriptors

Popular detectors:
- Maximally Stable Extremal Regions (MSER)
  Matas, Chum, Urban, Pajdla, “Robust wide-baseline stereo from maximally stable extremal regions”, BMVC’02.
- Difference of Gaussians (DoG)
- Harris-Affine and Hessian-Affine
  Mikolajczyk, Schmid, “Scale and affine invariant interest point detectors”, IJCV’04.
  → See also Mikolajczyk et al., “A comparison of affine region detectors”, IJCV’05.

Dense descriptors are also possible
- Mainly for classification → let the classifier decide
  Leung, Malik, “Representing and recognizing the visual appearance of materials using 3D textons”, IJCV’01.
- But also for image.scene/object retrieval
Feature Descriptors

Description of a patch after orientation/scale/photometric normalization

Most widely-used patch descriptor: SIFT
- 8 orientations of the gradient  → 128 dimensions
- 4x4 spatial grid

Many descriptors derive from SIFT:
- **More efficient:** SURF
  Bay, Tuytelaars, Van Gool, “SURF: speeded up robust features”, ECCV’06.
- **More compact:** CHOG, DAISY
  Tola, Lepetit, Fua, “DAISY: an efficient dense descriptor applied to wide baseline stereo”, TPAMI’10.
- **With color:** color SIFT
  Van de Weijer, Schmid, “Coloring local feature extraction”, ECCV’06.
  Burghouts and Geseborek, “Performance evaluation of local colour invariants”, CVIU’09.
Video Google

Feature Detectors / Descriptors

Harris-Affine & Hessian Affine as the feature detectors
SIFT as the feature descriptor

Video Google: A Text Retrieval Approach to Object Matching in Videos, Josef Sivic and Andrew Zisserman, ICCV 2003
Video Google
Bag of Visual Words

Images

Affine Invariant Feature Detectors
Harris-Affine & MSER

SIFT

~300K Feature Descriptors

k-means

2,237 Visual Words

A wheel of an airplane

Back of a motorbike or tip of the wings (Polysemy)

A motorbike handle

Set of SIFT descriptors

[Sivic03] sparse frequency vector

Discovering objects and their location in images, Sivic et. al., ICCV 2005
Video Google: A Text Retrieval Approach to Object Matching in Videos, Josef Sivic and Andrew Zisserman, ICCV 2003
http://www.robots.ox.ac.uk/~vgg/publications/2012/Arandjelovic12/presentation.pdf
Video Google
Large scale object instance retrieval

- Find all instances of the query object in a large scale dataset

- Do it instantly (< 1 sec), and be robust to scale, viewpoint, lighting, partial occlusion

Video Google: A Text Retrieval Approach to Object Matching in Videos, Josef Sivic and Andrew Zisserman, ICCV 2003
http://www.robots.ox.ac.uk/~vgg/publications/2012/Arandjelovic12/presentation.pdf
Video Google

Particular object retrieval - Bag of visual words

Object retrieval with large vocabularies and fast spatial matching, Philbin, Chum, Isard, Sivic, Zisserman, CVPR 2007

Video Google: A Text Retrieval Approach to Object Matching in Videos, Josef Sivic and Andrew Zisserman, ICCV 2003

http://www.robots.ox.ac.uk/~vgg/publications/2012/Arandjelovic12/presentation.pdf
Video Google

BOW + Inverted File Indexing

\[ \text{score} = \frac{q^\top x}{\|x\|} \]

Slide Credit - Chum, LSVR tutorial at CVPR'13
Spatial Verification

1. Image Query

2. Initial Retrieval Set

3. Spatial Verification

Object retrieval with large vocabularies and fast spatial matching, Philbin, Chum, Isard, Sivic, Zisserman, CVPR 2007

http://www.robots.ox.ac.uk/~vgg/publications/2012/Arandjelovic12/presentation.pdf
Query Expansion

Query image

Spatial verification

New results

New query

Results

Slide Credit - Chum, LSVR tutorial at CVPR'13
Query Expansion
Query Expansion
Video Google - Object Instance Retrieval

http://www.robots.ox.ac.uk/~vgg/demo/
Immediate, scalable object category detection
Motivation: Object Detection

• Running a detector fast on a single image (Cascades, PQ, etc.) [Felzenszwalb-CVPR10, Vedaldi-CVPR12, Sadeghi-NIPS13].

• Running multiple detectors fast on a single image (Sparselets, etc.) [Song-ECCV12, Dean-CVPR13].

• Running a detector fast (~1sec) on a large-scale image dataset, similar to Video Google [Sivic03] but for category detection.
Large Scale Object Instance Retrieval

- Retrieve instantly (< 1 sec)
- Robust to: scale, viewpoint, lighting, partial occlusion

Large Scale Object Category Detection

- Retrieve instantly (~ 1 sec)
- Robust to: scale, viewpoint, lighting, partial occlusion and Intra-class variance

[Arandjelovic-CVPR12] - query versus query
Overview

Uses the three stages of **Video Google** revamped for object category detection

Classifier Part (CP) Dictionary

1. Indexing and inverted file

Query HOG Template

Reconstructed Template

2. Shortlisting

3. Reranking
   (a) Spatial Reranking
   (b) HOG Scoring
Query Representation

Query (HOG template) is first represented as a sparse combination of CPs.

\[ \min_{\alpha} \left\| w - \sum_j \alpha_j u_j \right\|^2 + \gamma \| \alpha \|_1 \quad \text{st: } \alpha_j \geq 0 \]

\[ w^{rt} = \sum_j \alpha_j u_j \]

\[ = \alpha_1 \times d_1 + \alpha_2 \times d_2 + \alpha_3 \times d_3 \]

\( \alpha \) and the spatial layouts of CPs define the reconstructed template \( w^{rt} \).
**Image Representation:** Image $I$ is represented with vector $r = [\Psi(d_1, I) \; \Psi(d_2, I), \ldots, \Psi(d_M, I)]$, where the $i^{th}$ component is the maximum response of the CP $d_i$ sliding over $I$. 
Shortlisting: For the given template $w^{rt}$, images are shortlisted using the score $\alpha^T r$ which is an upper bound on the maximum score of $w^{rt}$ obtained by sliding over $I$.

$$\alpha^T r = \sum_j \alpha_j \Psi(d_j, I) \geq \Psi(w^{rt}, I)$$
Spatial Reranking

\[ I^k \]

\[ d_1 \rightarrow \alpha_1 \times \]

\[ d_2 \rightarrow \alpha_2 \times \]

\[ d_3 \rightarrow \alpha_3 \times \]

\[ w = \sum_j \alpha_j u_j \]

\[ s^k = \sum_j \alpha_j r_j^k \]

\[ \downarrow \text{Scoring Images} \]

\[ \downarrow \text{Spatial Reranking} \]

\[ \downarrow \text{Final Results} \]
Reranking (Spatial + Original Template)

Shortlisted images are reranked via fast Hough-like voting of bounding box candidates suggested by each CP.

Retrieved bounding box candidates are re-scored using the original HOG template with fast and memory efficient PQ compression.
Dictionary & Dataset

10K dictionaries of sizes 3x3 - 7x7 HOG cells are extracted from DPMs trained from 1000 ImageNet categories.

Tests are performed on PASCAL VOC07 test set (5K images) and validation sets (100K images) of ImageNet 2011 and 2012 challenges.
Detection Results

Top 3 retrievals

Person Template
Reconstructed Person Template

TV/Monitor Template
Reconstructed TV/Monitor Template

Cow Template
Reconstructed Cow Template

Motorbike Template
Reconstructed Motorbike Template

Immediate, scalable object category detection, Y. Aytar, A. Zisserman, CVPR 2014
Large Scale Image Search
Large Scale Image Search

- Find similar images in a large database

Fast & Accurate
Internet contains billions of images

The Challenge:
Need way of measuring similarity between images
(distance metric learning)
Needs to scale to Internet (How?)
Requirements for image search

- Search must be both **fast**, **accurate** and **scalable to large data set**
  
  - Fast
    - Kd-trees: tree data structure to improve search speed
    - Locality Sensitive Hashing: hash tables to improve search speed
    - Small code: binary small code (010101101)
  
  - Scalable
    - Require very little memory, enabling their use on standard hardware or even on handheld devices

- Accurate
  - Learned distance metric
Categorization of existing large scale image search algorithms

• Tree Based Structure
  – Spatial partitions (i.e. kd-tree) and recursive hyper plane decomposition provide an efficient means to search low-dimensional vector data exactly.

• Hashing
  – Locality-sensitive hashing offers sub-linear time search by hashing highly similar examples together.

• Binary Small Code
  – Compact binary code, with a few hundred bits per image
Tree Based Structure

- **Kd-tree**
  - The kd-tree is a binary tree in which every node is a k-dimensional point.
  - They are known to break down in practice for high-dimensional data, and cannot provide better than a worst case linear query time guarantee.
Locality Sensitive Hashing

- Take random projections of data $r^T x$
- Quantize each projection with few bits

No learning involved

Feature vector

Slide Credit - Fergus et al
How to search from hash table?

A set of data points

Hash function $h_{r_1...r_k}$

Hash table

New query $h_{r_1...r_k}$

Search the hash table for a small set of images

Results

Slide Credit - Kristen Grauman et al.
Binary codes for images

• Want images with similar content to have similar binary codes

• Use Hamming distance between codes
  – Number of bit flips
  – E.g.:  \( \text{Ham}_\text{Dist}(10001010,10001110)=1 \)
  \( \text{Ham}_\text{Dist}(10001010,11101110)=3 \)

• Semantic Hashing [Salakhutdinov & Hinton, 2007]
  – Text documents
Semantic Hashing

- Query
- Semantic Hash Function
- Binary code
- Address Space
- Images in database
- Query address
- Semantically similar images

Quite different to a (conventional) randomizing hash

- Find neighbors by exploring Hamming ball around query address
- Lookup time depends on radius of ball, NOT on # data points

Slide Credit - Fergus et al

Small Codes and Large Databases for Recognition, A. Torralba, R. Fergus, and Y. Weiss, CVPR 2008
Compact Binary Codes

- Google has few billion images ($10^9$)
- PC has ~10 Gbytes ($10^{11}$ bits)
- Codes must fit in memory (disk too slow)

Budget of $10^2$ bits/image

- 1 Megapixel image is $10^7$ bits
- 32x32 color image is $10^4$ bits

Semantic hash function must also reduce dimensionality
RBM architecture

- Network of binary stochastic units
- Hinton & Salakhutdinov, Science 2006

\[ E(v, h) = - \sum_{i \in \text{visible}} b_i v_i - \sum_{j \in \text{hidden}} b_j h_j - \sum_{i, j} v_i h_i w_{ij} \]

Parameters: Weights \( w \)  Biases \( b \)

Convenient conditional distributions:

\[ p(h_j = 1|v) = \sigma(b_j + \sum_i w_{ij} v_i) \]
\[ p(v_i = 1|h) = \sigma(b_i + \sum_j w_{ij} h_j) \]
\[ \sigma(x) = 1/(1 + e^{-x}), \text{ the logistic function} \]

Learn weights and biases using Contrastive Divergence

Hidden units: \( h \)

Symmetric weights \( w \)

Visible units: \( v \)
Examples of LabelMe retrieval

<table>
<thead>
<tr>
<th>Input image</th>
<th>Ground truth neighbors</th>
<th>L2-Pixels</th>
<th>Gist</th>
<th>32-RBM</th>
<th>16-RBM</th>
<th>8-RBM</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Input image" /></td>
<td><img src="image2" alt="Ground truth neighbors" /></td>
<td><img src="image3" alt="L2-Pixels" /></td>
<td><img src="image4" alt="Gist" /></td>
<td><img src="image5" alt="32-RBM" /></td>
<td><img src="image6" alt="16-RBM" /></td>
<td><img src="image7" alt="8-RBM" /></td>
</tr>
<tr>
<td><img src="image8" alt="Input image" /></td>
<td><img src="image9" alt="Ground truth neighbors" /></td>
<td><img src="image10" alt="L2-Pixels" /></td>
<td><img src="image11" alt="Gist" /></td>
<td><img src="image12" alt="32-RBM" /></td>
<td><img src="image13" alt="16-RBM" /></td>
<td><img src="image14" alt="8-RBM" /></td>
</tr>
<tr>
<td><img src="image15" alt="Input image" /></td>
<td><img src="image16" alt="Ground truth neighbors" /></td>
<td><img src="image17" alt="L2-Pixels" /></td>
<td><img src="image18" alt="Gist" /></td>
<td><img src="image19" alt="32-RBM" /></td>
<td><img src="image20" alt="16-RBM" /></td>
<td><img src="image21" alt="8-RBM" /></td>
</tr>
</tbody>
</table>

12 closest neighbors under different distance metrics