Large Scale Image Retrieval

Jiří Matas

Center for Machine Perception, Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University in Prague

work done as collaboration of:
Ondrej Chum, Andrej Mikulik, Filip Radenovic,
Michal Perdoch James Pritts, Dmytro Mishkin,
Johannes Schönberger, Jan-Michael Frahm, Jan Čech

Supported by the Czech Science Foundation Project GACR P103/12/G084
Outline

1. The standard retrieval problem:
   - The visually most similar
   - All visually similar

2. The standard method - Bag of Words with refinements:
   query expansion, context, geometric verification

3. Beyond similarity retrieval – obtaining new information
   - What is this? Where is this?
   - What is interesting here?
   - Where should I look?
   - Show me different appearances of the same scene

4. Application in 3D reconstruction.

5. The challenges in visual object retrieval.

6. Conclusions.
Standard Image Retrieval
CMP Image Retrieval 2.0 Live Demo
Standard Image Retrieval Evaluation

Query:
- 10 database images
- 5 relevant images

Ranking:

- 10 database images
- 5 relevant images

Precision = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}
Recall = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}

Area under the curve
Average Precision (AP)
1. recall and precision

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

Query

\[
\text{precision} = \frac{\#\text{relevant}}{\#\text{returned}}
\]
\[
\text{recall} = \frac{\#\text{relevant}}{\#\text{total relevant}}
\]

Results (ordered):

![Images of results ordered by precision and recall]
Bag-of-Words (BoW): Off-line Stage

Bag-of-Words is the standard image representation for specific object retrieval...
Building a Visual Vocabulary

Partition the feature space

- Feature distance
  0 : features in the same cell
  \( \infty \) : features in different cells

- most of the features are not considered (infinitely distant)
- near-by descriptors accessible instantly – storing a list of features for each cell

Feature distance

Partition the feature space

(k – means clustering)

- extremely time consuming (k=10^7, |dataset|>10^{10})
- approximate k-means used (FLANN)
- \( L_2 \) most common (for computational convenience), other options \( L_1 \), Hellinger,
Efficient Scoring

bag of words representation (up to 1,000,000 D)

$$\cos \varphi = \frac{x \cdot y}{||x|| \cdot ||y||} = \frac{1}{||x|| \cdot ||y||} \sum_{i=1}^{N} x_i y_i$$

$$\sum_{x_i \neq 0, y_i \neq 0} x_i y_i$$

Database Query Score

\[\begin{array}{cccc}
\alpha_1 & 1 & 0 & 0 & 2 \\
\alpha_2 & 0 & 2 & 0 & 1 \\
\alpha_3 & 1 & 0 & 0 & 0 \\
\vdots \\
A & 0 & \alpha_q & B & 3 \\
C & 0 & & C \\
D & 1 & & D \\
\end{array}\]

s_1 \quad s_2 \quad s_3 \quad \vdots
Bags of Words – Image Representation

Term-frequency (tf) – visual word D is twice in the image.

Images are represented by sparse vector / histogram of visual words present in them.
Bag of Words: Weighting words

- Words (in text) common to many documents are less informative – ‘the’, ‘and’, ‘or’, ‘in’, ...
- Words that are too frequent (virtually in every document) can be put on a stop list (ignored as if they were not in the document)
- Images are represented by weighted histograms $tf_x \cdot idf_x$ rather than just a histogram of $tf_x$.

$$idf_x = \log \frac{\# \text{ documents}}{\# \text{ docs containing } x}$$

1. Inverted file: posting list per visual word

<table>
<thead>
<tr>
<th>word</th>
<th>image ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1510…</td>
</tr>
<tr>
<td>2</td>
<td>2712…</td>
</tr>
<tr>
<td>3</td>
<td>1415… 7200190</td>
</tr>
</tbody>
</table>

2. Image ranking

<table>
<thead>
<tr>
<th>score</th>
<th>image ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.87</td>
<td>5</td>
</tr>
<tr>
<td>0.75</td>
<td>1573</td>
</tr>
<tr>
<td>0.52</td>
<td>11202</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>0.001</td>
<td>32</td>
</tr>
</tbody>
</table>

3. Spatial verification

<table>
<thead>
<tr>
<th>#inliers</th>
<th>image ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>247</td>
<td>1573</td>
</tr>
<tr>
<td>105</td>
<td>5</td>
</tr>
<tr>
<td>17</td>
<td>11202</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
</tr>
</tbody>
</table>

4. Re-ranked shortlist

<table>
<thead>
<tr>
<th>query</th>
<th>image 1573</th>
<th>image 45</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1510…</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>7200190</td>
<td></td>
</tr>
<tr>
<td>…</td>
<td>1599956</td>
<td></td>
</tr>
</tbody>
</table>

5. Query expansion

<table>
<thead>
<tr>
<th>query</th>
<th>image 1573</th>
<th>image 45</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1510…</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>7200190</td>
<td></td>
</tr>
<tr>
<td>…</td>
<td>1599956</td>
<td></td>
</tr>
</tbody>
</table>
Re-rank top ranked images (removing false positives)

- RANSAC

**NOTE:** Standard BoW score ranking performed without geometric information

**IMPORTANT:** Geometric verification crucial for query expansion

Sivic, Zisserman: Video Google, ICCV 2003

Philbin, Chum, Isard, Sivic, Zisserman: Object retrieval with large vocabularies and fast spatial matching, CVPR’07
Spatial Verification and Query Expansion

Results

Query image

New query

Chum, Philbin, Sivic, Isard, Zisserman: Total Recall..., ICCV 2007
Query Expansion Step by Step

Query Image

Retrieved image

Originally not retrieved
Query Expansion Step by Step
Query Expansion Results

Original results (good)

Query image

Expanded results (better)
Is this what we want?

- Visually most similar
  Results identical to query for large datasets

- All visually similar
  Output of varying length
  Ground truth hard to obtain

- Users won’t look at tens of near-duplicate images!
Beyond Similarity Retrieval
New Retrieval Problems

What is this?

... and what is that?

Let’s zoom-in!
Standard Retrieval (for browsing, ...)

Query 1

Query 2

Standard retrieval: similar results for both queries, no new information
Standard Retrieval and Details

query

rank: 1 2 32 64 65

query

rank: 1 2048 16384 81368

DIFFICULT

EASY
Zoom-in: On-line Stage

1. Inverted file: posting list per visual word

<table>
<thead>
<tr>
<th>word</th>
<th>image ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1510</td>
</tr>
<tr>
<td>2</td>
<td>2712</td>
</tr>
<tr>
<td>3</td>
<td>1415</td>
</tr>
</tbody>
</table>

2. Image ranking

<table>
<thead>
<tr>
<th>score</th>
<th>image ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.87</td>
<td>5</td>
</tr>
<tr>
<td>0.75</td>
<td>1573</td>
</tr>
<tr>
<td>0.52</td>
<td>11202</td>
</tr>
</tbody>
</table>

3. Spatial verification

3.8 inliers zoom image ID

| 8    | 37x    | 1573 |
| 105  | 17x    | 5    |
| 17   | 7x     | 11202|

4. Re-ranked shortlist

5. Query expansion

<table>
<thead>
<tr>
<th>query</th>
<th>image 1573</th>
<th>image 45</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>3 15</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1489021</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>1567892</td>
<td></td>
</tr>
</tbody>
</table>

Geometry compressed in inverted file taken into account during scoring

Problem specific ranking function, e.g. maximize scale change

Query expansion from already zoomed images
“What is this?” examples
Zoom-in: Example
Zoom-in: Query Expansion
Zoom-in: Example
Zoom in examples
Zoom-out: On-line Stage

1. Inverted file: posting list per visual word

<table>
<thead>
<tr>
<th>word</th>
<th>image ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5 10</td>
</tr>
<tr>
<td>2</td>
<td>7 12</td>
</tr>
<tr>
<td>3</td>
<td>4 15</td>
</tr>
</tbody>
</table>

2. Image ranking

<table>
<thead>
<tr>
<th>score</th>
<th>image ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.87</td>
<td>5</td>
</tr>
<tr>
<td>0.75</td>
<td>1573</td>
</tr>
<tr>
<td>0.52</td>
<td>11202</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>0.001</td>
<td>32</td>
</tr>
</tbody>
</table>

3. Spatial verification

<table>
<thead>
<tr>
<th>#inliers</th>
<th>zoom image ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>1/37x</td>
</tr>
<tr>
<td>105</td>
<td>1/17x</td>
</tr>
<tr>
<td>17</td>
<td>1/7x</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

4. Re-ranked shortlist

<table>
<thead>
<tr>
<th>query image ID</th>
<th>image 1573</th>
<th>image 45</th>
</tr>
</thead>
</table>

5. Query expansion

Geometry compressed in inverted file taken into account during scoring

Problem specific ranking function, e.g. maximize scale change

Query expansion from already zoomed images
Zoom-out: Iterate
Zoom-out: Iterate
Zoom-out: Iterate
Context expansion

- the model of the object is grown beyond the boundaries of the initial query,
- a feature added into the model that is not inside the context is inactive until confirmed by feature(s) from another image with the same visual word and similar geometry.
- Once a feature is confirmed, it adds the neighbourhood around its center to the context.
Learning the Context

Feature patches back-projected into the context from spatially verified images.
How Much Do We Need to See?

Oxford landmarks – 3 queries
100%, 50%, and 10% of the query bounding box

Context learned from the full bounding box

Context learned from 50% of the bounding box

Context learned from 10% of the bounding box
Retrieval with large viewpoint change

- Image retrieval is not an efficient execution of two view matching.
- Significant part is about finding paths, \textit{sequences of matches}
What is interesting here?
What should you not miss?
Highest Resolution Transform

Given a query and a dataset, for every pixel in the query image:
Find the database image with the maximum resolution depicting the pixel

Mikulík, Radenović, Chum, Matas: Efficient Image Detail Mining, ACCV 2014
What most people find interesting?

Most commonly photographed parts
Given a query and a dataset, for every pixel in the query image:
Find the frequency with which it is photographed in detail

Mikulík, Radenović, Chum, Matas: **Efficient Image Detail Mining**, ACCV 2014
All Details: On-line Stage

1. Inverted file: posting list per visual word

<table>
<thead>
<tr>
<th>word</th>
<th>image ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7350125</td>
</tr>
<tr>
<td>2</td>
<td>7399121</td>
</tr>
<tr>
<td>3</td>
<td>7200190</td>
</tr>
</tbody>
</table>

2. Image ranking

<table>
<thead>
<tr>
<th>score</th>
<th>image ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.87</td>
<td>5</td>
</tr>
<tr>
<td>0.75</td>
<td>1573</td>
</tr>
<tr>
<td>0.52</td>
<td>11202</td>
</tr>
</tbody>
</table>

3. Spatial verification

<table>
<thead>
<tr>
<th>#inliers</th>
<th>zoom</th>
<th>image ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>247</td>
<td>7x</td>
<td>1573</td>
</tr>
<tr>
<td>105</td>
<td>2x</td>
<td>5</td>
</tr>
<tr>
<td>17</td>
<td>37x</td>
<td>11202</td>
</tr>
</tbody>
</table>

4. Re-ranked shortlist

OUT: R

5. Query expansion

IN: q

BOW geometries

query image 1573 image 45
All Details: Hierarchical Query Expansion

**IN: R**

- BOW
- Geometries

1. Grouped images
   - Group $G_i$
   - Group $G_n$

2. Geometric consistency
   - $A_{q,i} \approx A_{q,j} A_{j,i}$

**OUT: $q_1, q_2, ..., q_n$**

- Query $q_1$
- Query $q_n$

- Image 1573 + ... + Image 1761
- Image 45 + ... + Image 33

2016.07.14 ICIAR J. Matas: Visual Retrieval
Image Retrieval for 3D Reconstruction
SfM - 3D Reconstruction

- **Few thousand images**
  - Exhaustive matching of all image pairs
    - [Snavely, Seitz, Szeliski: Photo tourism, SIGGRAPH 2006]
    - High level of details reconstructed
    - Infeasible for larger photo collections

- **Few million images**
  - Matching images through standard image retrieval
    - [Heinly, Schonberger, Dunn, Frahm: Reconstructing the World in Six Days, CVPR 2015]
    - Efficient and scalable image matching
    - Details not reconstructed
Retrieval for 3D Reconstruction

- Visually most similar search
  - Many near duplicates
  - Details lost

- Zoom-in and details search
  - Details retrieved
  - Transition images to match the details

- Zoom-out search
  - Viewpoint change
  - More context

- Sideways crawl
  - Significant viewpoint change
  - More context

Schoenberger, Radenović, Chum, Frahm:
*From Single Image Query to Detailed 3D Reconstruction*, CVPR 2015
Sideways image crawl

Schoenberger, Radenović, Chum, Frahm: *From Single Image Query to Detailed 3D Reconstruction*, CVPR 2015
Sideways crawl: On-line Stage

**1. Inverted file: posting list per visual word**

<table>
<thead>
<tr>
<th>word</th>
<th>image ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1 5 10 ...</td>
</tr>
<tr>
<td></td>
<td>7350125</td>
</tr>
<tr>
<td>2</td>
<td>2 7 12 ...</td>
</tr>
<tr>
<td></td>
<td>7399121</td>
</tr>
<tr>
<td>3</td>
<td>1 4 15 ...</td>
</tr>
<tr>
<td></td>
<td>7200190</td>
</tr>
</tbody>
</table>

**2. Image ranking**

<table>
<thead>
<tr>
<th>score</th>
<th>image ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.87</td>
<td>5</td>
</tr>
<tr>
<td>0.75</td>
<td>1573</td>
</tr>
<tr>
<td>0.52</td>
<td>11202</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>0.001</td>
<td>32</td>
</tr>
</tbody>
</table>

**Shortlist: top N images**

<table>
<thead>
<tr>
<th>#inliers</th>
<th>zoom</th>
<th>image ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>247</td>
<td>7x</td>
<td>1573</td>
</tr>
<tr>
<td>105</td>
<td>2x</td>
<td>5</td>
</tr>
<tr>
<td>17</td>
<td>37x</td>
<td>11202</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

**OUT: R**

Using geometry to find adequate features for expansion (left-right)

Building an expanded query using only sideways features

**3. Spatial verification**

**4. Re-ranked shortlist**

**5. Query expansion**
Sideways Left: Step by Step
Sideways Left: Step by Step
Summary

Visually most similar

Zoom-in / details

Zoom-out

Sideways right
Application of Zoom in, out, ...
From Single Image Query to Detailed 3D Reconstruction

Johannes L. Schönerberger\textsuperscript{1}, Filip Radenović\textsuperscript{2}, Ondrej Chum\textsuperscript{2}, Jan-Michael Frahm\textsuperscript{1}

\textbf{Input}: Single query image
7.4 million images downloaded from the Internet

\textbf{Output}: Detailed 3D model, Video, YouTube

\textbullet Recursive retrieval and Structure-from-Motion for detail reconstruction and model extension

![Diagram showing the process of retrieving a detailed 3D model from a single query image.](Diagram.png)
Retrieval with illumination change
Illumination constraints

Video, YouTube

Conclusions

- For a certain class of objects and problems, image retrieval is mature -- reliable, fast, ...

- There is more than the “specific most-similar object retrieval”, i.e. the nearest neighbor problem. E.g. Find:
  - the most geometrically different, yet the same scene
  - spots in the scene that attract attention
  - a representative collection of appearances of the scene – find the database image most different to already retrieved images of the scene, most similar in illumination, viewpoint, arrangement) to another image, a secondary query
Retrieval with occlusion
Retrieval with large time difference
Retrieving different modalities
Retrieving non-overlapping images
Retrieving *sequences* helps in complex cases
Thank you!

Questions?