Beyond Vanilla Visual 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 Cech

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

1. The standard image representation and retrieval method

2. Refinement: query expansion, context, geometric verification

3. Retrieving the “most different“ Application in 3D reconstruction.

4. The challenges in visual object retrieval.

5. Visual retrieval in changing conditions.

6. Conclusions.
- both the query and the retrieved regions may only (small) parts of the query and DB images
- images with the same scale and appearance retrieved first as they are the most similar. For very large datasets, virtually identical content retrieved.
1. Memory footprint (e.g. VLAD, FisherVectors).

2. Speed. Speed and memory closely related via the memory hierarchy.

3. Precision and recall.

4. Generality of changes the retrieval engine handles.

The Bag of Words provides a good compromise
Bag-of-Words (BoW): Off-line Stage

Bag-of-Words is the standard image representation for specific object retrieval.
Bag-of-Words (BoW) : off-line stage

Bag-of-Words is the standard image representation for specific object retrieval Core issues: robustness, memory footprint v. performance, speed
Building a Visual Vocabulary

Feature distance
0 : features in the same cell
∞ : 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

Partition the feature space
(k – means clustering)
- extremely time consuming (k=10⁷, |dataset|>10¹⁰)
- approximate k-means used (FLANN)
- L₀, most common (for computational convenience), other options L₁, Hellinger,
Bags of Words

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 $tf_x$

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

Efficient Scoring

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

\[
\sum_{x_i \neq 0, y_i \neq 0} x_i y_i
\]

Database

\[\begin{array}{cccc}
\alpha_1 & (1 & 0 & 0 & 2) \\
\alpha_2 & (0 & 2 & 0 & 1) \\
\alpha_3 & (1 & 0 & 0 & 0) \\
\vdots
\end{array}\]

Query

\[\begin{array}{c}
\bullet \alpha_q \\
\alpha_q & (0 & 3) \\
\vdots
\end{array}\]

Score

\[\begin{array}{c}
s_1 \\
s_2 \\
s_3 \\
\vdots
\end{array}\]

bag of words representation
(up to 1,000,000 D)
BoW and the Inverted File

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

query visual word 1

query visual word 2

query visual word 3

2015.06.12 CBMI Prague J. Matas: Visual Retrieval
Large Scale Object Retrieval

From the Standard Method to The State of the Art
Fine vocabulary (16 million visual words)
Using wide-baseline stereo matches on 6 million images to learn what is similar

Mikulik, Perdoch, Chum, and Matas: Learning a Fine Vocabulary, ECCV 2010
Mikulik, Perdoch, Chum, Matas: Learning Vocabularies over a Fine Quantization, IJCV 2012

- over 5 million images
- almost 20k clusters of 750k images (visual word based)
- 733k successfully matched in WBS matching (raw descriptor based)
- over 111 M feature tracks established (12.3 M with 6+ features)
- 564 M features in the tracks (319.5 M in tracks of 6+ features)
Bag-of-Words: On-line Stage

Inverted file

+TF-IDF weights

TF-IDF ranking

Final ranking

Geometry

LO-RANSAC

$T = A_j^{-1} A_i$
Jégou, Douze, Schmid: Hamming Embedding and Weak Geometric consistency for large-scale image search, ECCV’08, IJCV’10; ...

(USE marginals over rotation and scale rather than the whole table)
Exhaustive Geometric Verification

Stewénius, Gunderson, Pilet:
Size matters: exhaustive geometric verification for image retrieval, ECCV 2012

query visual word 1

query visual word 2

query visual word 3
Exhaustive Geometric Verification

- Costs time: $\log$ (number of posting lists involved)
- Efficient merging with fixed tree structure
- Any kind of spatial verification can be used
- Images with too few matches are rejected directly

Combined with small number of features per image and very large vocabularies gives impressive massive-scale results

Stewéníus et al., ECCV 2012
Bag-of-Words: On-line Stage

Inverted file

<table>
<thead>
<tr>
<th>word</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>948534</th>
<th>998125</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>100000</td>
<td></td>
</tr>
</tbody>
</table>

+TF-IDF weights

TF-IDF ranking

<table>
<thead>
<tr>
<th></th>
<th>0.85</th>
<th>0.81</th>
<th>0.013</th>
<th>0.001</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>graffiti</td>
<td>graffiti2</td>
<td>bark</td>
<td>all_souls</td>
</tr>
</tbody>
</table>

Final ranking

<table>
<thead>
<tr>
<th></th>
<th>210</th>
<th>138</th>
<th>3</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>graffiti2</td>
<td>graffiti</td>
<td>bark</td>
<td>all_souls</td>
</tr>
</tbody>
</table>
In: $U = \{x_i\}$ set of data points, $|U| = N$

$f(S): S \rightarrow p$ function $f$ computes model parameters $p$ given a sample $S$ from $U$

$\rho(p, x)$ the cost function for a single data point $x$

Out: $p^*$ $p^*$, parameters of the model maximizing the cost function

$k := 0$

Repeat until $P\{\text{better solution exists}\} < \eta$ (a function of $C^*$ and no. of steps $k$)

$k := k + 1$

I. Hypothesis

(1) select randomly set $S_k \subset U$, sample $|S_k| = m$

(2) compute parameters $p_k = f(S_k)$

II. Verification

(3) compute cost $C_k = \sum_{x \in U} \rho(p_k, x)$

(4) if $C^* < C_k$ then $C^* := C_k$, $p^* := p_k$

end
Point
Point + scale
Point + scale + orientation
Ellipse + orientation

Translation
Translation and isotropic scale
Similarity
Affine transformation

Ellipse correspondence and a gravity vector

Philbin et al.: Object retrieval with large vocabularies and fast spatial matching, CVPR’07 (using rotationally invariant descriptors)

Perdoch, Chum, Matas: Efficient Representation of Local Geometry for Large Scale Object Retrieval, CVPR 2009 (using rotationally variant descriptors)
Gravity Vector Assumption

- Performance comparison

<table>
<thead>
<tr>
<th>mAP</th>
<th>Oxford5K vocab.</th>
<th></th>
<th>Paris vocab.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ox5K</td>
<td>Ox105K</td>
<td>Ox5K</td>
</tr>
<tr>
<td>Dominant orient.</td>
<td>0.772 0.887*</td>
<td>0.687 0.844*</td>
<td>0.592 0.733*</td>
</tr>
<tr>
<td>Gravity vector</td>
<td>0.786 0.900*</td>
<td>0.723 0.852*</td>
<td>0.635 0.782*</td>
</tr>
</tbody>
</table>

*with query expansion.

- SIFT dominant orientation [Lowe’04] produces 50% more descriptors.

- Robustness of gravity vector assumption

- Robustness of SIFT to small imprecision in orientation.
- Correct final geometry due to LO-step in RANSAC.
Query Expansion

Inverted file

+TF-IDF weights

Geometry

LO-RANSAC

Final ranking

New expanded query
Query Expansion

Results

Spatial verification

New results

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

Query Image

Retrieved image

Originally not retrieved
Query Expansion Results

Query image
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.

Chum, Mikulik, Perdoch, Matas:
Total Recall II: Query Expansion Revisited, CVPR 2011
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.

The query
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
Browsing Image Collections

Do we really want the most similar image?
What is this?

... and what is that?

Let’s query!
Retrieval for Browsing

Query 1

Query 2

Standard retrieval: similar results for both queries, no new information
Retrieval for Browsing: Zoom in

Query 1

Query 2
Zoom: Key differences

- Geometry compressed in inverted file taken into account during TF-IDF scoring
- Problem specific ranking function \( \text{zoom in}_{z_{in}} = \sqrt{\frac{A_r}{A_q}} \)
- Query expansion from different images

Algorithm 1 Overview of the zooming algorithm. Note that step 5 represents a trade-off between the query time and output quality.

**Input**: Bag-of-words of the query image  
**Output**: Ranked list of images

1. Fetch posting lists for query visual words and score in DAAT order for each scale band separately.
2. Re-weight scores in scale bands to prefer desired change in scale and create a shortlist.
3. Spatially verify images in the shortlist, incrementally building an expanded query.
4. Rank images according to the desired goal (zoom-in/zoom-out)
5. Return the result or form the expanded query with context learning and goto 1
“What is this?” examples
Dataset Oxford buildings: Christ Church
Dataset Oxford buildings: Bodleian

QVOD FELICITER VORTAT
ACADEMICI OXONIENS
BIBLIOTHECAM HANC
VOBIS REIPVBLCAEQVE
LITERATORVM
T. B. P.
Zoom in examples
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

![Image of the clock tower with zoom-in and hierarchical QE graph]
“Find all details” examples

QUERY

HRT
“Where is this” example

DEMO
Application of Zoom in / Zoom out: From Single Image Query to Detailed 3D Reconstruction

Johannes L. Schönberger¹, Filip Radenović², Ondrej Chum², Jan-Michael Frahm¹

• Input: Single query image
  7.4 million images downloaded from the Internet

• Output: Detailed 3D model

• Recursive retrieval and Structure-from-Motion for detail reconstruction and model extension
Application of Zoom in / Zoom out:
From Single Image Query to Detailed 3D Reconstruction

3D reconstruction video
Retrieval challenges: large viewpoint change

- Image retrieval is not an efficient execution of two view matching.
- Significant part is about finding paths, *sequences of matches*
Challenges: viewpoint change
Challenges: illumination change
Challenges: large time difference
Challenges: occlusion

triangular inequality?
\[ d(a,c) \leq d(a,b) + d(b,c) \]
Challenges: different modalities
More difficult: non-overlapping images
And it can all happen at once!
Deep Convolutional Neural Networks? state-of-the-art in similar computer vision problems:


- e.g. 60M parameters, trained on 1.2M images
Experiments: Category Recognition

http://cmp.felk.cvut.cz/~zechj/tmp/Tristan.jpg

- Doberman, Doberman pinscher
- Rottweiler
- black-and-tan coonhound
- Great Dane
- bluetick
(In)Sensitivity to Image Rotation

Score vs. Angle for Various Dog Breeds

- Doberman, Doberman pinsch
- Rottweiler
- Black-and-tan coonhound
- miniature pinscher
- muzzle
- kelpie
- German shepherd, German shepherd
- toy terrier
- Chihuahua
- French bulldog
- Tench, Tinca tinca
- Chimpanzee, chimp, Pan troglodytes

2015.06.12 CBMI Prague J. Matas: Visual Retrieval
(In)Sensitivity to Image Blur
It is not a texture only...

- tiger, Panthera tigris
- tiger cat
- tabby, tabby cat
- lynx, catamount
- jaguar, panther, Panthera onca, Felis onca

- Saint Bernard, St Bernard
- Welsh springer spaniel
- Blenheim spaniel
- Irish setter, red setter
- Leonberg
• so far, no impact
• for specific objects (landmarks) spatial pooling detrimental
• often only single isolated image available (not enough data)
• occlusion handling, query expansion, spatial verification with CNNs unclear
• but:

CNN viewpoint regression (U. of Cambridge)
• Illumination and appearance

• Sensor

• Viewpoint

• Occlusions
Place Recognition System Structure

Reference images

"Memory"

Affine view synthesis

WxBS feature detection & description

BoVW engine

Query image

"Live"

Adjacency models pool

Adjacency model selection

Affine view synthesis

WxBS feature detection & description

BoVW engine

Retrieved candidates

MODS-WxBS matching, RANSAC verification

Candidates by adjacency model

Corresponding memory image
Handling Extreme Viewpoints

MODS – matching with on-demand synthesis

MODS handles angular viewpoint difference up to:
• $85^\circ$ for planar scenes
• $30^\circ$ for structured

D. Mishkin. J. Matas and M. Perdoch.
MODS: Fast and Robust Method for Two-View Matching, arXiv 2015,
Illumination and sensor

WxBS features:

- **Adaptive thresholding** for detect point in low-contrast regions: if \#MSERs < \( \theta_{MSER} \), or \#HesAffs < \( \theta_{HA} \), lower a detection threshold.

- Use HalfRootSIFT for description.


Why not use CNN?

Possible, but needs learning for each dataset

---

Query

BoW

WXBS

ImageNet

AlexNet

pool5
<table>
<thead>
<tr>
<th>Method</th>
<th>Prec</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAPIR (CNN)</td>
<td>0.747</td>
<td>0.836</td>
<td>0.789</td>
</tr>
<tr>
<td>Bonn (CNN)</td>
<td>0.726</td>
<td>0.758</td>
<td>0.741</td>
</tr>
<tr>
<td>BoW HalfRootSIFT</td>
<td>0.530</td>
<td>0.890</td>
<td>0.665</td>
</tr>
<tr>
<td>BoW Half &amp; RootSIFT</td>
<td>0.538</td>
<td>1.000*</td>
<td>0.700</td>
</tr>
<tr>
<td>BoW Half &amp; RootSIFT &amp; MODS + adj. model</td>
<td>0.821</td>
<td>0.825</td>
<td>0.823</td>
</tr>
</tbody>
</table>
Success and failure cases

query

output

ground truth

No match

No match

No match

No match
Conclusions

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

- There is more than the “specific most-similar object retrieval” (the nearest neighbour) problem:
  - most geometrically different.
  - return a representative collection of differing images.

- The field of Computer vision is in a turmoil. Deep nets and huge commercial investments are changing the landscape. The problems will stay, the state-of-the-art might not.
Thank you for attention!