Efficient Representation of Local Geometry for Large Scale Object Retrieval

Michal Perďoch
Ondřej Chum and Jiří Matas

Center for Machine Perception
Czech Technical University in Prague
Large Scale Object Retrieval

- Large (web) scale “real-time” search involves millions(billions) of images
- Indexing structure should fit into RAM, failing to do so results in an order of magnitude increase in response time
- 1 bit per feature requires ~1.16GB in a retrieval system with 5 million images, ~2000 features per image
- Scalability and the cost of the search engine is determined by memory footprint -> memory is the limiting factor

Methods based on local features

- Are able to retrieve objects, not only images
- Robust to occlusions and background changes
Image vs. Object Retrieval
Object Retrieval vs. Global Methods

Query
Image Representation by Local Features

- Local features = appearance + geometry
- Bag of words approaches focus on appearance.
- Geometry benefits object retrieval (in spatial verification) [Philbin’07]

The problem: How to efficiently store geometric information of local features?
- Currently, the straightforward representation of local geometry requires roughly 4-5x more memory than the appearance
- I.e. ~32bits for appearance vs. ~160bits for geometry.

Our solution: a learnt geometric vocabulary
Image Representation by Local Features

Keypoint Detection → Local Appearance

Bag of words, Video Google [Šivic’03]

SIFT Description [Lowe’04]

Visual Vocabulary

Local Geometry → Visual Words

word₁, word₂, word₈, ...

word₉₄₈₅₃₄, word₉₉₈₁₂₅
Object Retrieval with Spatial Verification

Inverted file

1
4
8
11
79
...
948534
998423

+TF-IDF weights

word

1
2
3
...
1000000

TF-IDF ranking

0.85
graffiti
0.81
graffiti2
...
0.013
bark
0.001
all_souls
...

Geometry

LO-RANSAC

A_i
A_j

T = A_j^{-1}A_i

Final ranking

210
graffiti2
138
graffiti
...
3
bark
1
all_souls
Local Geometry of Affine Features

- Coordinates \( x_0 \) and an ellipse \( E \)

\[
(x - x_0)^\top E (x - x_0) = 1,
\]

ellipse \( E \) can be decomposed up to an arbitrary rotation \( Q \)

\[
E = A^\top A = (QA)^\top (QA),
\]

- We fix ambiguity in \( Q \) by choosing

\[
A = \begin{pmatrix} a & 0 \\ b & c \end{pmatrix},
\]

- \( R \) is given by a dominant orientation as in [Lowe’04]

- Local affine transformation is defined by a single correspondence

\[
T = A_j^{-1} R_j^\top R_i A_i
\]
Geometry Compression

Representing transformation $A_i$

- $A_i$ (independent of rotation $R$), captures the shape of the ellipse

![Diagram showing transformation $A_i$ to $B$]

- Let a transformation $B_i$ be a representant of $A_i$, ideally

$$H = B_i^{-1} A_i = I.$$  

- However, to compress the representation, $B$ will be chosen from a vocabulary and shared by multiple $A_i$, so that

$$H = B^{-1} A_i = I + \varepsilon.$$  

- How to measure and minimize error $\varepsilon$ of $H$?
Minimizing Geometric Error of $H$

\[ H = B^{-1}A_i \]

We capture the quality of $H$ by integrating reprojection error over unit circle

\[
e = \int_{\|x\|^2=1} \|Ix - (I + E)x\|^2 = \int_{\|x\|^2=1} \|Ex\|^2,
\]

using simple manipulations

\[
e = \int_{\alpha=0}^{2\pi} \left\| E \begin{pmatrix} \cos \alpha \\ \sin \alpha \end{pmatrix} \right\|^2 = \pi \cdot \|E\|_F^2, \quad E = B^{-1}A_i - I.
\]

Frobenius norm can be used in space of transformations as a similarity
K-means-like Clustering

For a set of transformations $\mathbf{A}$ of cardinality $N$ we are looking for a best set $\mathbf{B}$ of $K$ representative candidates $\mathbf{B}_j$

$$\mathcal{A} = \{\mathbf{A}_1, \ldots, \mathbf{A}_N\}, \quad \mathcal{B} = \{\mathbf{B}_1, \ldots, \mathbf{B}_K\}, \quad K \ll N$$

that minimizes error $E$

$$E = \sum_{j \in K} \sum_{\mathbf{A}_i \in \mathcal{A}_j} \|\mathbf{B}_j^{-1}\mathbf{A}_i - \mathbf{I}\|_F^2$$

over all possible assignments $\mathcal{A}_j$.

1. assignment step

$$f(i) = \arg\min_j \|\mathbf{B}_j^{-1}\mathbf{A}_i - \mathbf{I}\|_F^2 \quad \mathcal{A}_j = \{\mathbf{A}_i, f(i) = j\}$$

2. refinement step

$$\mathbf{B}_j = \arg\min_{\mathbf{B}} \sum_{\mathbf{A}_i \in \mathcal{A}_j} \|\mathbf{B}^{-1}\mathbf{A}_i - \mathbf{I}\|_F^2$$

which leads to a closed form solution.

3. go to 1, until no reassignments occurred or error is below threshold
Geometric “vocabularies”

- Assuming that the scale and shape of the ellipse are independent, scale can be extracted and quantized separately.
- Each geometric vocabulary is characterized by two numbers \((x, y)\) and denoted \(SxEy\)
  
  \[ x \text{ – number of bits for encoding scale} \]
  
  \[ y \text{ – number of bits for encoding ellipse shape, } K = 2^y \]

- Geometric vocabularies learnt on a set of local geometries of 10M affine covariant points (for visualisation scale was quantized separately)

\[
\begin{align*}
y = 2, & \quad K = 2^2 \\
y = 3, & \quad K = 2^3 \\
y = 4, & \quad K = 2^4 \\
y = 5, & \quad K = 2^5
\end{align*}
\]
### Geometric “vocabularies”

<table>
<thead>
<tr>
<th>Identifier</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>S0E8</td>
<td>256 ellipses</td>
</tr>
<tr>
<td>S4E4</td>
<td>16 scales, 16 ellipses</td>
</tr>
<tr>
<td>S8E0</td>
<td>256 scales, circles</td>
</tr>
<tr>
<td>S4E12</td>
<td>16 scales, 4k ellipses</td>
</tr>
</tbody>
</table>
Efficient Geometry Representation

Coordinates $x_i, y_i$ are quantized separately and encoded using 16bits
Gravity Vector Assumption

- Transformation $A$ has one eigenvector $(0,1)^\top$ pointing down in the image – we call it a gravity vector.
- Surprisingly, upright direction is often preserved ($R = I$).
- Gravity vector assumption was used solely in spatial verification [Philbin’07].
Gravity Vector Assumption

Performance comparison

<table>
<thead>
<tr>
<th>mAP</th>
<th>Oxford5K vocab.</th>
<th></th>
<th>Paris vocab.</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ox5K</td>
<td>Ox105K</td>
<td>Ox5K</td>
<td>Ox105K</td>
</tr>
<tr>
<td>Dominant orient.</td>
<td>0.772</td>
<td>0.887*</td>
<td>0.687</td>
<td>0.844*</td>
</tr>
<tr>
<td>Gravity vector</td>
<td>0.786</td>
<td>0.900*</td>
<td>0.723</td>
<td>0.852*</td>
</tr>
<tr>
<td></td>
<td>0.592</td>
<td>0.733*</td>
<td>0.501</td>
<td>0.637*</td>
</tr>
<tr>
<td></td>
<td>0.635</td>
<td>0.782*</td>
<td>0.572</td>
<td>0.725*</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
Experiments – Datasets and Protocol

Oxford Buildings Dataset (Ox5K) [Philbin’07]

- 5062 images of buildings collected from Flickr, 55 queries with ground truth, 11 landmarks each with 5 different queries
- mAP – mean Average Precision, average area below the precision-recall curves for all queries
- Additional 100k of distractor images, together dataset (Ox105K)
Experiments – Geometric Vocabularies

- Two different visual vocabularies, trained on Oxford5K and ~6000 images of tourist spots in Paris, protocol from [Philbin’07]

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td>bits</td>
<td>Ox5K</td>
</tr>
<tr>
<td>w/o spatial</td>
<td>0</td>
<td>0.717</td>
</tr>
<tr>
<td>S0E4</td>
<td>16+4</td>
<td>0.766</td>
</tr>
<tr>
<td>S8E0</td>
<td>16+8</td>
<td>0.782</td>
</tr>
<tr>
<td>S4E4</td>
<td>16+8</td>
<td>0.787</td>
</tr>
<tr>
<td>S0E8</td>
<td>16+8</td>
<td>0.788</td>
</tr>
<tr>
<td>S4E12</td>
<td>16+16</td>
<td>0.789</td>
</tr>
<tr>
<td>Exact</td>
<td>160</td>
<td>0.786</td>
</tr>
</tbody>
</table>

- Spatial verification with “8bit per feature” geometric vocabulary for ellipse shape performs almost as exact geometry

- It might be necessary to use more challenging datasets to highlight the benefit of affine shape. In Oxford buildings dataset
  - 86% of all GT pairs have anisotropy under 1.5 (72% is under 1.25)
  - 90% of all GT pairs have skew under 20° (76% is under 10°)
Experiments – Performance Comparison

Oxford Buildings Dataset

- Proposed method outperforms all state of the art methods
- Geometry representation S0E8 was used

<table>
<thead>
<tr>
<th>Method</th>
<th>Oxford5K vocab.</th>
<th>Paris/Other*</th>
</tr>
</thead>
<tbody>
<tr>
<td>S0E8</td>
<td>0.788</td>
<td>0.725</td>
</tr>
<tr>
<td>S0E8+Soft Assign.</td>
<td>0.846</td>
<td>0.779</td>
</tr>
<tr>
<td>S0E8+Query Exp.</td>
<td>0.901</td>
<td>0.856</td>
</tr>
<tr>
<td><strong>S0E8+SA+QE</strong></td>
<td><strong>0.916</strong></td>
<td><strong>0.885</strong></td>
</tr>
<tr>
<td>Philbin’08 SPatial</td>
<td>0.653</td>
<td>0.565</td>
</tr>
<tr>
<td>Philbin’08 SP+QE</td>
<td>0.801</td>
<td>0.708</td>
</tr>
<tr>
<td><strong>Philbin’08 SP+SA+QE</strong></td>
<td><strong>0.825</strong></td>
<td><strong>0.719</strong></td>
</tr>
<tr>
<td>Jégou’08</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Jégou’08,Tech.Report</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>Jégou CVPR’09</strong></td>
<td><strong>0.747</strong></td>
<td><strong>0.692</strong></td>
</tr>
</tbody>
</table>

*results with different visual vocabulary
Conclusions

Contributions

- We have proposed a method for learning an efficient, compact representation of local geometry.
- Method achieved state of the art results (0.916 mAP) on Oxford buildings dataset and competitive results on INRIA Holidays dataset.
- Proposed method requires only 45.2 bits/feature on average.
  - Local geometry in 24 bits (16 bits for position, 8 bits for shape).
  - Visual word labels and TF/IDF weights in 21.2 bits.
- We have shown that the use of gravity vector assumption in feature description significantly improves recognition rate and further reduces the memory footprint.

Future work

- Compare performance of scale vs. affine covariant points on standard dataset.