INSCRIPTIONS VISUAL RECOGNITION
A comparison of state-of-the-art object recognition approaches

Giuseppe Amato
Fabrizio Falchi
Fausto Rabitti
Lucia Vadicamo

ISTI-CNR, Pisa
lucia.vadicamo@isti.cnr.it
Why inscriptions visual recognition?

Objective:

- Offering the possibility of retrieving information on inscriptions from images
- Both, using traditional Web interfaces and mobile devices
Why inscriptions visual recognition?

Objective:

- Offering the possibility of retrieving information on inscriptions from images
- Both, using traditional Web interfaces and mobile devices
Why inscriptions visual recognition?

Objective:

- Offering the possibility of retrieving information on inscriptions from images
- Both, using traditional Web interfaces and mobile devices
Why inscriptions visual recognition?

Objective:
- Offering the possibility of retrieving information on inscriptions from images
- Both, using traditional Web interfaces and mobile devices
Where is the inscription?
Where is the inscription?
Where is the inscription?
Where is the inscription?

Images need appropriate mathematical representation.
Image recognition

Database images

Query image

Mathematical representation
Image recognition

Database images

Query image

Mathematical representation
Using **Local features**: mathematical representation of visual appearance of specific areas (keypoints) of images

Images are compared by analyzing matching local feature pairs

- Matching pairs are selected by computing pairwise similarity among local features in different images
Local features

- Local features extraction:
  - **Keypoints detection**
    - Uses strategies to decide which points in the image are relevant to be represented
  - **Descriptor building**
    - Build a descriptor (e.g. vector) for each keypoint in the image

- **SIFT** = Scale Invariant Feature Transform
  - The most important and cited local feature
  - [Lowe 1999]
Local features

- Local features extraction:
  - **Keypoints detection**
    - Uses strategies to decide which points in the image are relevant to be represented
  - **Descriptor building**
    - Build a descriptor (e.g. vector) for each keypoint in the image

- SIFT= Scale Invariant Feature Transform
  - The most important and cited local feature
  - [Lowe 1999]
Comparing images

_matching local features is expensive_

- Too many keypoints (thousands) per image
Comparing images

Matching local features is expensive
- Too many keypoints (thousands) per image

Solution: aggregate local feature

ONE descriptor for each image!!
Local Features Aggregation

Aggregation approaches:

- **Bag-of-Features (BoF)**
  - [Sivic & Zisserman 2003]

- **Vector of Locally Aggregated Descriptor (VLAD)**
  - [Jégou et al. 2010]

Both use a visual vocabulary in the aggregation phase

- BoF vocabulary ~ 100,000 – 400,000 elements
- VLAD vocabulary ~ 64-512 elements
Experiments
Dataset: EDR

17,155 photos
14,560 inscriptions

Epigraphic Database Roma (EDR)
EAGLE project
70 queries selected from the whole dataset

- represent various types of inscriptions
- for each query we select the associated images using database consistency
- only inscriptions that have more than one photo are used
Quality Measure

To evaluate the quality of the results we use the probability $p$ of finding an image of the same query object between the first $r$ result ($1 \leq r \leq 100$)

<table>
<thead>
<tr>
<th>Query</th>
<th>Retrieved Image</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.jpg" alt="Image 1" /></td>
<td><img src="image2.jpg" alt="Image 2" /></td>
</tr>
<tr>
<td><img src="image8.jpg" alt="Image 8" /></td>
<td><img src="image9.jpg" alt="Image 9" /></td>
</tr>
</tbody>
</table>
| ... | ... | ... | ... | ... | ... | ... | ...


<table>
<thead>
<tr>
<th>Approach</th>
<th>avg SIFTs</th>
<th>vocabulary size</th>
<th>Bytes</th>
<th>p r =1</th>
<th>p r =10</th>
<th>p r =100</th>
</tr>
</thead>
<tbody>
<tr>
<td>VLAD</td>
<td>235</td>
<td>256</td>
<td>131,072</td>
<td>.69</td>
<td>.74</td>
<td>.84</td>
</tr>
<tr>
<td>BoF / cos TF-IDF</td>
<td>235</td>
<td>400,000</td>
<td>940</td>
<td>.64</td>
<td>.76</td>
<td>.87</td>
</tr>
<tr>
<td>VLAD</td>
<td>235</td>
<td>128</td>
<td>65,536</td>
<td>.64</td>
<td>.73</td>
<td>.87</td>
</tr>
<tr>
<td>BoF / cos TF-IDF</td>
<td>235</td>
<td>200,000</td>
<td>940</td>
<td>.60</td>
<td>.71</td>
<td>.81</td>
</tr>
<tr>
<td>VLAD</td>
<td>1591</td>
<td>256</td>
<td>131,072</td>
<td>.56</td>
<td>.71</td>
<td>.90</td>
</tr>
<tr>
<td>VLAD</td>
<td>1591</td>
<td>128</td>
<td>65,536</td>
<td>.56</td>
<td>.69</td>
<td>.87</td>
</tr>
<tr>
<td>BoF / cos TF-IDF</td>
<td>235</td>
<td>100,000</td>
<td>940</td>
<td>.56</td>
<td>.69</td>
<td>.79</td>
</tr>
<tr>
<td>VLAD</td>
<td>235</td>
<td>64</td>
<td>32,768</td>
<td>.53</td>
<td>.70</td>
<td>.86</td>
</tr>
<tr>
<td>VLAD</td>
<td>1591</td>
<td>64</td>
<td>32,768</td>
<td>.50</td>
<td>.61</td>
<td>.79</td>
</tr>
<tr>
<td>VLAD-PCA (d'=512)</td>
<td>1591</td>
<td>128</td>
<td>2,048</td>
<td>.44</td>
<td>.59</td>
<td>.79</td>
</tr>
</tbody>
</table>

- BoF approach obtained good results when a large vocabulary (400,000) was used
- The best accuracy was obtained by using the VLAD approach with a vocabulary size of 256
Conclusions

► We tested the state-of-the-art object recognition techniques on Epigraphic Database Roma
  - 17,155 photos related to 14,560 inscriptions

► The best accuracy was obtained by using the Vector of Locally Aggregated Descriptor (VLAD)
  - The obtained accuracy was of .69, which is good considering the difficulties of the task and the few images available for each inscription in the dataset

We plan to use VLAD approach in the official EAGLE inscription recognition engine
Thanks

Lucia Vadicamo
lucia.vadicamo@isti.cnr.it