A Large-Scale Image Dataset Collected via Google Image Search

Yu Su & Frédéric Jurie
GREYC, University of Caen, France
Background

The Quaero Still Images Dataset

- Corpus project, WP 8.1
- Objectives
  - Provide training and evaluation data for CTC.WP8, e.g. scene annotation and object recognition.
  - Should be useful in the application where image content has to be searched.
  - Allow to evaluate automatic annotations tools.
Background

Some existing image datasets

- CalTech-256
  - 256 object categories containing 30,607 images.
- PASCAL VOC
  - 20 object categories containing 14,743 images.
- ImageNet
  - 14,841 concepts organized by WordNet hierarchy.
  - About 10 millions images.

...
Background

Differences with previous datasets

- Addressed task: improve the ranking of existing text-based image search engines
  - In large scale problem using text to retrieve an initial set of images is reasonable
  - Images can be re-ranked according to their visual content
- Previous dataset does not include any textual description (contextual information)
Background

Features of Quaero Still Images Dataset

- Large scale
  - 518 concepts, ~1,000 images per concepts
- Diverse concepts
  - Objects, people, scenes, events, …
- Collection source
  - Google Image Search
- Contextual text and images
  - URLs, HTML tags and surrounding words
  - Other images on web pages
Outline

- Background
- Collection
- Statistics
- Annotation
Dataset Collection

- 518 pre-defined concepts*

  indoor/outdoor, day/night, ...

  beach, forest, street, Alffel-tower, ...

  soldier, nurse, ...
  M.Jackson, J.Reno, ...

  dog, cat, ...
  rose, mushroom, ...

  earthquake, meeting, traffic jam, ...

  man-made object 44%
  animal&plant 18%
  scene 18%
  people 4%

*cooperating with R. Landais and G. Quénot, some concepts are the same as those in TRECVID
Dataset Collection

- Exclude some “bad” concepts

- rice
  - cooked rice
  - rice plant

- apple
  - fruit
  - trademark

non-consistent visual contents
Collection Flowchart

concept

Google images

imgtype=photo
omit unrealistic images

filter=1
omit similar images

image1 page1
contextual text & images

image2 page2
contextual text & images

image N page N
contextual text & images
The image shows a dog and a cat, with the text:

doors and never leave our houses again. Life is risky. Terribly risky.

Dataset Structure

518 concepts

- sky
  - unit 1
  - unit 2
  - ... unit N

- dog
  - unit 1
  - unit 2
  - ... unit N

- mug
  - unit 1
  - unit 2
  - ... unit N

...
Dataset Structure

518 concepts

- sky
  - unit 1
  - unit 2
  - unit N

- dog
  - unit 1
  - unit 2
  - unit N

- mug
  - unit 1
  - unit 2
  - unit N

image
contextual text
- page & image url
- title & alg html tag
- surrounding words

contextual images
Outline

- Background
- Collection
- Statistics
- Annotation
Dataset Statistics

Overview
- 518 concepts containing 484,747 images
- 482,007 pages and 1,543,766 contextual images
- Data volume: ~170G bytes
- Time cost: ~1500 hours
Dataset Statistics

- 935 returned images per concept
  - For each concept, returned images are less than 1,000 due to bad URLs, unrealistic images and duplications.
94% returned images have contextual text

Not all the returned images have contextual text due to the bad URLs of some web pages

Example:

dog: 928 images, 892 with text
text percent is 892/928 = 96%
59% returned images have contextual images.

Averagely, each of them have 5 contextual images.

![Histogram showing the distribution of the number of contextual images per returned image. The x-axis represents the number of contextual images, ranging from 1 to 10, and the y-axis represents the proportion. The histogram peaks at 9 contextual images.]
Annotation (not finished yet)

query: dog

<table>
<thead>
<tr>
<th>Image</th>
<th>Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Image 1" /></td>
<td>✔️</td>
</tr>
<tr>
<td><img src="image2.png" alt="Image 2" /></td>
<td>✔️</td>
</tr>
<tr>
<td><img src="image3.png" alt="Image 3" /></td>
<td>✔️</td>
</tr>
<tr>
<td><img src="image4.png" alt="Image 4" /></td>
<td>✔️</td>
</tr>
<tr>
<td><img src="image5.png" alt="Image 5" /></td>
<td>✔️</td>
</tr>
<tr>
<td><img src="image6.png" alt="Image 6" /></td>
<td>✔️</td>
</tr>
<tr>
<td><img src="image7.png" alt="Image 7" /></td>
<td>✔️</td>
</tr>
<tr>
<td><img src="image8.png" alt="Image 8" /></td>
<td>✔️</td>
</tr>
<tr>
<td><img src="image9.png" alt="Image 9" /></td>
<td>✔️</td>
</tr>
<tr>
<td><img src="image10.png" alt="Image 10" /></td>
<td>✔️</td>
</tr>
<tr>
<td><img src="image11.png" alt="Image 11" /></td>
<td>✔️</td>
</tr>
<tr>
<td><img src="image12.png" alt="Image 12" /></td>
<td>✔️</td>
</tr>
<tr>
<td><img src="image13.png" alt="Image 13" /></td>
<td>✔️</td>
</tr>
<tr>
<td><img src="image14.png" alt="Image 14" /></td>
<td>✔️</td>
</tr>
<tr>
<td><img src="image15.png" alt="Image 15" /></td>
<td>✔️</td>
</tr>
</tbody>
</table>

~2 hours to annotate each concept
Thanks for your attention!