The ImageCLEF 2013 Scalable Concept Image Annotation Subtask

Mauricio Villegas,† Roberto Paredes† and Bart Thomee‡

† ITI/DSIC, Universitat Politècnica de València
{mvillegas, rparedes}@iti.upv.es

‡ Yahoo! Research
bthomee@yahoo-inc.com

Valencia, 25th of September 2013
Outline

1. Introduction
   - Motivation

2. Subtask Description
   - Lines of work
   - Web training dataset

3. Evaluation
   - Participation
   - Results

4. Conclusions and Future Work
Automatic image annotation is the process by which a computer assigns to an image, metadata that describes its content.

In this work the metadata considered is only the presence or absence of concepts in the images, e.g.

- Dog
- Table
- Rural
- Grass
- Daytime
- Tree
- ...

The ImageCLEF 2013 Scalable Concept Image Annotation Subtask

CLEF 2013 (September 25, 2013)
Image annotation research has mostly relied on manually labeled training data. Examples of available datasets are:

- **ImageNet**: $\approx 1.2$M images, 1000 concepts, but only one concept per image.
- **NUS-WIDE**: $\approx 269$k images, multiple concepts per image, but only 81 concepts.

Even though crowdsourcing has proved to be very useful, it is expensive and difficult to scale to a large amount of concepts.

Are there alternatives that do scale concept-wise?

- Millions of images and corresponding related text can be cheaply crawled from the Internet for practically any topic.
Introduction – Motivation

How to effectively use Web data for image annotation?

- The text in websites is noisy and the degree of relationship to the images varies greatly.

- The types of images also varies. Take for example images from a search query of “rainbow”: 
1 Introduction
   - Motivation

2 Subtask Description
   - Lines of work
   - Web training dataset

3 Evaluation
   - Participation
   - Results

4 Conclusions and Future Work
Objective: To use only automatically gathered data for developing concept scalable image annotation systems.

- Any data could be used as training, except for hand labeled images, e.g. crawled data, WordNet, dictionaries, stemmers, etc.

Participants were provided with:

- Crawled dataset (250,000 images and respective webpages).
- Development set (1,000 images, labeled for 95 concepts).
- Implementation of a baseline system and code for computing the performance measures.

Test set: 2,000 images, the participants had to label them for 116 concepts (max. 6 runs could be submitted per group).

Concepts: Were defined as WordNet synsets and for most of them, also a Wikipedia article was associated.
In contrast to traditional image annotation tasks, the proposed one involves more lines of work:

- Which representation to use for the images (visual features).
- How to use unsupervised web data as training.
  - Automatically assign concepts to the images using the textual data?
  - How to preprocess and clean the textual data?
  - Use other resources:
    - Ontologies
    - Language dictionaries
    - Automatic translation
- Which method to use for modeling the concepts.
- What strategy to use for deciding how many and which concepts are assigned to an image.
Subtask description – Web training dataset

- Web training dataset\(^1\) composed of 250,000 images, 7 visual features types and 4 textual feature types.

- Images found by querying Google, Bing and Yahoo using the words from the English dictionary.

- Precautions taken to avoid “message images”, duplicates and near-duplicates.

- To ease data download and handling by participants, the subset of 250,000 images was selected using 158 concepts (including the concepts for the task).

\(^1\)Dataset available at [http://risenet.iti.upv.es/webupv250k](http://risenet.iti.upv.es/webupv250k)
Visual Features:

<table>
<thead>
<tr>
<th>Feature</th>
<th>Dimensionality</th>
<th>Training data size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thumbnails</td>
<td>Max. 200 pixels high</td>
<td>15 GB</td>
</tr>
<tr>
<td>GIST</td>
<td>480</td>
<td>810 MB</td>
</tr>
<tr>
<td>Color Hist.</td>
<td>576</td>
<td>170 MB</td>
</tr>
<tr>
<td>GETLF</td>
<td>256</td>
<td>30 MB</td>
</tr>
<tr>
<td>SIFT</td>
<td>5,000 BoW</td>
<td>770 MB</td>
</tr>
<tr>
<td>C-SIFT</td>
<td>5,000 BoW</td>
<td>660 MB</td>
</tr>
<tr>
<td>RGB-SIFT</td>
<td>5,000 BoW</td>
<td>750 MB</td>
</tr>
<tr>
<td>OPP-SIFT</td>
<td>5,000 BoW</td>
<td>720 MB</td>
</tr>
</tbody>
</table>
Textual Features:

1. Words used to find the images (3MB).
2. Relative URLs of images in webpages (25MB).
3. Image webpages as valid XML (2.3GB).
4. Webpage text (110M):

```
<html>
<head>
<title>Dogs can tell size of another dog by listening to its growls | Science / Technology</title>
</head>
<body>
<h2>Dogs can tell size of another dog by listening to its growls</h2>
<img src="img/dogs.jpg" alt="dogs in the park" />
<p>Washington, Dec 21: A new study has shown that dogs can tell the size of another dog by listening to its growls. Peter Pongracz and his team recruited 96 dogs of various breeds ...</p>
</body>
</html>
```
Subtask description – Web training dataset

Textual Features:

1. Words used to find the images (3MB).
2. Relative URLs of images in webpages (25MB).

Dogs can tell size of another dog by listening to its growls

Washington, Dec 21: A new study has shown that dogs can tell the size of another dog by listening to its growls. Peter Pongracz and his team recruited 96 dogs of various breeds ...

<html>
<head>
<title>Dogs can tell size of another dog by listening to its growls | Science / Technology</title>
</head>
<body>
<h2>Dogs can tell size of another dog by listening to its growls</h2>
<p><img src="img/dogs.jpg" alt="dogs in the park" /></p>
<p>Washington, Dec 21: A new study has shown that dogs can tell the size of another dog by listening to its growls. Peter Pongracz and his team recruited 96 dogs of various breeds ...</p>
</body>
</html>

3. Image webpages as valid XML (2.3GB).

4. Webpage text (110M):

dogs 0.09 of 0.0422 by 0.0336 growls 0.33 to 0.0326 dog 0.0321 can 0.0309 size 0.0307 ...
Evaluation – Participation

<table>
<thead>
<tr>
<th>Groups that registered</th>
<th>104</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total submitted runs</td>
<td>58</td>
</tr>
<tr>
<td>Groups that participated</td>
<td>13</td>
</tr>
<tr>
<td>Groups that submitted working notes paper</td>
<td>9</td>
</tr>
</tbody>
</table>

Participants:

- **CEA LIST**: Vision & Content Engineering group of CEA LIST (Gif-sur-Yvettes, France).
- **INAOE**: Instituto Nacional de Astrofísica, Óptica y Electrónica (Puebla, Mexico).
- **KDEVIR**: Computer Science and Engineering department of the Toyohashi University of Technology (Aichi, Japan).
- **LMCHFUT**: Hefei University of Technology (Hefei, China).
- **MICC**: Media Integration and Communication Center of the Università degli Studi di Firenze (Florence, Italy).
- **MIL**: Machine Intelligence Lab of the University of Tokyo (Tokyo, Japan).
- **RUC**: School of Information of the Renmin University of China (Beijing, China).
- **SZTAKI**: Datamining and Search Research Group of the Hungarian Academy of Sciences (Budapest, Hungary).
- **THSSMPAM**: Jile Zhou (Beijing, China).
- **TPT**: CNRS TELECOM ParisTech (Paris, France).
- **UNED&UV**: Universidad Nacional de Educación a Distancia and Universitat de València (Spain).
- **UNIMORE**: University of Modena and Reggio Emilia (Modena, Italy).
- **URJC&UNED**: Universidad Rey Juan Carlos and Universidad Nacional de Educación a Distancia (Spain).
## Evaluation – Some of the submitted systems

<table>
<thead>
<tr>
<th>System</th>
<th>Visual Feat.</th>
<th>Training Data Processing</th>
<th>Annotation Technique</th>
</tr>
</thead>
</table>
| **TPT #6** | Provided by organizers | Tr. images selected/labeled by appearance of concept in webpage (+morphological expansions) | - Multiple SVMs with context dependent kernels  
- Annotation based on threshold |
| **MIL #4** | Fisher Vectors (SIFT, C-SIFT, LBP, GIST) | Tr. images selected/labeled by appearance of concept in webpage (+synonyms and hyponyms with a single meaning) | - Linear multilabel classifier learned by PAAPL  
- Annotation of the top 5 concepts |
| **UNIMORE #2** | Multiv. Gauss. Distrib. of local desc. (HSV-SIFT, OPP-SIFT, RGB-SIFT) | Tr. images selected/labeled by appearance of concept in webpage (+stopwords, stemming, synonyms, hyponyms and negative context disambiguation) | - Linear SVMs learned by stochastic gradient descent  
- Annotation based on threshold |
| **RUC #6** | Provided by organizers | Positive Tr. images selected by a combination of text feat. and Flicker based weighted search engine keywords. Negative examples selected by Negative Bootstrap. | - Multiple staked hikSVMs and kNNs  
- Annotation of the top 6 concepts |
The ImageCLEF 2013 Scalable Concept Image Annotation Subtask
Evaluation – Concept $F_1$ boxplots for all runs
Evaluation – Concept $F_1$ boxplots for all runs
1 Introduction
   - Motivation

2 Subtask Description
   - Lines of work
   - Web training dataset

3 Evaluation
   - Participation
   - Results

4 Conclusions and Future Work
Conclusions and Future Work

- Participation was excellent, and the teams presented diverse approaches to address the proposed challenge.

- The results indicate that the web data can be effectively used for training practical and scalable annotation systems.

- The performances improved from a baseline below 10% to over 40% for both MAP and $MF_1$ measures.

- The performance for the concepts not seen during development demonstrates potential for scalability of the systems.

- Comparing the systems, several of the proposed ideas are complementary, thus future improvements are expected.
This task has attracted considerable interest, so we decided to continue it for ImageCLEF 2014.

Ideally more testing data should be used to obtain more conclusive results related to the performance of unseen concepts.

Modifications for the task, e.g. use both supervised and unsupervised data.

Try the same ideas in other tasks, e.g. video.
Thank you for your attention!

Questions? Comments?