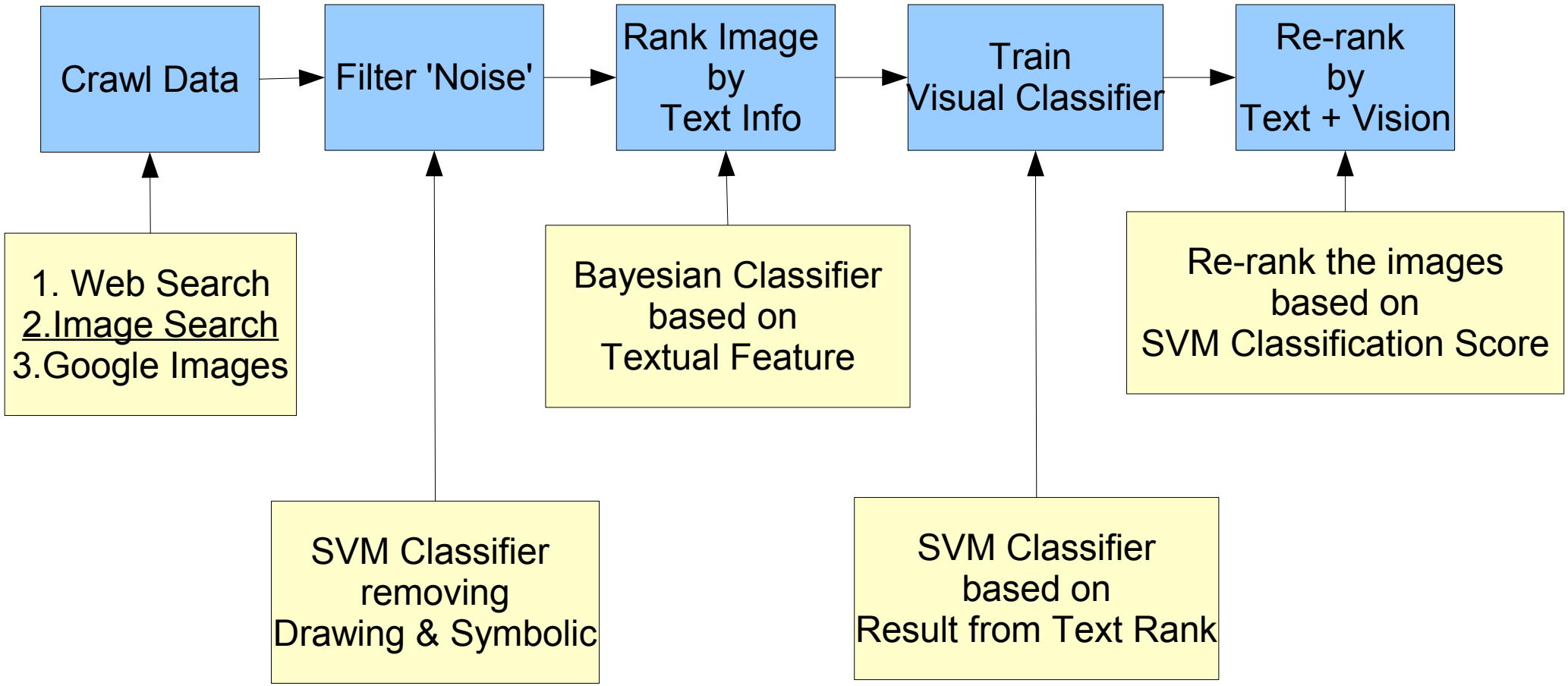
Harvesting Image Databases from the Web

- Dongliang Xu

- 15th.2.2008

Overview of Text-Vision Image Harvesting Algorithm

Flowchart of Original Version



Crawl Images

- **WebSearch:** Submits the query word to Google web search and all images that are linked within the returned web pages are downloaded. (limit 1000 pages)
- **GoogleImages:** Download images directly returned by Google image search.
- **ImageSearch:** Each of the returned Google Image Search is treated as a “seed” - further images are downloaded from the web page from where the seed image originated.

Crawl Images

- **in-class-good:** Images that contain one or many class instances in a clearly visible way (without major occlusion, lighting deterioration or background clutter and of sufficient size).
- **in-class-ok:** Images that show parts of a class instance, or obfuscated views of the object due to lighting, clutter, occlusion and the like.
- **non-class:** Images not belonging to in-class.
-
- **The good and ok sets are further divided into two subclasses:**
-
- **abstract:** Images that don't look like realistic natural images (e.g. drawings, non realistic paintings, comics, casts or statues).
- **non-abstract:** Images not belonging to the previous class.

Crawl Images

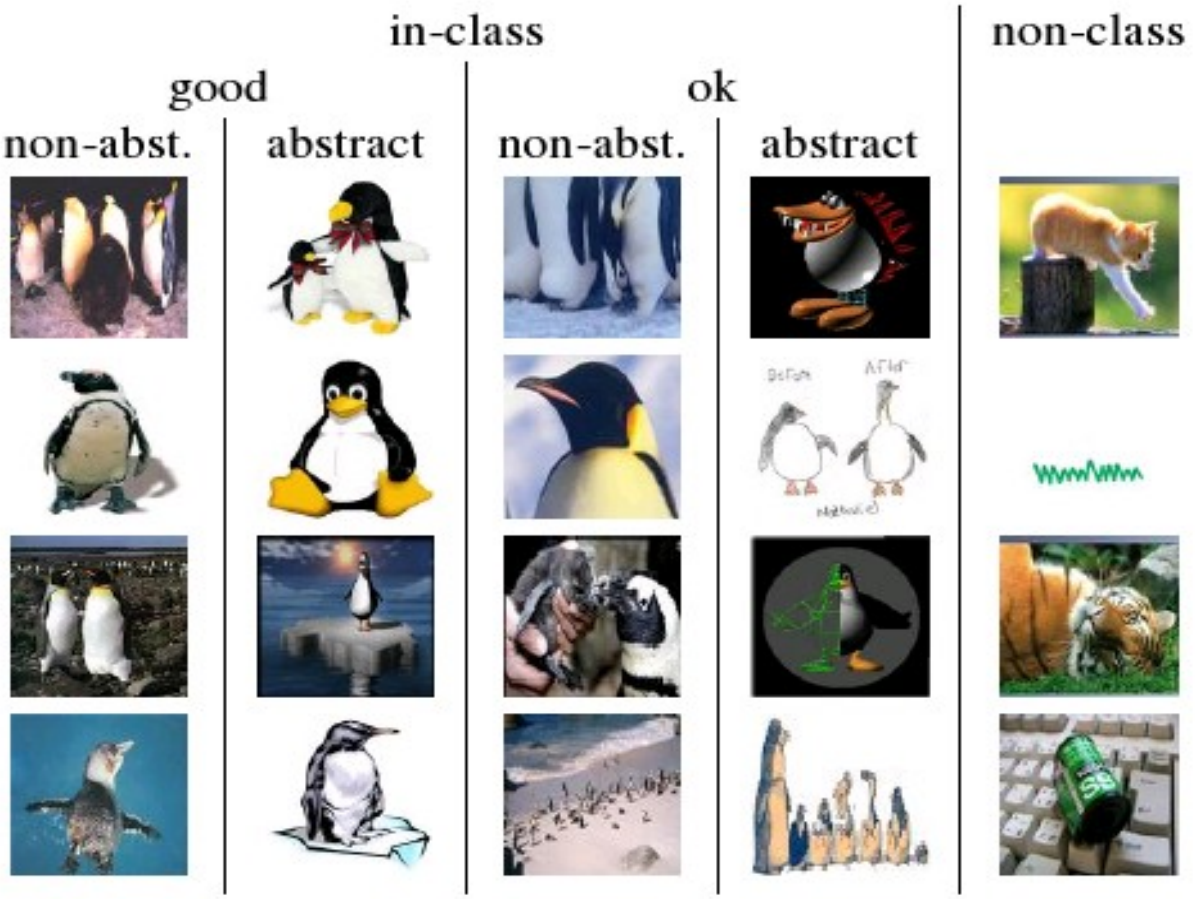


Figure 2. **Image annotations:** Example images corresponding to annotation categories for the class penguin.

Crawl Images

Service	in-class	non-class	precision
WebSearch	8773	25252	26%
ImageSearch	5963	135432	4%
GoogleImages	4416	6766	39%

Table 1. **Statistics by source**: The statistics of downloaded images for different retrieval techniques.

Removing Drawing & Symbolic Images

- These images include: comics, graphs, plots, maps, charts, drawings and sketches.

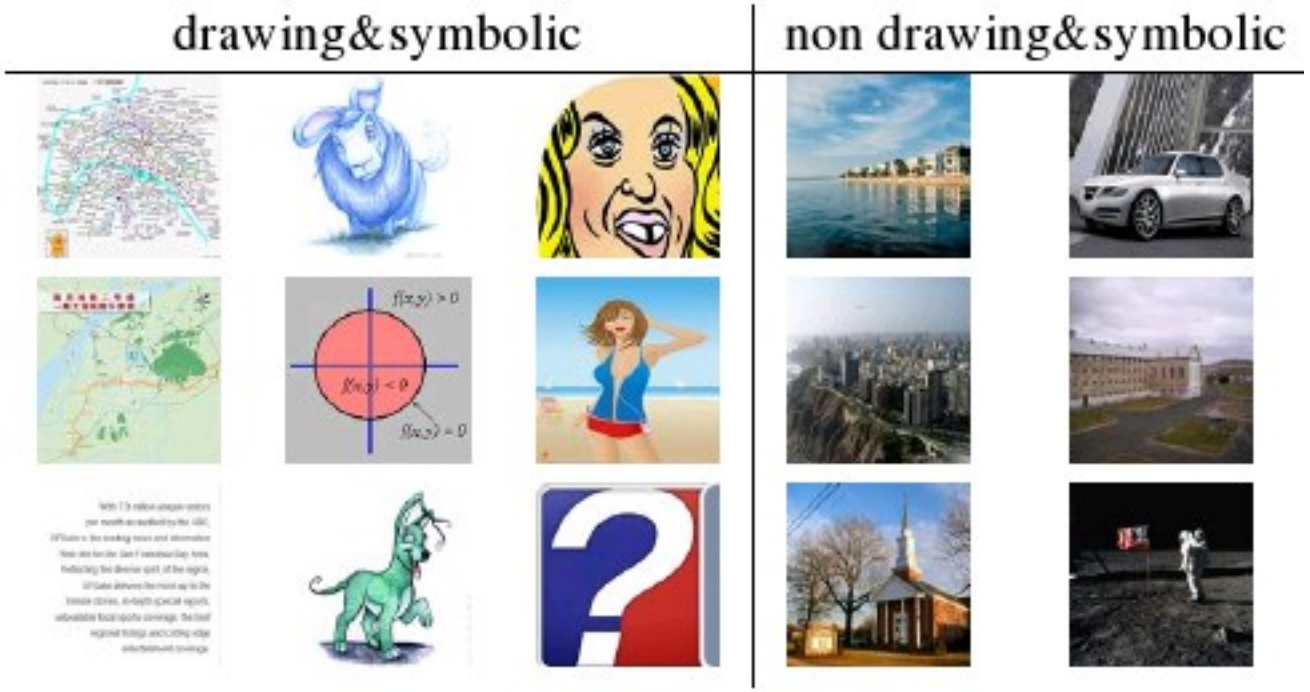


Figure 3. **Drawings&symbolic images:** Examples of positive and negative training images.

Removing Drawing & Symbolic Images

- **Vector**(1000 equally spaced bins)
  - a color histogram
  - a histogram of the L2-norm of the gradient
  - a histogram of the angles (0... π) weighted by the L2-norm of the corresponding gradient
- **Classifier**
  - A radial basis function Support Vector Machine(SVM)

Removing Drawing & Symbolic Images

- **Positive Samples(2000):** any non drawings&symbolic images

- **Negative Samples(1400):** images downloaded from queries 'sketch','drawing' or 'draft'.

The method achieves around 90% classification accuracy on the drawing&symbolic images using two-fold cross-validation

Removing Drawing & Symbolic Images

- Removing an average of 42% **non-class** images
- Removing an average of 60%(123 images) **in-class abstract** images with a range between 45% and 85%
- Removing an average of 13%(90 images) **in-class non-abstract** images

Ranking on Textual Features

- Textual Features
  - *filedir*
  - *filename*
  - *imagealt*
  - *imagetitle*
  - *websitetitle*
  - *context10*: includes the ten words on either side of the image-link
  - *contextR*: describes the words on the web-page between eleven and 50 words away from the image-link

Ranking on Textual Features

- **Structure**

.....

I offer some worthwhile advice this time. If you are going to purchase (moderately)

.....

- **The seven features define a binary feature vector for each image**

a=(a1,.....,a7)

(a stop list and a stemmer used in this process. Word Breaker?)

Ranking on Textual Features

- A simple Bayesian posterior estimation

$$P(y|\mathbf{a}) = P(\mathbf{a}|y)P(y)/P(\mathbf{a})$$

$$P(\mathbf{a}|y) = P(a_1, \ldots, a_4|y) \prod_{5}^{7} P(a_i|y)$$

$$y \in \{in\text{-}class, non\text{-}class\}$$

where $P(a_1, \ldots, a_4|y)$ is the joint probability of the first four textual features (*contextR*, *context10*, *filedir*, *filename*).

Ranking on Visual Features

- Vector
  - Build Visual Words Histogram from all images crawled.
- Classifier(for each class)
  - A radial basis function Support Vector Machine(SVM)

(SVM light)

Ranking on Visual Features

- Positive Samples: Top 250/150 images from text rank
- Negative Samples: Any images(250/500/1000) from other class
- Re-rank based on SVM classification score

Ranking on Visual Features

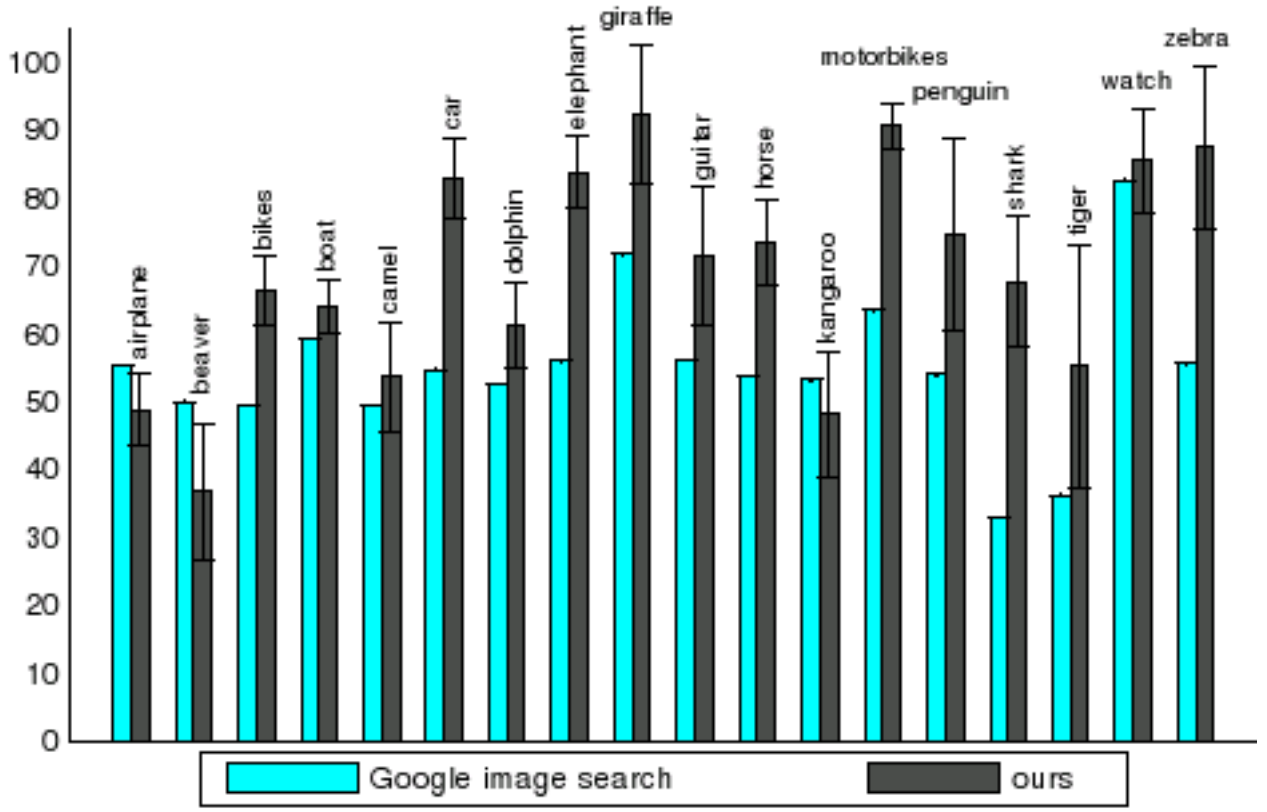


Figure 7. **Comparison with Google image search.** Precision at 100 image recall.

Ranking on Visual Features

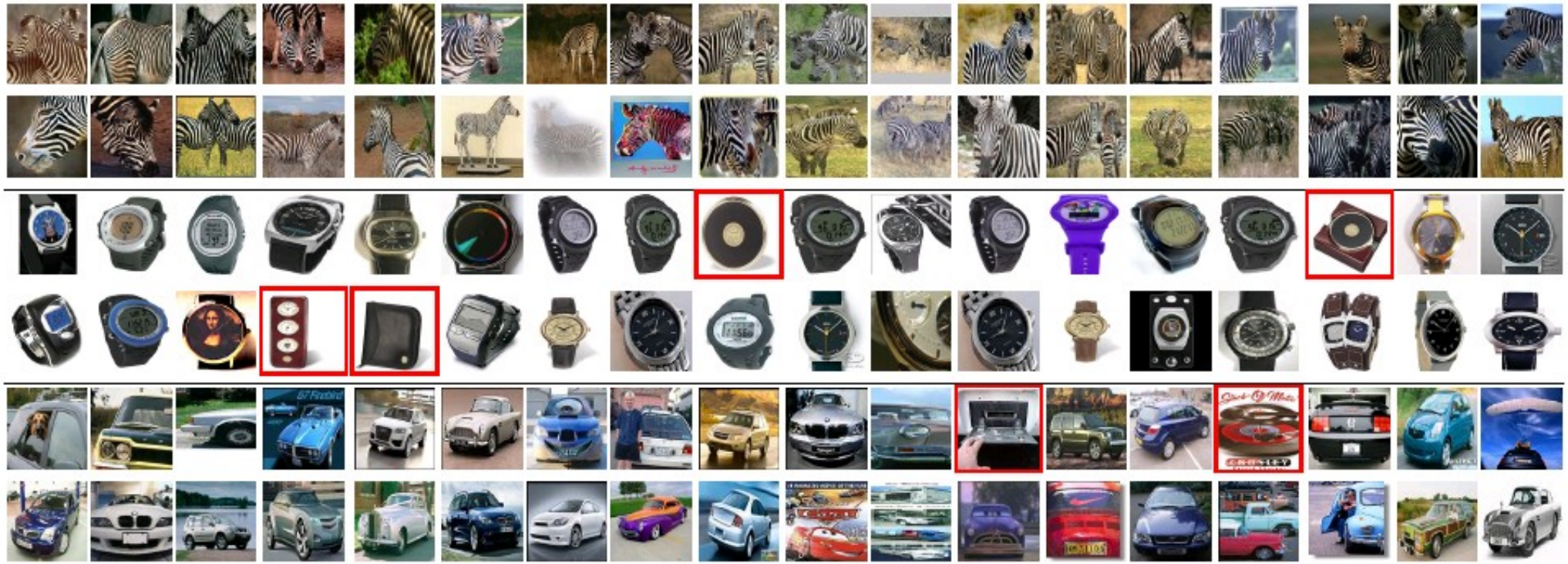
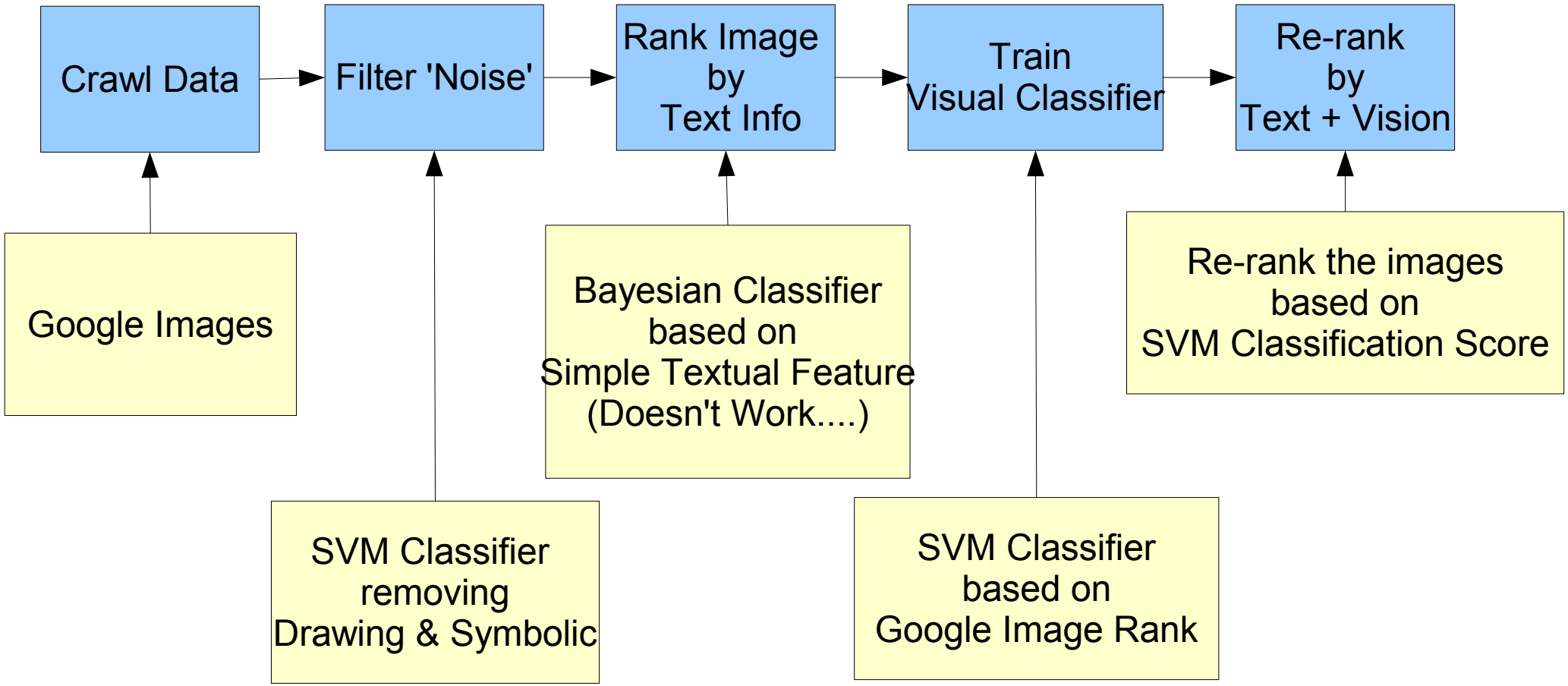


Figure 6. **Top ranked 36 images** of zebra, wristwatch and car using the text+vision algorithm of figure 5. Red boxes indicate false positives.

Overview of Text-Vision Image Harvesting Algorithm

Flowchart of Distilled Version



Crawl Data

- Goal: Images are crawled from Google Image Search, when info and related data are stored in MYSQL.
- Tools: Perl Module Package(WWW::Google::Images, WWW::Mechanize)
- Problems:

1. Fail to crawl part of data due to temporary connection failure or IP block.
2. 1000 Image Limitation

Ground Truth Annotation

- Images are divided into three categories: in-class-good, in-class-ok , non-class(by myself.....)

URL:http://www.v-flyer.com/pages.asp%3Fpageid%3D206
ImageIndex:1
Good:non-abstract None Good:non-abstract Good:abstract OK:non-abstract OK:abstract non-class

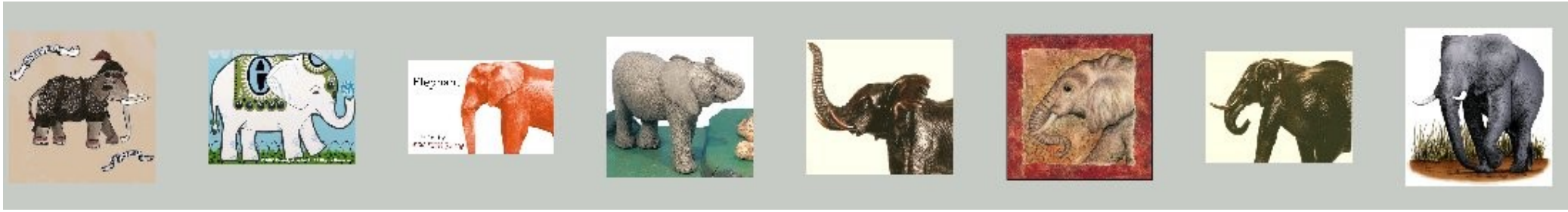
URL:http://animals.nationalgeographic.com/animals/mammals/african-elephant.html
ImageIndex:2
Good:non-abstract None Good:non-abstract Good:abstract OK:non-abstract OK:abstract non-class

URL:http://fohn.net/elephant-pictures-facts/

Ground Truth Annotation

- in-class-real

- in-class-abstract



Ground Truth Annotation

- Statistics

Keyword	IN-CLASS	NON-CLASS	REAL/ABSTRACT	Prec.
elephant	323	433	3.82	0.43
car	367	395	6.64	0.48
panda	302	504	5.57	0.37
tiger	199	680	5.03	0.22
teapot	526	208	6.41	0.72
zebra	236	575	5.05	0.29

Keyword	IN-CLASS	NON-CLASS	REAL/ABSTRACT	Prec.
elephant	326	430	3.66	0.43

- Problems:

1. Labeling should be performed by individual who has no knowledge about the algorithm.(I do it by myself...)

2. many ambiguous images

3. more specific query? (such as '2008 Honda Civic', you can try it in home)

Removing Drawing & Symbolic Images

- Vector: A histogram of the angles(0..2π) weighted by the L2-norm of the corresponding gradient.
- Classifier: A radial basis function SVM on a hand-selected dataset
- (1800)Negative samples from 'draft','cartoon','animation','sketch' and 'drawing'.
- (1200)Positive samples from 'photo','realphoto','shot' and 'real'.
- Tools(OPENCV, LIBSVM)

Removing Drawing & Symbolic Images

- Statistics

Keyword	IN-CLASS	NON-CLASS	REAL/ABSTRACT	Prec.
elephant	263(323)	277(433)	5.57(3.82)	0.487(0.43)
car	277(367)	239(395)	16.3(6.64)	0.536(0.48)
panda	269(302)	307(504)	6.47(5.57)	0.467(0.37)
tiger	141(199)	428(680)	9.07(5.03)	0.247(0.22)
teapot	326(526)	116(208)	8.88(6.41)	0.737(0.72)
zebra	158(236)	322(575)	9.53(5.05)	0.329(0.29)

- Problems:

1. Typical failure on the static object (teapot, wristwatch, see figure 6).

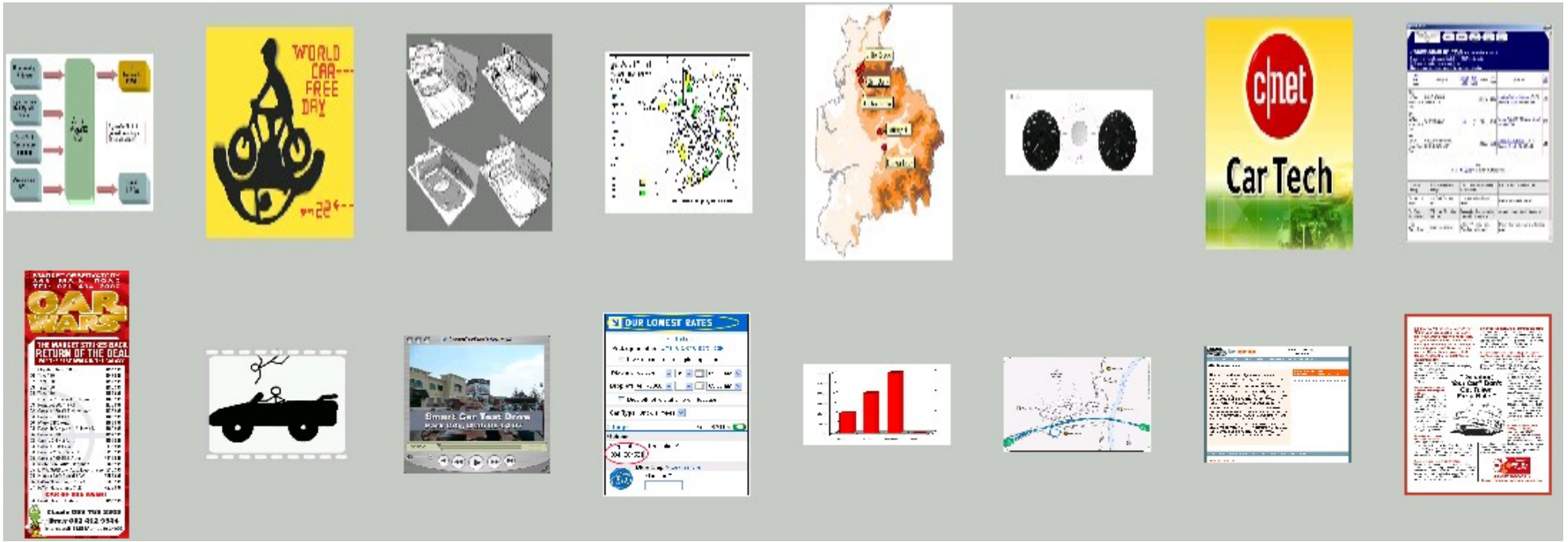
Removing Drawing & Symbolic Images

Keyword	in-cl-real	in-cl-abstract	non-cl
motorbikes	615	89	981
wristwatch	903	13	982
panda	256	46	504
teapot	455	71	208

filter

in-cl-real	in-cl-abstract	non-cl
522	49	593
656	2	478
233	36	307
293	33	116

Removing Drawing & Symbolic Images



Rank Image by Text Information

- Vector: 6-dimension binary vector

(filedir, filename, websitetitle, context, alt, title)

- Classifier: Naïve Bayes, all are i.i.d.
- No Stop List Used (a, the, however.....)
- No Word Breaker Used (realphoto, real-photo -> real photo)
- No Stemmer Used(bikes -> bike, further -> far)
- Tools: Perl Module Package (WWW::Mechanize)

Rank Image by Text Information

- Structure

.....

I offer some worthwhile advice this time. If you are going to purchase (moderately)

.....

Problems:

1. My rank performance is definitely worse than Google Image Rank. (As I expect.....)
2. I really want to know text rank performance respectively on

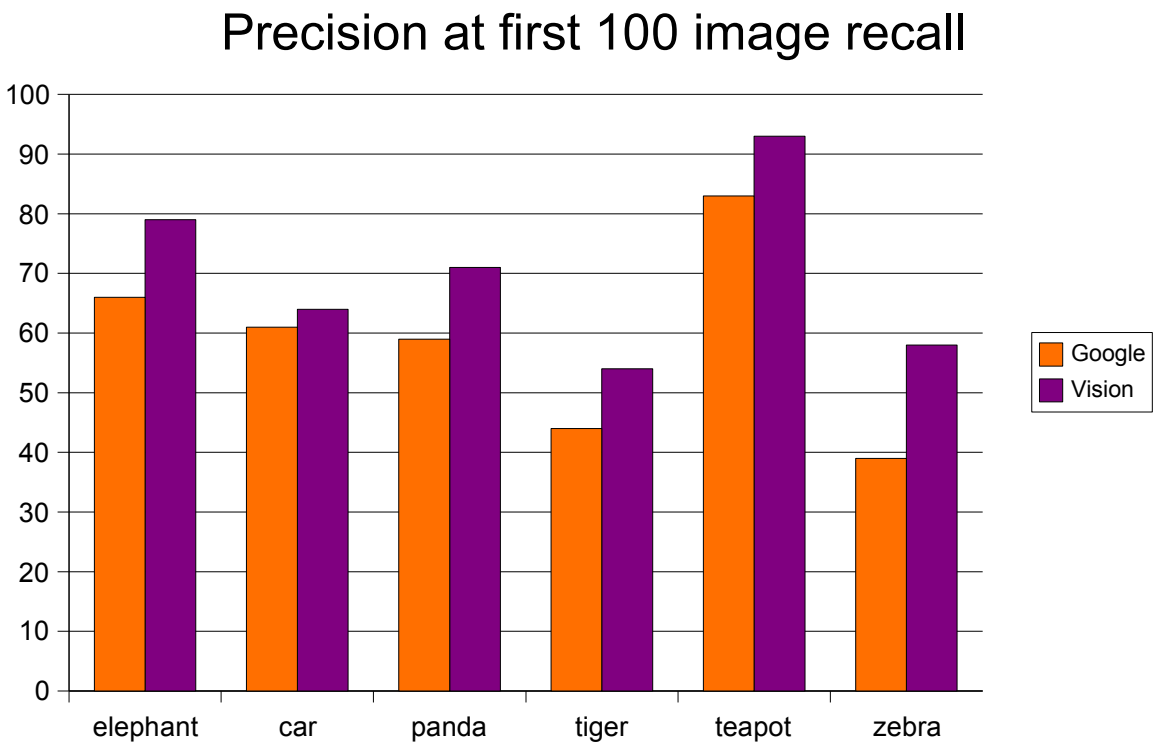
Web Search VS. Google Image Search

Ranking on Visual Features

- Top 50 Google images results are good enough?
- 400 Visual Words obtained from the whole image set.
- Vector: Histogram of Visual Words
- Classifier: A radial basis function SVM with probability estimates
- Re-rank based on the probability value from SVM prediction.

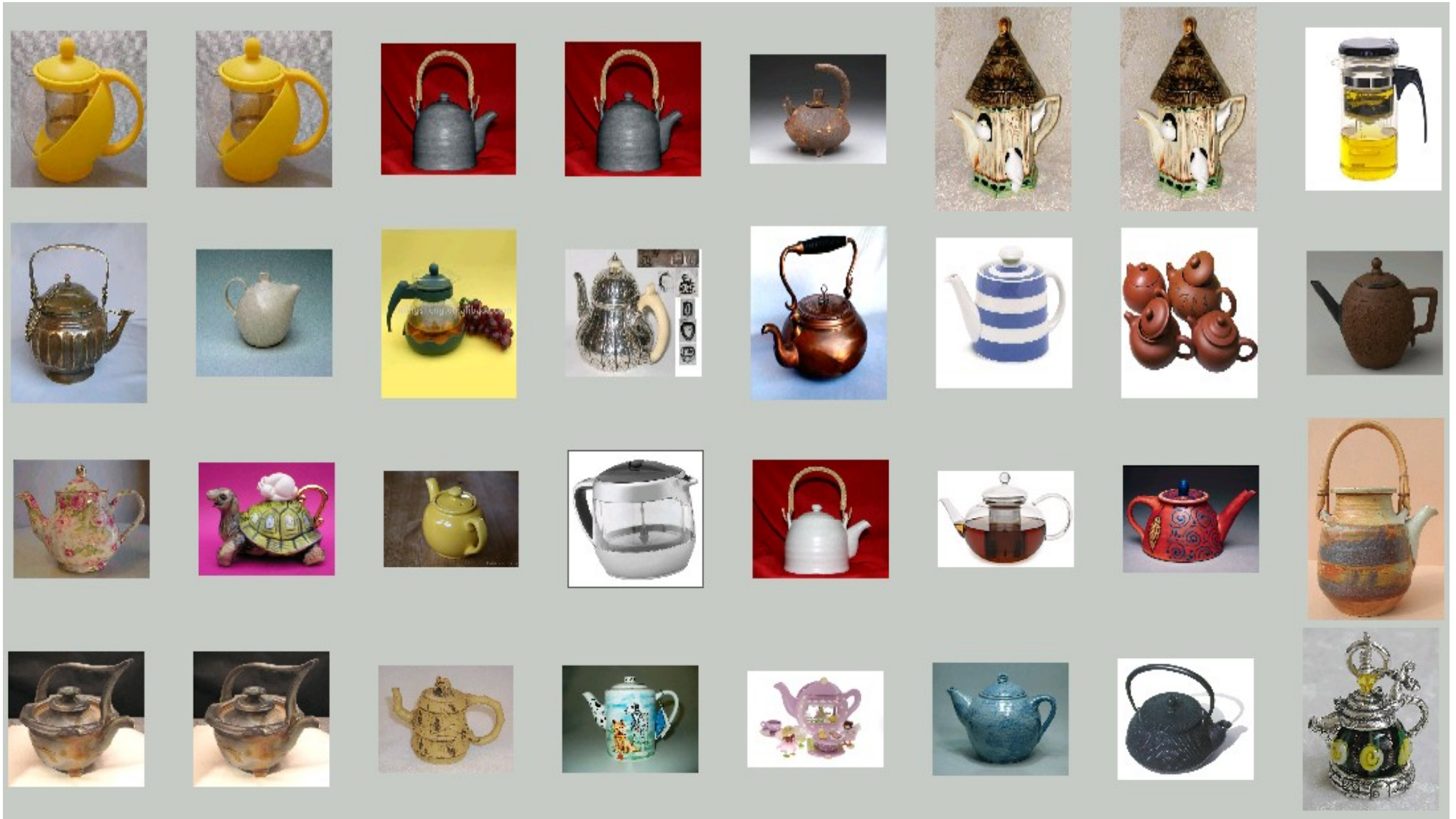
Ranking on Visual Feature

- Statistics

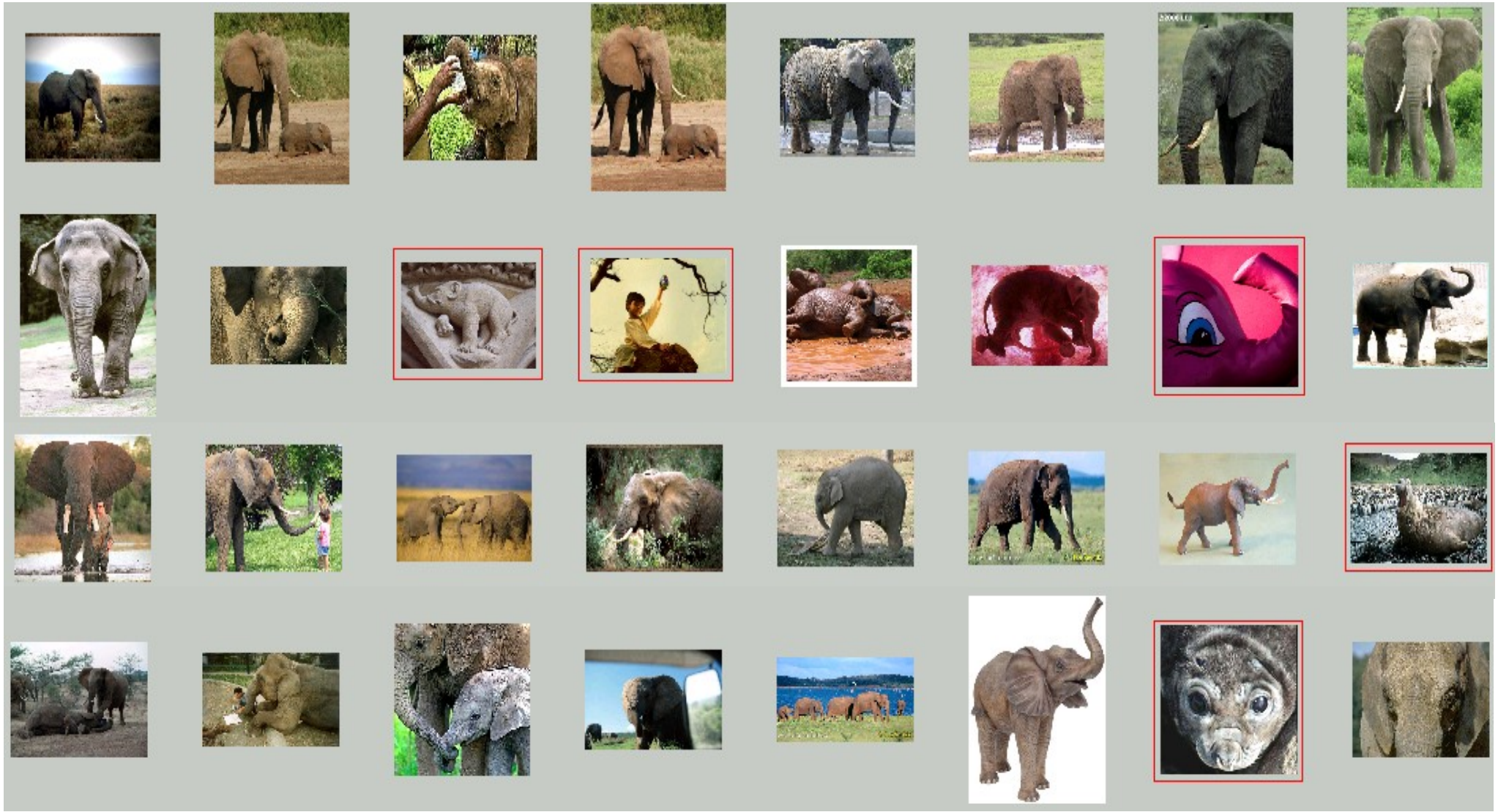


Ranking on Visual Feature

Ranking on Visual Feature



Ranking on Visual Feature



Tools

- MySQL 5.0
- Perl Module
  - GoogleImage
  - Mechanize
  - PerlMagick
- OPENCV
- Affine Covariant Region Detectors
- Comparison of Affine Region Detectors
- LIBSVM

Summary

- Add new image source
- Reverse part of the sequence
- Add other step into the whole structure
- Mining the knowledge from query

  http://adlab.microsoft.com/
- Mining the knowledge from the webs
- New method to combining text and visual features

Thank You!