Evaluation Strategies for Image Understanding and Retrieval

Keiji Yanai

(Univ. Electro-Communications, Japan & former Visiting Scholar at Univ. Arizona)

Nikhil Shirahatti, Prasad Gabbur, and Kobus Barnard

(Univ. Arizona, USA)

Our Three Evaluation Projects

(1) Evaluation of low and mid-level algorithms in the word-image-translation [CVPR 01]

Algorithms to extract features

(2) Evaluation of image retrieval methods [CVPR 05]

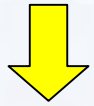
CBIR systems

(3) Evaluation of "visualness" of words [sp]

Words to be annotated with

Introduction: We need “evaluation” !

- Now we have a huge number of images.



- In MM, CV and IR , we are eagerly developing methods to infer semantics from them.

- We need “Evaluation Strategies”
  to compare many methods
  in the comprehensive points of view.

Projects-1

Evaluation of low and mid-level algorithms in the word-image-translation

[CVPR 01]

Word-image-translation

Input

Output

grass
cat
tiger
rock

Word to image

[ICCV01]

Words to regions

[ECCV02]

We proposed two types of annotation.

word to image

word to regions

Segmentation & Image features

Input

sun sky waves sea

Segmentation

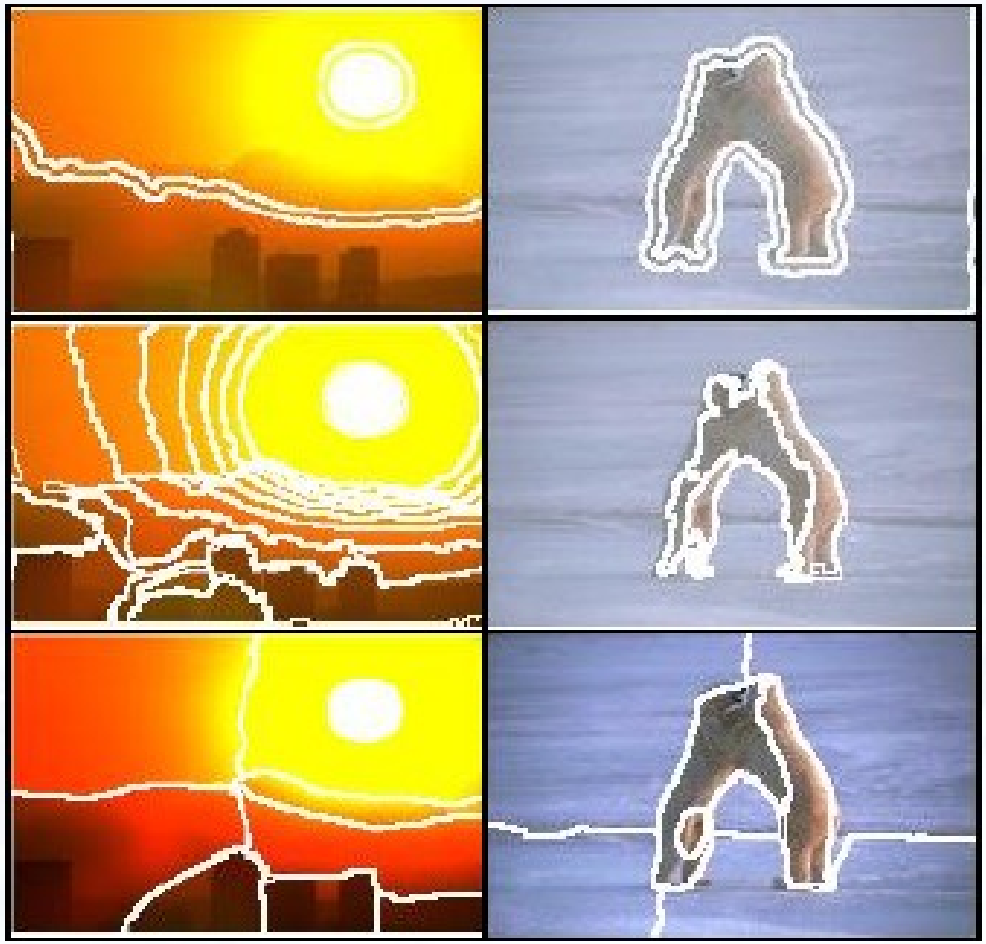
Blobworld
Mean-shift
N-Cut

Each regions is a large vector of features

- Region size
- Position
- Color
  - ✓ RGB, L*a*b* or rgS
- Texture
  - ✓Oriented energy (12 filters)
  - ✓Response to DOG (4 filters)
- Shape features

We have many combinations of segmentation and features.

Sample Segmentations



Blobworld
[UCB 02]

Mean-Shift
[Rutger Univ. 02]

Normalized cuts
[UCB 00]

Measuring Annotation Performance

predicted words

CAT HORSE
GRASS WATER

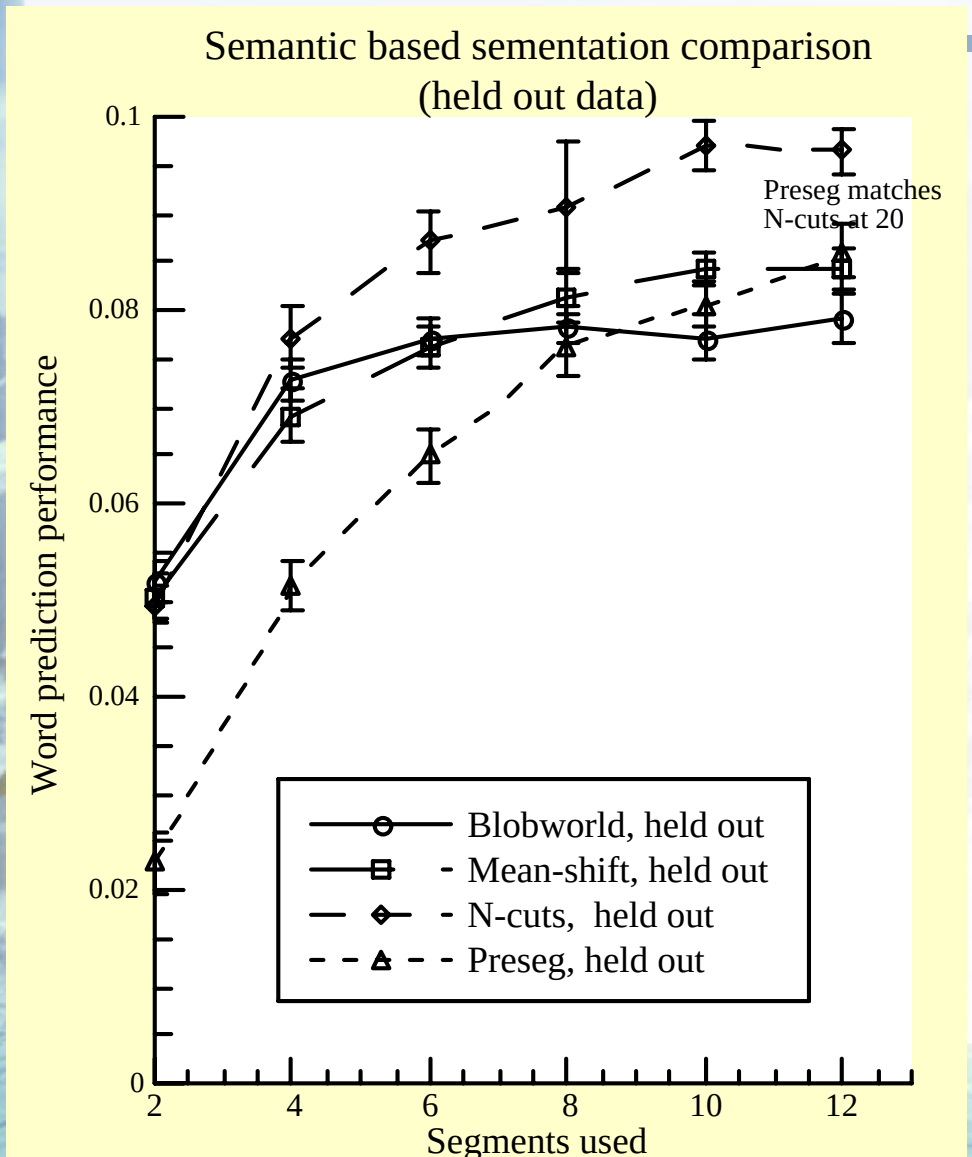
Compare the ratio of overlap

actual keywords

GRASS TIGER
CAT FOREST

We compare segmentation methods and combinations of features with this annotation performance.

Segmentation evaluation



The performance varied from 0.075 to 0.095

- N-cuts *outperformed* Mean-shift & Blobworld.

Feature evaluation

Feature set	Performance
Base set	0.020
Base set, RGB	0.057
Base set, L*a*b	0.085
Base set, rgS	0.092
Base, rgS, color context	**0.094**
Base set, texture	0.048
Base, rgS, texture	0.072
Base, RGB, color context, texture	0.073
Base set, shape	0.016
Base set, rgS, shape	0.029
Base,rgS, texture, shape	0.043
Everything	0.055

Base set:

- Size
- Location
- First moment
- Area / $(\text{Perimeter})^2$

Varied from 0.016 to 0.094

Color is the most important.

La*b* and rgS are better than RGB.

12 types of combinations of feature sets

Projects-2

Evaluation of image retrieval methods

[CVPR 05]

Objective

- **Develop a comprehensive method** and **provide ground truth data** *to evaluate image retrieval algorithms or systems.*
  - **Human-centered evaluation**
  - Fully automatic evaluation
  - Independent of image retrieval system
  - Open calibration/evaluation software and ground truth data available at our Web site **http://kobus.ca/research/projects/cbir-eval/**

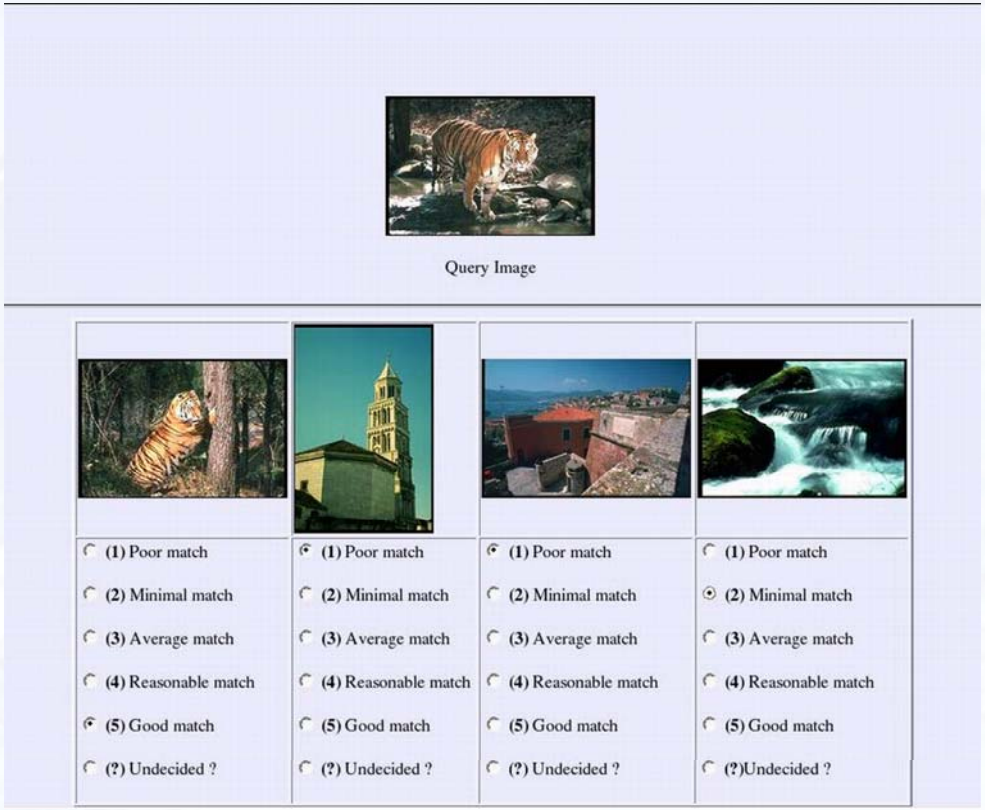
Preparation for evaluating your system

- You need:
  - CBIR system to be evaluated
  - COREL image data sets
  - System output: many query-result pairs with score
    (assuming query by one image and no feedback)
- We provide (at our Web site) :
  - Human-evaluation data
    - 20000 pairs of
      <query COREL image id> <result image id><human score>
  - Calibration and evaluation software
    - to measure the performance of your CBIR system

Collecting human-eval. data

- Collected 20,000 query-result pairs
- 32 participants
- Calibrated for participants variance.

Common ground truth data

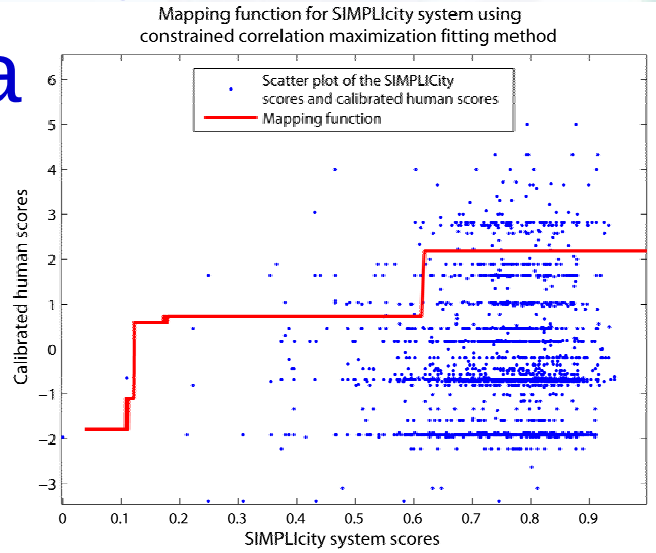


Web interface for collecting data
(4 pairs on one page)

5-step human evaluation

Mapping CBIR to human score & calculating their correlation

- Estimate mapping function with common ground truth data
  - Monotonic constraint
  - 3 methods to map
    - Least Mean Square
    - Correlation Maximization
    - Bayesian Inference



- Calculate correlation of GT and estimated human score

Correlation between real and estimated human score

What we want to measure

System performance

Case-study:

- Evaluate 4 CBIR systems
  - GNU Image Finding Tool (GIFT)
  - SIMPLIcity [J.Z.Wang 01]
  - Our Translation Model [ICCV01]
  - Corel keyword-based search (text search)

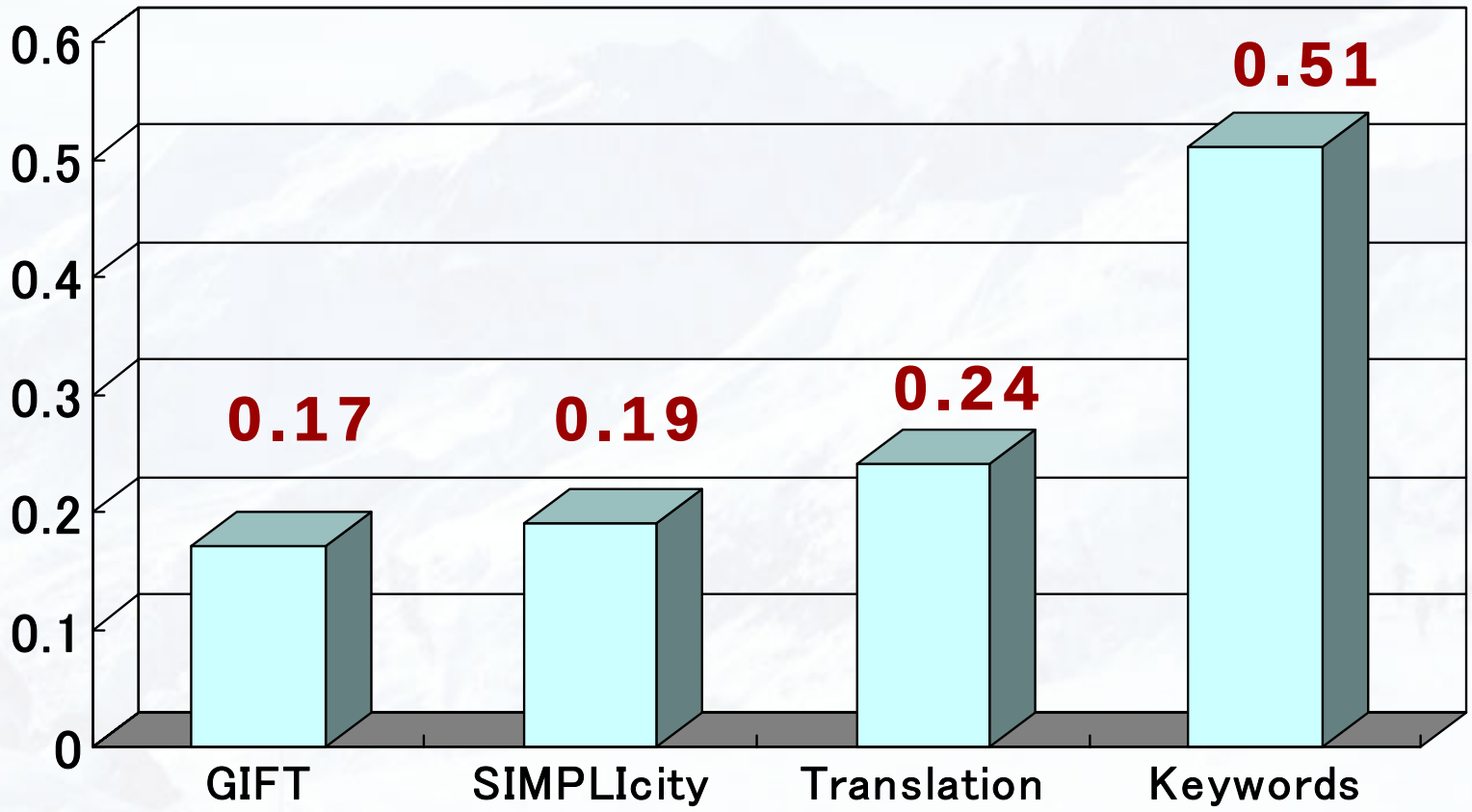
e.g. bear, river, animal → fox, river, animal → 2/3=0.667

query

result

system score

Results: correlation between human and system scores



Keywords can represent semantics much better than image features.

Projects-3

Evaluation of "visualness" of words

[ACM MM 05 short paper]

Motivation

- A lots of words for annotation of images

Corel ID 108041

tiger feline cat
mammal animal wildlife
grass forest

8 words [Corel image gallery 1,000,000]

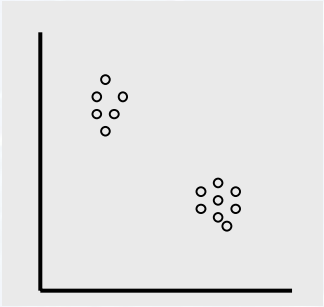
e.g. “Mammal” is classified based on the way of their birth, not based on their appearance.

Some words are not appropriate for image recognition.
Words related to “visual properties” are good for that.

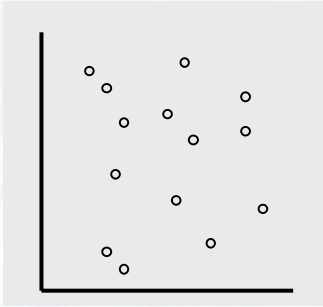
We need to evaluate “visualness” of words.

Image region entropy

- A measure of “visualness” of words (concepts)
- Represent the property of the distribution of image features



Biased / uneven:
low entropy
having “visualness”



Random/uniform:
high entropy
not having “visualness”

- Need no ground truth data unlike rec.-prec. diagram
- To get images associated to the given word, use **images on the Web with Google**
  - Enables us examine about anv words automaticallv

To examine “image region entropy”,
we have to provide only a concept keyword at first.

Method: prepare generic model

- To make “entropy” meaningful, select “X” regions, excluding backgrounds with a probabilistic method. (same as prob. Web gathering)
- Calculate the entropy of the “X” regions with respect to a generic model
  - Build a generic distribution model of region features of randomly collected images in advance

random Web images

Case study:

Finding “visual” adjectives

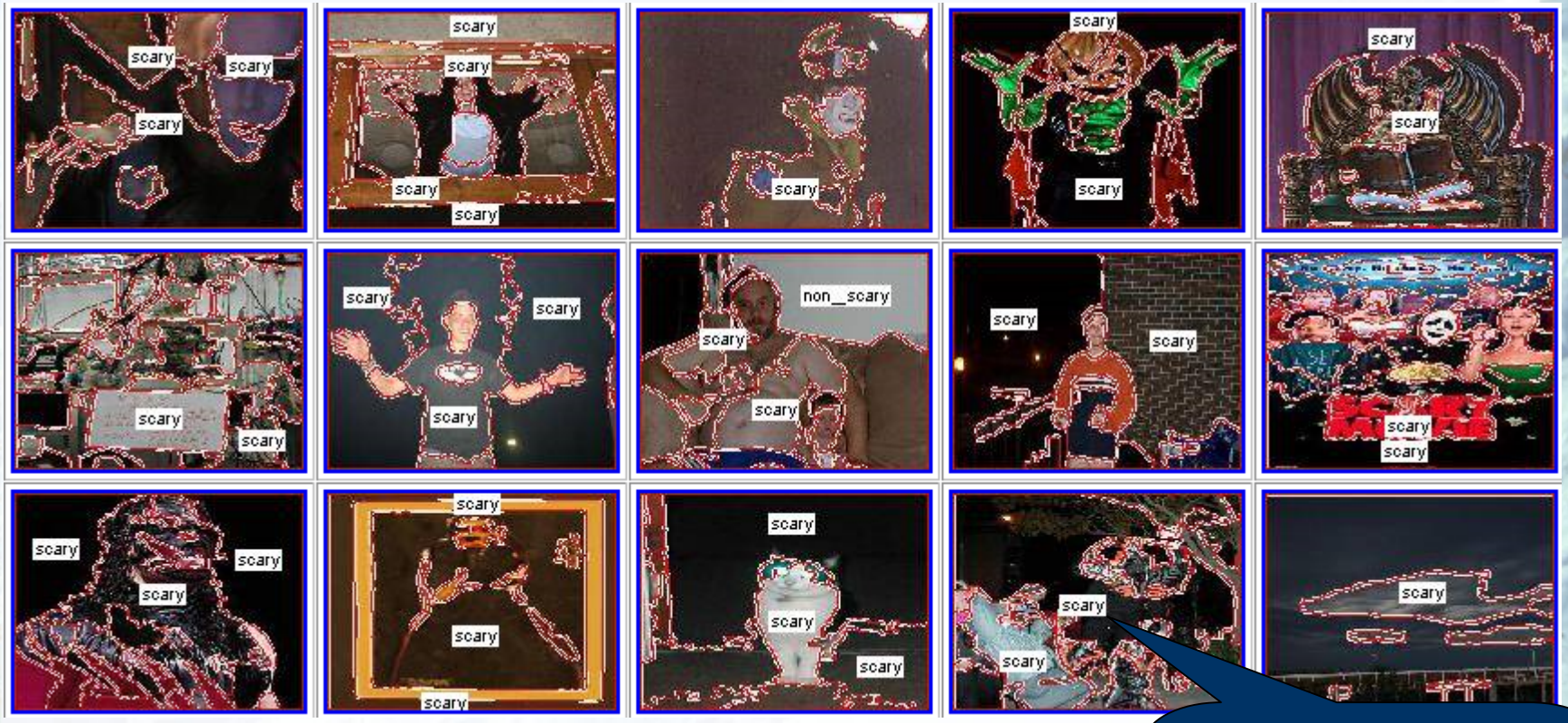
- Collect 250 images per word for 150 adjectives using Google Image Search
- Our model detects regions related to the concept of the given word without any prior knowledge

images with “yellow” regions

Experimental results

- Low entropy (“visual” adjectives)
  - *dark visual rusted purple black shiny scary…*
- High entropy (“non-visual” adjectives)
  - *medical famous angry large open acoustic religious…*

Low entropy: “scary”



“Visual” adjective

High entropy: “famous”

“Non-visual” adjective

Conclusion

- We introduced our three projects related to evaluation briefly:
  - Segmentation algorithms and combinations of image features
  - Image retrieval systems
  - Words to be annotated with

Thank you!

If you are interested in our projects, please visit

http://kobus.ca/