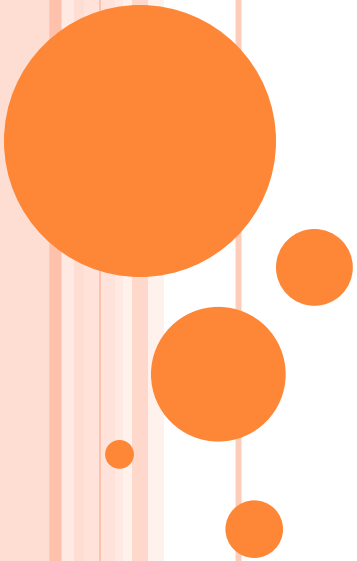


CS598 Visual Info. Retrieval

Words and Pictures

Slides tailored from Cees

Lecture IX

Part I: Matching Words and Pictures

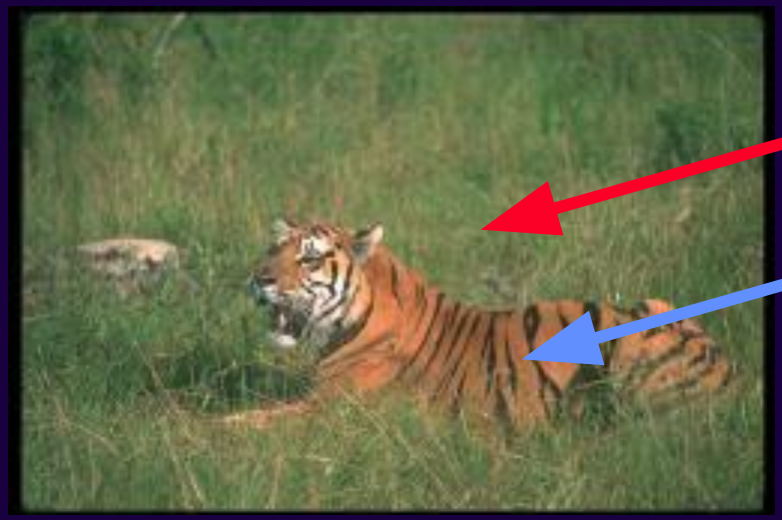
By Kobus Barnard

University of Arizona

Visual Representation ➡ Semantic Representation

➡ A tiger lying in the grass

Visual Representation ➡ Semantic Representation



grass

tiger

Recap: concept detection

Finding words for the images

tiger grass cat

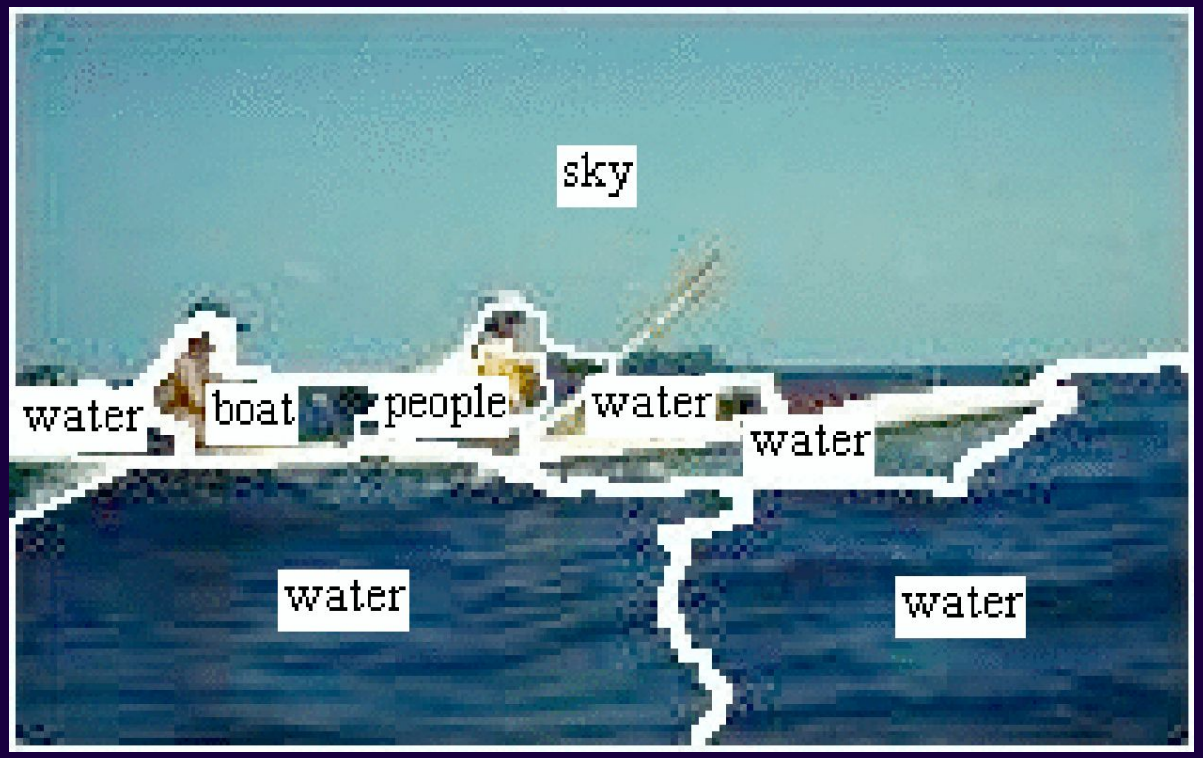
See lecture on concept detection

Concept detection vs Recognition

?

tiger cat grass

Recognition



Semantic representation includes not only what is there, but **where** it is

General Approach

Learn models for annotation and recognition from large image data sets with associated text

Key Point

Learn from data without explicit correspondence between image components

Data with correspondence ambiguity is common

- Images with associated text
- Video (which frame goes with which speech or text)

Key Point (cont)

Trade quality for quantity (and realism)

Sources of information

- A word (tiger) is much more likely than chance to have something to do with the image
- If a word refers to something in the image (tiger), it is less likely to refer to something else
- Relationship between visual information and words has structure across images

Statistical Machine Translation

Data: Aligned sentences, but word correspondences are unknown

“the beautiful sun”

“le soleil beau”

Brown, Della Pietra, Della Pietra & Mercer 93

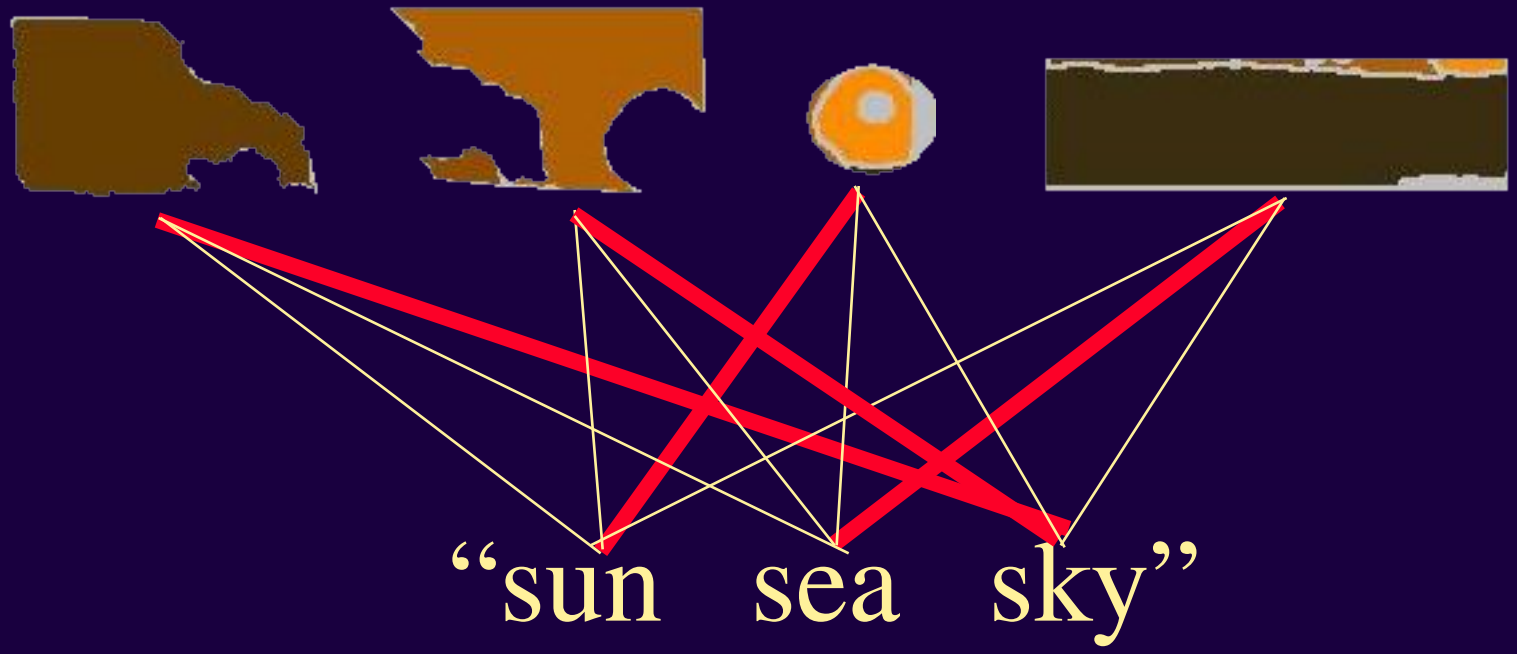
Quiz: How to learn word correspondences?

With one sentence we cannot say anything. But if we have

The beautiful sun, beautiful dog, beautiful girl…

Then, we can learn that beautiful goes to beau not to soleil or le.

Basic idea: Multimedia Translation



Approaches

Discretize (tokenize) blobs [Duygulu, Barnard, de Freitas, Forsyth, ECCV 02]

Simultaneously learn blob models and translation [Barnard et al, JMLR 03]

Multiple instance learning with support vector machines [Andrews et al, NIPS 02]

Integrate context into features [Barnard et al. CVPR 03] and into the model [Carbenetto et al. 03]

Composite models [Barnard et al. CVPR 03, Wachsmuth et al, 03]

Corel Database

118011
WATER HARBOR
SKY CLOUDS

TIGER CAT WATER GRASS

1090
SUN CLOUDS
WATER SKY

1015
SUN TREE
PLAIN SKY

143078
MOUNTAINS TREES
aspens VALLEY

102042
MUSEUM memorial
FLAGS GRASS

119094
GARDEN BUILDING
FLOWERS TREES

131007
GARDEN FLOWERS
HOUSE TREES

392 CD's, each consisting of 100 annotated images.

Input

Image processing*

sun sky waves sea

Each region is described by a set of features

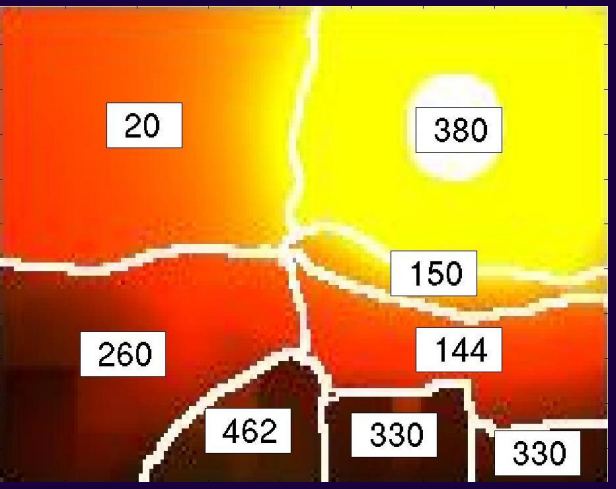
- Region size
- Position
- Color
- Oriented energy (12 filters)
- Simple shape features

*Thanks to Blobworld team [Carson, Belongie, Greenspan, Malik], N-cuts team [Shi, Tal, Malik]

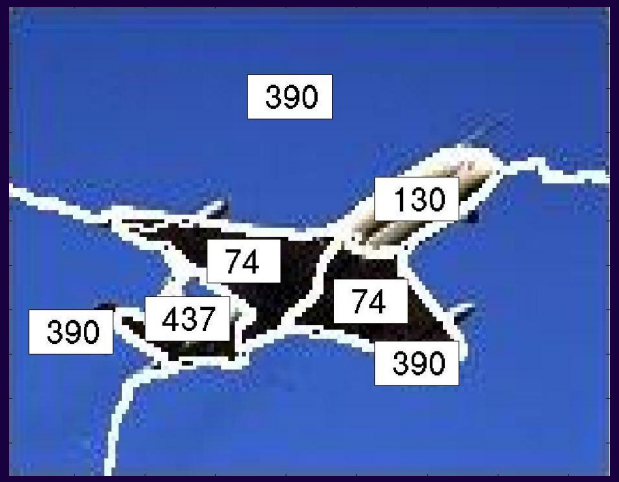
Discrete Model [ECCV 02]

Straightforward adaptation of machine translation

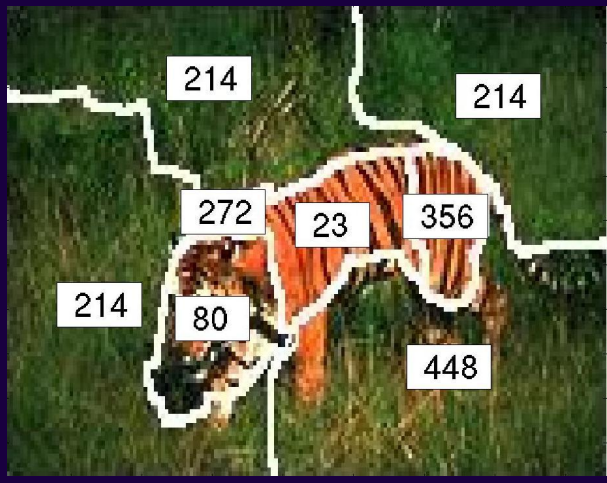
Need to vector quantize blobs (simple but better to simultaneously learn blob model)



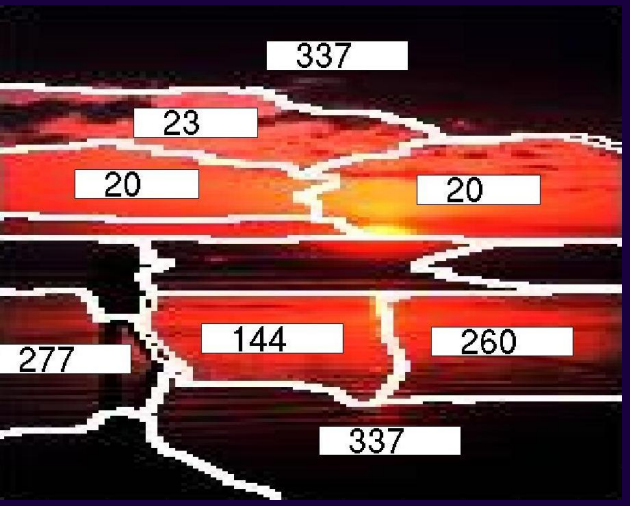
city mountain sky sun



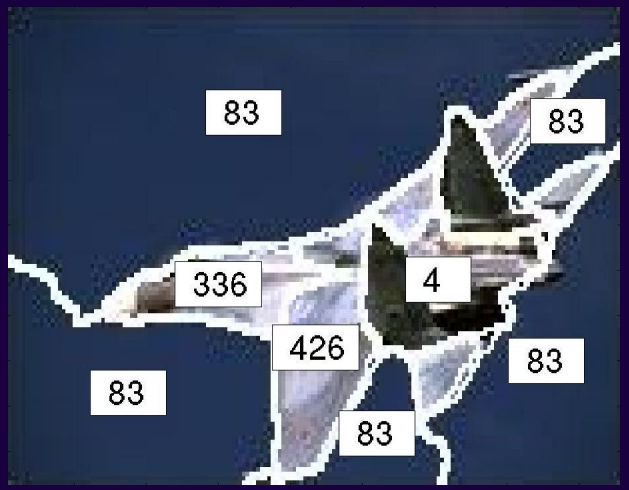
jet plane sky



cat forest grass tiger



beach people sun water



jet plane sky

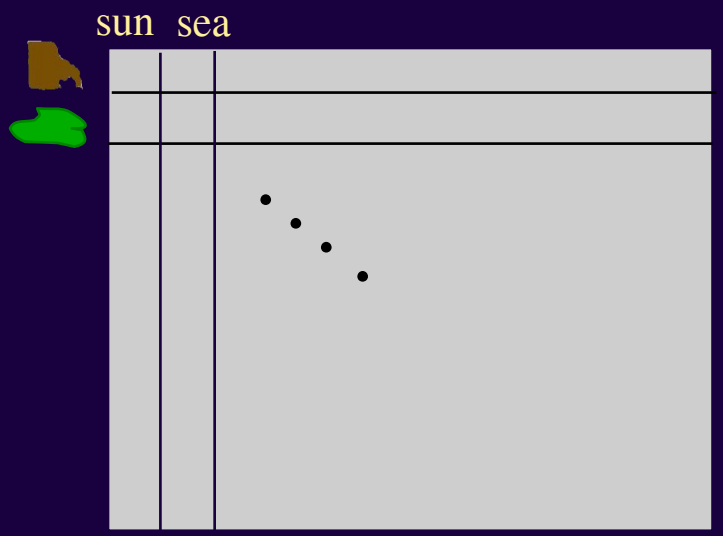
cat grass tiger water

Dictionary

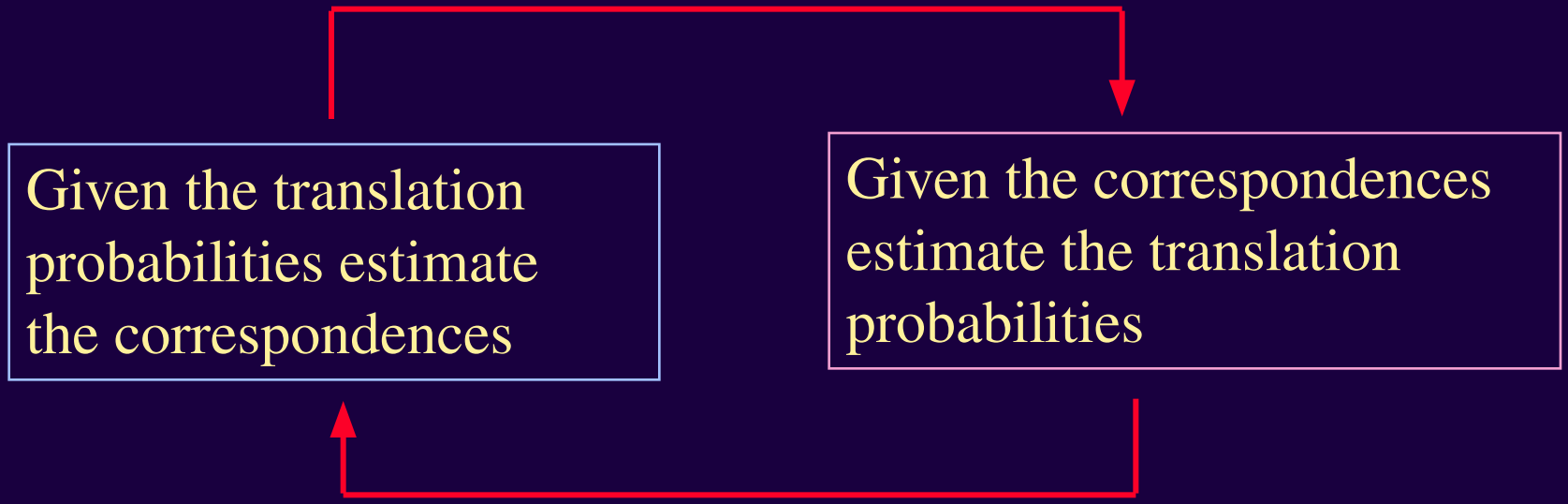
Blobs for three blob tokens	Most probable word
	sun
	sky
	cat

Initialization

Initialize translation table
to blob-word
co-occurrences
(empirical joint distribution
of blobs and words)



Expectation Maximization



Dempster et al., 77

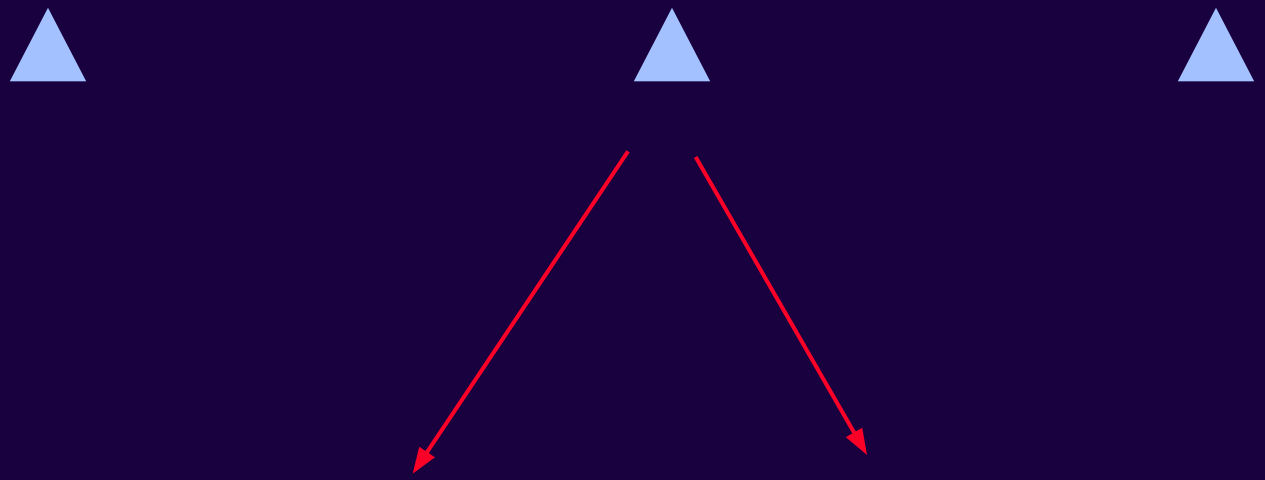
Why does this work?

Co-occurrence is a sensible starting point

EM process sharpens probabilities by integrating dictionary with constrained choices

More general models

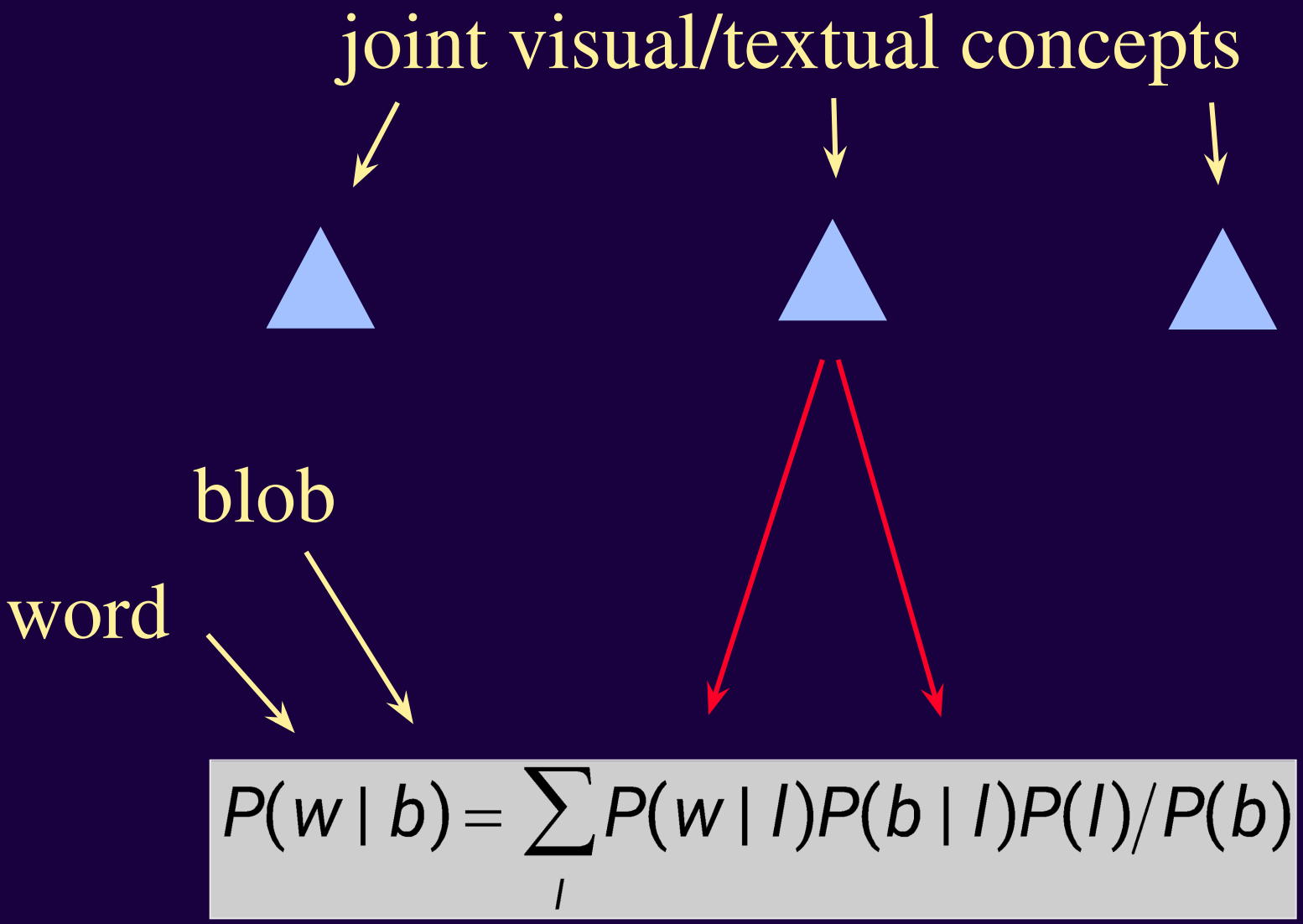
More general models



Generate words by frequency table

Generate blobs by Gaussian over features

(Conditionally independent given node)



$$P(w \mid b) = \sum_{l} P(w \mid l) P(b \mid l) P(l) / P(b)$$

Learn p(wll), p(bll), and p(l) from data using EM

More general models

sky

sea waves

sun

More general models

sky

sea waves

sun

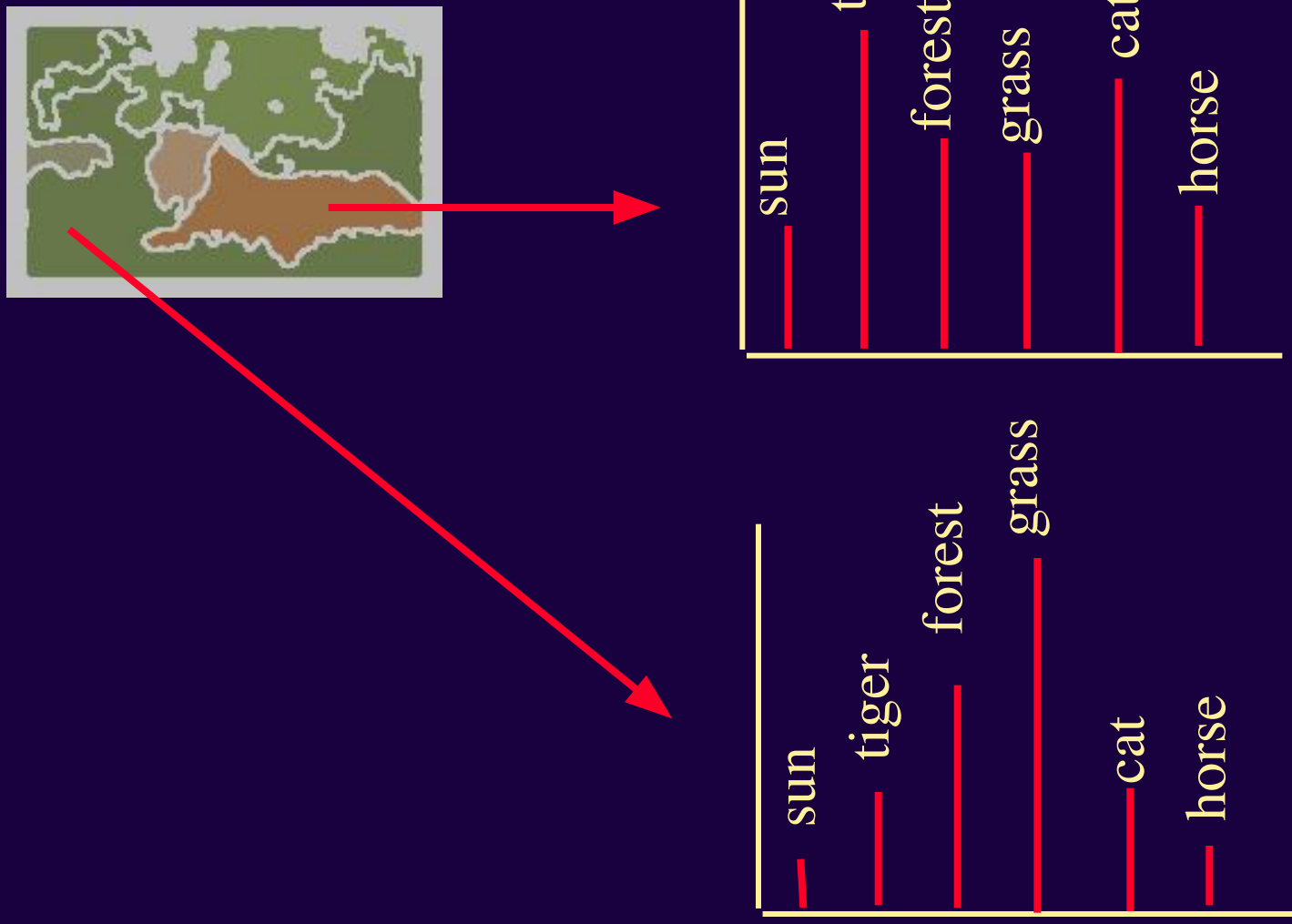
sky sea waves sun

Labeling Regions

Segment the image

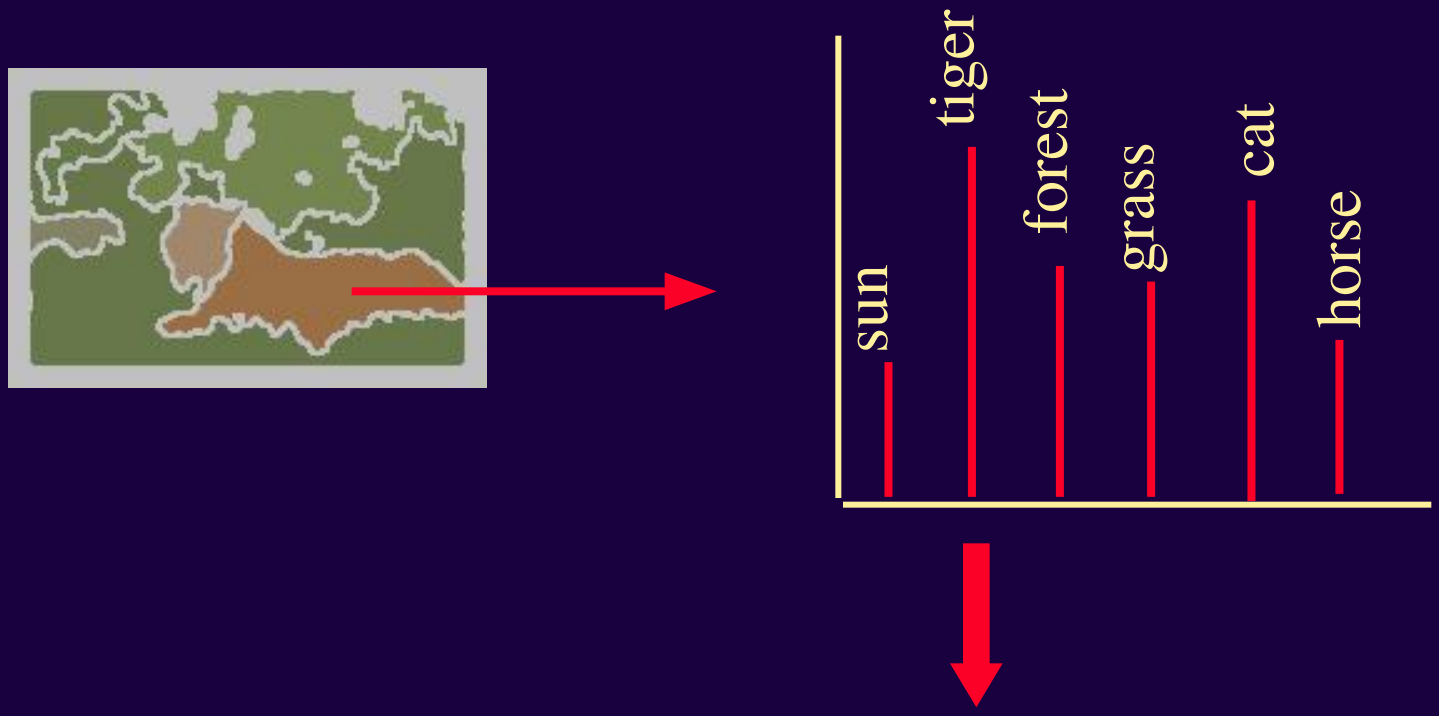
Use model to compute P(word | region)

Labeling Regions



Labeling Regions

Display only maximal probable word



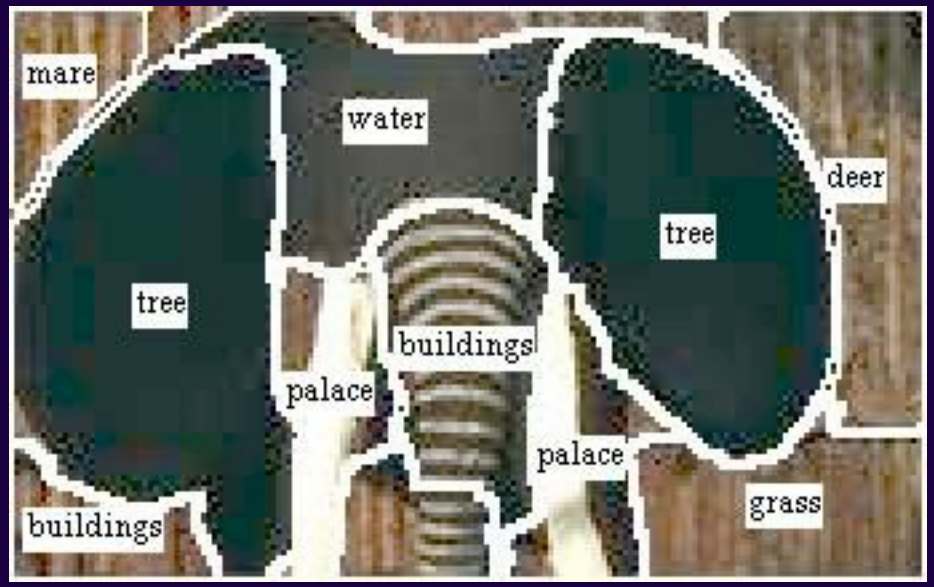
Results


grass
cat
horses
buildings
tiger
grass
mare
mare


aterfalls
sky
row
buildings
building
buildings
house
people
tree
tree


ground
flowers
ground
antlers
shops
town
crab
tiger
town
ground

Failure case



Measuring Performance

First strategy--score by hand

Second strategy--use annotation performance as a proxy.

First Strategy
Score by hand

Average performance is four times better than guessing the most common word

(“water”)

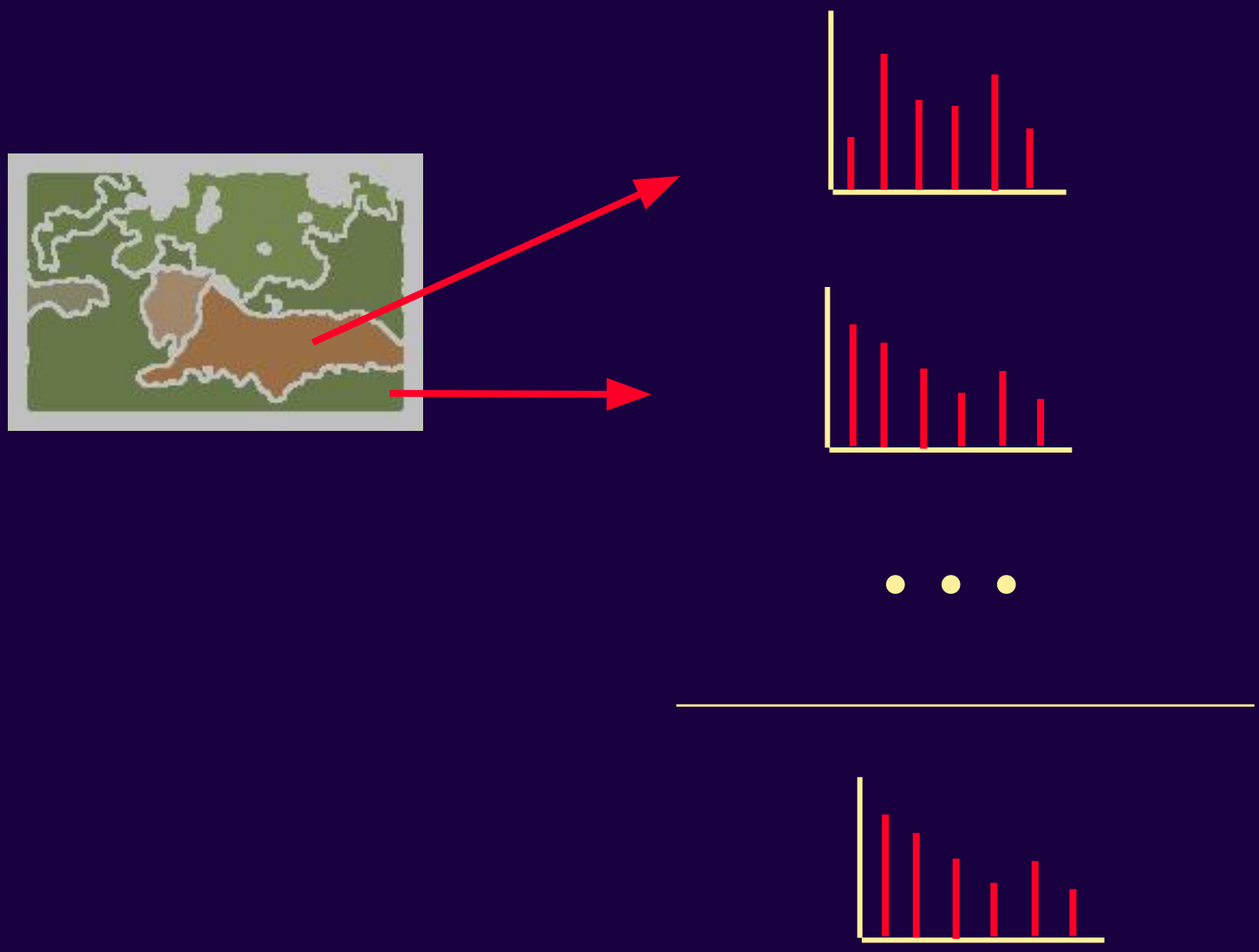
Not very scalable….

Second Strategy
Use Annotation

tiger cat grass water

Automatic : Don’t need to do by hand

Annotating Images



Measuring Annotation Performance

Actual Keywords

GRASS TIGER CAT FOREST

Predicted Words

CAT HORSE GRASS WATER

Measuring Annotation Performance

Actual Keywords

GRASS TIGER CAT FOREST

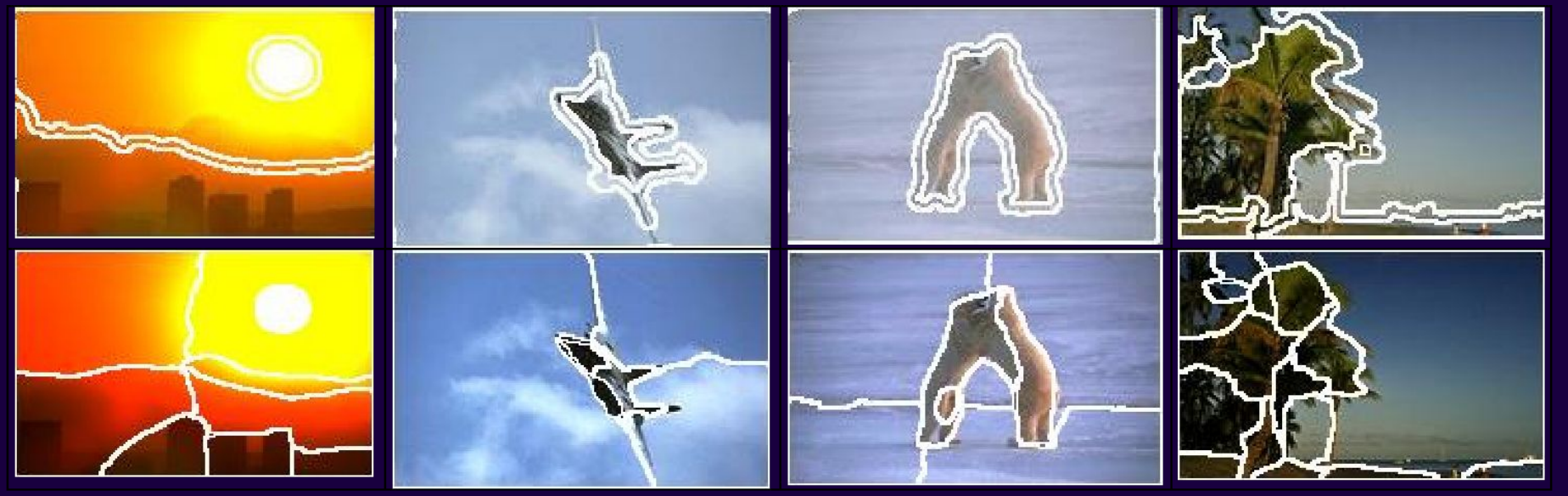
Predicted Words

CAT HORSE GRASS WATER

Exploiting Word Prediction

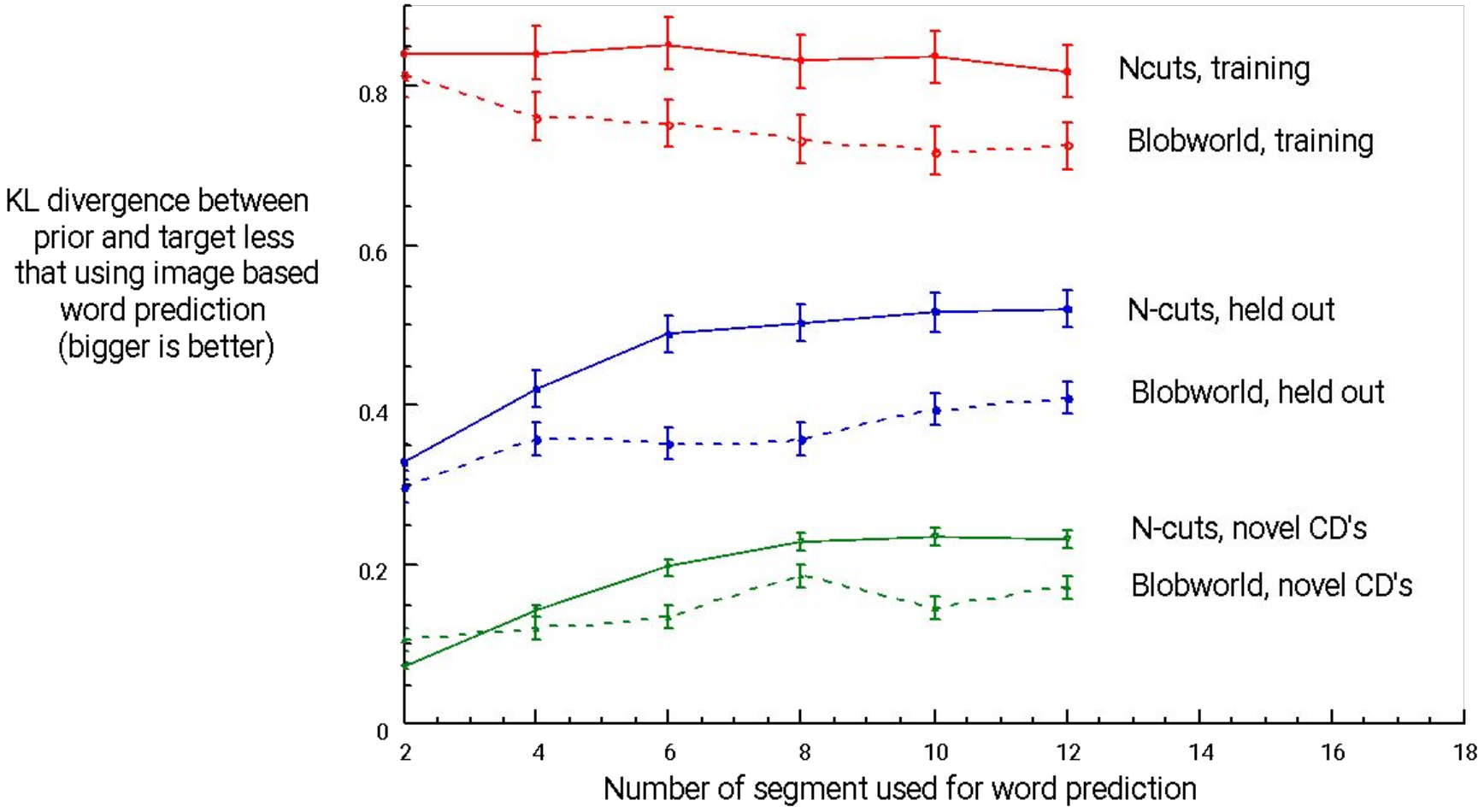
Model Selection
Segmentation
Feature choices

Blobworld segmentations



N-cuts segmentations

A comparison of two segmentation algorithms using word prediction performance



Some Applications

Using text search engines to find images without keywords (P(image|word))

Browsing and searching image parts

More Applications

Building a face gazetteer from news photos

Indexing and linking news stories using pictures

Linking news articles and news video using pictures

More Applications

Augmenting/combining information across the domains

Learning relationships across domains

Learning Relationships

Plausible tag learnt using U.S. news corpus

Jacques Chirac

Learning Relationships

Plausible tag learnt using U.S. news corpus

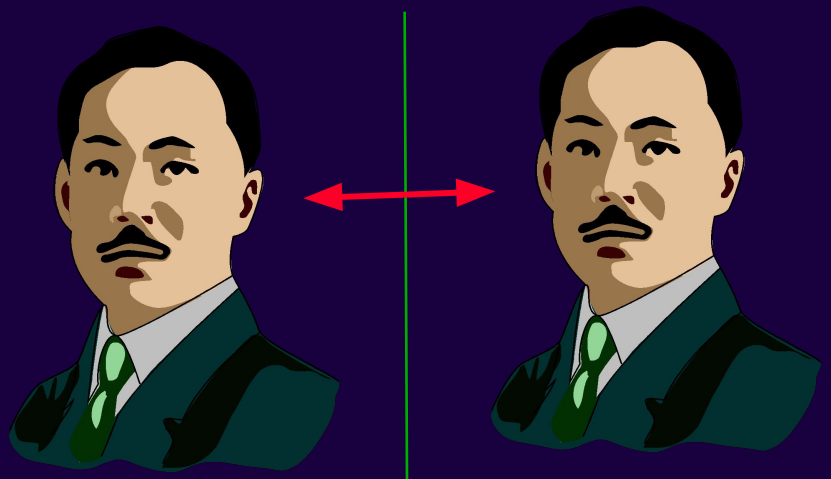
Jacques Chirac

Plausible tag learnt using French news corpus

Monsieur le president

Learning Relationships

Plausible tag learnt using U.S. news corpus



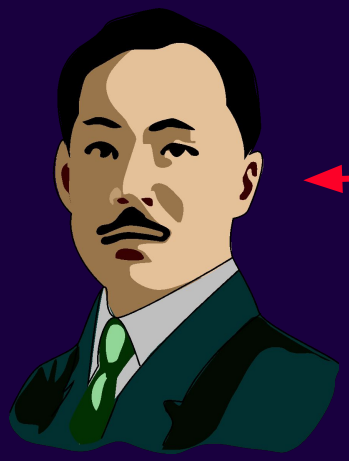
Plausible tag learnt using French news corpus

Jacques Chirac

Monsieur le president

Learning Relationships

Plausible tag learnt using U.S. news corpus



Plausible tag learnt using French news corpus

Jacques Chirac

Monsieur le president

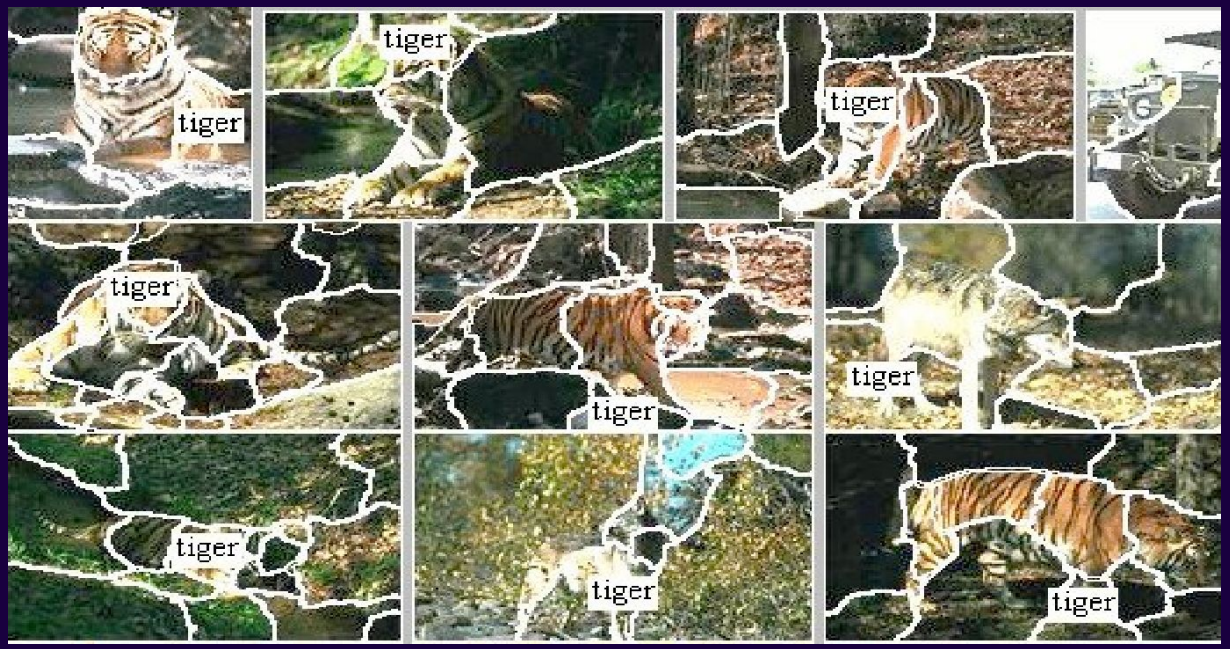
Useful ways to ‘cheat’

For some applications, good absolute performance may not be needed.

1) Image retrieval: Give the user multiple results

2) If words are available for test image, labeling improves

Held out data, no words allowed (pure vision task).



Held out data, no words allowed (useful for browsing/retrieval).

Useful ways to ‘cheat’

3) Word sense disambiguation with pictures [NAACL/HLT-WSM 2003]

The Word Sense Disambiguation Problem

212001 **bank** buildings trees city

125090 **bank** machine money currency bills

125084 piggy **bank** coins currency money

26078 water grass trees **bank**s

173044 mink rodent **bank** grass

151096 snow **bank**s hills winter

Conjecture

Images illustrating text can help

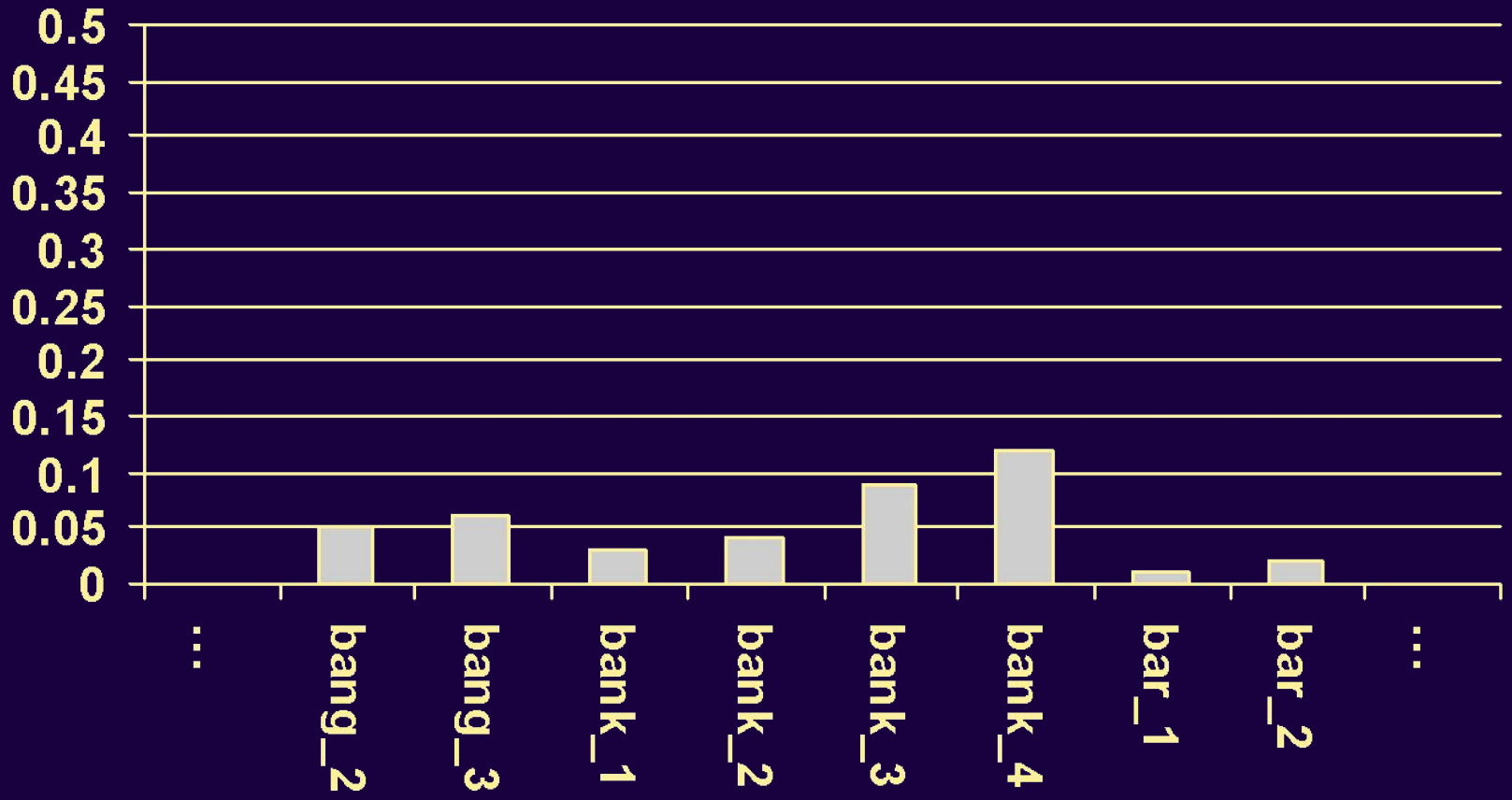
The man ate his lunch down by the **bank**.

Combining Sources

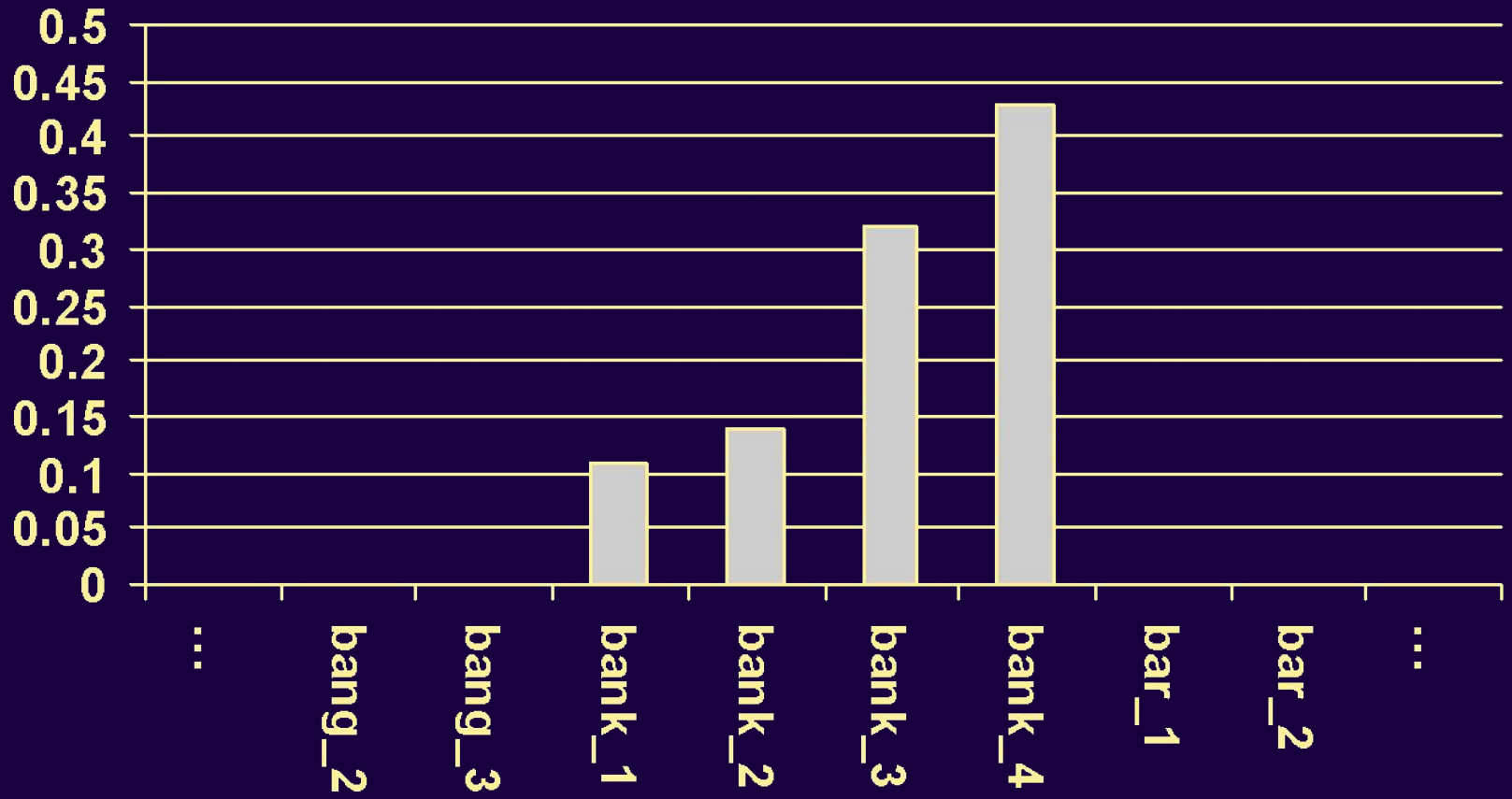
$$P(s \mid w, C, I) \propto P(s \mid w, I) P(s \mid w, C)$$

where $s \in S, w \in W$, S is the set of senses for the dictionary W, C is the textual context for w, and I is an image context for w.

P(s|I)


0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
…
bang_2
bang_3
bank_1
bank_2
bank_3
bank_4
bar_1
bar_2
…

P(s|w,I) (the ‘cheat’)



Preliminary Testing

Problems with Corel

- Not much ambiguity
- Keywords not disambiguated

Culled the database to a subset

- Images with one or more ambiguous keywords
- Hand-labeled the subset

More criticism on Corel in lecture on benchmarking….

Results

Word sense disambiguation strategy	Score
Naïve text based method	0.858 (0.008)
Random sense choice	0.875 (0.012)
Image and text method	0.948 (0.015)

However…

- Dataset (culled Corel) not very ambiguous
- WSD algorithm simplistic
  - Better algorithm might not need help from images

Comments on recognition vs annotations

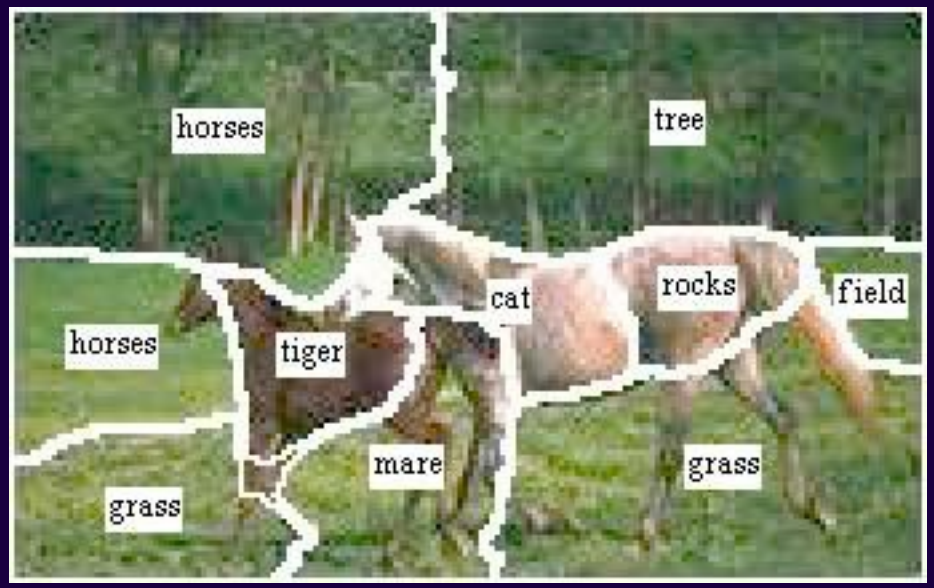
Learning on data without correspondence is a good trick BUT there are fundamental problems

Intuitively the words are generated through the parts (regions, groups), but the error function refers to the whole.

Need a better theory of how to link the two.

Integrating Supervision

Estimate where a minimal amount of supervision can be most helpful.



Integrating Feature Selection

Propose good features to differentiate words that are not distinguishable (e.g., eagle and jet)

Integrating Vision Levels

Word prediction gives a new way to think about integrating high and low level vision processes

Region Merging

Region Merging

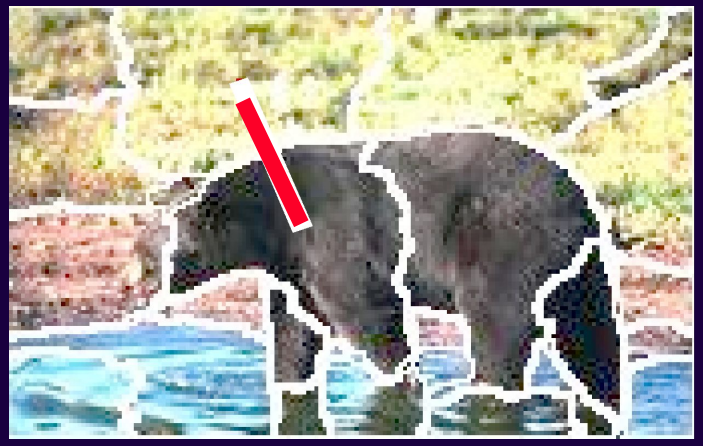
Use word posteriors to propose region merges

Recompute descriptors for the conglomerate object (color histograms, shape descriptors)

Have the system learn what kinds of “familiar configurations” are useful (i.e. lead to better word prediction)

Preliminary Experiment [CVPR, 03]

Good merge



Poor merge

More Complex Semantics

Current system links uniform blobs to simple nouns

Working towards linking groups of blobs to nouns, relations to prepositions, and attributes to adjectives

Summary

~~Recognition on the large scale~~

Unsupervised - label without labeled training data

Learn what to recognize

Semantic evaluation of vision tools

Integrating vision processing levels

Bottom Line

Recognition as machine translation

Machine vision as data-mining

Lecture IX: Words and Pictures

Part II: Learning social tag relevance

Xirong Li, Cees G.M. Snoek, and Marcel Worring
Intelligent Systems Lab Amsterdam
University of Amsterdam

Today's roadmap

- Matching words and pictures
- Names and faces in the news
- Learning social tag relevance

Learning social tag relevance

Xirong Li, **Cees G.M. Snoek**, and **Marcel Worring**
Intelligent Systems Lab Amsterdam
University of Amsterdam

Worldwide social-tagged visual data

- Emerging data
  - Billions of social-tagged images

flickr facebook … Picasa Zooomr

- Emerging research
  - Broad interest in multiple fields*
    - MM,CVPR,CIVR,MIR,ICME,WWW,KDD,…
  - CIVR'10: one keynote, two oral sessions

*Query google scholar with "social image retrieval"

Demo: What is on Flickr?

But, image retrieval results are poor

- Key problem: **Social tagging is subjective**
  - User tags might be irrelevant to the visual content

Query ***tiger***

But, image retrieval results are poor

- Key problem: **Social tagging is subjective**
  - User tags might be irrelevant to the visual content

Query ***classroom***

kindergarten classroom layout of ...

view details

stone age stone age ...

view details

365days me of me ...

view details

cfc minnesotawoods minnesota forest ...

view details

12 favorite classroom incubator ...

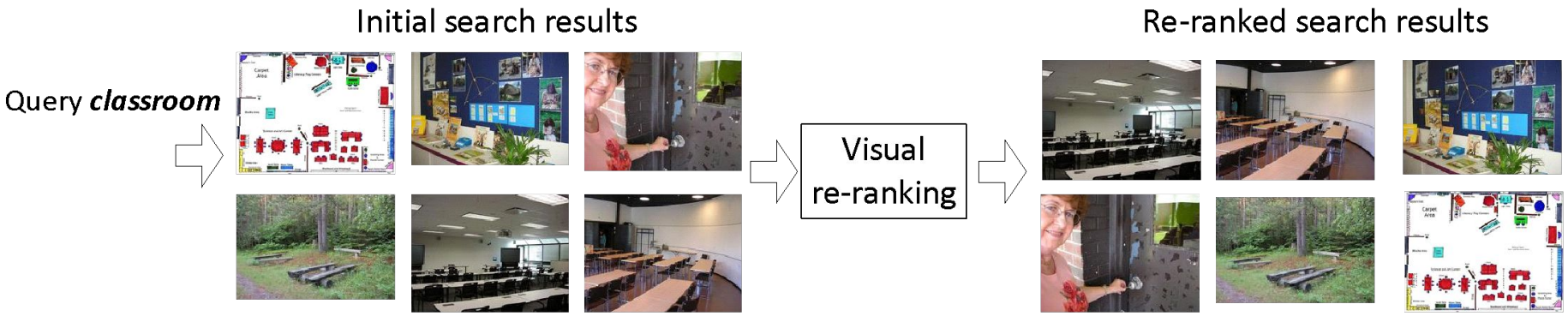
view details

tour tampere church upload ...

view details

Query-dependent methods

- Search results re-ranking [Hsu'07,…]



- Combining multiple textual features [Olivares'08,…]
  - tags, titles, notes,…

Query-independent methods

- Tag relevance learning [Our work Li'08,Li'09] and later [Liu'09,Wu'09,Kennedy'09,Xu'09,…]
  - Estimating the relevance of a tag with respect to the visual content

bridge ✓

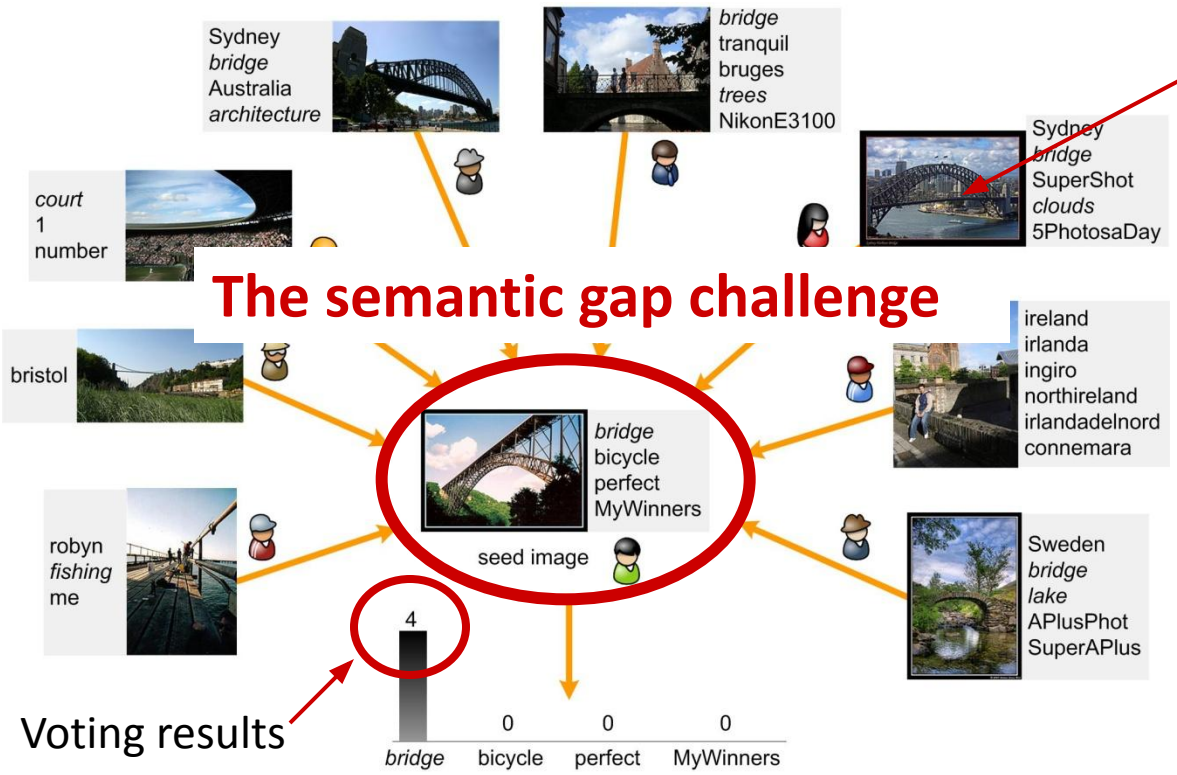
bicycle ✗

perfect

MyWinners

Learning tag relevance by neighbor voting [Li'08,Li'09]

- Exploiting the wisdom of multiple users in social image tagging



Need visual features to describe the content and to measure visual similarity.

The semantic gap: Mismatch between visual similarity and semantic similarity.

Intuition

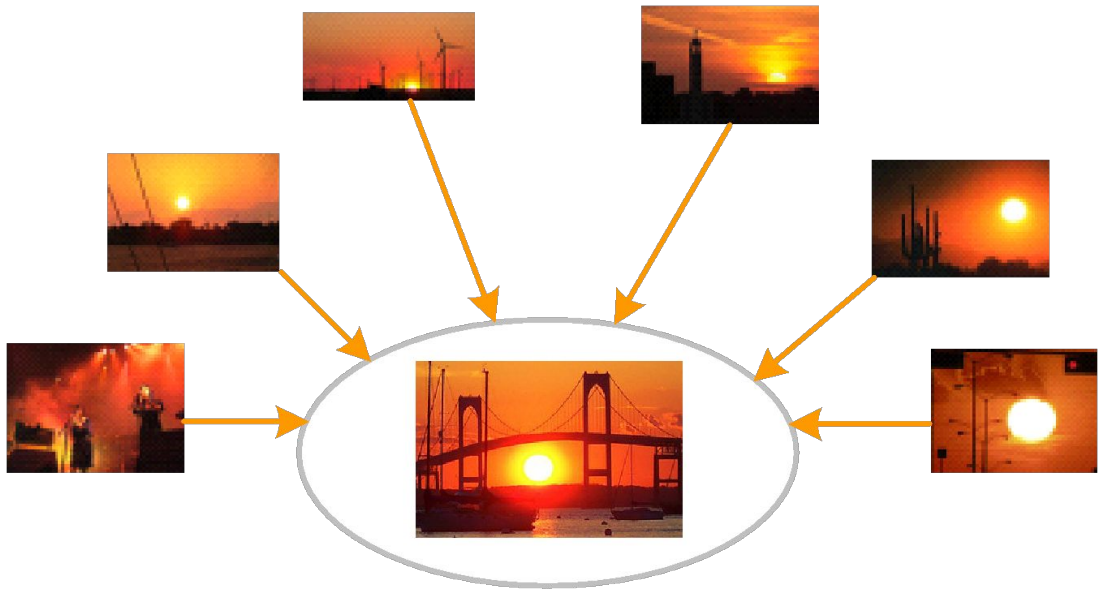
- The semantic gap is tag-dependent and feature-dependent[Lu'10]

sunset ← Objective tags
ocean
bridge ← Objective tags
nature
instantfave
scenery

Intuition cont.

- The semantic gap is tag-dependent and feature-dependent

  For color features, sunset has a smaller semantic gap than other tags



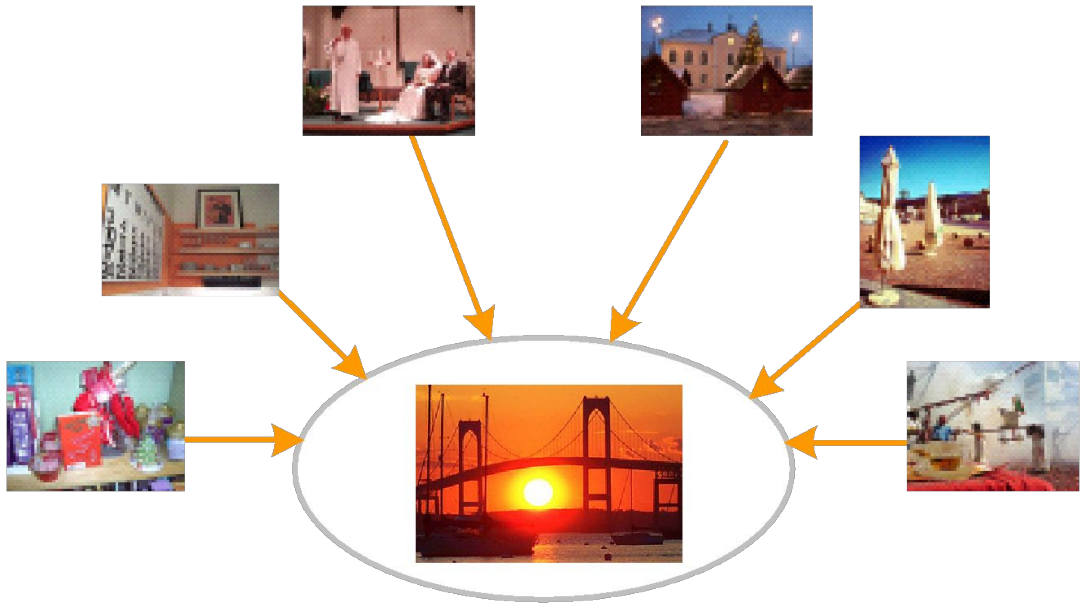
sunset

ocean
bridge
nature
instantfave
scenery

Intuition cont.

- The semantic gap is tag-dependent and feature-dependent

  The spatial structure feature does not work in this example

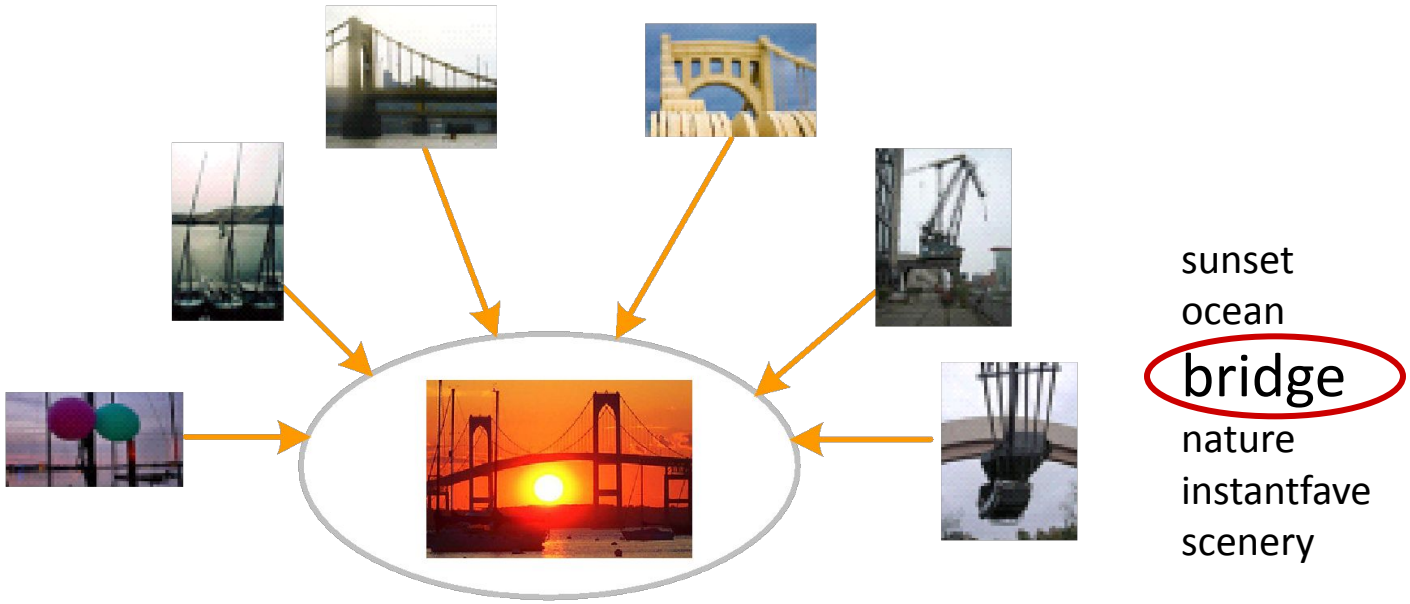


sunset
ocean
bridge
nature
instantfave
scenery

Intuition cont.

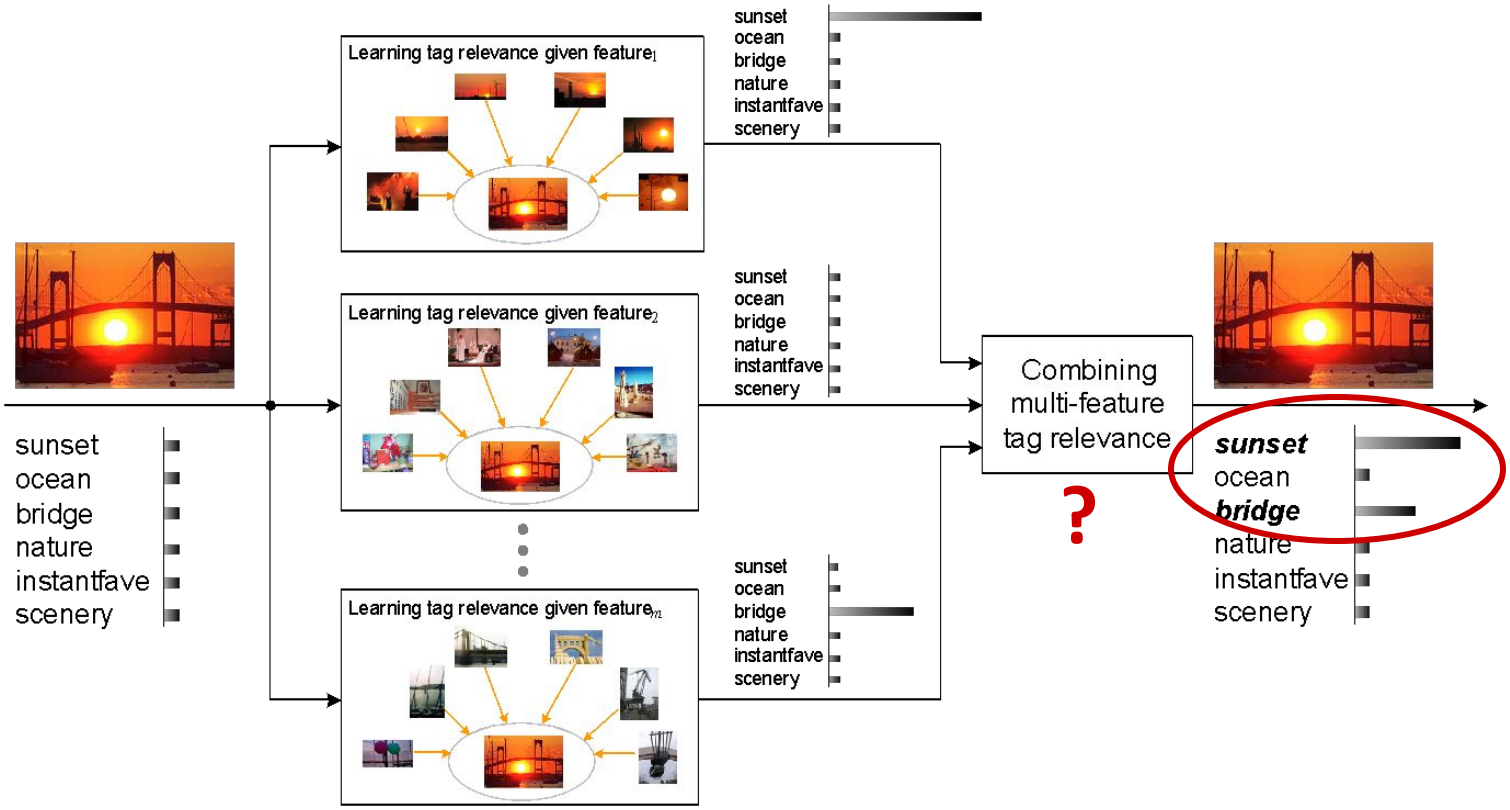
- The semantic gap is tag-dependent and feature-dependent

For bag-of-words features, bridge has a smaller semantic gap than other tags



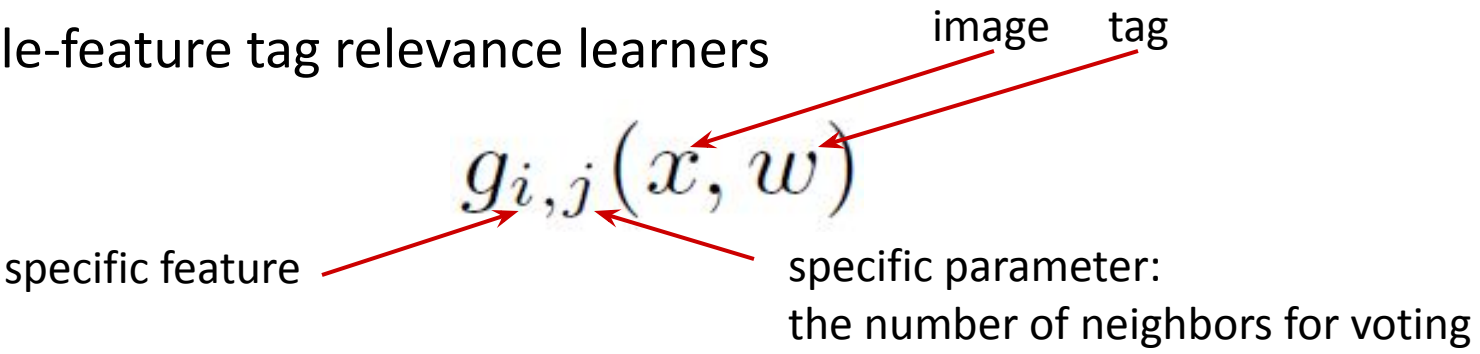
From intuition to algorithm

- Multi-feature tag relevance learning



General framework

- Single-feature tag relevance learners



- We seek a convex combination
  - A good choice for combining classification/ranking models [Freund'03,Hastie'01]

$$G(x, w) = \sum_{i=1}^{m} \sum_{j=1}^{n} \alpha_{i,j} \cdot g_{i,j}(x, w)$$

Combination weights

Unsupervised combination methods

- Uniform weights
  - Following the principle of maximum entropy[Jaynes'03]
  - We term the method ***UniformTagger***

$$\frac{1}{m \cdot n} \sum_{i=1}^{m} \sum_{j=1}^{n} g_{i,j}$$

- Borda Count
  - A well-known rank aggregation algorithm[JC de Borda, 1770]

$$\frac{1}{m \cdot n} \sum_{i=1}^{m} \sum_{j=1}^{n} (N_w - rank(g_{i,j}))$$

Supervised combination methods

- Supervised
  - Best Single Learner[Li'09]
  - Weighted Borda Count[Aslam'01]
  - RankBoost[Freund'03]

- The objective function to maximize

$$r(G) = \sum_{x_+} \sum_{x_-} \lambda_{x_+, x_-} \cdot (G(x_+, w) - G(x_-, w))$$

pairwise weight

ranking gain
the larger the better

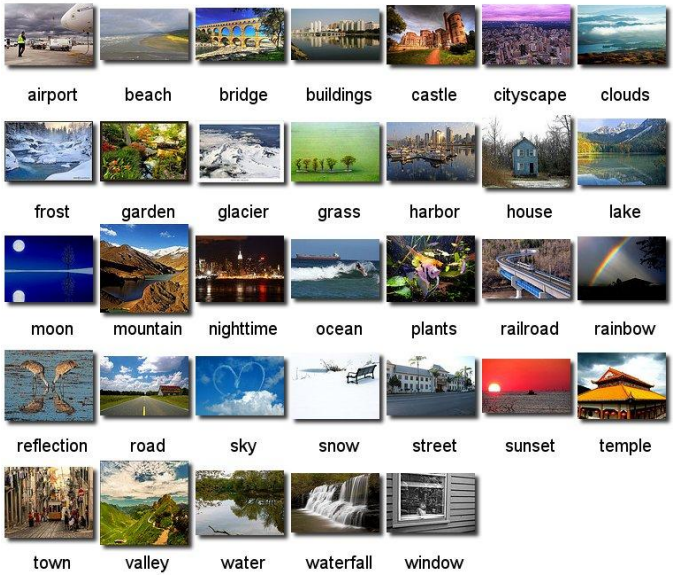
Questions to answer

- Is **multi-feature** better than single-feature for learning tag relevance?
- Compared to the supervised combination methods, how much do we lose when using the **unsupervised** combination methods?

Experimental setup

- Two ground truth query sets for evaluation

NUS-SCENE[1]: 33 concepts, ~34,000 images



Social20[2]: 20 concepts, ~20,000 images

- Neighbor voting set
  - **3.5 million** social-tagged images randomly collected from Flickr[2]

[1]http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm

[2]http://staff.science.uva.nl/~xirong/tagrel/

Experimental setup cont.

- Visual features

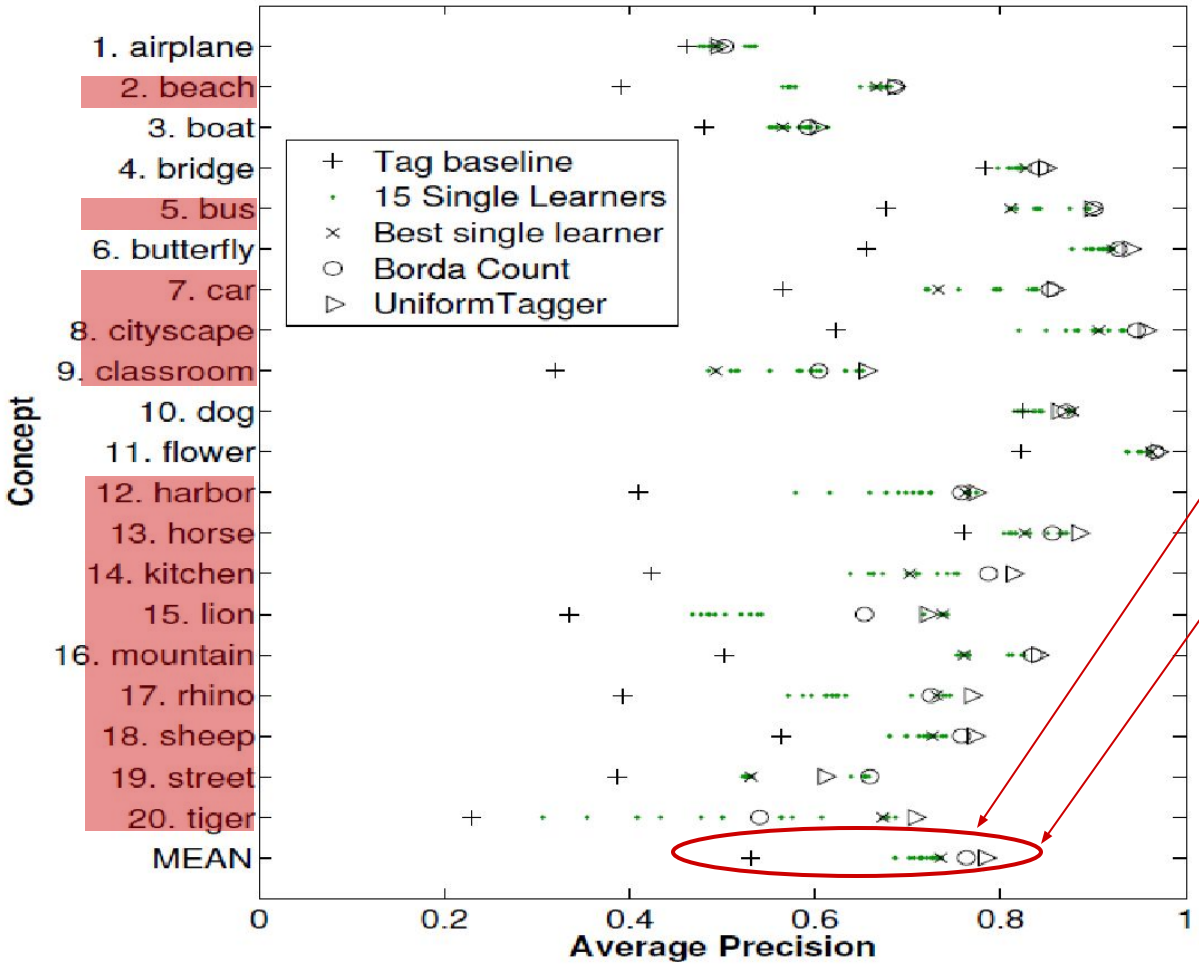
Feature	Granularity	Dim	Descriptions
Color64[Huang'99,Yu'02]	global	64	color correlogram + texture moment + color moment
GIST[Oliva'01]	global	960	dominant spatial structure of a scene
Dense-SURF[Bay'08,Uijlings'09]	local	4000	bag-of-keypoints: dense sampling + SURF descriptor

- Parameters of base tag relevance learners
  - Number of neighbors: {500,1000,1500,2000,2500} □ 3x5=15 base learners
- Approximate visual neighbor search on large data sets
  - K-means based indexing
- Evaluation criteria
  - average precision, mean average precision

Experiment 1: Unsupervised image retrieval

- Tag Baseline
  - Given a concept, rank images according to the concept's occurrence frequency in descending order.
- Single-feature tag relevance baseline
  - Update tag frequency with learned tag relevance values
- Two unsupervised combination methods
  - UniformTagger
  - Borda Count

Results of experiment 1 (on Social20)



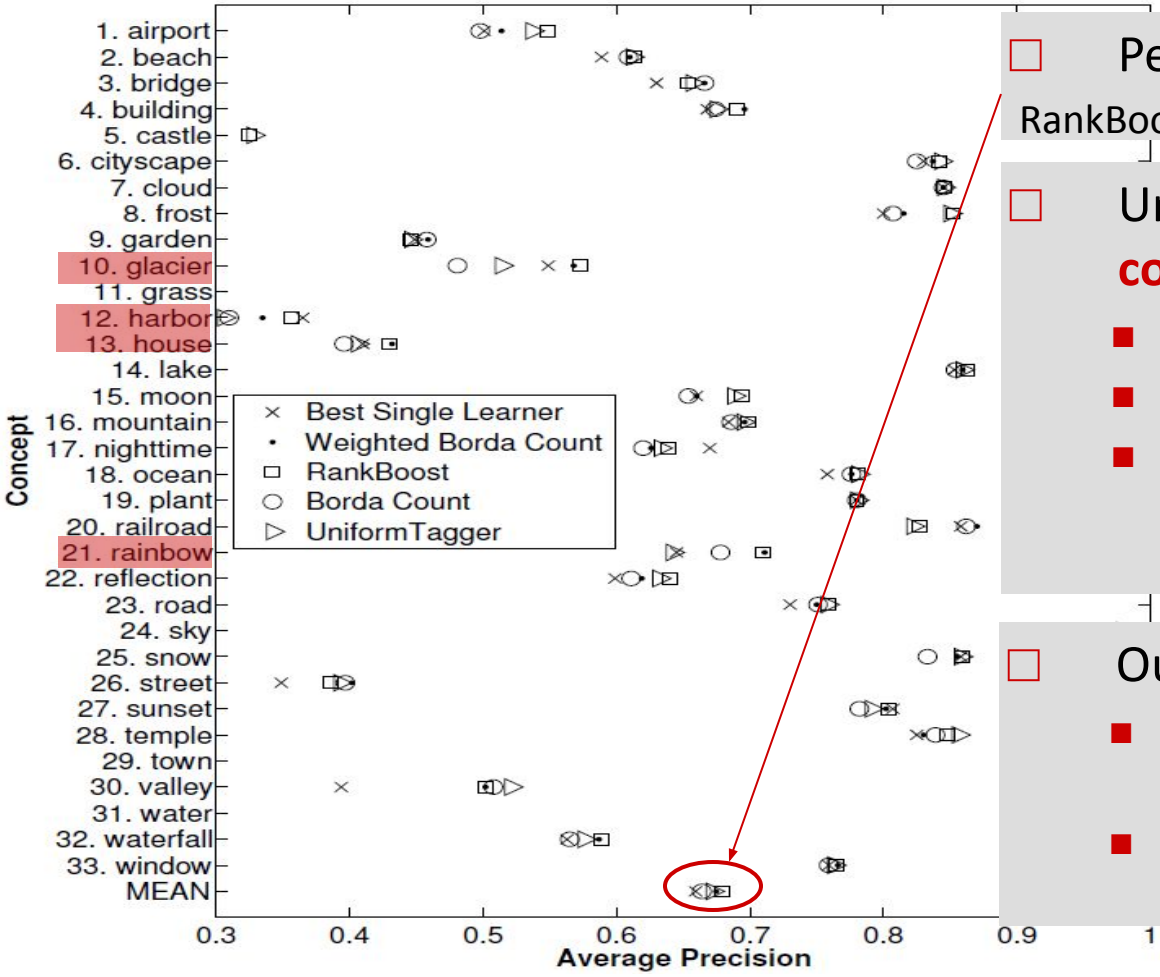
- UniformTagger beats tag baseline and single-feature tag relevance learning.
  - 47.6% relative improvement over tag baseline.
  - 8.6% relative improvement over single-feature.

  14 out of 20 concepts have more than 5% relative improvement.

Experiment 2: Supervised image retrieval

- For each concept, we train the combination weights $\{\alpha_{i,j}\}$ using
  - Best Single Learner
  - Weighted Borda Count
  - RankBoost

Results of experiment 2 (on NUS-SCENE)



- Performance comparison

RankBoost > Weighted BordaCount > Best Single

- UniformTagger is **surprisingly comparable** to RankBoost.
  - RankBoost MAP: 0.679
  - UniformTagger MAP: 0.672
  - For only 4 out of 33 concepts, RankBoost outperforms UniformTagger, with over 5% relative improvement.
- Our explanation
  - RankBoost might over-emphasize over-confident base learners.
  - Base learners complement each other and perform reasonably well in general.

Image search after tag relevance learning

- Objective social tagging Accurate image retrieval results

Query ***tiger***

tiger 24
zoo 24
animal 14
chi a 4
...

view details

zoo 46
animals 29
tiger 22
siber ntiger 2
...

view details

zoo 37
animals 28
tiger 22
lion 16
...

view details

nature 41
zoo 25
cat 23
tiger 21
...

view details

zoo 37
tiger 20
cat 11
cats 4
...

view details

tiger 18
cat 9
nationalzoo 4
cats 3
...

view details

Image search after tag relevance learning

- Objective social tagging Accurate image retrieval results

Query ***classroom***

classroom 24

view details

classroom 23

view details

classroom 22
school 12
2007 5
3 1
...

view details

classroom 20
school 8
china 4
2004 3
...

view details

classroom 19
school 13
japan 9
japanese 3
...

view details

classroom 18
japan 14

view details

Failure case

- Many users label these images with the tag “airplane”….

Query ***airplane***

clouds 130
airplane 93
plane 50
mountains 36
...

view details

sky 134
airplane 76

view details

airplane 71
flight 20
japan 3

view details

airplane 71

view details

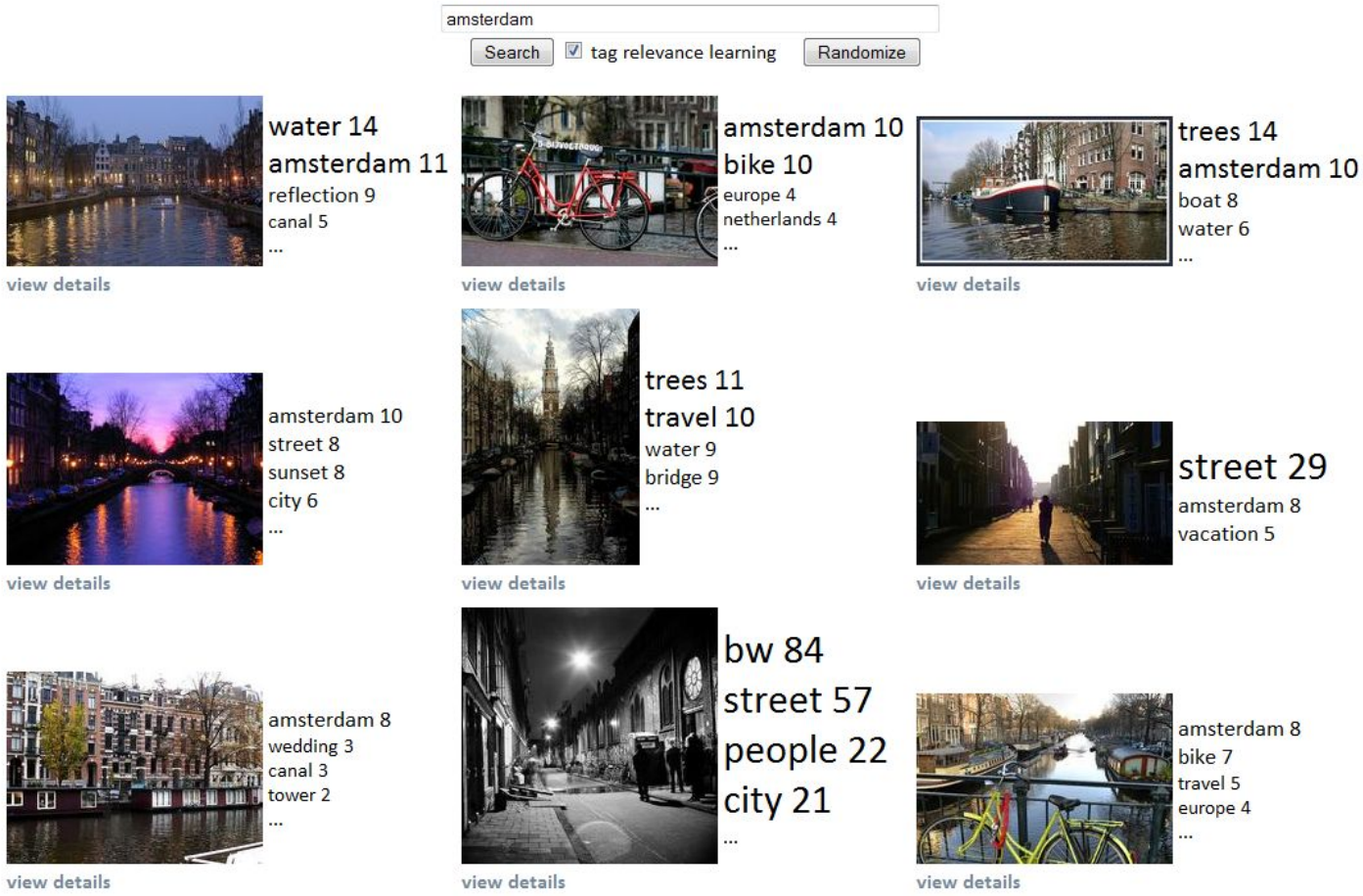
sky 138
clouds 135
airplane 70
flying 42
...

view details

sky 101
clouds 82
airplane 65
blue 42
...

view details

Demo: multi-feature tag relevance



Quiz: what is the downside of multi-feature tag-relevance learning

- While the accuracy is significantly improved, the visual diversity in the results is lost to some extent.

Conclusions

- We study **multi-feature** tag relevance learning to conquer the subjective social tagging problem for social image retrieval.
- Multi-feature outperforms single-feature for tag relevance learning.
- We propose **UniformTagger** to combine multi-feature tag relevance
  - Solid theoretical foundation
  - Simple to implement
  - NO labeling required at all
  - Competitive performance to supervised combination methods