Computer Vision as Multimedia Translation and Data Mining

Kobus Barnard

University of Arizona

Visual Representation → Semantic Representation

→ A tiger lying in the grass

Visual Representation ⇒ Semantic Representation

grass

tiger

Auto-Annotating Images

Finding words for the images

tiger grass cat

Barnard, Forsyth (ICCV 2001) , Barnard, Duygulu, Forsyth (CVPR 2001)

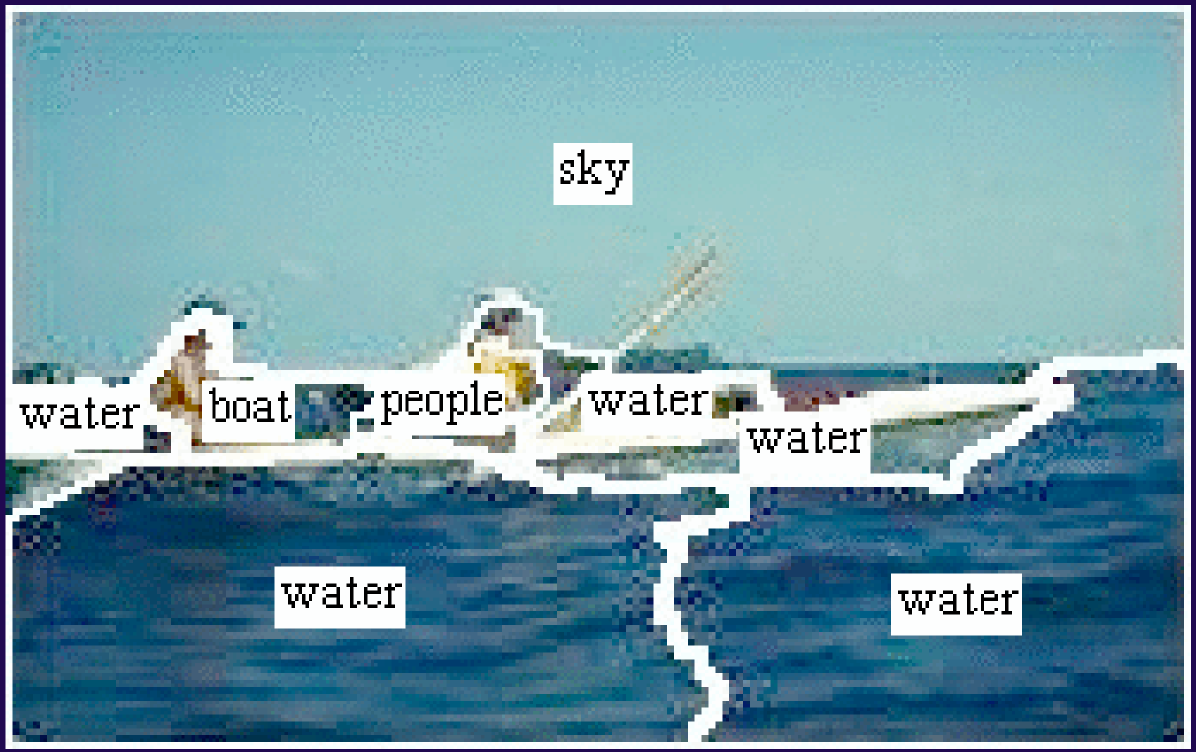
Other related work : Maron 98, Mori 99

Annotation 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

[ICCV 2000, ECCV, 2002, JMLR 2003]

Key Point

Learn from data without explicit correspondence between image components

Data with correspondence ambiguity is common

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

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

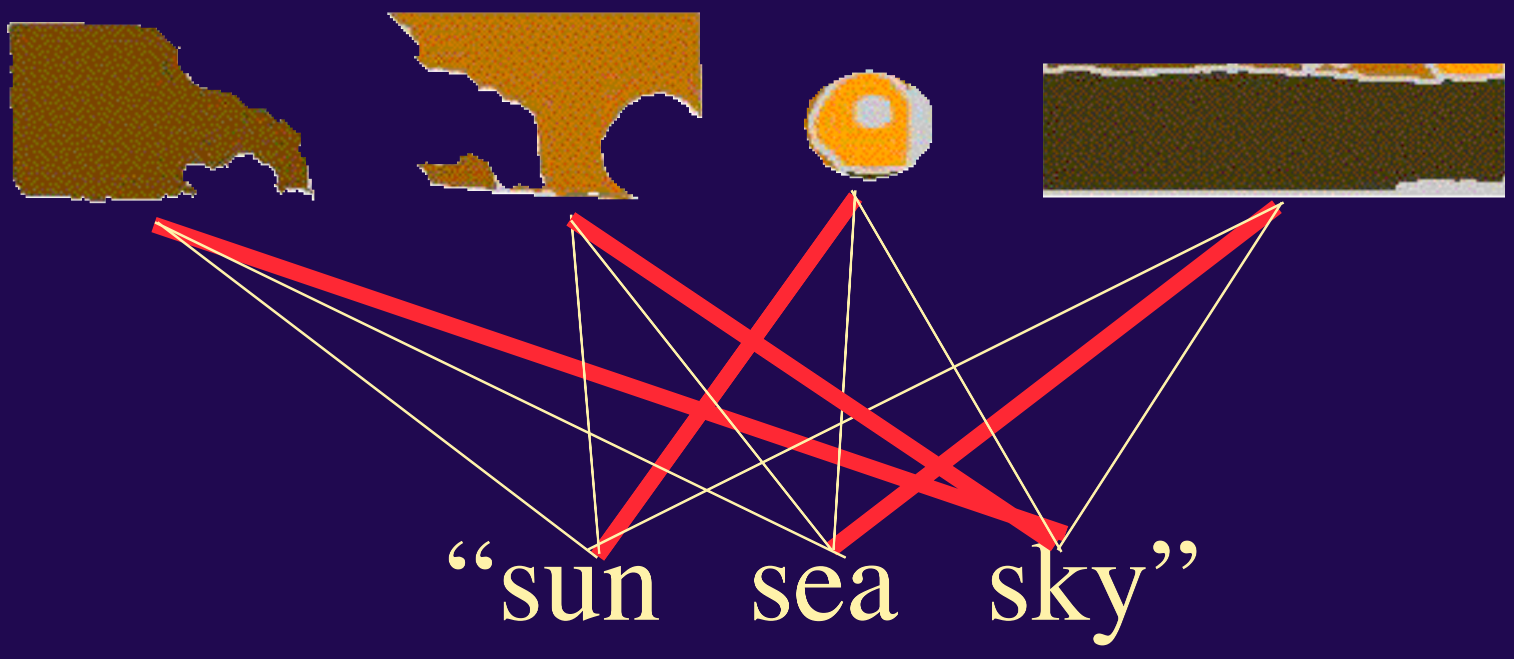
Data: Aligned sentences, but word correspondences are unknown

"the beautiful sun"

"le soleil beau"

Brown, Della Pietra, Della Pietra & Mercer 93

Multimedia Translation



Approaches

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

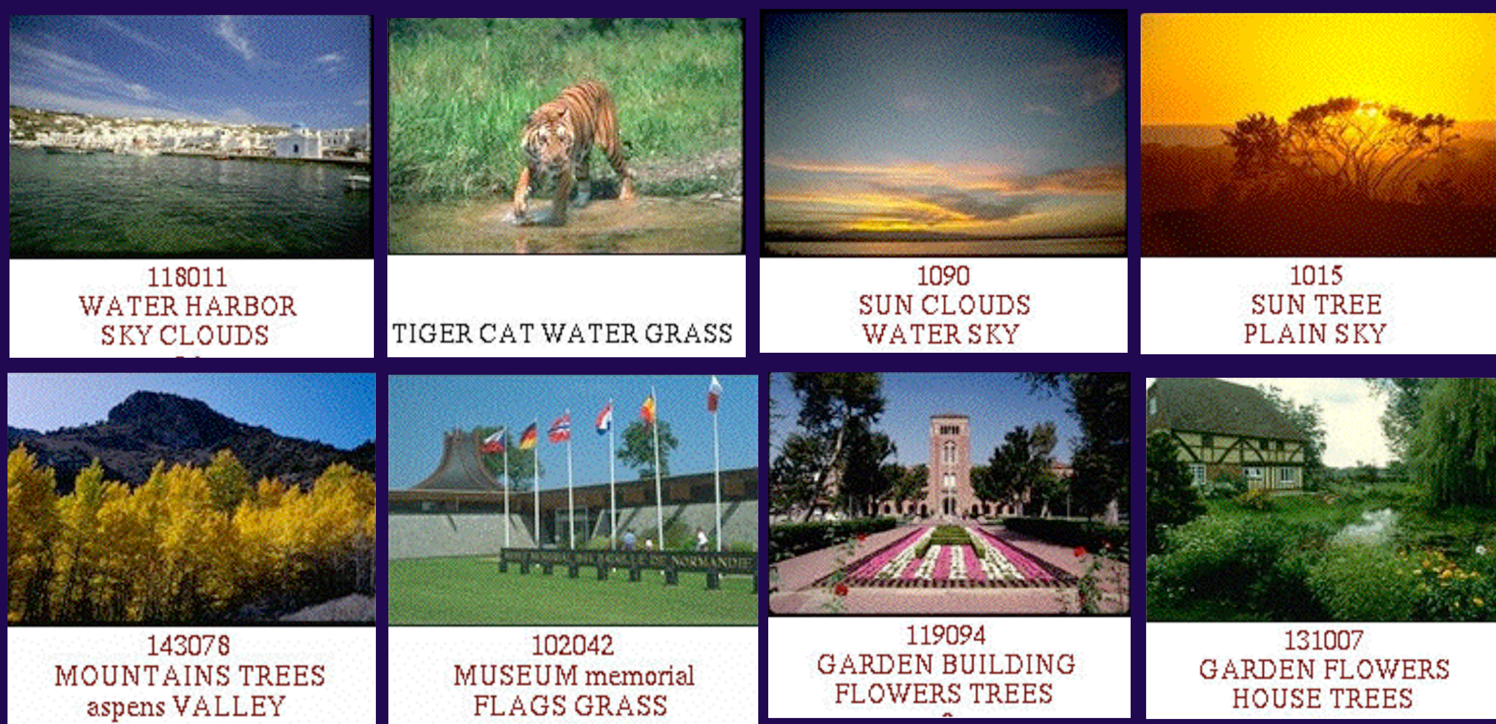
Simultaneously learn blob models and translation [Barnard et al, JMLR 03]

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



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

Input

sun sky waves sea

Image processing*

Each region is described by a set of features

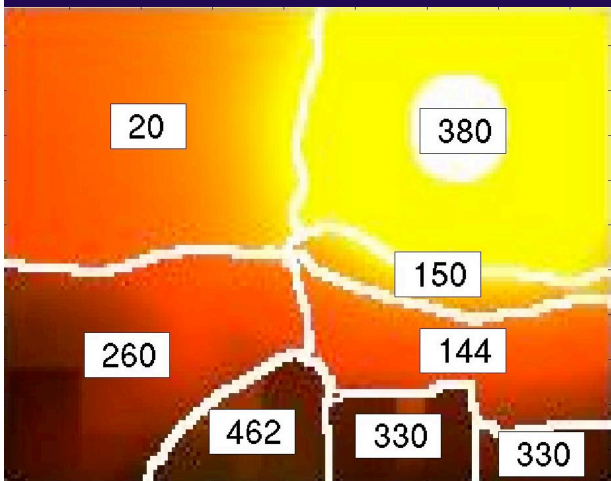
- Region size
- Position
- Color
- Oriented energy (12 filters)
- Simple shape features

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

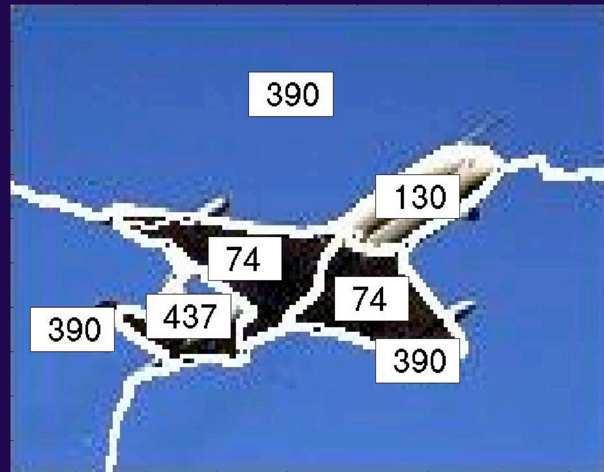
Discrete Model [ECCV 02]

Straightforward adaptation of machine translation

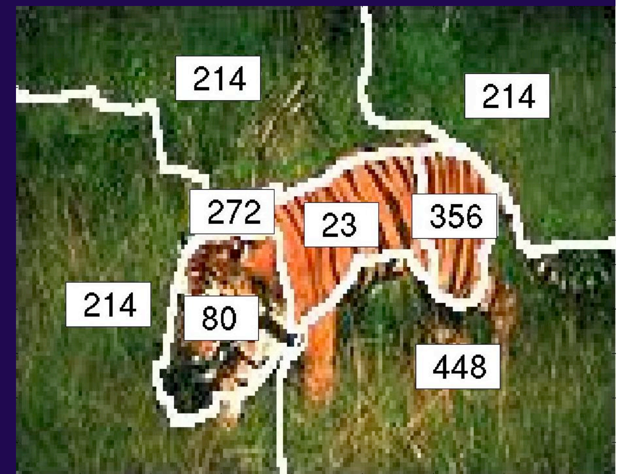
Need to vector quantize blobs (simple but better to simultaneously learn blob model)



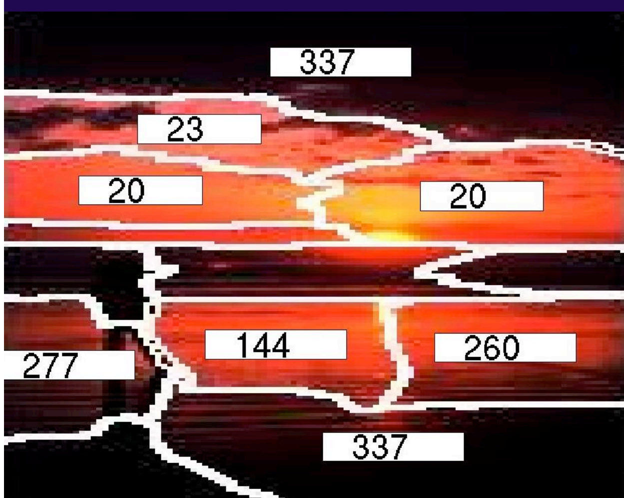
city mountain sky sun



jet plane sky

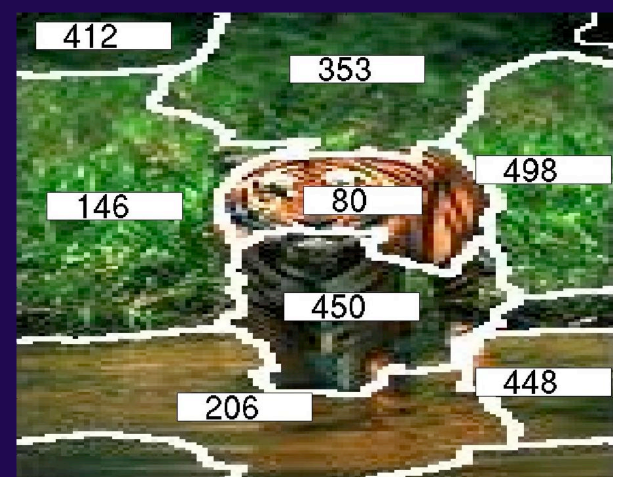


cat forest grass tiger



beach people sun water

jet plane sky



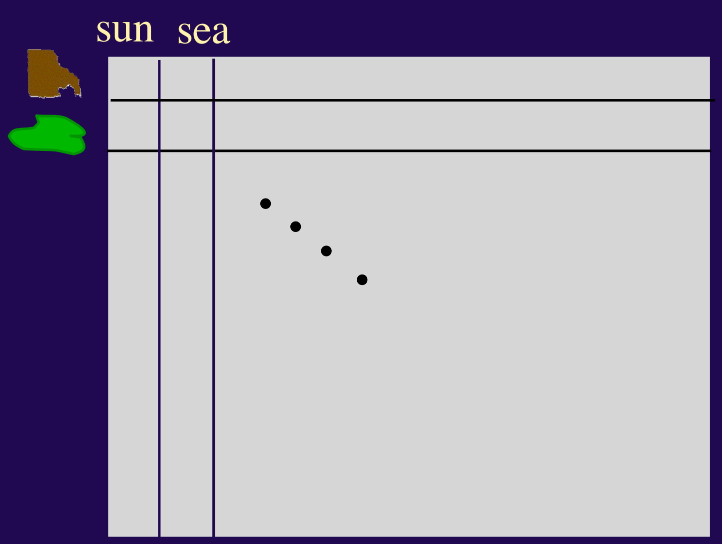
cat grass tiger water

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

Given the translation probabilities estimate the correspondences

Given the correspondences estimate the translation probabilities

Dempster et al., 77

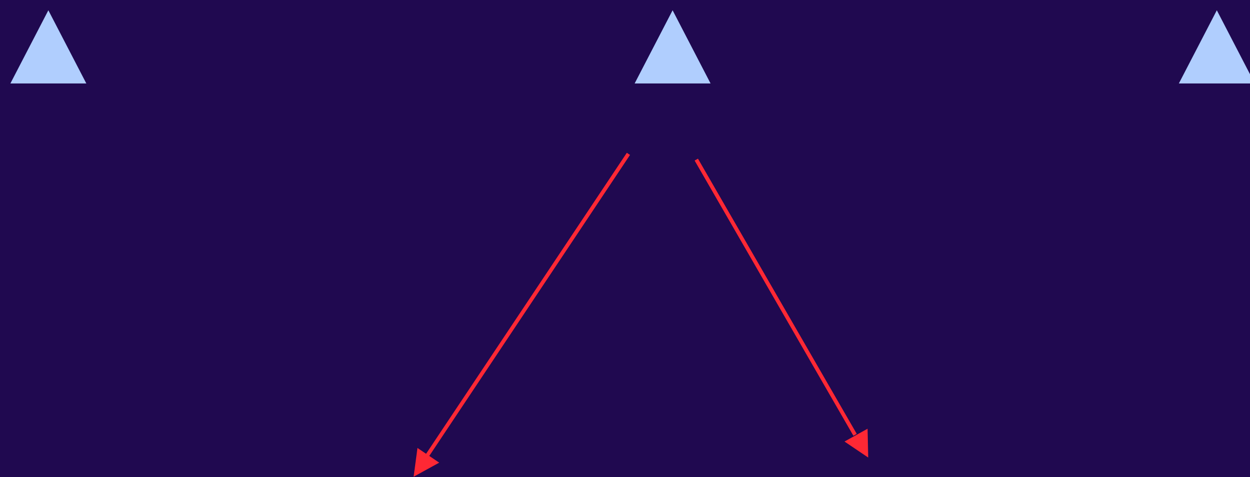
Why does this work?

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

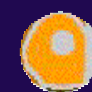
Generate blobs by Gaussian over features

(Conditionally independent given node)

More general models

sky

sea waves

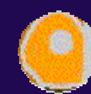


sun

More general models

sky

sea waves



sun

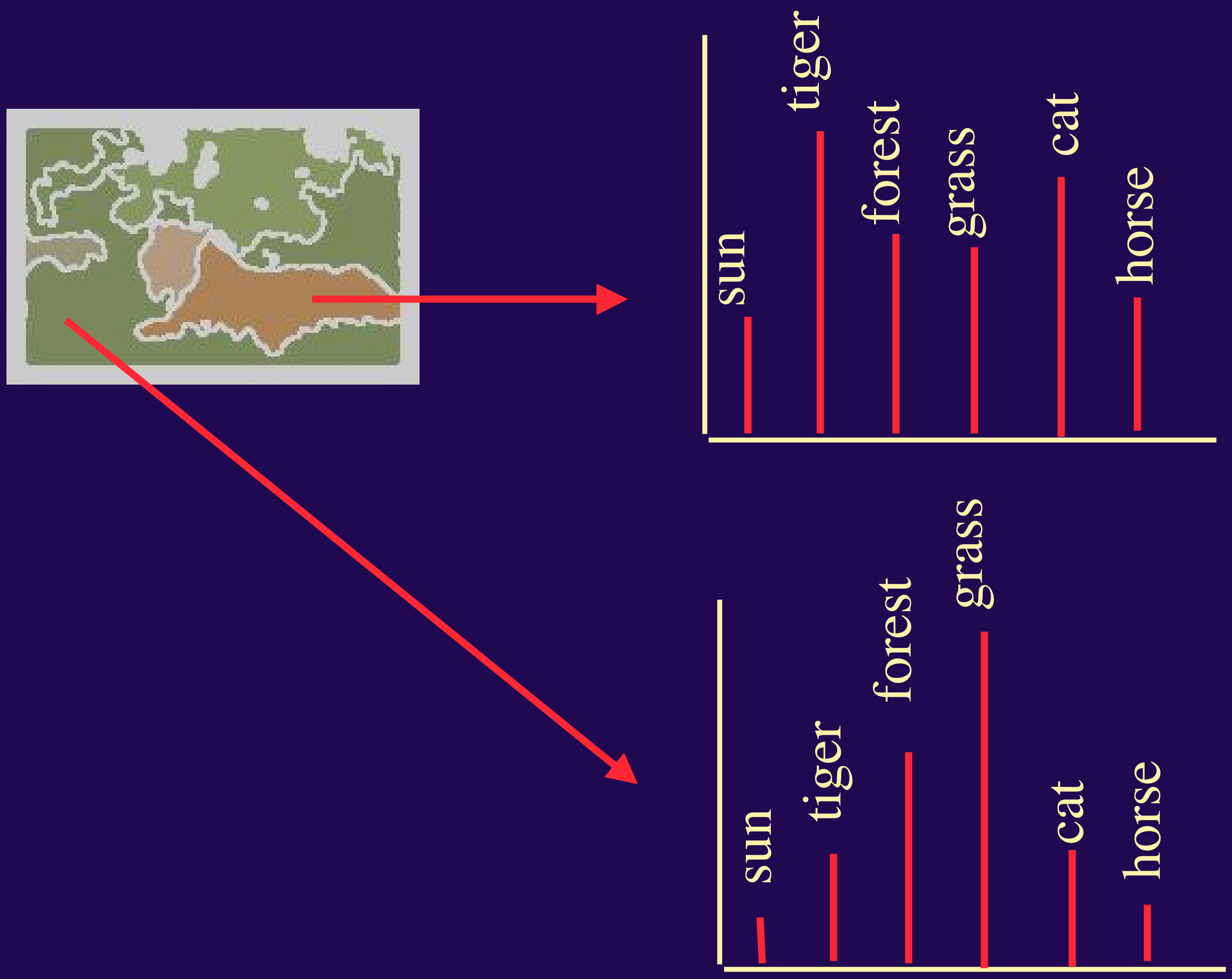
sky sea waves sun

Labeling Regions

Segment the image

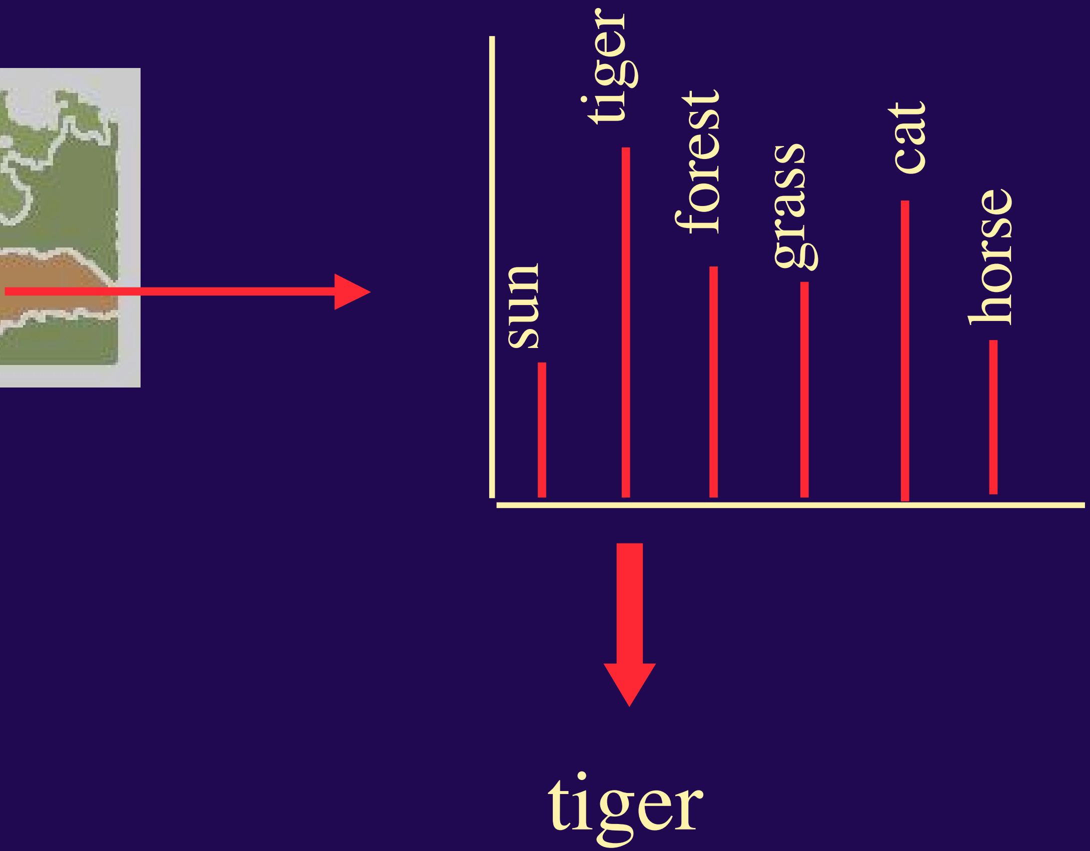
Use model to compute P(word | region)

Labeling Regions

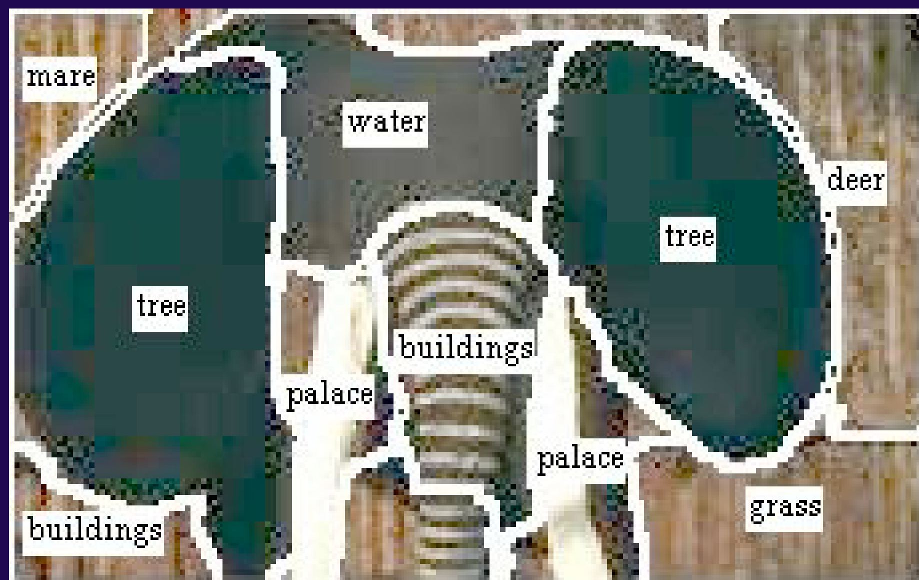


Labeling Regions

Display only maximal probable word



grass
horses
cat
buildings
tiger
grass
mare
mare
aterfalls
sky
row
building
buildings
buildings
house
people
tree
tree
ground
flowers
ground
antlers
shops
town
crab
tiger
town
ground

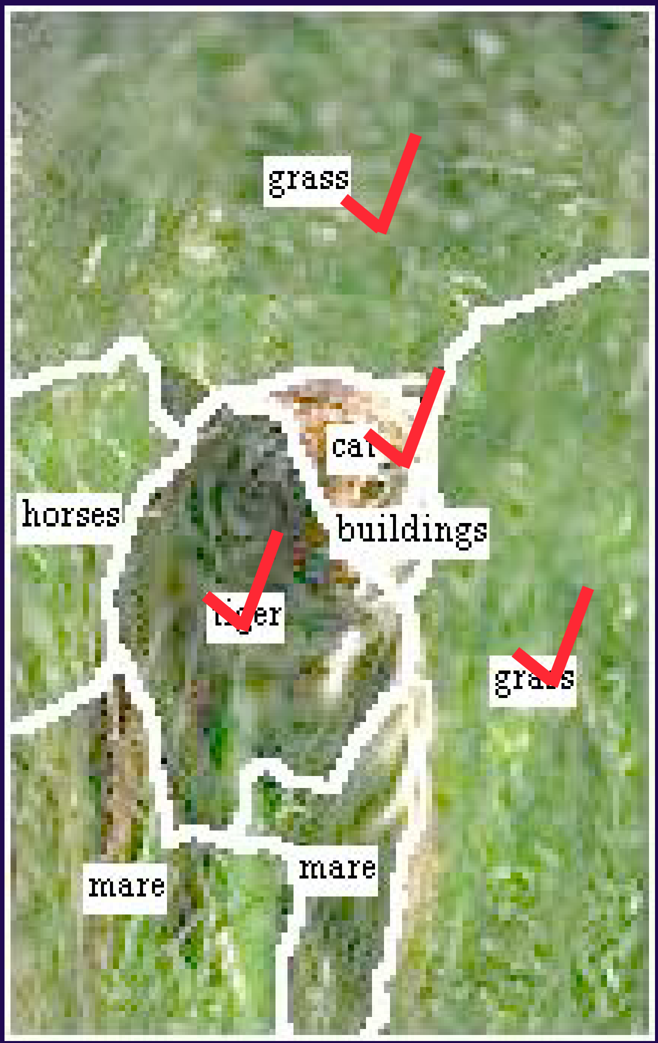

mare
water
deer
tree
tree
buildings
palace
palace
grass
buildings

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

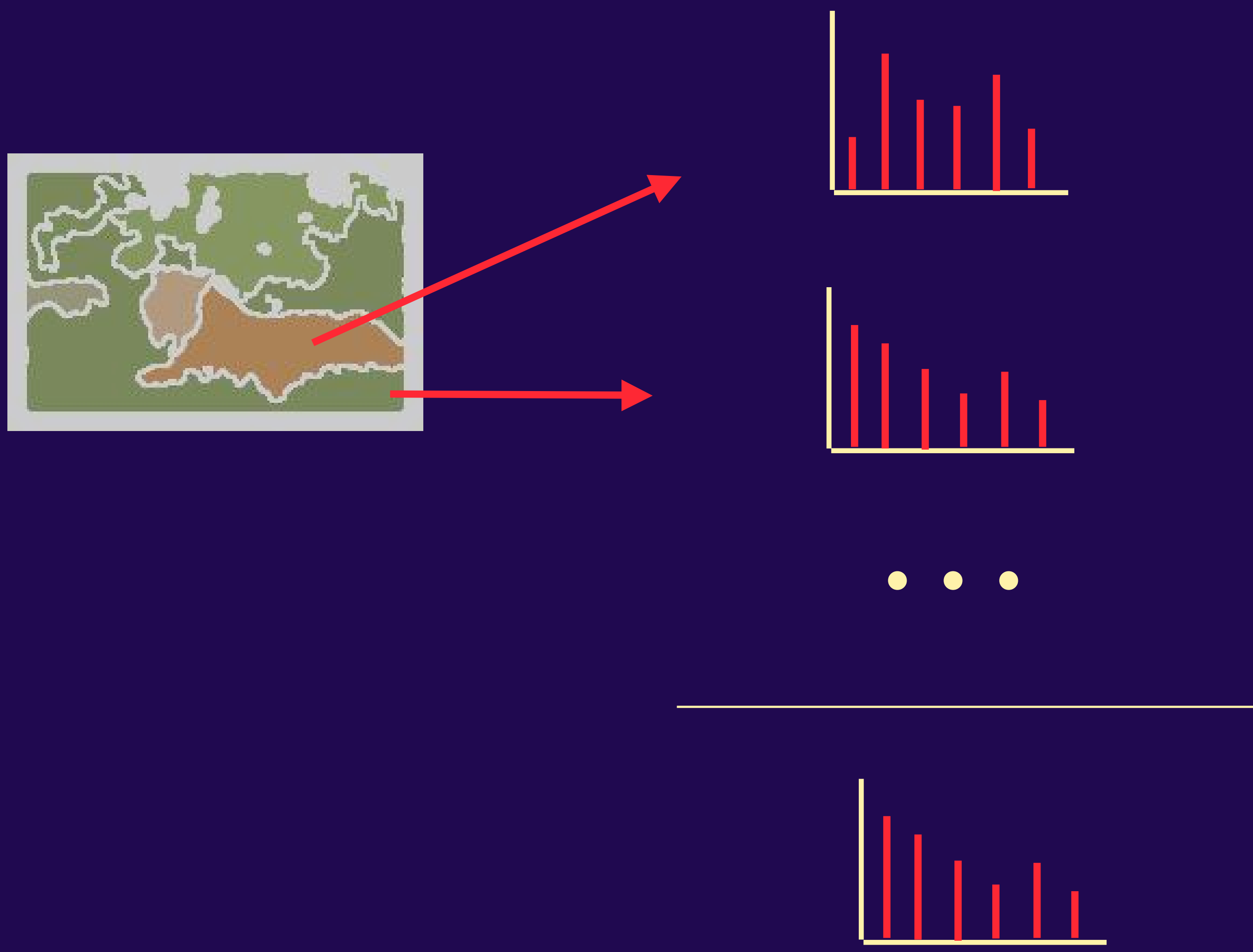
("water")

Second Strategy
Use Annotation

tiger cat grass water

Automatic : Don’t need to do by hand

Annotating Images



Measuring Annotation Performance

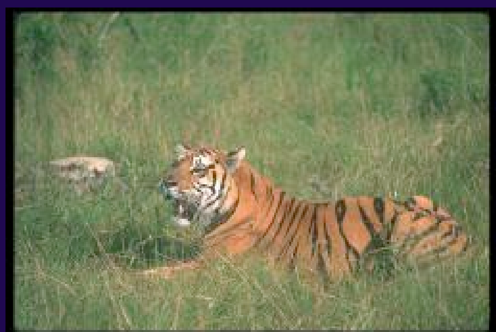
Actual Keywords

GRASS TIGER CAT FOREST

Predicted Words

CAT HORSE GRASS WATER

Measuring Annotation Performance



Actual Keywords

GRASS TIGER CAT FOREST

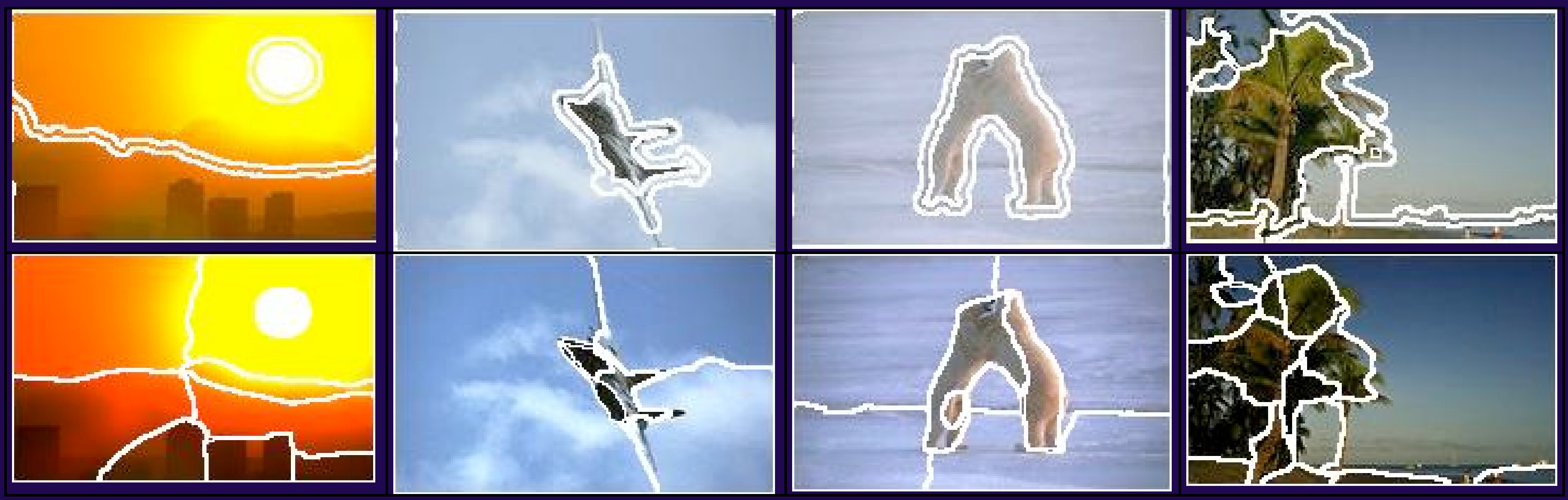
Predicted Words

CAT HORSE GRASS WATER

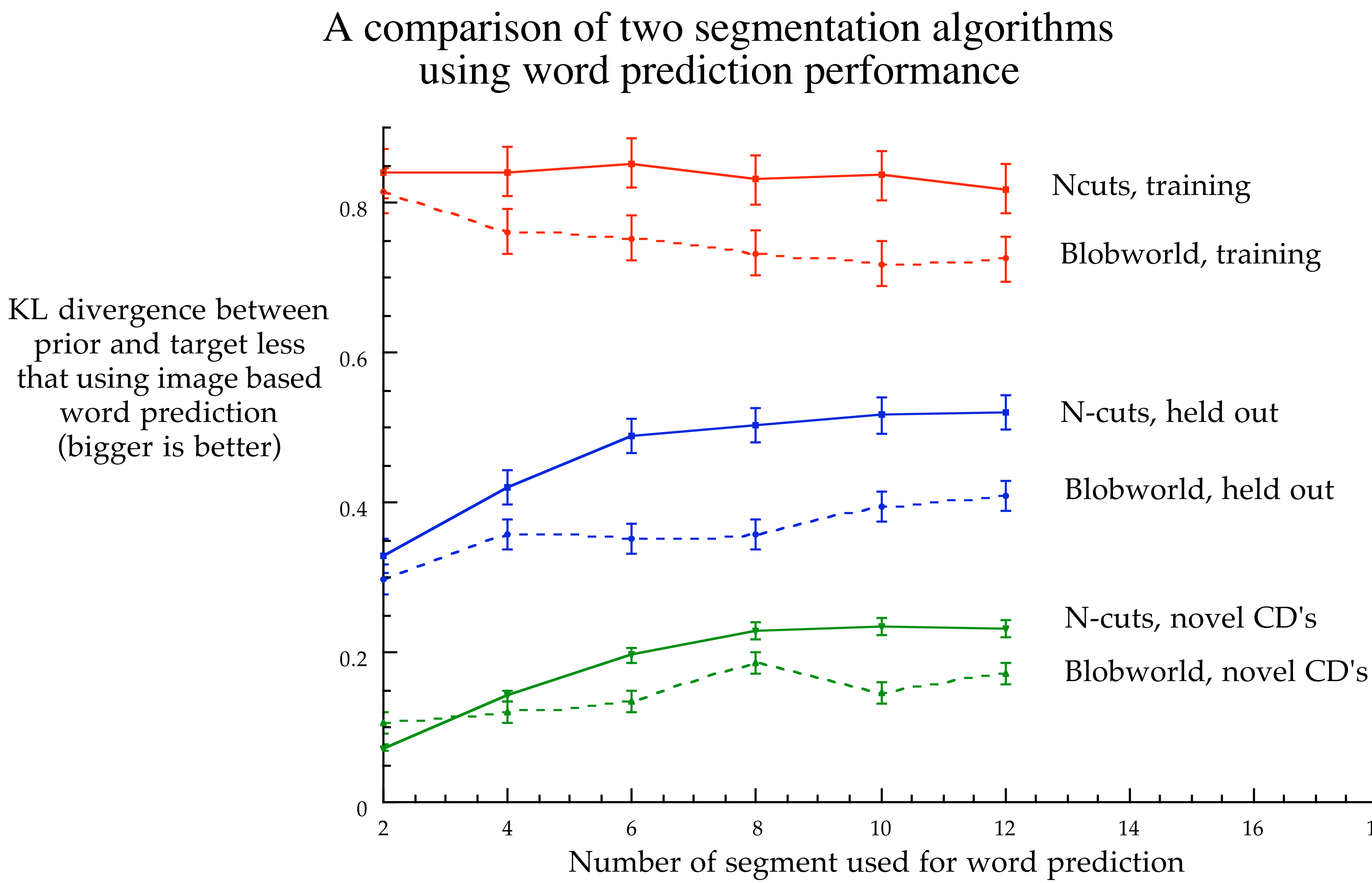
Exploiting Word Prediction

Model Selection
Segmentation
Feature choices

Blobworld segmentations



N-cuts segmentations



A comparison of two segmentation algorithms
using word prediction performance
KL divergence between
prior and target less
that using image based
word prediction
(bigger is better)
Ncuts, training
Blobworld, training
N-cuts, held out
Blobworld, held out
N-cuts, novel CD's
Blobworld, novel CD's
0
0.2
0.4
0.6
0.8
2
4
6
8
10
12
14
16
18
Number of segment used for word prediction

Comments on recognition vs annotations

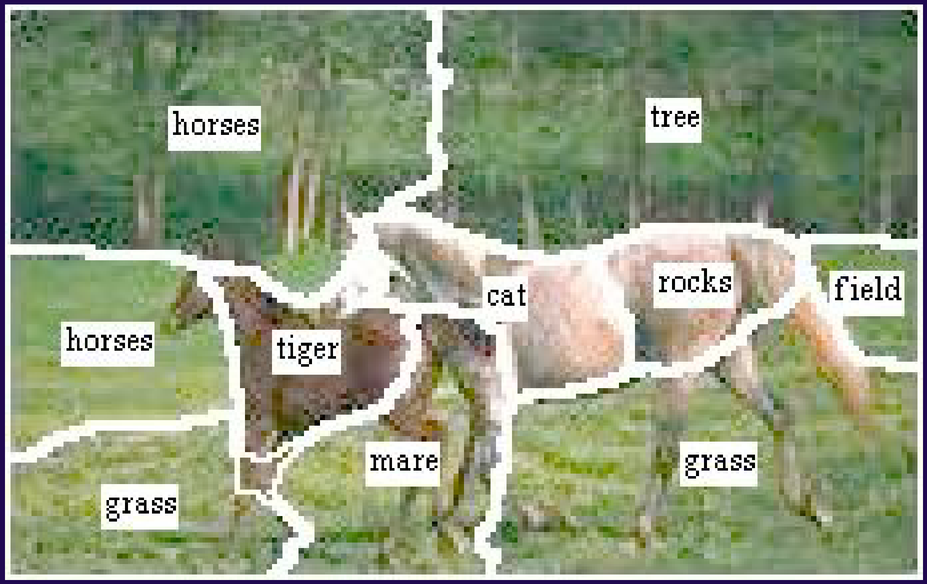
Learning on data without correspondence is a good trick BUT there are fundamental problems

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

Need a better theory of how to link the two.

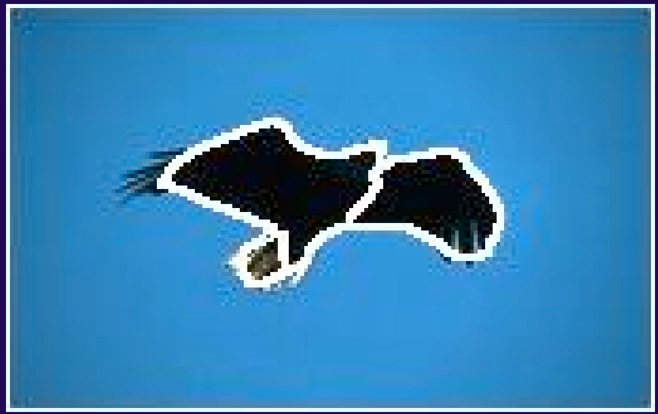
Integrating Supervision

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

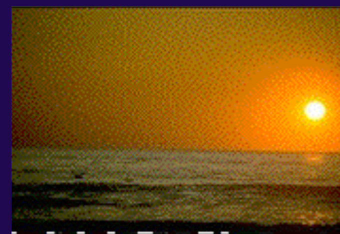
Semantic evaluation of vision tools

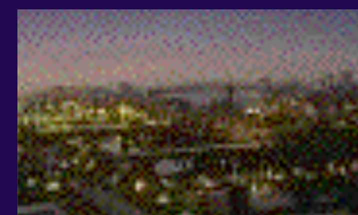
Integrating vision processing levels

Bottom Line

Recognition as machine translation

Machine vision as data-mining





The End