

# **Recognition as Machine Translation:**

## **Labeling Objects and Faces Using Large Image and Video Collections**

**Pinar Duygulu**

**Bilkent University, Turkey**

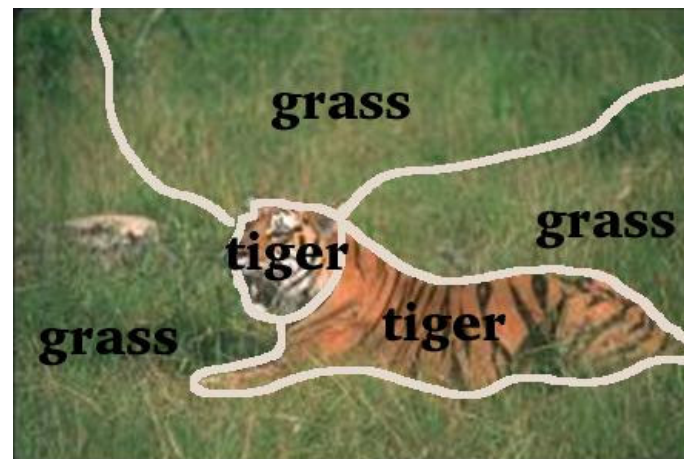
# Outline

- Object Recognition as Machine Translation
  - Joint work with Kobus Barnard, Nando de Freitas and David Forsyth
- Preliminary work on naming faces and objects in news videos

# A novel approach for object recognition

Object recognition on large scale is linking image regions with words

Use joint probability of words and Images in large data sets.



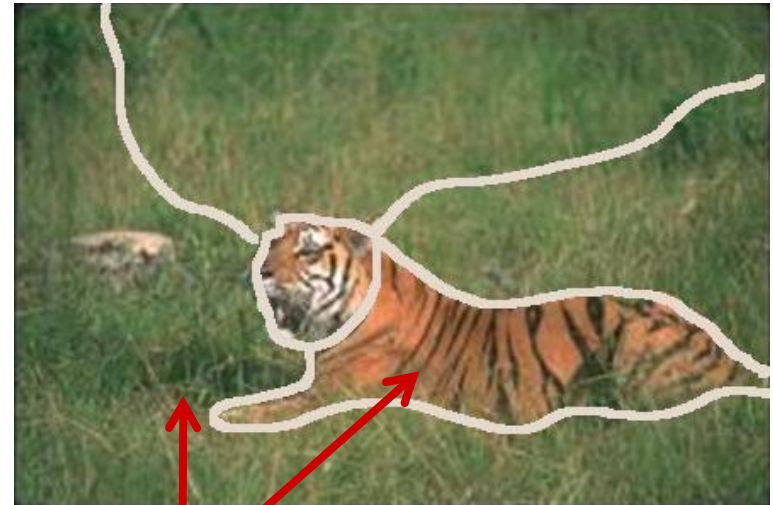
tiger grass cat

# Annotation vs. Recognition



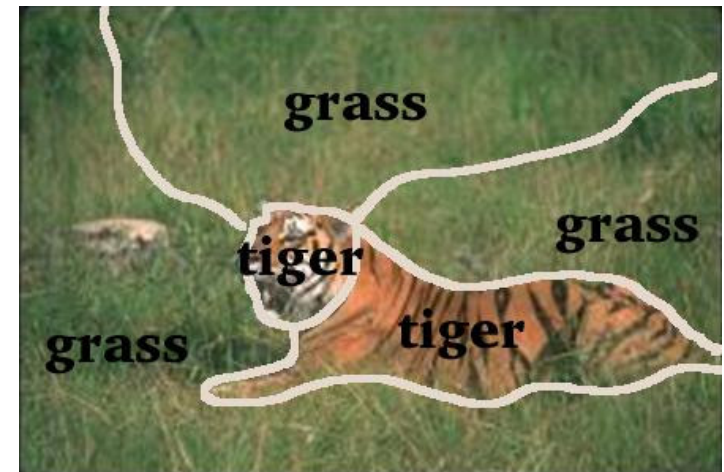
tiger grass cat

Cannot be learned from  
a single image



tiger grass cat

# Learning recognition from large data sets



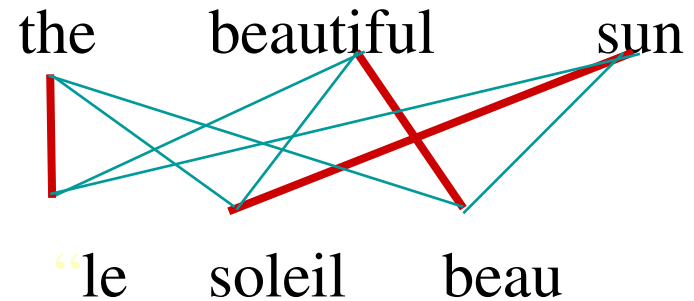
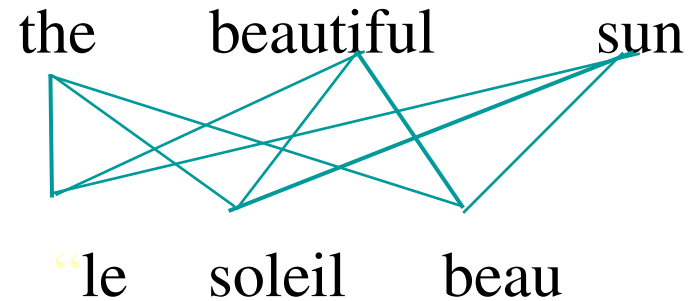
# Statistical Machine Translation

**Data :** aligned sentences  
But word correspondences  
are unknown

- Given the correspondences, we can estimate the translation  $p(\text{sun} \mid \text{soleil})$
- Given the probabilities, we can estimate the correspondences

**Solution:** enough data + EM

Brown et. al 1993



# Multimedia Translation

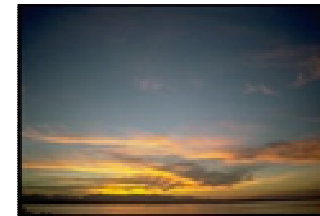
Data :



118011  
WATER HARBOR  
SKY CLOUDS



TIGER CAT WATER GRASS



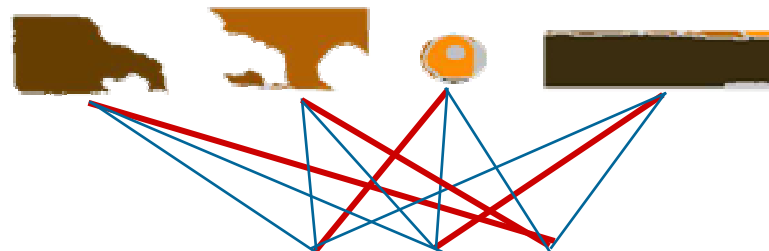
1090  
SUN CLOUDS  
WATER SKY

Words are associated with the images

But correspondences between image regions and words are unknown



“sun sea sky”



“sun sea sky”

# Input Representation



sun sky waves sea

word tokens

↓ segmentation



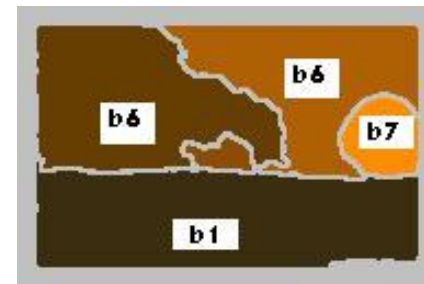
Each blob is a large vector  
of features

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

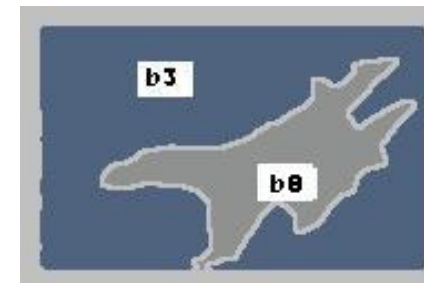
k-means to cluster features

For each blob label of the

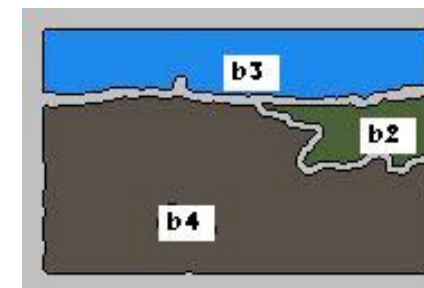
Closest cluster → blob tokens



w6 w7 w8 w1



w3 w4 w5 w1



w12 w2 w1

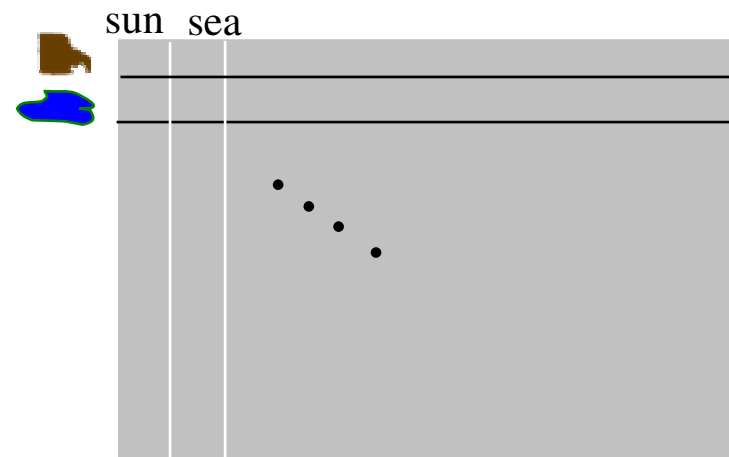


# Method

Given the translation probabilities  
estimate the correspondences

Given the correspondences  
estimate the translation probabilities

## Initialization



Initialize to co-occurrences

# Dictionary

sun



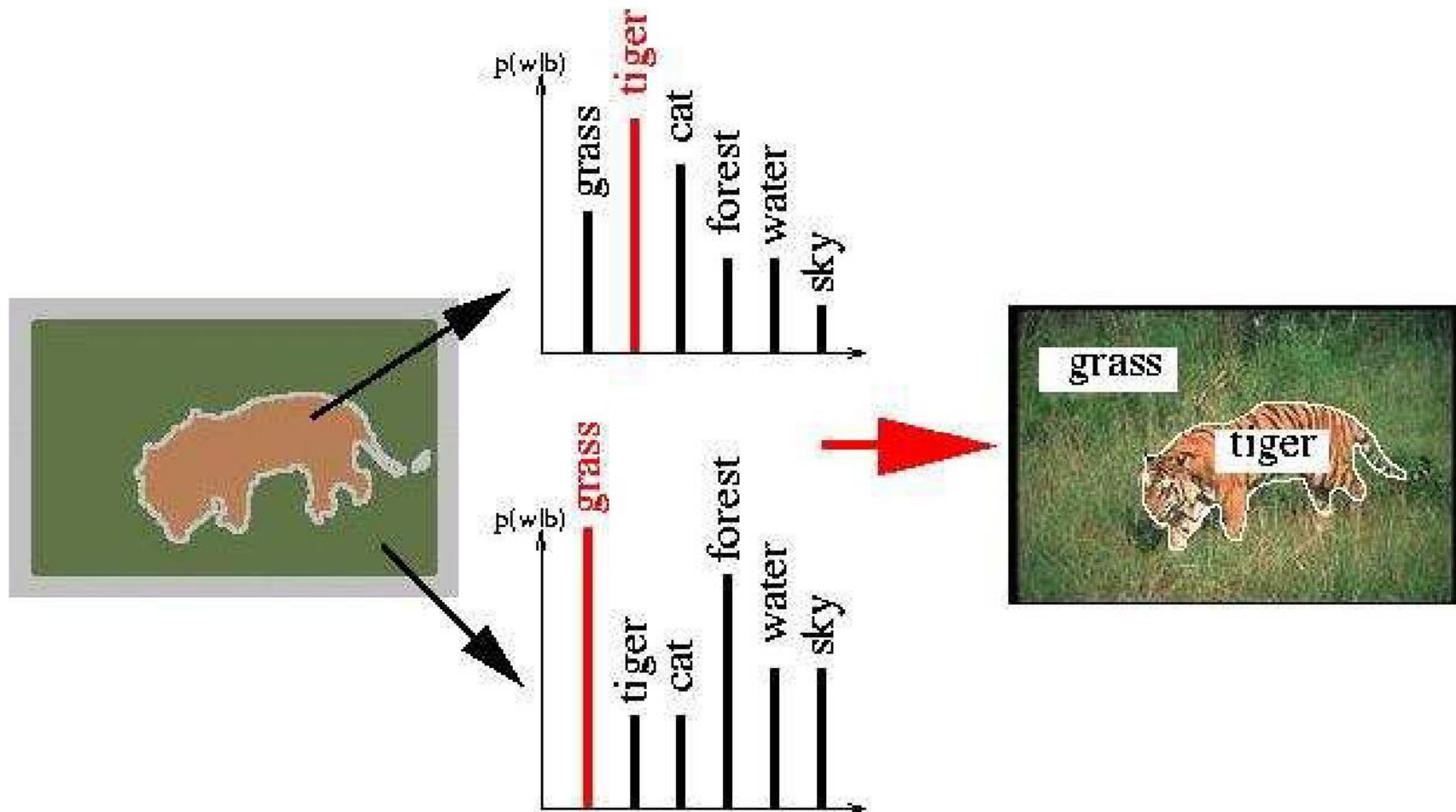
sky



cat



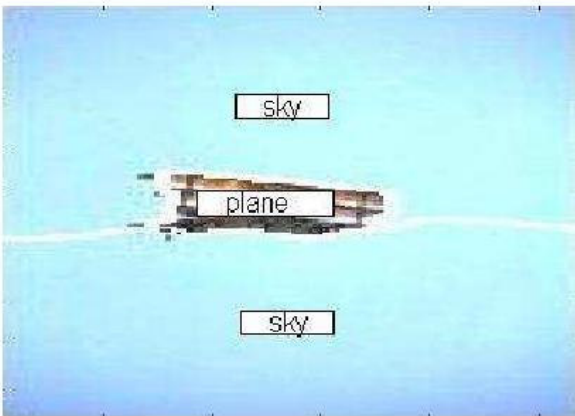
# Region Naming



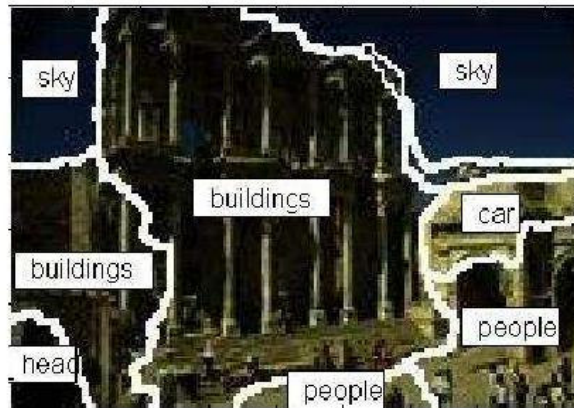
# Results



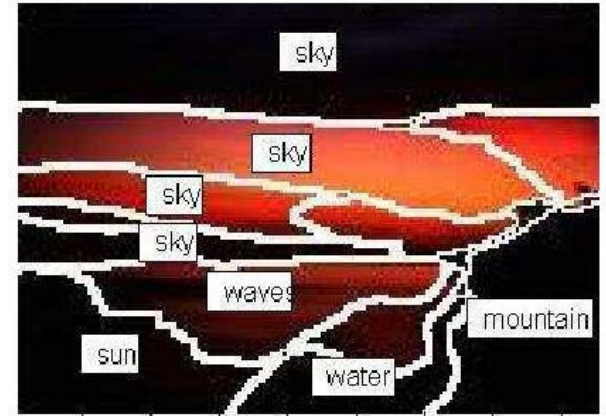
plane sky



people ruins stone



sunset tree water

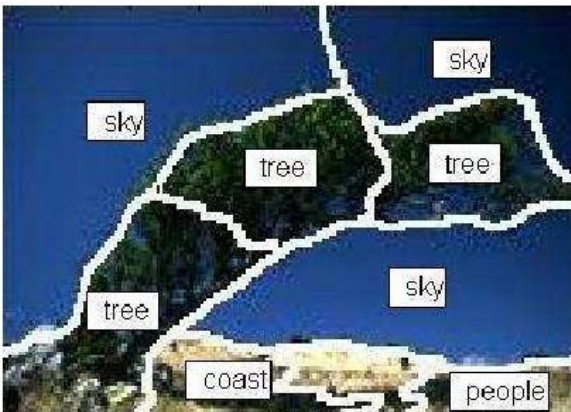




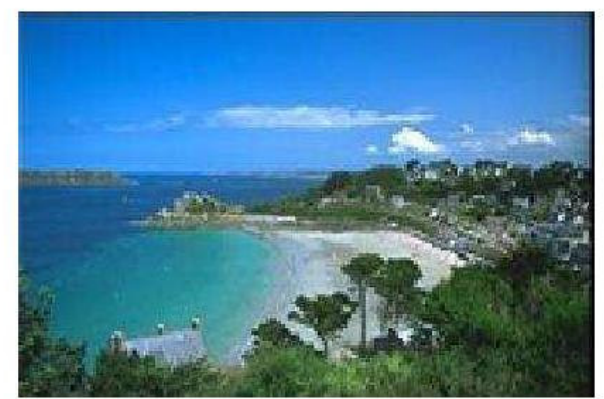
# Results



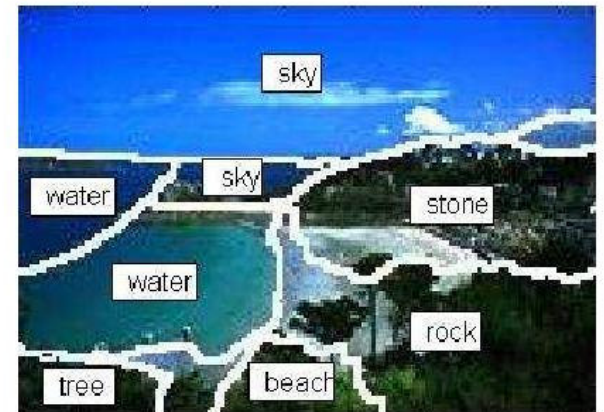
hills sky tree



mountain tree water



beach sky tree water



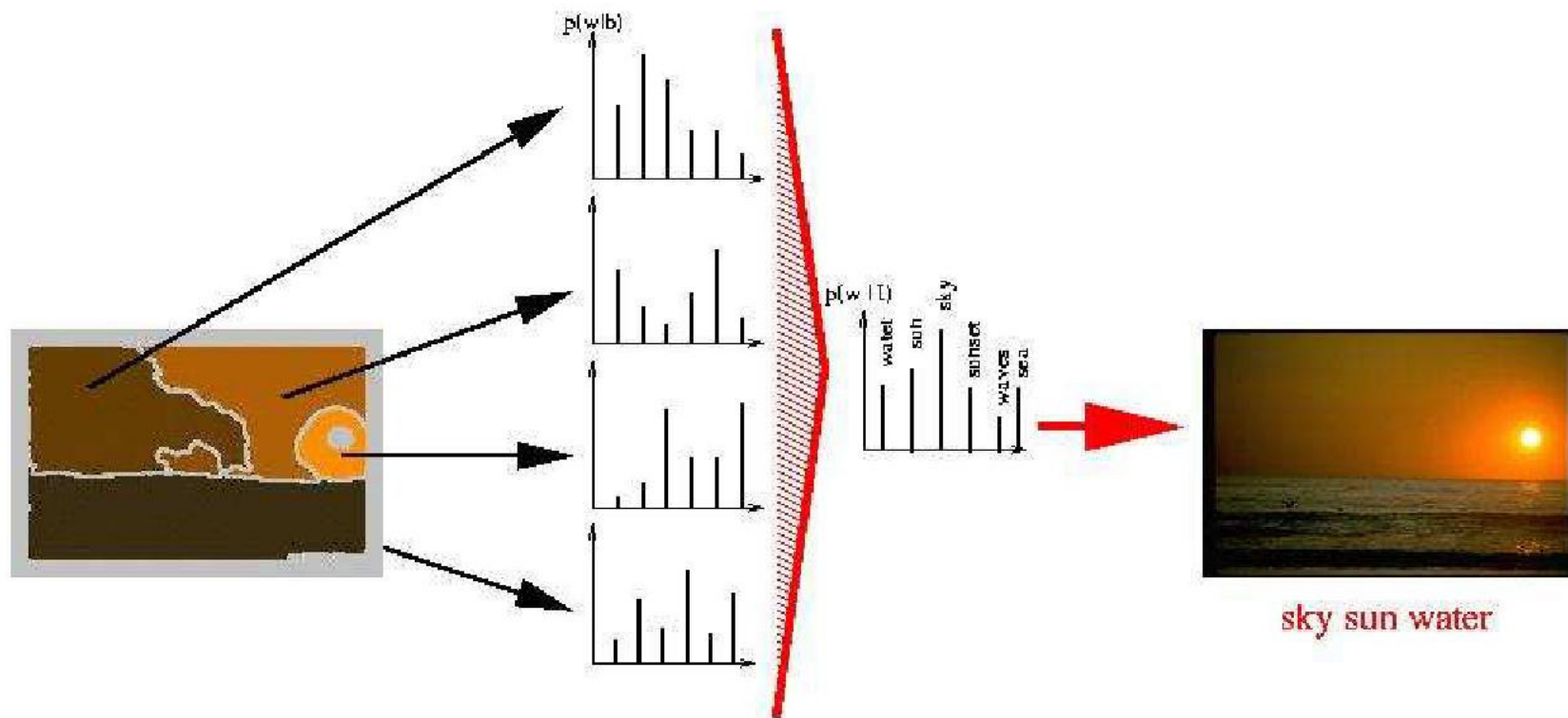
# Measuring the performance



- Do we predict the right words?
- Are they on the right blob?

Visual inspection answers both of the questions, but it is not possible to do for a large number of images

# Auto-Annotation



# Measuring Annotation Performance



Actual Keywords

GRASS TIGER CAT FOREST



Predicted Words

CAT HORSE GRASS WATER



# Measuring Annotation Performance



Actual Keywords

✓  
GRASS TIGER ✓  
CAT FOREST



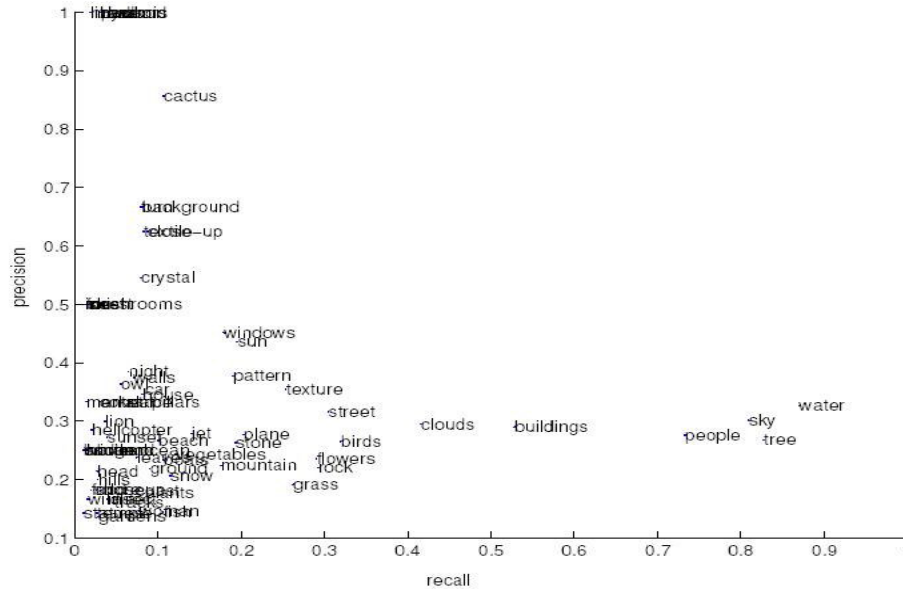
Predicted Words

✓  
CAT HORSE ✓  
GRASS WATER

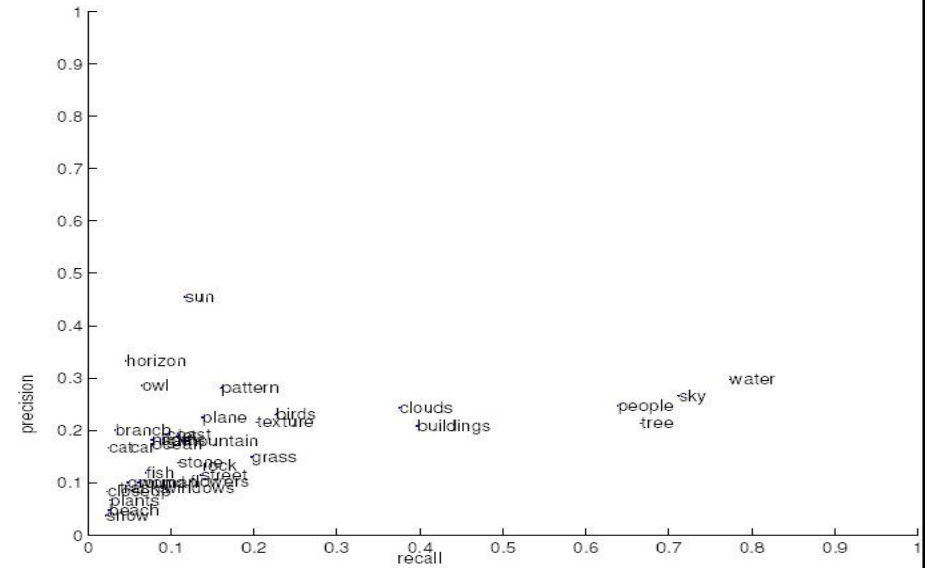
# Word Prediction Measure

set	training	standard test	novel test
001	0.2708	0.2171	0.2236
002	0.2799	0.2262	0.2173
003	0.2763	0.2288	0.2095
004	0.2592	0.1925	0.2172
005	0.2853	0.2370	0.2059
006	0.2776	0.2198	0.2163
007	0.2632	0.2036	0.2217
008	0.2799	0.2363	0.2102
009	0.2659	0.2223	0.2114
010	0.2815	0.2297	0.1991

# Recall versus Precision



training



test

# Refusing to predict

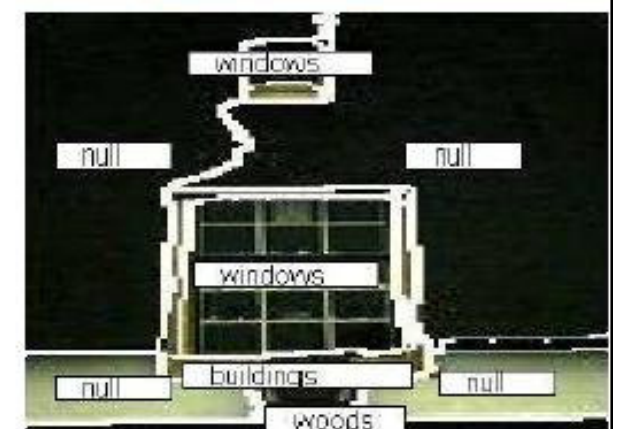
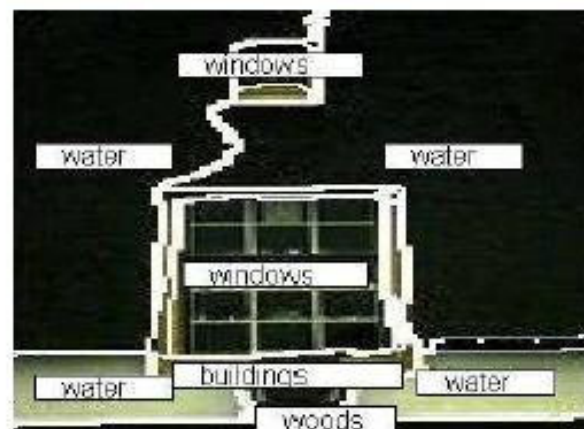
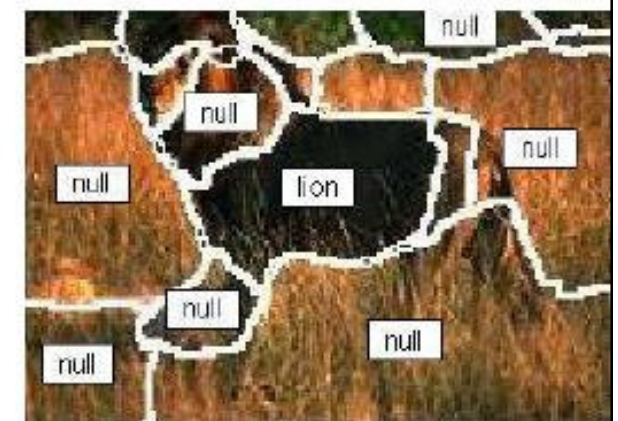
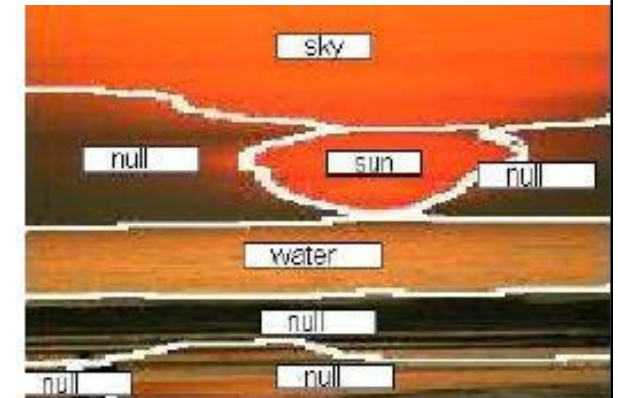
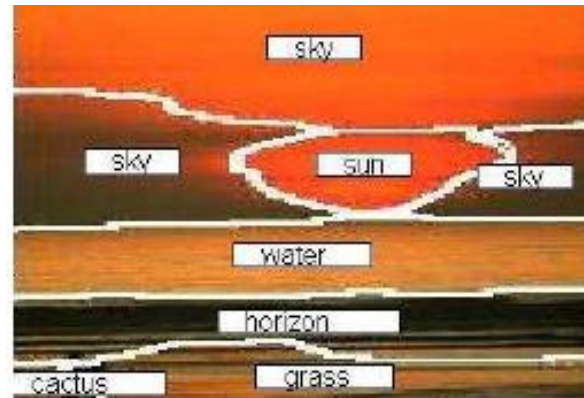
Null and fertility problems

simple solution to null - refusing to predict

If  $\text{prob}(\text{word} \mid \text{blob}) > \text{threshold}$  then  
    predict the word

else

    assign NULL



# Merging Indistinguishable words

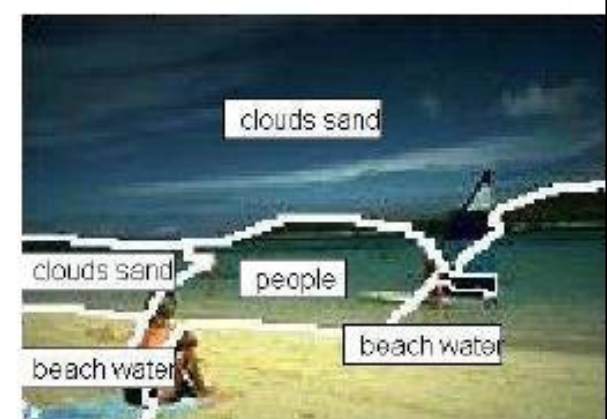
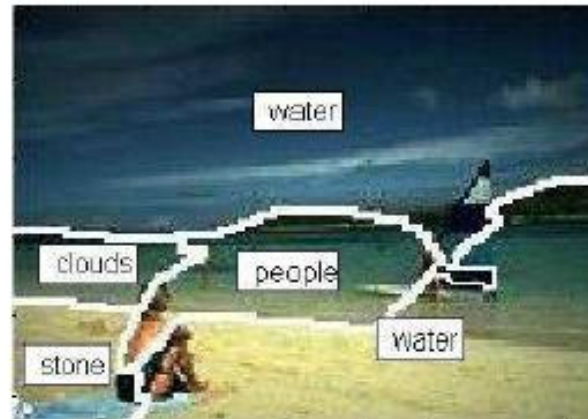
Some words cannot be set apart

either they are synonyms  
(e.g. locomotive and train)

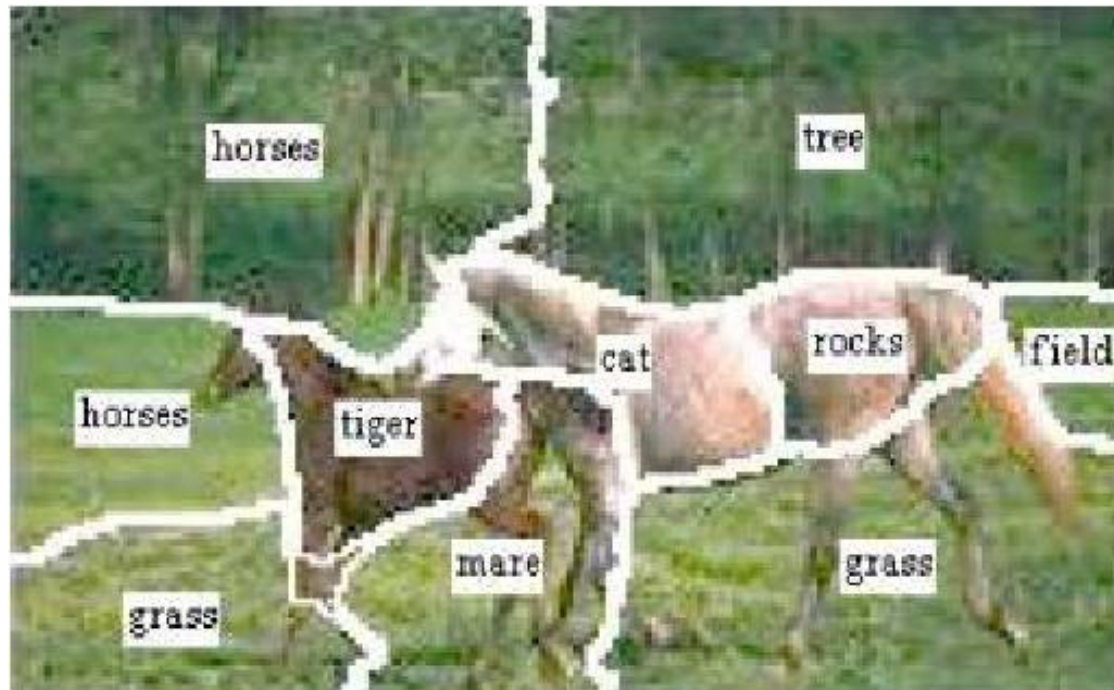
or they are indistinguishable using the current feature set  
(e.g. eagle and jet)

construct a similarity matrix based on the posterior probabilities  
Then, use N-cuts for clustering





# Integrating supervised data



a small amount of supervised data can be helpful  
for breaking symmetries  
for a better clustering



# Integrating labeled data

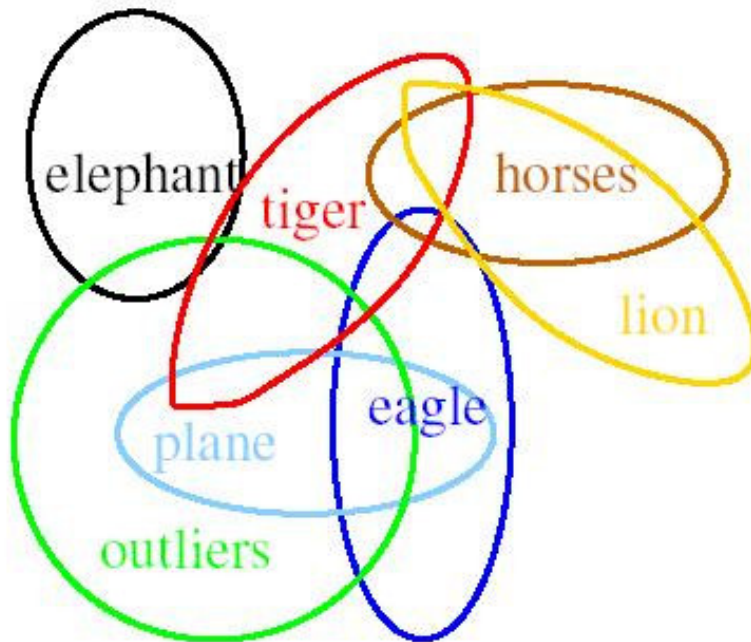
A set of regions are labeled manually

6 CDs , 10 images from each : eagles, elephants, tigers, horses, planes, lions

## For improving clustering

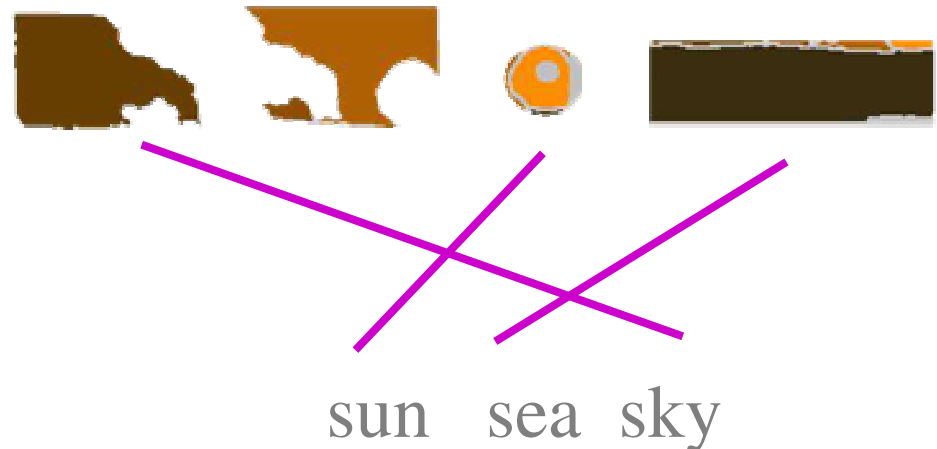
Apply linear discriminant analysis

21 label words + outlier = 22 labeled classes



## For fixing correspondences

Set the alignments between the labeled regions and the corresponding words to 1, and the others to 0



# Integrating supervised data

## first three words predicted

label	method 1	method 2	method 4
tiger	elephant horses field	tiger null water	tiger null water
plane	sky plane forest	plane sky null	plane null sky
runway	null sky eagle	runway plane eagle	runway plane eagle
field	plane null sky	null horses field	field null elephant
horses	tiger null forest	null tiger tree	horses null tiger
sky	forest sky tiger	sky eagle null	sky null eagle
elephant	sky null grass	tree elephant null	elephant null tree
grass	horses null plane	grass horses null	grass horses field
tree	plane sky runway	elephant horses null	tree field horses
water	tiger plane water	water null sky	water null sky
lion	tiger null plane	grass lion tiger	lion grass tiger

	clustering	training
1	k-means	unlabeled data + EM
2	labeled data	unlabeled data + EM
3	labeled data	nearest neighbor classifier
4	labeled data	labeled data + EM

Method 3 (supervised method)  
has more false positives

# Associating video frames with text

President visited agency the chief eavesdrops agency for the intelligence community. Implici...

Obviously, this is all awkward for the President. Whether he'd like to, or not, he can't come out and oppose an investigation. That's not realistic. At the same time, he's made it clear he does not want an independent commission, the Warren Commission of this tragedy, and today he made quite clear that one investigation is enough, there is other work to be done. And, as if to underscore the point, he made the comments at the National Security Agency. We go back to the White House and our senior White House correspondent John King with the President side of things. On, good evening, President visited agency the chief eavesdrops agency for the intelligence community. Private pep talk for the employees there. Came out to speak to reporters. Clearly looking to put his stamp in the investigation. President saying for all the revelation, some would say outrages about what the government knew but did not share between agencies between—before September 11th. Convinced no evidence at all that he has seen the government could have done anything to stop the attacks. President said there's also no question that the FBI and the CIA failed to share clues and suspicions about the terrorist threat. President Bush: In terms of whether or not the FBI and the CIA were communicating properly, I think it is clear that they weren't, now we've addressed that issue, CIA and FBI are now in close communications. There's better sharing of intelligence. One of the things that essential to win this war is to have the best intelligence possible. When we get the best intelligence, to be able to share it throughout our government. Implicit in President's remarks both of CIA director and FBI director as investigation get under way. President putting popularity and prestige in the line insisting Congress stick to one investigation into what the government knew before 9/11. President publicly playing down the finger point of late between the FBI and CIA. We're told by several senior fishes at the White House. President is quite furious, he and other top officials plied—made clear to Mueller, and deputies President execs and end to sniping at one senior official said tonight we don't need self-inflicting wounds. Aaron: Institutional question that gone on longer than any of us has gone along, does the President accept—I think the word I want is responsibility—whatever went wrong in the CIA and FBI happened on his watch and that he's the guy and he's responsible. Yes and no. If you like President certainly says he's the man in charge of the government. He's responsible but also has said and his senior aides are saying in more detail, many of the communication break down go back in the Clinton administration, this is not a new issue if you will. Brent National Security advise her former Bush's administration, he is not—now the add migs point man, whether all the agencies need to be reformed. Whether defense agencies could be pulled from the Pentagon and matched to the CIA. Now decade, getting more attention because the tragedy President say he's responsible for the government? Yes. Will he say the problems predate him, he'll say yes too. Aaron: Thank you, senior correspondent John King tonight.

Query on  
“president”

Association  
problem



# Associating video frames with text

Solving correspondences in broadcast news for better retrieval & sense disambiguation



..tanks on the street ...



..start attacking on houses  
by helicopters and tanks...



..fuel tank...

Face Recognition by resolving correspondences between named entities and faces



# Data Set

- TRECVID2003 video retrieval evaluation data provided by NIST
- 120 hours of news videos (ABC and CNN) from 1998



# Concepts



man-made object



car  
transportation  
vehicle  
outdoors  
non-studio setting  
nature-non-vegetation  
snow

138 concepts

15 hours of video is manually annotated by around 100 people

# Representation



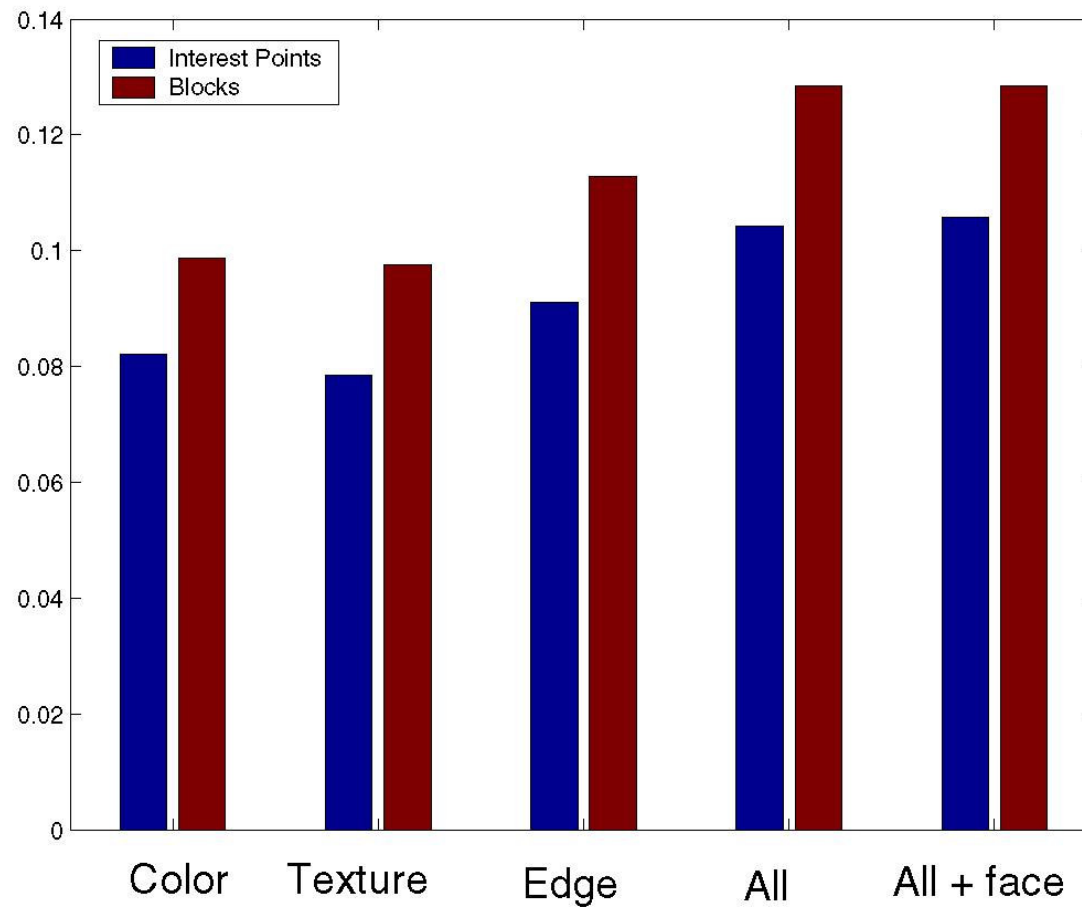
Features extracted

- Lab Moments
- Smoothed Edge Orientation Histogram
- Grey-level Co-occurrence matrix

1000 visual tokens

138 concepts

# Feature comparison



mAP values for different features

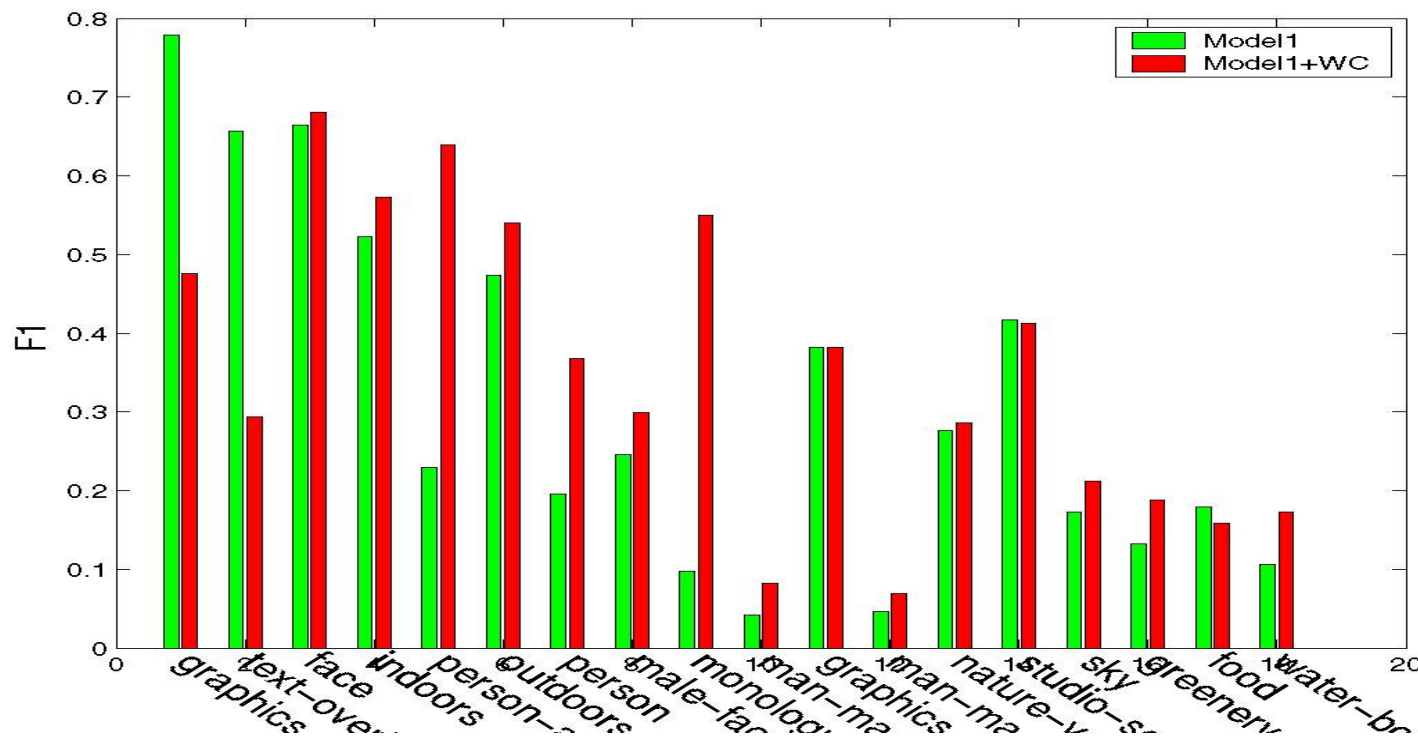


# Integrating Language Modeling

Word co-occurrences as a simple language model

$$P_1(c_i | v) = \sum_{j=1}^{|C|} P(c_i | c_j) P_0(c_j | v)$$

v: visual tokens, c: concepts



Training:  
9413 shots,  
Test :  
4787 shots

Lab Moments

$F1 = (\text{recall} + \text{precision}) / 2$

# Concepts versus Speech Transcripts

## Concepts

- Requires manual annotation

- Noisy

- Limited set of vocabulary

## Speech transcripts and closed captions

- Available for almost all the videos

- Free text which usually does not correspond to the visual cues

- Text is not associated with the frames



...despite heroic efforts many of the worlds wild creatures are doomed the loss of species is now the same as when the great dinosaurs become extinct will these creatures become the dinosaurs of our time today...

# Associating key-frames with surrounding text



...efforts many | of the worlds | wild creatures | are doomed | the loss of | species ...

Brill's tagger is used to extract nouns  
Stop words and rare words are eliminated

# News videos - structured

Taking the surrounding words are problematic  
Use structure to obtain story segments

*anchor*

*anchor – reporter dialogs*

*logos*

*overview*



*News story*

*weather*

*commercials*

*sports*

# Using delimiters to obtain story segments

## Remove commercials



## Remove graphics



## Remove anchor frames but use text

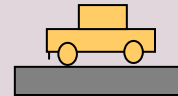


- Idea : A story segment starts or ends with a delimiter or with an anchor/reporter shot

# Associating text with frames

w1 w2 w10 w1 w5 w6 w2 w1 w4 w10 w5 w3 w11

...

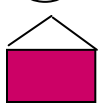


...

Color tokens : 1-230 (quantized using G-means)



Num faces (1 / 2 /  $\geq 3$ )



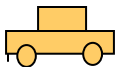
building



road



outdoor



car

# Semantic retrieval

!! only single occurrence per segment

Search on clinton



20 / 130 (15%)

27 / 133 (20%)

Search on fire



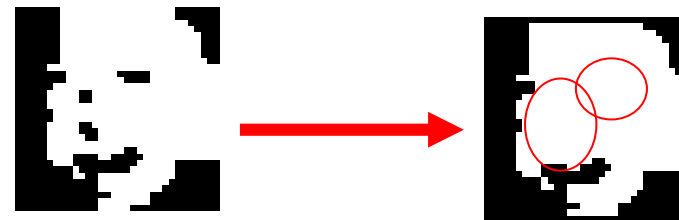
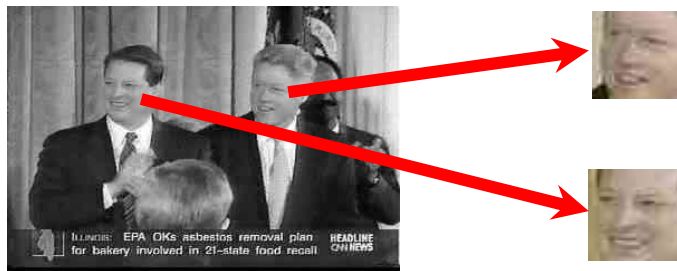
11 / 44 (25%)

15 / 38 (40%)

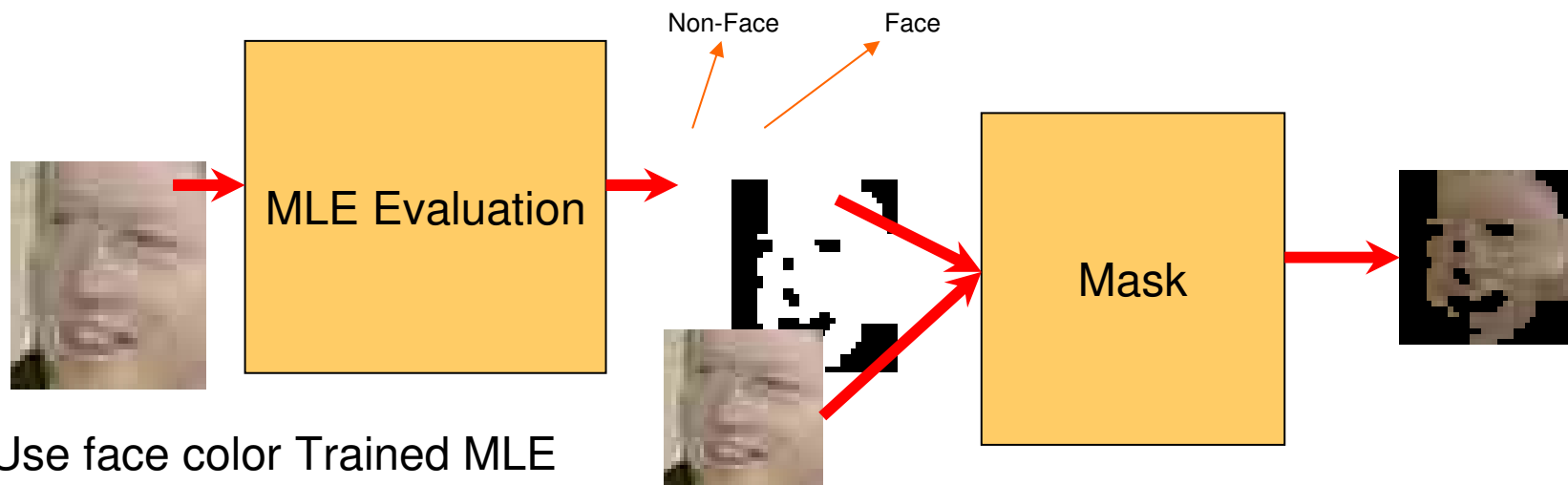


# Detecting Faces

Detect Faces



Improve amount of “face” by filling regions of the mask

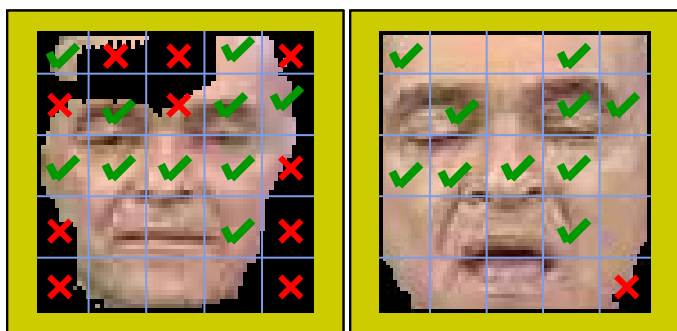


Use face color Trained MLE to eliminate background

Mikolajczyk, Schmid



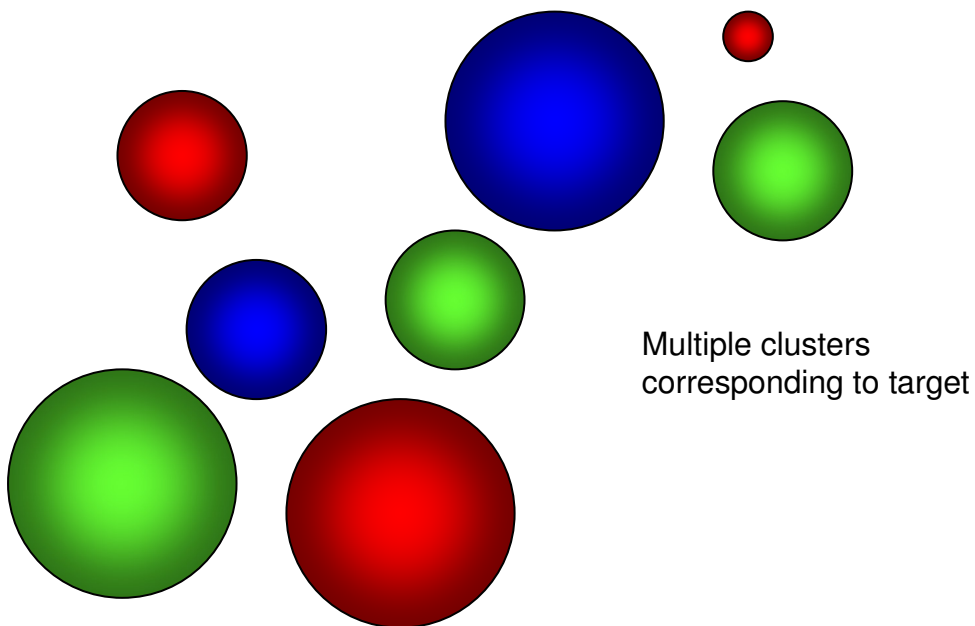
# Methodology



Similarity < Threshold  
✓ Good Match  
✗ Not enough Data

Join Similar faces  
together to obtain  
clusters of images.

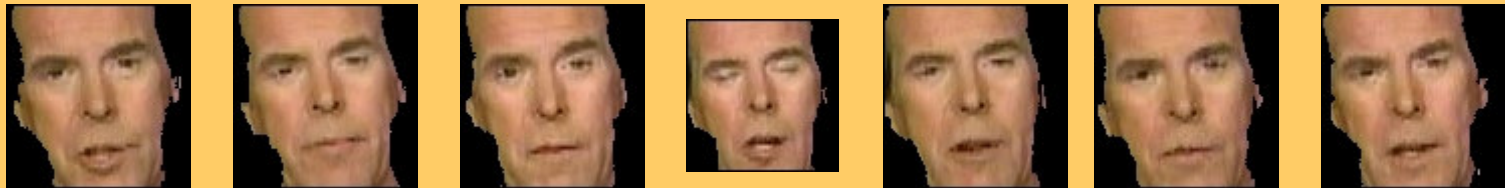
Label the cluster  
with most possible  
name extracted  
from speech data.



Multiple clusters  
corresponding to target

# Results

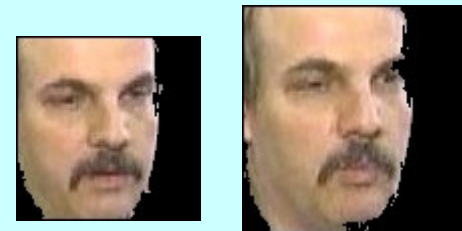
Sam Donaldson



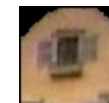
Anchor (ABC)



Unknown



Anchor (CNN)



# Summary & Future Directions

- When text and visual features are combined it is possible to do many interesting tasks
- Object recognition on the very large scale can be viewed as translation of regions to words
- Important objects
  - People
  - Objects that move
- Use temporal information and associate the motion with words
- There are many other available multi-modal data sets
  - Recognizing words in Ottoman documents