Recognition as Machine Translation:

Labeling Objects and Faces Using Large Image and Video Collections

Pinar Duygulu

Bilkent University, Turkey

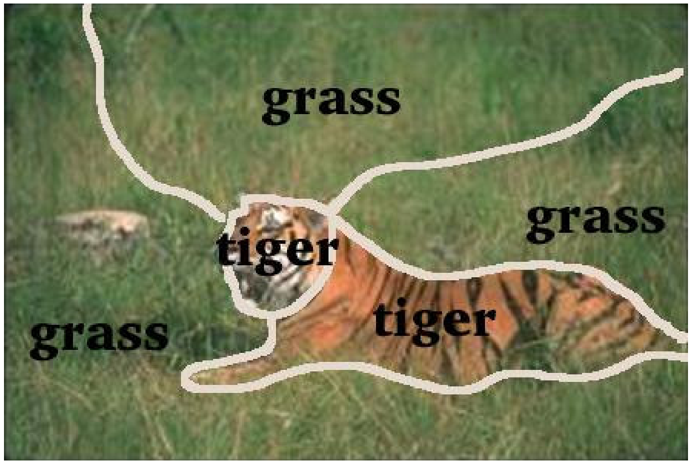
Outline

- 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

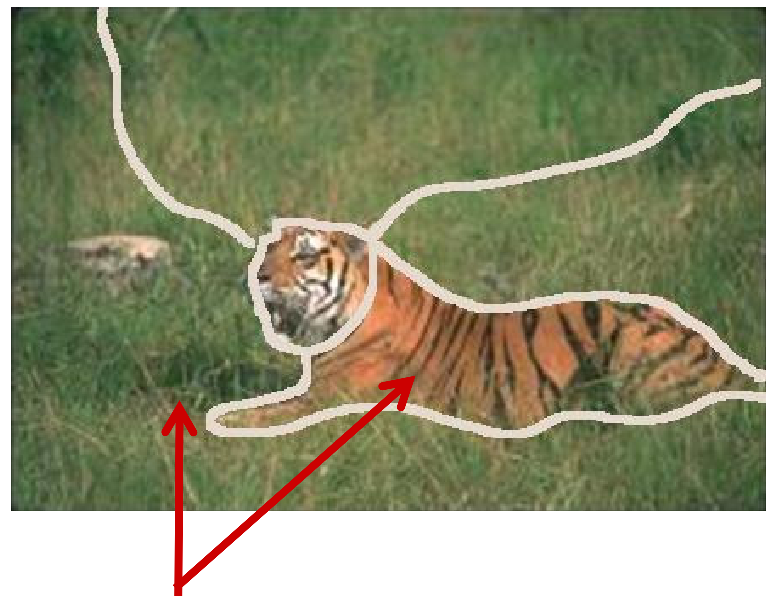


Use joint probability of words and Images in large data sets.

tiger grass cat

Annotation vs. Recognition

tiger grass cat

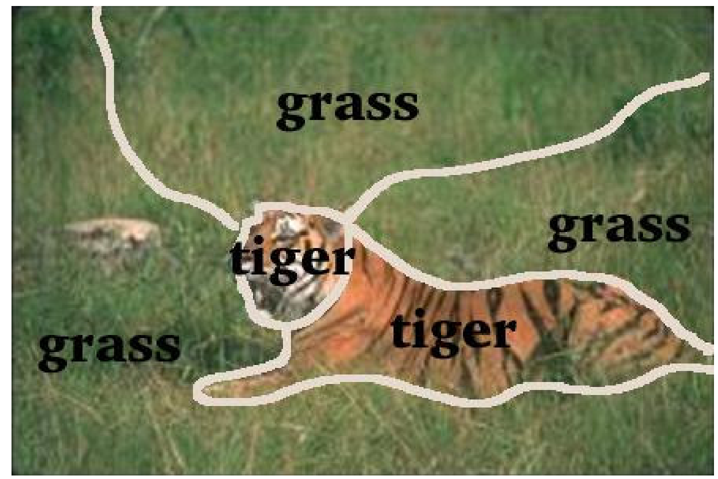


?

tiger grass cat

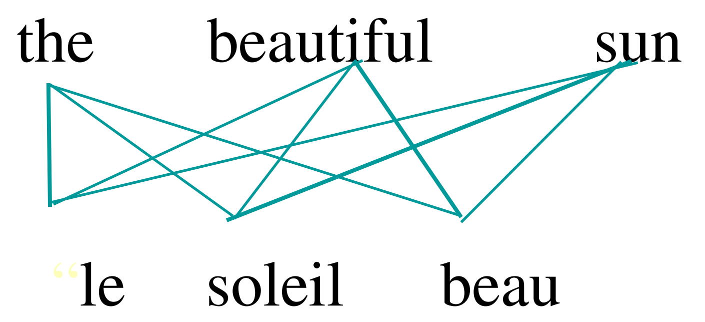
Cannot be learned from
a single image

Learning recognition from large data sets

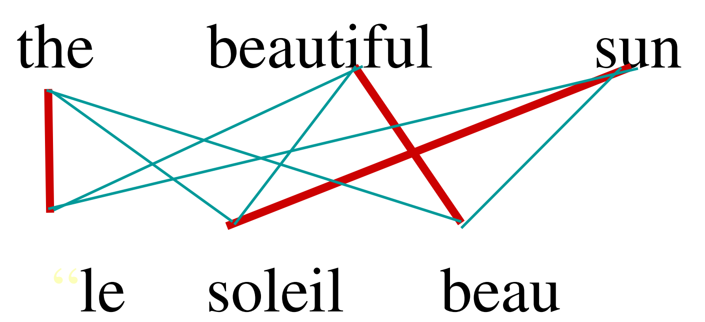


Statistical Machine Translation

Data : aligned sentences
But word correspondences
are unknown



•Given the correspondences,
we can estimate the translation
p(sun | soleil)
•Given the probabilities, we can
estimate the correspondences



Solution: enough data + EM

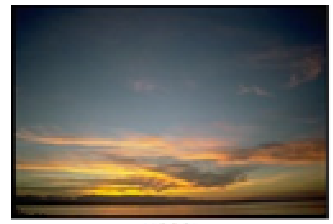
Brown et. al 1993

Multimedia Translation

Data :

118011
WATER HARBOR
SKY CLOUDS

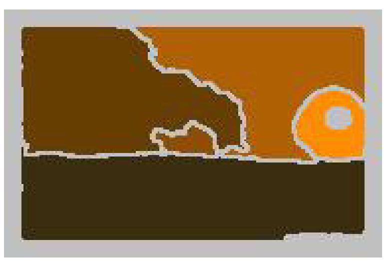
TIGER CAT WATER GRASS



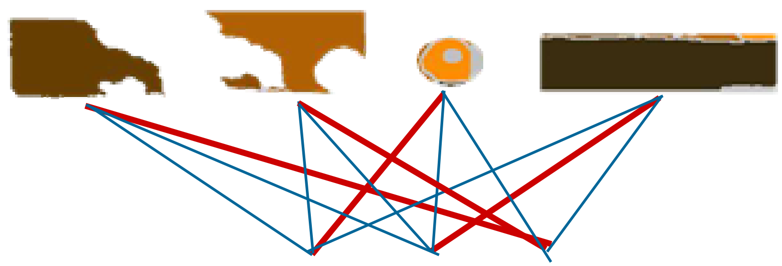
1090
SUN CLOUDS
WATER SKY

Words are associated with the images

But correspondences between image regions and words are unknown



"sun sea sky"



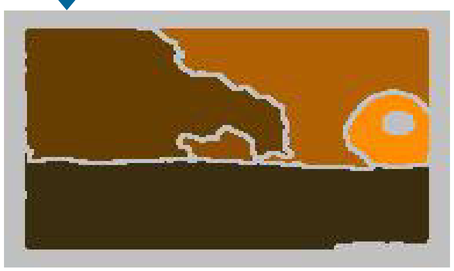
"sun sea sky"

Input Representation

sun sky waves sea

word tokens

segmentation



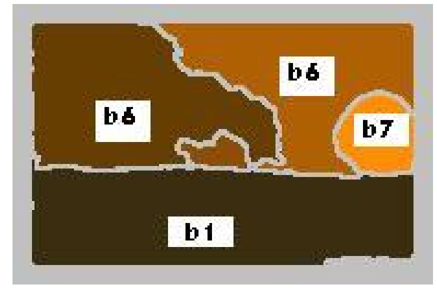
Each blob is a large vector of features

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

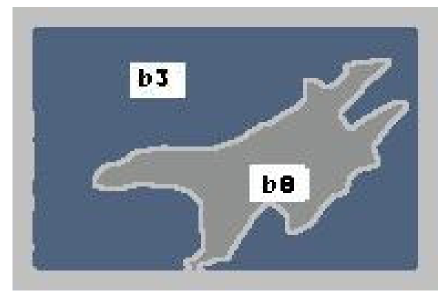
k-means to cluster features

For each blob label of the

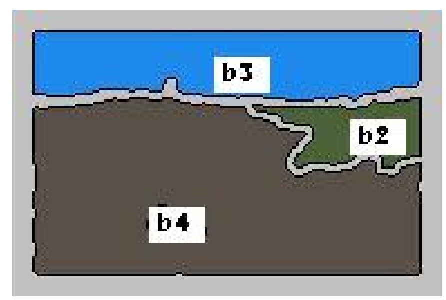
Closest cluster→ blob tokens



w6 w7 w8 w1

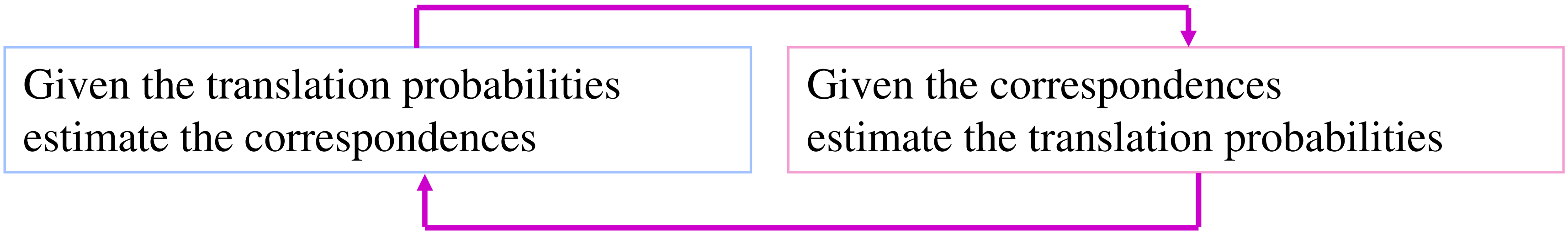


w3 w4 w5 w1

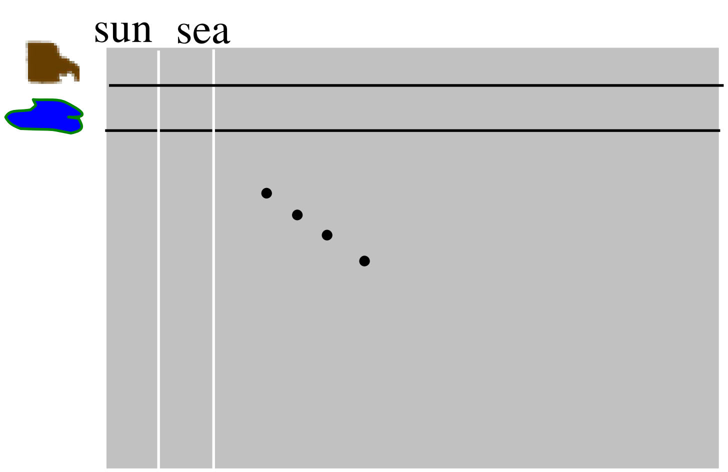


w12 w2 w1

Method



Initialization

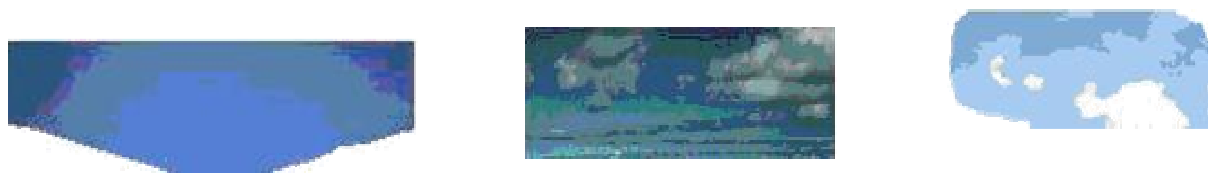


Initialize to co-occurrences

Dictionary

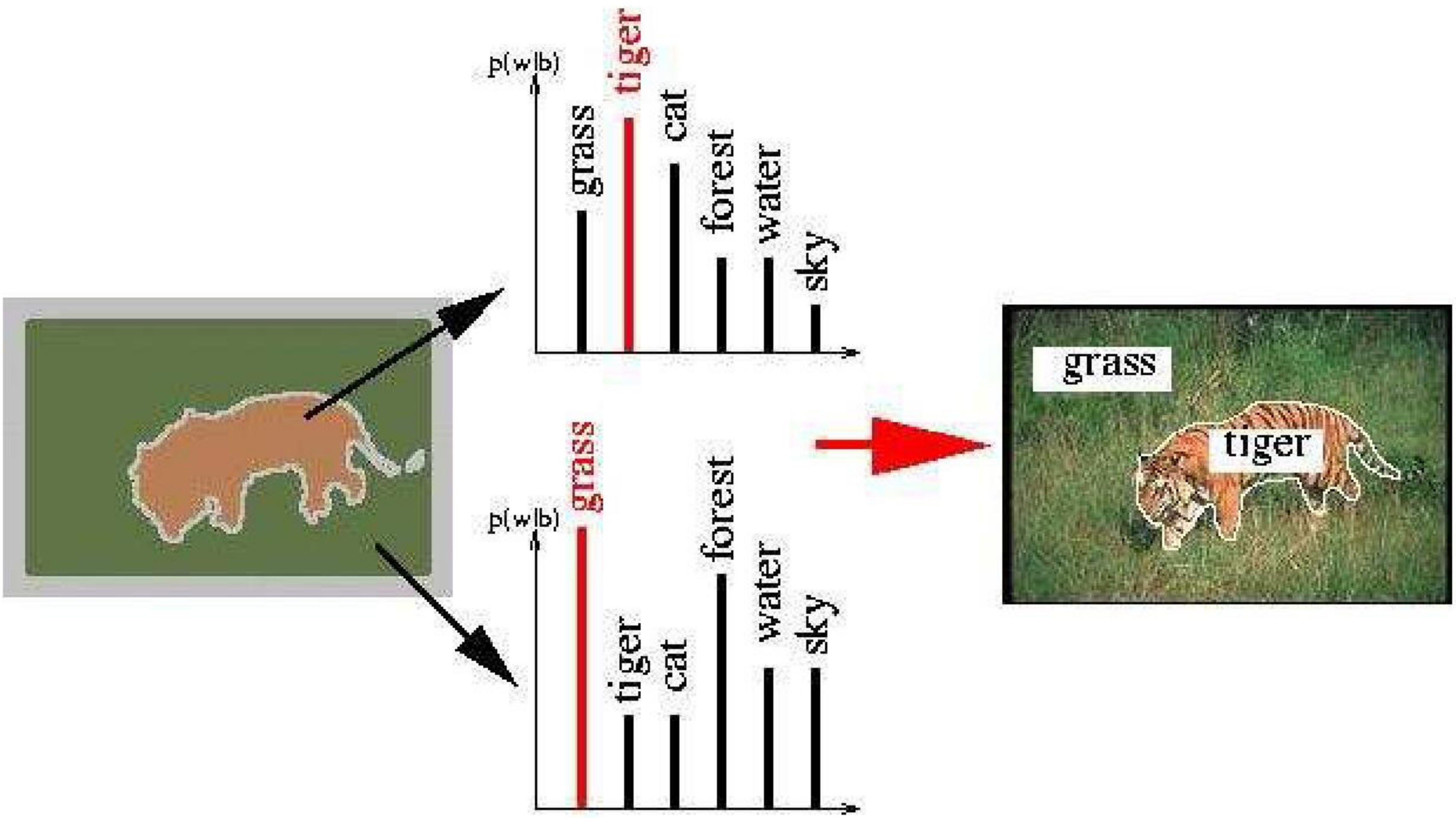
sun

sky



cat

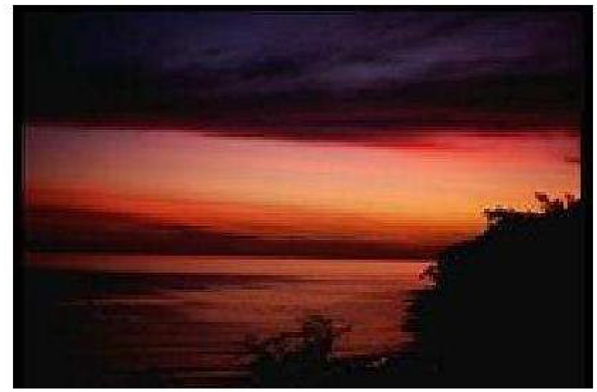
Region Naming



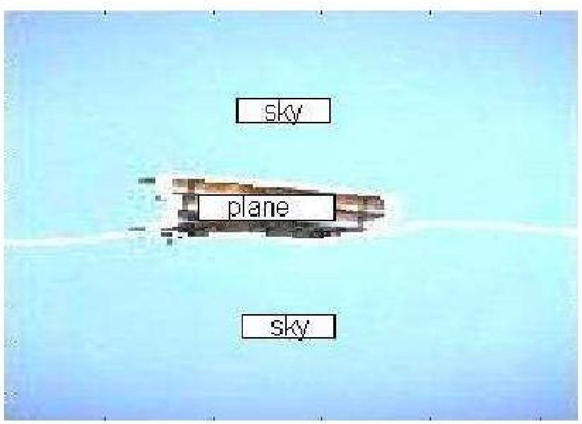
Results

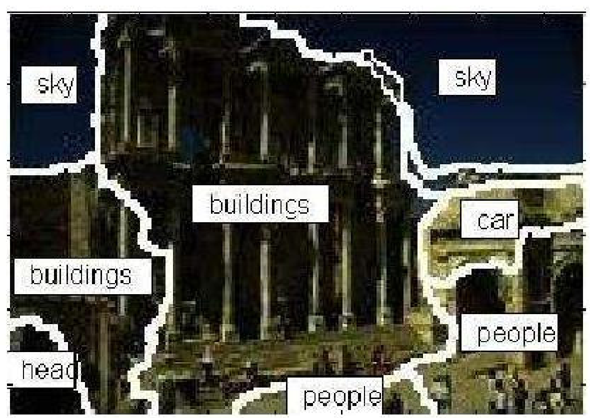
plane sky

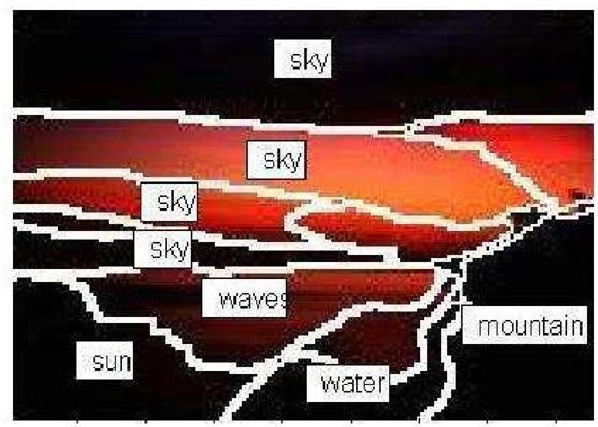
people ruins stone



sunset tree water







Results

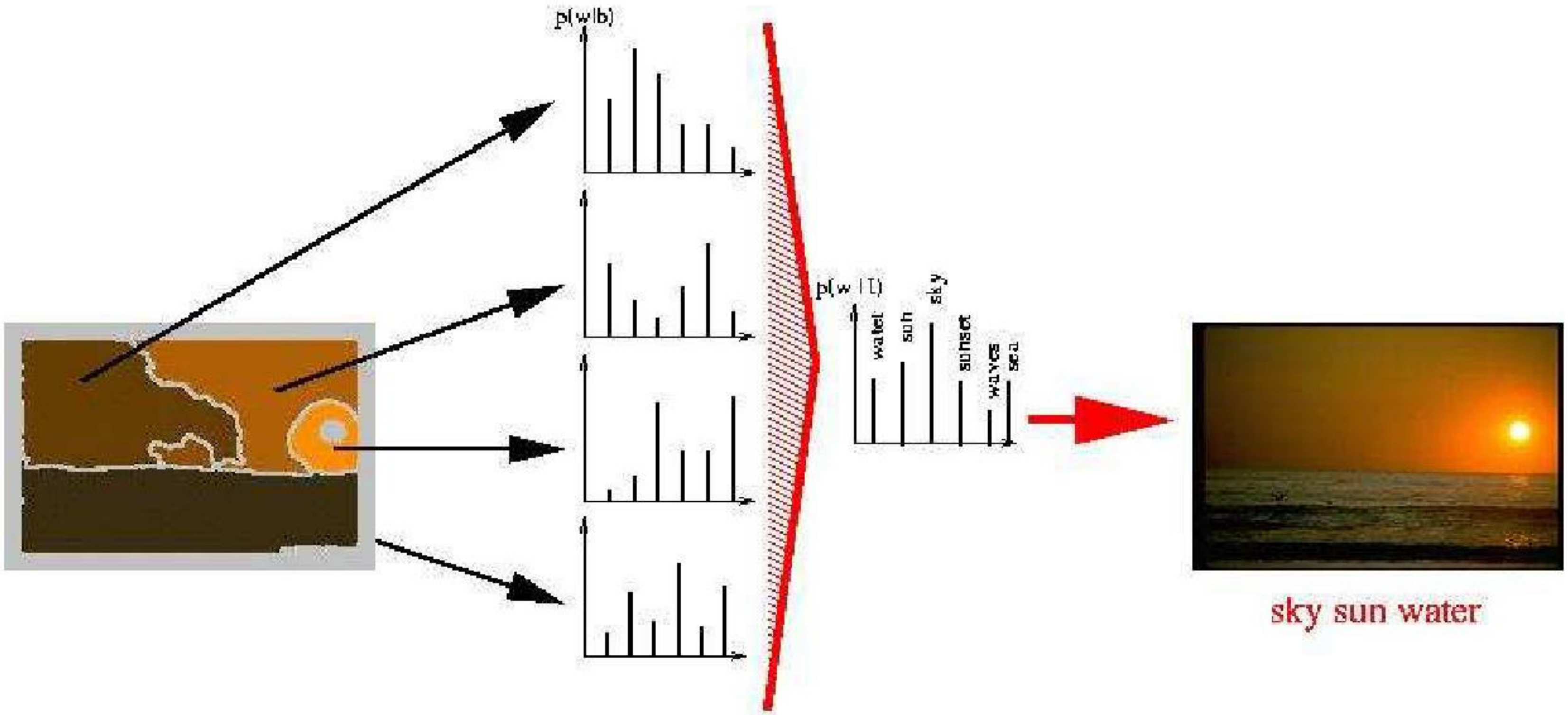
hills sky tree | mountain tree water | beach sky tree water

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

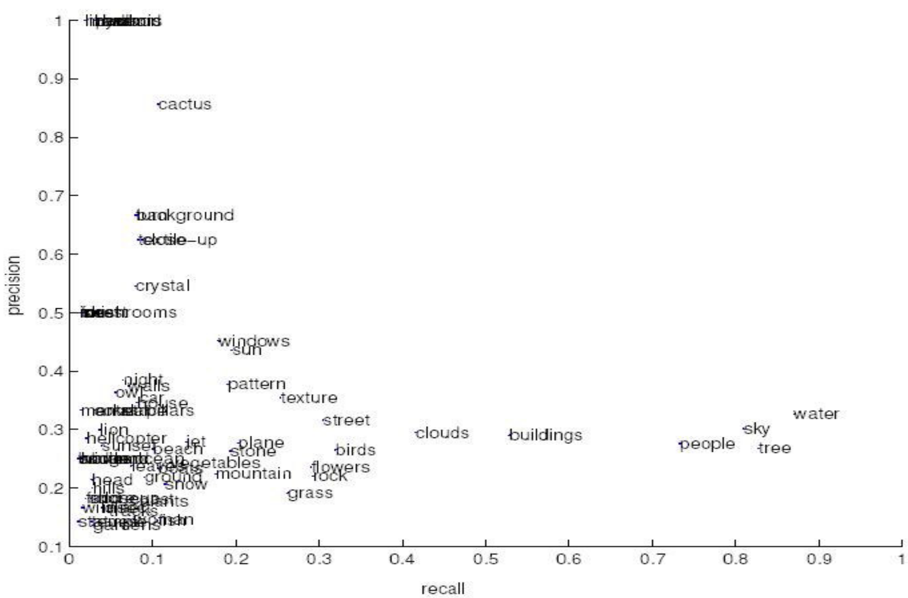
Predicted Words

CAT HORSE GRASS WATER

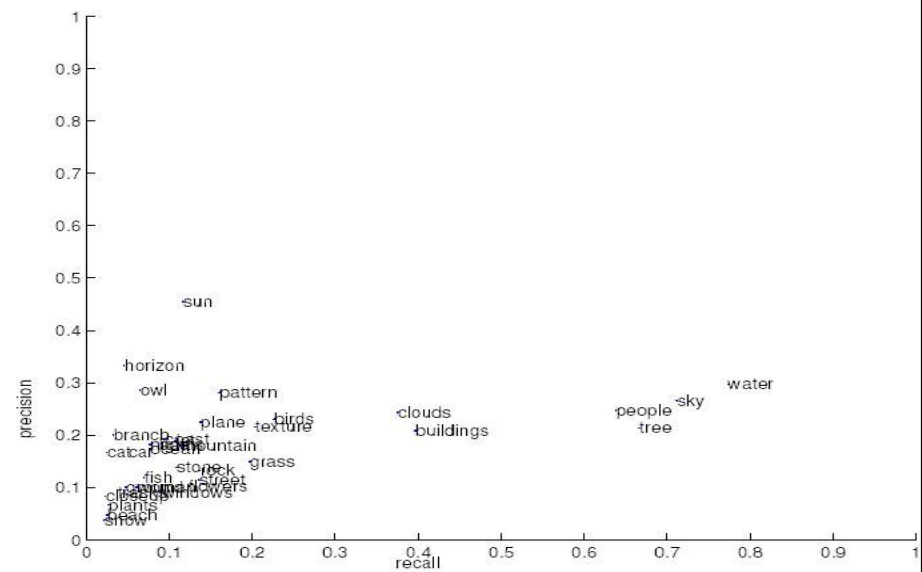
Word Prediction Measure

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

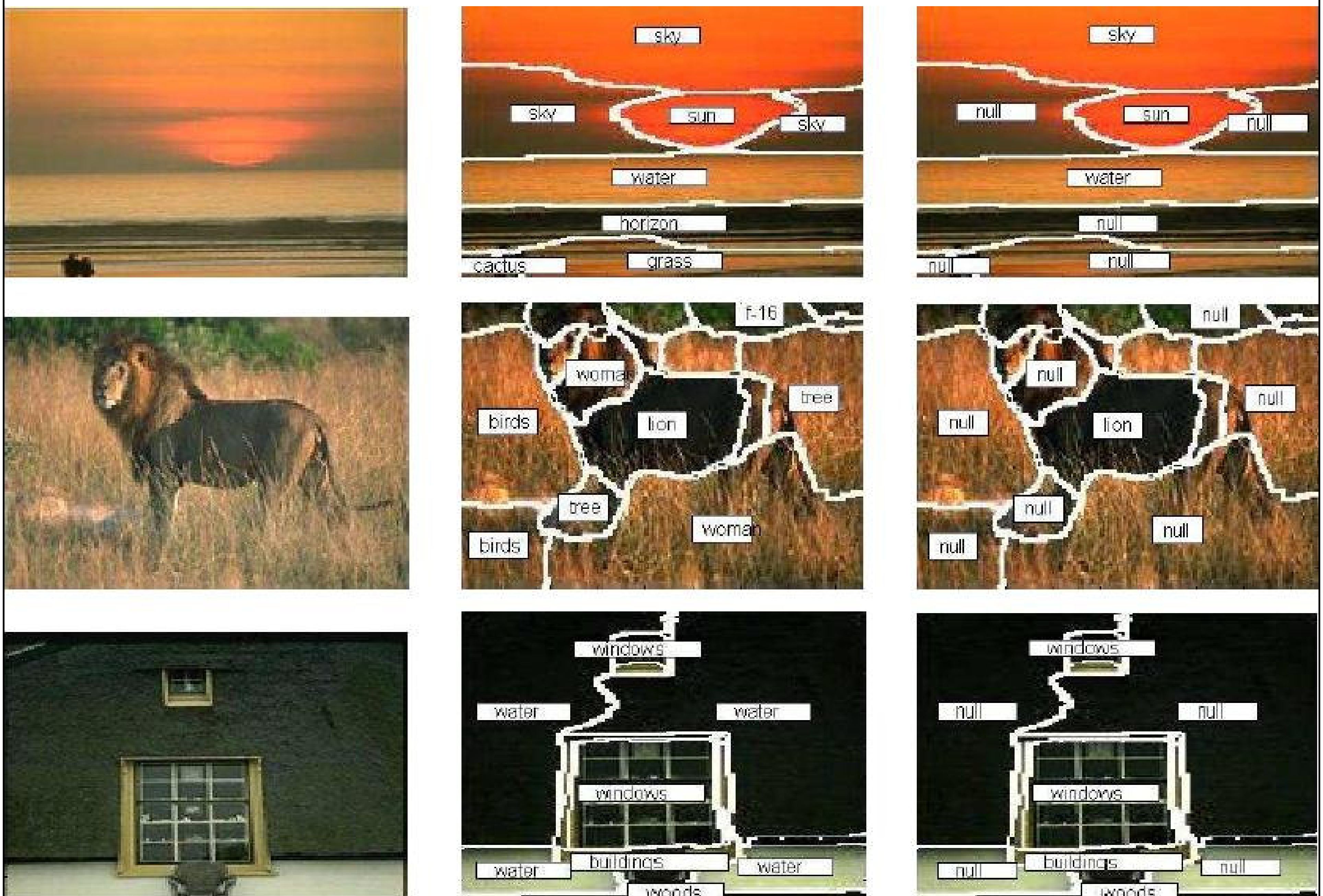


test

Refusing to predict

Null and fertility problems
simple solution to null - refusing to predict

If prob(word | blob) > threshold then
    predict the word
else
    assign NULL


sky
sky
sun
sky
water
horizon
cactus
grass
sky
null
sun
null
water
null
null
null
f-16
woman
birds
lion
tree
tree
birds
woman
null
null
null
lion
null
null
null
null
windows
water
water
windows
water
buildings
water
woods
windows
null
null
windows
null
buildings
null
woods

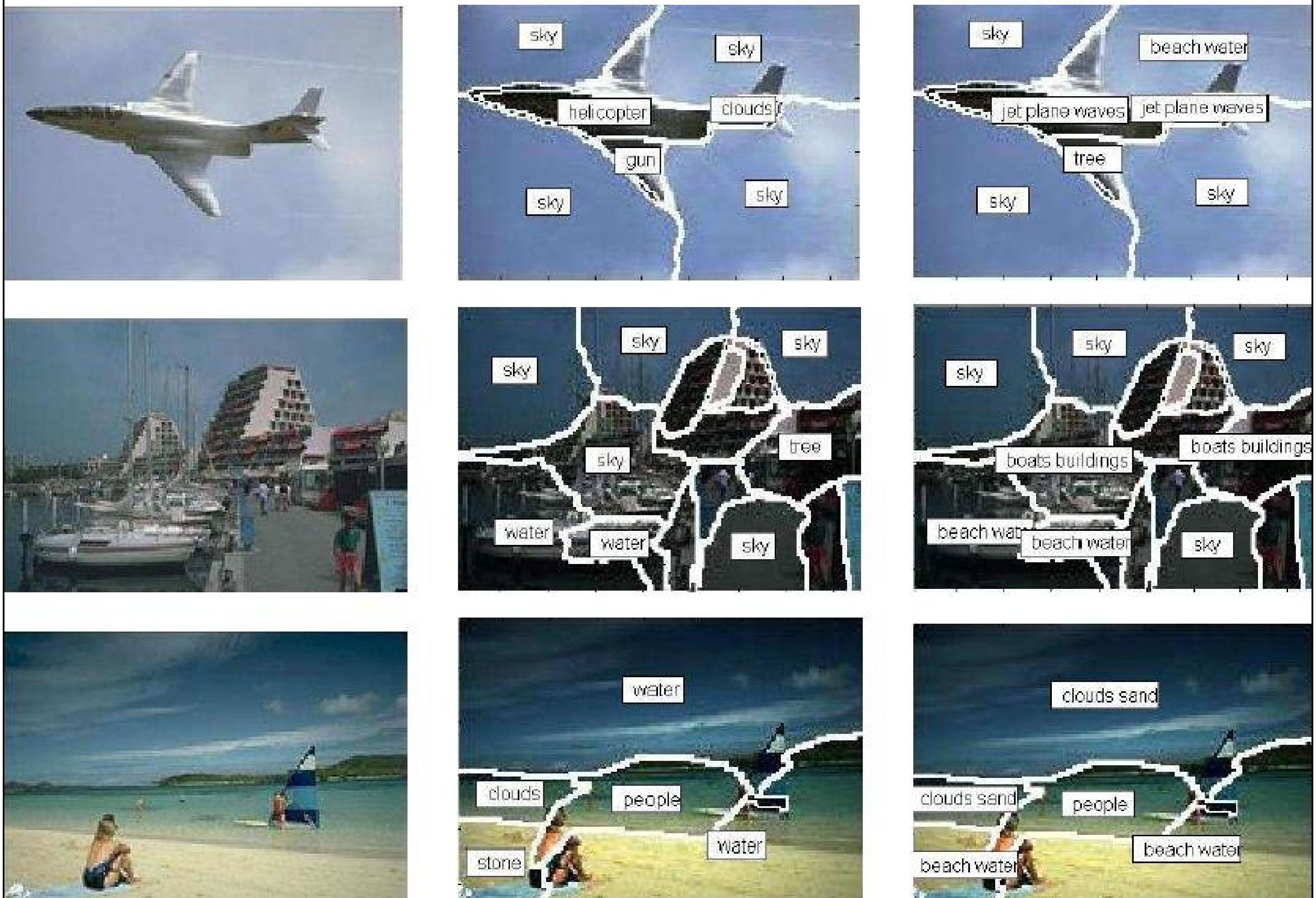
Merging Indistinguishable words

Some words cannot be set apart

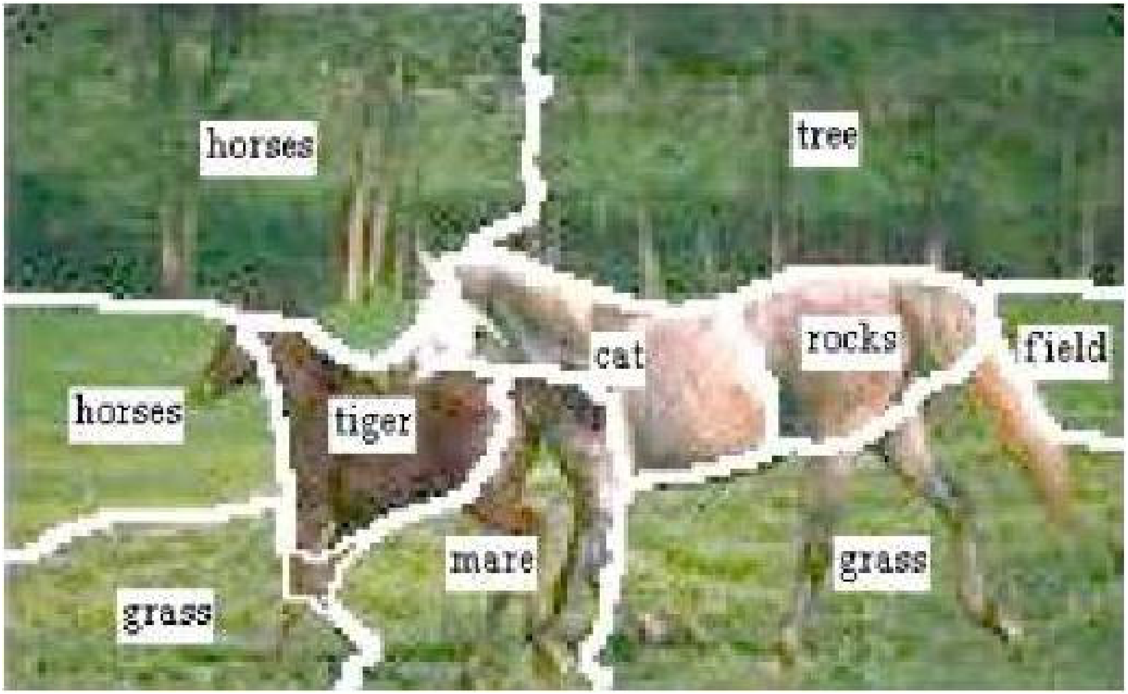
either they are synonyms
(e.g. locomotive and train)

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


sky
sky
helicopter
clouds
gun
sky
sky
sky
beach water
jet plane waves
jet plane waves
tree
sky
sky
sky
sky
sky
sky
tree
water
water
sky
sky
sky
sky
boats buildings
boats buildings
sky
water
clouds
people
water
stone
clouds sand
clouds sand
people
beach water

Integrating supervised data



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

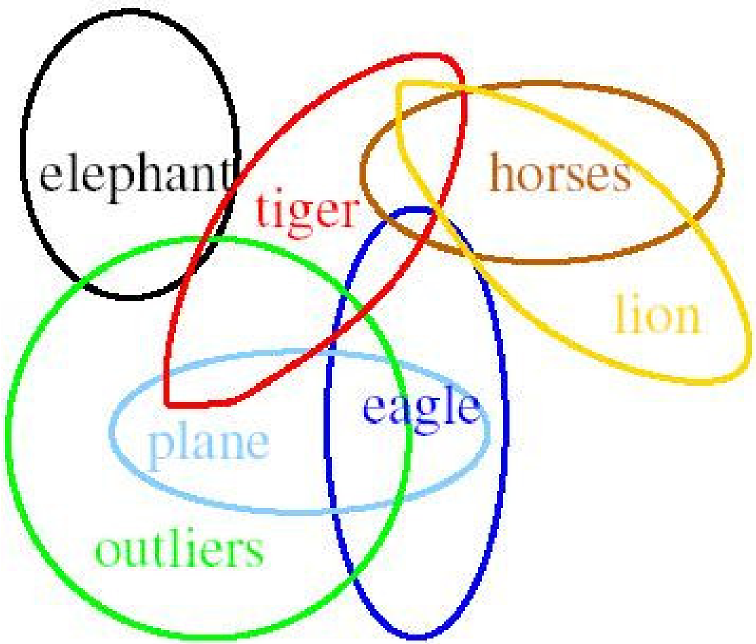
Integrating labeled data

A set of regions are labeled manually

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

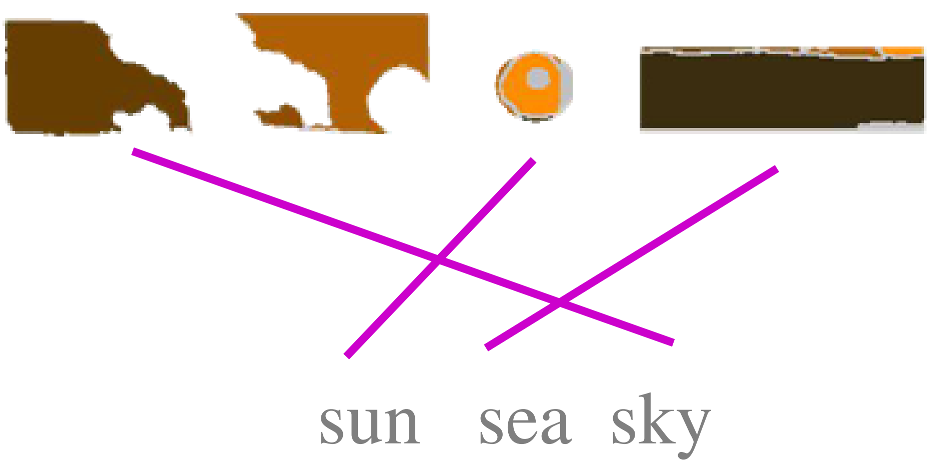
For improving clustering

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

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

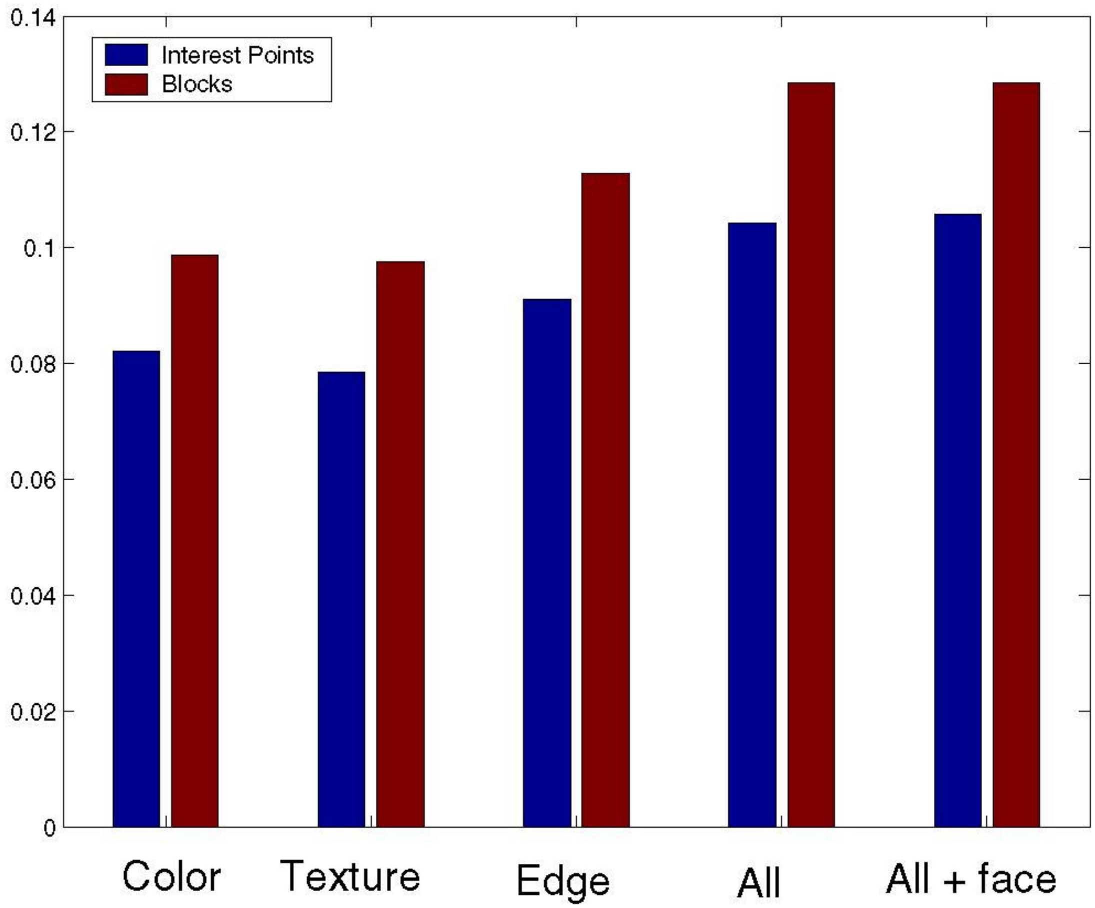
Features extracted

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

1000 visual tokens

138 concepts

Feature comparison



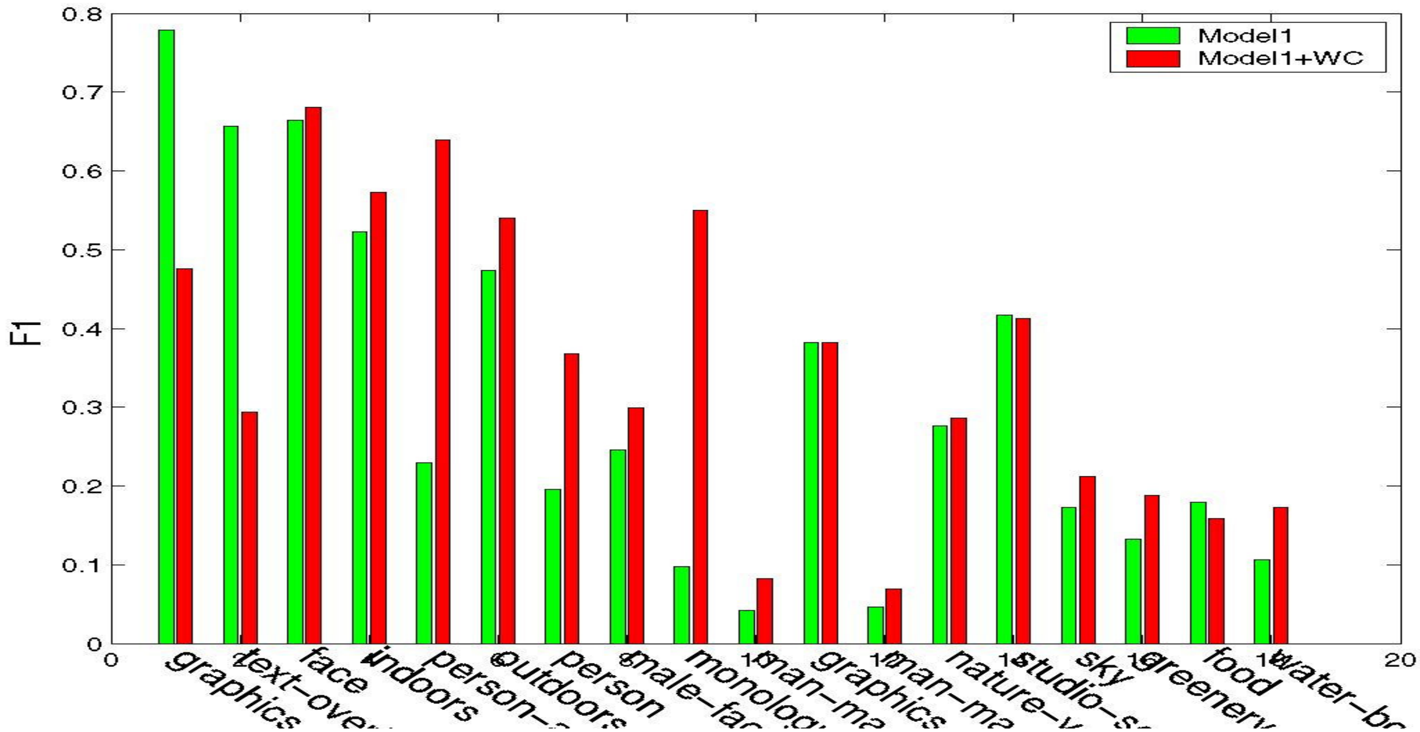
mAP values for different features

Integrating Language Modeling

Word co-occurrences as a simple language model

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

v: visual tokens, c: concepts



Training:
9413 shots,
Test :
4787 shots

Lab Moments

F1 = (recall + 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

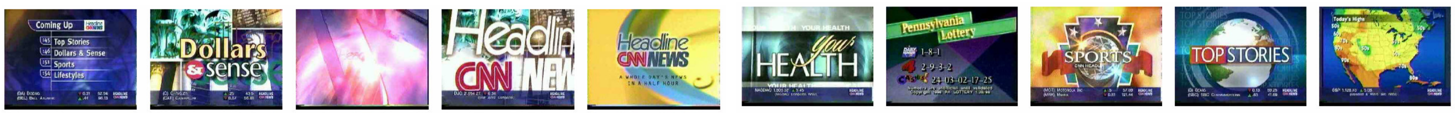
anchor *anchor – reporter dialogs* *logos* *overview*

News story *weather* *commercials* *sports*

Using delimiters to obtain story segments

Remove commercials

Remove graphics

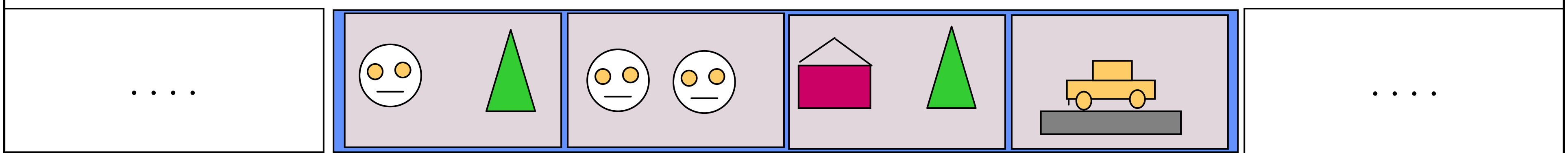


Remove anchor frames but use text

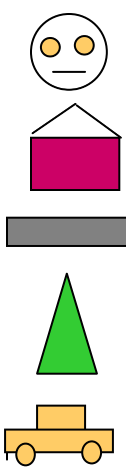
•Idea : A story segment starts or ends with a delimiter or with an anchor/reporter shot

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 / >=3)

building

road

outdoor

car

Semantic retrieval

!! only single occurrence per segment

Search on clinton

20 / 130 (15%)

27 / 133 (20%)

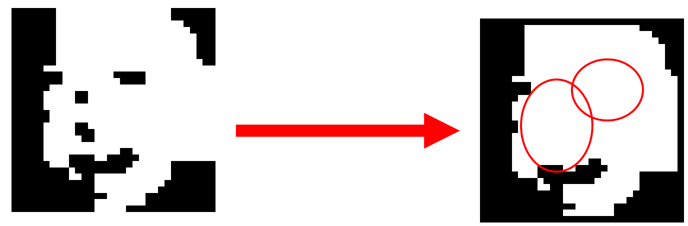
Search on fire

11 / 44 (25%)

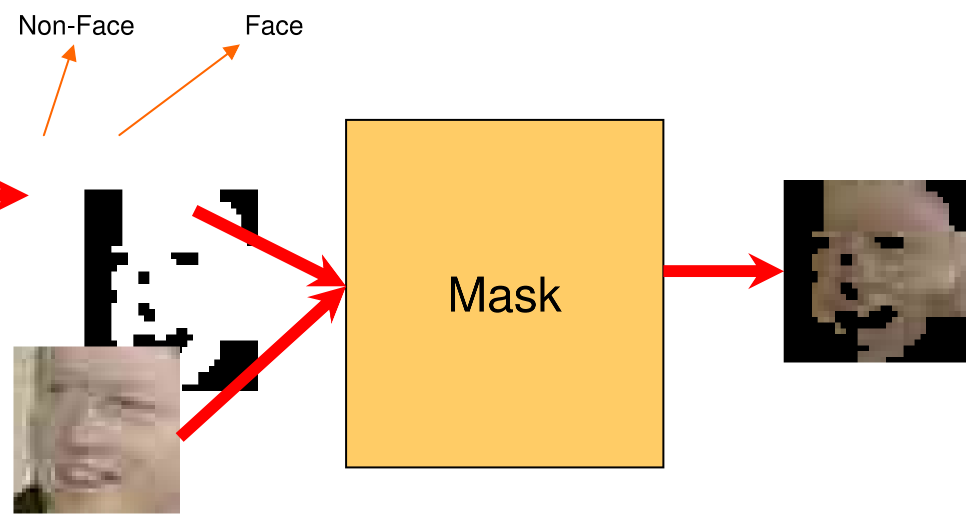
15 / 38 (40%)

Detecting Faces

Detect Faces



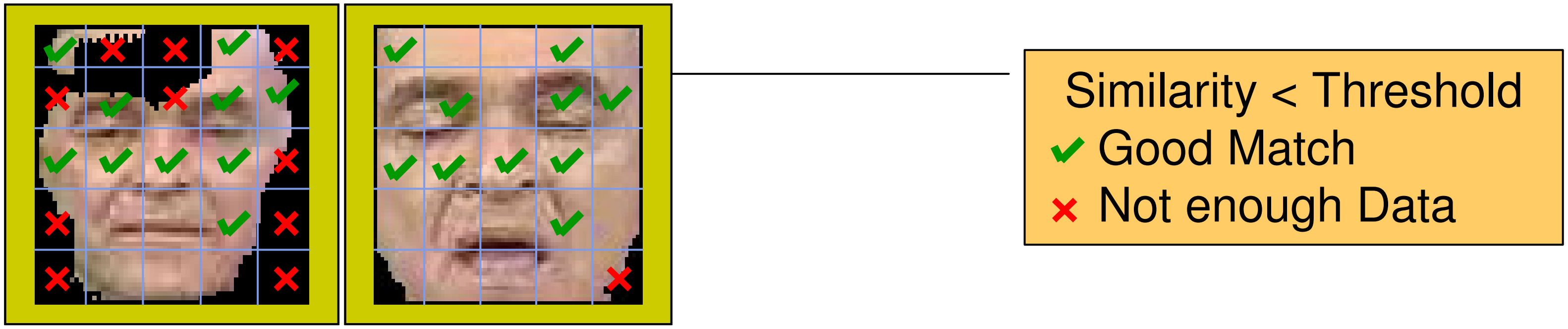
Improve amount of “face” by filling regions of the mask



Use face color Trained MLE to eliminate background

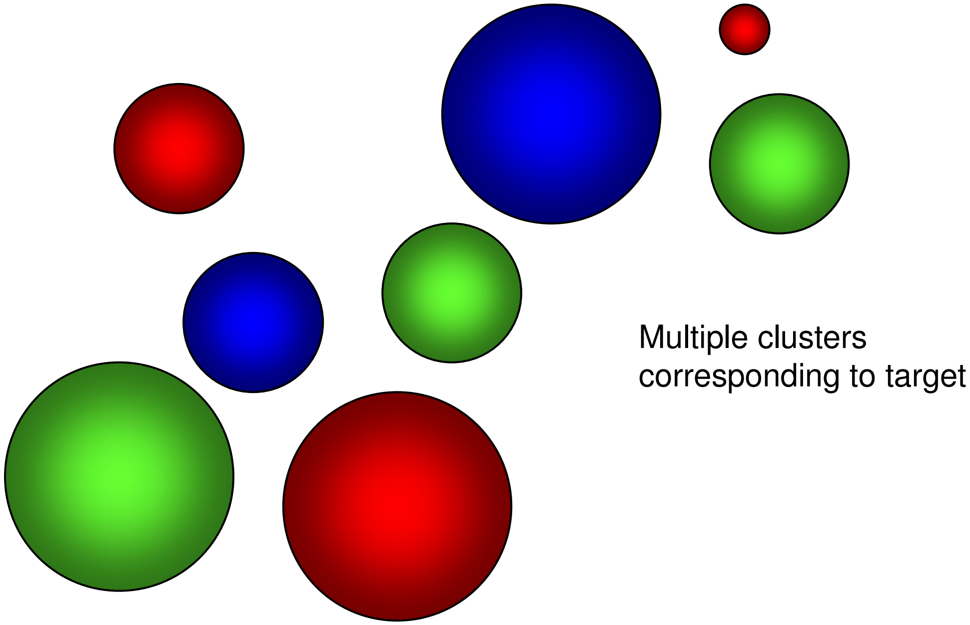
Mikolajcyzk, Schmid

Methodology



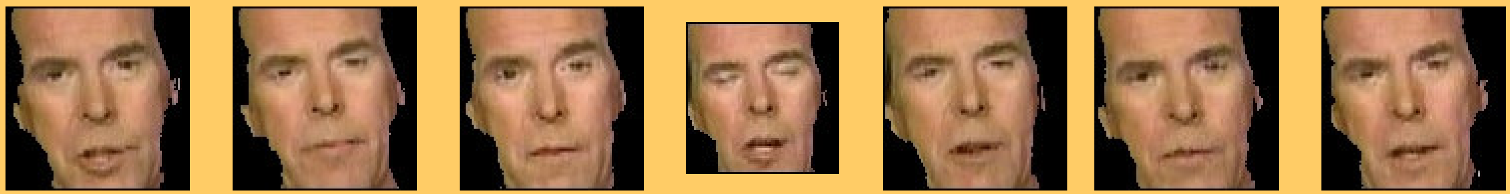
Join Similar faces together to obtain clusters of images.

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



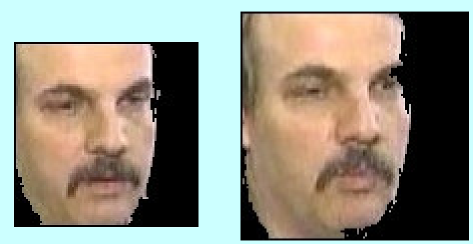
Results

Sam Donaldson

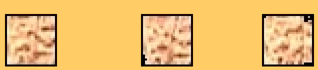


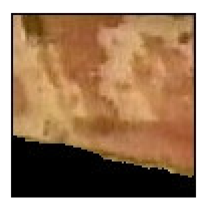
Anchor (ABC)

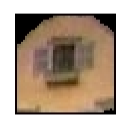
Unknown



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