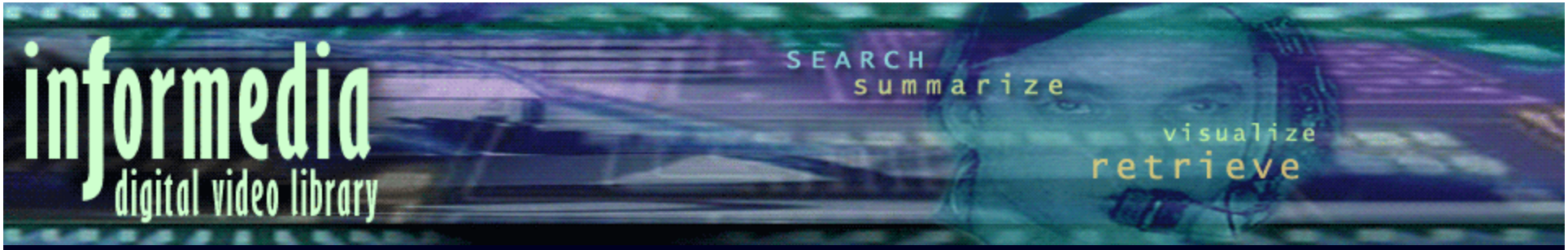
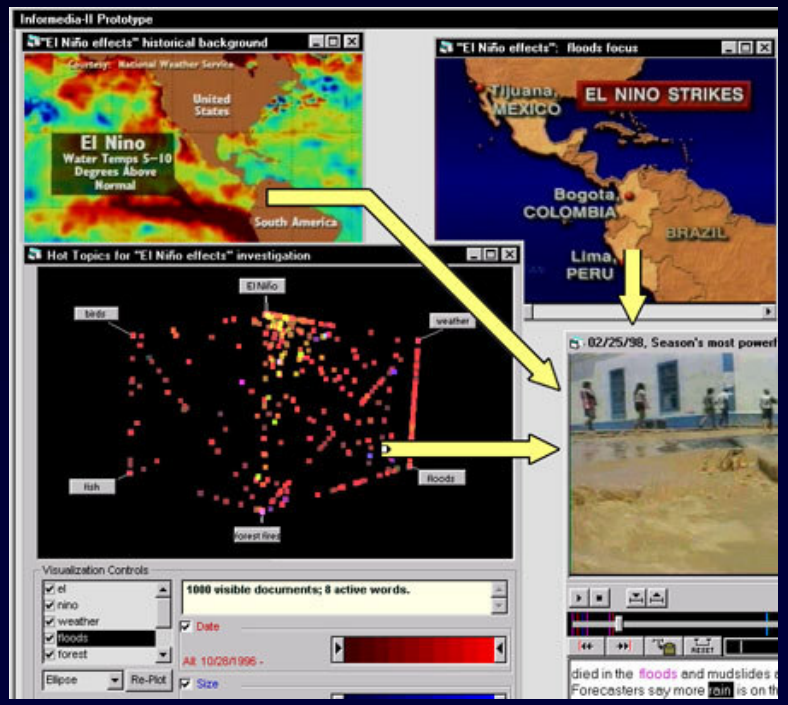


Associating video frames with text

Pinar Duygulu and Howard Wactlar

Informedia Project
School of Computer Science
Carnegie Mellon University

Informedia Digital Video Understanding Project



IDVL interface returned for "El Nino" query along with different multimedia abstractions from certain documents.

Informedia Digital Video Understanding Project

IDVL interface returned for "bin ladin" query

The results can be tuned using many classifiers

Query on “president” : association problem

Obviously, this is all awkward for the President. Whether he'd like to, or not, he can't come out an 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 an suspicions about the terrorist threat. President Bush: In terms of whether or not the FBI and the cia were communicating properly, I think it as 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 kches in 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 plaed – maid clear to Mueller, and deputies President execs and end to sniping at one senior official said tonight we don't need self-in flibting 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 Ichlt President certainly says he's the man in charge of the government. He's responsible but also has said and his senior AIDS 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 ner former bush's administration, he is not – now the add mirgs 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.

Approach

•Combine textual and visual features
to understand how to link semantics with appearance

•Model joint statistics of visual features and words using
a large collection of visual data with annotated text

Outline

- Correspondence problem between image regions and text using annotated image collections
- Correspondence problem between video frames and video
- Experiments on TREC 2001 data set
- Experiments on Chinese Cultural data set

Outline

- Correspondence problem between image regions and text using annotated image collections
- Correspondence problem between video frames and video
- Experiments on TREC 2001 data set
- Experiments on Chinese Cultural data set

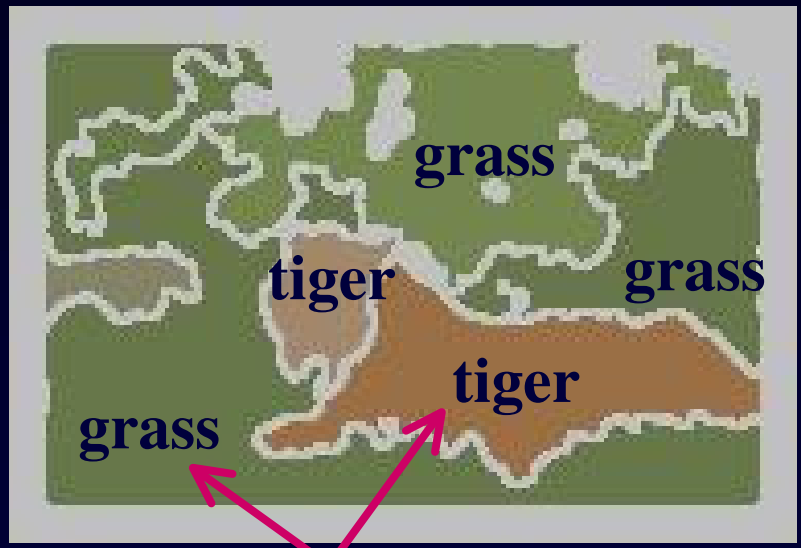
Associating image regions with words

tiger grass cat

?

tiger cat grass

Associating image regions with words

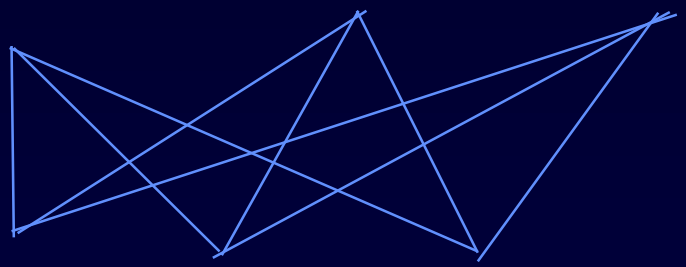


tiger cat grass

Statistical machine translation

Data: Aligned sentences, but word correspondences are unknown

“the beautiful sun”



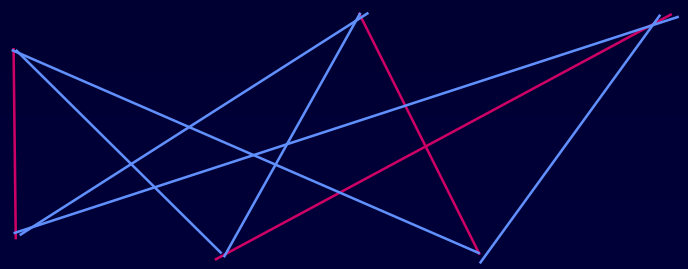
“le soleil beau”

Brown1993

Statistical machine translation

- Given the correspondences we can estimate p(sun|soleil)
- Given the probabilities we can estimate the correspondences

"the beautiful sun"

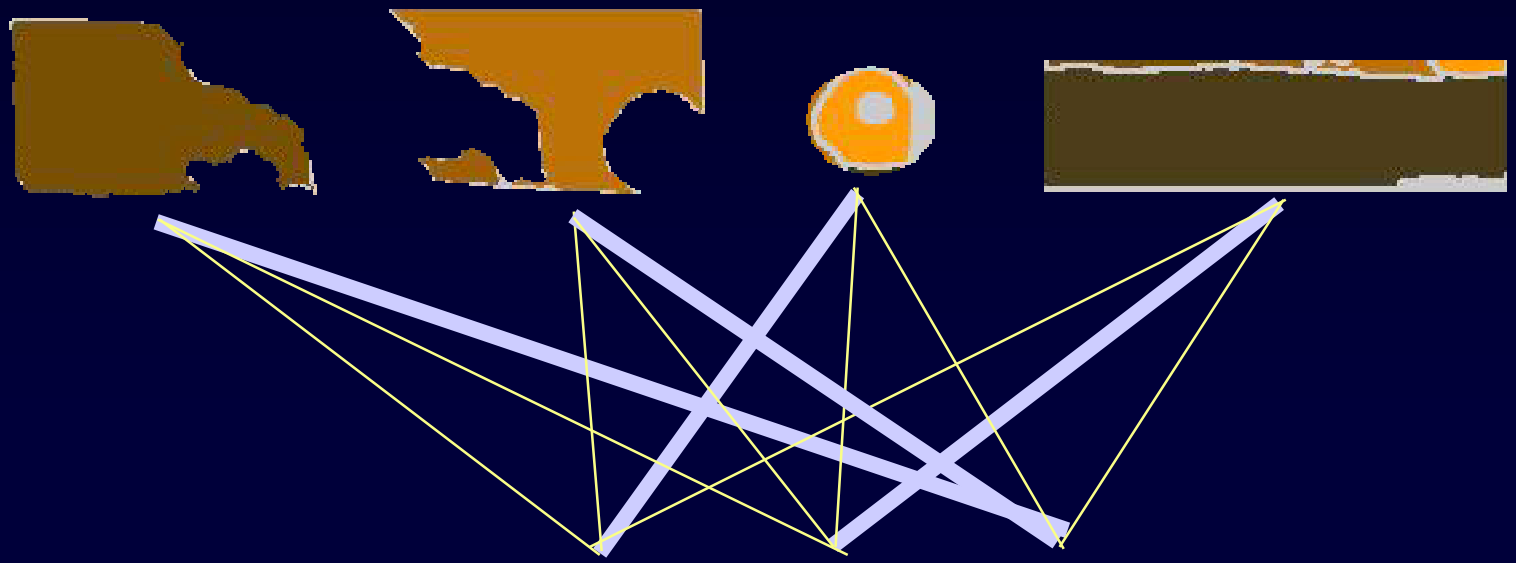


"le soleil beau"

Enough data + EM, we can obtain the translation p(sun|soleil)=1

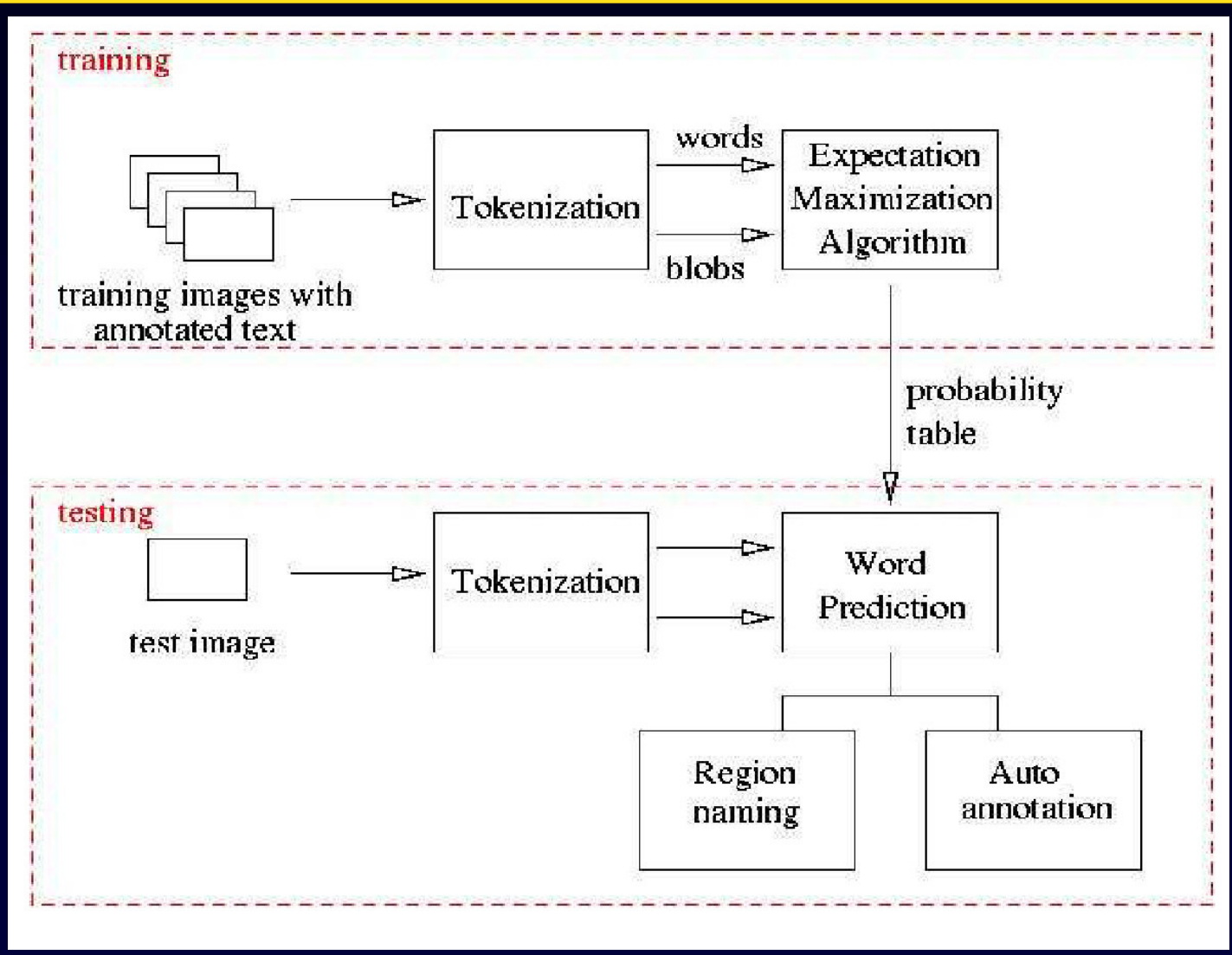
Multimedia translation

“sun sea sky”



“sun sea sky”

Overview


training
training images with annotated text
Tokenization
words
blobs
Expectation Maximization Algorithm
probability table
testing
test image
Tokenization
Word Prediction
Region naming
Auto annotation

Input Representation

sun sky waves sea

Image processing*

Each blob is a large vector of features

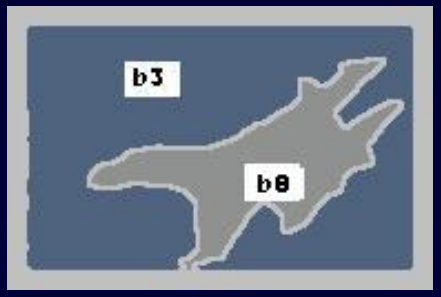
- Region size
- Position
- Colour
- Oriented energy (12 filters)
- Simple shape features

Tokenization

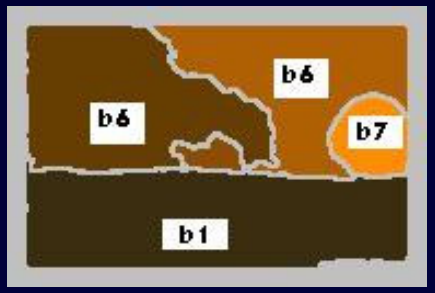
- Words → word tokens

- Image segments

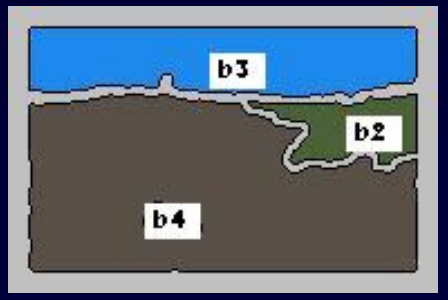
  •represented by 40 features
  (size, position, color, texture and shape)
  k-means to cluster features
  •best cluster for the blob → blob tokens



w3 w4 w5 w1

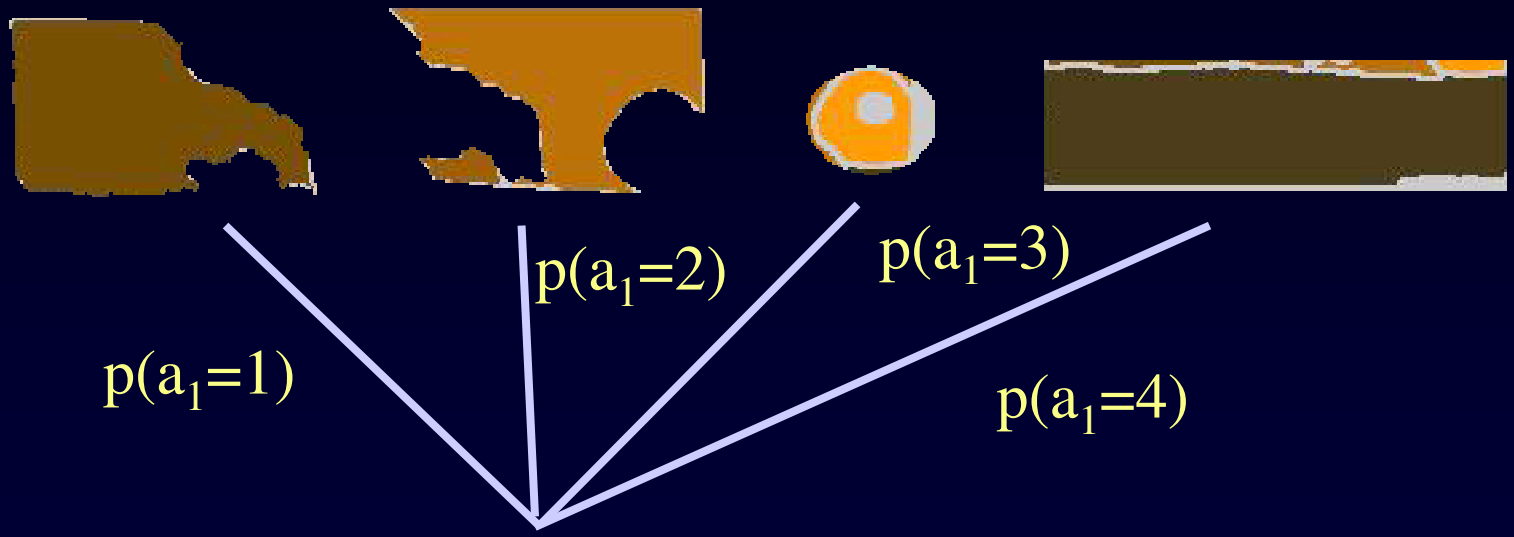


w6 w7 w8 w1



w12 w2 w1

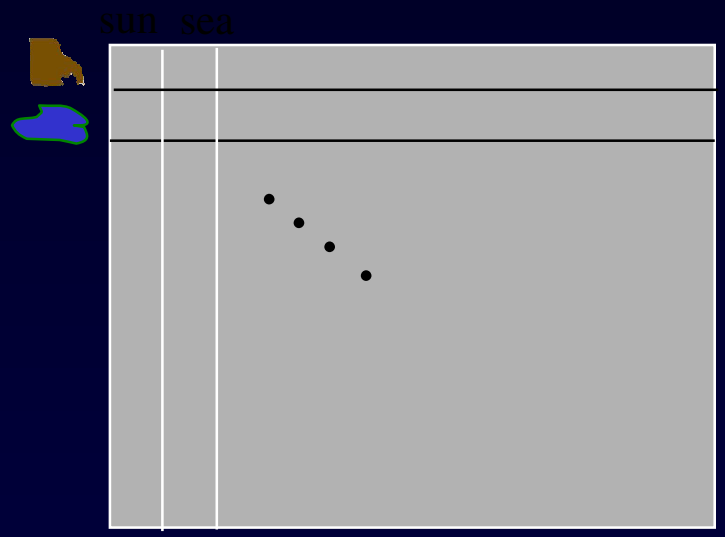
Associations



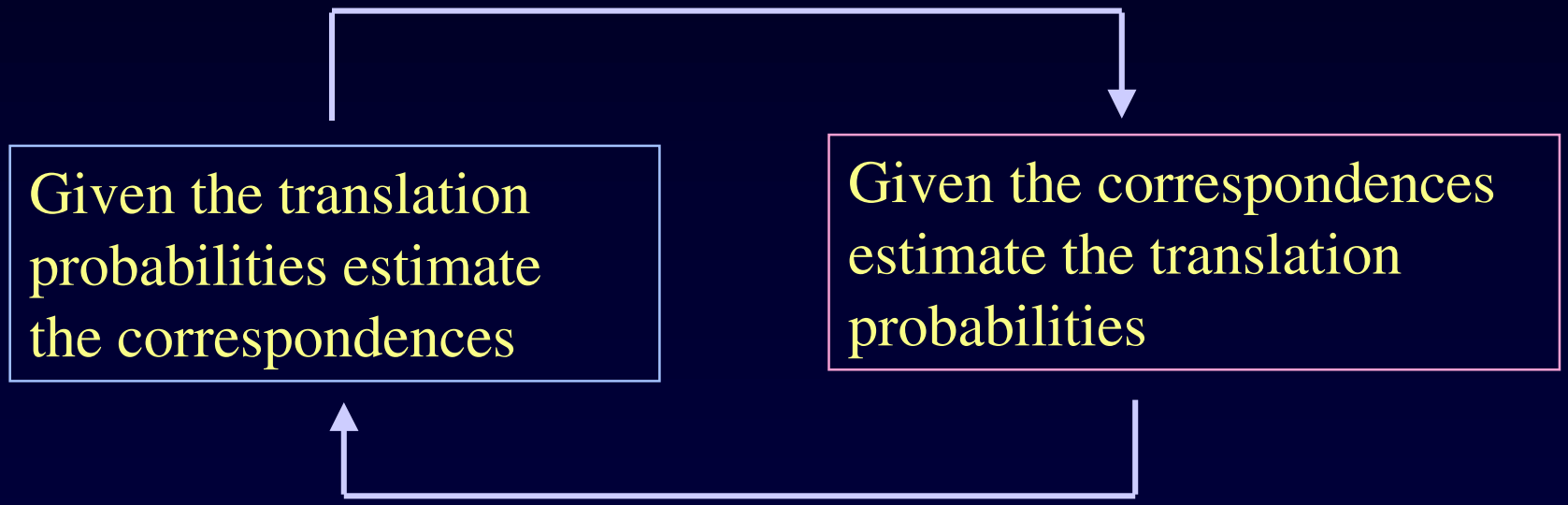
$$\sum_{i=1}^{B_n} p(a_1 = i) = 1$$

Initialization

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



Expectation Maximization Algorithm

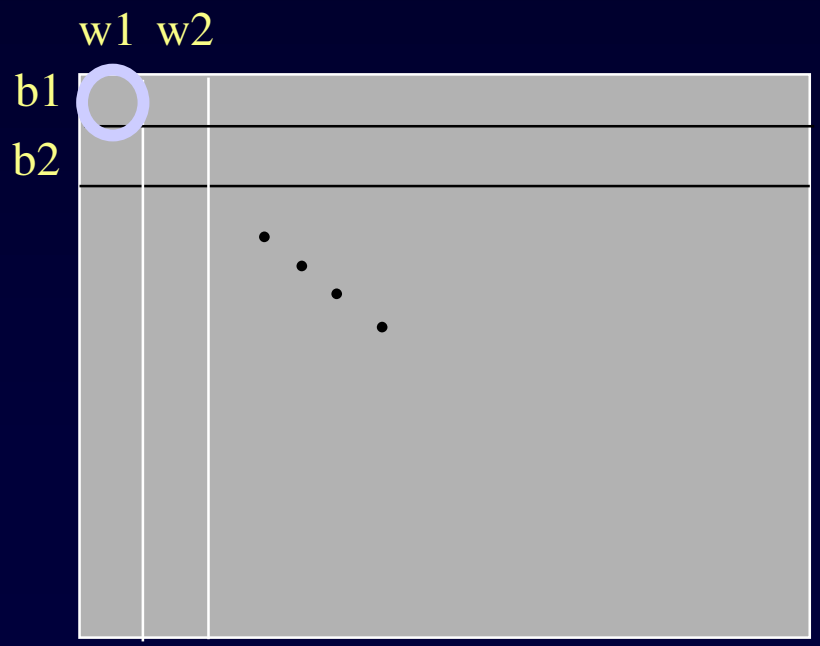


EM Algorithm

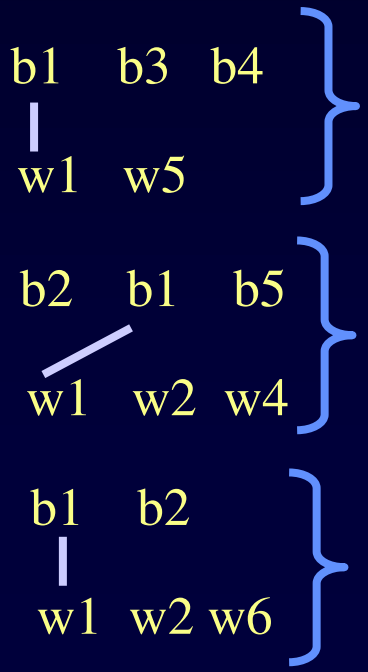
E step : Predicting correspondences from translation probabilities

(for one pair)

translation probabilities



correspondences

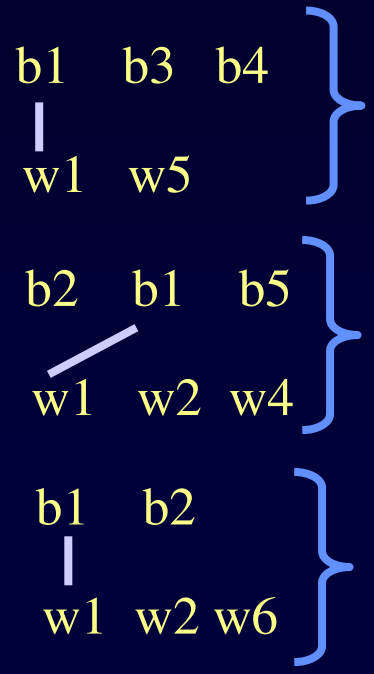


EM Algorithm

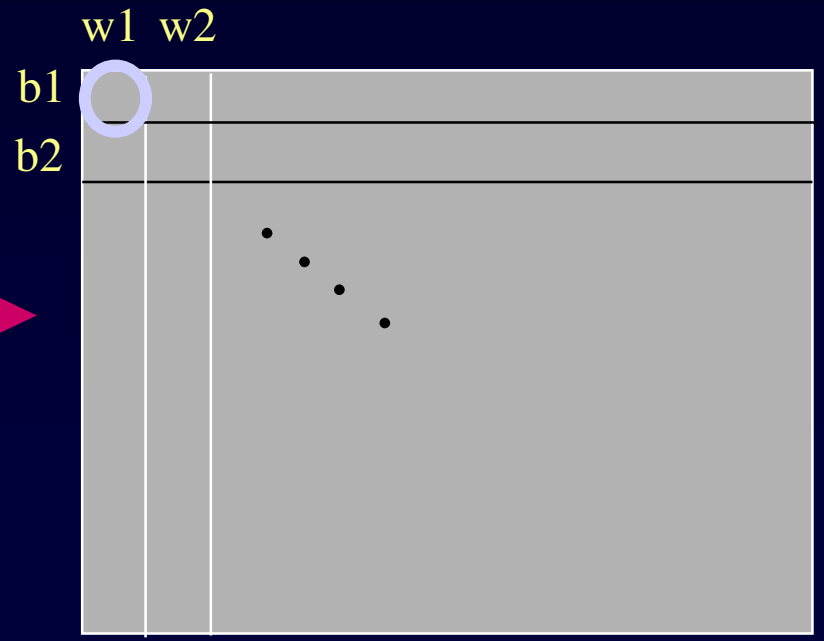
M step (for one pair)

Predicting translation probabilities from correspondences

correspondences



translation probabilities

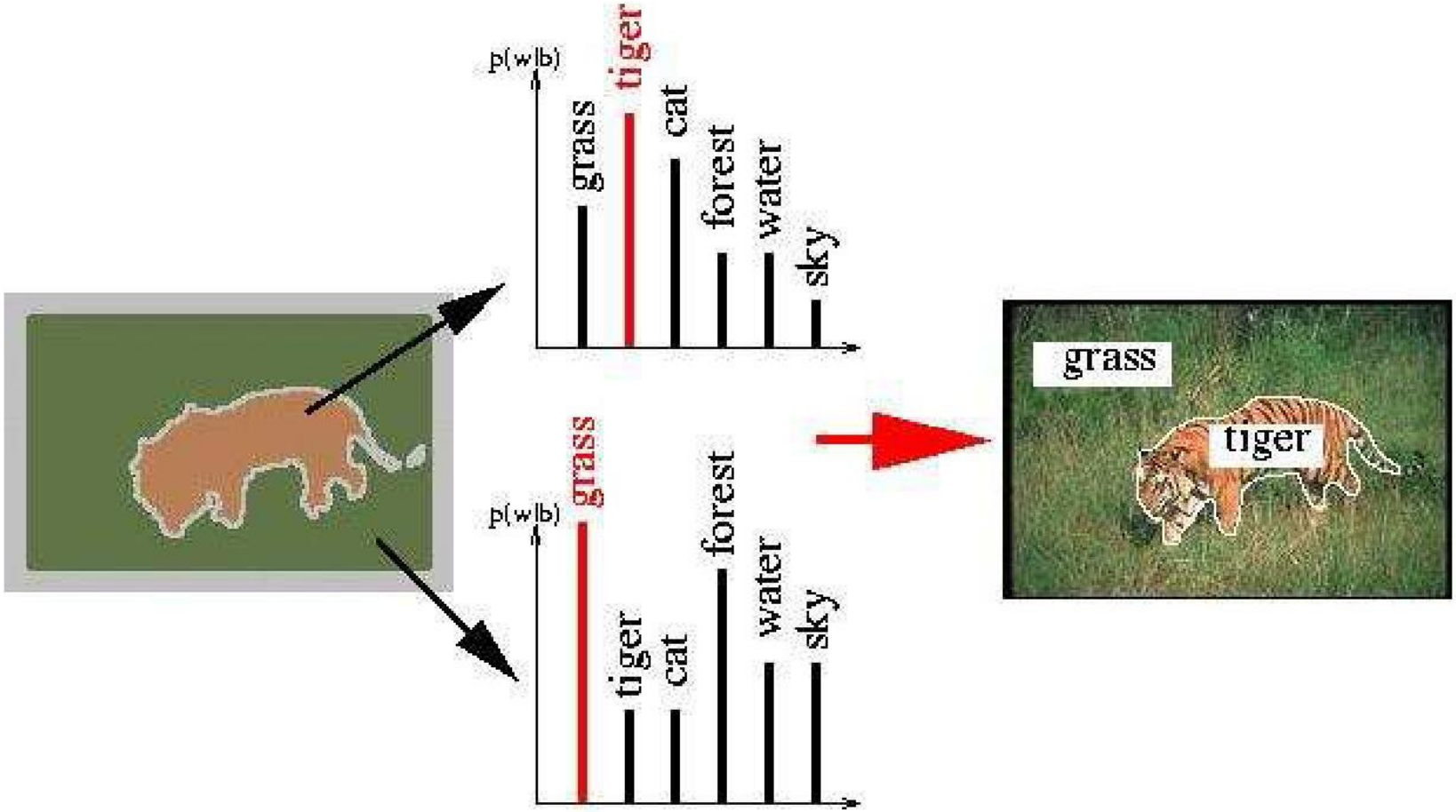


Word prediction

On a new test image
- segment the image
- extract the features from the regions
- then, for each region
  - find the corresponding blob token b
  - use word posterior probabilities p(w|b) for predicting words

Use predicted words for
- region naming
- auto-annotation

Region naming

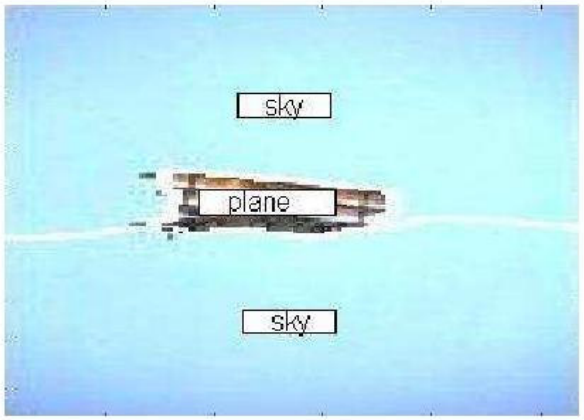


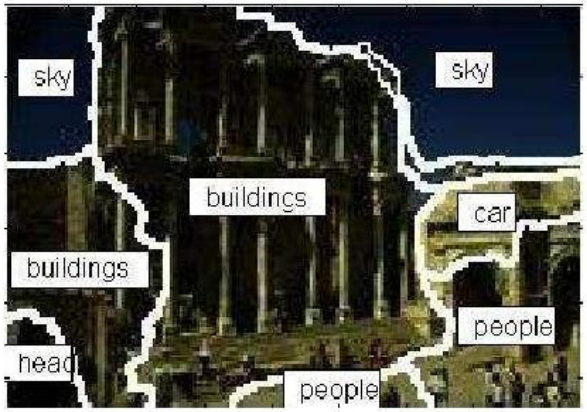
Results

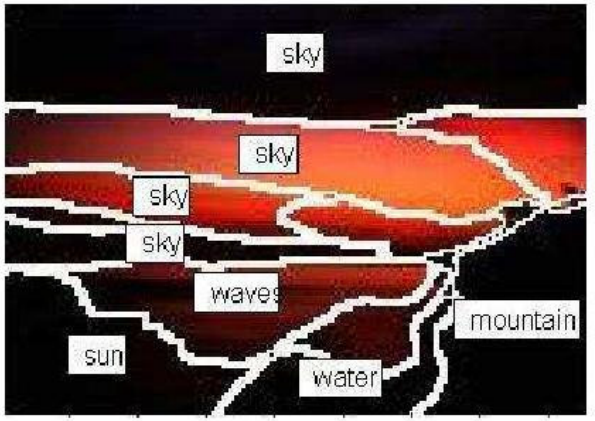
plane sky

people ruins stone

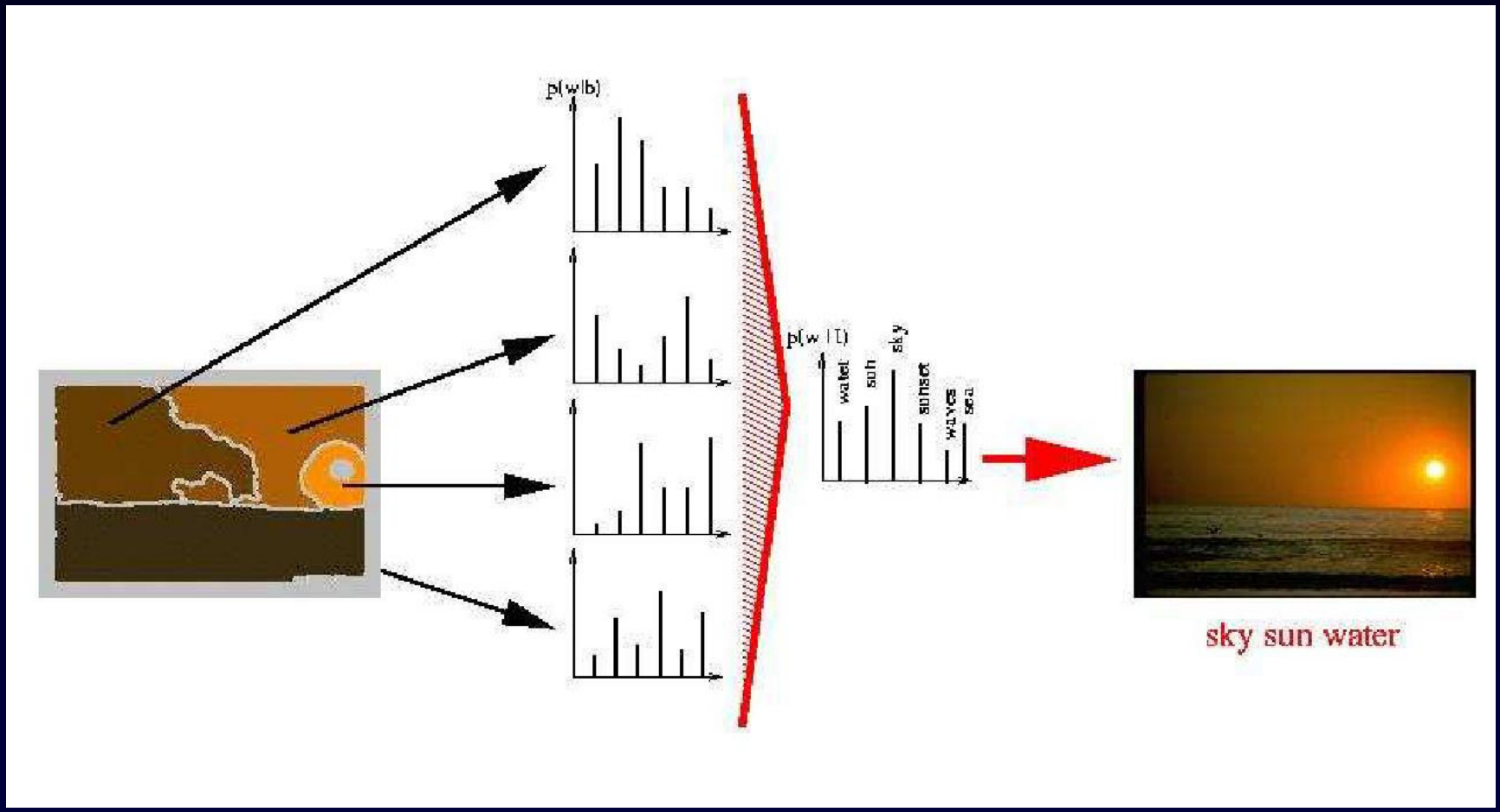
sunset tree water







Auto-annotation



Using annotation performance as a proxy

Actual Keywords → GRASS TIGER CAT FOREST

Predicted Words → CAT HORSE GRASS WATER

Using annotation performance as a proxy

Actual Keywords

GRASS TIGER CAT FOREST

Predicted Words

CAT HORSE GRASS WATER

Outline

- Correspondence problem between image regions and text using annotated image collections
- Correspondence problem between video frames and video
- Experiments on TREC 2001 data set
- Experiments on Chinese Cultural data set

Video

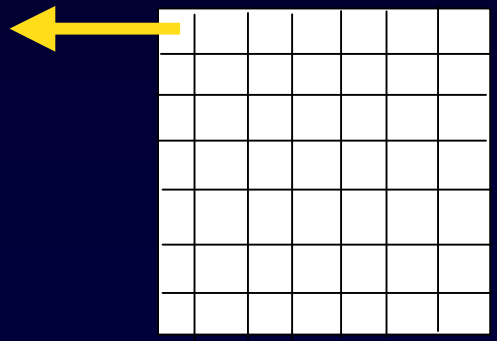
…despite heroic efforts many of the worlds wild creatures are doomed the loss of species is now the same as when the great dinosaurs become extinct will these creatures become the dinosaurs of our time today…

Input

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

Input

Position,
Color
(RGB and Lab, mean and std)
Texture
(Oriented energy filters, DoG)



Input

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

Brill's tagger is used to extract nouns

The text only corresponding to the shot can be used
Or also the surrounding text within a window size can be used

Outline

- Correspondence problem between image regions and text using annotated image collections
- Correspondence problem between video frames and video
- Experiments on TREC 2001 data set
- Experiments on Chinese Cultural data set

TREC 2001 data

Number of images = 2232

7 x 7 blocks
56 features extracted from each block
Number of blob tokens = 500

Number of words = 1938

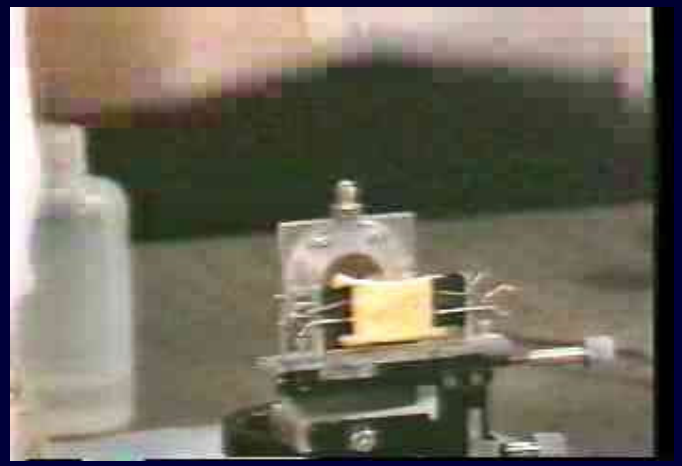
Window size for surrounding text = 5

Auto-annotations

statue(1) liberty(2)

plane(2)

space (1), telescope(10)



robot(5)

Auto-annotations

space (6), astronaut(7)

plane(2)

water(1) research(3)

space(1), world(6)

Query for “Statue of Liberty”

Associating frames and text

statue(1) liberty(2)

statue(1) liberty(3)

statue(1) liberty(2)

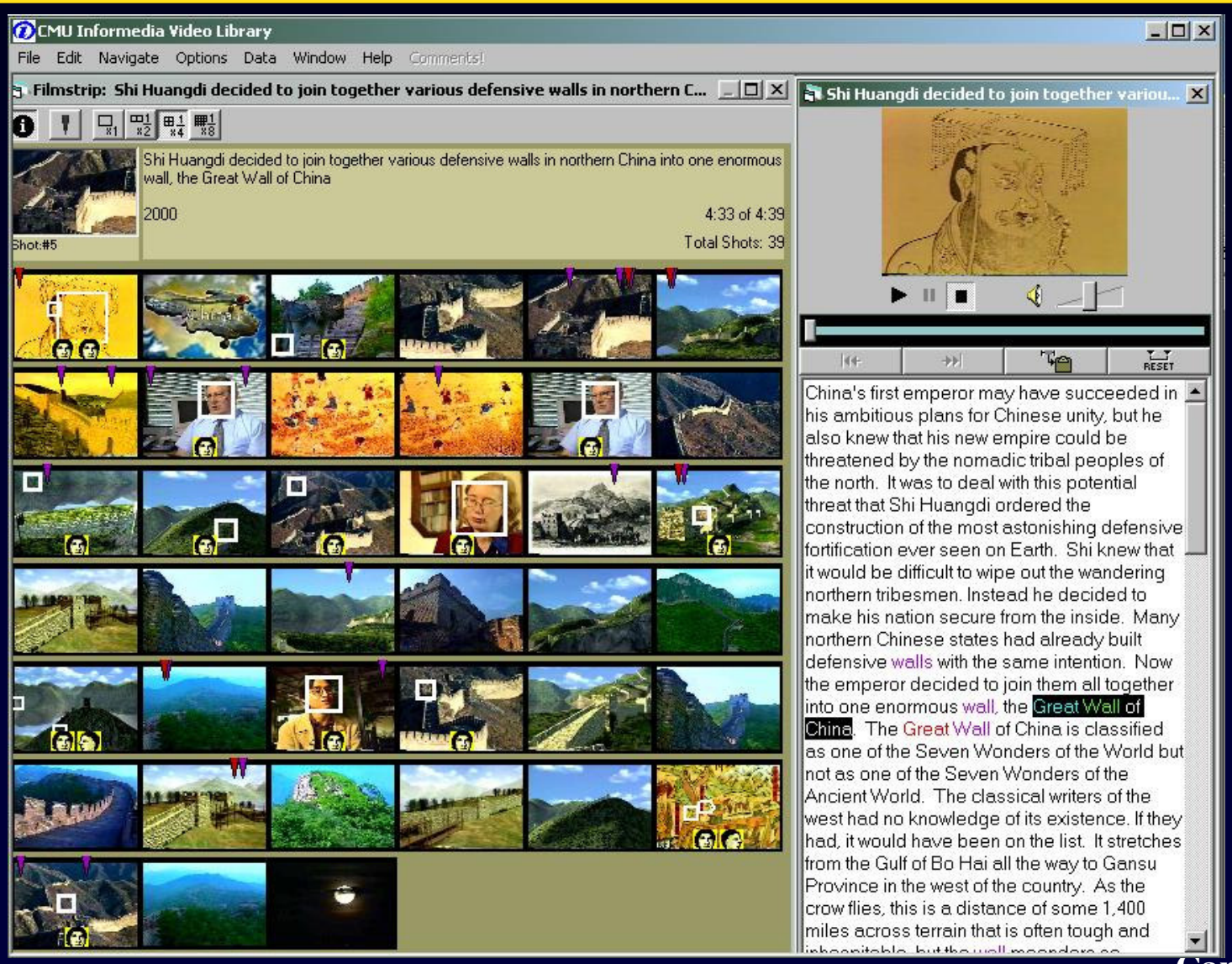
statue(1) liberty(3)

Outline

- Correspondence problem between image regions and text using annotated image collections
- Correspondence problem between video frames and video
- Experiments on TREC 2001 data set
- Experiments on Chinese Cultural data set

Query on “panda”

Query on “Great Wall”



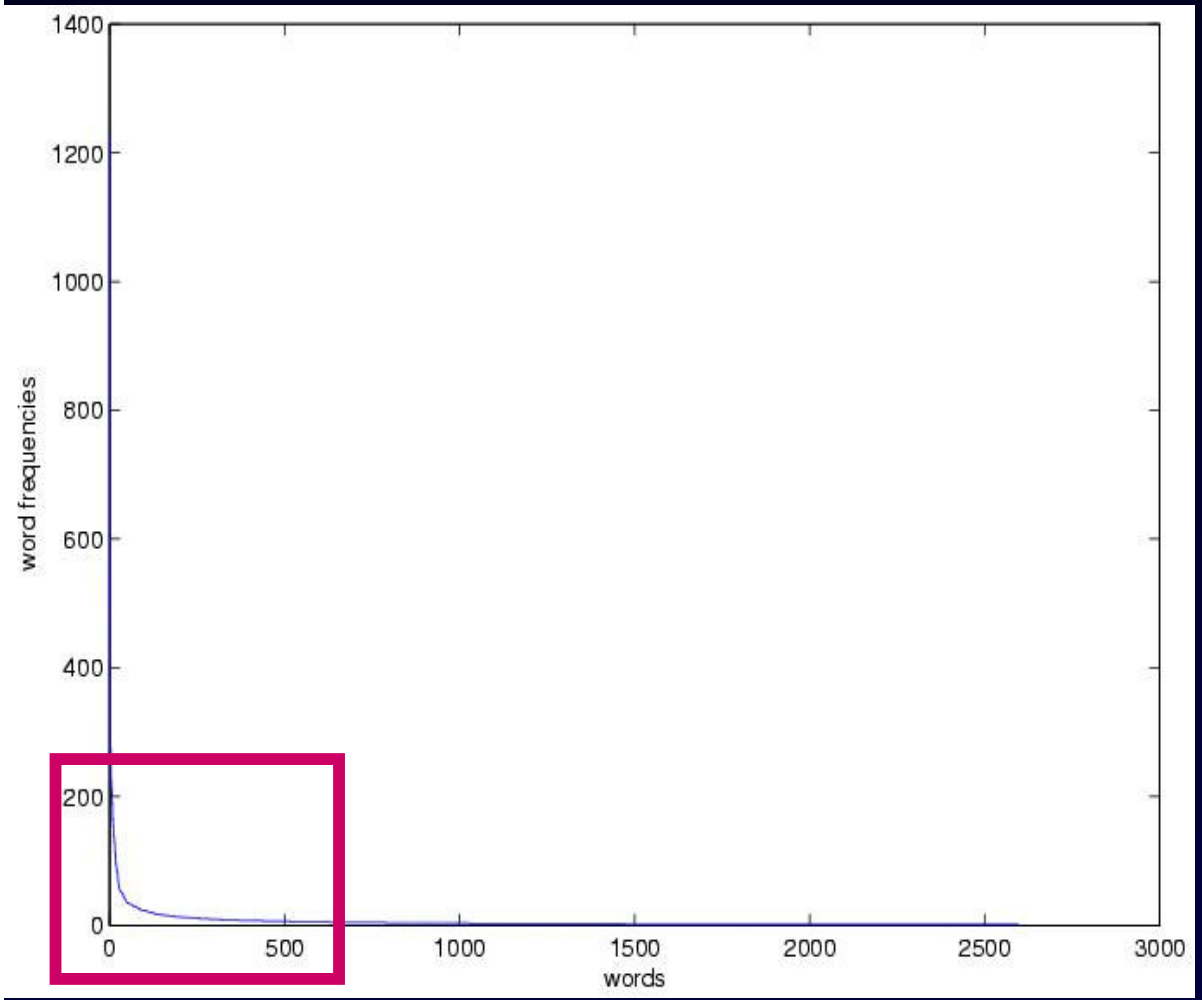
Chinese Culture data set

Number of images = 3745

5 x 5 blocks
56 features extracted from each block
Number of blob tokens = 1000

Number of words = 2597

Pruning data

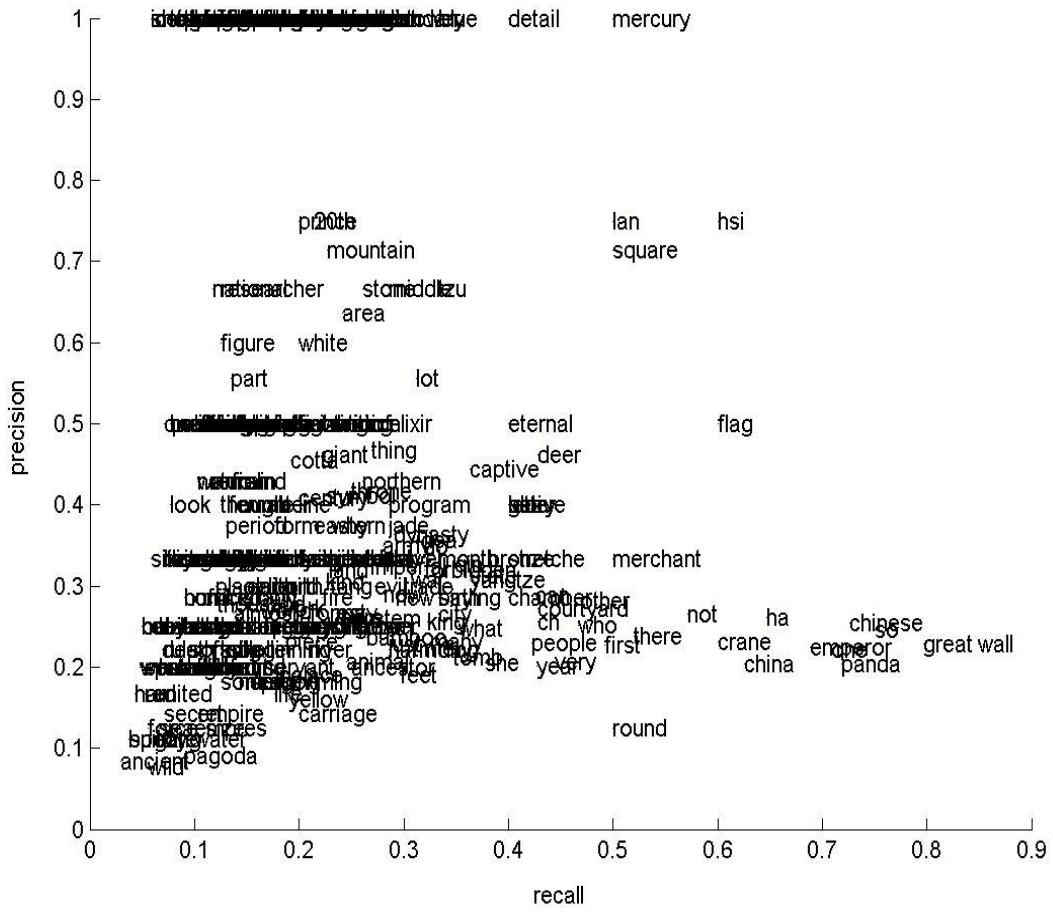


Eliminate the words
if frequency <= 5 or
frequency > 250

Number of words = 626

Number of images = 2785

Recall and Precision



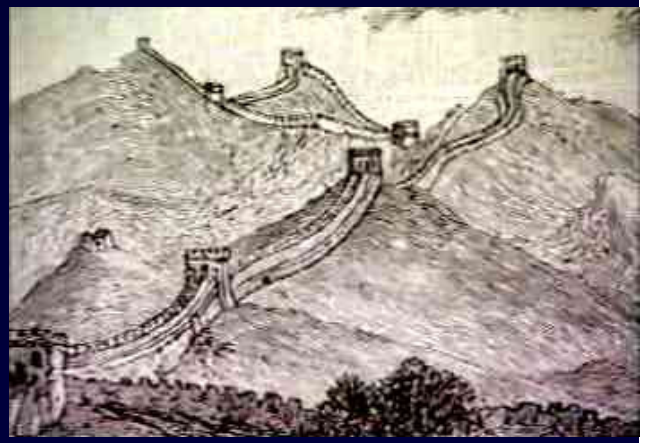
When only the words that are associated with the frame are taken

Recall :
number of correct predictions / number of actual occurrence

Precision :
number of correct predictions / number of all predictions

Panda (when predicted in the first 3)

Wall (when predicted in the first 3)



Emperor (when predicted in the first 3)



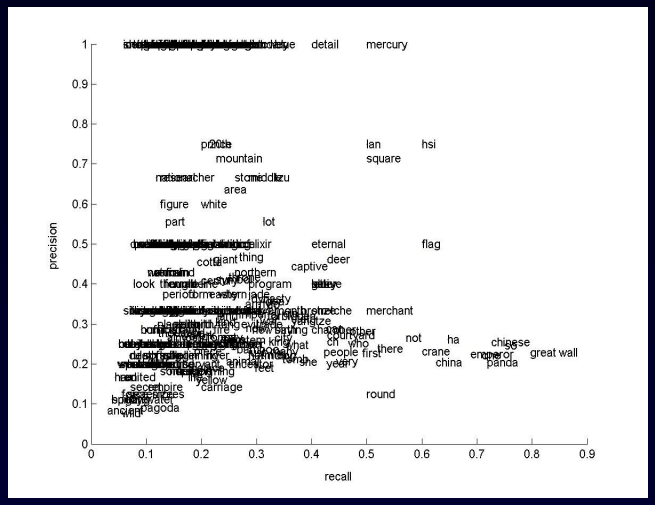



Correspondence results for “panda”

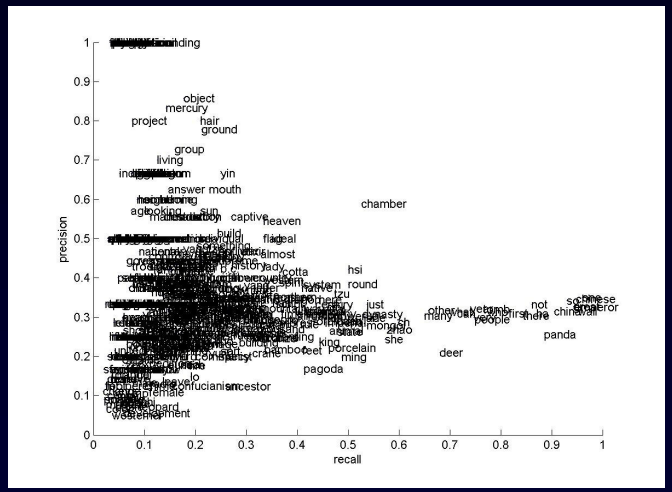
number of annotations	127
number of correct annotations	42
number of incorrect annotations	85
number of predictions	121
number of correct predictions	51
number of incorrect predictions	70
annotation present not predicted	64
annotation correct not predicted	11
annotation incorrect not predicted	53
not annotated but predicted	61
not annotated but correctly predicted	25
not annotated and incorrectly predicted	36
annotated and predicted and correct	26
annotated and predicted but incorrect	31

Correspondence results for “panda”

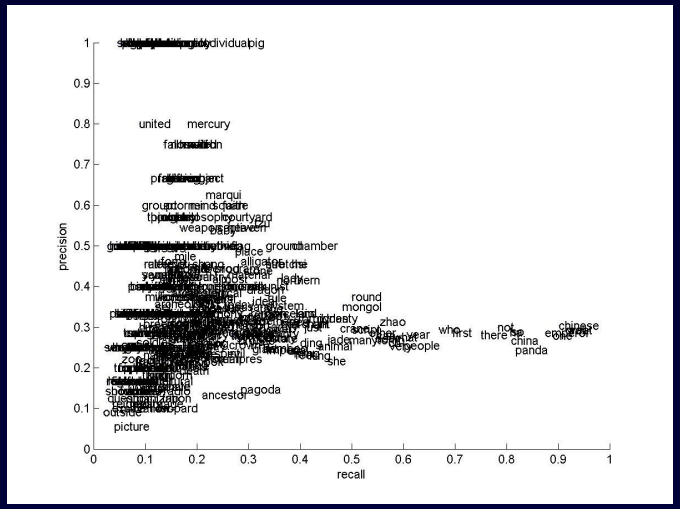
Recall and precision as a function of window size



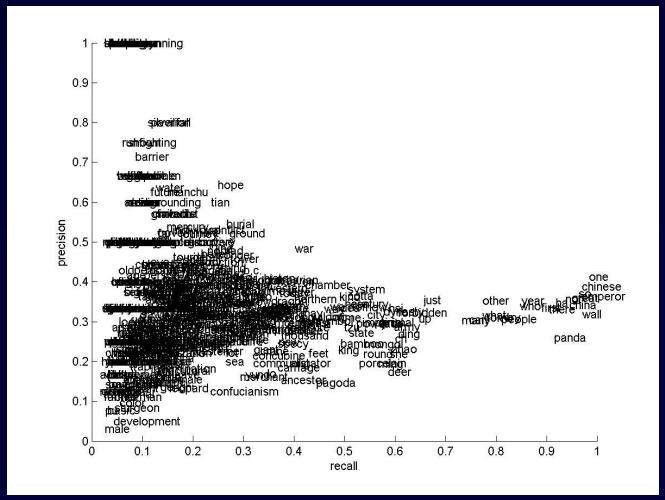
Single frame



Window size = 2



Window size = 1



Window size = 3

Recall and precision for some selected words as a function of window size

	panda	wall	emperor
single frame	0.7183 - 0.2044	0.8492 - 0.2281	0.6885 - 0.2242
wsize = 1	0.8182 - 0.2432	0.9180 - 0.2923	0.8744 - 0.2860
wsize = 2	0.8857 - 0.2560	0.9517 - 0.3149	0.9446 - 0.3257
wsize = 3	0.9154 - 0.2604	0.9720 - 0.3186	0.9693 - 0.3631

Conclusions

While text and images are separately ambiguous, jointly they tend not to be.

Linking visual information with text improves performance

The proposed method
- Can learn correspondence between visual data and text
- Unsupervised – uses the available large data sets efficiently

Can be used
- For region naming – object recognition on the large scale
- Auto-annotation – predicting words
- Associating frames and text

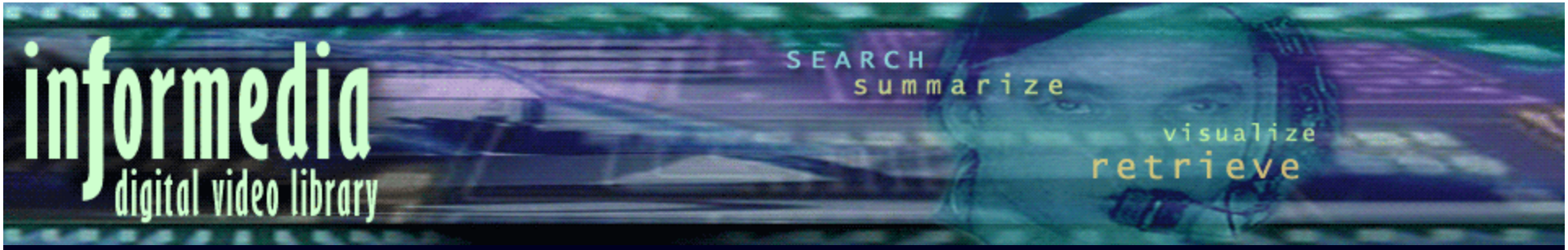
Future Directions

Applying on broadcast news

A better set of features including face detectors, and moving objects

Better linguistic analysis (e.g. taking noun phrases)

Finding the best window size by statistical analysis



Thank you

http://www.informedia.cs.cmu.edu/

Prediction measure as a function of window size

	PR
single frame	0.1851
wsize = 1	0.2469
wsize = 2	0.2783
wsize = 3	0.2975

Prediction measure (PR):

$$1/N \sum_{N} (\#\text{correct} / \#\text{ actual})$$