Cross-modal disambiguation

Kobus Barnard

The University of Arizona

In collaboration with

Arnon Amir (IBM), Pinar Duygulu, Alon Efrat (UA), Sandiway Fong (UA),

David Forsyth (UIUC), Matt Johnson (Nokia), Ranjini Swaminathan (UA),

Keiji Yanai (U. Electro-Communications,Tokyo),

Overview

- Ambiguity challenges for multimedia applications
- Context for disambiguation
- Cross-modal disambiguation
- Multi-modal modeling
- Cross-modal disambiguation examples
  - Correspondence ambiguity in image labels
  - Parts of speech for disambiguating annotation reference
  - Word sense disambiguation with pictures
  - Electronic slides for improving speech transcripts

Ambiguity is everywhere

- Words
  - river bank versus money bank
- Images pieces
  -  water? sky? blue shirt?
- Annotated images
  - Which part of the image does a keyword refer to?
- Spoken words
  - “she a” versus “chia”

Reducing ambiguity using context

- Vision
  - Isolated pixels or regions do not mean much

Reducing ambiguity using context



Reducing ambiguity using context



Reducing ambiguity using context

- Vision
  - Image interpretation relies on context
  - Much recent work on spatial context models
    - (See for example, Torralba, Belongie, Zemel, Heitz)

Reducing ambiguity using context

- Language
  - The surrounding words tell you which “sense” of the word is meant

    He face-planted into the snow bank

  - Much effort has been expended on automated “word sense disambiguation”

Cross modal disambiguation

- Exploit context from multiple modalities

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



Cross-modal analysis

- Instead of
  - p(X | o, C)

(e.g., o is the observation "bank", C is text context)

- Consider
  - p(X | o, C_1, C_2, …)

(C1, C2 are co-occurring contexts of different modes)

Cross-modal analysis requirements

- Some mutual information between cues and query
- Some independence between the the cues

Cross-modal disambiguation examples

- Region-word correspondence
  - X = keyword, o = region, C1 = NULL, C2 = keyword list
- Word sense disambiguation with pictures
  - X = sense, o = word, C1 = text, C2 = image
- Geotags helping image understanding
  - X = object, o = image, C1 = Geotag

Multi-modal modeling for automated image annotation *

* Barnard & Forsyth, ICCV 2001; Duygulu, Barnard, de Freitas, and Forsyth, ECCV 2002; Barnard et al., JMLR 2003.

Training data



TIGER CAT WATER GRASS



1090
SUN CLOUDS
WATER SKY



143078
MOUNTAINS TREES
aspens VALLEY

Automated image annotation



tiger grass cat

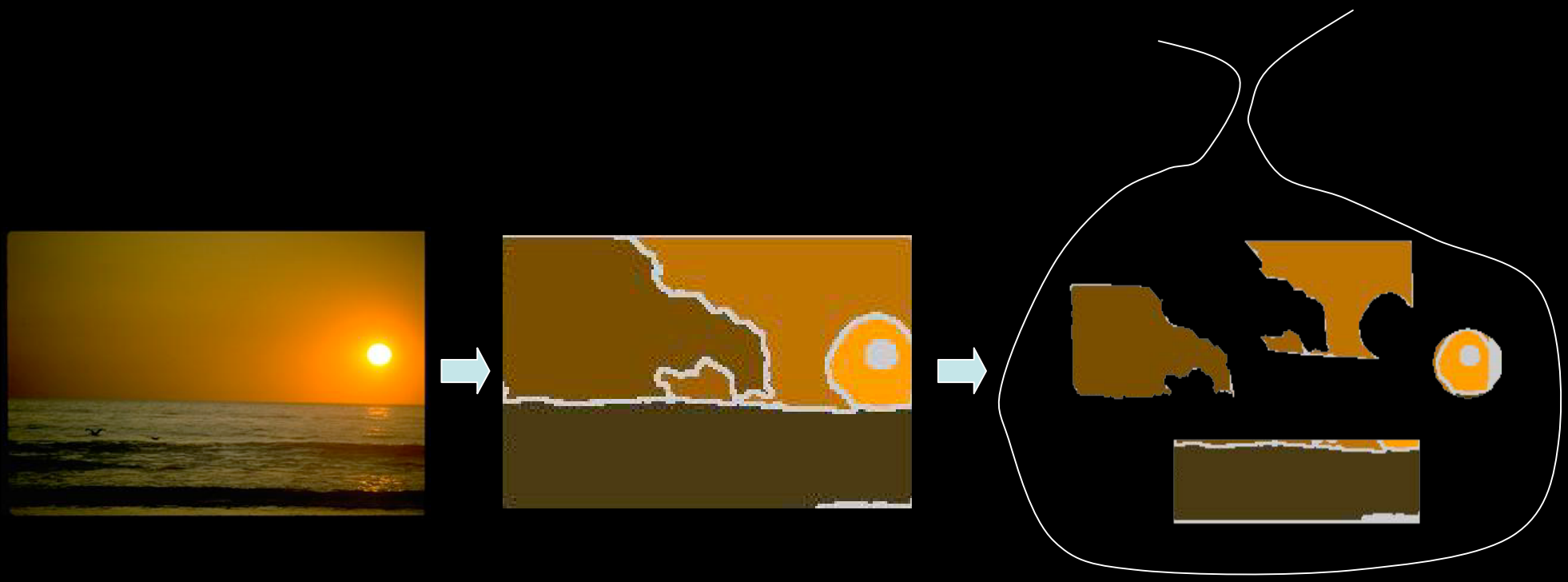
Unknown correspondence between words and image sections in training data



Construct models that implicitly resolve correspondence ambiguity



Preliminaries



W = {sky,water,sun}

R =

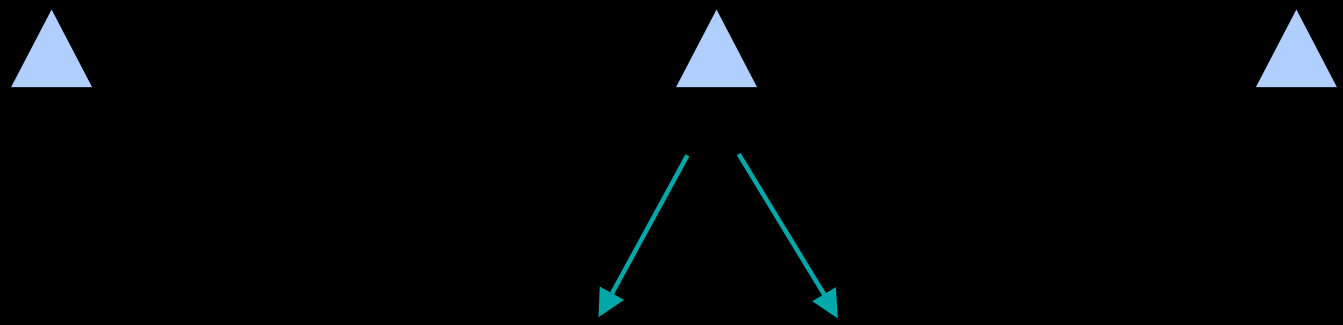
Multi-modal distribution

We want the joint distribution p(W,R)

We assume that words are associated with image parts

So p(W,R) comes from p(w,r)

Image semantic clusters



Generate words by frequency table

e.g., “sun”

Generate blobs by Gaussian over features

e.g. 

multinomial

Gaussian

$$p(w,r) = \sum_{l} p(w \mid l)\, p(b \mid l)\, p(l)$$

prior

and

$$P(w \mid R) = \frac{1}{|R|} \sum_{r} p(w \mid r)$$

(under one modeling assumption)

$$p(w \mid R) = \sum_l p(w \mid l) p(l \mid R)$$

$$p(l \mid R) = \frac{1}{|R|} \sum_r p(l \mid r)$$ (modeling choice)

$$p(l \mid r) = p(r \mid l) p(l) / p(r)$$

$$p(w \mid R) = \frac{1}{|R|} \sum_l p(w \mid l) \sum_r p(r \mid l) p(l) / p(r)$$

$$p(w \mid R) = \frac{1}{|R|} \sum_l \sum_r p(w \mid l) p(r \mid l) p(l) / p(r)$$

$$p(w \mid R) = \frac{1}{|R|} \sum_r p(w \mid r)$$



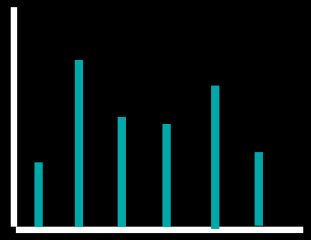
sun
tiger
forest
grass
cat
horse

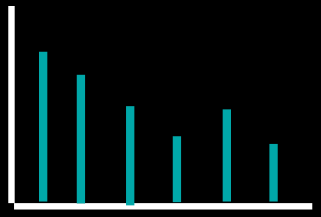
sun
tiger
forest
grass
cat
horse

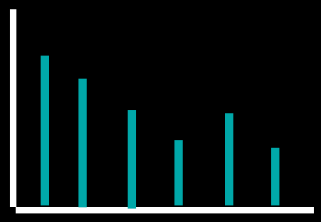
Labeling regions

Annotating images









Examples of Cross-Modal Disambiguation

Reducing Correspondence Ambiguity in Annotated Images

Given an image and its keywords, which part of the image corresponds to the keyword?

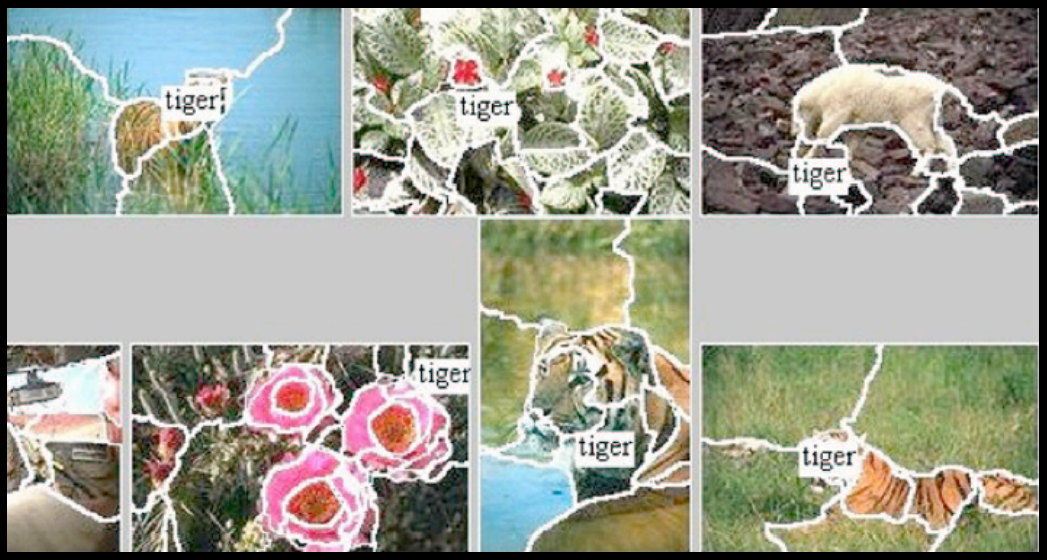


Basic idea --- restrict image annotation to the keywords

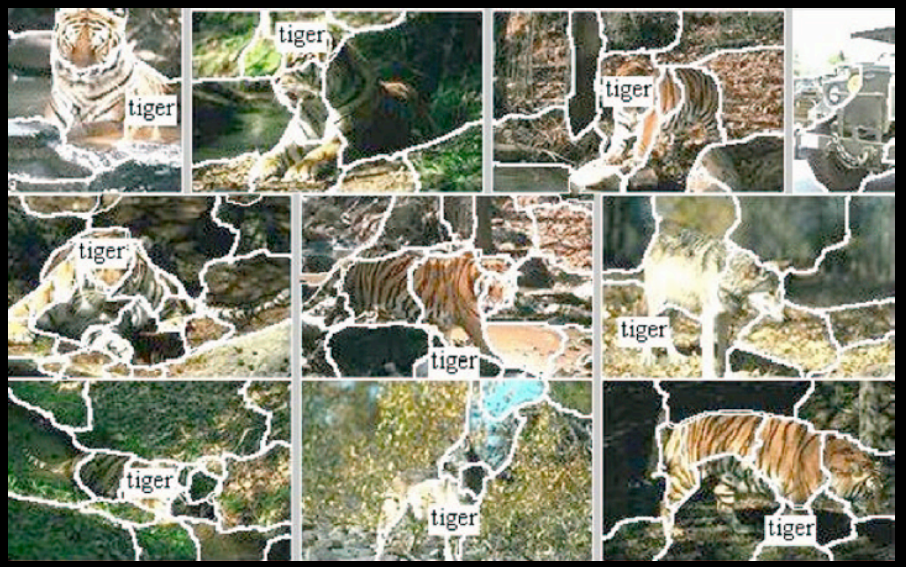
Related idea --- focus on labeling the training data*

*Barnard and Fan, CVPR 07

Reducing Correspondence Ambiguity in Annotated Images



Good tiger regions using only annotation model



Restricting region annotation to images with keyword “tiger”

Reducing Correspondence Ambiguity in Annotated Images

Our image annotation model is rewarded for getting common words right

Simply restricting image annotation overly promotes popular words

If you are given the words to choose from, popularity is irrelevant

Thus we developed a different approach (LTD) designed to learn rare words better.

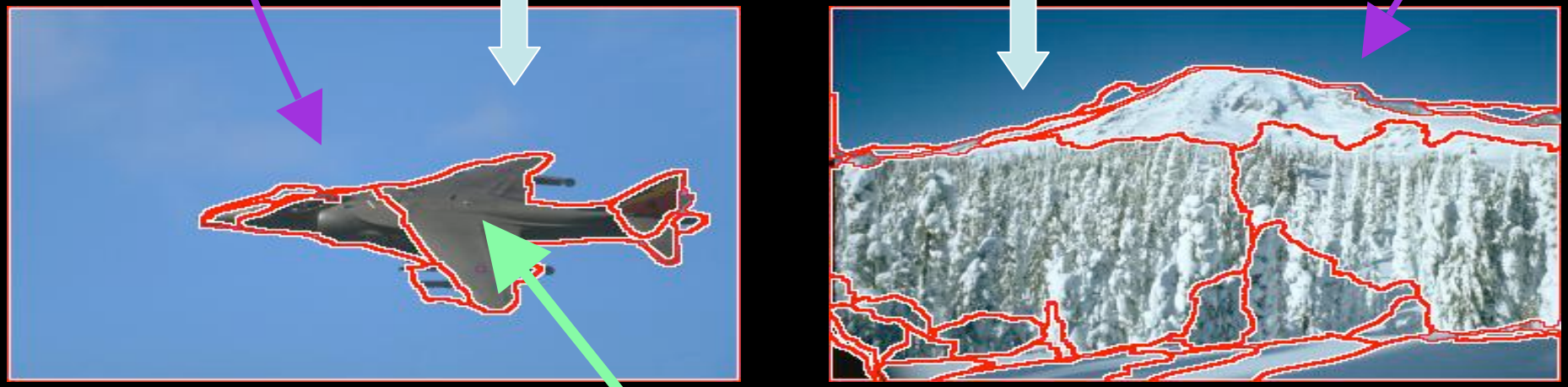
Labeling the training data (LTD)

We first focus on the correspondence ambiguity in the training data

Supports non-parametric inference on held out data

We insist that all image keywords are represented when we label a training image

A better way to learn rare words



Sky is accounted for, so perhaps plane

Affinity matrix specifies for each pair of regions, the probability that they have the same label. Affinity is zero for regions from images with different words.

$$A_{ij} \propto \hat{A} = \begin{cases} \exp\left(-\sum_k \left(x_k^i - x_k^j\right) w_k \left(x_k^i - x_k^j\right)\right) & i \neq j,\ i \text{ and } j \text{ share words} \\ 0 & \text{otherwise} \end{cases}$$

Feature relevance, w_k, is estimated from labeled data.

Each row, $\mathbf{u_i}$, of an assignment matrix, U, holds the estimate of probabilities for each label for a given region.

Estimate the probability that two regions are both labeled with a word, *k*, by $U_{ik} * U_{jk}$.

Estimate the probability that two regions are both labeled with any word by: $\mathbf{u_i} \bullet \mathbf{u_j}$

Combining, $UU^T = A = \alpha \hat{A}$

Constraints: $0 \le U_{iw} \le 1$ and $\sum_w U_{iw} = 1$

$$\sum_{i \in I} U_{iw} \ge \min\left(1, \frac{num_regions_in_I}{num_words_for_I}\right)$$

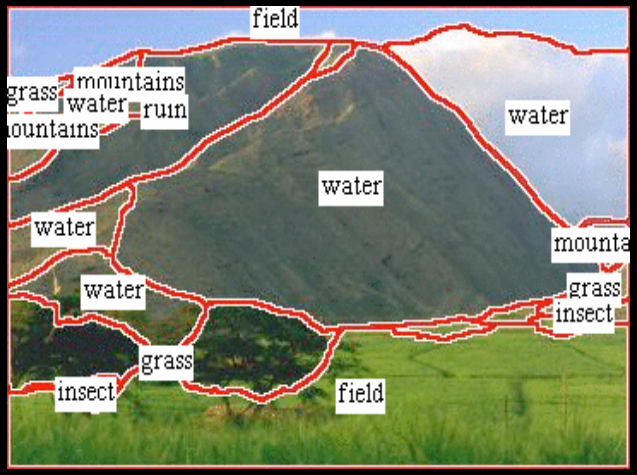
$U_{ik} = 0$, for k $\notin$ words(image(region))

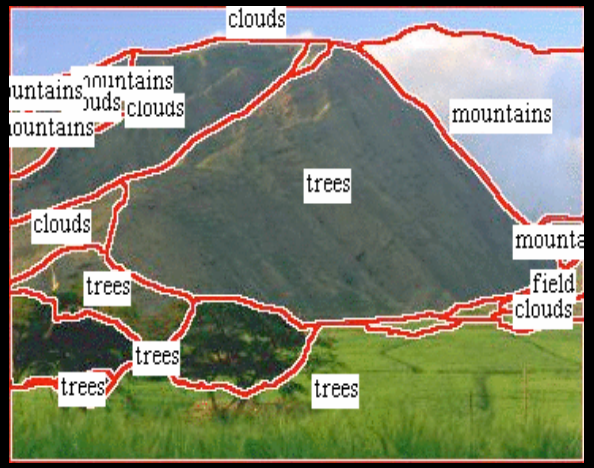
Finding U

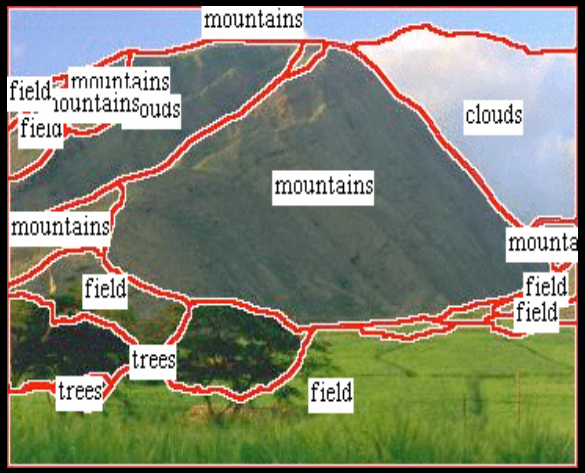
Initialize U as uniform over image words for region

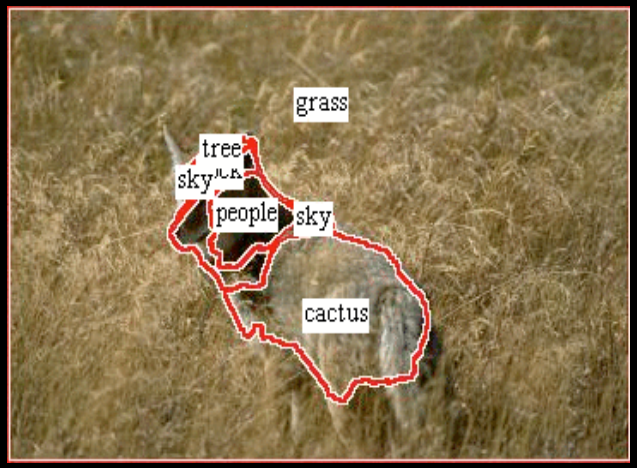
Improve estimate in least squares sense by constrained gradient descent

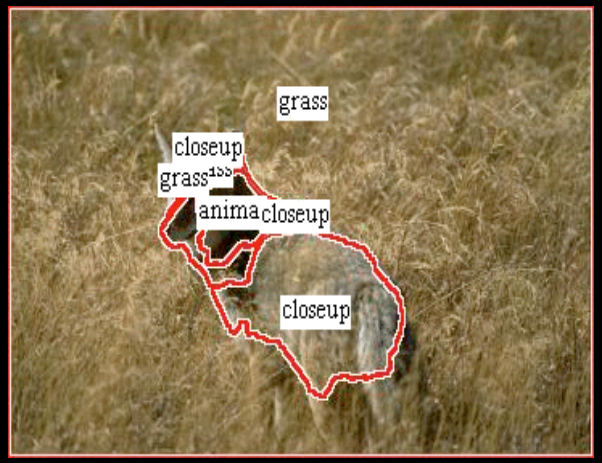
Exploit sparseness of A for efficiency

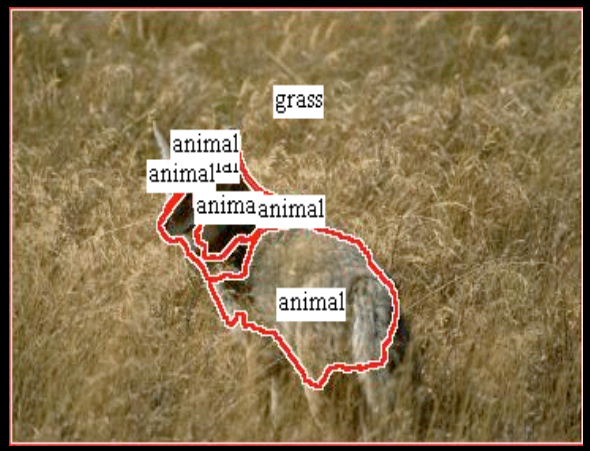












Annotate

Restricted to keywords

Correspondence ambiguity reduction approach

Algorithm	Freq	Range
Empirical distribution	2.05 (1.8)	0.74 (0.3)
Basic annotation	4.89 (1.2)	2.50 (0.5)
Restricted annotation	8.00 (1.3)	5.65 (1.1)
New method	**8.67** (0.9)	**7.40** (1.0)
Theoretical maximum	22.3 (3.3)	20.3 (3.2)

Frequency correct rewards common words

Semantic range normalizes so that each word contributes equally

Dataset and metrology from Barnard et al., IJCV 2008.

The Search for Visual Adjectives*

When can we resolve the correspondence ambiguity?

Some adjectives are visual (e.g. “blue”), but many are not (e.g., “interesting”).

Consider building a model to isolate “pink” regions of a images with the keyword “pink”.

If the entropy of the concept is low, then the word is considered visual.

* Yanai and Barnard, 2005

Testing the Visualness of an adjectives

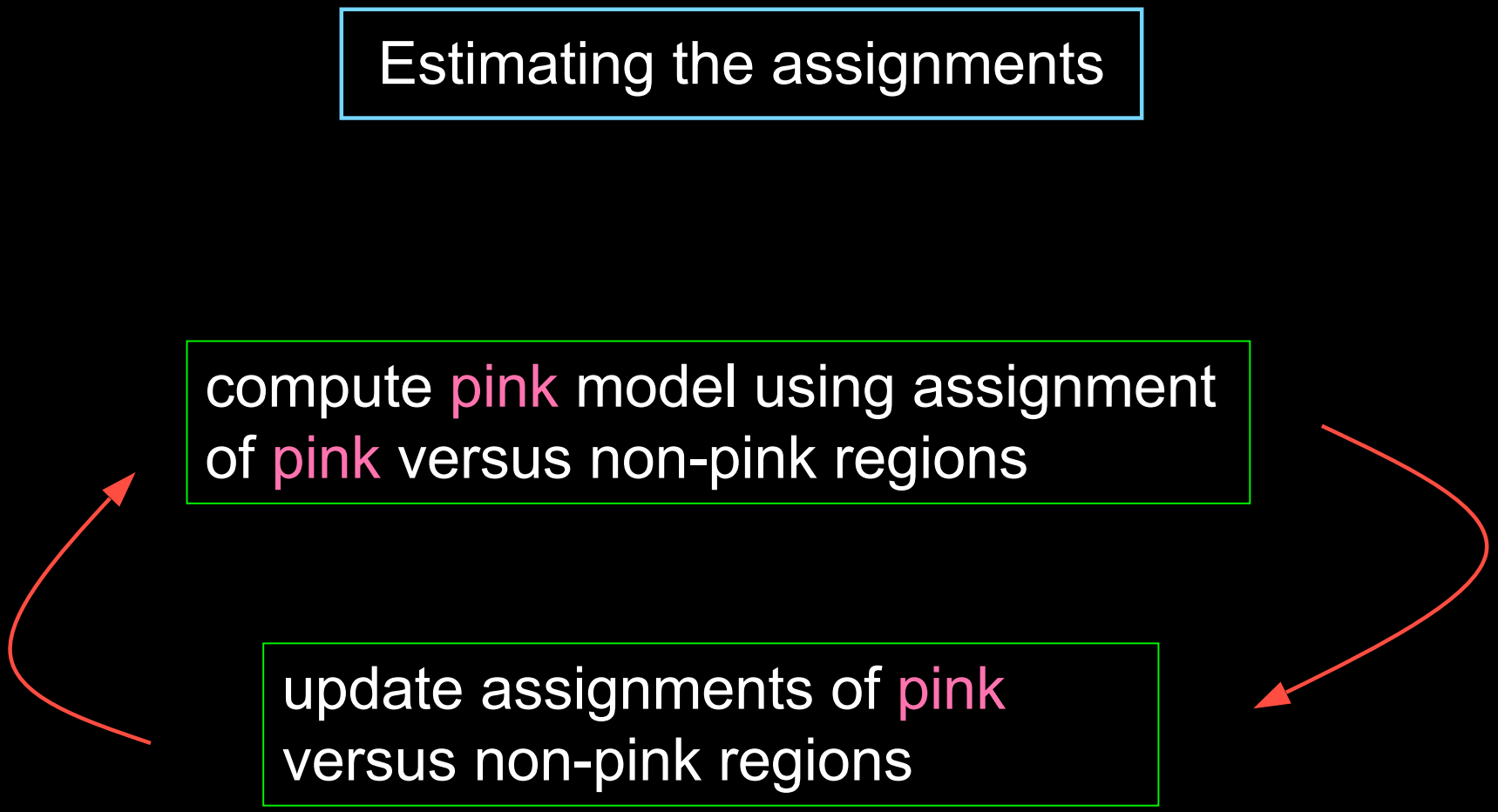
Query Google image finder with an adjective.

Make a “background” model using images without the adjective.

Segment the images

Learn which regions should be associated with the adjective.

Look at the entropy of the regions with respect to a generic model for image regions. If it is relatively low, keep the (visual) adjective .


Estimating the assignments
compute pink model using assignment of pink versus non-pink regions
update assignments of pink versus non-pink regions

Web images found from Google API and “pink” (first iteration)



Images with “pink” are not completely pink!

Best pink regions marked, but this is essentially random at this stage

Web images found from Google API with “pink” (after 10 iterations)



Now best pink regions have meaning

Exploiting parts of speech

If car related keywords have the adjective “red”, we can further reduce the correspondence ambiguity.

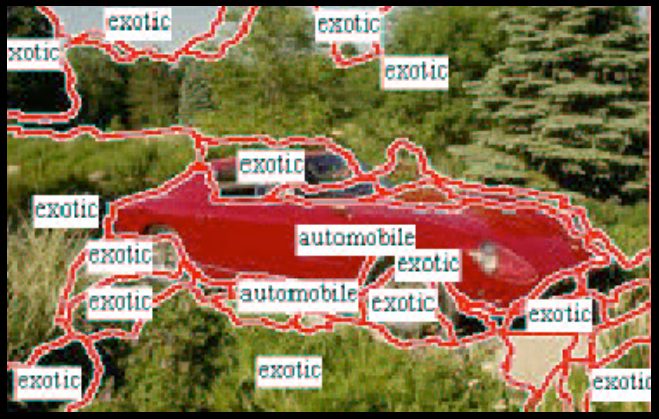


Exploiting parts of speech

$$p(n \mid r) = p(n, a \mid r) \propto p_n(n \mid r) p_a(a \mid r)$$

(under some assumptions)

Exploiting parts of speech



adjectives only
(restricted)

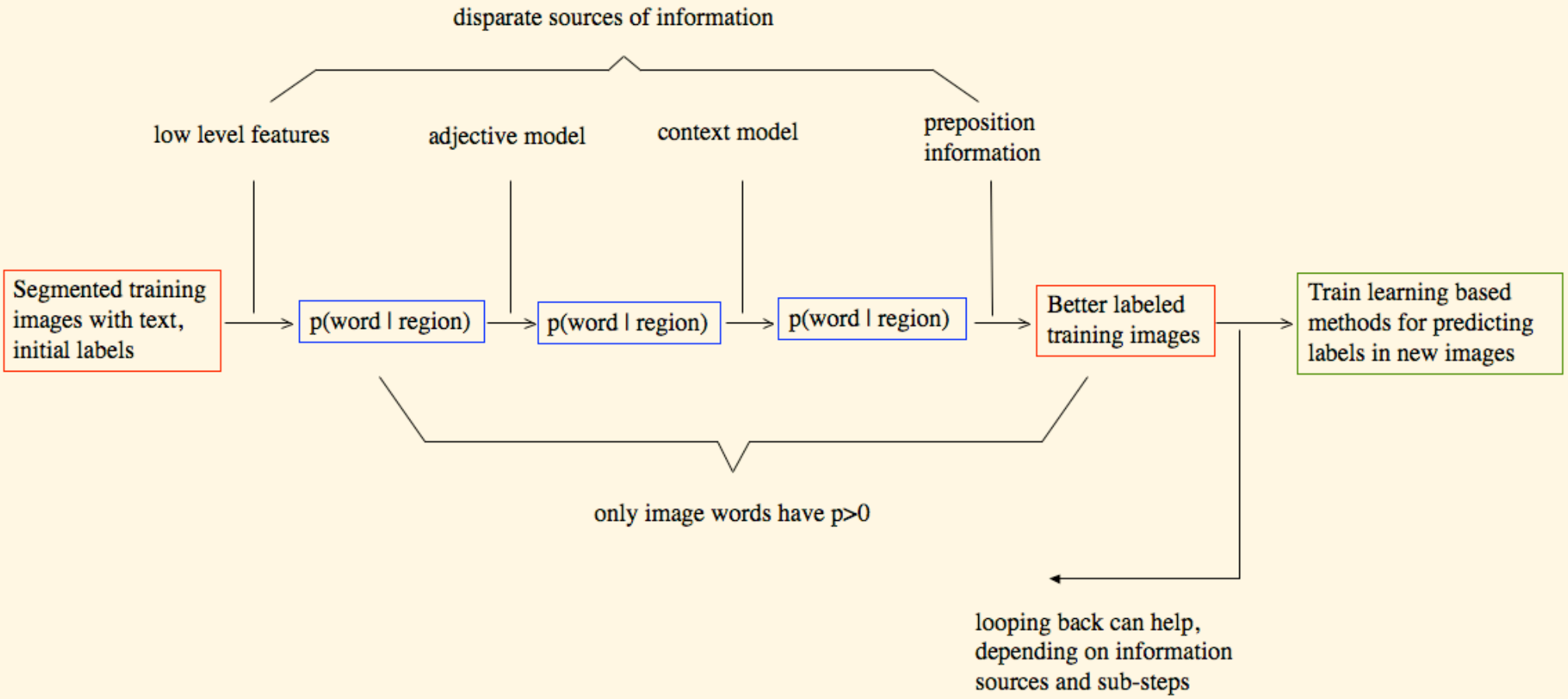


nouns only



both

The image labeling pipeline



Exploiting parts of speech

Bibliographic note --- Recently Gupta and Davis (ECCV 2008) substantively expanded on using parts of speech other than nouns in image annotation.

Word Sense Disambiguation with Pictures*

Using images to help ground language

*Barnard and Johnson, AIJ 05

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

Can Images illustrating the text help?

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



Preliminaries

- Text tools such as WordNet catalog senses, s, for words, w.
- The word, w = “bank” can correspond to one of several senses, “s = bank_1” or s = “bank_2”
- Word sense disambiguation can be considered as estimating p(s|w,…)

Can Images illustrating the text help?

T

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

p (s | w, T)

I



p (s | w, I)

Combining Sources

Assuming conditional independence and uniform prior on s,

$$p(s \mid w, T, I) \propto p(s \mid w, T) \bullet p(s \mid w, I)$$

Cross-modal contributions

For $p(s \mid w, T)$ we implemented the winner of the SENSEVAL-2 competition of text based methods

For $p(s \mid w, I)$ we use automated image annotation but using training data with word senses.

ImCor

Training and testing requires an annotated image collection which is sense disambiguated (does not exist)

We linked the Corel data set to a standard, sense disambiguated corpus (SemCor)

About 23,000 images could be linked

ImCor collection tool



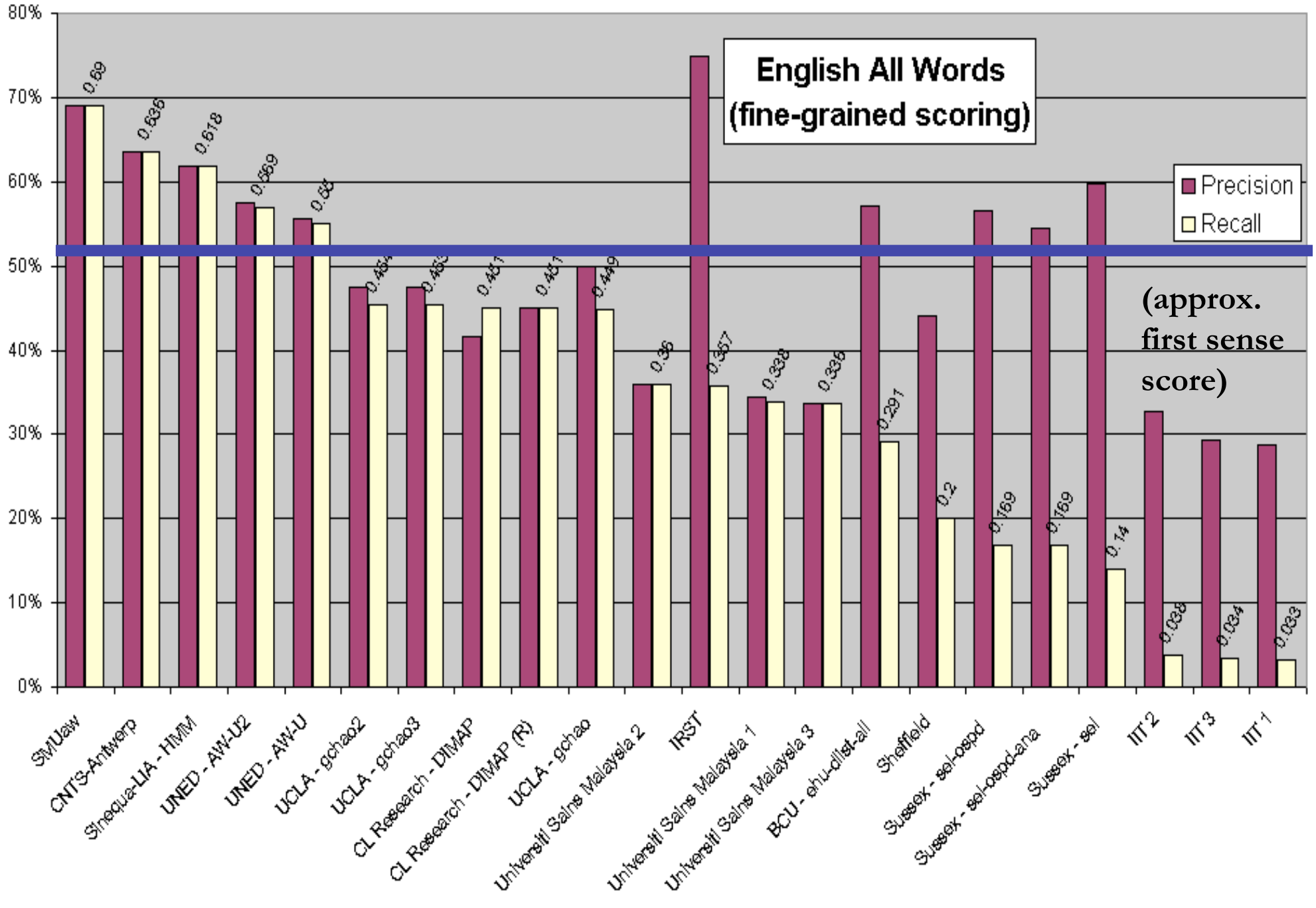
Results

Word sense disambiguation strategy	Score*
Baseline (most common sense)	0.615 (0.002)
Text based method alone	0.683 (0.002)
Images alone	0.791 (0.002)
Image and text together	0.817 (0.002)

* Percentage of senses correctly predicted

Results

- Images can disambiguate text in this data set (beats the text based method!)
- Combining the sources further increases performance
  - But not as much as expected if independence assumption holds completely
  - Plausible explanation
    - Image model also embodies empirical distribution of text and Traupman and Wilensky suggests that text based WSD has difficulty doing much better than this.



[SENSEVAL-2 Proceedings, 2001; Traupman and Wilensky]

Disambiguating lecture speech using slides

- SLIC (semantically linked instructional content) exploits multimodal data in recorded lectures
- Data modalities
  - electronic slides
  - video
    - with extracted laser pointer and speaker from video
  - audio
    - with speech recognition
- A key technology is robust matching of slides to video frames*

*Fan et al. 2006, 2007, 2008, 2009

Info | Search | Options | Help

THE UNIVERSITY OF ARIZONA

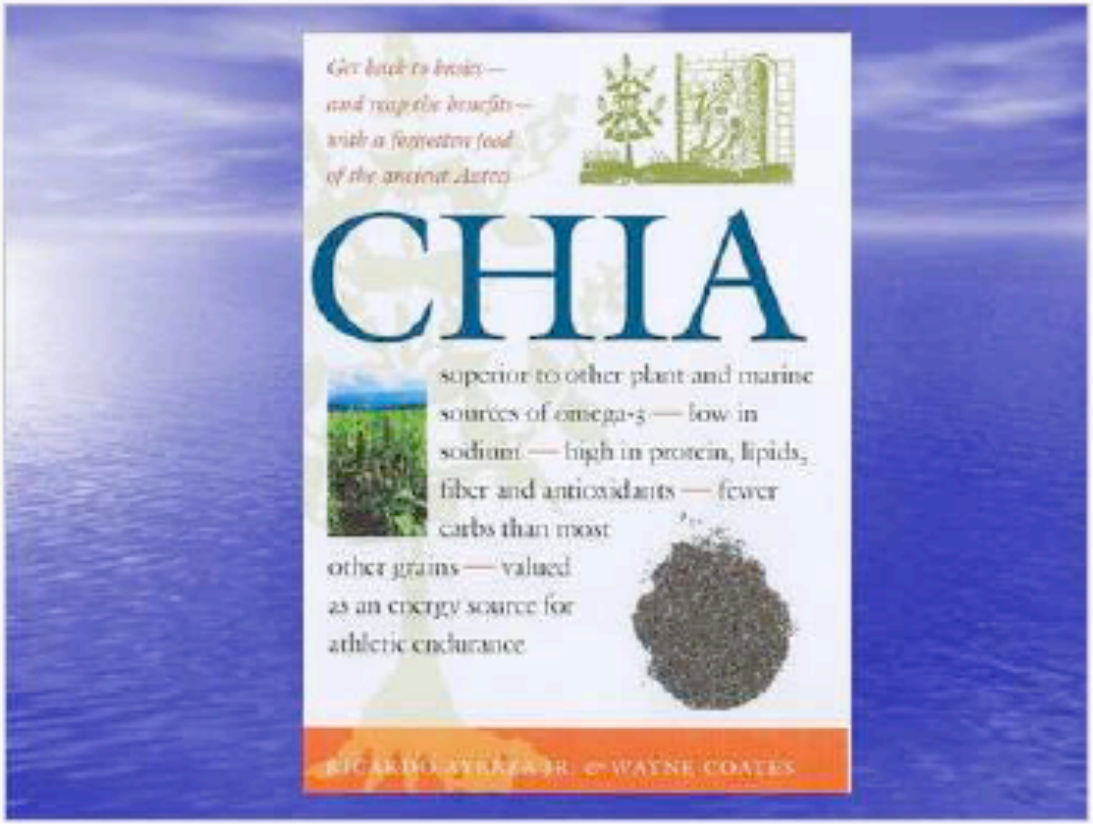
Makes a case for the consumption and production of the Chia plant. Overview of the history of Chia beginning with the Aztecs. Describes change in human diet since the evolutionary period with concentration on declining levels of omega-3



Chia: Rediscovering a Forgotten Crop of the Aztecs

Faculty Fellows Speaker Series, Unknown

Wayne Coates (00/00/0000)

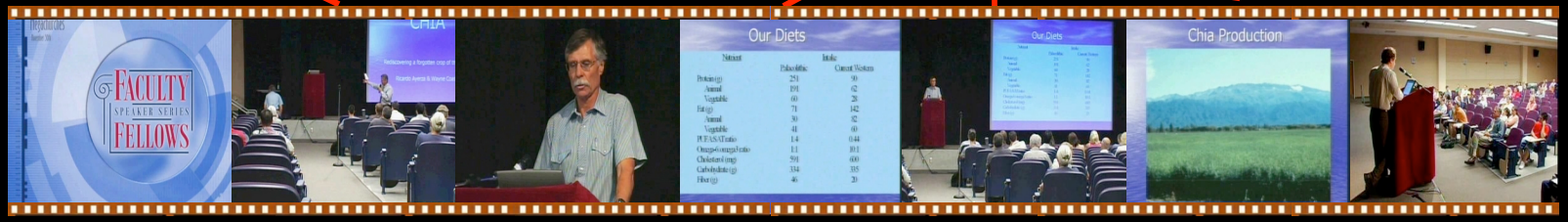




Slides



Frames



Intricate combination of known techniques gives error rate of only a few percent.

Also can get an accurate location of the slide in the frame --- very useful for user interface tricks.

Disambiguating lecture speech using slides

- Given slide matching, we know what slide was covered for a given time period.
- Can the slide words help improve the speech transcript?
  - Slide words are more likely to be technical, and either rare, or out of vocabulary
  - Many slide words are perhaps important to get right
- Example
  - “chia” --> “she a”

Possible approach

- Assume we only get a transcript from a speech recognizer.
- Convert transcript to phonemes.
- Align phoneme stream to slide using
  - phoneme confusion
  - priors on slide word forward jumps and backward jumps
- Results due soon!







The End









