Words and Pictures

D.A. Forsyth

U Illinois Urbana-Champaign
and UC Berkeley

With Kobus Barnard, Pinar Duygulu, Nando de Freitas,
R. Manmatha, Tamara Miller, Alex Berg, JT Edwards, Ryan White,
Michael Maire, Yee-Whye Teh, Erik Learned-Miller,
David Blei, Roger Bock and Deva Ramanan

Thanks to NSF Digital Library Initiative II.
Kobus Barnard also receives funding from NSERC (Canada)
Pinar Duygulu is also supported by TUBITAK (Turkey)
Tamara Miller and Deva Ramanan have NSF Fellowships
Ryan White, Roger Bock and Michael Maire have Department of
Homeland Security Fellowships (US)
Yee Whye Teh and Erik Learned-Miller are partially supported under an ONR MURI

LOTS of BIG collections of images

Corel Image Data	40,000 images
Fine Arts Museum of San Francisco	83,000 images online
Cal-flora	20,000 images, species information
News photos with captions (yahoo.com)	1,500 images per day available from yahoo.com
Hulton Archive	40,000,000 images (only 230,000 online)
internet.archive.org	1,000 movies with no copyright
TV news archives (televisionarchive.org, informedia.cs.cmu.edu)	Several terabytes already available
Google Image Crawl	>330,000,000 images (with nearby text)
Satellite images (terrarserver.com, nasa.gov, usgs.gov)	(And associated demographic information)
Medial images	(And associated with clinical information)

* and the BBC is releasing its video archive, too;
and we collected 500,000 captioned news images;
and it's easy to get scanned mediaeval manuscripts;
etc., etc.,

Imposing order

- Iconic matching
  - child abuse prosecution
  - managing copyright (BayTSP)

Current, practical applications

- Clustering
  - Browsing for:
    - web presence for museums (Barnard et al, 01)
    - home picture, video collections
    - selling pictures

Maybe applications

- Searching
  - scanned writing (Manmatha, 02)
  - collections of insects

Maybe applications

- Building world knowledge
  - a face gazetteer (Miller et al, 04)

Search is well studied

- Metadata indexing
  - keywords, date of photo, place, etc.
- Content based retrieval
  - query by example with
    - global features
      - (e.g. Flickner et al. 95, Carson et al. 99, Wang 00, various entire conferences)
    - local features
      - (e.g. Photobook - Pentland et al 96; Blobworld - Carson et al, 98)
    - relevance feedback
      - (e.g. Cox et al 00; Santini 00; Schettini 02; etc.)
  - query by class
    - naughty pictures
      - (eg. Forsyth et al. 96, 99; Wang et al. 98; Chan et al 99)

What will users pay for?

- Work by Peter Enser and colleagues on the use of photø movie collections
  (Enser McGregor 92; Ornager 96; Armitage Enser 97; Markkula Sormunen 00; Frost et al 00; Enser 00)
- Typical queries:

What is this about?

"… smoking of kippers…"

"The depiction of vanity in painting, the depiction of the female figure looking in the mirror, etc."

"Cheetahs running on a greyhound course in Haringey in 1932"

Annotation results in complementary words and pictures

Query on

“Rose”

Example from Berkeley Blobworld system

Annotation results in complementary words and pictures

Query on

Example from Berkeley
Blobworld system

Annotation results in complementary words and pictures

Query on

“Rose”

and

Example from Berkeley
Blobworld system

Exploiting complementary information

- A probability model linking images and annotations
  - exploit co-occurence
  - better estimates of “meaning” for clustering and browsing
  - soft search, auto illustration, auto annotation
- Predicting words from image regions
  - explicitly encode and infer correspondence
    - aligned bitext
    - no alignment
  - rather like recognition
  - pinch techniques from statistical natural language processing
- Linking face images with names
  - an important special case
  - datasets of an epic scale available
  - like face recognition, but easier
  - breaking correspondence by clustering

Browsing

- Searching big, unknown collections is hard for naive user
  - skilled users don’t benefit from vision-based tools
  - problem of overrated significance
- Browsing?
  - seems to be preferred by naive users (Frost et al, `00)
  - but browsing requires organization too
  - generally underrated problem

*Notable exceptions ---Sclaroff, Taycher, and La Cascia, 98; Rubner, Tomasi, and Guibas, 00; Smith Kanade, 97.

Clustering words and pictures

- Build a joint probability model linking words and pictures
- Use Hoffman's hierarchical aspect model
  - which is a form of clusterer [Hofmann 98; Hofmann & Puzicha 98]
- Lay out and browse the clusters

Input

"This is a picture of the sun setting over the sea with waves in the foreground"

Image processing*

Each blob is a large vector of features

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

Language processing

sun sky waves sea

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

FAMSF Data

Web number: 4359202410830012

rec number: 2

Title: Le Matin

Primary class: Print

Artist: Tissot

Description:
serving woman stands in a dressing room, in front of vanity with chair, mirror and mantle, holding a tray with tea and toast

Display date: 1886

Country: France

83,000 images online, we clustered 8000

Natural Language Processing

- Parts of speech* (prefer nouns for now)
- Expand semantics using WordNet†
- Sense Disambiguation

* We use Eric Brill's parts of speech tagger (available on-line)

† WordNet is an on-line lexical reference system from Princeton (Miller et.al)

Node Behavior

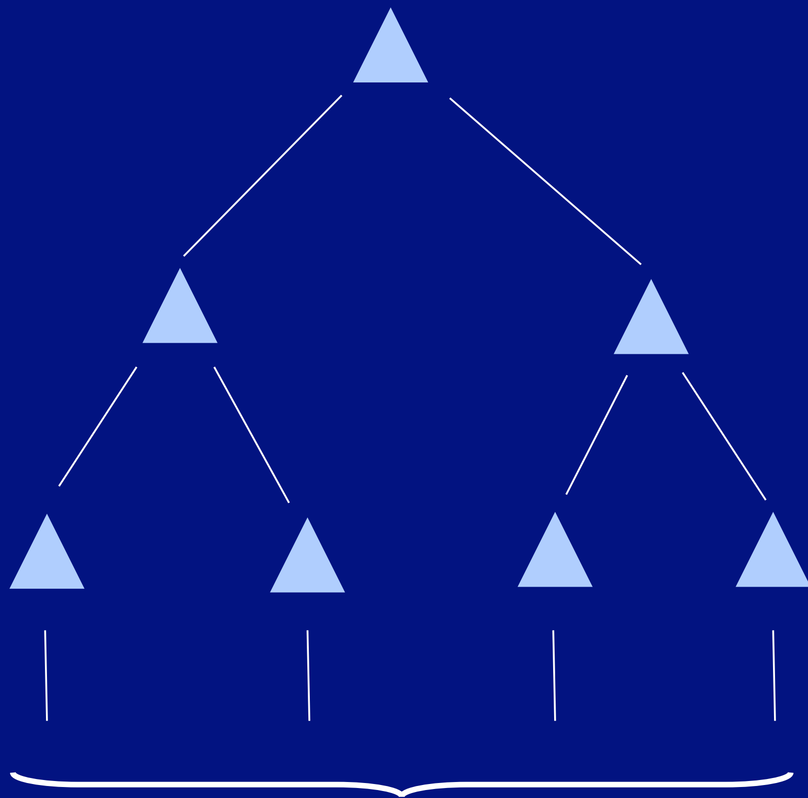
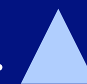


Image Clusters

Each node .... 

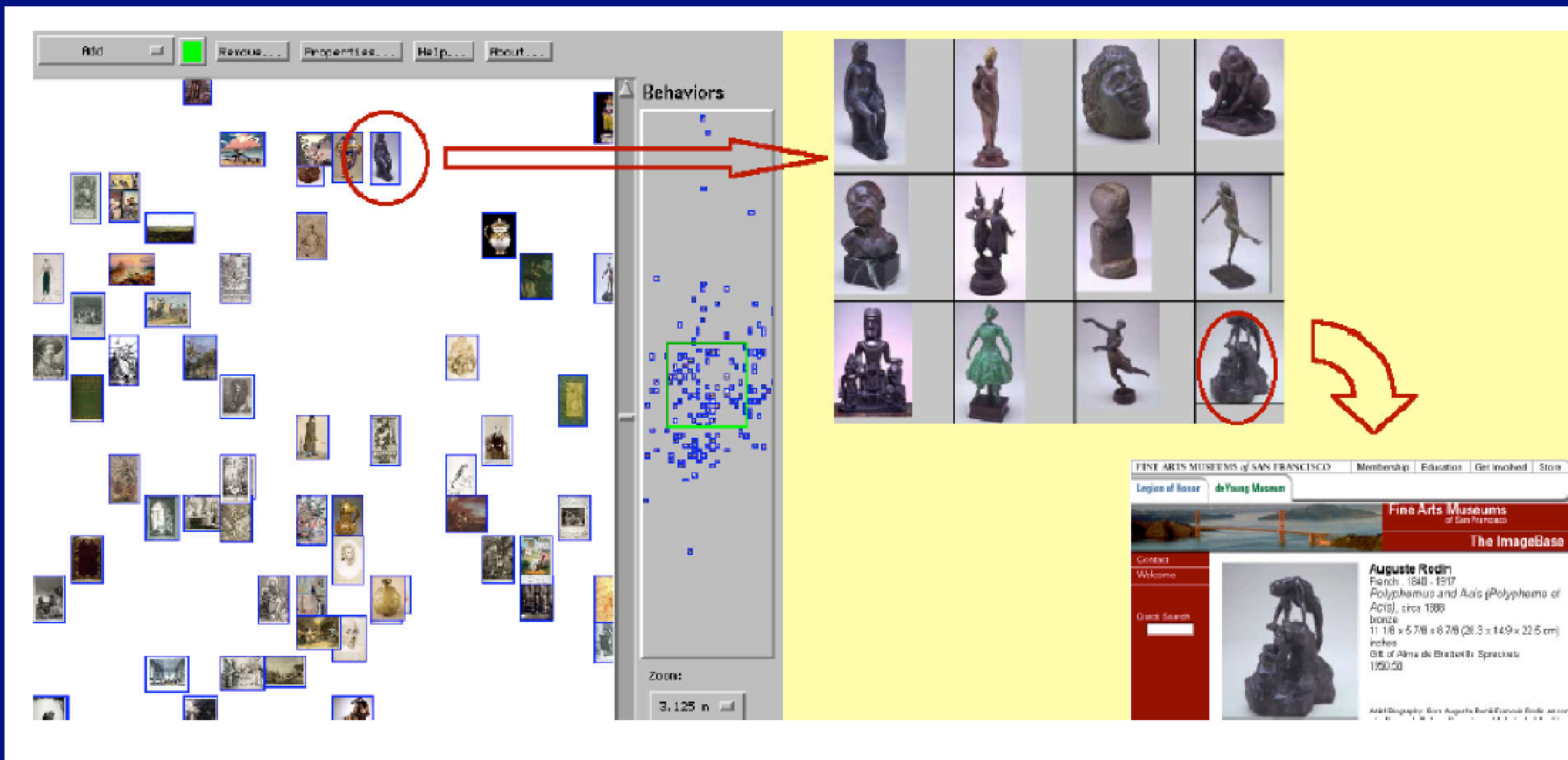
Emits each modeled word,
W , with some probability

Generates blobs according
to a Gaussian distribution
(parameters differ for each node).

- Estimation
  - Straightforward missing data problem
  - EM
    - If path, node known for each data element, easy to get estimate of parameters
    - given parameter estimate, path, node easy to figure out

FAMSF Demo

(Based on GIS Viewer from UC Berkeley
digital library project)


Add
Remove...
Properties...
Help...
About...
Behaviors
Zoom:
FINE ARTS MUSEUMS of SAN FRANCISCO
Membership
Education
Store
Fine Arts Museums
The ImageBase
Contact
Welcome
Quick Search
Auguste Rodin
bronze
inches

Pictures from Words (Auto-illustration)

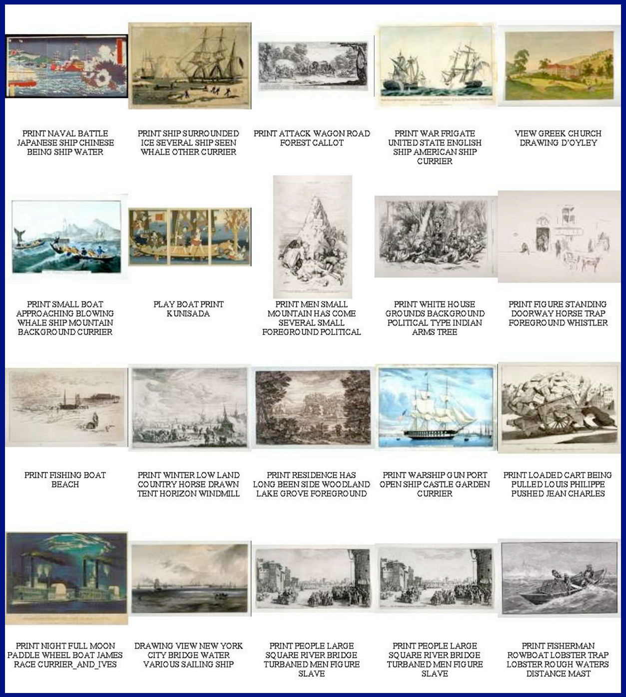
Text Passage (Moby Dick)

"The large importance attached to the harpooneer's vocation is evinced by the fact, that originally in the old Dutch Fishery, two centuries and more ago, the command of a whale-ship …"

Extracted Query

large importance attached fact old dutch century more command whale ship was person was divided officer word means fat cutter time made days was general vessel whale hunting concern british title old dutch ...

Retrieved Images



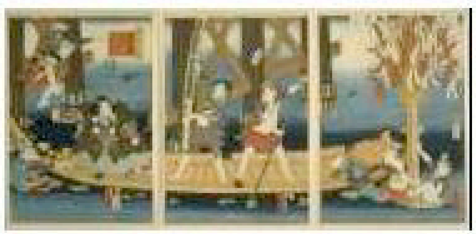
PRINT NAVAL BATTLE JAPANESE SHIP CHINESE BEING SHIP WATER

PRINT SHIP SURROUNDED ICE SEVERAL SHIP SEEN WHALE OTHER CURRIER

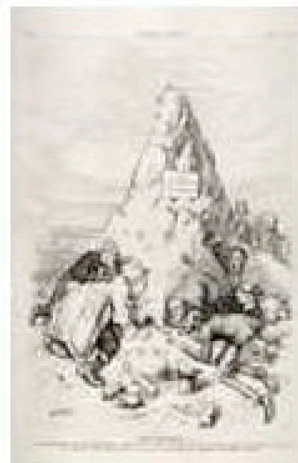
PRINT ATTACK WAGON ROAD FOREST CALLOT

PRINT WAR FRIGATE UNITED STATE ENGLISH SHIP AMERICAN SHIP CURRIER

PRINT SMALL BOAT APPROACHING BLOWING WHALE SHIP MOUNTAIN BACKGROUND CURRIER



PLAY BOAT PRINT KUNISADA



PRINT MEN SMALL MOUNTAIN HAS COME SEVERAL SMALL FOREGROUND POLITICAL

PRINT WHITE HOUSE GROUNDS BACKGROUND POLITICAL TYPE INDIAN ARMS TREE

Auto-annotation

- Predict words from pictures
  - Obstacle:
    - Hoffman’s model uses document specific level probabilities
  - Dodge
    - smooth these empirically
- Attractions:
  - easy to score
  - large scale performance measures (how good is the segmenter?)
  - possibly simplify retrieval (Li+Wang, 03)

Keywords
GRASS TIGER CAT FOREST
Predicted Words (rank order)

tiger cat grass people water bengal
buildings ocean forest reef

Keywords
HIPPO BULL mouth walk
Predicted Words (rank order)

water hippos rhino river grass
reflection one-horned head
plain sand

Keywords
FLOWER coralberry LEAVES
PLANT

Predicted Words (rank order)
fish reef church wall people water
landscape coral sand trees

To do

- Package up software for clustering and drop on various museums
- Experiment with other image representations, segment fusing, etc. (some already in Barnard et al, ‘03)
- Better layout

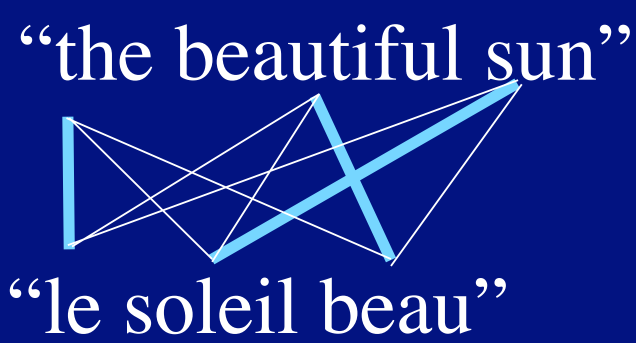
Exploiting complementary information

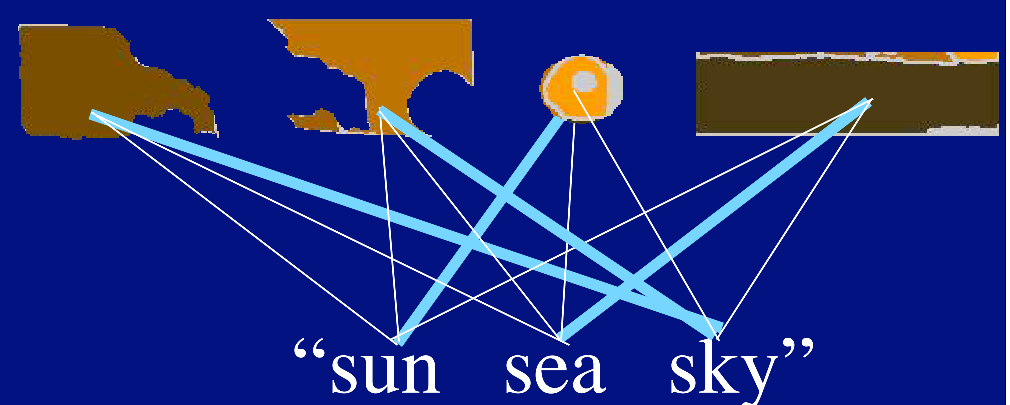
- A probability model linking images and annotations
  - exploit co-occurence
  - better estimates of “meaning” for clustering and browsing
  - soft search, auto illustration, auto annotation
- Predicting words from image regions
  - explicitly encode and infer correspondence
    - aligned bitext
    - no alignment
  - rather like recognition
  - pinch techniques from statistical natural language processing
- Linking face images with names
  - an important special case
  - datasets of an epic scale available
  - like face recognition, but easier
  - breaking correspondence by clustering

Annotation vs Recognition

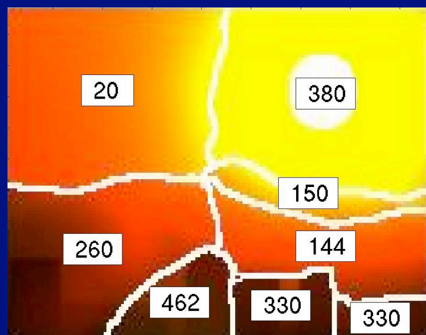
Lexicon building

- In its simplest form, missing variable problem
- Pile in with EM
  - given correspondences, conditional probability table is easy (count)
  - given cpt, expected correspondences could be easy
- Caveats
  - might take a lot of data; symmetries, biases in data create issues

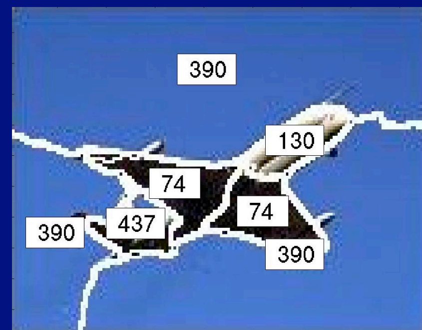




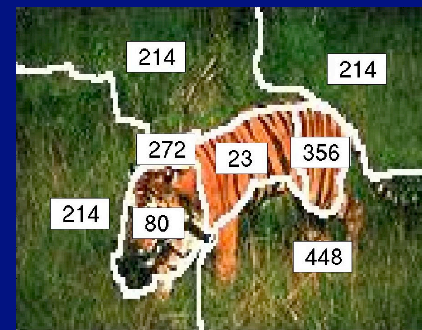
Brown, Della Pietra, Della Pietra & Mercer 93; Melamed 01



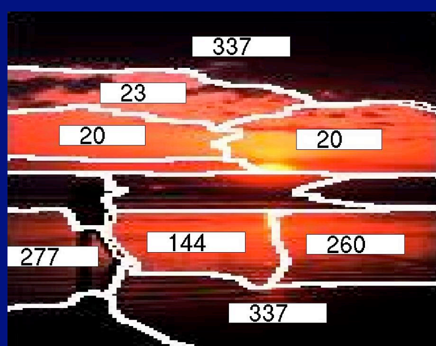
city mountain sky sun



jet plane sky

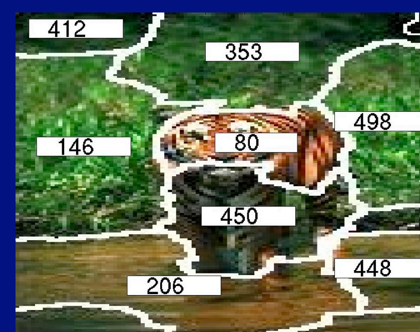


cat forest grass tiger

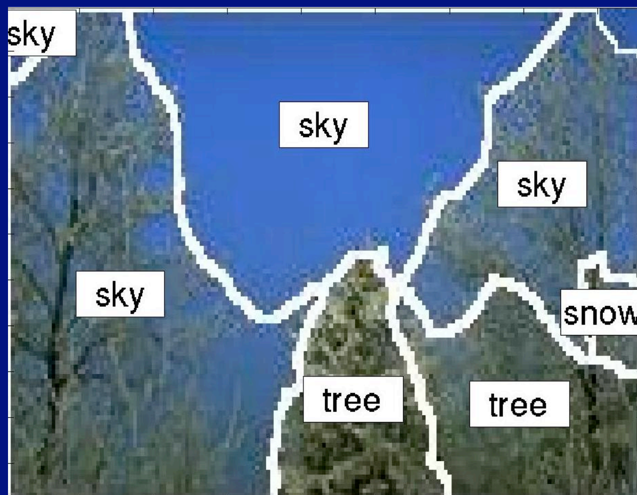


beach people sun water

jet plane sky

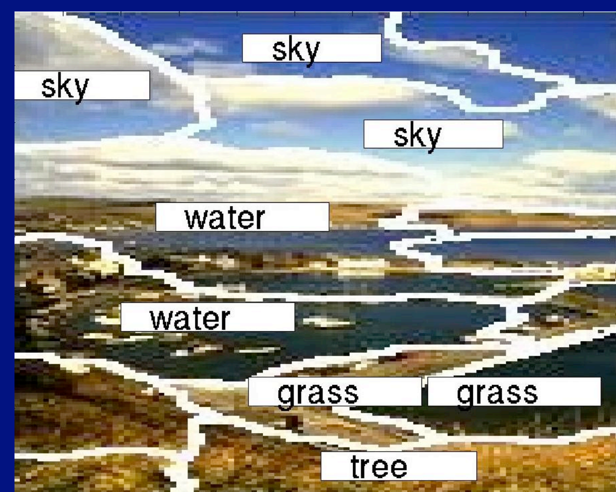


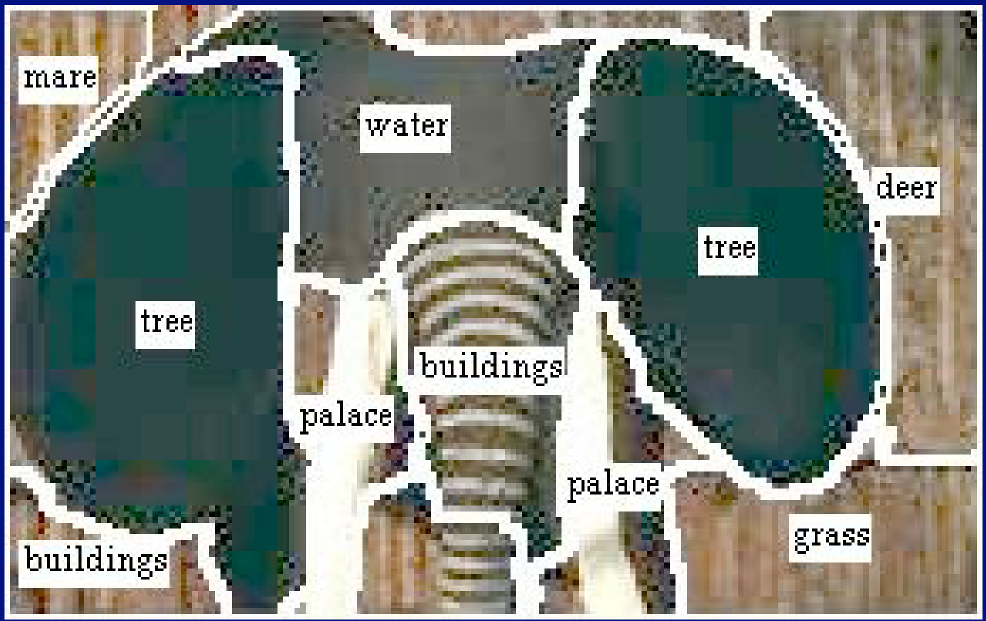
cat grass tiger water


sky
sky
sky
sky
snow
tree
tree


sky
sky
water
buildings
sky
tree
buildings
buildings


rocks
sky
beach
water
tree
people
street
cars


sky
sky
sky
water
water
grass
grass
tree

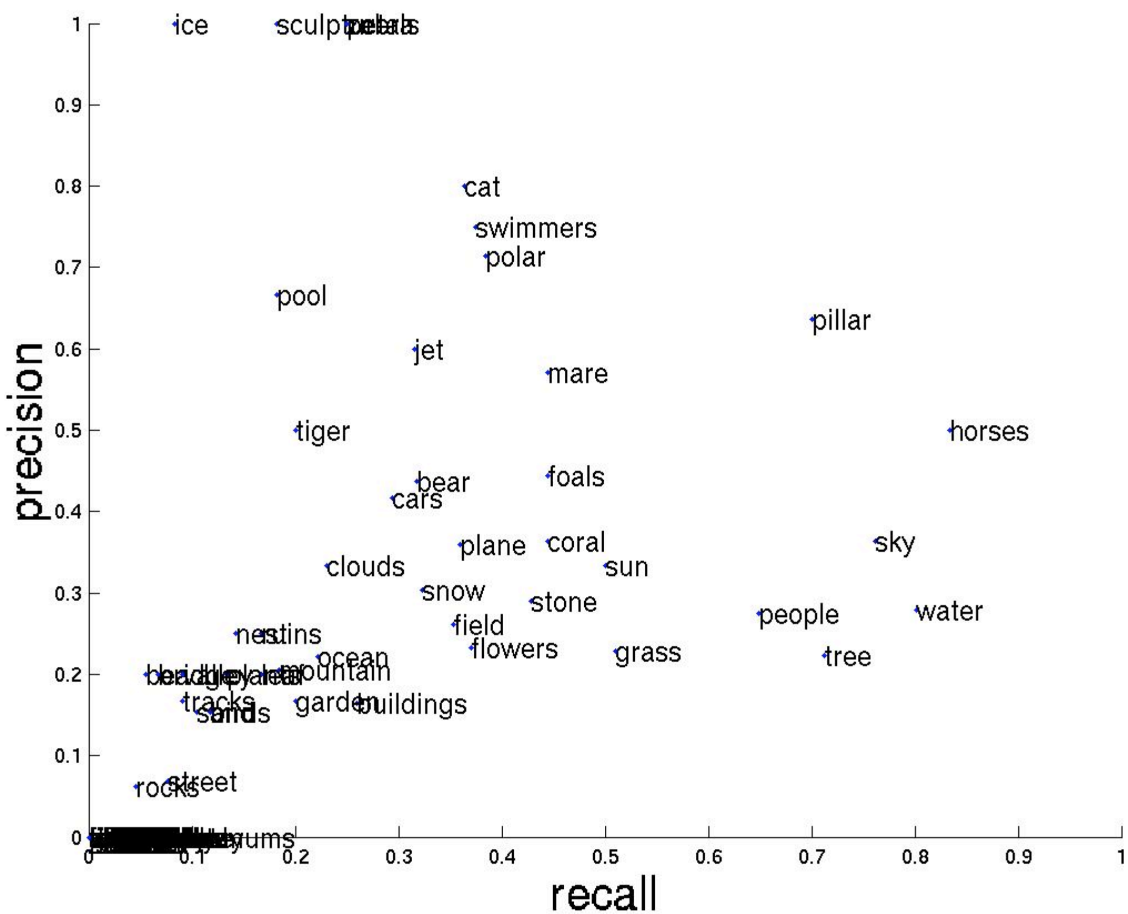

mare
water
deer
tree
tree
buildings
palace
palace
grass
buildings

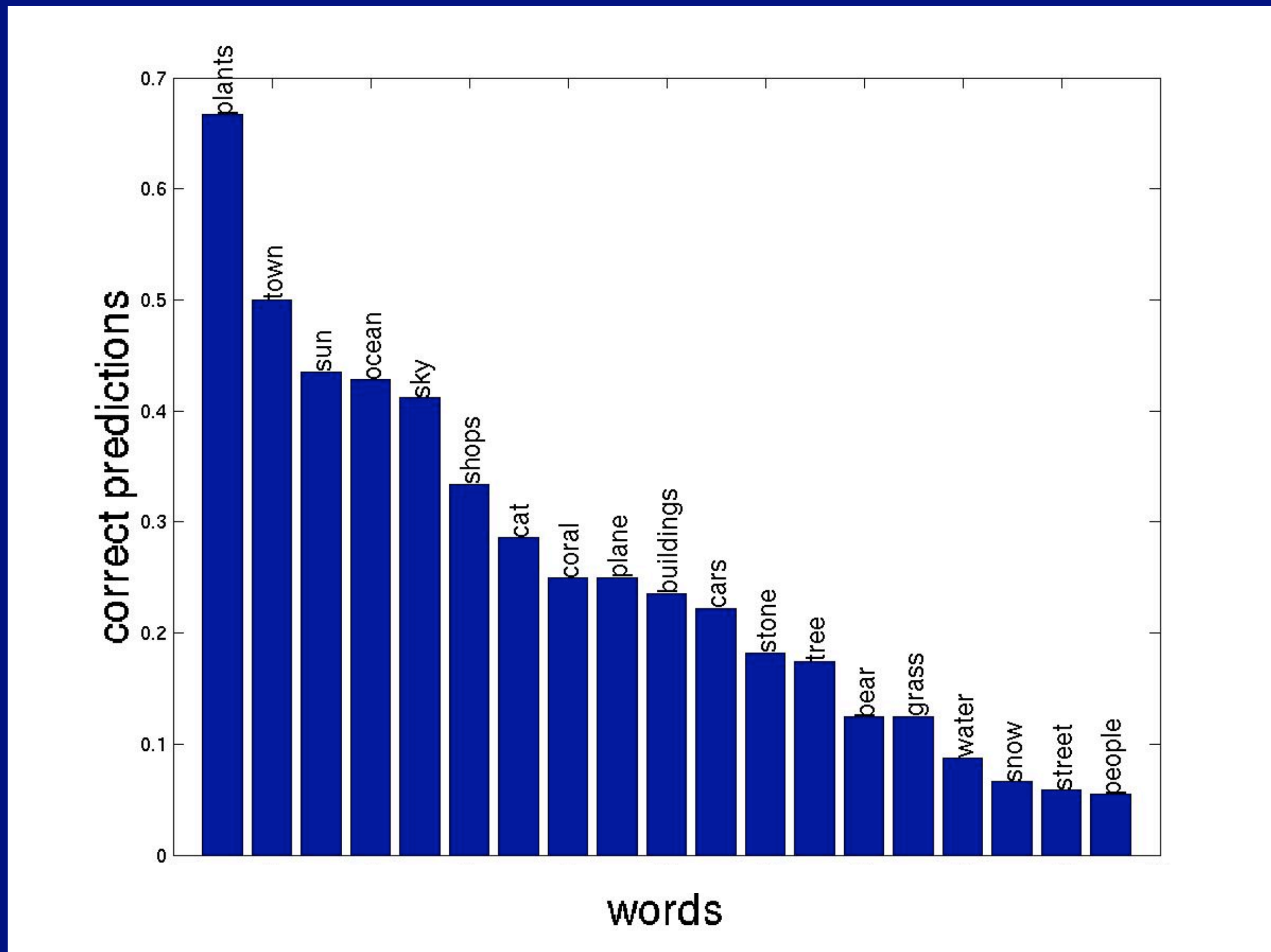
Performance measurement

By hand

By proxy

Grass Cat Buildings
Horses Tiger Mare


precision
recall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
ice
cat
swimmers
polar
pool
pillar
jet
mare
tiger
horses
bear
cars
foals
plane
coral
sky
clouds
sun
snow
stone
people
water
field
flowers
grass
tree
ocean
mountain
tracks
garden
buildings
street
rocks



correct predictions
words
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
plants
town
sun
ocean
sky
shops
cat
coral
plane
buildings
cars
stone
tree
bear
grass
water
snow
street
people

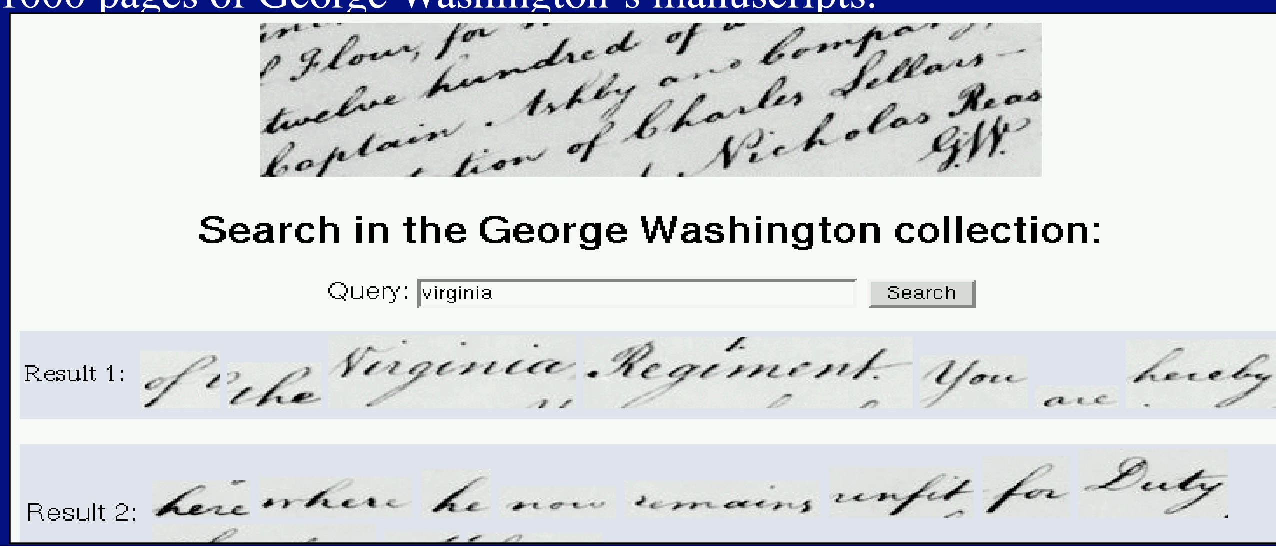
Scanned handwriting

- Special case of words and pictures
- Important applications
  - military
  - climate change
- Various versions
  - aligned training set
    - scanned hw + transcription=supervised data
    - uncommon
  - no aligned training data
    - but letter and word frequencies are preserved
    - extremely useful

Word spotting

- Large collections of scanned handwritten documents are common; handwriting recognition doesn't work
  - make documents searchable with free text ascii queries
    - scanned text is pictures, transcription is words
    - do auto annotation
    - e.g. – T. Rath, R. Manmatha and V. Lavrenko, A Search Engine for Historical Manuscript Images, To Appear Proc. SIGIR'04.
    - Dataset 1000 pages of George Washington's manuscripts.

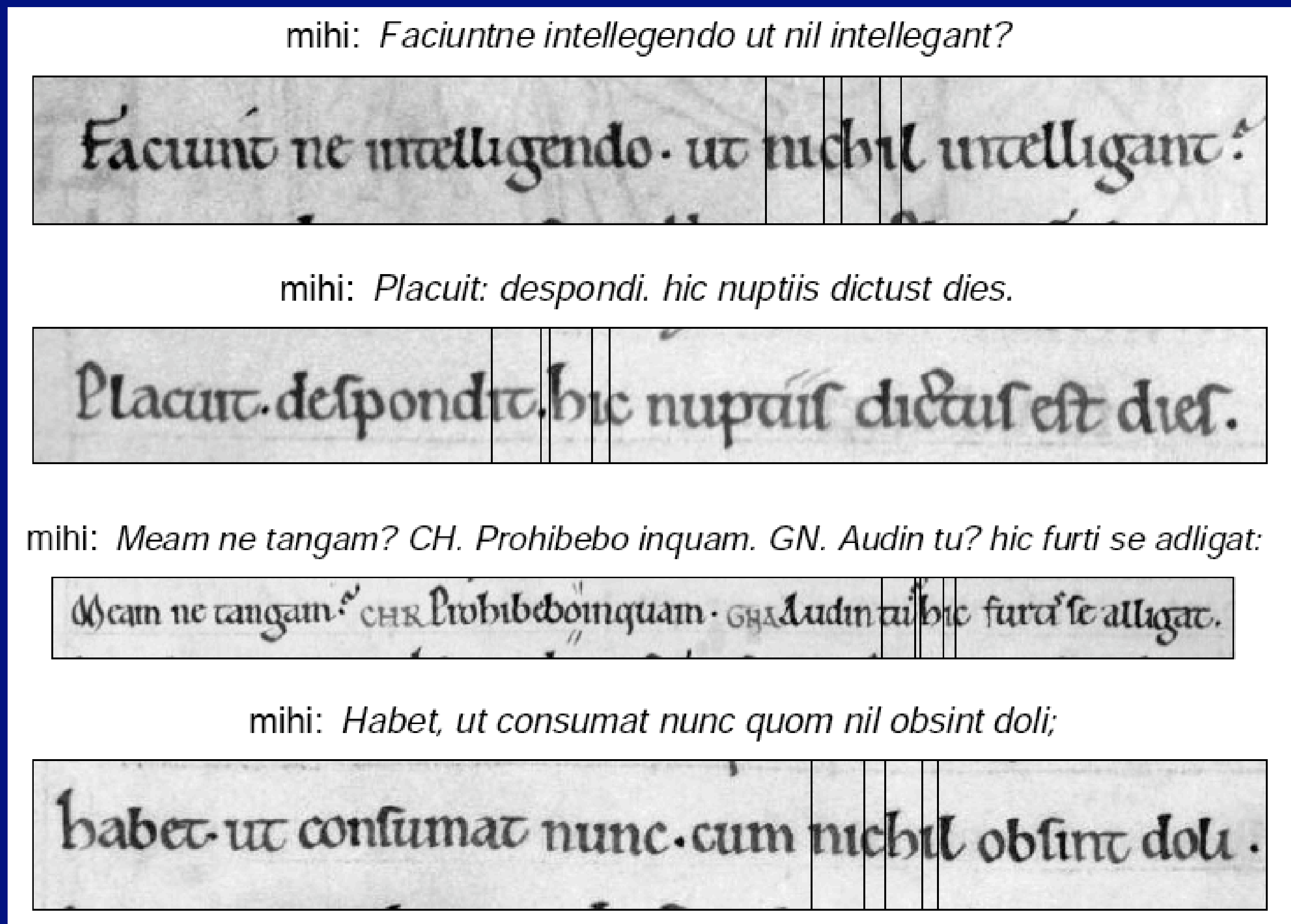
Line Based retrieval example



Strategy

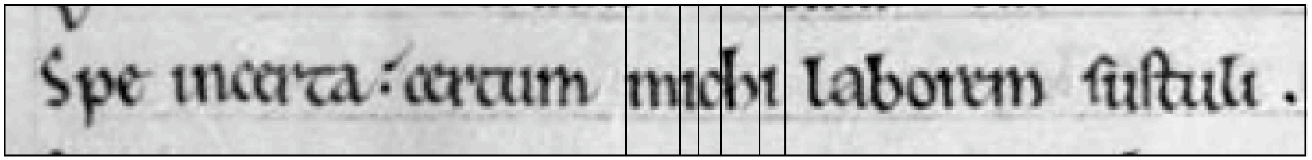
- Handwriting=substitution cipher

- Find lines with elementary methods
- Use vertical bars to quantize direction along lines
- Model text with generalized hidden markov model
  - hidden states can emit several tokens
  - accommodates templates of variable width
  - DP still applies
  - Dynamics from electronic latin
- Use letters
  - should do better, but
  - one example glyph per letter -- TOTAL 22 example glyphs
- Should be unsupervised
  - letters look like themselves

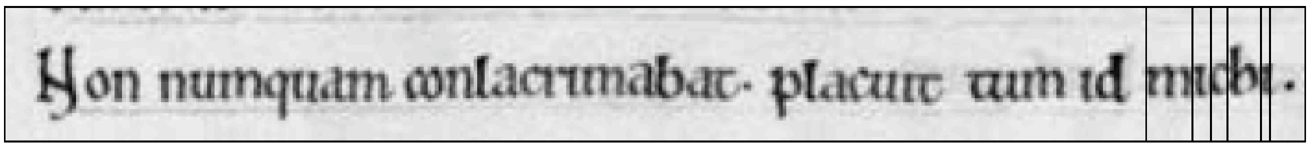


Edwards et al, NIPS, ‘04

michi: *Spe incerta certum mihi laborem sustuli,*

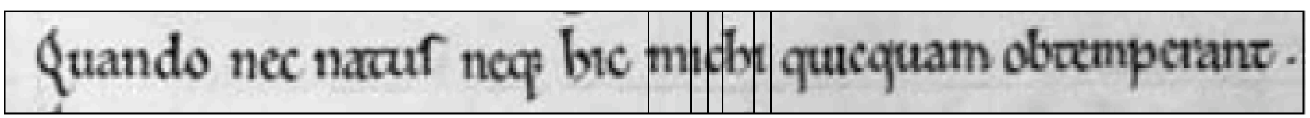


michi: *Nonnumquam conlacrumabat. placuit tum id mihi.*



michi: *Sto exspectans siquid mi imperent. venit una, "heus tu" inquit "Dore,*

michi: *Quando nec gnatu' neque hic mi quicquam obtemperant,*



Edwards et al, NIPS, '04

More to do

- Comparing models
  - Voluminous data on different models in JMLR paper (Barnard et al., 03)
  - More recently, Blei and Jordan's correspondence LDA (Blei Jordan 03)
- Image representation
  - e.g. point feature based models
- Vocabulary management
  - fuse visually equivalent words (train=locomotive)
- The effects of supervision
  - funny problems caused by near symmetries in likelihood (mare, grass)
  - small inputs should give very large outputs
- words aren't independent
  - e.g. Li and Wang, 03

Exploiting complementary information

- A probability model linking images and annotations
  - exploit co-occurence
  - better estimates of “meaning” for clustering and browsing
  - soft search, auto illustration, auto annotation
- Predicting words from image regions
  - explicitly encode and infer correspondence
    - aligned bitext
    - no alignment
  - rather like recognition
  - pinch techniques from statistical natural language processing
- Linking face images with names
  - an important special case
  - datasets of an epic scale available
  - like face recognition, but easier
  - breaking correspondence by clustering

News dataset

- Approx 5e5 news images, with captions
  - Easily collected by script from Yahoo over the last 18 months or so
- Mainly people
  - politicians, actors, sportsplayers
  - long, long tails distribution
- Face pictures captured “in the wild”
- Correspondence problem
  - some images have many (resp. few) faces, few (resp. many) names (cf. Srihari 95)

President George W. Bush makes a statement in the Rose Garden while Secretary of Defense Donald Rumsfeld looks on, July 23, 2003. Rumsfeld said the United States would release graphic photographs of the dead sons of Saddam Hussein to prove they were killed by American troops. Photo by Larry Downing/Reuters

Data examples

Doctor Nikola shows a fork that was removed from an Israeli woman who swallowed it while trying to catch a bug that flew in to her mouth, in Poriah Hospital northern Israel July 10, 2003. Doctors performed emergency surgery and removed the fork. (Reuters)

President George W. Bush waves as he leaves the White House for a day trip to North Carolina, July 25, 2002. A White House spokesman said that Bush would be compelled to veto Senate legislation creating a new department of homeland security unless changes are made. (Kevin Lamarque/Reuters)

Process

- Extract proper names
  - rather crudely, at present
- Detect faces
  - with Cordelia Schmid's face detector, (Vogelhuber Schmid 00)
- Rectify faces
  - by finding eye, nose, mouth patches, affine transformation
- Kernel PCA rectified faces
- Estimate linear discriminants
- Now have (face vector; name_1,...., name_k)

Scale

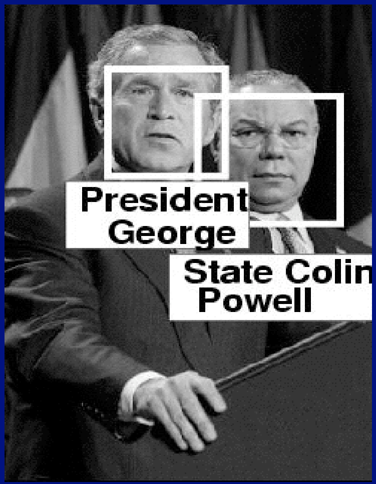
44773 big face responses

34623 properly rectified

27742 for k<=4

Building a face dictionary

- Compute linear discriminants
  - using single name, single face data items
  - we now have a set of clusters
- Now break correspondence with modified k-means
  - assign face to cluster with closest center,
    - chosen from associated names
  - recompute centers, iterate
  - using distance in LD space
- Now recompute discriminants, recluster with modified k-means



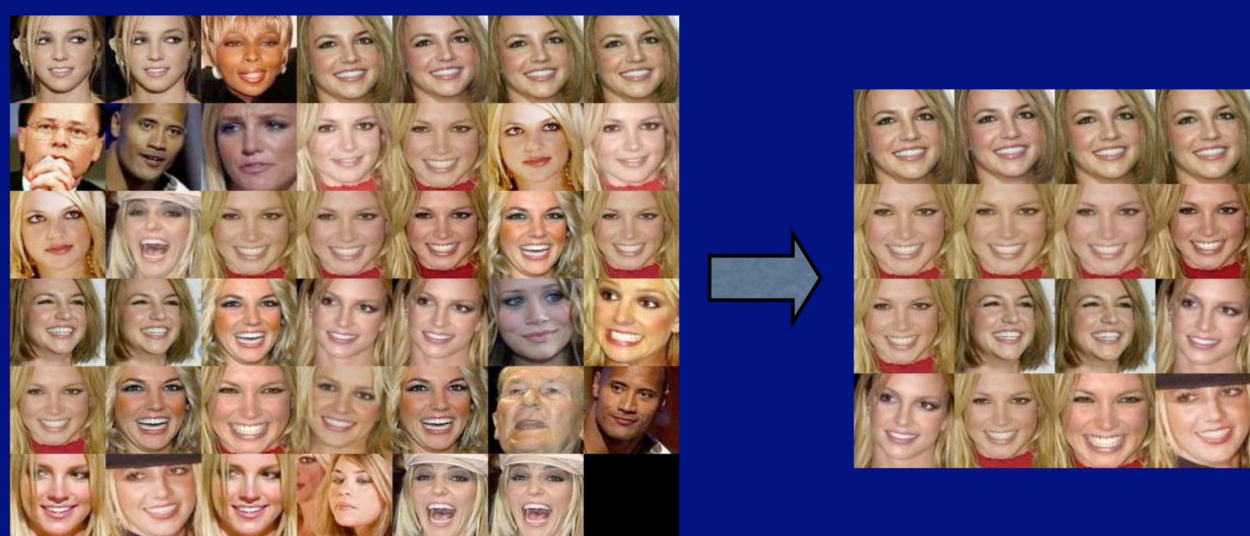
US President George W. Bush (L) makes remarks while Secretary of State Colin Powell (R) listens before signing the US Leadership Against HIV /AIDS , Tuberculosis and Malaria Act of 2003 at the Department of State in Washington, DC. The five-year plan is designed to help prevent and treat AIDS, especially in more than a dozen African and Caribbean nations(AFP/ Luke Frazza)

German supermodel Claudia Schiffer gave birth to a baby boy by Caesarian section January 30, 2003, her spokeswoman said. The baby is the first child for both Schiffer, 32, and her husband, British film producer Matthew Vaughn, who was at her side for the birth. Schiffer is seen on the German television show 'Bet It...?!' ('Wetten Dass...?!') in Braunschweig, on January 26, 2002. (Alexandra Winkler/Reuters)

British director Sam Mendes and his partner actress Kate Winslet arrive at the London premiere of 'The Road to Perdition', September 18, 2002. The films stars Tom Hanks as a Chicago hit man who has a separate family life and co-stars Paul Newman and Jude Law. REUTERS/Dan Chung

Pruning

- Using a likelihood model
- Tradeoff: size vs accuracy



Merging

Venezuelan
President Chavez

Hugo Chavez

Ryan's clean demo http://www.eecs.berkeley.edu/~ryanw/clustersFull_theta15/index.html

Tamara's demo http://www.cs.berkeley.edu/~millert/faces/faceDict/starClust/

How well does it work?

- Draw a cluster from the list, and an image from that cluster
  - frequency that that image is of someone else

#Images	#Clusters	error rate
19355	2357	26%
7901	1510	11%
4545	765	5.2%
3920	725	7.5%
2417	328	6.6%

- How many bits are required to fix result?

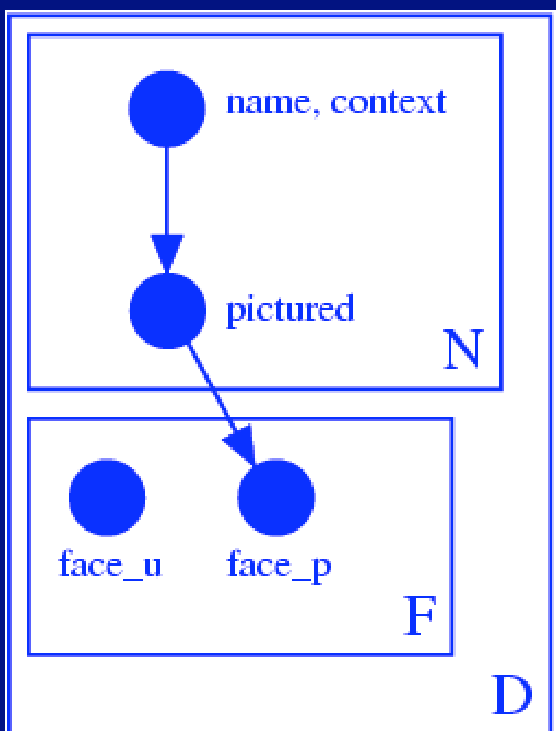
Works - but

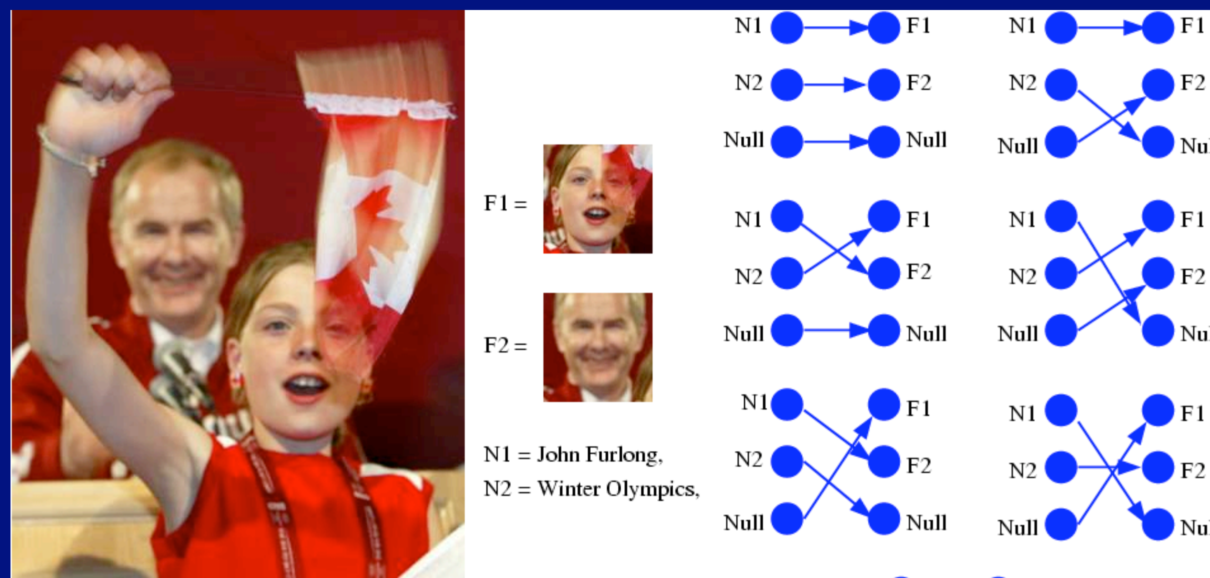
- We are missing language cues

Sahar Aziz, left, a law student at the University of Texas, hands the business card identifying Department of the Army special agent Jason D. Treesh to one of her attorneys, Bill Allison, right, during a news conference on Friday, Feb. 13, 2004, in Austin, Texas. In the background is Jim Harrington, director of the Texas Civil Rights Project. (AP Photo Harry Cabluck)

Training a language module

- Idea:
  - a set of named faces is supervised training data for a “who’s in the picture” module
  - actually, do EM (or maximize?) over missing correspondences



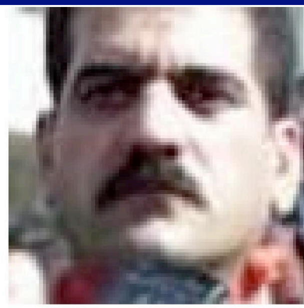


President and Chief Operating Officer of the Vancouver, British Columbia 2010 Bid Corporation John Furlong (rear) smiles while celebrating with compatriots their victory in obtaining the 2010 Winter Olympics bid on late July 2, 2003 in Prague. Vancouver won with 56 votes against 53 votes for Pyeonchang in the second round of balloting at an IOC gathering in Prague. REUTERS/Petr Josek

All possible name/face assignments

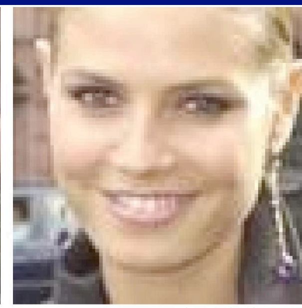
Language improves naming,

before – CEO Summit
after – Martha Stewart

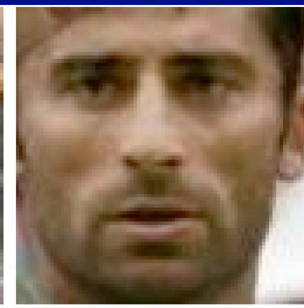


before – U.S. Joint
after – Null

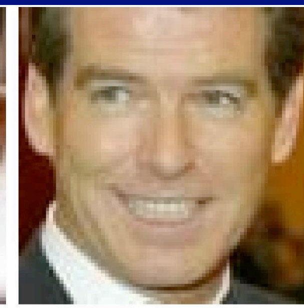
before – Angelina Jolie
after – Jon Voight



before – Ric Pipino
after – Heidi Klum

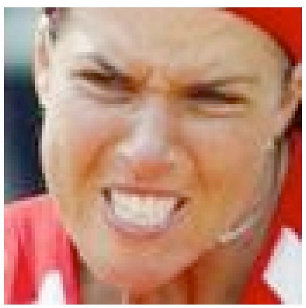


before – U.S. Open
after – David Nalbandian



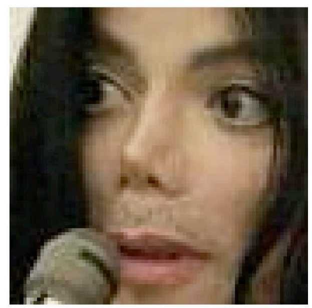
before – James Bond
after – Pierce Brosnan

before – U.S. House
after – Andrew Fastow

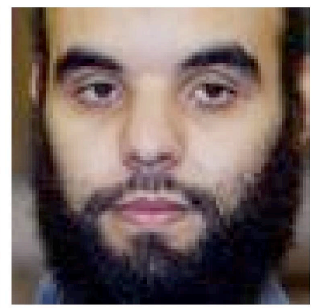


before – Julia Vakulenko
after – Jennifer Capriati

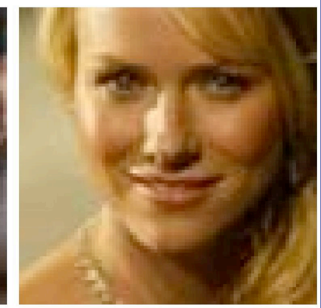
before – Vice President Dick Cheney
after – President George W.



before – Marcel Avram
after – Michael Jackson



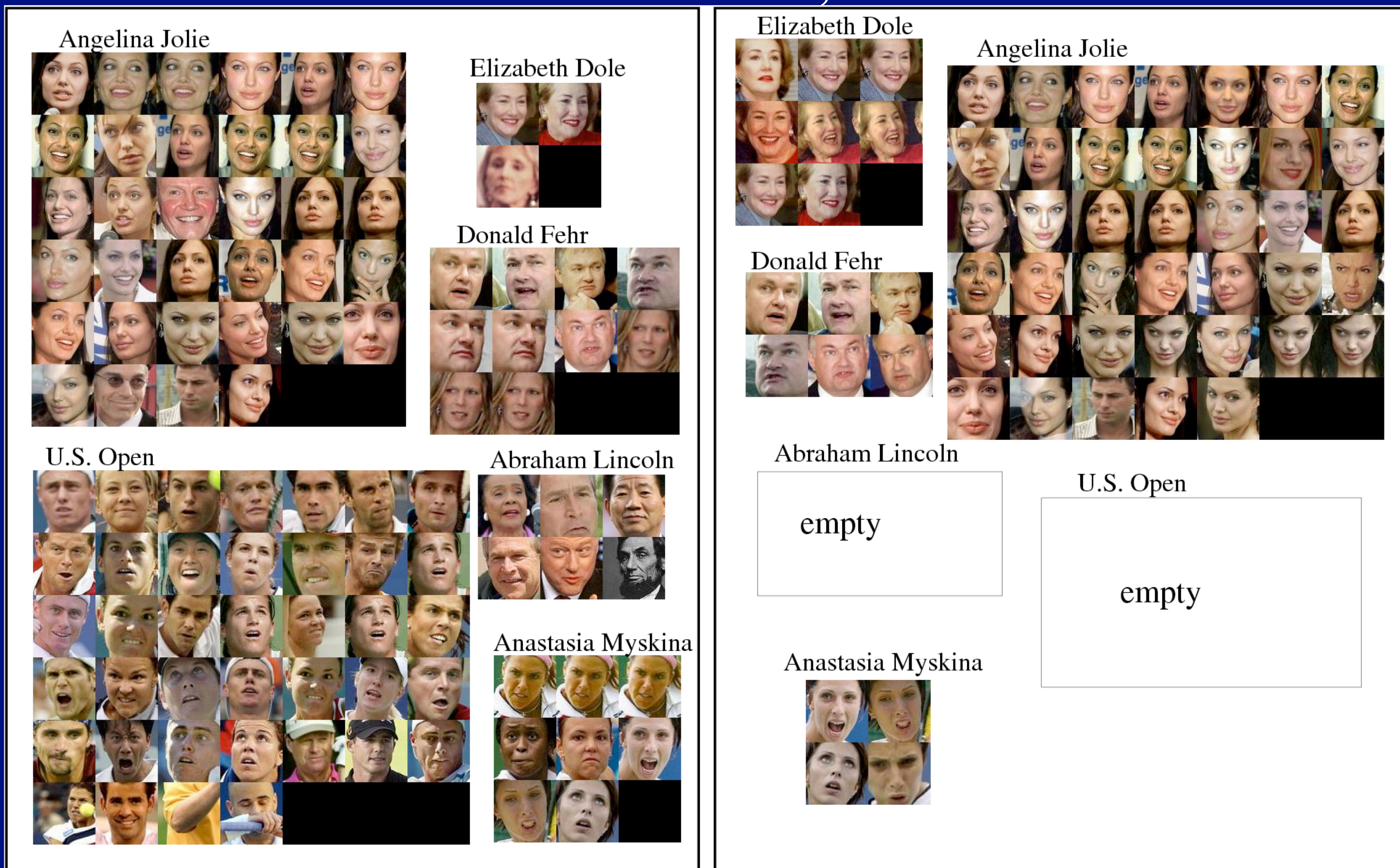
before – al Qaeda
after – Null



before – James Ivory
after – Naomi Watts

Model	EM	MM
Appearance Model, No Lang Model	56%	67%
Appearance Model + Lang Model	72%	77%

Clusters,



Without language model

With language model

and yields a useful little NLP module, too

IN Pete Sampras IN of the U.S. celebrates his victory over Denmark's **OUT Kristian Pless OUT** at the **OUT U.S. Open OUT** at Flushing Meadows August 30, 2002. Sampras won the match 6-3 7- 5 6-4. REUTERS/Kevin Lamarque
Germany's **IN Chancellor Gerhard Schroeder IN**, left, in discussion with France's **IN President Jacques Chirac IN** on the second day of the EU summit at the European Council headquarters in Brussels, Friday Oct. 25, 2002. EU leaders are to close a deal Friday on finalizing entry talks with 10 candidate countries after a surprise breakthrough agreement on Thursday between France and Germany regarding farm spending.(AP Photo/European Commission/HO)
'The Right Stuff' cast members **IN Pamela Reed IN**, (L) poses with fellow cast member **IN Veronica Cartwright IN** at the 20th anniversary of the film in Hollywood, June 9, 2003. The women played wives of astronauts in the film about early United States test pilots and the space program. The film directed by **OUT Philip Kaufman OUT**, is celebrating its 20th anniversary and is being released on DVD. REUTERS/Fred Prouser
Kraft Foods Inc., the largest U.S. food company, on July 1, 2003 said it would take steps, like capping portion sizes and providing more nutrition information, as it and other companies face growing concern and even lawsuits due to rising obesity rates. In May of this year, San Francisco attorney **OUT Stephen Joseph OUT**, shown above, sought to ban Oreo cookies in California – a suit that was withdrawn less than two weeks later. Photo by Tim Wimborne/Reuters REUTERS/Tim Wimborne

Classifier	labels correct	IN correct	OUT correct
Baseline	67%	100%	0%
EM Labeling with Language Model	76%	95%	56%
MM Labeling with Language Model	84%	87%	76%

Faces - To do

- Better image features
- More sophisticated probability model, EM
- Estimate P (no pic | name) using EM
- Better named entity recognition
- Co-reference resolution (across languages?) using faces
- Use non-parametric face model (animation?)
- Start looking at face recognition

Partially supervised data == Missing correspondence

- Supervised data, but with a little bit missing
  - There's not all that much unsupervised data but lots of semi-supervised
- Linking and association
  - picture is labelled, but object not segmented
    - Faces (Leung, Burl, Perona, 95); Faces and cars (Weber Perona 01); Faces,cars,motorbikes,planes,tigers (Fergus Zisserman Perona 03); Animal pix (Schmid 01); Clustering (Barnard et al, 01, 01); word prediction (Barnard et al 03; Wang et al, 02; Lia et al, 03;); album cover-music (Brochu et al; 02); objects (Duygulu et al, 02; Barnard et al 03); names and faces (Miller et al 04); speech and pictures (Fleck et al, 04 patent).
  - Words, metadata should be linked to picture
    - Face pix (Srihari, 95); Corel (Barnard et al 01; Li+Wang 03); Art (Barnard et al. 01);
- Coherence
  - Objects of interest look coherent from frame to frame in video
    - People tracking (Ramanan+Forsyth '03); Animals (Ramanan+Forsyth '03)
  - Picture posesses noisy label; which labels are right?
    - Image search results (Fergus et al 04)
- missing data tends to be correspondence

Conclusions

- There’s more data out there about the visual world than immediately meets the eye
- Visual information should be linked with other forms of information
  - so one can work where it’s easiest
- Doing so may yield useful artifacts and insights