Joint Visual-Text Modeling for Multimedia Retrieval

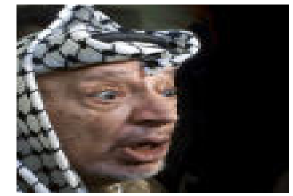
JHU CLSP Workshop 2004 – Final Presentation, August 17 2004

Team

- Undergraduate Students
  - Desislava Petkova (Mt. Holyoke), Matthew Krause (Georgetown)
- Graduate Students
  - Shaolei Feng (U. Mass), Brock Pytlik(JHU), Paola Virga (JHU)
- Senior Researchers
  - Pinar Duygulu, Bilkent U., Turkey
  - Pavel Ircing (U. West Bohemia)
  - Giri Iyengar, IBM Research
  - Sanjeev Khudanpur, CLSP, JHU
  - Dietrich Klakow, Uni. Saarland
  - R. Manmatha, CIIR, U. Mass Amherst
  - Harriet Nock, IBM Research (external participant)

Big Picture: Multimedia Retrieval Task

Find clips showing Yasser Arafat → **Multimedia Retrieval System** →



" … Palestinian leader Yasser Arafat today said …"

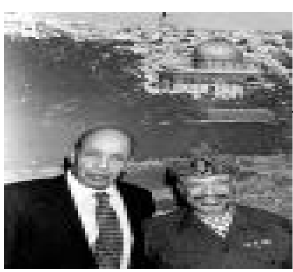
VIDEO CLIPS

Most research has addressed:

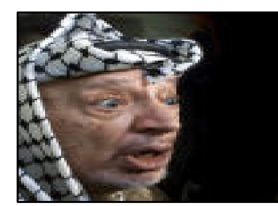
I. Text queries, text (or degraded text) documents

II. Image queries, image data

Yasser Arafat → Process Query Text → Spoken Document Retrieval → Combine Scores



→ Process Query Image → Image Content-based Retrieval → Combine Scores



" … Palestinian leader Yasser Arafat today said …"

Joint-Visual Text Models!

Joint Visual-Text Modeling

Document of words → Query of words

Retrieve documents using p(Document|Query)

Document of Words and Visterms → Query of Words and Visterms

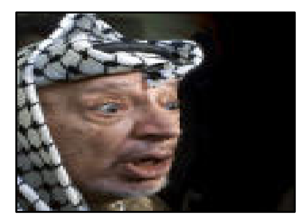
Retrieve documents using p(dw,dv | qw,qv)

Yasser Arafat → Process Query Text

→ Process Query Image

Process Query Text, Process Query Image → Joint word-visterm retrieval →

VIDEO CLIPS



" … [Yes sir, you're fat today said]…

Joint Visual-Text Modeling: KEY GOAL

- Show that joint visual-text modeling improves multimedia retrieval
  - Demonstrate and Evaluate performance of these models on TRECVID2003 corpus and task

Key Steps

- Automatically annotate video with concepts (meta-data)
  - E.g. Video contains a face, in a studio-environment …
- Retrieve video
  - Given a query, select suitable meta-data for the query and retrieve
  - Combine with text-retrieval in a unified Language Model-based IR setting

TRECVID Corpus and Task

- Corpus
  - Broadcast news videos used for Hub4 evaluations (ABC, CNN, CSPAN)
  - 120 Hours of video
- Tasks
  - Shot-boundary detection
  - News Story segmentation (multimodal)
  - Concept detection (Annotation)
  - Search task

Alternate (development) Corpus

- COREL photograph database
  - 5000 high-quality photographs with captions
- Task
  - Annotation

TRECVID Search task definition

NIST

Evaluation

Statement of Information need + Examples

Manual Selection of System Parameters

Ranked list of video shots

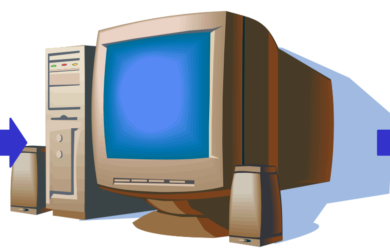
Manual
Interactive

Our search task definition

NIST

Evaluation

Statement of Information need + Examples



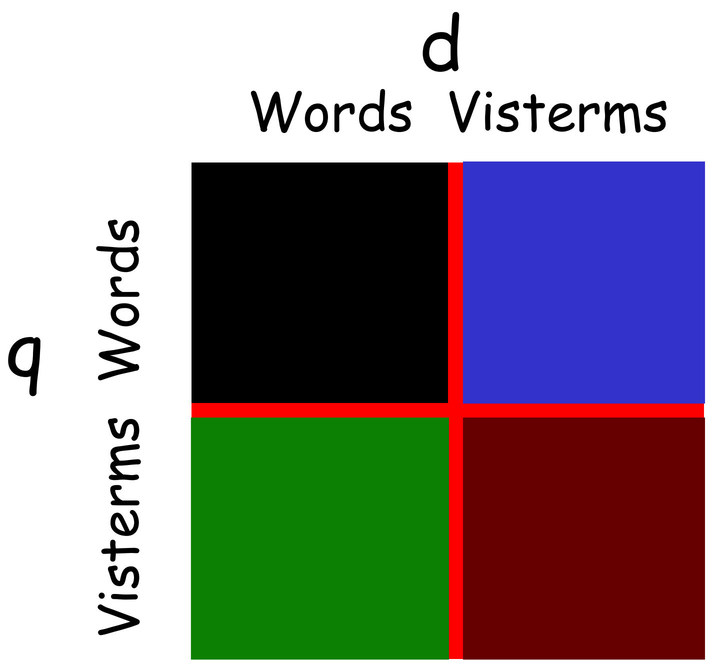
Automatic Selection of System Parameters

Ranked list of video shots

Isolate Algorithmic issues from interface and user issues

Language Model based Retrieval

Rank documents with p(qw,qv|dw,dv)



Baseline model

Relating document visterms to query words (MT,Relevance Model,HMMs)

Relating document words to query images (Text Classification experiments)

Visual-only retrieval models

Evaluation

- Concept annotation performance
  - Compare against manual ground truth
- Retrieval task performance
  - Compare against NIST relevance judgements
- Both measured using Mean Average Precision (mAP)

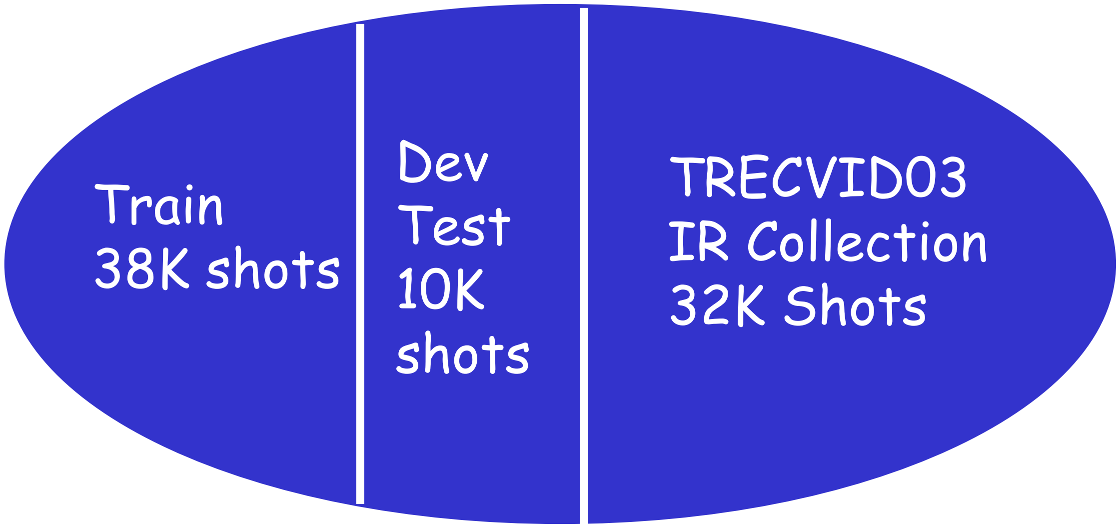
Mean Average Precision (mAP)

$$S(t) = \sum_{i \in \{relevant\}} precision(i)$$

$$AP(t) = \frac{S(t)}{|rel(t)|}$$

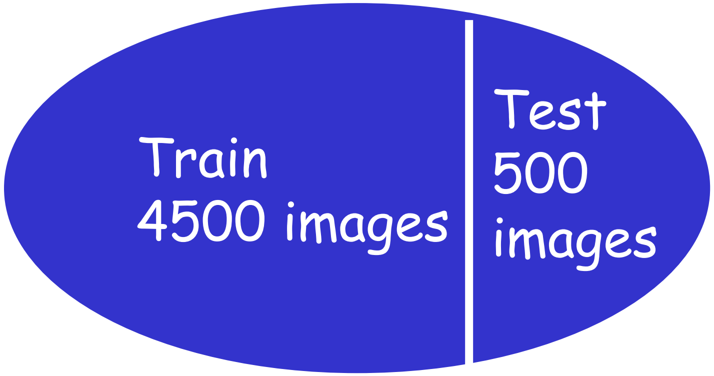
$$mAP = \frac{\sum_{t \in T} AP(t)}{|T|}$$

Experimental Setup: Corpora



TRECVID03 Corpus
120 Hours
Ground Truth on Dev data

COREL Corpus
5000 images



Experimental Setup: Visual Features

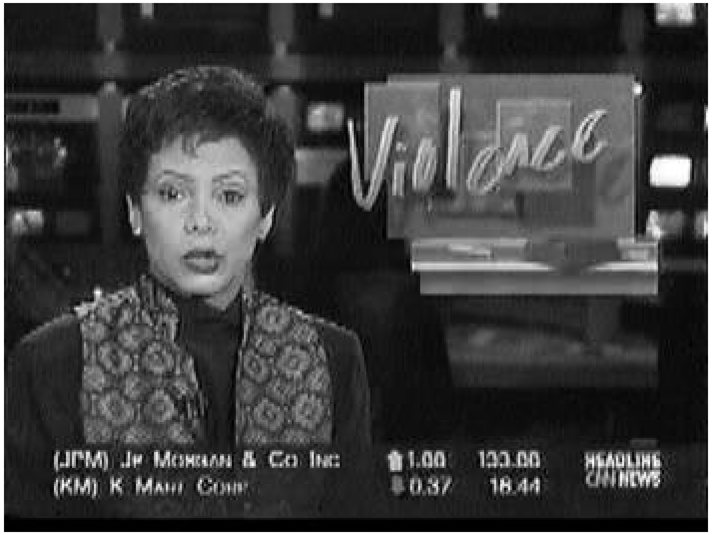
Original

L*a*b

Edge Strength

Co-occurrence

Interest Point Neighborhoods (Harris detector)



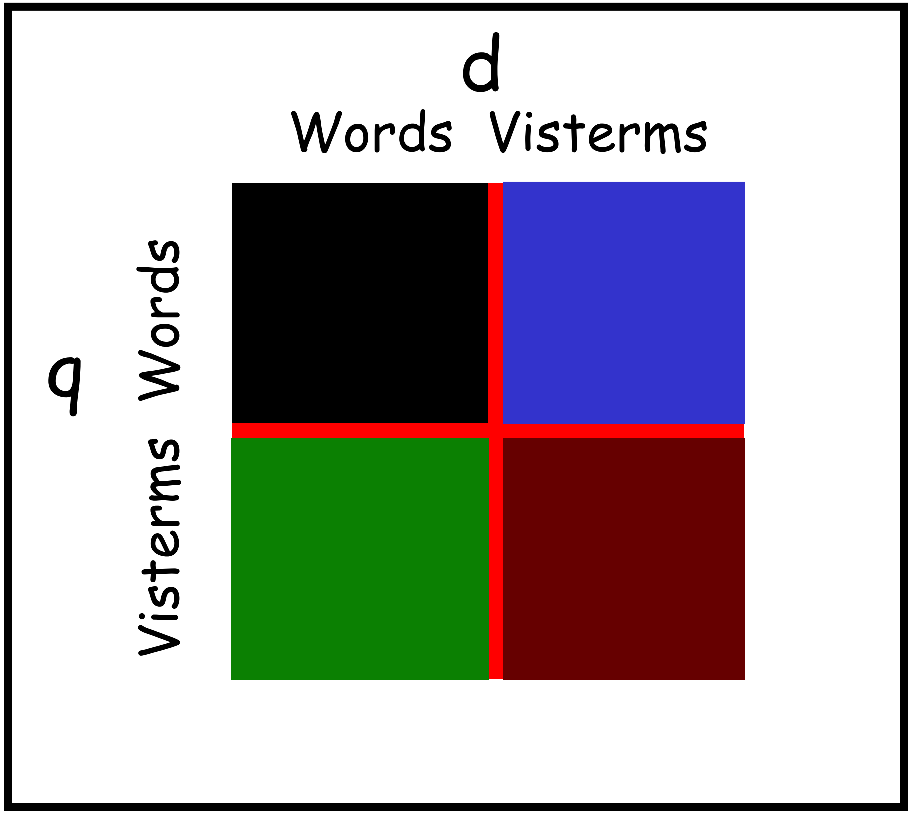
Greyscale image

Interest points

Experimental Setup: Visual Feature list

- Regular partition
  - L*a*b Moments (COLOR)
  - Smoothed Edge Orientation Histogram (EDGE)
  - Grey-level Co-occurrence matrix (TEXTURE)
- Interest Point neighborhood
  - COLOR, EDGE, TEXTURE

Presentation Outline



Translation (MT) models
(Paola),

Relevance Models
(Shao Lei,Desislava),

Graphical Models
(Pavel, Brock)

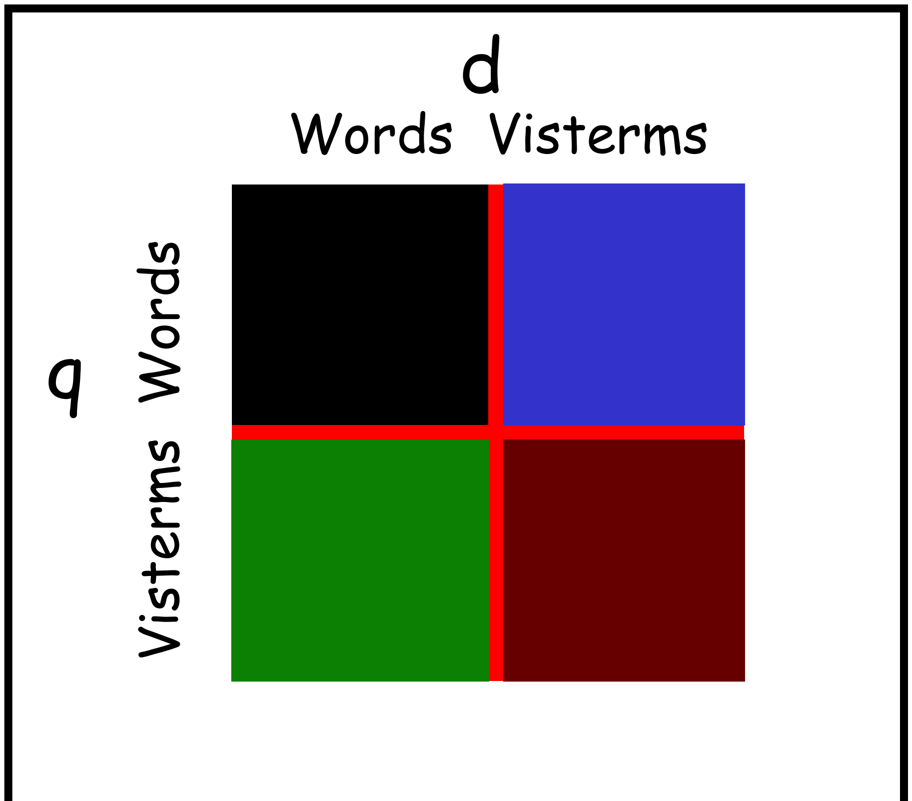
Text classification models
(Matt)

Integration & Summary
(Dietrich)

A Machine Translation Approach to Image Annotation

Presented by Paola Virga

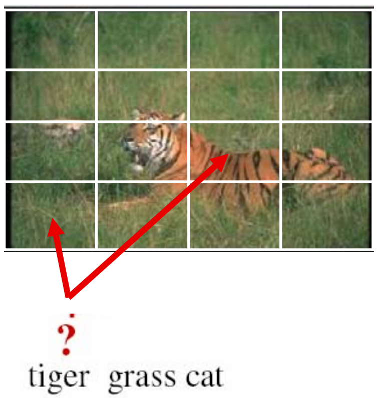
Presentation Outline

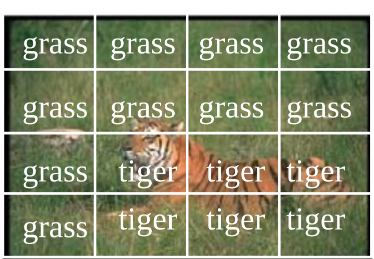


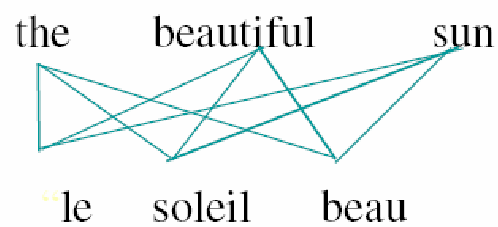
Translation (MT) models

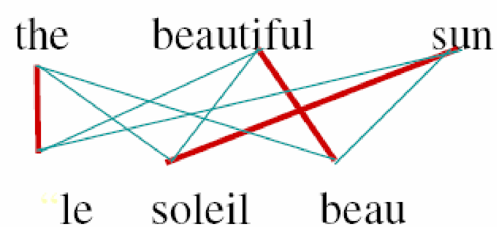
$$p(q_w \mid d_V) = \sum_c p(q_w \mid c)\boxed{p(c \mid d_V)}$$

Inspiration from Machine Translation









$$p(f|e) = \sum_{a} p(f,a|e)$$

$$p(c|v) = \sum_{a} p(c,a|v)$$

Direct translation model

Discrete Representation of Image Regions (visterms) to create analogy to MT

In Machine Translation → discrete tokens

In our task

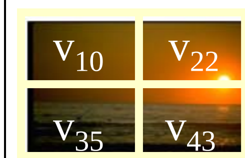
sun sky waves sea

concepts ✓

However, the features extracted from regions are continuous

Solution : Vector quantization → visterms ✓

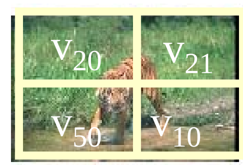
$\{f_{n1}, f_{n2}, \ldots f_{nm}\} \rightarrow v_k$



sun sky sea waves

$v_{10}\ v_{22}\ v_{35}\ v_{43}$

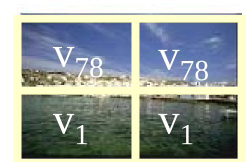
$c_5\ c_1\ c_{38}\ c_{71}$



tiger water grass

$v_{20}\ v_{21}\ v_{50}\ v_{10}$

$c_{15}\ c_{21}\ c_{83}$

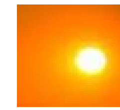


water harbor sky clouds sea

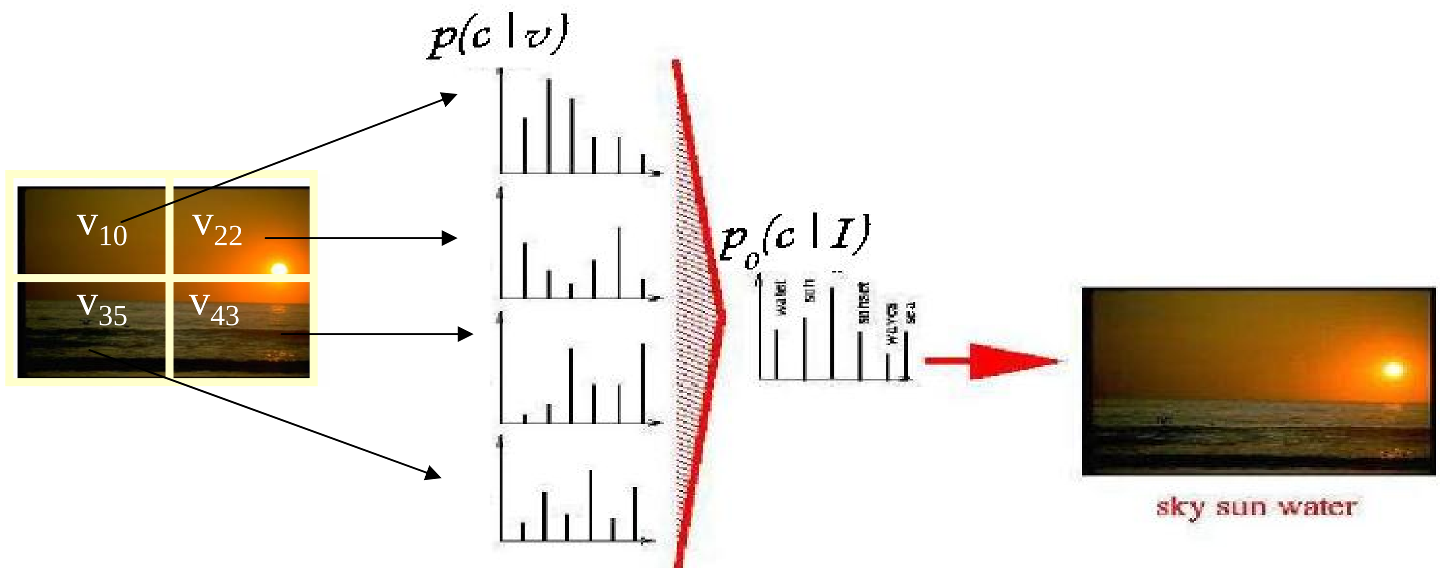
$v_{78}\ v_{78}\ v_1\ v_1$

$c_{21}\ c_{19}\ c_1\ c_{56}\ c_{38}$

Image annotation using translation probabilities

p(c|v) : Probabilities obtained from direct translation   p (sun | )

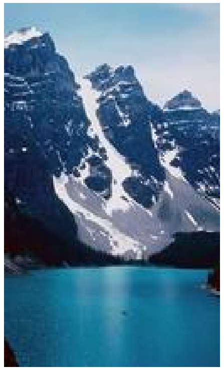
$$P_0(c \mid d_V) = \frac{1}{|d_V|} \sum_{v \in d_V} P(c \mid v)$$



Annotation Results (Corel set)

field foals horses mare
tree horses foals mare field

flowers leaf petals stems
flowers leaf petals grass tulip



mountain sky snow water
sky mountain water clouds snow

people pool swimmers water
swimmers pool people water sky

jet plane sky
sky plane jet tree clouds

people sand sky water
sky water beach people hills

Top: manual annotations, bottom : predicted words (top 5 words with the highest probability)
Red : correct matches

Feature selection

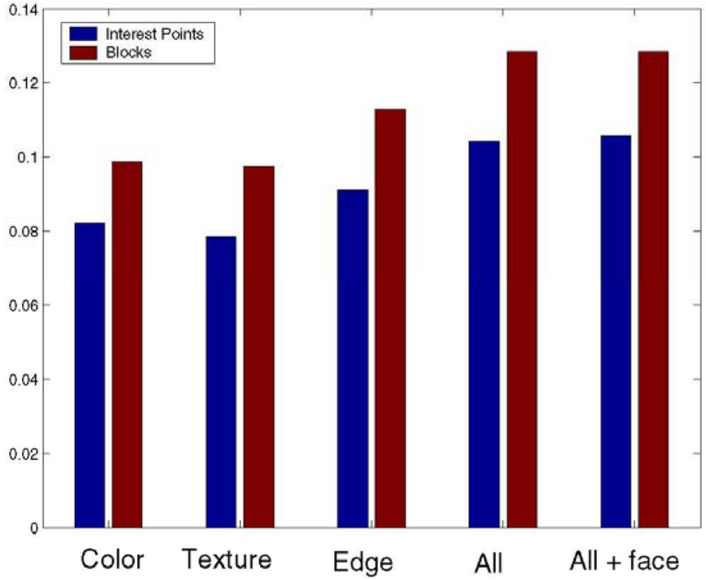
Features : color, texture, edge

Extracted from blocks, or around interest points

Observations

- Features extracted from blocks give better performance than features extracted around interest points
- When the features are used individually

  Edge features give the best performance
- Training using all is the best
  - Using Information Gain to select visterms vocabulary didn't help
- Integrating number of faces, increases the performance slightly



mAP values for different features

Model and iteration selection

Strategies compared

(a) IBM Model 1 $p(c|v)$

(b) HMM on top of (a)

(c) IBM Model 4 on top of (b)

-> Observation : IBM Model 1 is the best

Corel	TREC
0.125	0.124

Number of iterations in Giza training affects the performance

-> Less iterations give better annotation performance
but cannot produce rare words

Integrating word co-occurrences

- Model 1 with word co-occurrence

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

- Integrating word co-occurrences into the model helps for Corel but not for TREC

	Corel	TREC
Model 1	0.125	0.124
Model 1 + Word-CO	0.145	0.124

Inspiration from CLIR

- Treat Image Annotation as a Cross-lingual IR problem
  - Visual Document comprising visterms (target language) and a query comprising a concept (source language)

$$p(c \mid d_V) = \lambda \left(\sum_{v \in V} p(c \mid v)\, p(v \mid d_V) \right) + \underbrace{(1-\lambda)\, p(c \mid G_C)}_{\text{same } \forall d_V}$$

Inspiration from CLIR

- Treat Image Annotation as a Cross-lingual IR problem
  - Visual Document comprising visterms (target language) and a query comprising a concept (source language)

$$p(c \mid d_V) = \sum_{v \in d_V} p(v \mid d_V)\, p(c \mid v)$$

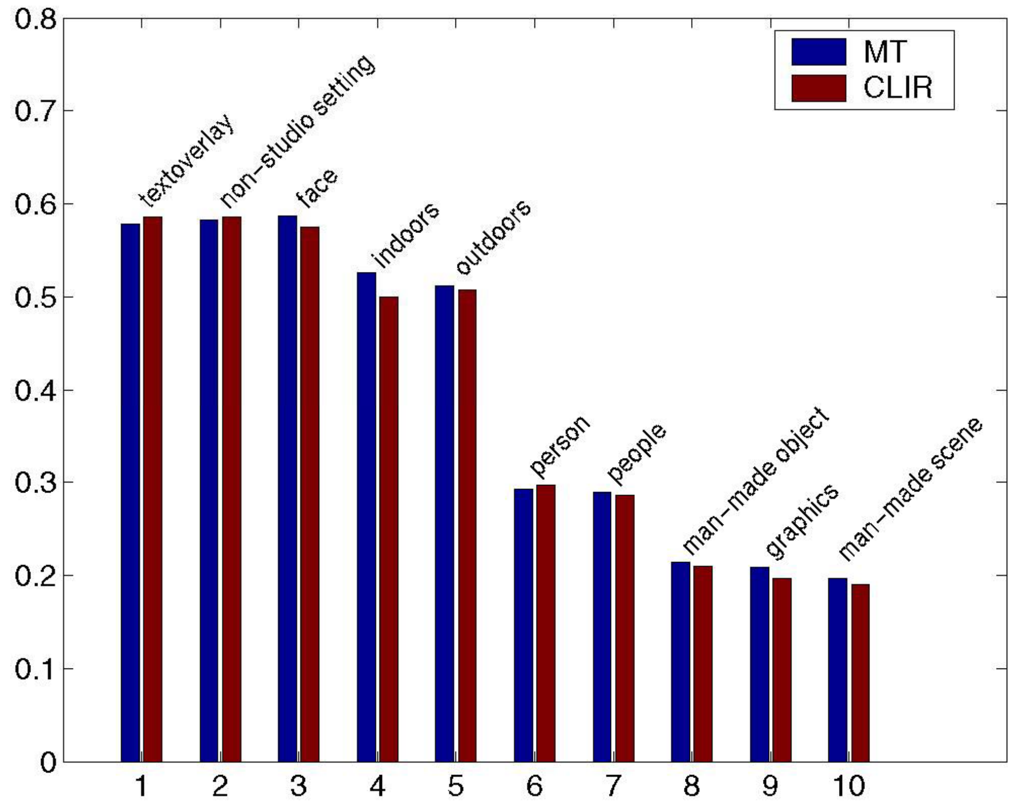
  - Image does not provide a good estimate of $p(v|d_v)$
  - Tried p(v) and DF(v), DF works best

$$score(c \mid d_V) = \sum_{v \in d_V} DF_{Train}(v)\, p(c \mid v)$$

Annotation Performance on TREC

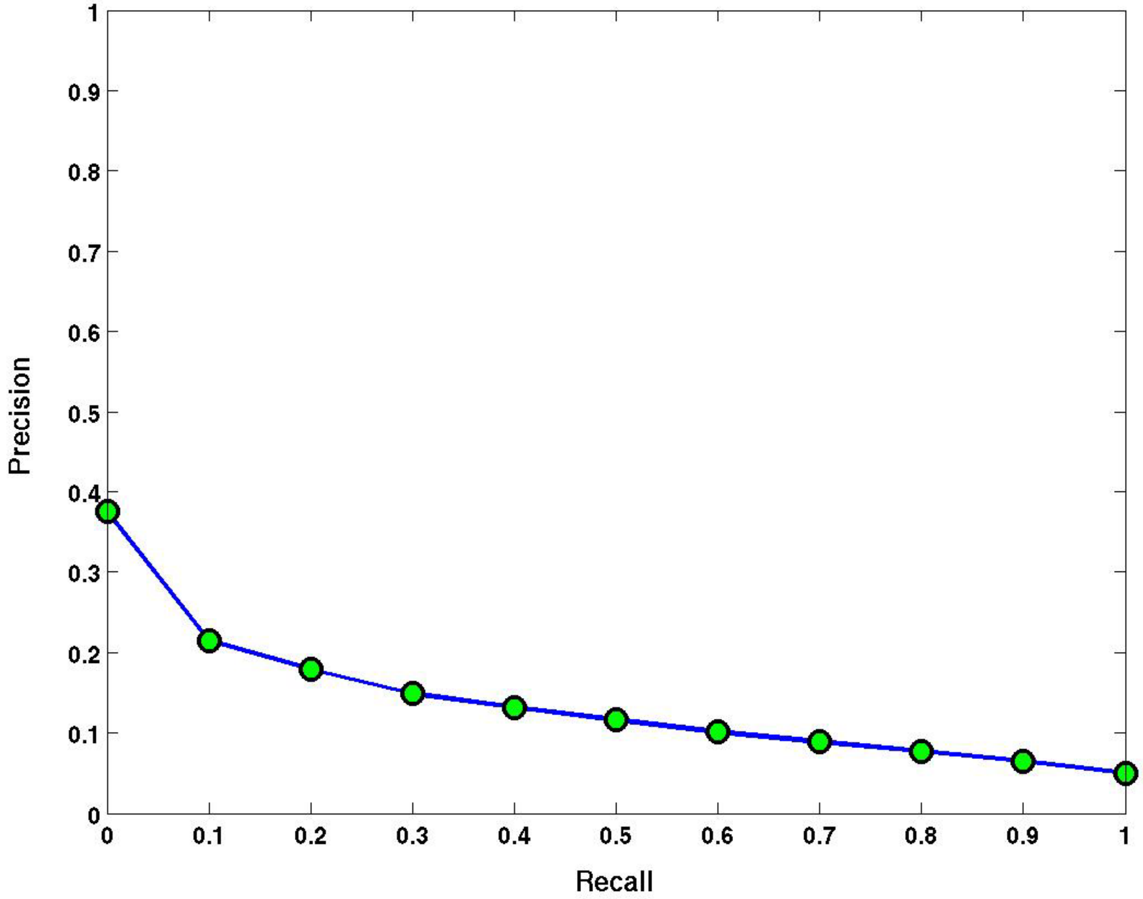
Model 1	0.124
CLIR using Model 1	0.126

Significant at p=0.04



Average Precision values for the top 10 words
For some concepts we achieved up to 0.6

Annotation Performance on TREC

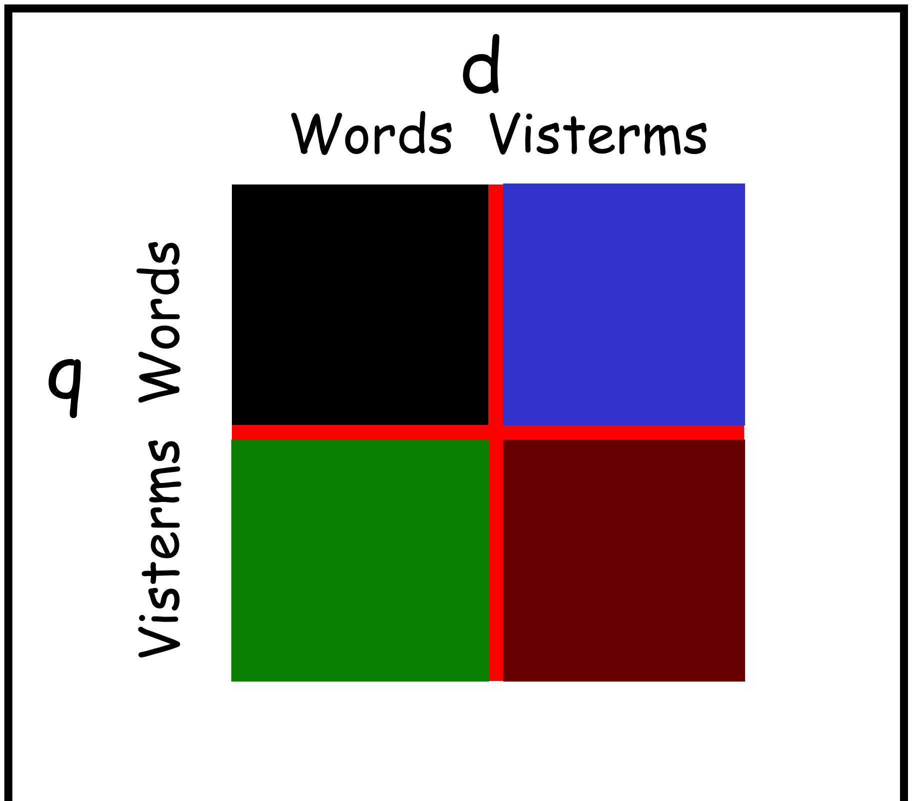


Questions?

Relevance Models for Image Annotation

Presented by Shaolei Feng
University of Massachusetts, Amherst

Relevance Models as Visual Model



$$p(q_w \mid d_v) = \sum_c p(q_w \mid c) \boxed{p(c \mid d_v)}$$

Goal:

Use Relevance Models to estimate the probabilities of concepts given test keyframes

Intuition

- Images are defined by spatial context.
  - Isolated pixels have no meaning.
  - Context simplifies recognition/retrieval.
  - E.g.Tiger is associated with grass, tree, water forest.
    - Less likely to be associated with computers.



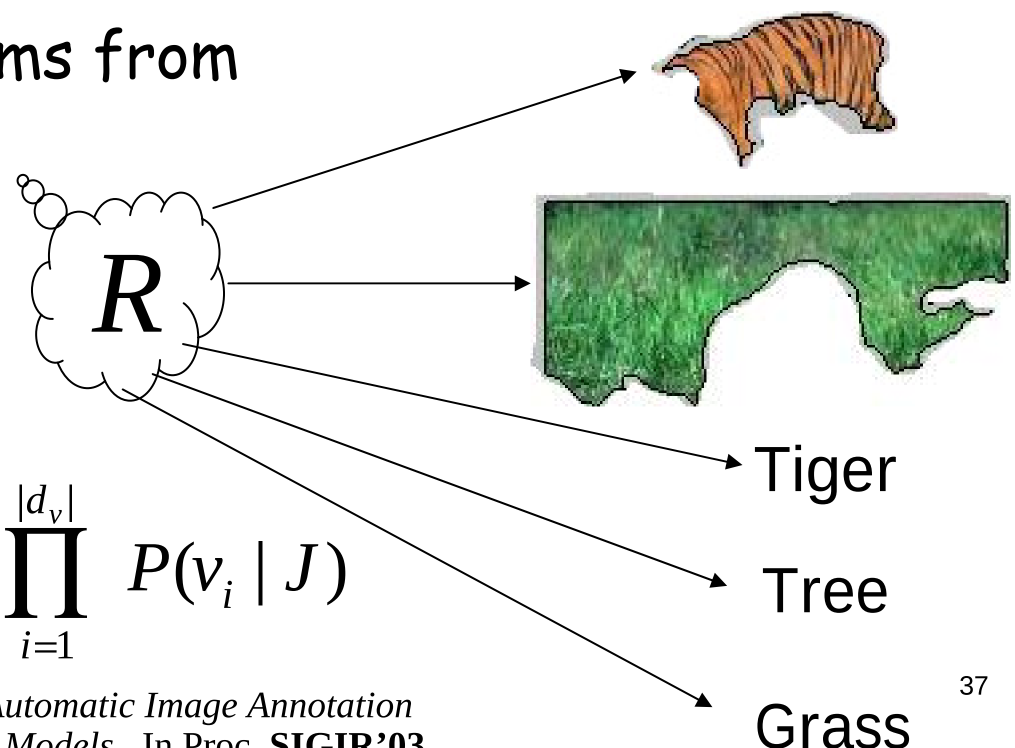

Introduction to Relevance Models

- Originally introduced for text retrieval and cross-lingual retrieval
  - Lavrenko and Croft 2001, Lavrenko, Choquette and Croft, 2002
  - A formal approach to query expansion.

- A nice way of introducing context in images
  - Without having to do this explicitly
  - Do this by computing the joint probability of images and words

Cross Media Relevance Models (CMRM)

- Two parallel vocabularies: Words and Visterms
- Analogous to Cross – lingual relevance models
- Estimate the joint probabilities of words and visterms from training images



$$P(c, d_v) = \sum_{J \in T} P(J) P(c \mid J) \prod_{i=1}^{|d_v|} P(v_i \mid J)$$

J. Jeon, V. Lavrenko and R. Manmatha, *Automatic Image Annotation and Relevance Using Cross-Media Relevance Models*, In Proc. **SIGIR'03.**

Continuous Relevance Models (CRM)

- A continuous version of Cross Media Relevance Model
- Estimate the P(v|J) using kernel density estimate

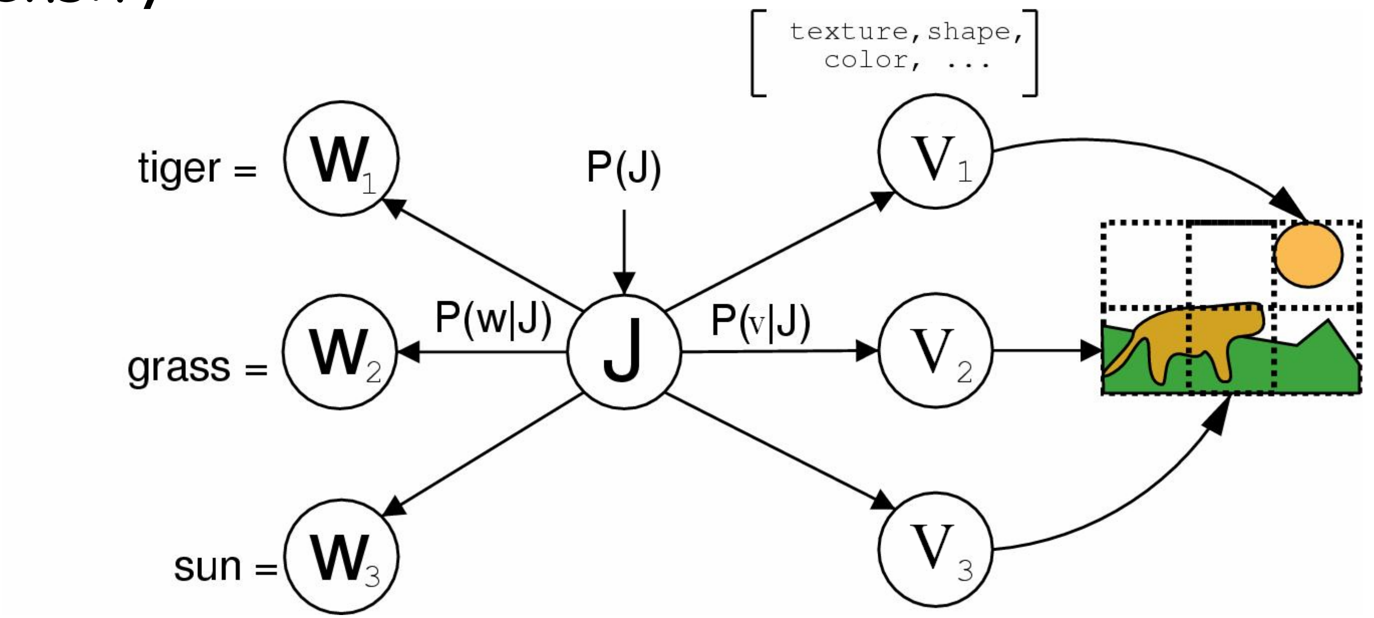
$$P(v \mid J) = \frac{1}{n} \sum_{i=1}^{|J|} K\left(\frac{\left\| v - v_{Ji} \right\|}{\beta} \right)$$

K: Gaussian Kernel

β: Bandwidth

Continuous Relevance Model

- A generative model
- Concept words w_j generated by an i.i.d. sample from a multinomial
- Visterms v_i generated by a multi-variate (Gaussian) density



Normalized Continuous Relevance Models

- Normalized CRM
  - Pad annotations to fixed length. Then use the CRM.
  - Similar to using a Bernoulli model (rather than a multinomial for words).
  - Accounts for length (similar to length of document in text retrieval).

S. L. Feng, V. Lavrenko and R. Manmatha, *Multiple Bernoulli Models for Image and Video Annotation*, in CVPR'04

V. Lavrenko, S. L. Feng and R. Manmatha, *Statistical Models for Automatic Video Annotation and Retrieval*, in ICASSP04

Annotation Performance

- On Corel data Set:

Models	CMRM	CRM	Normalized-CRM
Mean average Precision	0.14	0.23	0.26

- Normalized-CRM works best

Annotation Examples (Corel set)

Sky train railroad locomotive water

Cat tiger bengal tree forest

Snow fox arctic tails water

Tree plane zebra herd water

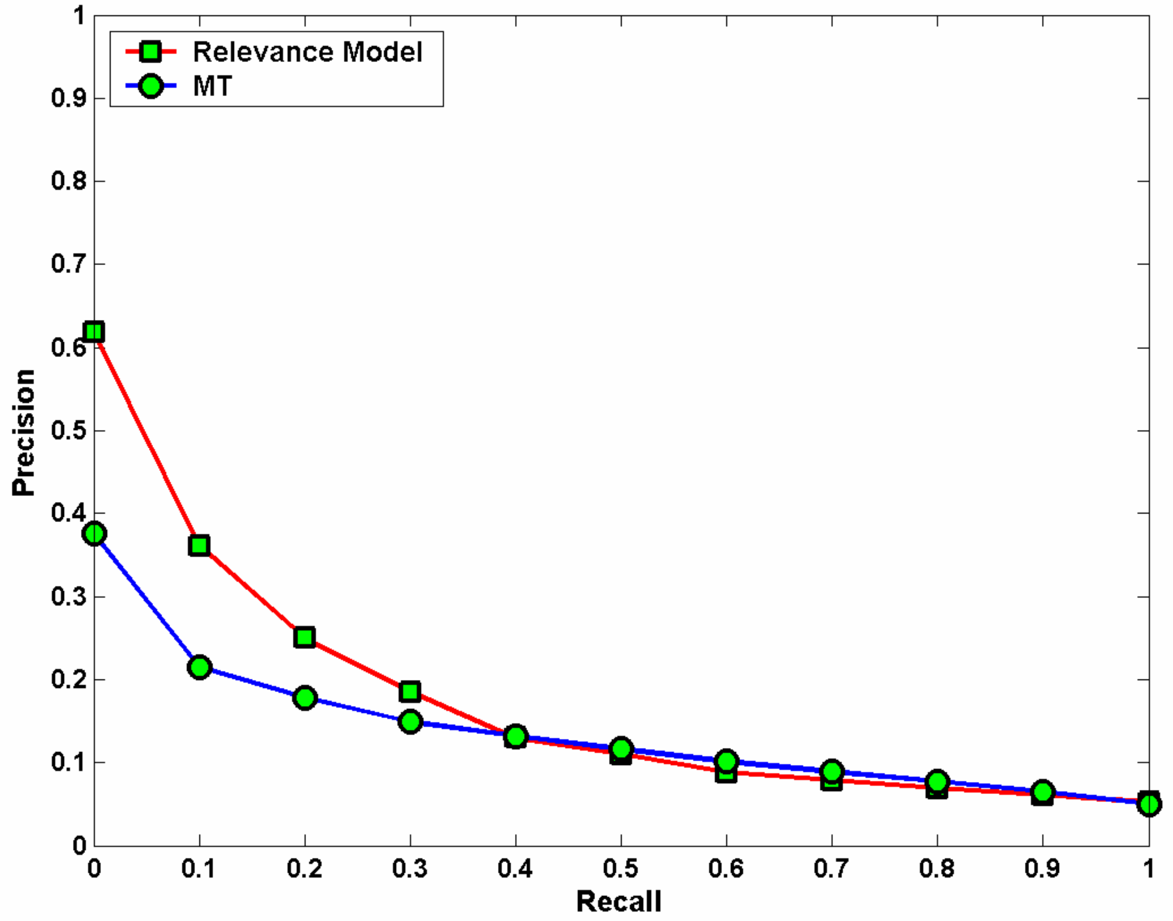
Birds leaf nest water sky

Mountain plane jet water sky

Results: Relevance Model on Trec Video Set

- Model: Normalized continuous relevance model
- Features: color and texture
  - Our comparison experiments show adding edge feature only get very slight improvement
- Evaluate annotation on the development dataset for annotation evaluation
  - mean average precision: 0.158

Annotation Performance on TREC



Proposal: Using Dynamic Information for Video Retrieval

Presented by Shaolei Feng
University of Massachusetts, Amherst

Motivation

- Current models based on single frames in each shot.
- But video is dynamic
  - Has motion information.
- Use dynamic (motion) information
  - Better image representations (segmentations)
  - Model events/actions

Why Dynamic Information

- Model actions/events
  - Many Trecvid 2003 queries require motion information. E.g.
    - find shots of an airplane ***taking off.***
    - find shots of a person ***diving into*** water.
  - Motion is an important cue for retrieving actions/events.
    - But using the optical flow over the entire image doesn't help.
    - Use motion features from objects.

- Better Image Representations
  - Much easier to segment moving objects from background than to segment static images.

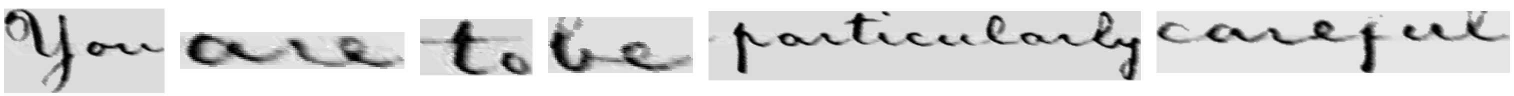
Problems with still images.

- Current approach
  - Retrieve videos using static frames.
- Feature representations
  - Visterms from keyframes.
  - Rectangular partition or static segmentation
    - Poorly correlated with objects.
  - Features – color, texture, edges.

- Problem: visterms not correlated well with concepts.

Better Visterms – better results.

- Model performs well on related tasks.
- Retrieval of handwritten manuscripts.
  - Visterms – word images.

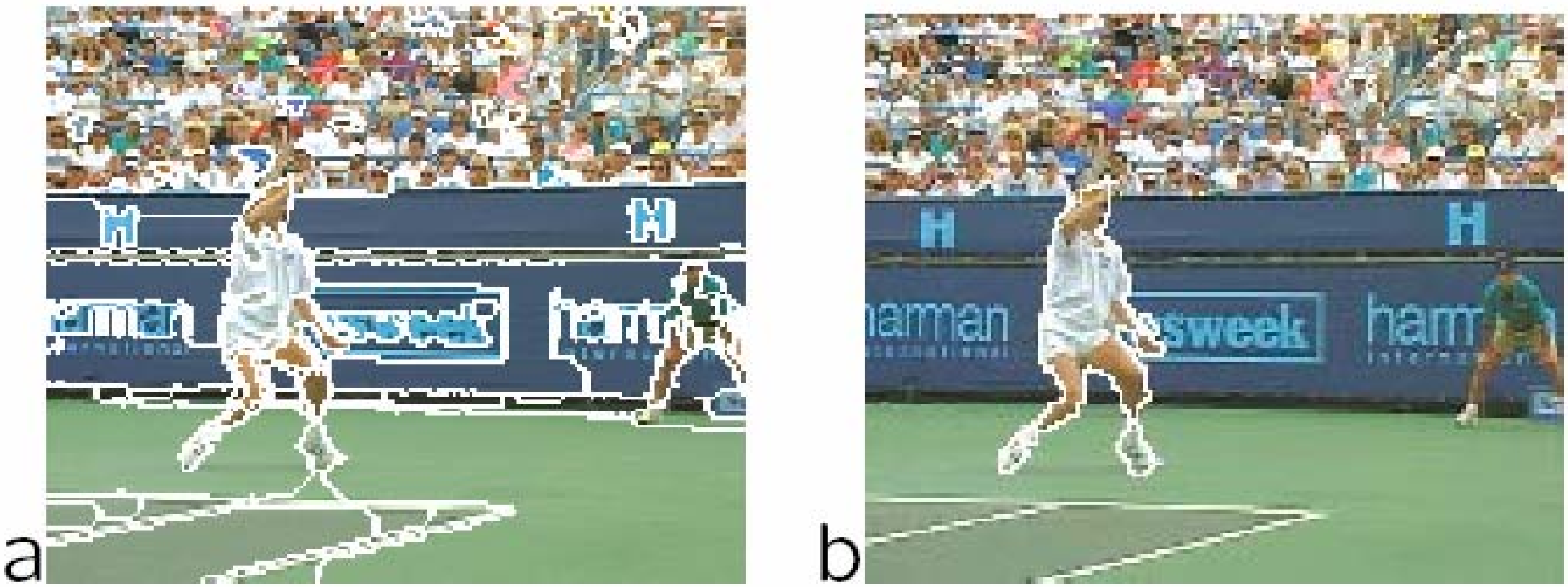


  - Features computed over word images.
  - Annotations – ASCII word.

    "you are to be particularly careful"

  - Segmentation of words easier.
  - Visterms better correlated with concepts.
- So can we extend the analogy to this domain…

Segmentation Comparison



a: Segmentation using only still image information

b: Segmentation using only motion information

Pictures from Patrick Bouthemy’s Website, INRIA

Represent Shots not Keyframes

- Shot boundary detection
  - Use standard techniques.
- Segment moving objects
  - E.g. By finding outliers from dominant (camera) motion.
- Visual features for object and background.
- Motion features for object
  - E.g Trajectory information,
- Motion features for background.
  - Camera pan, zoom …

Models

- One approach - modify relevance model to include motion information.
- Probabilistically annotate shots in the test set.

$$P(c,d_v) = \sum_{J \in T} P(J)P(c \mid J) \prod_{i=1}^{|d_v|} P(v_i \mid J)$$

$$P(c,(d_v,d_m)) = \sum_{S \in T} P(S)P(c \mid S) \prod_{i=1}^{|d|} P(v_i \mid S)P(m_i \mid S)$$

T: training set, S: shots in the training set

- Other models e.g. HMM also possible

Estimation $P(v_i|S)$, $P(m_i|S)$

- If discrete visterms use smoothed maximum likelihood estimates.
- If continuous use kernel density estimates.

- Take advantage of repeated instances of the same object in shot.

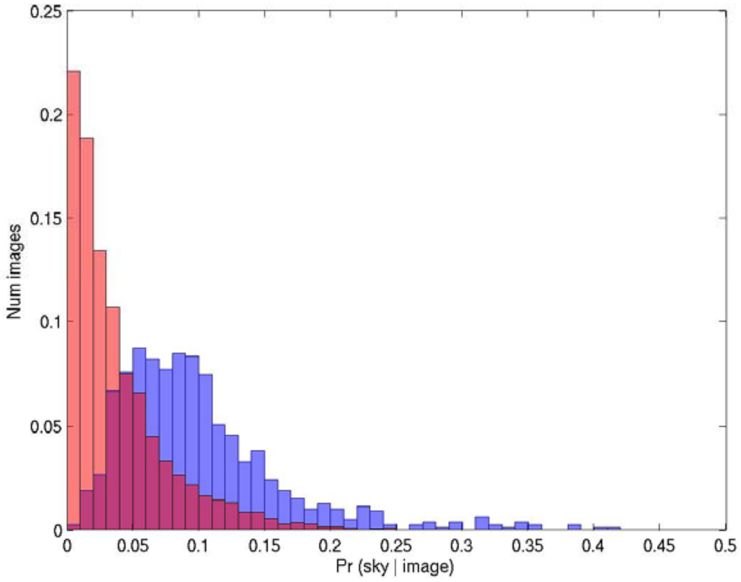
Plan

- Modify models to include dynamic information
- Train on TrecVID03 development dataset
- Test on TrecVID03 test dataset
  - Annotate the test set
  - Retrieve using TrecVID 2003 queries.
  - Evaluate retrieval performance using mean average precision

Score Normalization Experiments

Presented by Desislava Petkova

Motivation for Score Normalization



- Score probabilities are small
- But there seems to be discriminating power
- Try to use likelihood ratios

Bayes Optimal Decision Rule

$$r(s) = \frac{P(w|s)}{P(\bar{w}|s)}$$

$$= \frac{P(s)P(w|s)}{P(s)P(\bar{w}|s)}$$

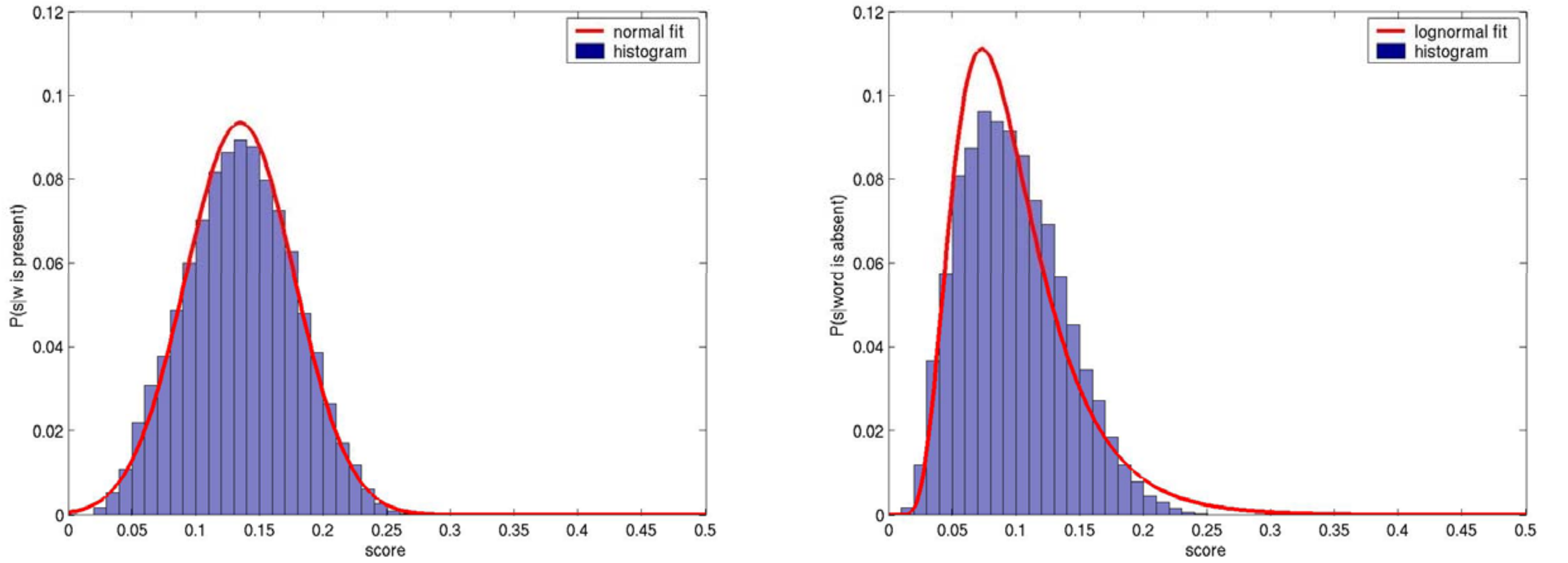
$$= \frac{P(w)P(s|w)}{P(\bar{w})P(s|\bar{w})}$$

$$= \frac{p(w)\,pdf_{w}(s|w)}{p(\bar{w})\,pdf_{\bar{w}}(s|\bar{w})} \Rightarrow P(w|s) = \frac{r(s)}{1+r(s)}$$

Estimating Class-Conditional PDFs

- For each word:
  - Divide training images into positive and negative examples
  - Create a model to describe the score distribution of each set
    - Gamma
    - Beta
    - Normal
    - Lognormal
- Revise word probabilities

Annotation Performance



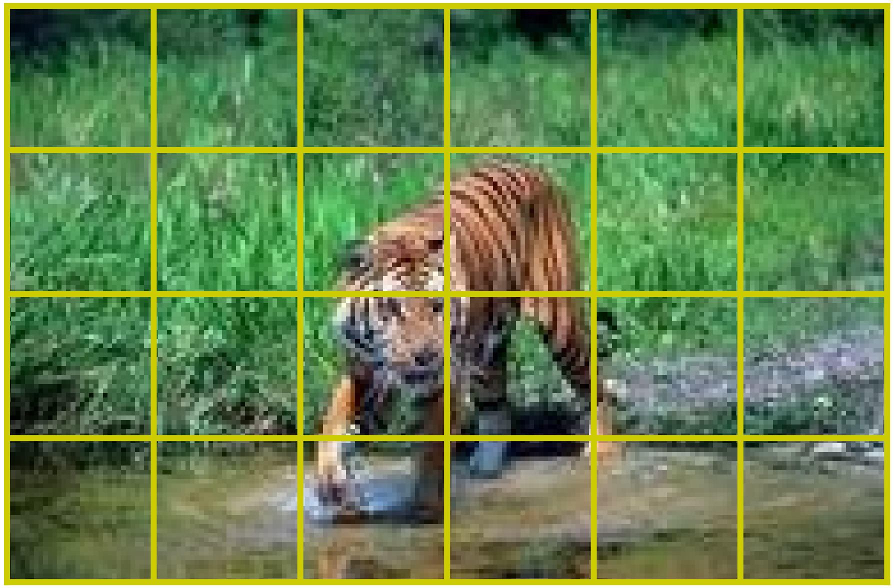
- Did not improve annotation performance on Corel or TREC

Proposal:Using Clustering to Improve Concept Annotation

Desislava Petkova
Mount Holyoke College
17 August 2004

Automatically annotating images

- Corel:
- 5000 images
  - 4500 training
  - 500 testing
- Word vocabulary
  - 374 words
- Annotations
  - 1-5 words
- Image vocabulary
  - 500 visterms



Relevance models for annotation

- A generative language modeling approach
- For a test image $I = \{v_1, \ldots, v_m\}$ compute the joint distribution of each word w in the vocabulary with the visterms of I
  - Compare I with training images J annotated with w

$$P(w, I) = \sum_{J \in T} P(J) P(w, I | J)$$

$$= \sum_{J \in T} P(J) P(w|J) \prod_{i=1}^{m} P(v_i | J)$$

Estimating P(w|J) and P(v|J)

- Use maximum-likelihood estimates
  - Smooth with the entire training set T

$$P(w|J) = (1-a)\frac{c(w,J)}{|J|} + a\frac{c(w,T)}{|T|}$$

$$P(v|J) = (1-b)\frac{c(v,J)}{|J|} + b\frac{c(v,T)}{|T|}$$

Motivation

- Estimating the relevance model of a single image is a noisy process
  - P(v|J): visterm distributions are sparse
  - P(w|J): human annotations are incomplete
- Use clustering to get better estimates

Potential benefits of clustering

{cat, grass, tiger, water}

{cat, grass, tiger}
{water}

{grass, tiger, water}
{cat}

{cat, grass, tiger, tree}

Words in red are missing in the annotation

Relevance Models with Clustering

- Cluster the training images using K-means
  - Use both visterms and annotations
- Compute the joint distribution of visterms and words in each cluster
  - Use clusters instead of individual images

$$P(w, I) = \sum_{C \in T} P(C) P(w|C) \sum_{i=1}^{m} P(v_i|C)$$

Preliminary results on annotation performance

	mAP
Standard relevance model (4500 training examples)	0.14
Relevance model with clusters (100 training examples)	0.128

Cluster-based smoothing

- Smooth maximum likelihood estimates for the training images based on clusters they belong to

$$P(w|J) = (1 - a_1 - a_2)\frac{c(w,J)}{|J|} + a_1 \frac{c(w,C_J)}{|C_J|} + a_2 \frac{c(w,T)}{|T|}$$

$$P(v|J) = (1 - b_1 - b_2)\frac{c(v,J)}{|J|} + b_1 \frac{c(v,C_J)}{|C_J|} + b_2 \frac{c(v,T)}{|T|}$$

Experiments

- Optimize smoothing parameters
  - Divide training set
    - 4000 training images
    - 500 validation images
- Find the best set of clusters
  - Query-dependent clusters
  - Investigate soft clustering

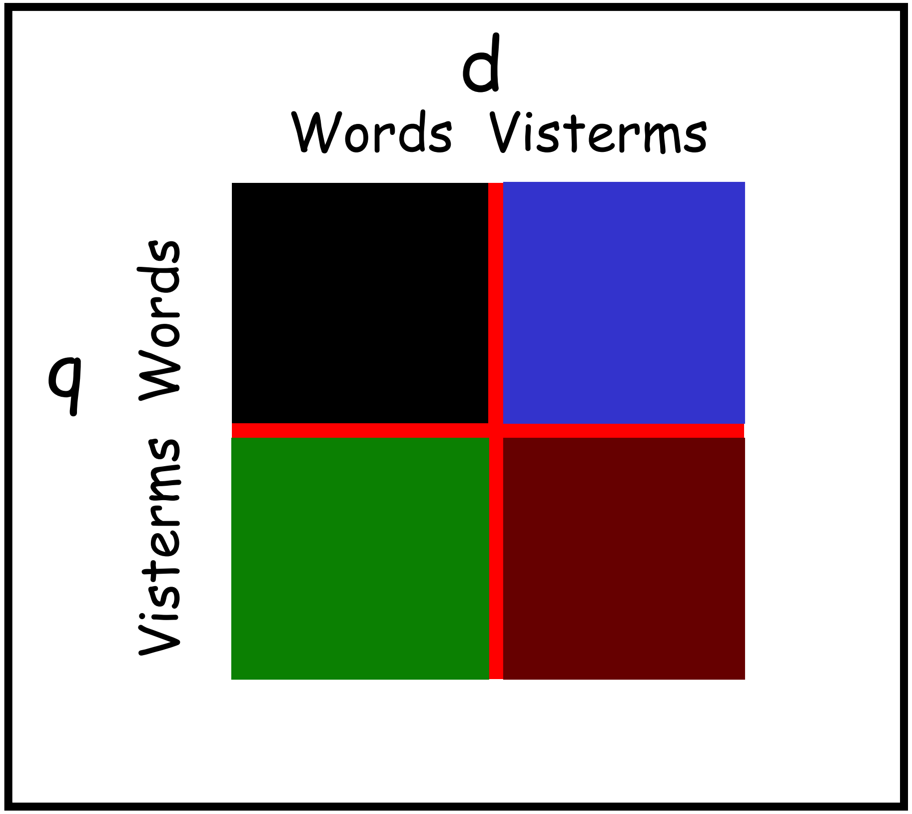
Evaluation plan

- Retrieval performance
  - Average precision and recall for one-word queries
    - Comparison with the standard relevance model

Hidden Markov Models for Image Annotations

Pavel Ircing
Sanjeev Khudanpur

Presentation Outline



Translation (MT) models
(Paola),

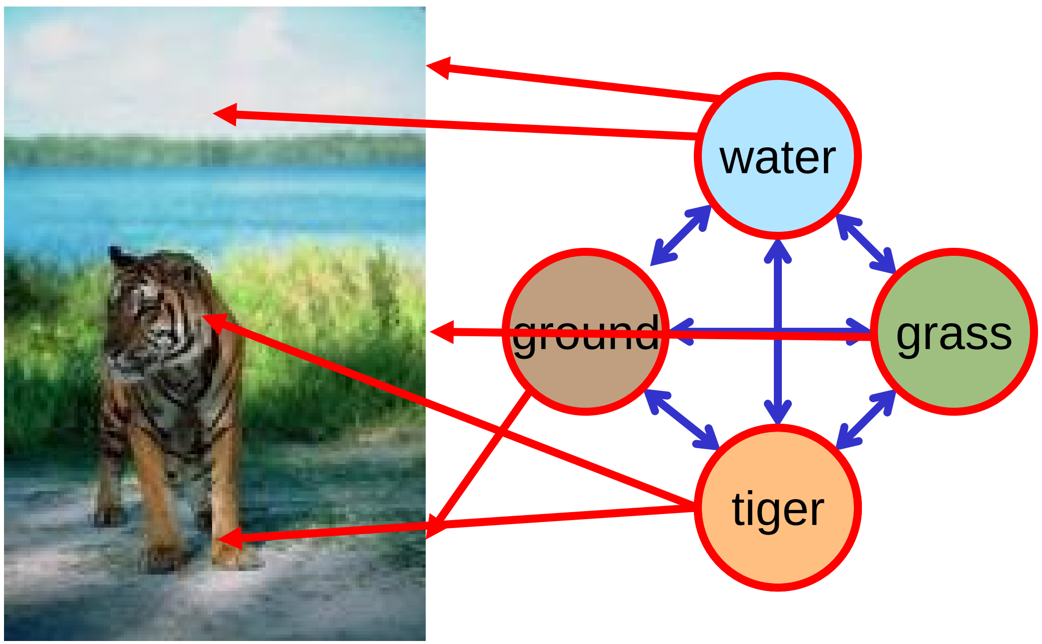
Relevance Models
(Shao Lei, Desislava),

Graphical Models
(Pavel, Brock)

Text classification models
(Matt)

Integration & Summary
(Dietrich)

Model setup



Training HMMs:

- separate HMM for each training image – states given by manual annotations.
- image blocks are "generated" by annotation words
- alignment between image blocks and annotation words is a hidden variable, models are trained using the EM algorithm (HTK toolkit)

Test HMM has |W| states, 2 scenarios: (a) **p(w'|w)** uniform

(b) **p(w'|w)** from co-occurrence LM

Posterior probability from forward-backward pass used for **p(w|Image)**

Challenges in HMM training

- Inadequate annotations
- There is no notion of **order** in the annotation words
  - Difficulties with automatic alignment between words and image regions
- No **linear** order in image blocks (assume raster-scan)
  - Additional spatial dependence between block-labels is missed
  - Partially addressed via a more complex DBN (see later)

Inadequacy of the annotations

- Corel database
  - Annotators often mark only interesting objects

beach
palm
people
tree

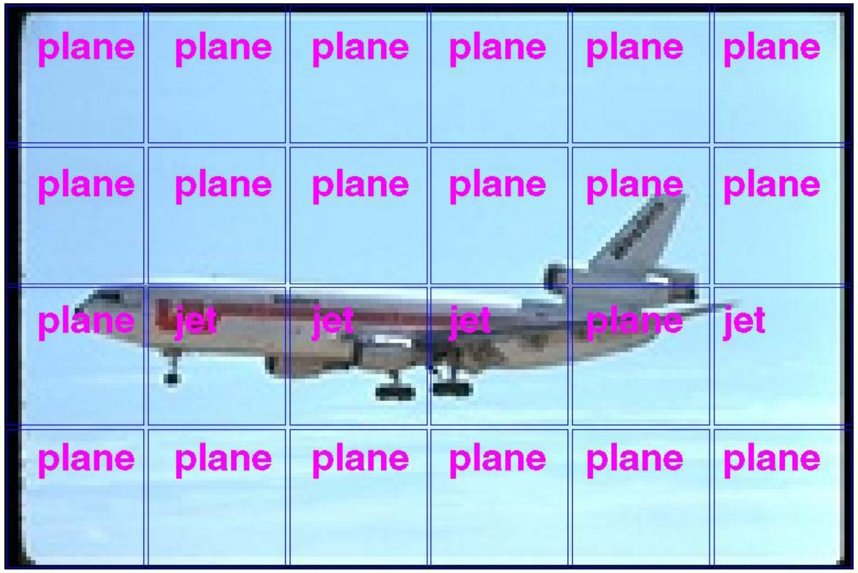
- TRECVID database
  - Annotation concepts capture mostly semantics of the image and they are not very suitable for describing visual properties

man-made object

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

Alignment problems

- There is no notion of **order** in the annotation words
  - Difficulties with automatic alignment between words and image regions



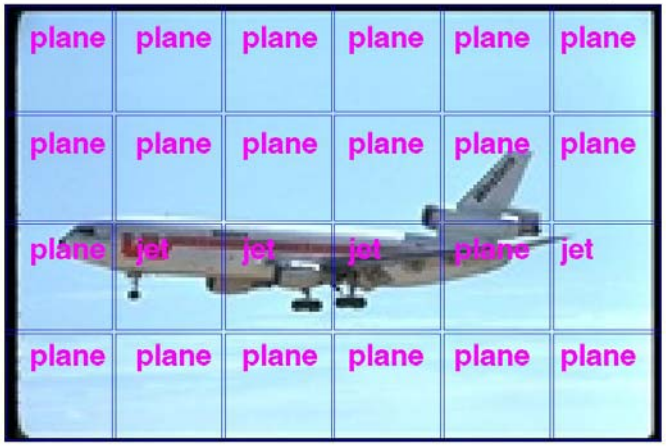
jet, plane, sky

Gradual Training

- Identify a set of "background" words (sky, grass, water,...)
- In the initial stages of HMM training
  - Allow only "background" states to have their individual emission probability distributions
  - All other objects share a single "foreground" distribution
- Run several EM iterations
- Gradually untie the "foreground" distribution and run more EM iterations

Gradual Training Results

Forced alignment – flat–start training



jet, plane, sky

Forced alignment – gradual training

jet, plane, sky

Results:

- Improved alignment of training images
- Annotation performance on test images did not change significantly

Another training scenarios

- models were forced to visit every state during training
  - huge models, marginal difference in performance
- special states introduced to account for unlabelled background and unlabelled foreground, with different strategies for parameter tying

Annotation performance - Corel

Image features	LM	mAP
Discrete	No	0.120
	Yes	0.150
Continuous (1 Gaussian per state)	No	0.140
	Yes	0.157

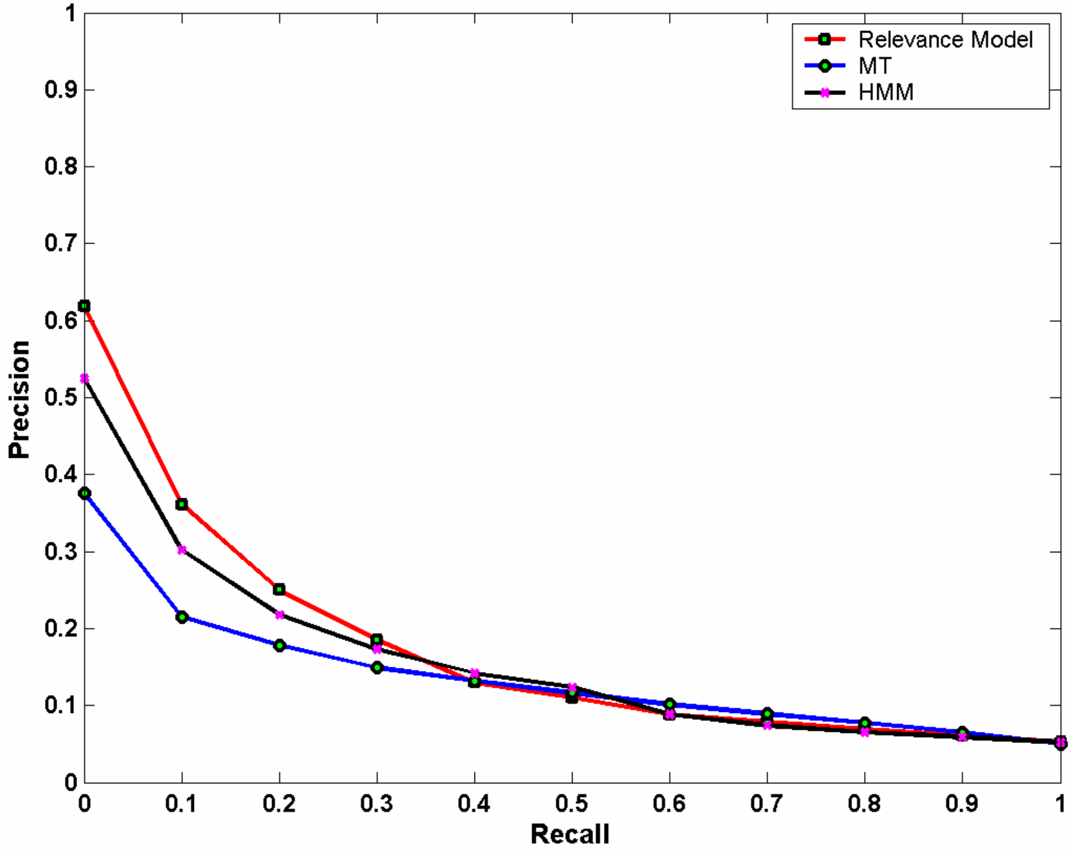
- Continuous features are better than discrete
- Co-ocurrence language model also gives moderate improvement

Annotation performance - TRECVID

- Continuous features only, no language model

Model	LM	mAP
1 Gaussian per state	No	0.094
	Yes	X
12 Gaussians per state	No	0.145
	Yes	X

Annotation Performance on TREC



Summary: HMM-Based Annotation

- Very encouraging preliminary results
  - Effort started this summer, validated on Corel, and yielded competitive annotation results on TREC
- Initial findings
  - Proper normalization of the features is crucial for system performance: bug found and fixed on Friday!
  - Simple HMMs seem to work best
    - More complex training topology didn't really help
    - More complex parameter tying was only marginally helpful
- Glaring gaps
  - Need a good way to incorporate a language model

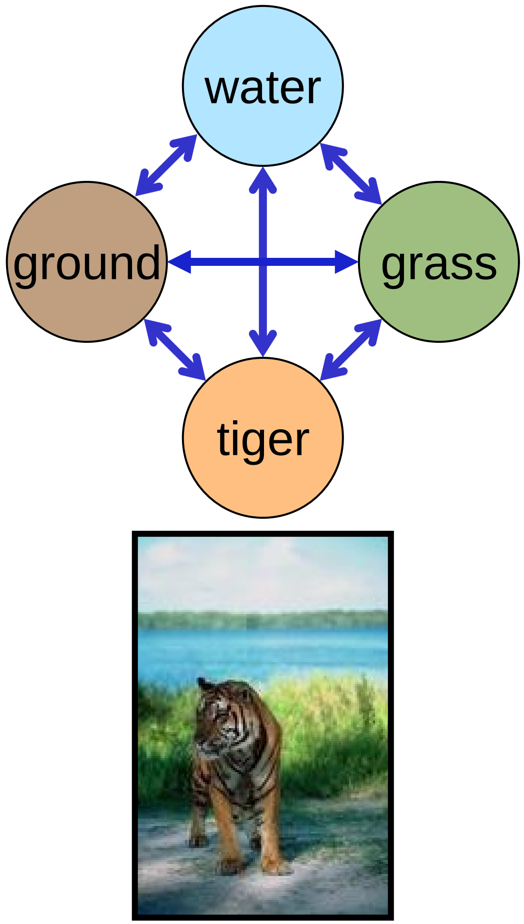
Graphical Models for Image Annotation
+
Joint Segmentation and Labeling for Content Based Image Retrieval

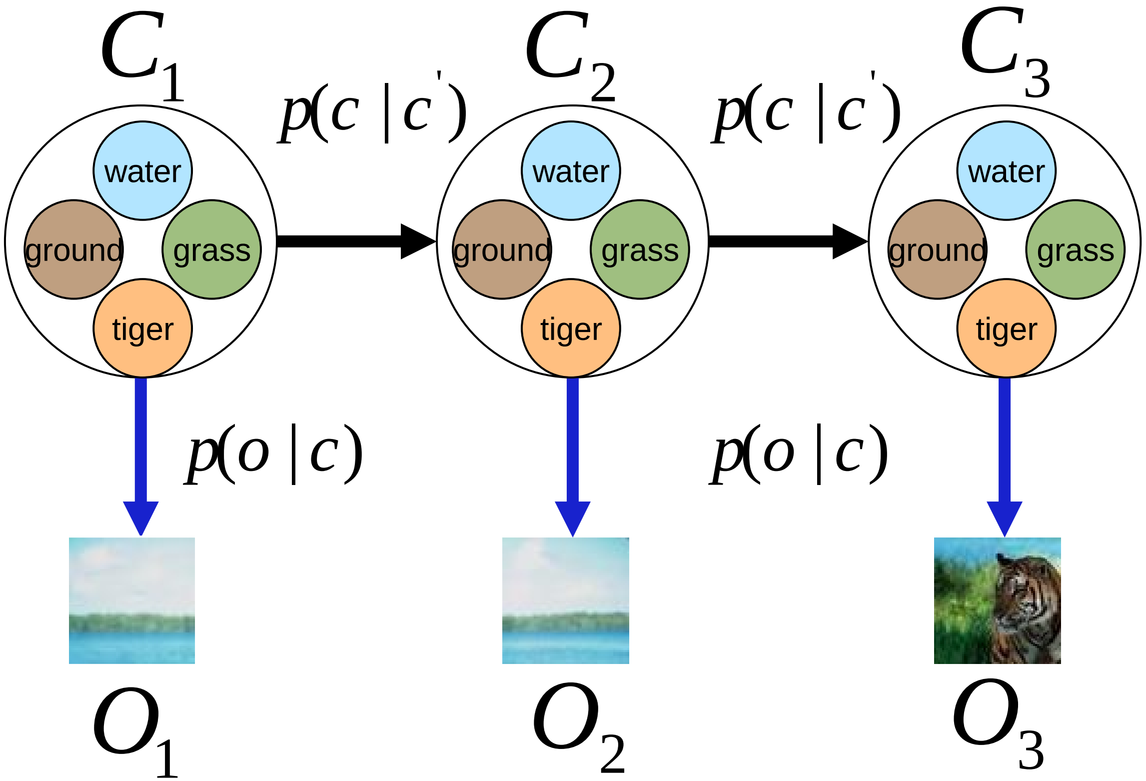
Brock Pytlik
Johns Hopkins University
bep@cs.jhu.edu

Outline

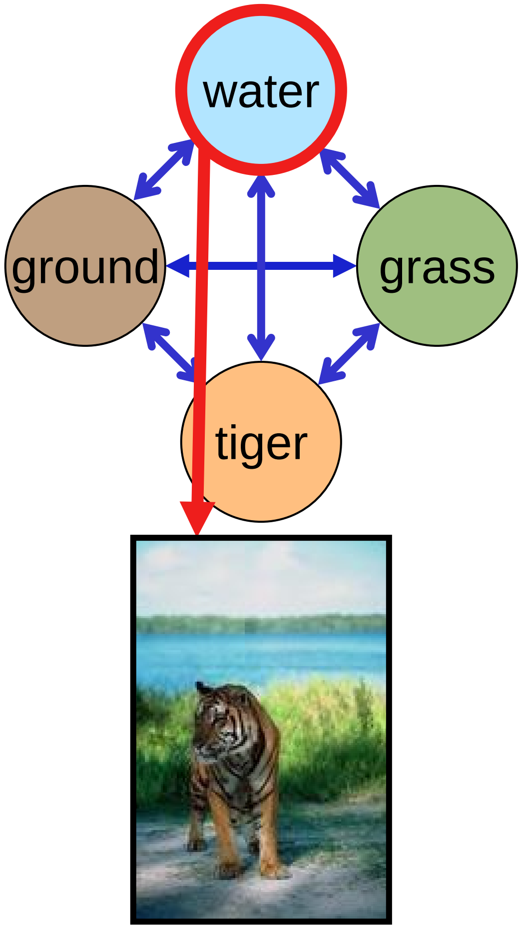
- Graphical Models for Image Annotation
  - Hidden Markov Models
    - Preliminary Results
  - Two-Dimensional HMM's
    - Work in Progress
- Joint Image Segmentation and Labeling
  - Tree Structure Models of Image Segmentation
    - Proposed Research

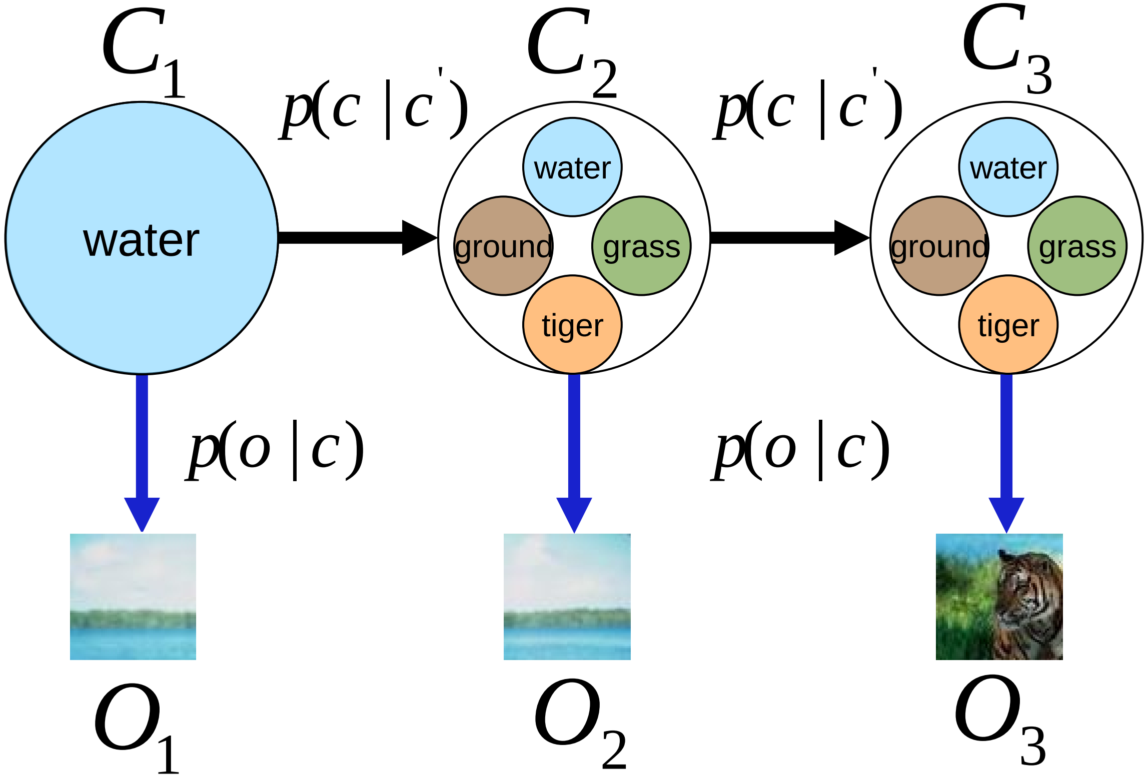
Graphical Model Notation



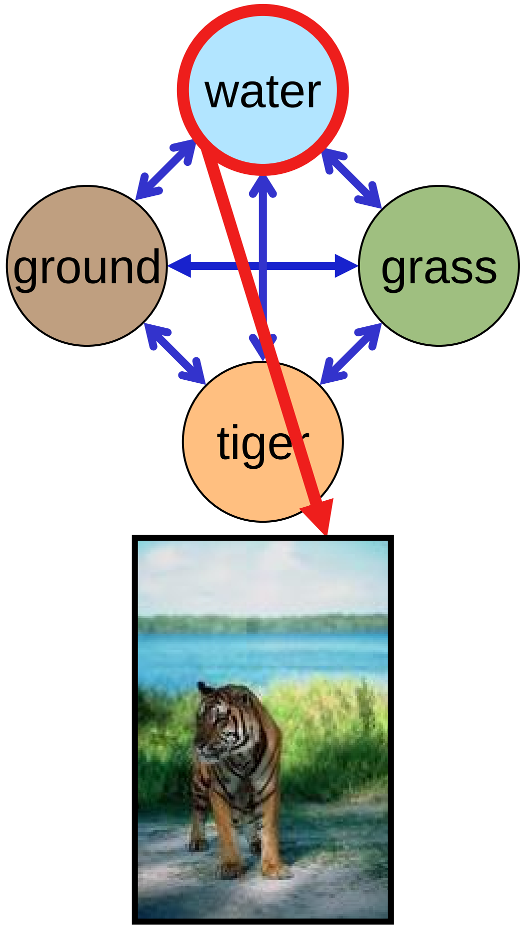


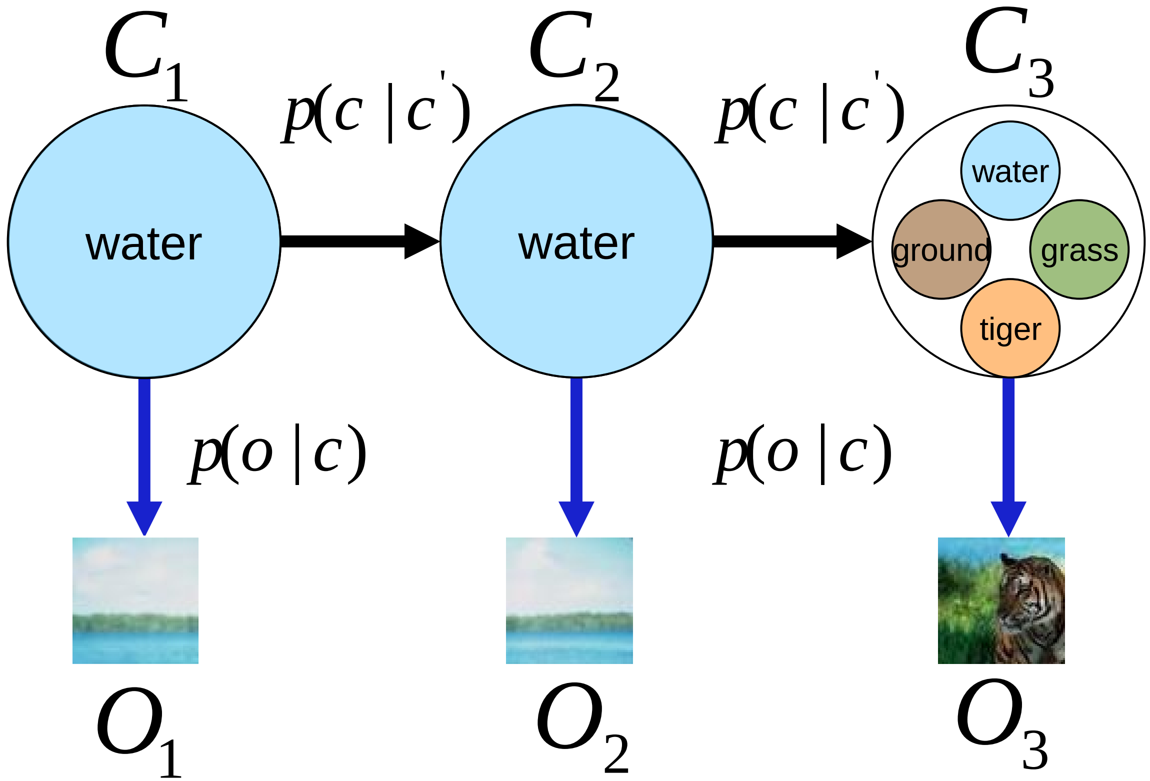
Graphical Model Notation



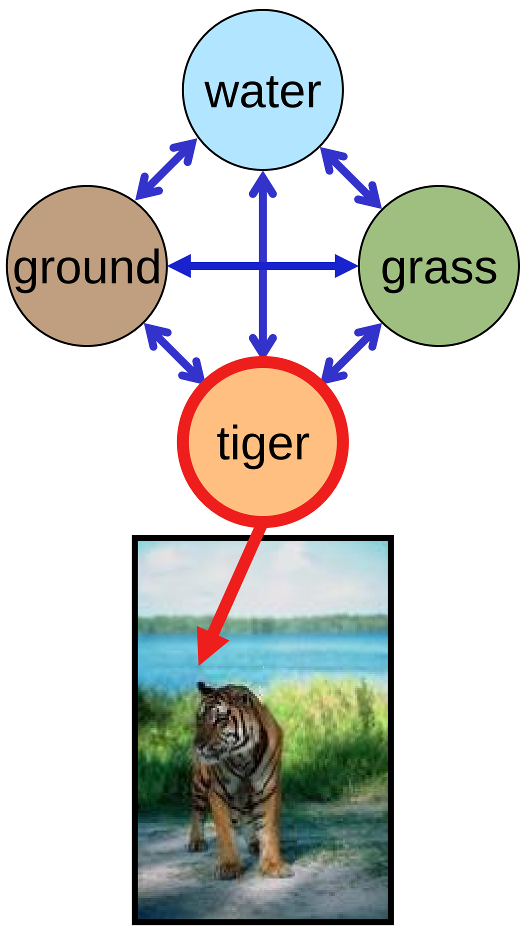


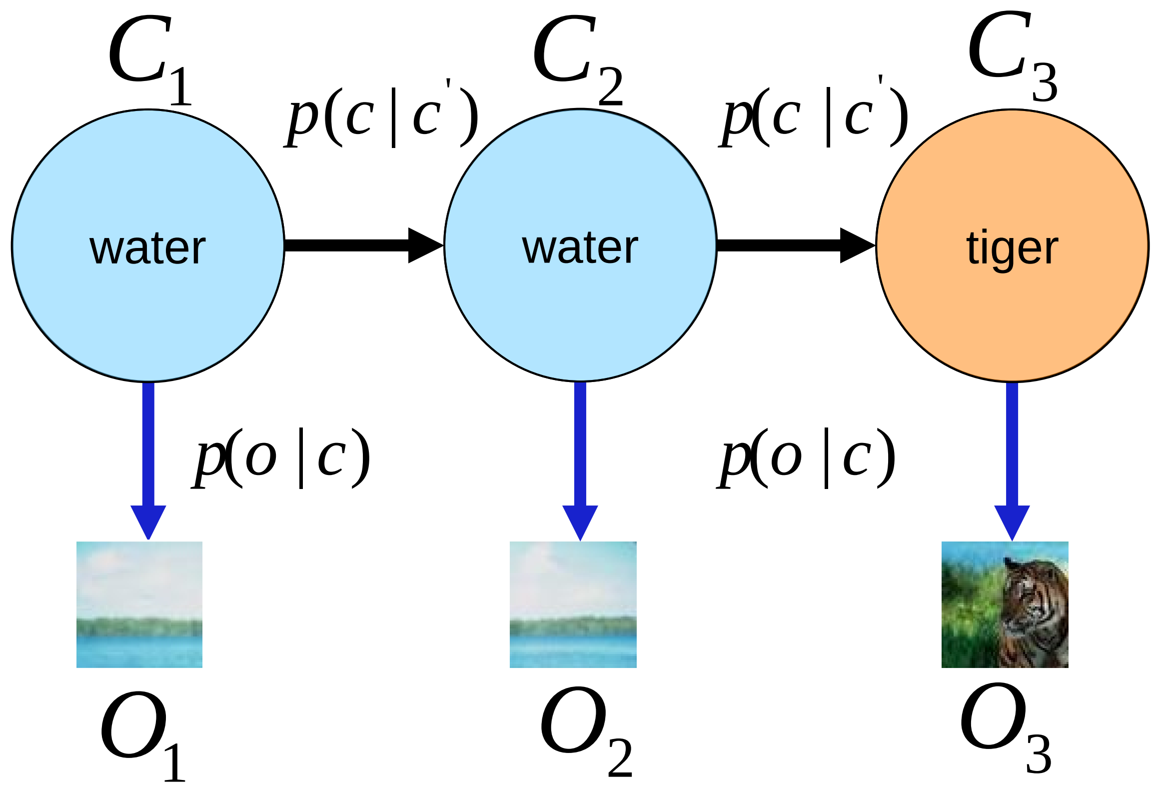
Graphical Model Notation



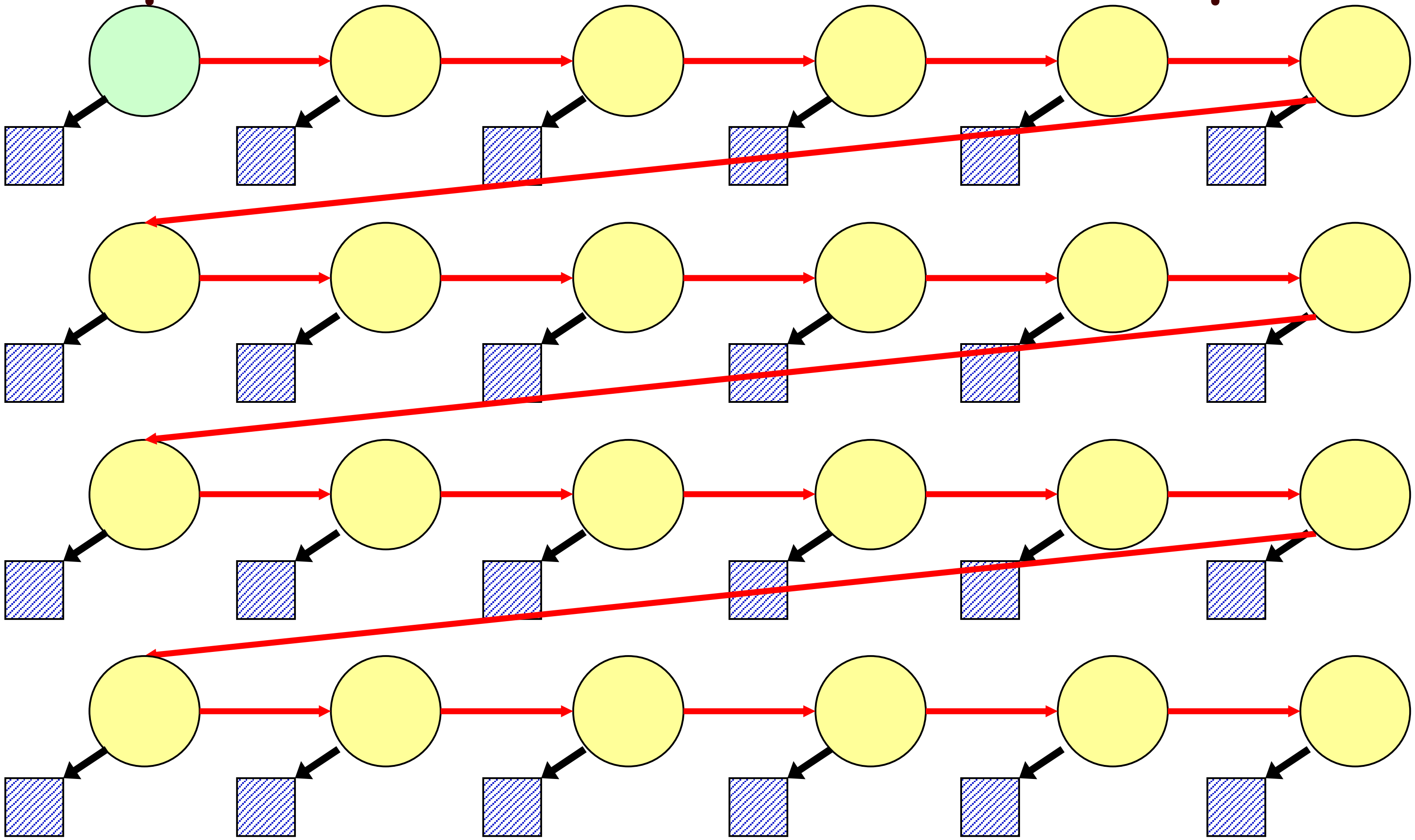


Graphical Model Notation



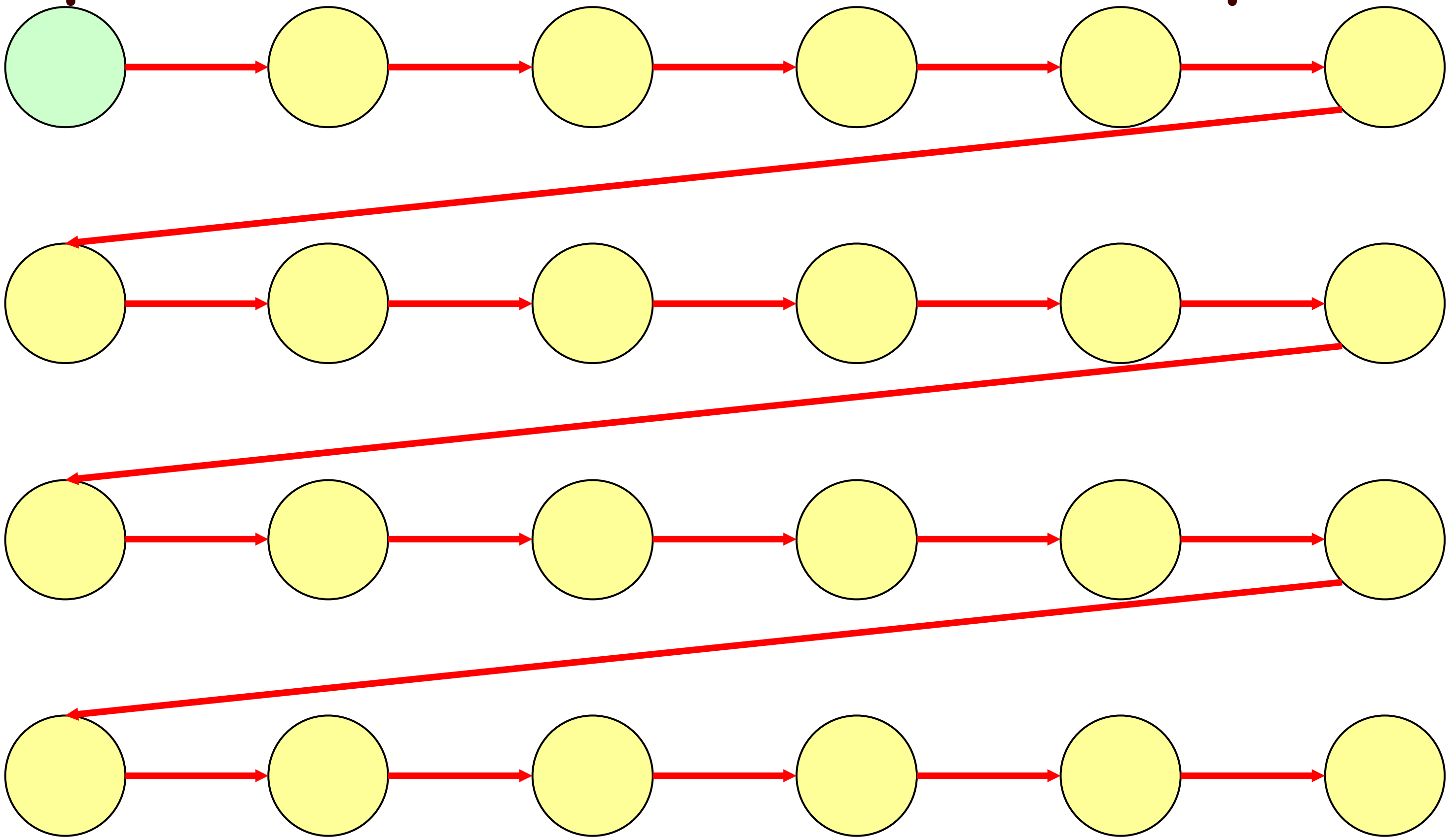


Graphical Model Notation Simplified



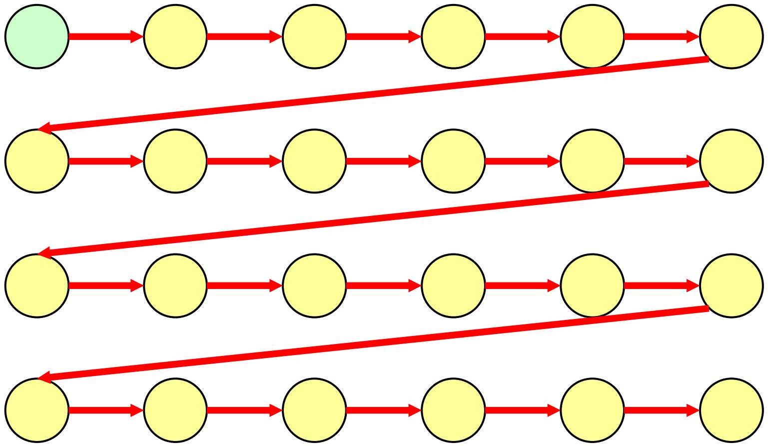
An HMM for a 24-block Image

Graphical Model Notation Simplified



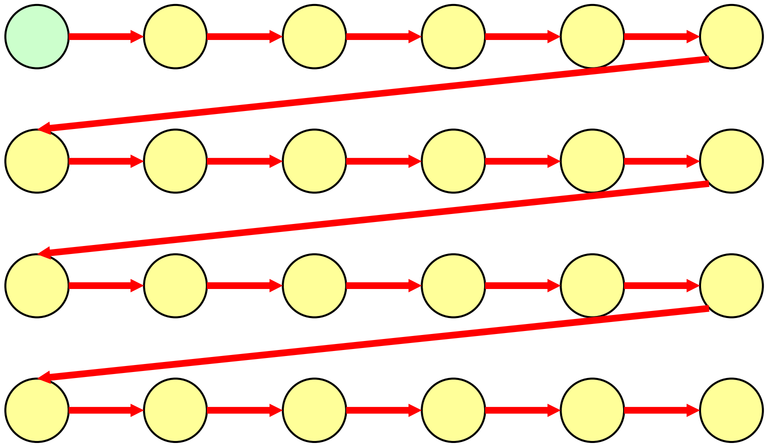
An HMM for a 24-block Image

Modeling Spatial Structure

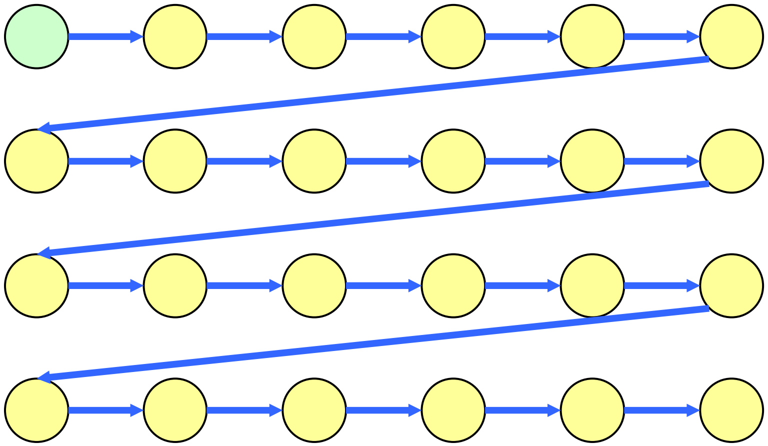


An HMM for a 24-block Image

Modeling Spatial Structure

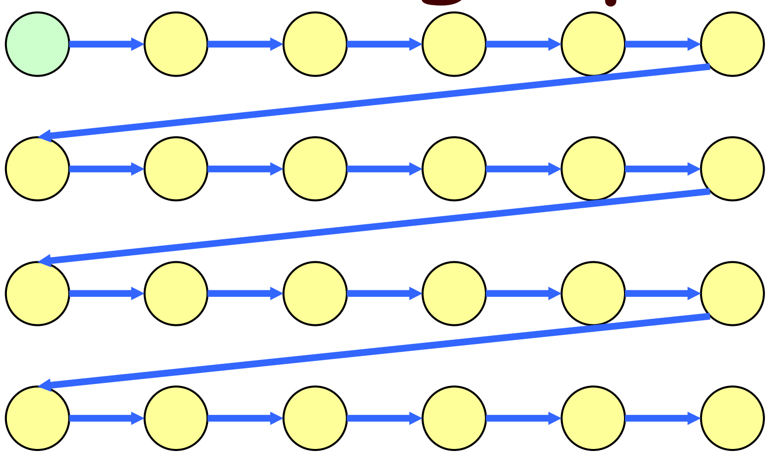


An HMM for a 24-block Image

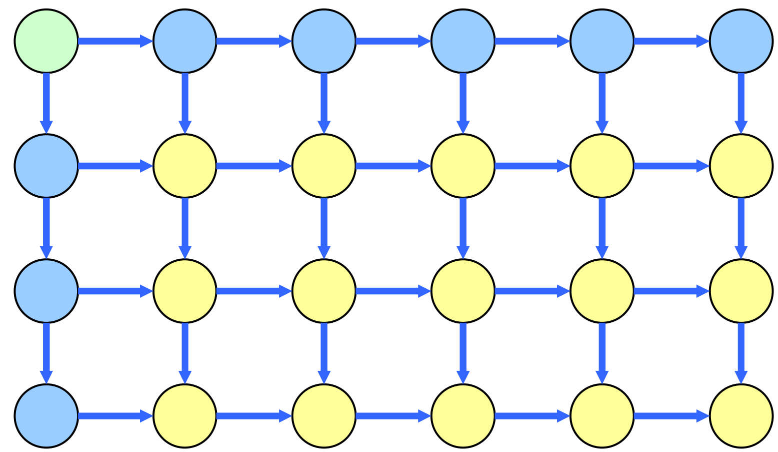


Transition probabilities represent spatial extent of objects

Modeling Spatial Structure

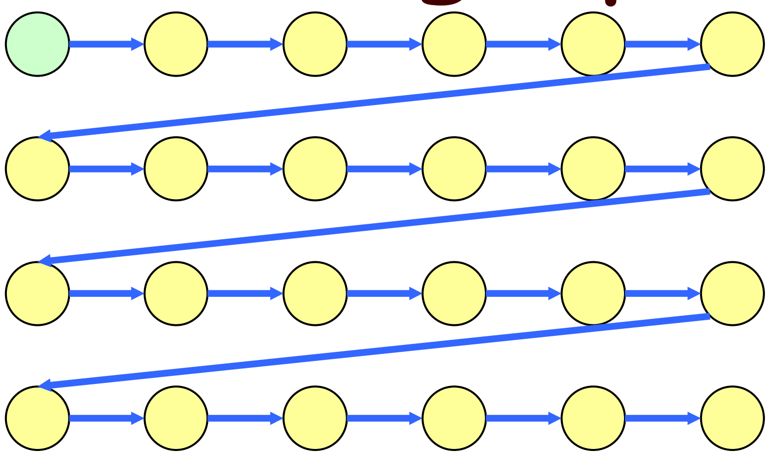


Transition probabilities represent spatial extent of objects

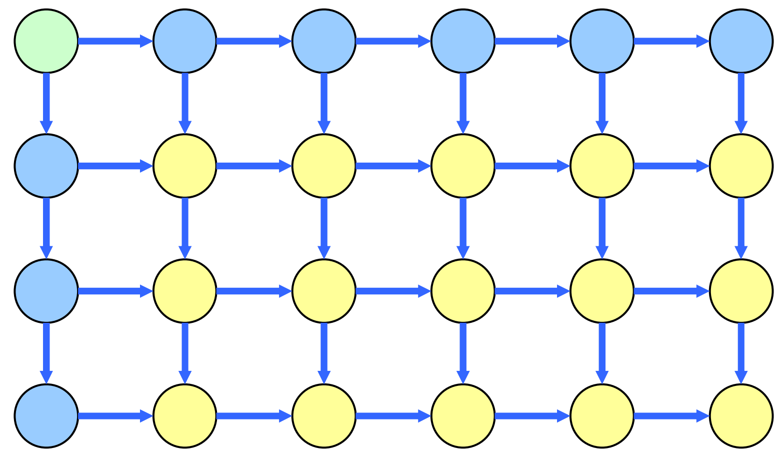


A Two-Dimensional Model for a 24-block Image

Modeling Spatial Structure



Transition probabilities represent spatial extent of objects

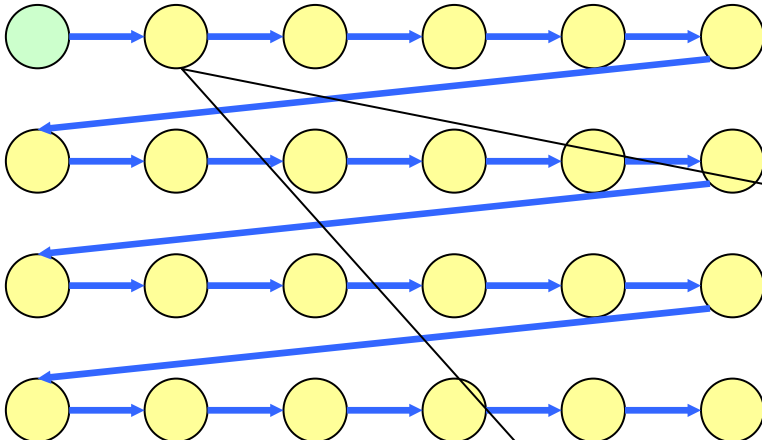


A Two-Dimensional Model for a 24-block Image

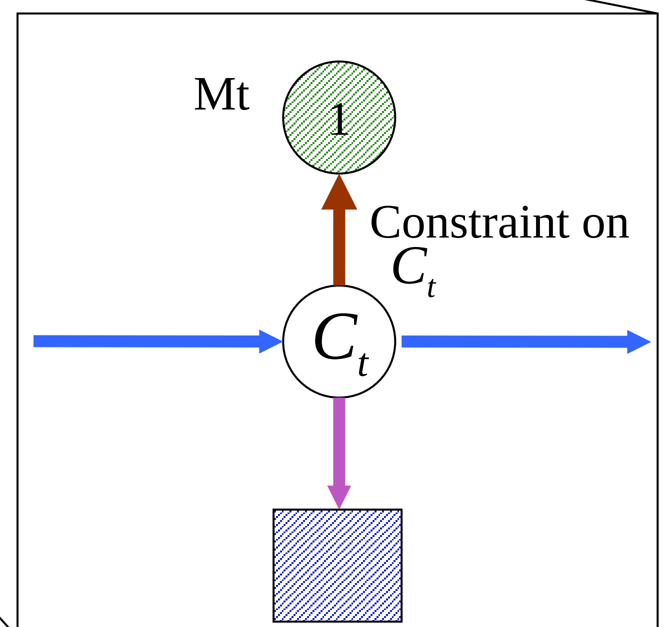
Model	Training Time Per Image	Training Time Per Iteration
1-D HMM	.5 sec	37.5 min
2-D HMM	110 sec	8250 min = 137.5 hr

Bag-of-Annotations Training

Unlike ASR Annotation Words are Unordered

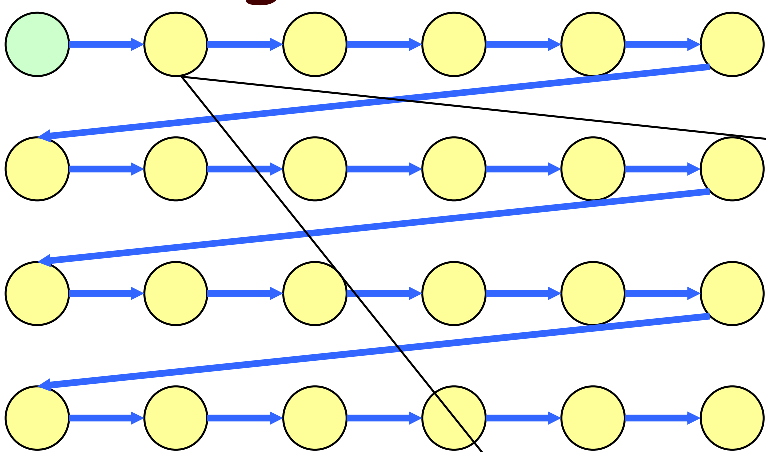


$$p(M_t = 1) = \begin{cases} 1 & 1 \text{ if } c_t \in \{\text{tiger, grass, sky}\} \\ 0 & \text{otherwise} \end{cases}$$

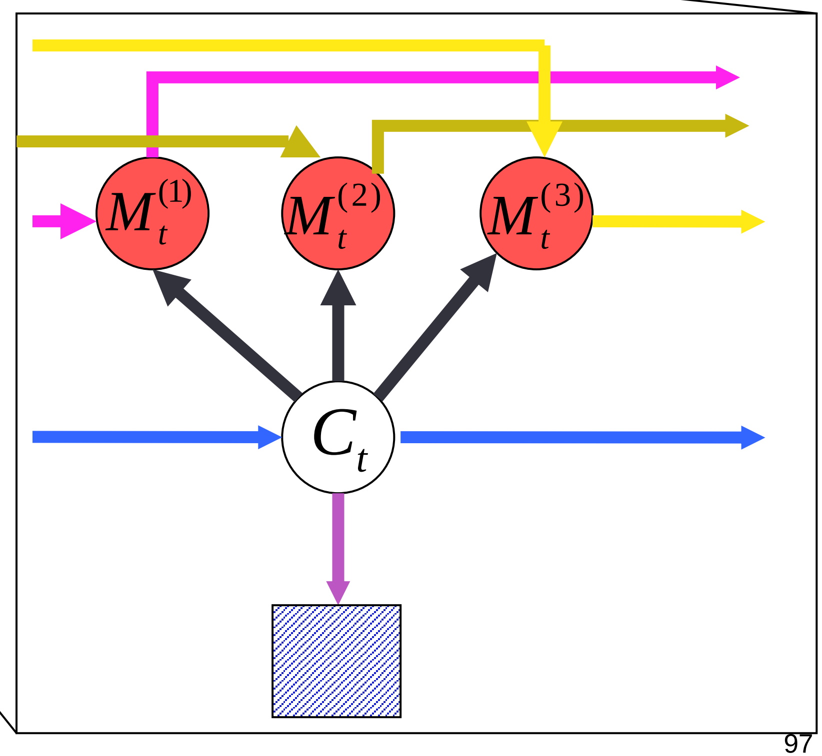


Tiger, Sky, Grass

Bag-of-Annotations Training (II)
Forcing Annotation Words to Contribute



Only permit paths that visit every annotation word.



$$M_t^{(1)} = M_{t-1}^{(1)} \vee (C_t = \text{tiger})$$

$$M_t^{(2)} = M_{t-1}^{(2)} \vee (C_t = \text{grass})$$

$$M_t^{(3)} = M_{t-1}^{(3)} \vee (C_t = \text{sky})$$

Inference on Test Images

- Forward Decoding

$$p(c \mid d_v) = \frac{p(c, d_v)}{p(d_v)}$$

Inference on Test Images

- Forward Decoding

$$p(c \mid d_v) = \frac{p(c, d_v)}{p(d_v)} = \frac{\sum_{S \ni c} \left[\prod_{i=1}^{N} p(v_i \mid s_i) \right] p(S)}{}$$

Inference on Test Images

- Forward Decoding

$$p(c \mid d_v) = \frac{p(c, d_v)}{p(d_v)} = \frac{\sum_{S \ni c} \left[\prod_{i=1}^{N} p(v_i \mid s_i) \right] p(S)}{\sum_{S} \left[\prod_{i=1}^{N} p(v_i \mid s_i) \right] p(S)}$$

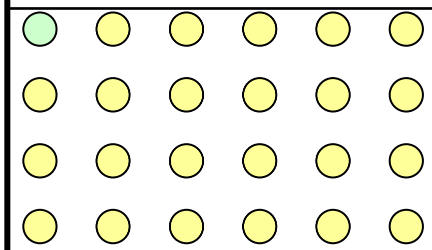
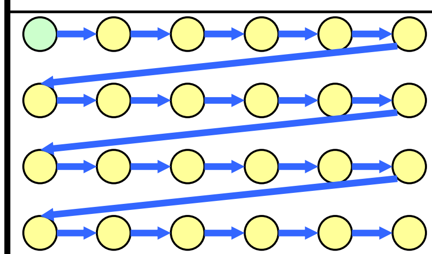
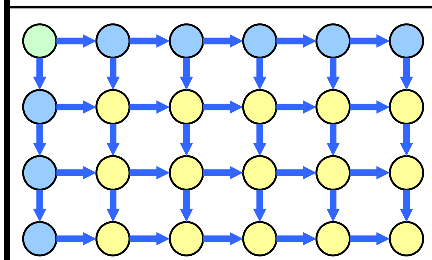
Inference on Test Images

- Forward Decoding

$$p(c \mid d_v) = \frac{p(c, d_v)}{p(d_v)} = \frac{\sum_{S \ni c} \left[\prod_{i=1}^{N} p(v_i \mid s_i) \right] p(S)}{\sum_{S} \left[\prod_{i=1}^{N} p(v_i \mid s_i) \right] p(S)}$$

- Viterbi Decoding
  - Approximate Sum over all Paths with the Best Path

Annotation Performance on Corel Data

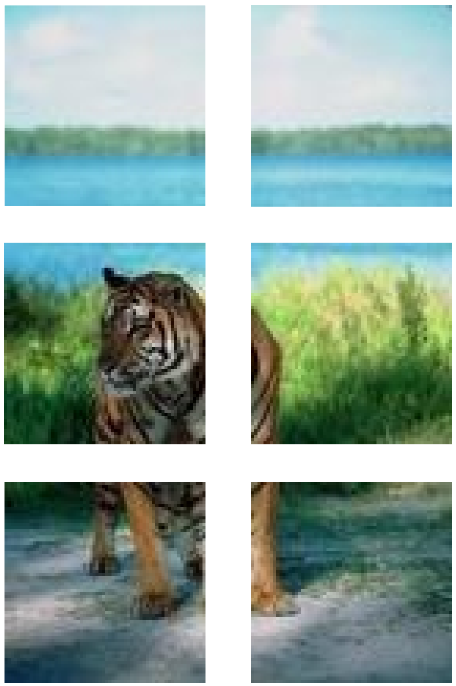
Model	Image Features	mAP
	Discrete	0.071
	Discrete Continuous	0.086 0.074
	Discrete Continuous	Training TBD

- Working with 2-D models needs further study
- mAP not yet on par with other models

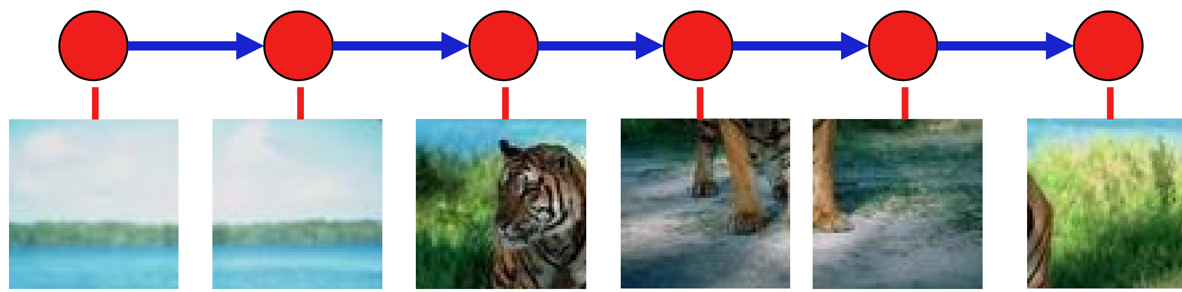
Future Work

- Improved Training for Two-Dimensional Models
  - $p(c_{i,j} \mid c_{i-1,j}, c_{i,j-1}) \propto p(c \mid c_{i-1,j})\, p(c \mid c_{i-1,j})$
    - Permits training horizontal and vertical chains separately
  - Other variations could be investigated
- Next Idea
  - Joint Image Segmentation and Labeling

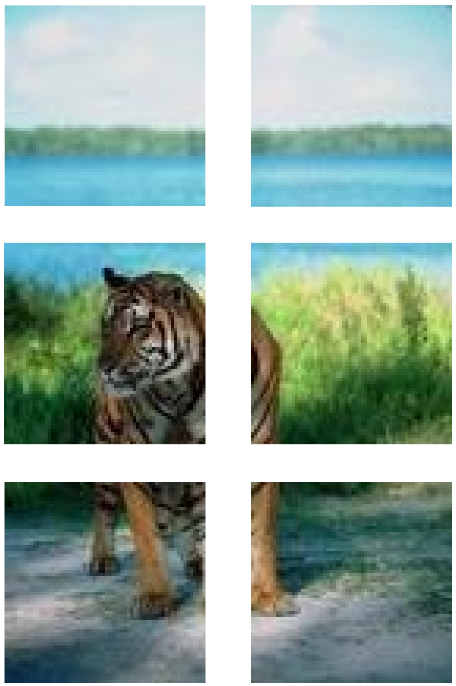
Joint Segmentation and Labeling



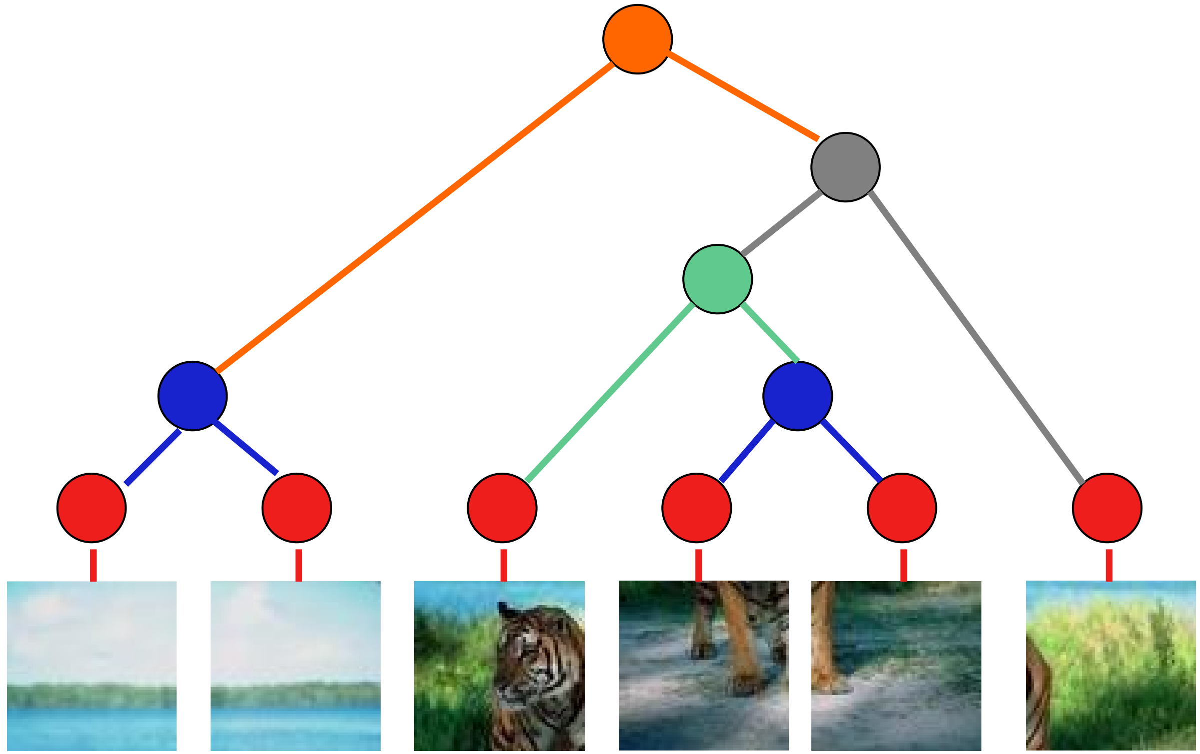
tiger, grass, sky



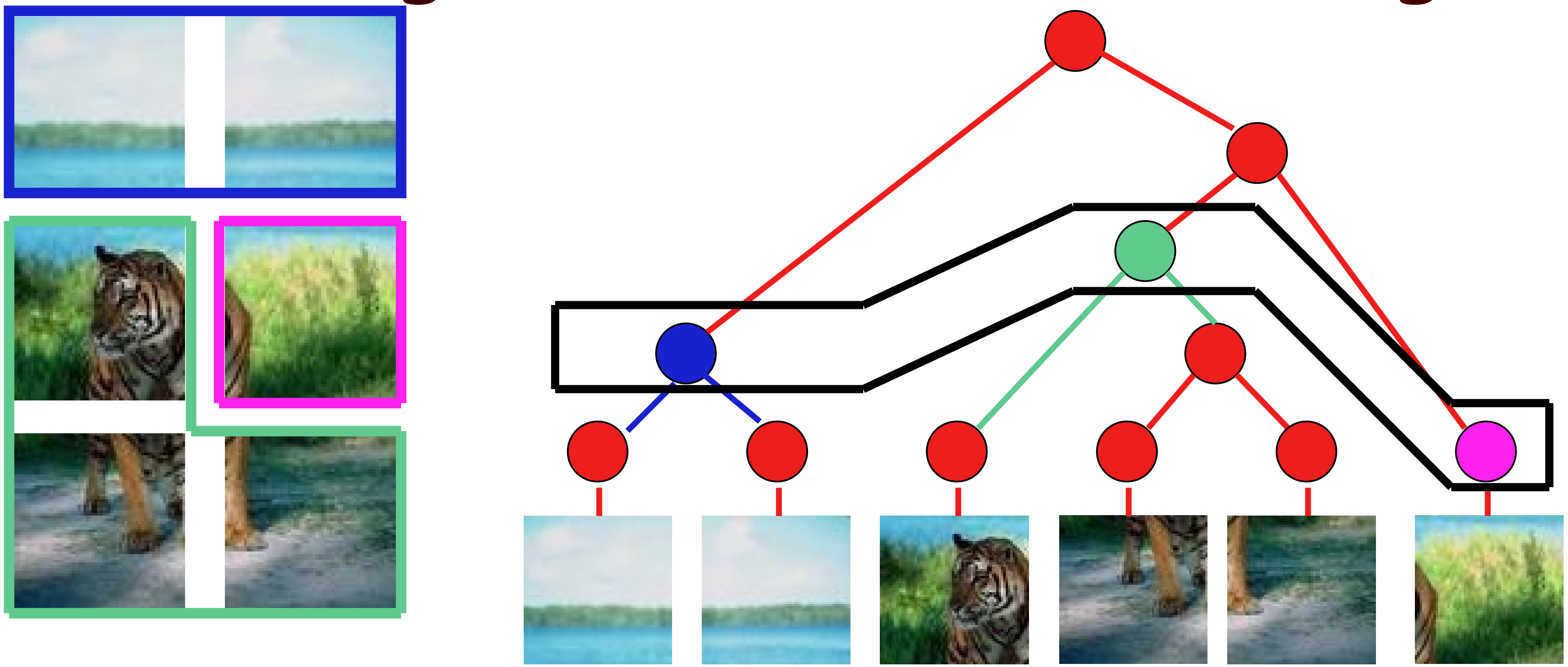
Joint Segmentation and Labeling



tiger, grass, sky

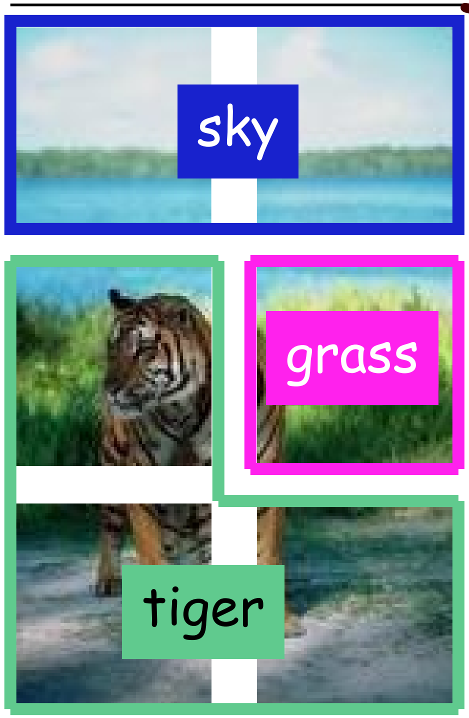


Joint Segmentation and Labeling

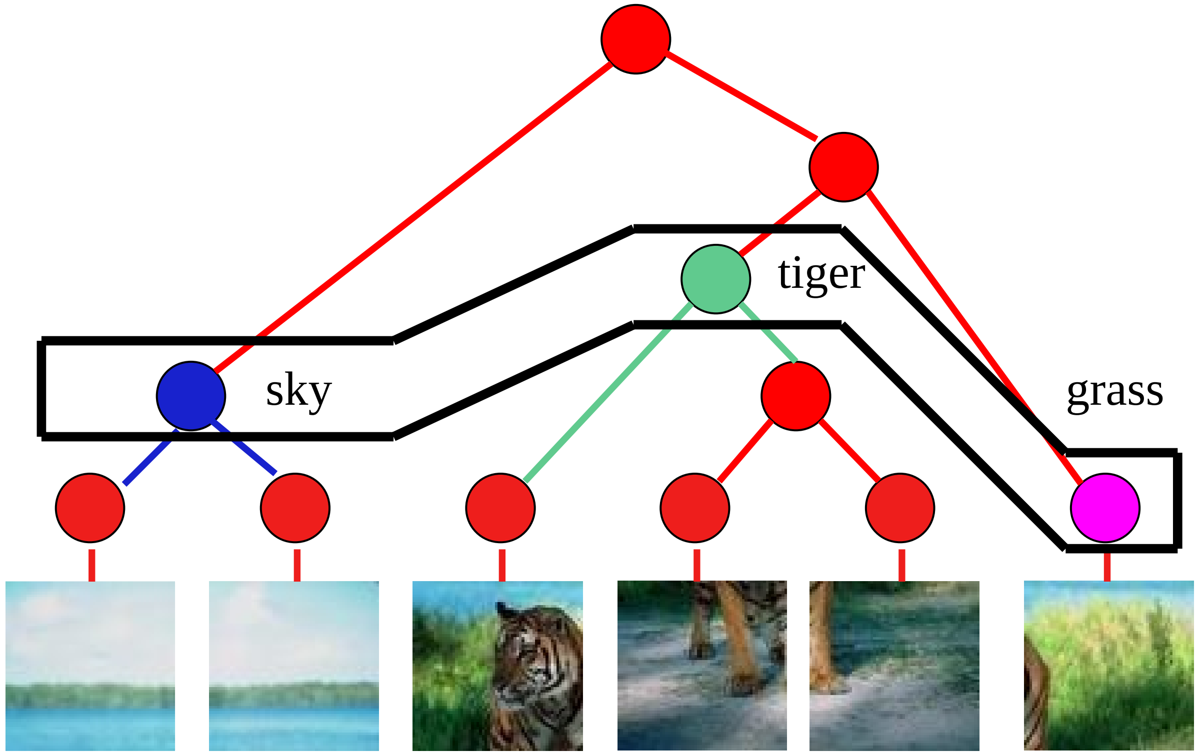


tiger, grass, sky

Joint Segmentation and Labeling



tiger, grass, sky



Research Proposal

- A Generative Model for Joint Segmentation and Labeling
  - Tree construction by agglomerative clustering of image regions (blocks) based on visual similarity
    - Segmentation = A cut across the resulting tree
    - Labeling = Assigning concepts to resulting leaves

Model

- General Model

$$p(\bar{c}, d_v) = \sum_{u \in \text{cuts}(\text{tree}(d_v))} p(u) \prod_{l \in \text{leaves}(u)} p(c_l \mid u, l)\, p(\text{obs}(l) \mid c_l)$$

Model

- General Model

$$p(\bar{c}, d_v) = \sum_{u \in \text{cuts}(\text{tree}(d_v))} \boxed{p(u)} \prod_{l \in \text{leaves}(u)} p(c_l \mid u, l)\, p(\text{obs}(l) \mid c_l)$$

Probability of Cut

Model

- General Model

$$p(\bar{c}, d_v) = \sum_{u \in \text{cuts}(\text{tree}(d_v))} p(u) \prod_{l \in \text{leaves}(u)} \boxed{p(c_l \mid u, l)}\, p(\text{obs}(l) \mid c_l)$$

Probability of Label Given
Cut and Leaf

Model

- General Model

$$p(\bar{c}, d_v) = \sum_{u \in \text{cuts}(\text{tree}(d_v))} p(u) \prod_{l \in \text{leaves}(u)} p(c_l \mid u, l) \boxed{p(\text{obs}(l) \mid c_l)}$$

Probability of
Observation Given Label

Model

- General Model

$$p(\bar{c}, d_v) = \sum_{u \in \text{cuts}(\text{tree}(d_v))} p(u) \prod_{l \in \text{leaves}(u)} p(c_l \mid u, l)\, p(\text{obs}(l) \mid c_l)$$

- Independent Generation of Observations Given Label

$$p(\bar{c}, d_v) = \sum_{u \in \text{cuts}(\text{tree}(d_v))} p(u) \prod_{l \in \text{leaves}(u)} p(c_l \mid u, l) \prod_{o \in \text{child}(l)} p(o \mid c_l)$$

Estimating Model Parameters

- Suitable independence assumptions may need to be made
  - All cuts are equally likely?
  - Given a cut, leaf labels have a Markov dependence
  - Given a label, its image footprint is independent neighboring image regions
- Work out EM algorithm for this model

Estimating Cuts given Topology

- Uniform
  - All cuts containing $|\bar{c}|$ leaves or more equally likely
- Hypothesize number of segments produced
  - Hypothesize which possible segmentation used
- Greedy Choice
  - Pick node with largest observation probability remaining that produces a valid segmentation
    - Repeat until all observations accounted for
  - Changes Model
    - No longer distribution over cuts
    - Affects valid labeling strategies

Estimating Labels Given Cuts

- Uniform
  - Like HMM training with fixed concept transitions
- Number of Children
  - Sky often generates a large number of observations
  - Canoe often generates a small number of observations
- Co-occurrence Language Model
  - Eliminates label independence given cut
  - Could do two-pass model like MT group did (not exponential)

$$p_2(c \mid u, l) = \sum_{a \in C} \left[\sum_{m \in \text{leaves}(u)} p_1(a \mid m) \right] p(c \mid a)$$

Estimating Observations Given Labels

- Label Generates its Observations Independently
  - Problem: Product of Children at least as high as Parent Score
- Label Generates Composite Observation at Node

Evaluation Plan

- Evaluate on Corel Image set using mAP
- TREC annotation task

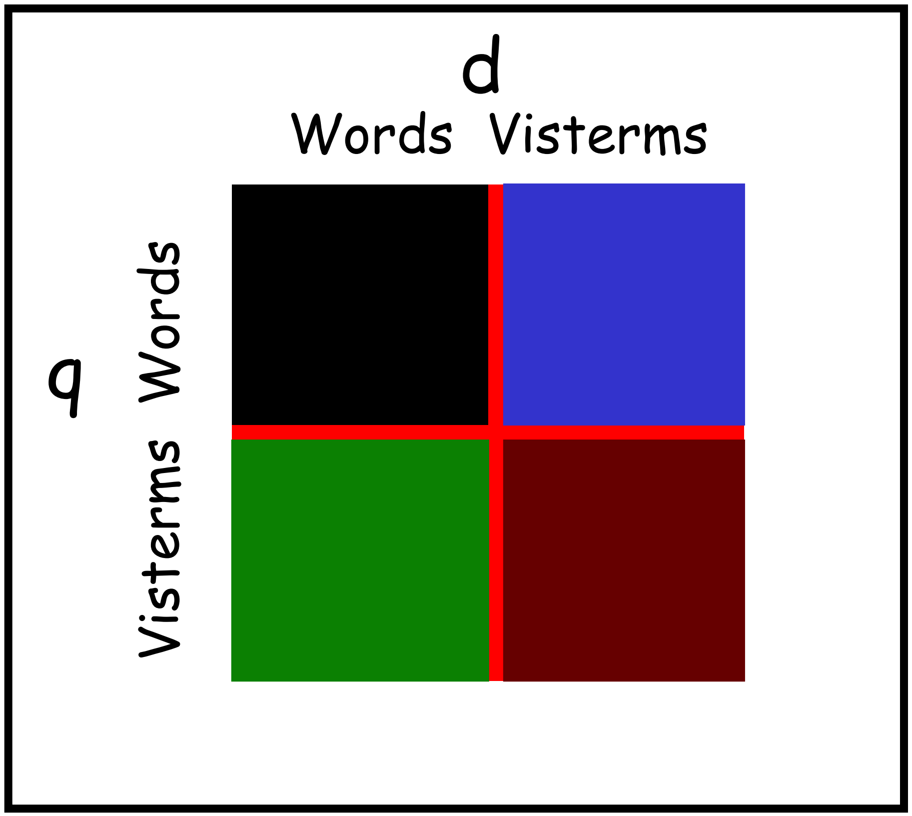
Questions?

Predicting Visual Concepts From Text

Presented by

Matthew Krause

Presentation Outline



Translation (MT) models (Paola),

Relevance Models (Shao Lei,Desislava),

Graphical Models (Pavel, Brock)

Text classification models (Matt)

Integration & Summary (Dietrich)

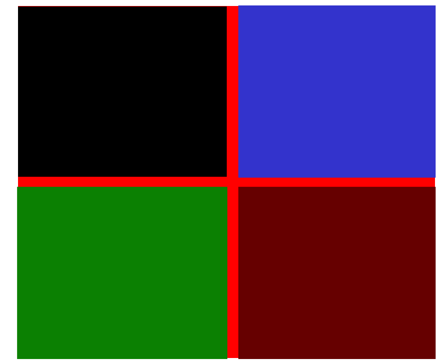
A Motivating Example

A Motivating Example

<Word stime="177.09" dur="0.22" conf="0.727"> IT'S </Word>
<Word stime="177.31" dur="0.25" conf="0.963"> MUCH </Word>
<Word stime="177.56" dur="0.11" conf="0.976"> THE </Word>
<Word stime="177.67" dur="0.29" conf="0.977"> SAME </Word>
<Word stime="177.96" dur="0.14" conf="0.980"> IN </Word>
<Word stime="178.10" dur="0.13" conf="0.603"> THE </Word>
<Word stime="178.38" dur="0.57" conf="0.953"> SUMMERTIME </Word>
<Word stime="178.95" dur="0.50" conf="0.976"> GLACIER </Word>
<Word stime="179.45" dur="0.60" conf="0.974"> AVALANCHE </Word>

Concepts

- Assume there is a hidden variable c which generates query words from a document's visterms.

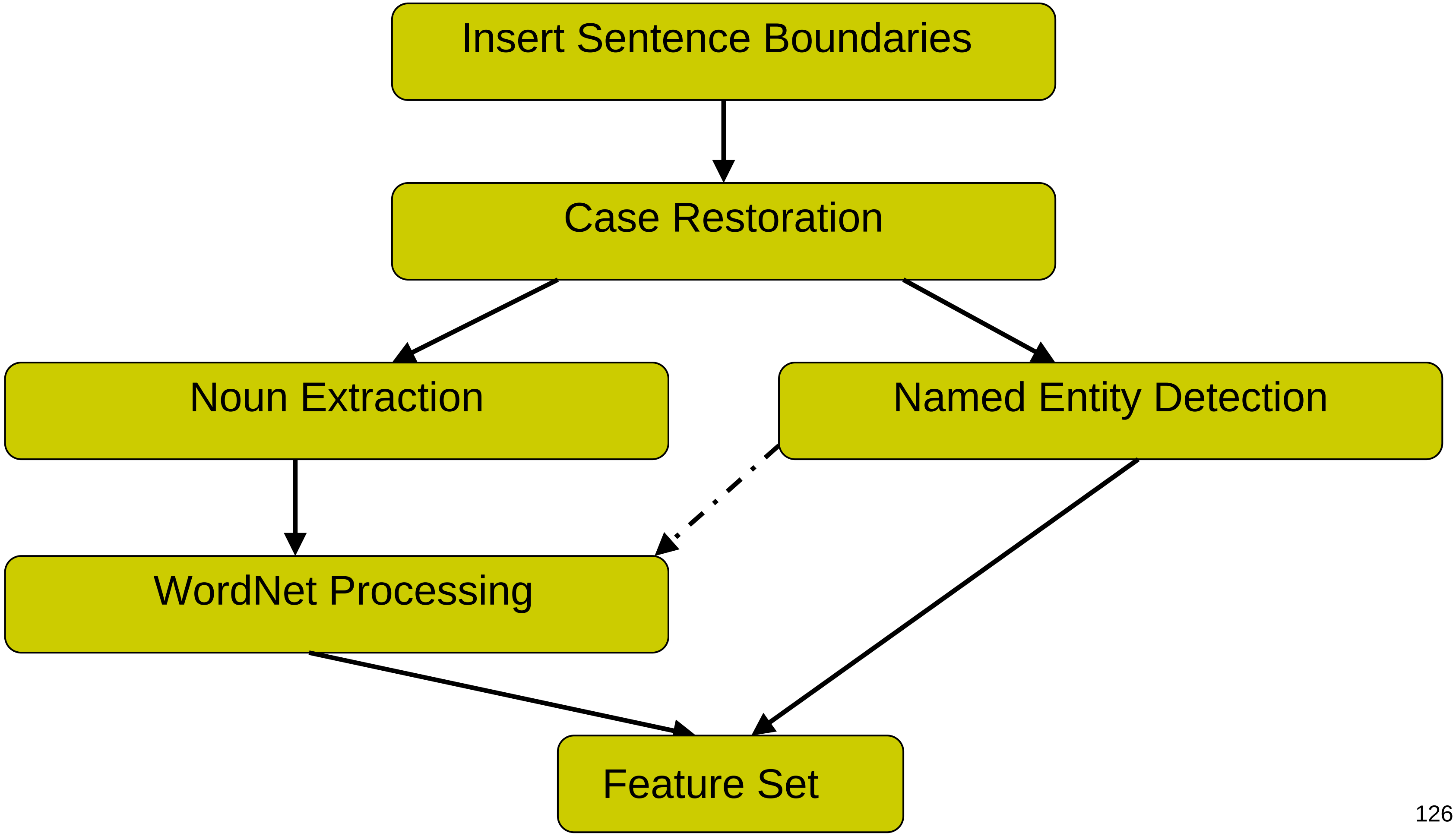


$$p(q_v \mid d_w) = \sum_C p(q_v \mid d_w, c)\, p(c \mid d_w) \cong \sum_C \underbrace{p(q_v \mid c)}_{}\, \underbrace{p(c \mid d_w)}_{}$$

ASR → Features Example

STEVE FOSSETT AND HIS BALLOON SOLO SPIRIT ARSENIDE OVER THE BLACK SEA DRIFTING SLOWLY TOWARDS THE COAST OF THE CAUCUSES HIS TEAM PLANS IF NECESSARY TO BRING HIM DOWN AFTER DAYLIGHT TOMORROW YOU THE CHECHEN CAPITAL OF GROZNY

Building Features



ASR → Features Example

STEVE FOSSETT AND HIS BALLOON SOLO SPIRIT ARSENIDE OVER THE BLACK SEA DRIFTING SLOWLY TOWARDS THE COAST OF THE CAUCUSES HIS TEAM PLANS IF NECESSARY TO BRING HIM DOWN AFTER DAYLIGHT TOMORROW YOU THE CHECHEN CAPITAL OF GROZNY

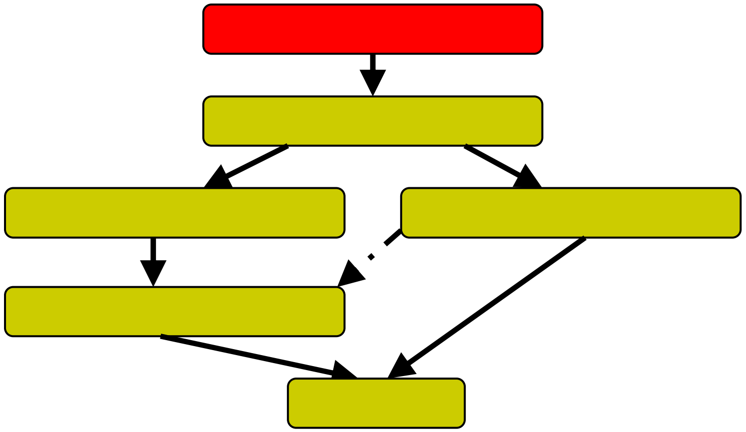
ASR → Features Example

STEVE FOSSETT AND HIS BALLOON SOLO SPIRIT ARSENIDE.

OVER THE BLACK SEA DRIFTING SLOWLY TOWARDS THE COAST OF THE CAUCUSES.

HIS TEAM PLANS IF NECESSARY TO BRING HIM DOWN AFTER DAYLIGHT TOMORROW.

YOU THE CHECHEN CAPITAL OF GROZNY



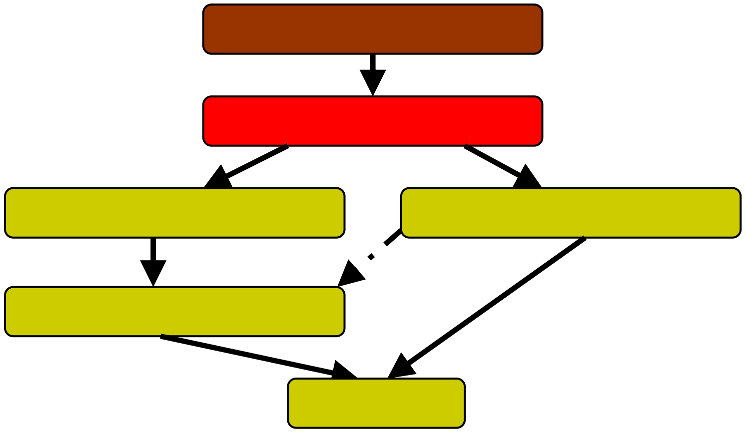
ASR → Features Example

Steve Fossett and his balloon Solo Spirit arsenide.

Over the Black Sea drifting slowly towards the coast of the caucuses.

His team plans if necessary to bring him down after daylight tomorrow.

you the Chechan capital of Grozny….



ASR → Features Example

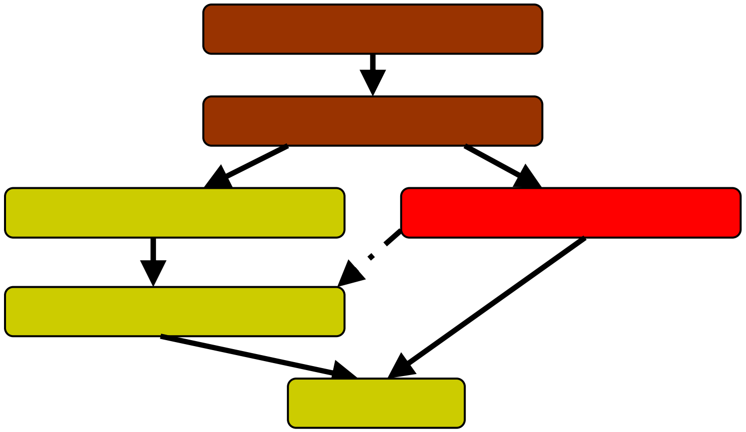
Steve Fossett and his balloon Solo Spirit arsenide.

Over the Black Sea drifting slowly towards the coast of the caucuses.

His team plans if necessary to bring him down after daylight tomorrow.

you the Chechan capital of Grozny.

- Named Entities
  - Male Person, Location (Region)



ASR → Features Example

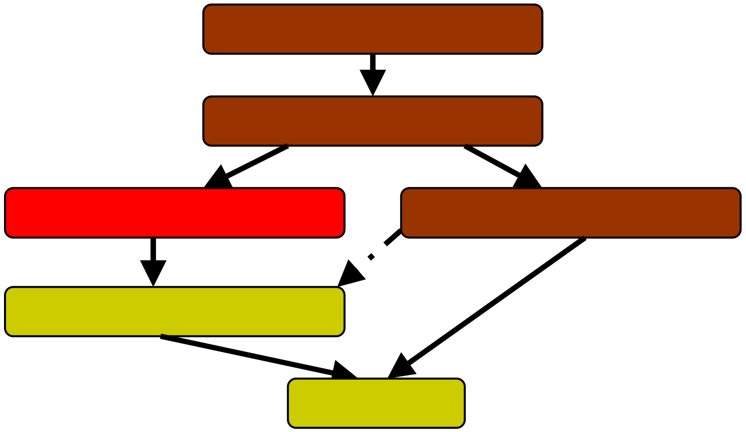
Steve Fossett and his balloon Solo Spirit arsenide.

Over the Black Sea drifting slowly towards the coast of the caucuses.

His team plans if necessary to bring him down after daylight tomorrow.

you the Chechan capital of Grozny.

- Named Entities
  - Male Person, Location (Region)



ASR → Features Example

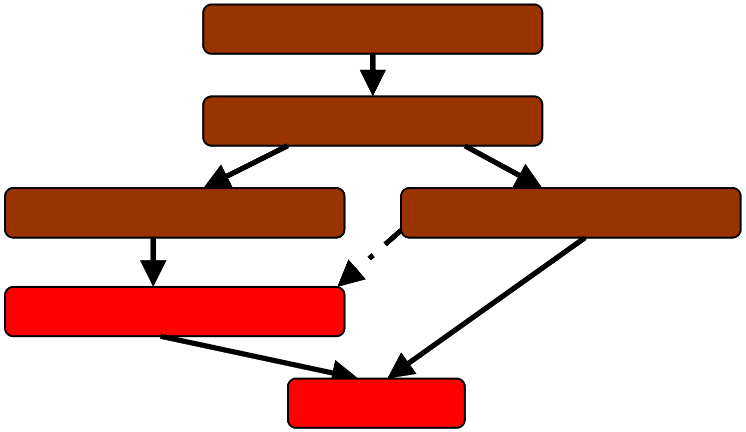
Steve Fossett and his balloon Solo Spirit arsenide.

Over the Black Sea drifting slowly towards the coast of the caucuses.

His team plans if necessary to bring him down after daylight tomorrow.

you the Chechan capital of Grozny.

- Named Entities
  - Male Person, Location (Region)
- Nouns
  - balloon, solo, spirit, coast, caucus, team, daylight, Chechan, capital, Grozny
- WordNet
  - nature



Feature Selection

- Basic feature set (nouns + NEs) has ~18,000 elements/shot
  - 6000 elements x {previous, this, next}
- Using only a subset of the possible features may affect performance.
- Two strategies for feature selection:
  - Remove very rare words (18,000 → 7902)
  - Eliminate low-value features

Information Gain

- Measures the change in entropy given the value of a single feature

$$Gain(C,F) = H(C) - \sum_{w \in Values(F)} p(w) H(C \mid F = w)$$

Information Gain Results

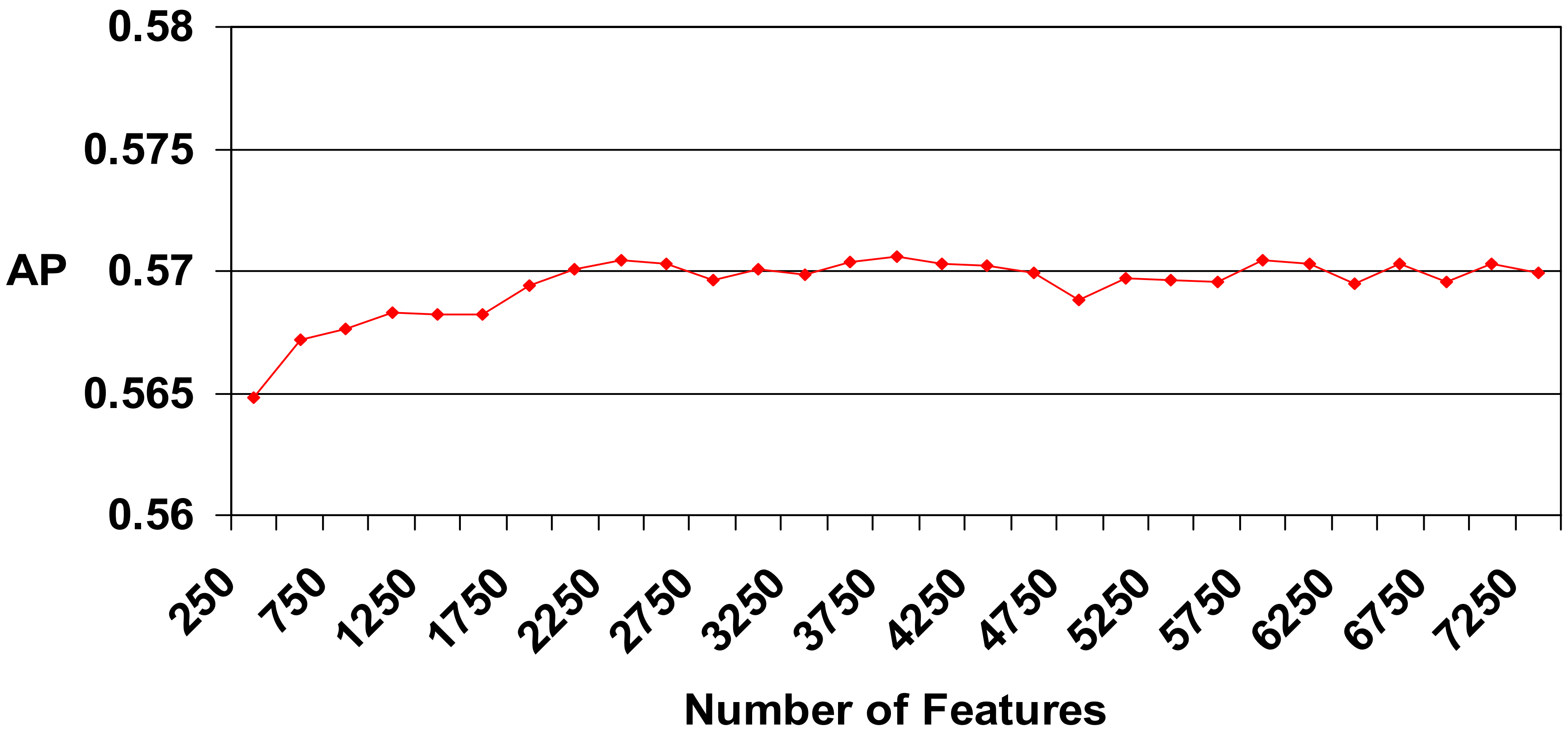
Basketball

1. (empty)
2. Location-city
3. (empty) (previous)
4. "game" (previous)
5. "game"
6. Person-male
7. "point" (previous)
8. "game" (next)
9. "basketball (previous)
10. "win"
11. (empty) (next)
12. "basketball"
13. "point"
14. "title" (previous)
15. "win" (previous)

Sky

1. Person-male (previous)
2. "car" (previous)
3. Person
4. Person-male
5. "jury"
6. Person (next)
7. (empty) (next)
8. "point"
9. "report"
10. "point" (next)
11. "change" (previous)
12. "research" (next)
13. "fiber" (previous)
14. "retirement" (next)
15. "look"

Choosing an optimal number of features



Classifiers

- Naïve Bayes
- Decision Trees
- Support Vector Machines
- Voted Perceptrons
- Language Model
- AdaBoosted Naïve Bayes & Decision Stumps
- Maximum Entropy

Naïve Bayes

- Build a binary classifier (present/absent) for each concept.

$$p(c \mid d_w) = \frac{p(d_w \mid c)\, p(c)}{p(d_w)}$$

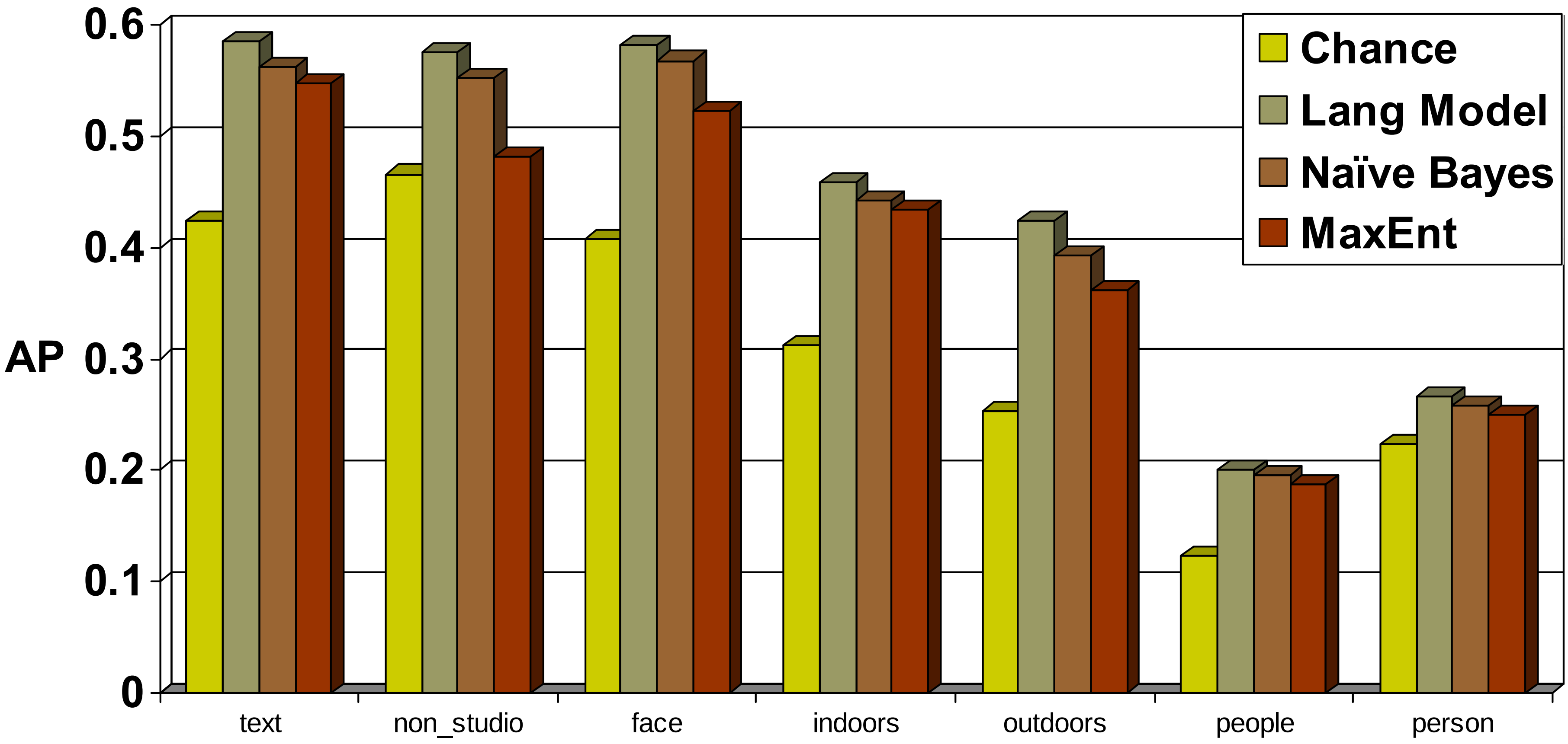
Language Modeling

- Conceptually similar to Naïve Bayes but
  - Multinomial
  - Smoothed distributions
  - Different feature selection

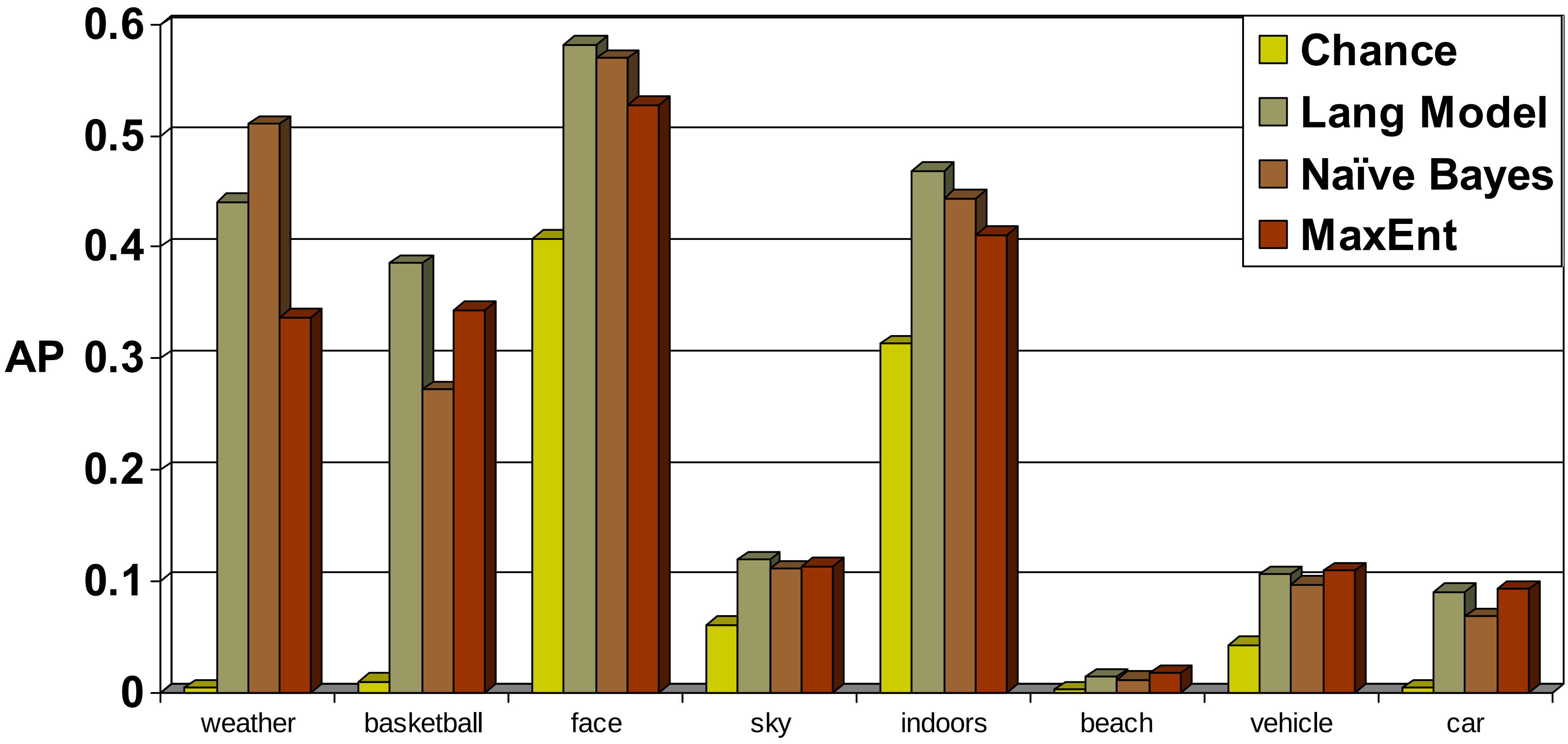
Maximum Entropy Classification

- Binary constraints
- Single 75-concept model
- Ranked list of concepts for each shot.

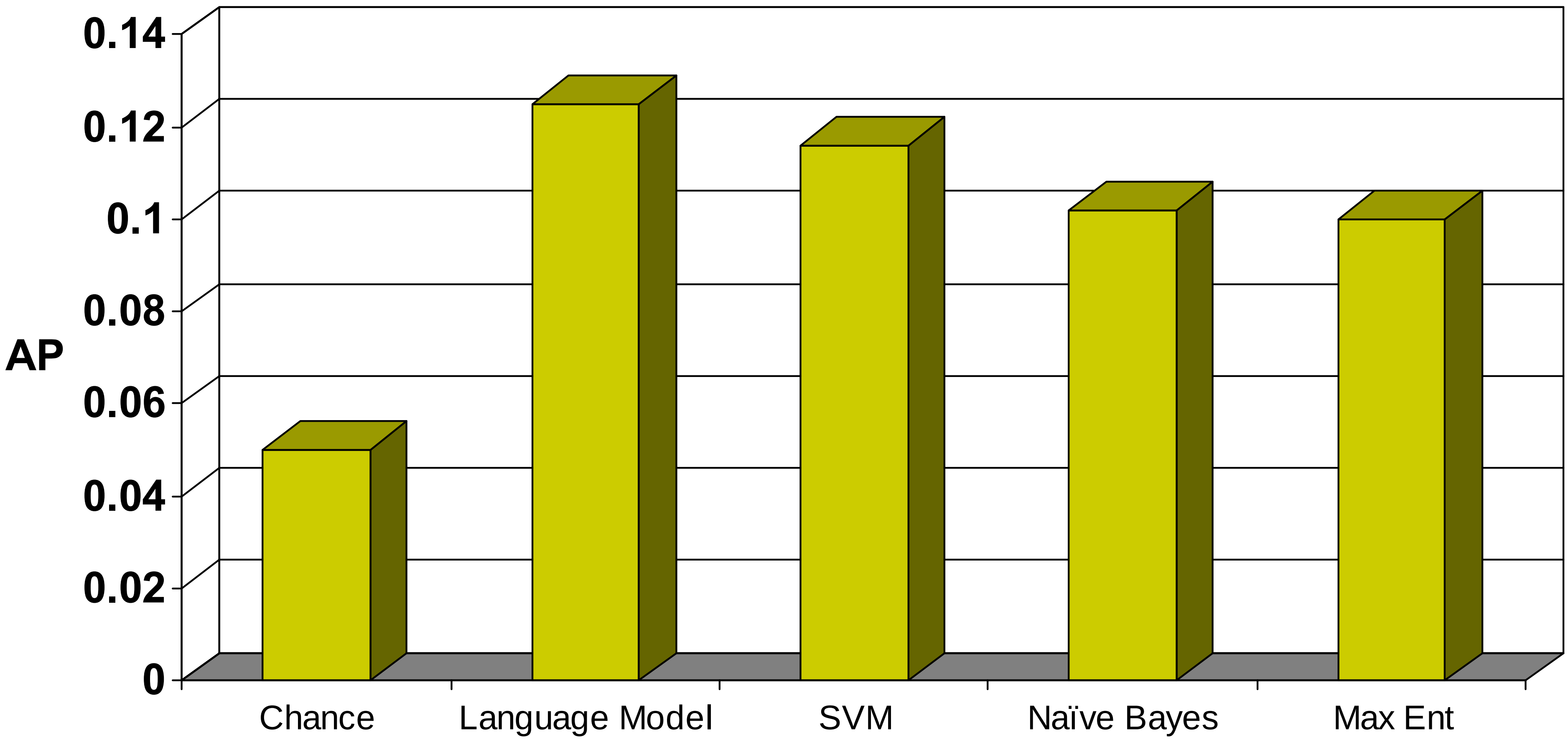
Results on the most common concepts



Results on selected concepts



Mean Average Precision



Will this help for retrieval?

- “Find shots of a person diving into some water.“
- “Find shots of the front of the White House in the daytime with the fountain running.“
- “Find shots of Congressman Mark Souder.“

Will this help for retrieval?

- "Find shots of a person diving into some water."
  - person, water_body, non-studio_setting, nature_non-vegetation, person_action, indoors
- "Find shots of the front of the White House in the daytime with the fountain running."
  - building, outdoors, sky, water_body, cityscape, house, nature_vegetation
- "Find shots of Congressman Mark Souder."
  - person, face, indoors, briefing_room_setting, text_overlay

Performance on retrieval-relevant concepts

Concept	Importance	AP	Chance
outdoors	0.68	0.434	0.270
person	0.48	0.267	0.227
sky	0.40	0.119	0.061
vehicle	0.36	0.106	0.043
face	0.28	0.582	0.414
man-made-obj.	0.28	0.190	0.156
building	0.24	0.078	0.042
road	0.24	0.055	0.037
transportation	0.24	0.151	0.065
indoors	0.24	0.459	0.317

Summary

- Predict visual concepts for ASR
- Tried Naïve Bayes, SVMs, MaxEnt, Language Models,...
- Expect improvements in retrieval

Joint Visual-Text Video OCR

Proposed by:
Matthew Krause
Georgetown University

Motivation

- TREC queries ask for:
  - specific persons
  - specific places
  - specific events
  - specific locations

Motivation

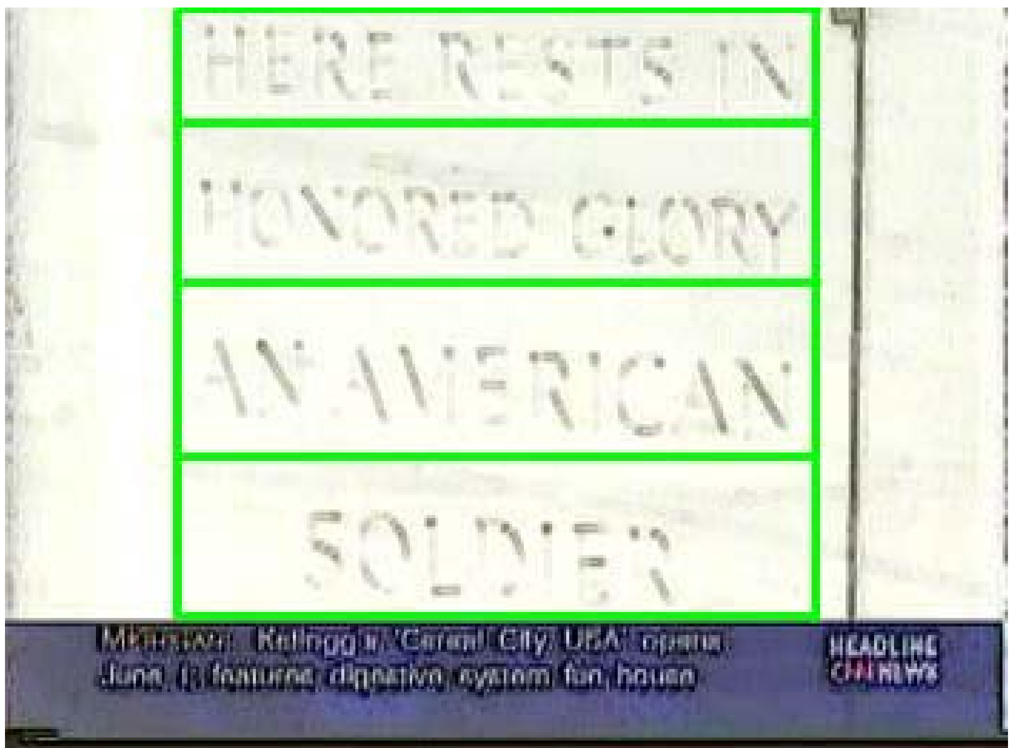
- "Find shots of Congressman Mark Souder"

Motivation

- "Find shots of a graphic of Dow Jones Industrial Average showing a rise for one day. The number of points risen that day must be visible."

Motivation

- Find shots of the Tomb of the Unknown Soldier in Arlington National Cemetery.



Motivation

WEIFll I1 NFWdJ TNNIF H

Joint Visual-Text Video OCR

- Goal: Improve video OCR accuracy by exploiting *other* information in the audio and video streams during recognition.

Why use video OCR?

- …. Sources tell C.N.N. there's evidence that links those incidents with the January bombing of a women's health clinic in Birmingham, Alabama. Pierre Thomas joins us now from Washington. He has more on the story in this live report…

Why use video OCR?

Why use video OCR?

Why use video OCR?

Those links are growing more intensive investigative focus toward fugitive Eric Rudolph who's been charged in the Birmingham bombing which killed an off-duty policeman…

Why use video OCR?

- Text overlays provide high precision information about query-relevant concepts in the *current* image.

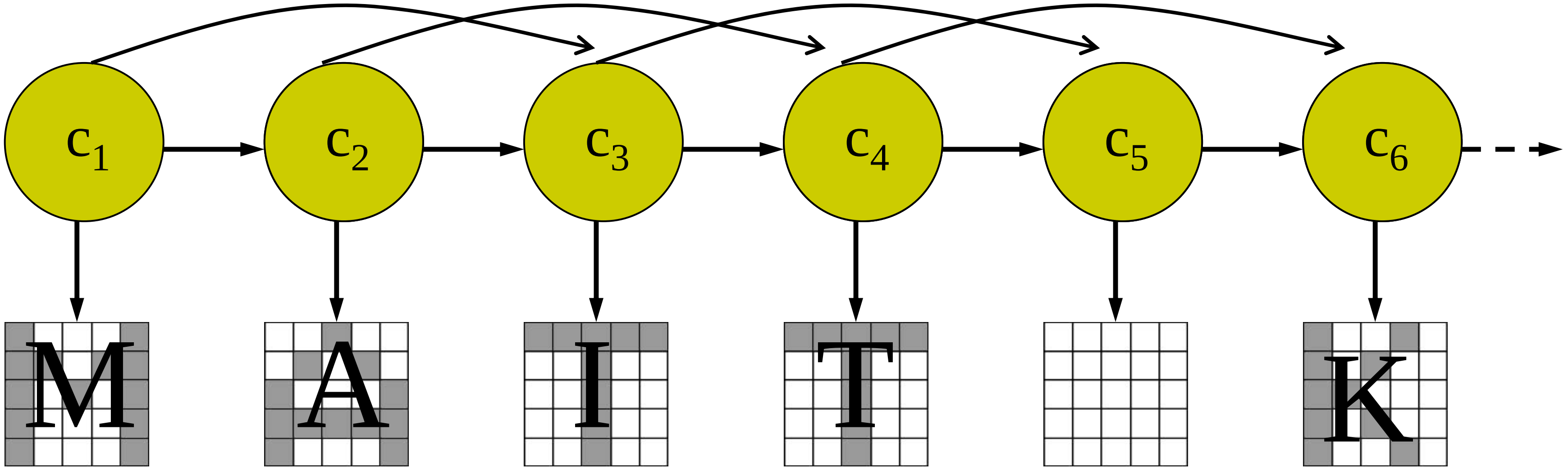
Finding Text

- Use existing tools and data from IBM/CMU.

Image Processing

- Preprocessing
  - Normalize the text region's height
- Feature extraction
  - Color
  - Edge Strength and Orientation

Proposal: HMM-based recognizer



Proposal: Cache-based LMs

- Augment the recognizers with an interpolation of language models
  - Background language model
  - Cache-based language model
    - ASR or closed caption text
  - "Interesting" words from the cache
    - Named Entities

$$p(c_i \mid h) = p_{bg}(c_i \mid h)^{\lambda_1} \, p_{cache}(c_i \mid h)^{\lambda_2} \, p_{interest}(c_i \mid h)^{\lambda_3}$$

Evaluation

- Evaluate on TRECVID data
- Character Error Rate
  - Compare vs. manual transcriptions
- Mean Average Precision
  - NIST-provided relevance judgments

Summary

- Information from text overlays appears to be useful for IR.
- General character recognition is a Hard problem.
- Adding in external knowledge sources via the LMs should improve accuracy.

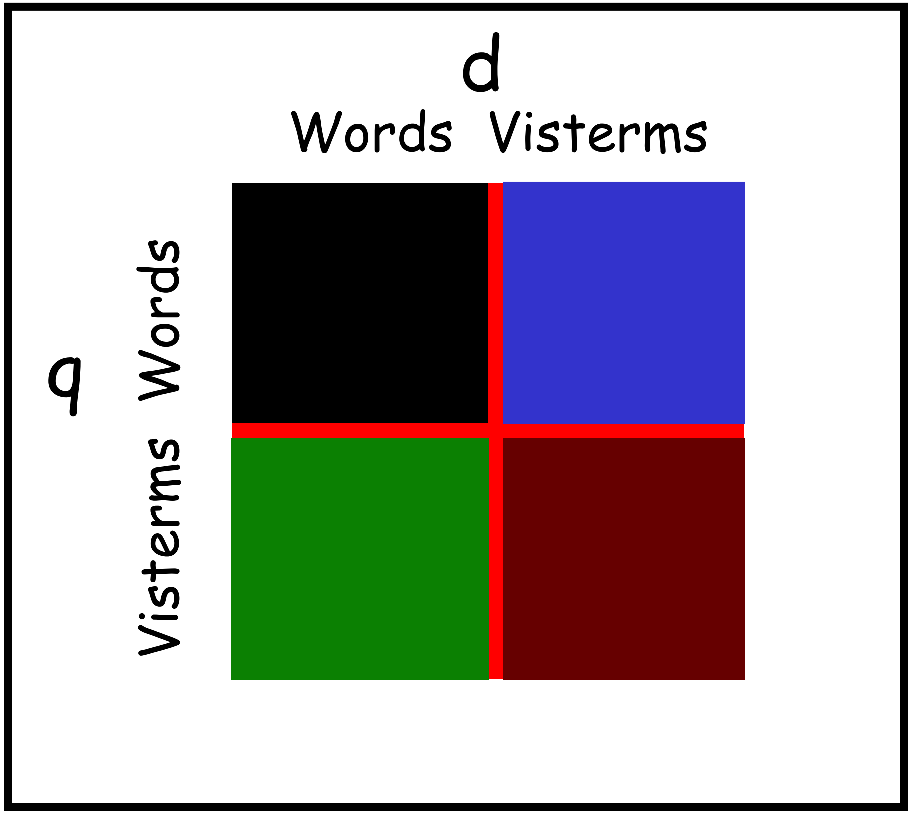
Work Plan

1. Text Localization
   - IBM/CMU text finders + height normalization
2. Image Processing & Feature Extraction
   - Begin with color and edge features
3. HMM-based Recognizer
   - Train using TREC data with hand-labeled captions
4. Language Modeling
   - Background, Cache, and "Interesting Words"

Retrieval Experiments and Summary

Presented by Dietrich Klakow

Presentation Outline



Translation (MT) models (Paola),

Relevance Models (Shao Lei, Desislava),

Graphical Models (Pavel, Brock)

Text classification models (Matt)

Integration & Summary (Dietrich)

The Matrix

Document

Words d_w | Visterms d_v

Query

Words q_w

Visterms q_v

$$p(q_w, q_v \mid d_w, d_v)$$

The Matrix

Document

Query		Words d_w	Visterms d_v
	Words q_w	$p(q_w \mid d_w)$	$p(q_w \mid d_v)$
	Visterms q_v	$p(q_v \mid d_w)$	$p(q_v \mid d_v)$

The Matrix

Query \ Document	Words d_w	Visterms d_v
Words q_w	$p(q_w \mid d_w)$	$p(q_w \mid d_v)$ •MT •Relevance Models •HMM
Visterms q_v	$p(q_v \mid d_w)$ •Naïve Bayes •Max. Ent •LM •SVM, Ada Boost, …	$p(q_v \mid d_v)$

Retrieval Model I: p(q|d)

$$p(q_w, q_v \mid d_w, d_v)$$

$$= p(q_w \mid d_w, d_v) \times p(q_v \mid d_w, d_v)$$

Text Query

Baseline: Standard text-retrieval

Image Documents

$$p(q_w \mid d_w, d_v)$$

$$= \lambda_w \boxed{p(q_w \mid d_w)} + (1-\lambda_w)\boxed{p(q_w \mid d_v)}$$

Retrieval Model I: p(q|d)

$$
\begin{aligned}
& p(q_w, q_v \mid d_w, d_v) = \\
& [\lambda_w \boxed{p(q_w \mid d_w)} + (1-\lambda_w) \boxed{p(q_w \mid d_v)}] \times \\
& [\lambda_v \boxed{p(q_v \mid d_w)} + (1-\lambda_v) \boxed{p(q_v \mid d_v)}]
\end{aligned}
$$

α Only minor improvements over baseline

Retrieval Model II: p(q|d)

- We want to estimate $p(q_w, q_v, d_w, d_v)$
- Assume pairwise marginals given:

$$\sum_{q_v, d_w} p(q_w, q_v, d_w, d_v) = p(q_w, d_v)$$

- Setting: Maximum Entropy problem
  - 4 constraints
  - 1 iteration of GIS:

$$p(q_w, q_v \mid d_w, d_v) \propto p(q_w \mid d_w)^{\lambda_1} p(q_w \mid d_v)^{\lambda_2} p(q_v \mid d_w)^{\lambda_3} p(q_v \mid d_v)^{\lambda_4}$$

Baseline TRECVID: Text Retrieval

Document

Query		Words d_w	Visterms d_v
	Words q_w	$p(q_w \mid d_w)$	
	Visterms q_v		

Retrieval mAP: 0.131

Combination with visual model

Query		Document: Words d_w	Document: Visterms d_v
	Words q_w	$p(q_w \mid d_w)$ **mAP: 0.131**	$p(q_w \mid d_v)$
	Visterms q_v		

Combination with visual model

Document

Query		Words d_w	Visterms d_v
	Words q_w	$p(q_w \mid d_w)$ **mAP: 0.131**	$p(q_w \mid d_v)$ MT
	Visterms q_v		

Concept Annotation on images
mAP on TRECVID

MT	0.126
Relevance Models	0.158
HMM	0.145

MT: Best overall performance so far

Retrieval mAP: 0.139

Combination with MT and ASR

Query	Document: Words d_w	Document: Visterms d_v
Words q_w	$p(q_w \mid d_w)$ **mAP: 0.131**	$p(q_w \mid d_v)$ MT
Visterms q_v	$p(q_v \mid d_w)$	

Concept Annotation
on images:
mAP on TRECVID

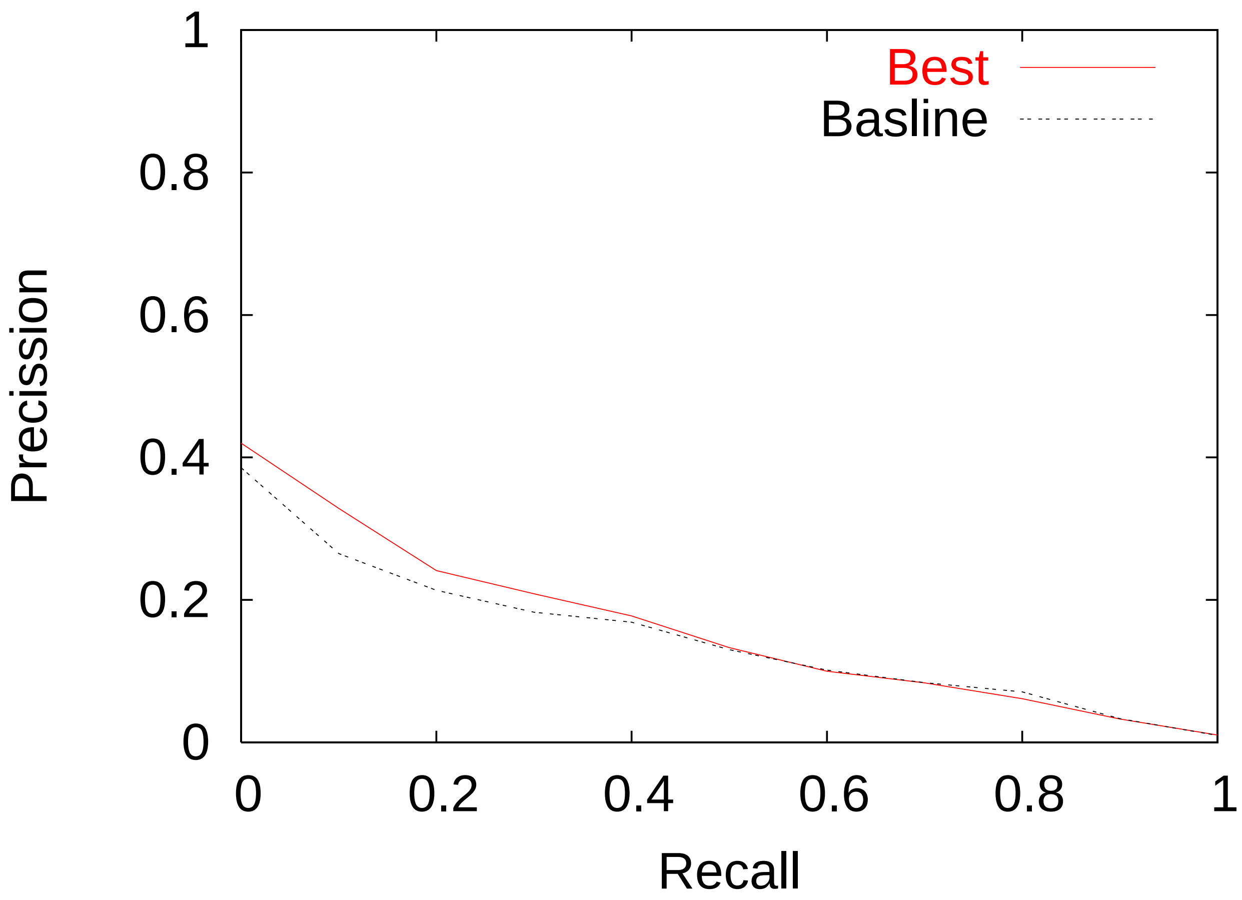
MT	0.126
Relevance Models	0.158
HMM	0.145

Concepts from ASR: mAP=0.125

Retrieval mAP: 0.149

Best results reported in literature: retrieval mAP=0.162

Recall-Precision-Curve



Improvements in high precision region

Difficulties and Limitations we faced

- Annotations are
  - Inconsistent, sometimes abstract, …
- Used plain vanilla features
  - Color, texture, edge on key-frames
  - No time for exploration of alternatives
- Uniform block segmentation of images
- Upper bound for concepts from ASR

Future Work

- Model
  - Incompletely labelled images
  - Inconsistent annotations

Desislava

- Get beyond the 75-concept bottleneck
  - Larger concept set (+training data)
  - Direct modelling
- Better model for spatial and temporal dependencies in video

Shaolei and Brock

- Query dependent processing
  - E.g. image features, combination weights, OCR-features

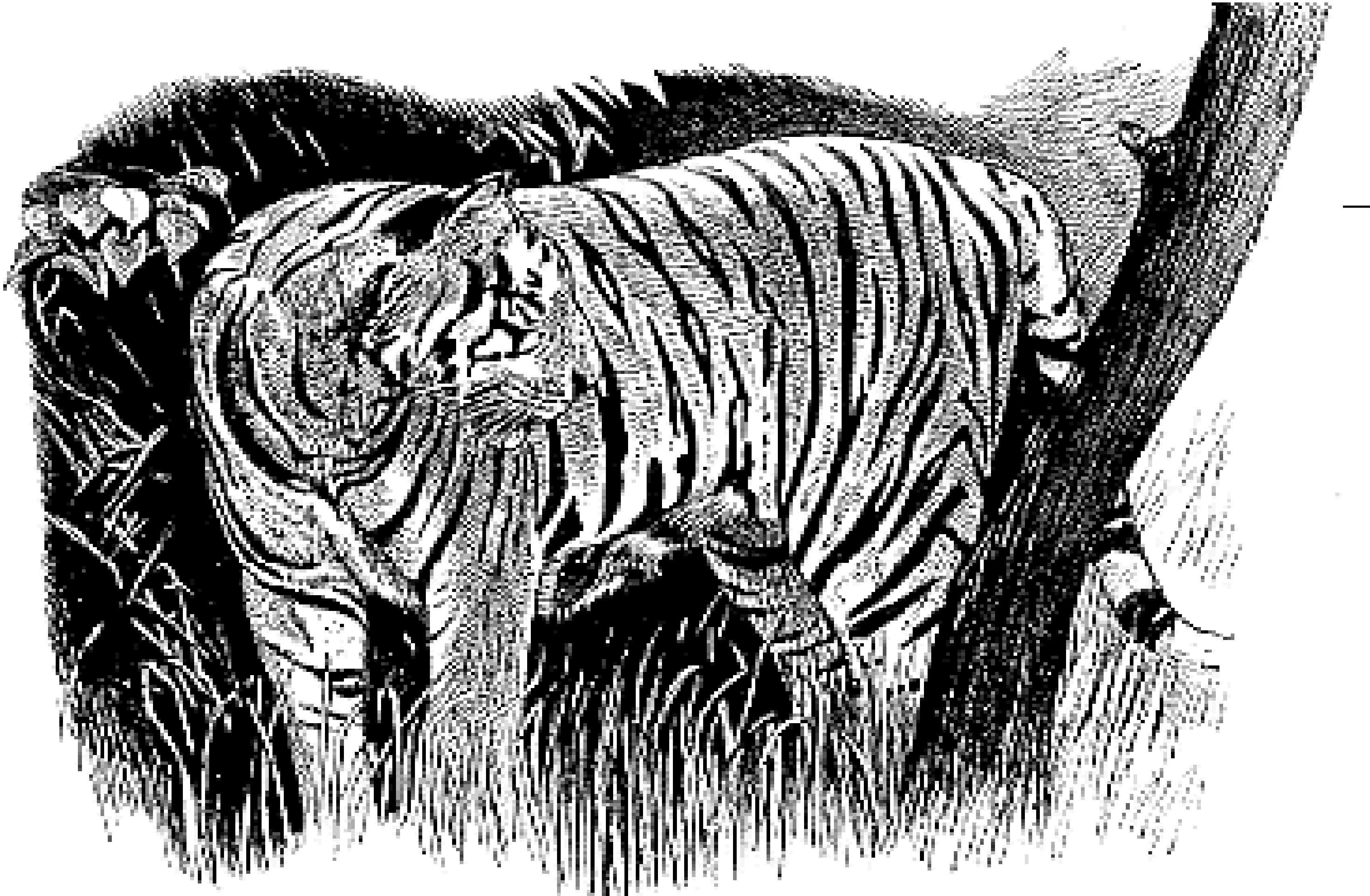
Matt

Overall Summary

- Concepts from image
  - MT: CLIR with direct translation works best
  - Relevance models: best numbers on development test
  - HMM: novel competitive approach for image annotation
- Concepts from ASR:
  - oh my god, it works
- Fusion:
  - adding multiple source in log-linear combination helped
- Overall: 14% improvement

Acknowledgments

- TREC for the data
- BBN for NE-tagging
- IBM:
  - for providing the features
  - Close captioning alignment (Arnon Amir)
- Help with GMTK: Jeff Bilmes and Karen Livescu
- CLSP for the capitalizer (WS 03 MT-team)
- INRIA for the face detector
- NSF, DARPA and NSA for the money
- CLSP for hosting
  - Laura, Sue, Chris
  - Eiwe, John, Peter
  - Fred



From: http://www.nature.ca/notebooks/english/tiger.htm