CIVR 2004

Pattern Mining in Large-Scale Image and Video Sources

Prof. Shih-Fu Chang

Digital Video and Multimedia Lab
Columbia University

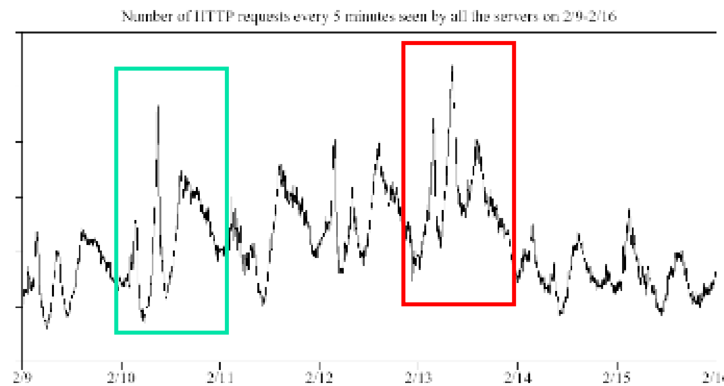
July 21, 2004

http://www.ee.columbia.edu/dvmm

- Joint work
  - Lexing Xie, Lyndon Kennedy (Columbia U.)
  - Ajay Divakaren and Huifan Sun (MERL)
- Supported in part by
  - MERL
  - ARDA VACE phase II
    IBM-Columbia Semantrix Project

Temporal Patterns Everywhere …

music motifs



Internet traffic patterns

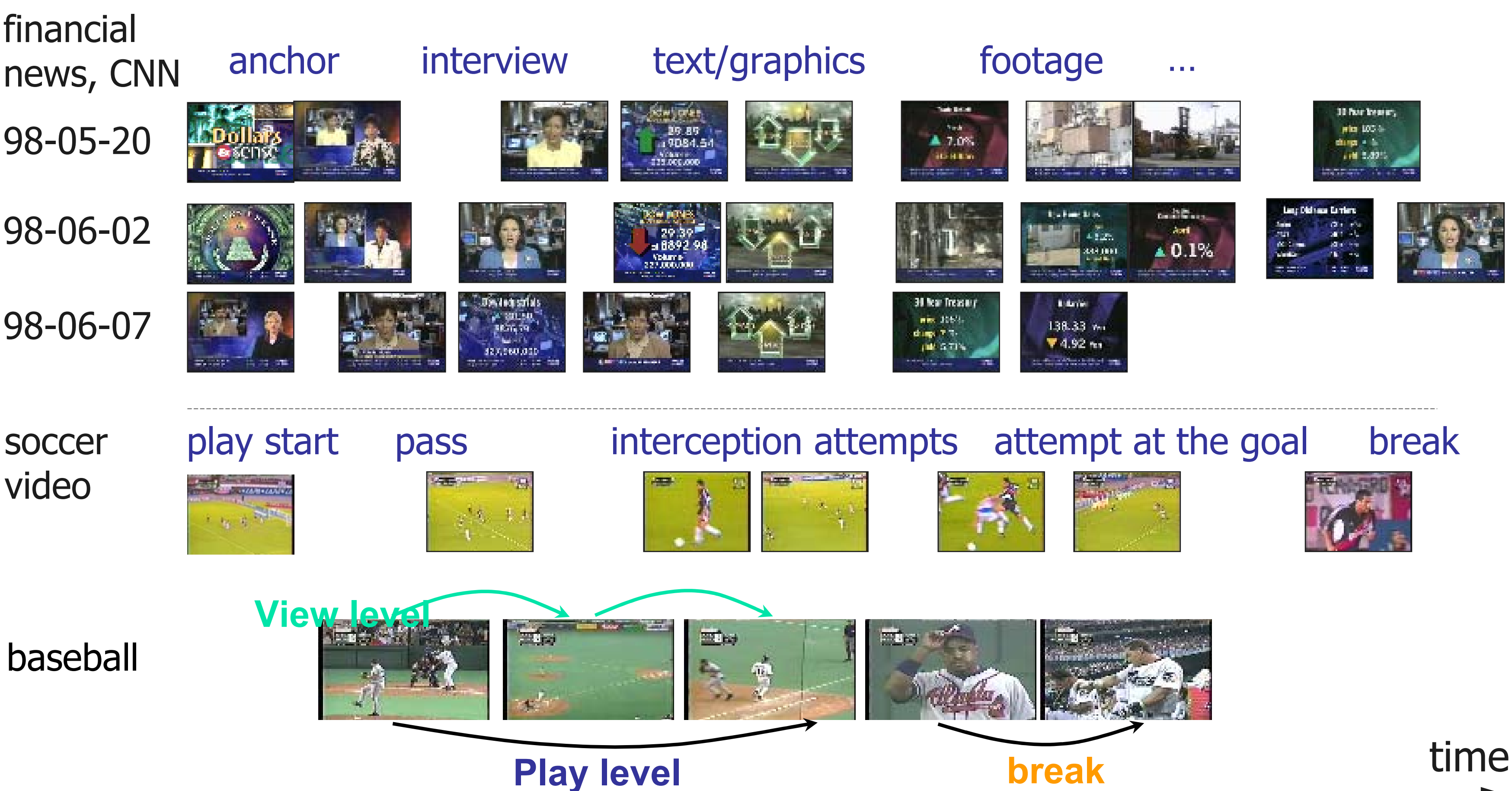
Fis Protein	QTRAALMM	GIN	RGTLRKKLK
λ Rep	QESVADKM	GMG	QSGVGALFN
λ Cro	QTKTAKDL	GVY	QSAINKAIH
434 Cro	QTELATKA	GVK	QQSIQLIEA

protein motifs

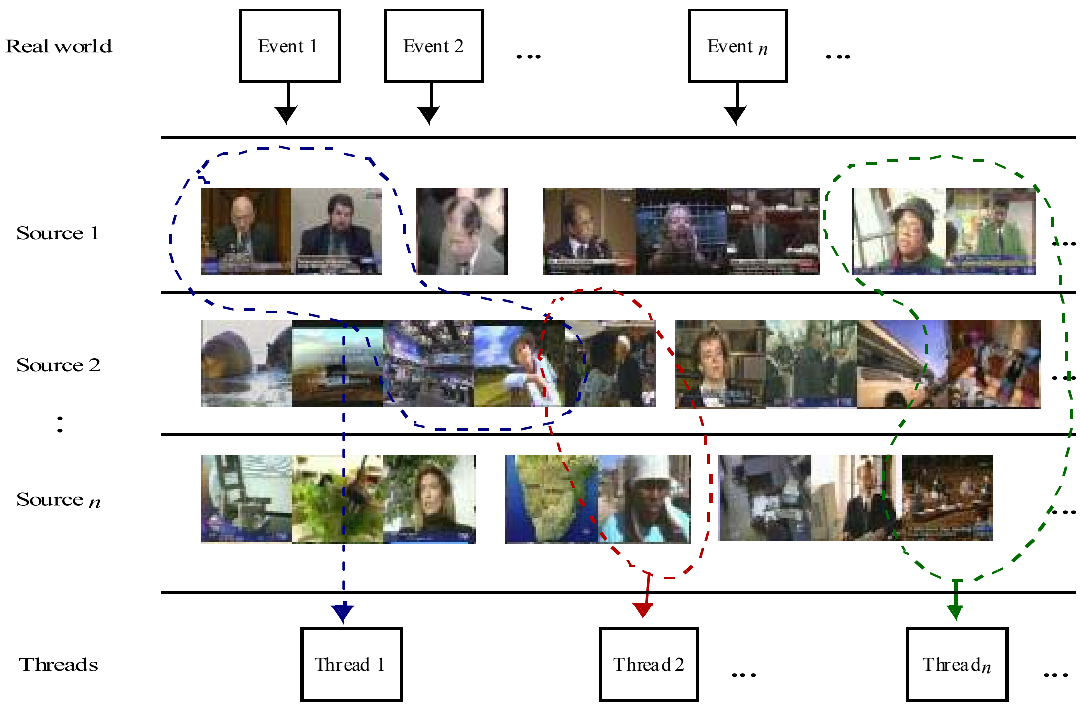
[Iyengar99]
[www.music-scores.com][Purdue Stat490B]

- There are interesting patterns indicating useful information in different domains.

Example Patterns in Video



Patterns Across News Sources



Example
Marines Extend Duty

CNN News

Chinese News

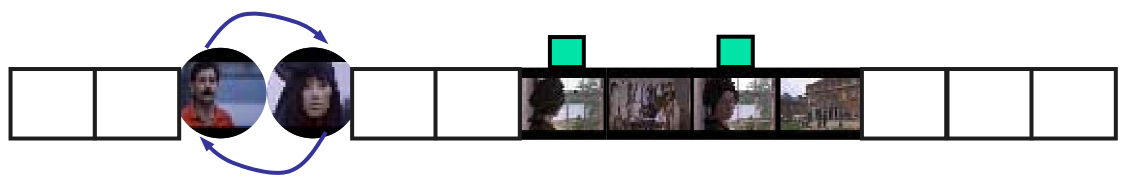
- Story/event often re-occur within or across channels
  - Semantic thread reconstruction (IBM-CU ARDA VACE II project)

Types of Temporal Patterns

- Dense and stochastic

Focus of the talk

- Sparse and deterministic



- Temporal association of events
  - If A occurs in channel 1, then B occurs within time T in channel 2,
    e.g. recurrent news topics
- Others …

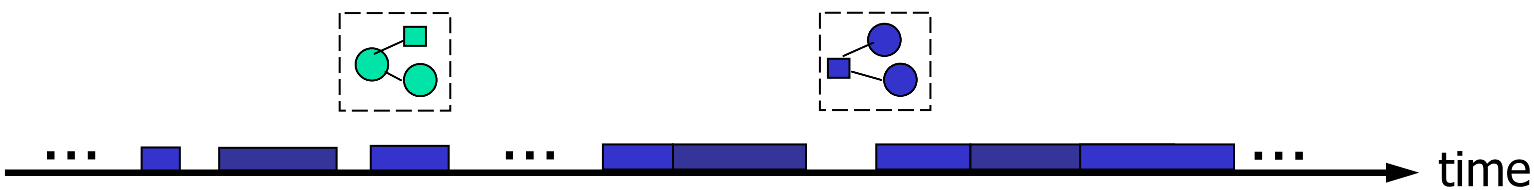
(Grand) Challenges for Pattern Mining

- **There are many patterns in video at different levels.**
  - **How do we discover them (semi)-automatically?**
  - **Do they correspond to any semantic meanings?**
  - **What are the underlying features/structures of each pattern?**
  - **Which patterns are more promising for developing classifiers?**

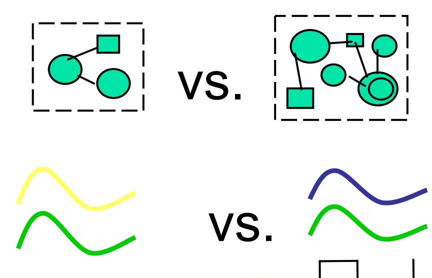
- **Why do these matter?**
  - **Unsupervised discovery of a large number of patterns**
    → discover interesting, meaningful concepts & events
    → help define normal states and novelty
  - **Scalability to new content and domains**

Challenge: Unsupervised Pattern Discovery

- Given a new domain/corpus, discover patterns automatically
  - E.g., News, consumer, surveillance, and personal life log
- Technical Issues:
  - Find appropriate spatio-temporal statistical models
  - Locate segments that fit such models

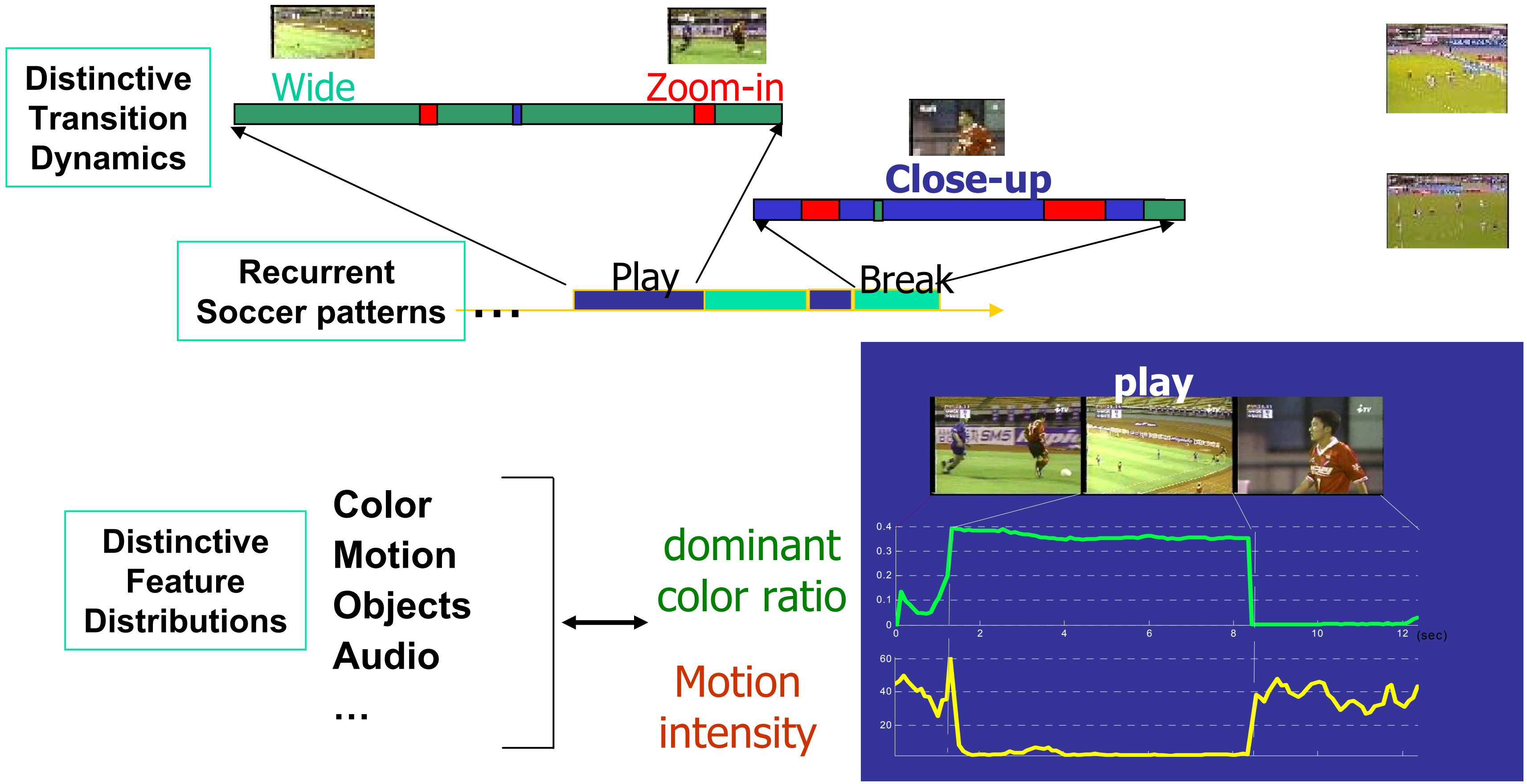


- **Issues**
  - **What's the adequate class of models?**
  - **How to determine model structures?**
  - **What are "good" features?**



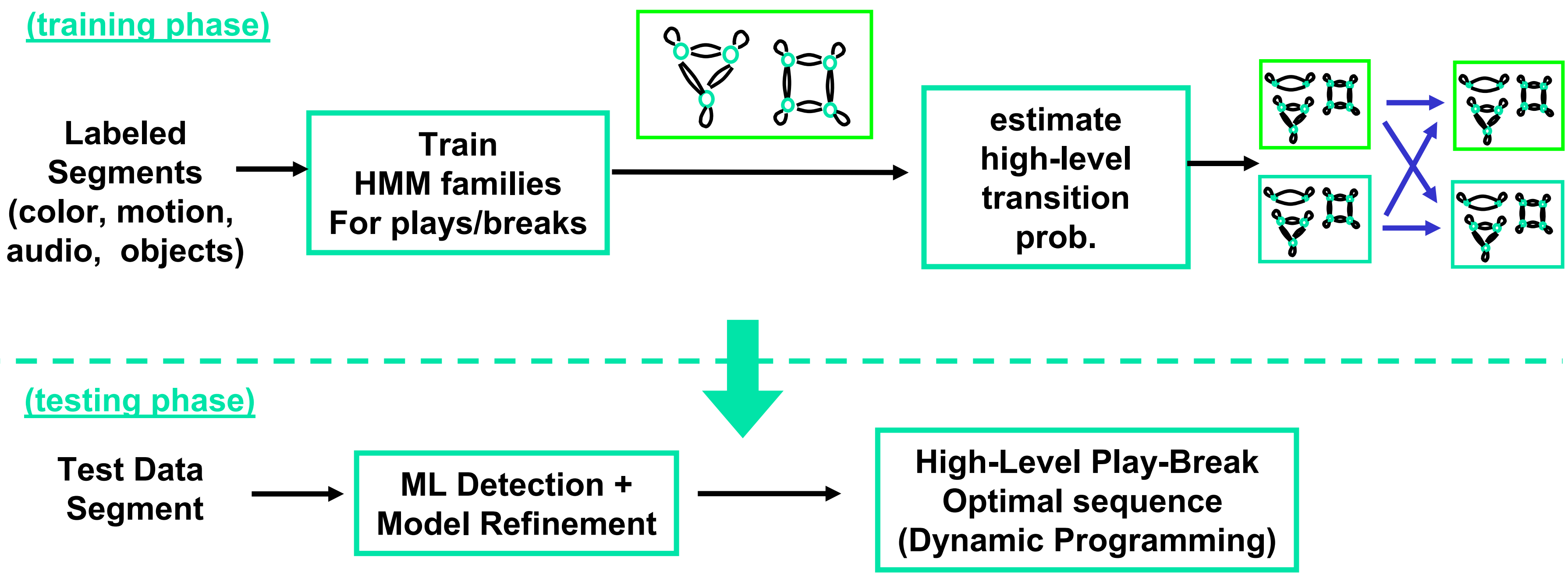
- Finding the right class of models …
  - Lessons from prior work
    supervised + unsupervised

Many video temporal patterns are characterized by *Dynamics* and *Features*



- **Distinctive patterns are characterized by state-dependent transitions and features.**
- **Like speech recognition, HMM and variants may serve as promising candidates.**

HMM Models of Temporal Patterns in Soccer



Supervised HMM models are effective for parsing structures in sports video

(Xie et al ‘02)

- Sports video (4 test clips, 15~25 min., various countries)
- Cross Validation

Test set	Training Set			
	Argentina	KoreaA	KoreaB	Espana
Argentina	87.2%	82.5%	82.5%	80.6%
KoreaA	78.1%	84.3%	84.3%	79.8%
KoreaB	79.9%	85.3%	85.3%	89.6%
Espana	79.9%	89.6%	89.6%	81.7%

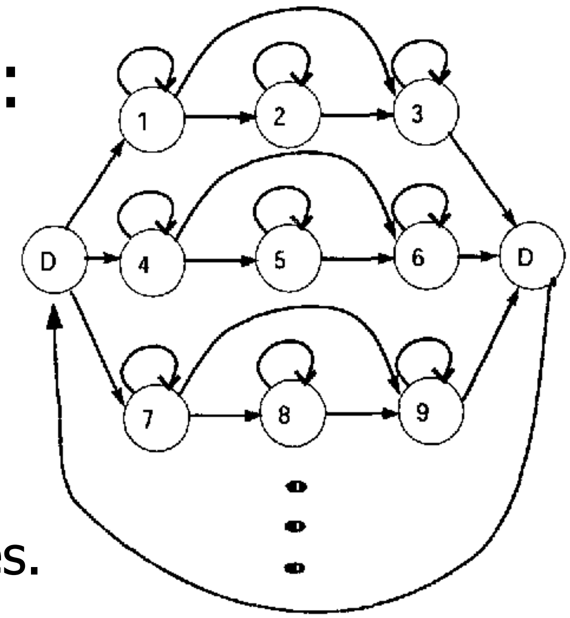
- **Avg. Play-Break Classification Rate: 83.5% vs. 60% of blind guessing**
- **Boundary timing accuracy: 62% within 3 seconds**
- **The good classification provides preliminary evidence supporting HMM model and the A-V features Demo**

Columbia University in the City of New York

dvmm DIGITAL VIDEO MULTIMEDIA LAB

Applying HMM to unsupervised mining

- With straightforward structures: multi-path left-right HMM



[Clarkson et al '99]

- Long ambulatory videos captured with wearable device
- Color histogram and MFCC features at 10Hz
- Cross-correlation coefficients 0.7~0.9 between ground-truth and likelihood sequences.

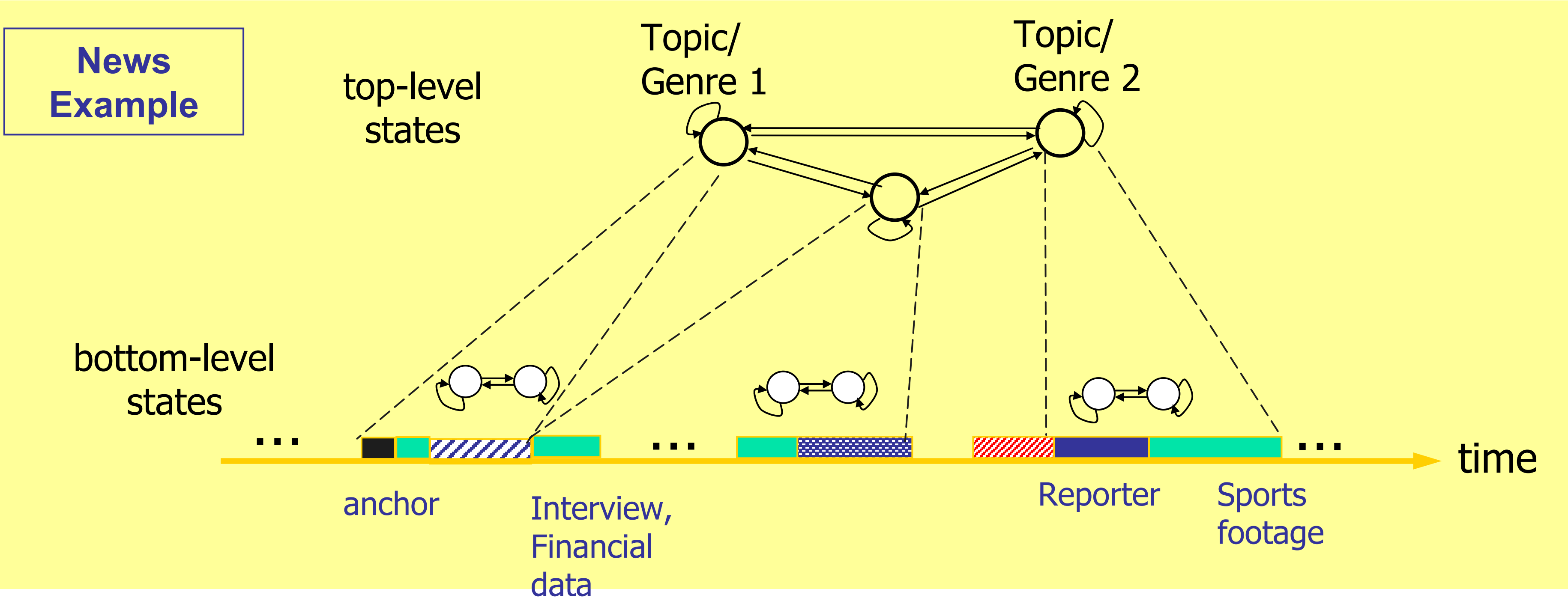
[Naphade et al '02]

- Films and talk shows
- Color and edge histogram, MFCC and energy features at 30Hz
- discover recurrent patterns of explosion and applause

Pattern Mining Using Hierarchical HMM

(Xie et al '02)

- Intuitive Representation for Video Patterns
  - Patterns occur at different levels following different transition models
  - States in each level may correspond to different semantic concepts

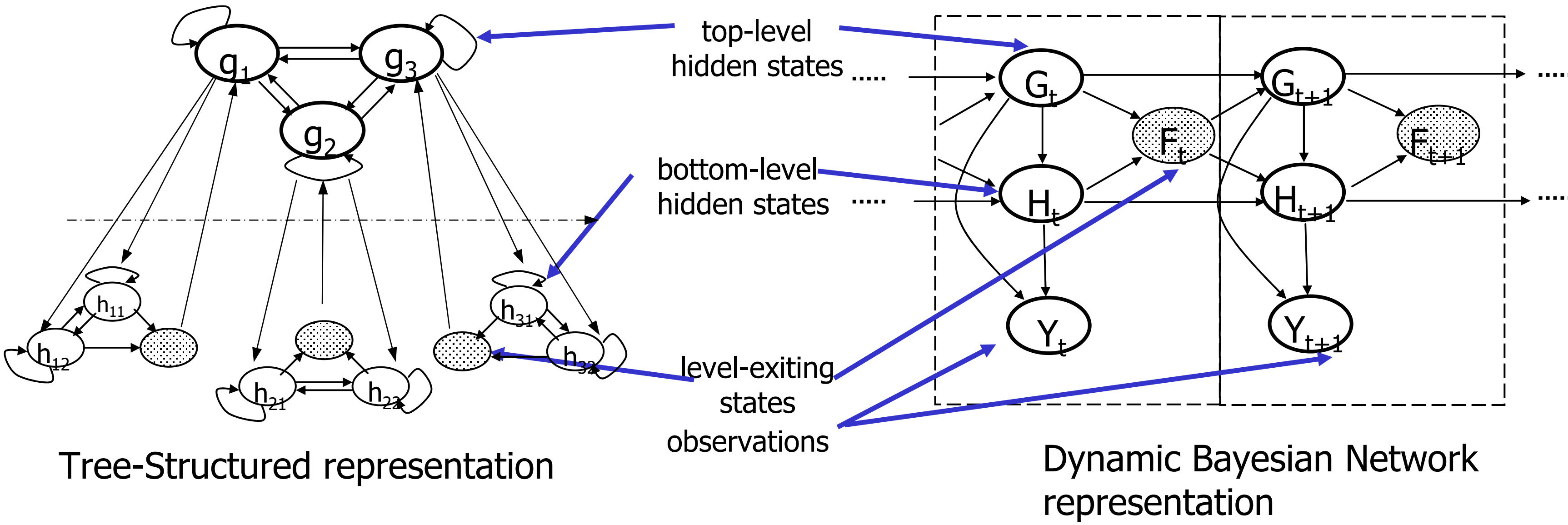


Columbia University in the City of New York

dvmm DIGITAL VIDEO MULTIMEDIA LAB

Hierarchical HMM

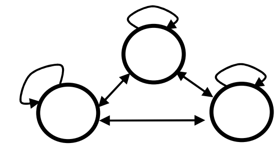
[Fine, Singer, Tishby '98] [K. Murphy, '01] [Xie et al '02]



- Flexible control structure (bottom-up control with exit state)
- Extensible to multiple levels and distributions
- Efficient inference technique available
  - Complexity $O(D \cdot T \cdot Q^{\alpha D})$, α=1.5 to 2
- Application in unsupervised discovery has not been explored
  - Questions: how to find right model structures and feature sets?

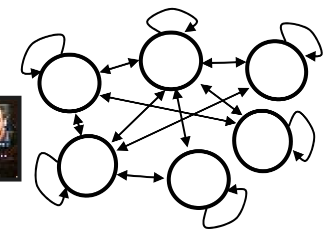
The Need for Model Selection

soccer



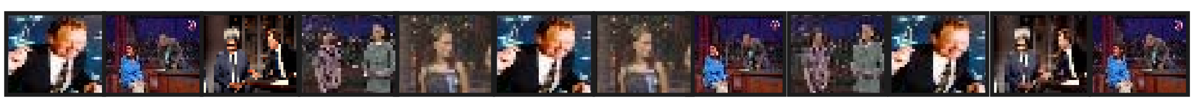
?

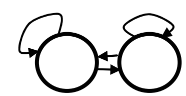
news



?

talk show





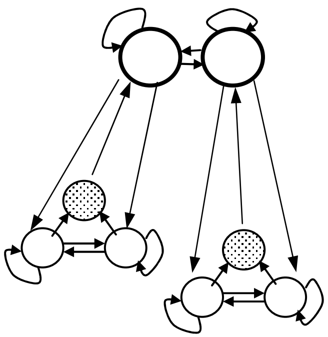
?

- Different domains have different descriptive complexities.

[Green95]
[Andrieu99]
[Xie ICME03]

Model Selection with RJ-MCMC

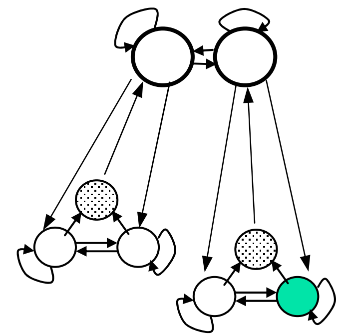
Default HHMM



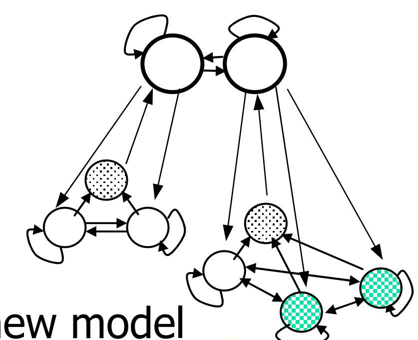
Possible Model Operations

- EM
- Split
- Merge
- Swap

(move, state)=(split, 2-2)



new model



Prob. Thresh. for accepting change = (e^{BIC} ratio)x (proposal ratio)xJ

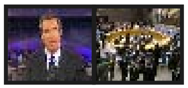
Accept proposal?

next iteration

stop

Optimum points: balance (data fitness) + (model complexity)

Which Features Shall We Use?



color histogram

edge histogram

Gabor wavelet descriptors

motion estimates

zernike moments

face?

people?

outdoors?

vehicle?

MFCC

zero-crossing rate

LPC coeff.

delta energy

Spectral rolloff

pitch

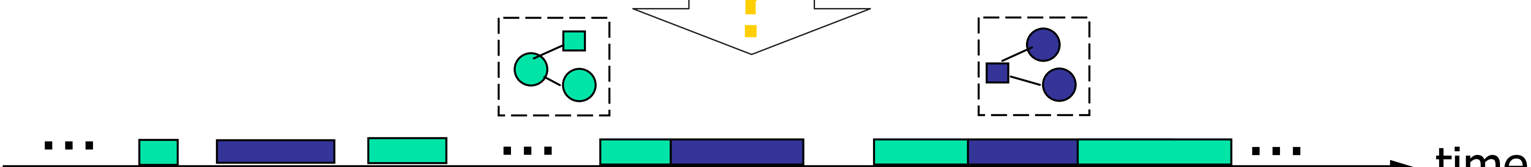
logtf-entropy

keywords

tf-idf

jennings nasdaq juri accus
tonight clinton lawyer

?



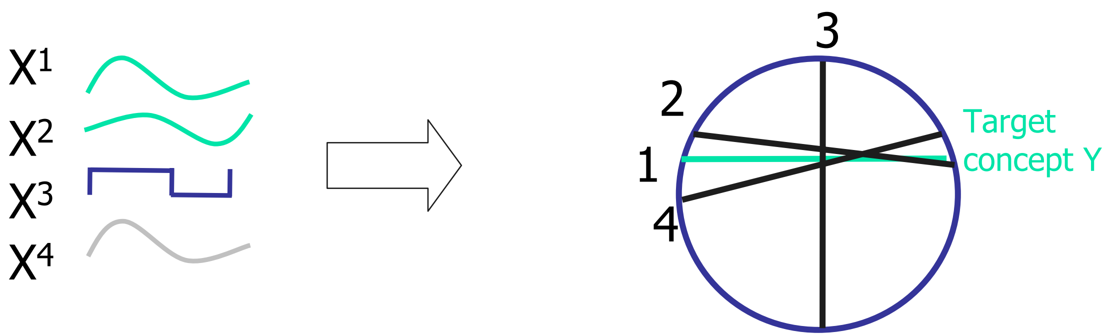
Columbia University in the City of New York

dvmm DIGITAL VIDEO MULTIMEDIA LAB

[Koller,Sahami'96] [Zhu et.al.'97]
[Xing, Jordan'01] [Ellis,Bilmes'00]…

Issues of Feature Selection

Goal: To identify a good subset of observations in order to improve model generalization and reduce computation.



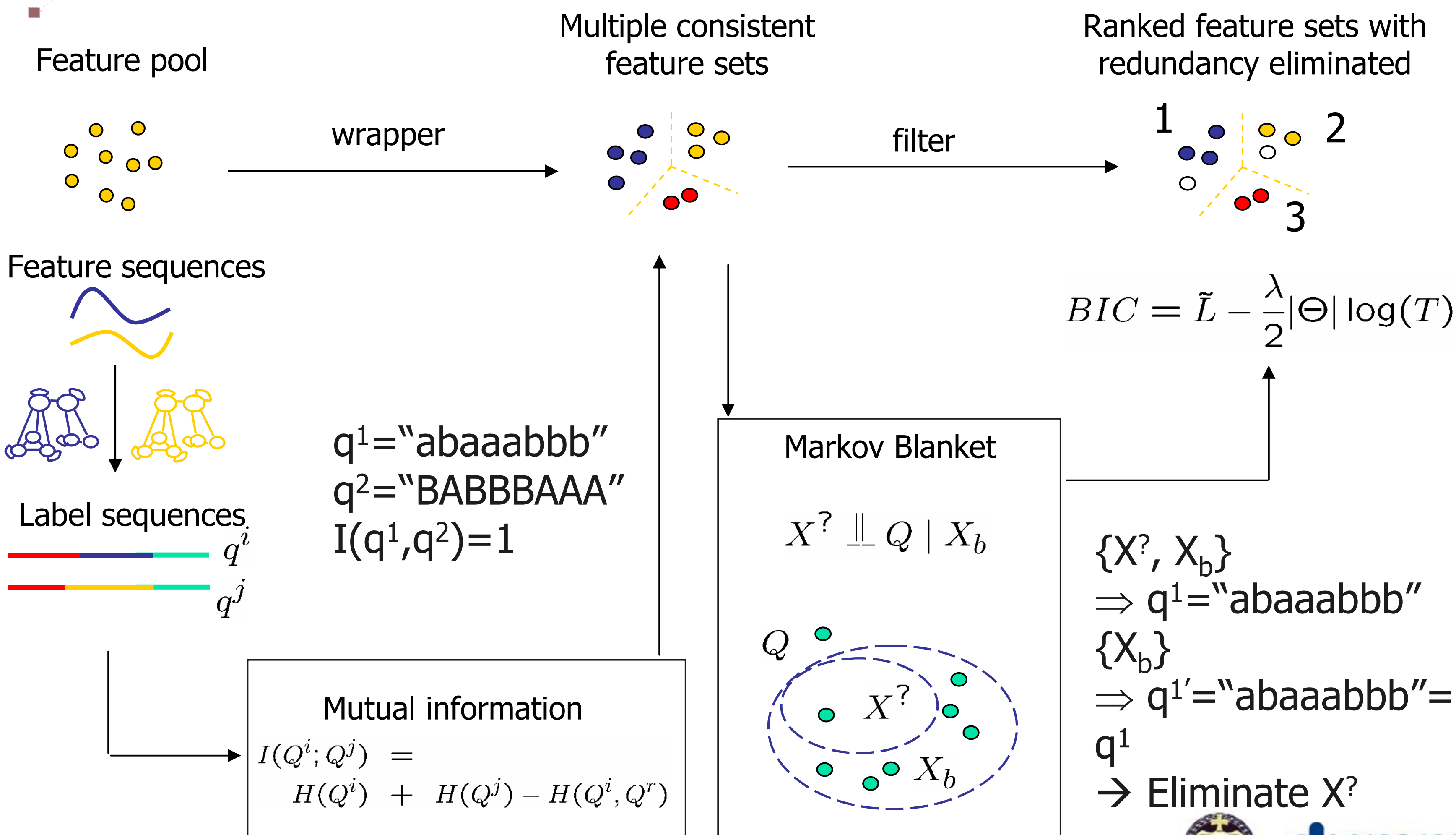
Criteria: (1) find the feature-feature or feature-concept relevance
(2) eliminate any redundancy

Unique problems for temporal sequence mining:

Unsupervised →	No target concept defined a priori
Temporal sequence →	Temporal samples not i.i.d.

Feature Selection for Temporal Pattern Mining

[Koller'96] [Xing'01]
[Xie et al. ICIP'03]



Results: on Sports Videos

videos

baseball

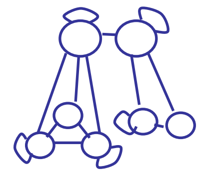
soccer

features

visual: dominant color ratio, camera translation motion

audio: energies, zero-crossing rate, spectral rolloff

HHMM



patterns

HHMM top-level label sequence

$q^*_{1:T}$

vs. play/break?

A Simple Test: Mining Baseball Videos

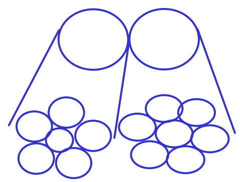
videos

baseball

Ranking based on BIC score

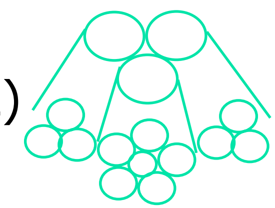
HHMM + feature selection

(1)



dominant color ratio
horizontal motion
(vertical motion)

(2)



audio volume
low-band energy

(3) … …

Semantics of patterns?

Correspond to play/break with 82.3% accuracy

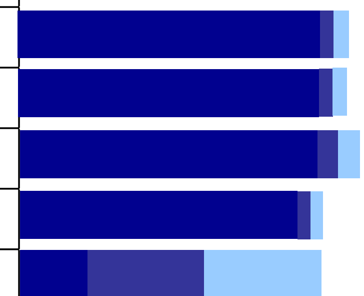
?

(demo)

Unsupervised Mining not less Effective than Supervised Learning

Fixed features {DCR, MI}, MPEG-7 Korean Soccer video

Model	Supervised?	Model Selection	Correspondence w. Play/Break
HHMM	N	Y	75.2±1.3%
HHMM	N	N	75.0±1.2%
HMM	Y	N	75.5±1.8%
LR-HHMM	N	N	73.1±1.1%
K-Means	N	N	64.0±10.%



Automatic selection of both model and features

Test clip	Feature Set	# "events"	Correspondence w. Play/Break
Korea	DCR,Mx	2~4	75.2%
Spain	DCR,Volume	2~3	74.8%
Baseball	DCR,Mx	2	82.3%

* DCR='dominant-color-ratio', MI='motion-intensity', Mx='horizontal-camera-pan'

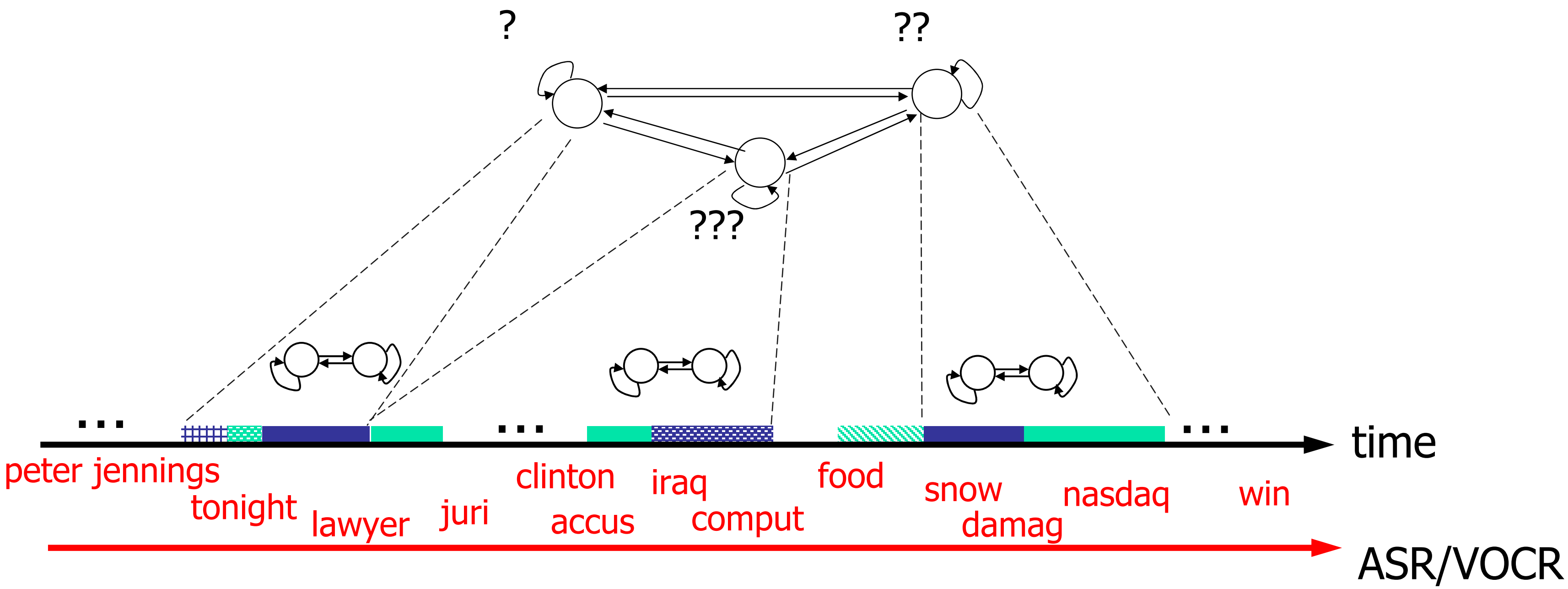
- HHMM seems promising in finding statistical temporal patterns.
- But how to find the meanings of the patterns?
- Approach → fuse the metadata streams when available.

Outline

- The problem
- Unsupervised pattern discovery with HHMM
- Finding meaningful patterns
  - With text association
  - By multi-modal fusion
- Summary

Towards Meaningful Patterns

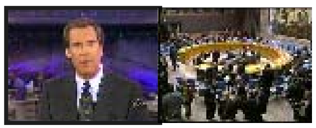
- Manual association feasible only if meanings are *few* and *known.*
- Metadata come to the rescue.



Associating Patterns with Text

videos

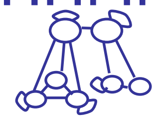
news



AV features and concepts

words (ASR/VOCR)

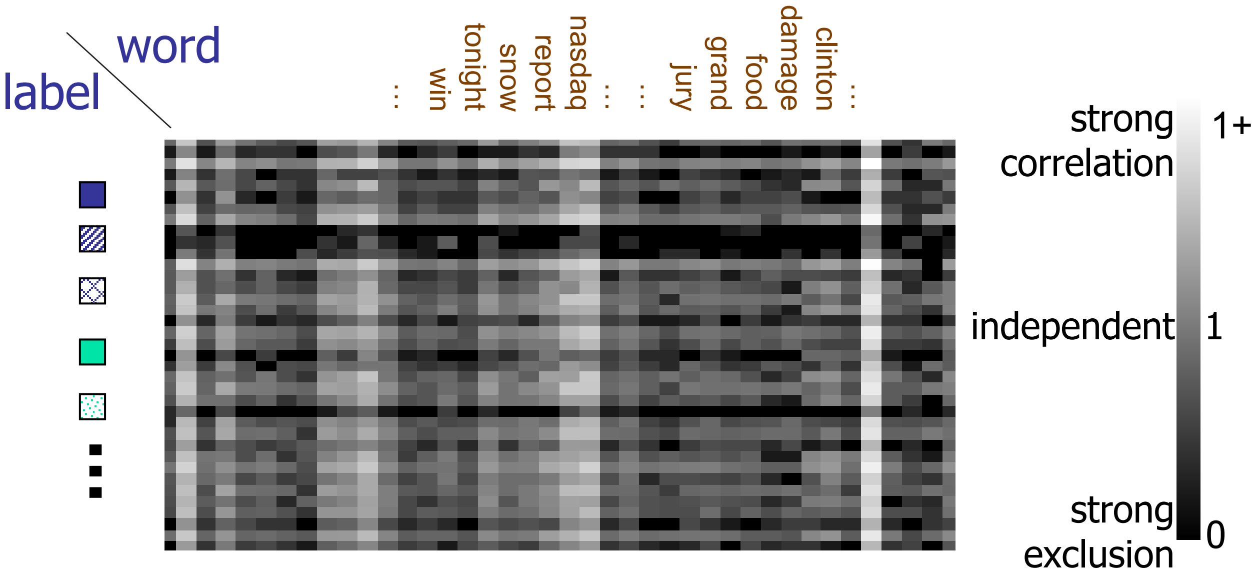
HHMM



patterns

label-word association

Co-Occurrence of HHMM Labels & Words



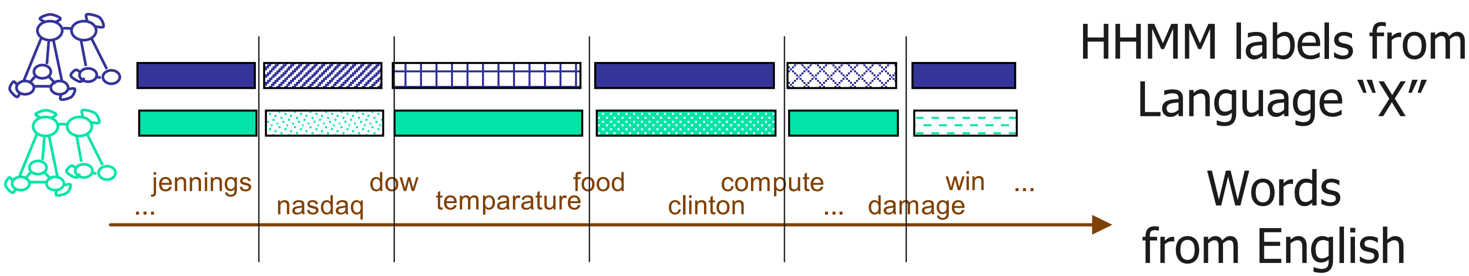
Conditional prob.

$$C(q \mid w) = C(q,w) / C(q,.)$$

$$C(w \mid q) = C(q,w) / C(.,w)$$

Likelihood ratio:

$$L^c(q,w) = \frac{C(q,w)}{C(q,\cdot)C(\cdot,w)}$$



"*correlation*" between HHMM labels and words
→ co-occurrence counts.

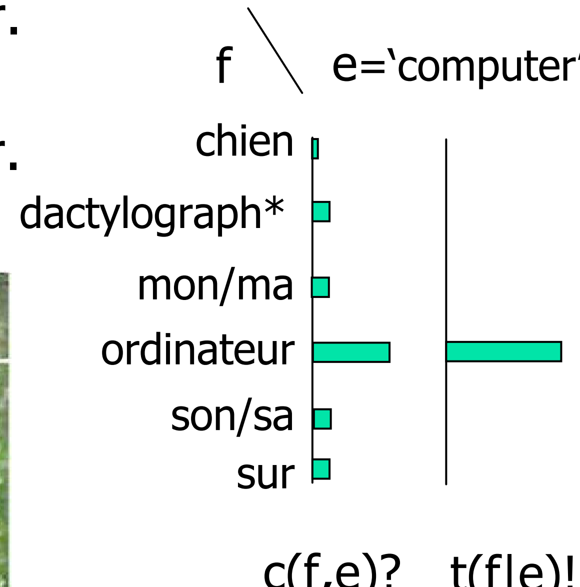
Refining the Co-occurrence Statistics

				"observed"	"true unsmoothed"
Story#	1	2	3		
News Video					
HHMM label	q_1	q_1 q_2	q_2	$\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$	$\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$
ASR token	w_1	w_2 w_1	w_2		

Machine Translation
[Brown'93]

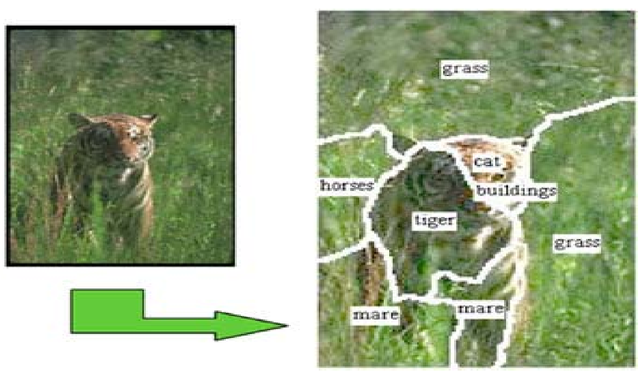
Her dog is typing on my computer.

Son chien dactylographie sur mon ordinateur.



[Dyugulu et. al. 2002]

image
$\sim \{b_1, \ldots, b_n\}$
$\sim \{w_1, \ldots, w_n\}$



Translation between AV Tokens & Words

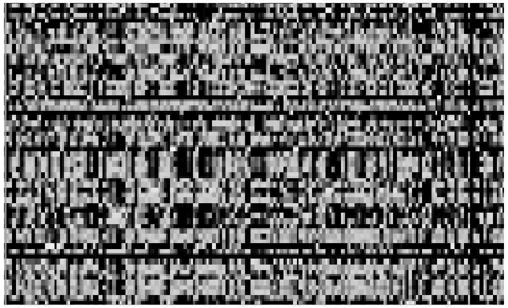
The problem:
Co-occurrence "un-smoothing".

know: $C(q,w)$;

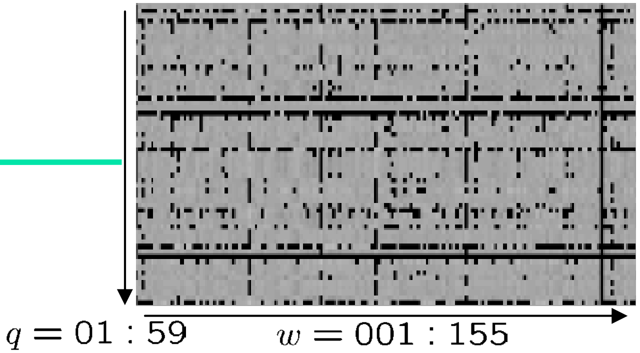
seek: $t(w|q)$, $t(q|w)$.

Solve with EM [Brown'93]

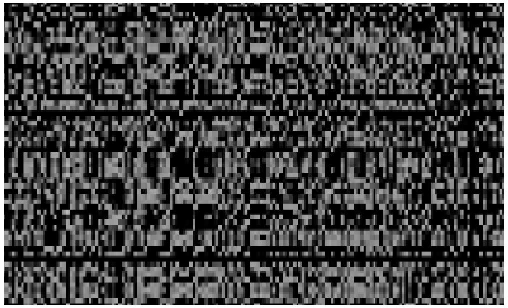
$L^t_q(q,w)$ (assume q's are ind.)



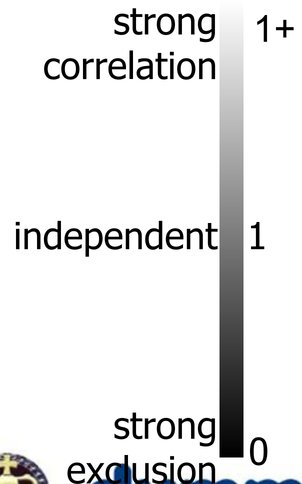
$$L^c(q,w) = \frac{C(q,w)}{C(q,\cdot)C(\cdot,w)}$$



$L^t_w(q,w)$ (assume w's are ind.)



L=



Experiments

- TRECVID2003 news
  - 44 half-hour videos, ABC/CNN
  - 12 visual concepts for each shot [IBM-TREC'03] (*weather, people, sports, non-studio, nature-vegetation, outdoors, news-subject-face, female speech, airplane, vehicle, building, road)*
  - ASR transcript
- HHMM on concept confidence scores
  - 10 models from hierarchical clustering in feature selection, size automatically determined
  - Co-occurrence with story boundaries

[Xie et al. ICIP'04]

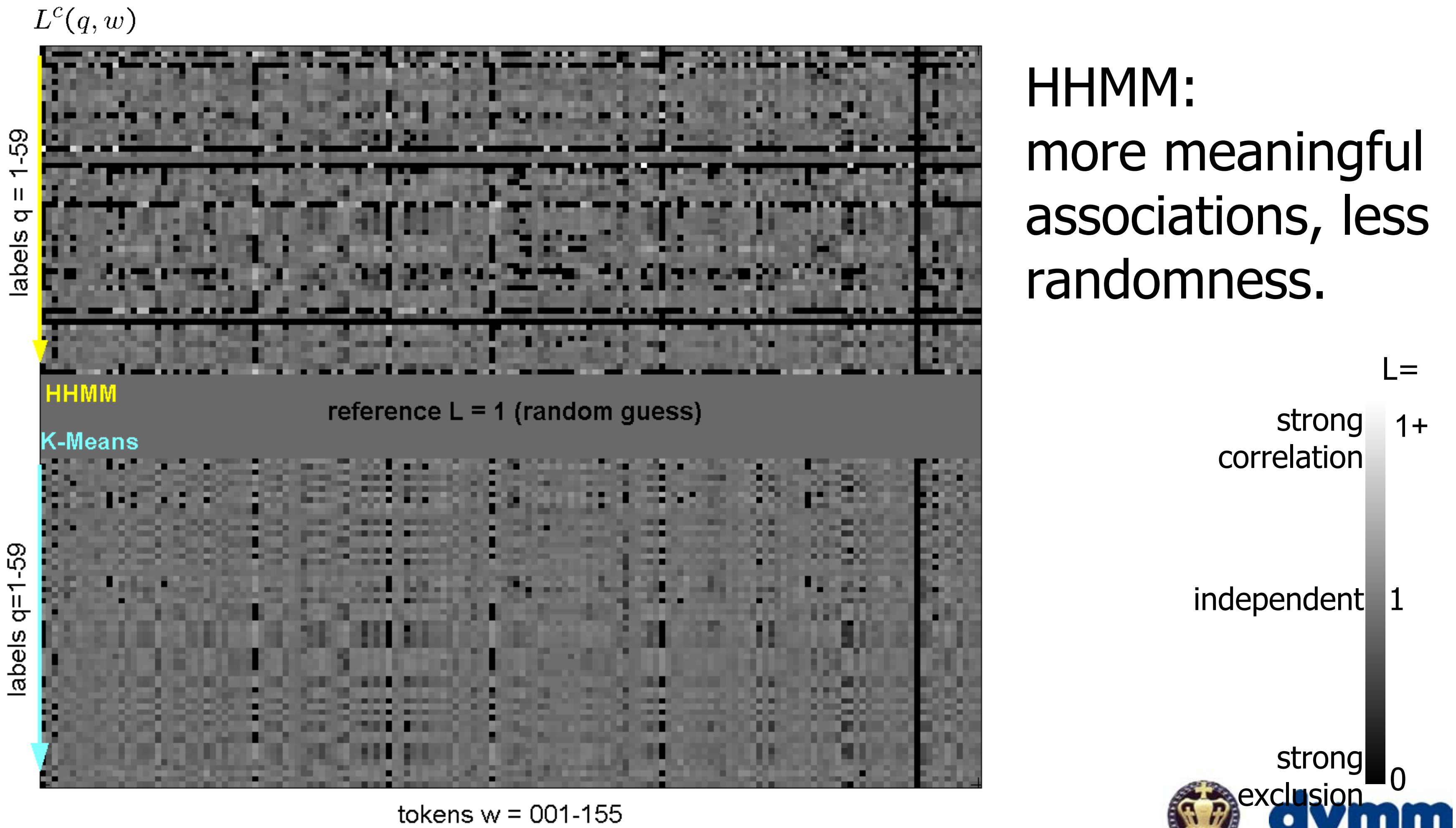
Example Correspondences

HHMM label	Visual Concept	Words	Topic groundtruth
(6,3)	people, non-studio-setting	storm, rain, forecast, flood, coast, el, nino, administer, water, cost, weather, protect, starr, north, plane, …	El-nino Storm '98 (recall 80%)
(9,1)	indoor, news-subject-face, building	murder, lewinski, congress, allege, jury, judge, clinton, preside, politics, saddam, lawyer, accuse, independent, monica, charge, …	Clinton-Jones (Recall 45%, Precision 15%) Iraqi-weapon (Recall 25%, Precision 15%)
(m, q): model # m state # q	Obtained with SVM classifiers [IBM'03]	Lexicon obtained by shallow parsing of keywords from speech recognition output.	

Automatic

Manual Inspection

Can't we find such patterns using conventional clustering? (HHMM vs. K-means comparison)

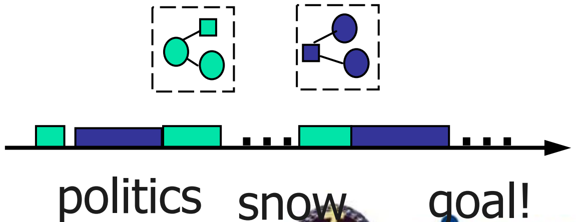


HHMM:
more meaningful associations, less randomness.

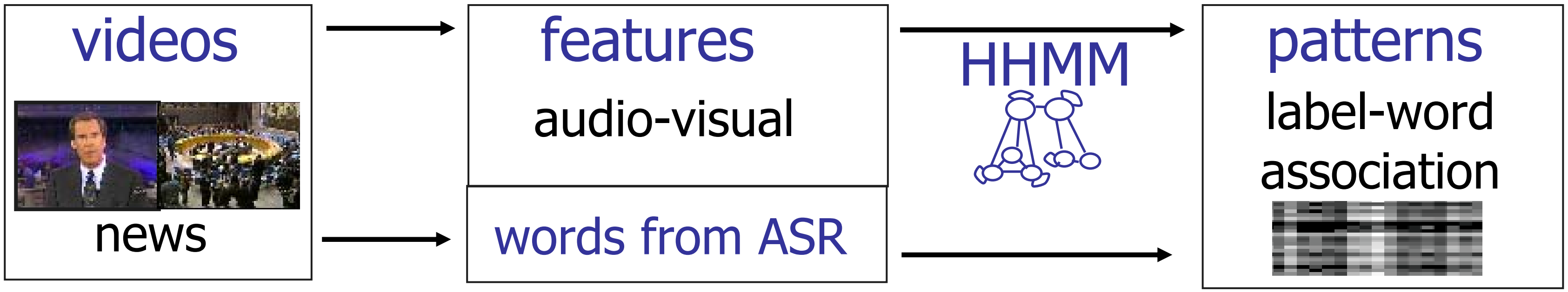
Outline

- The problem
- Unsupervised pattern discovery with HHMM
- Finding meaningful patterns
  - With text association
  - By multi-modal fusion
- Summary

- Versatile
- Multi-modal
- Meaningful
- Knowledge-adaptive

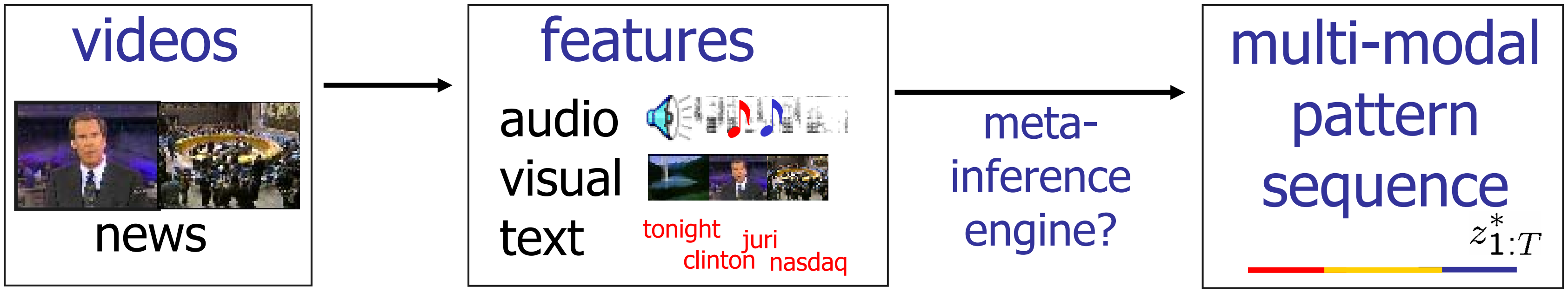


Is MT the right paradigm?



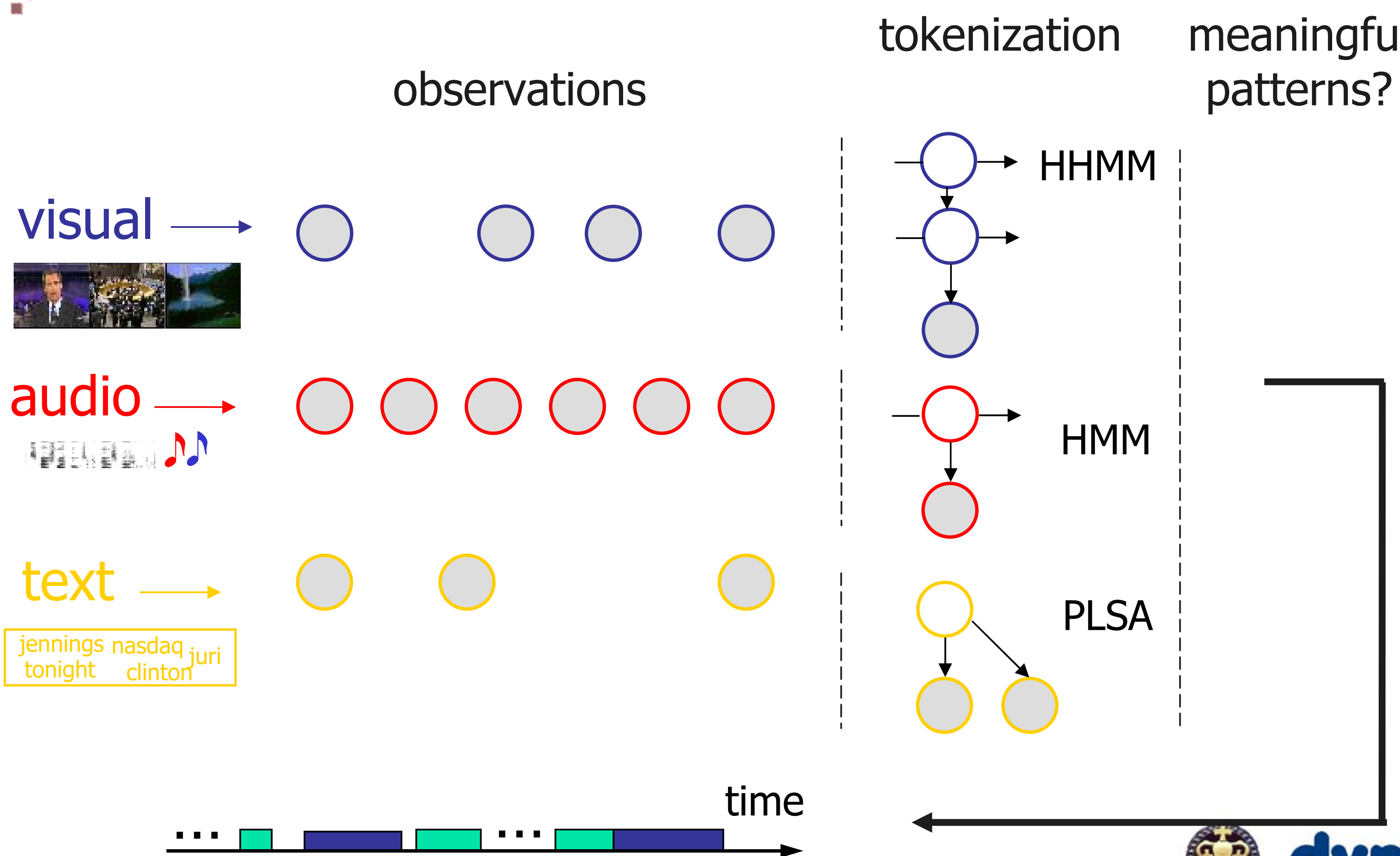
- Potential Problems:
  - Words $\neq$ meanings
- Temporal correspondence between words and pattern labels may not exist.

Change to Joint AVT Pattern Mining



- Instead, both the pattern labels and words jointly support some hidden semantic states.
- A multi-modal data fusion problem [AVSR, multi-modal interaction]
- Aiming at recovering the semantics [TDT 1998-2004]

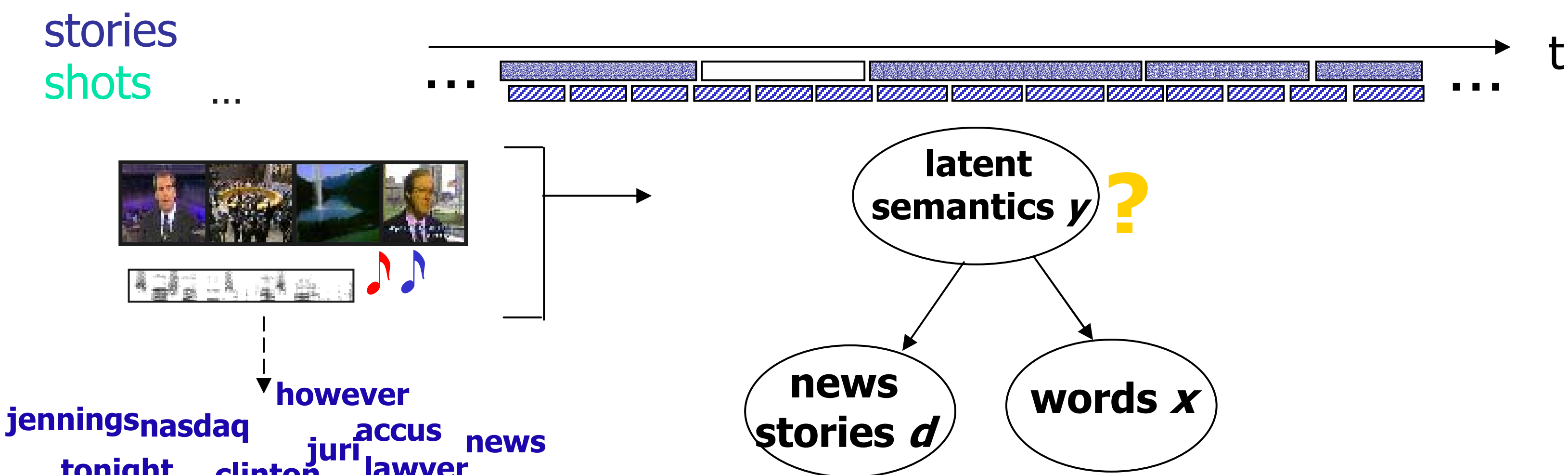
Multi-modal Fusion for Produced Videos

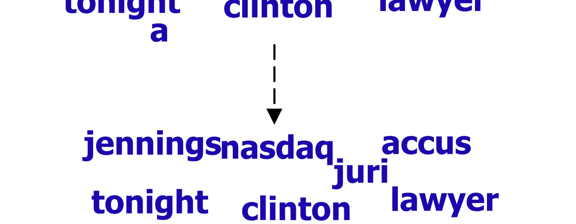


Columbia University in the City of New York
dvmm DIGITAL VIDEO MULTIMEDIA LAB

Extracting High-level Concepts from Text – pLSA

- **Use data-driven analysis to find concept association and latent semantic topics**
- **Use the syntactic story structures of video to define co-occurrence**
- **Use graphics model statistical inferencing to discover latent semantics**
  - **Not just counting occurrences**





$$p(y|x,d) \propto p(y)p(x|y)p(d|y)$$

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} p(y|x,d)$$

Some semantic clusters from text pLSA

"financial"
dow
nasdaq
industrial
average
wall
jones
gain
trade
… …

"olympics"
gold
olympics
… …

"iraq"
saddam
iraq
baghdad
weapon
hussein
strike
secure
… …

"investigation"
jury
lewinski
starr
grand
accusation
sexual
independent
water
monica
investigation
president
… …

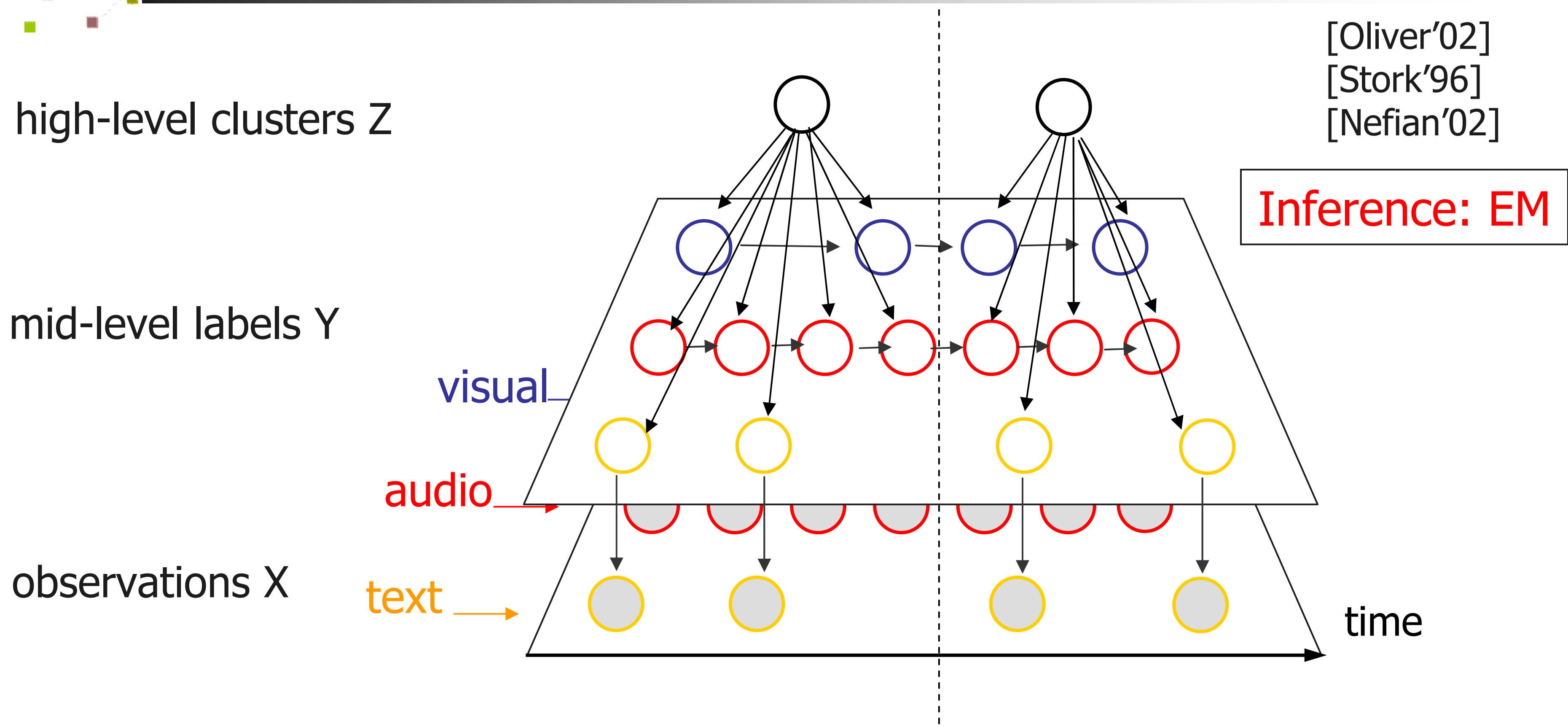
"weather"
temperature
rain
coast
snow
el
heavy
northern
storm
forecast
tornado
pressure
east
florida
nino
gulf
weather
… …

"bad" clusters
cancer
increase
secure
temperature
texas
accusation
chance
nasdaq
pressure
center
… …

cancer
africa
temperature
movie
coast
center
heavy
research
rain
strike
… …

Hierarchical Mixture Model

[Xie'TR04]



Example: z ~ "weather"; y ~ … ;

$x_{vconpcet}$ ~"graphics", x_{audio} ~ "speech", x_{text} ~"pressure, temperature …"

Experiments

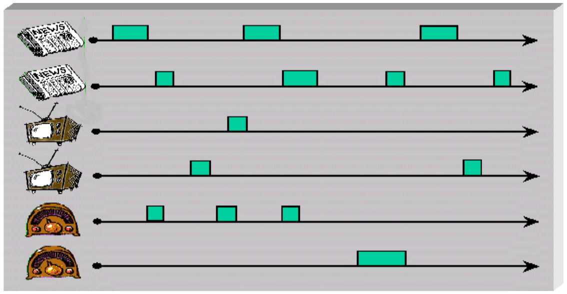
- News videos [TRECVID2003]
  - ABC, CNN : 30 min x 151 clips
  - Training/testing on same channels

modality	feature elements	granularity	bottom-level model
audio	pitch, pause, audio-class	.5 sec	HHMM
color	histogram (15-d)	1 sec	
motion	camera translation (2-d)	1 sec	
visual	22 concepts	every shot	
text	ASR word-stems tf-idf	every story	PLSA

Fusion Results Evaluated by Text TDT Topics & Metrics

- http://www.nist.gov/speech/tests/tdt/

•"NIST TDT research develops algorithms for discovering and threading together topically related material in streams of data such as newswire and broadcast news in both English and Mandarin Chinese. "



• Current TDT use text or ASR only.

TDT Tasks

1. **Story Segmentation - Detect changes between topically cohesive sections**
2. **Topic Tracking - Keep track of stories similar to a set of example stories**
3. **Topic Detection - Build clusters of stories that discuss the same topic**
4. **First Story Detection - Detect if a story is the first story of a new, unknown topic**
5. **Link Detection - Detect whether or not two stories are topically linked**

Hierarchical Mixture Model

[Xie'TR04]

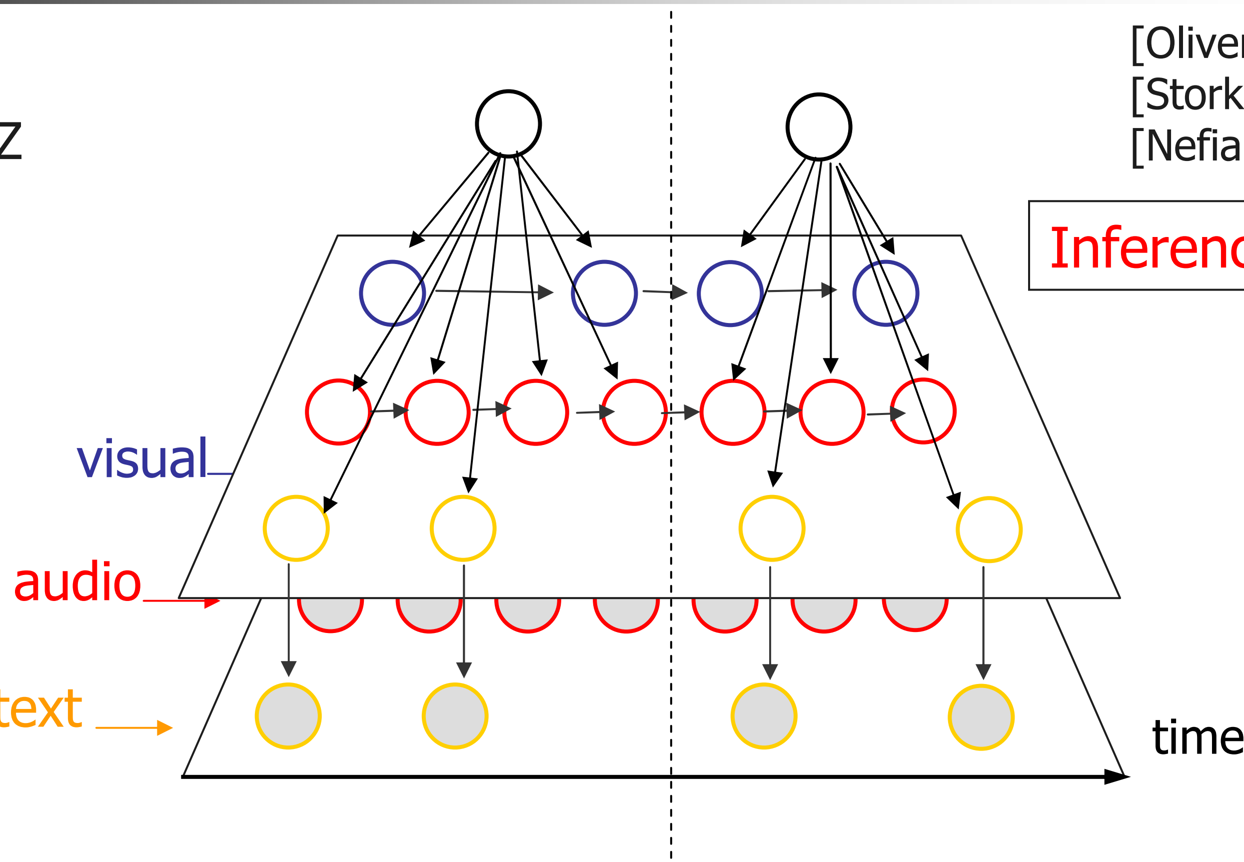
[Oliver'02]
[Stork'96]
[Nefian'02]

Inference: EM

high-level clusters Z

mid-level labels Y

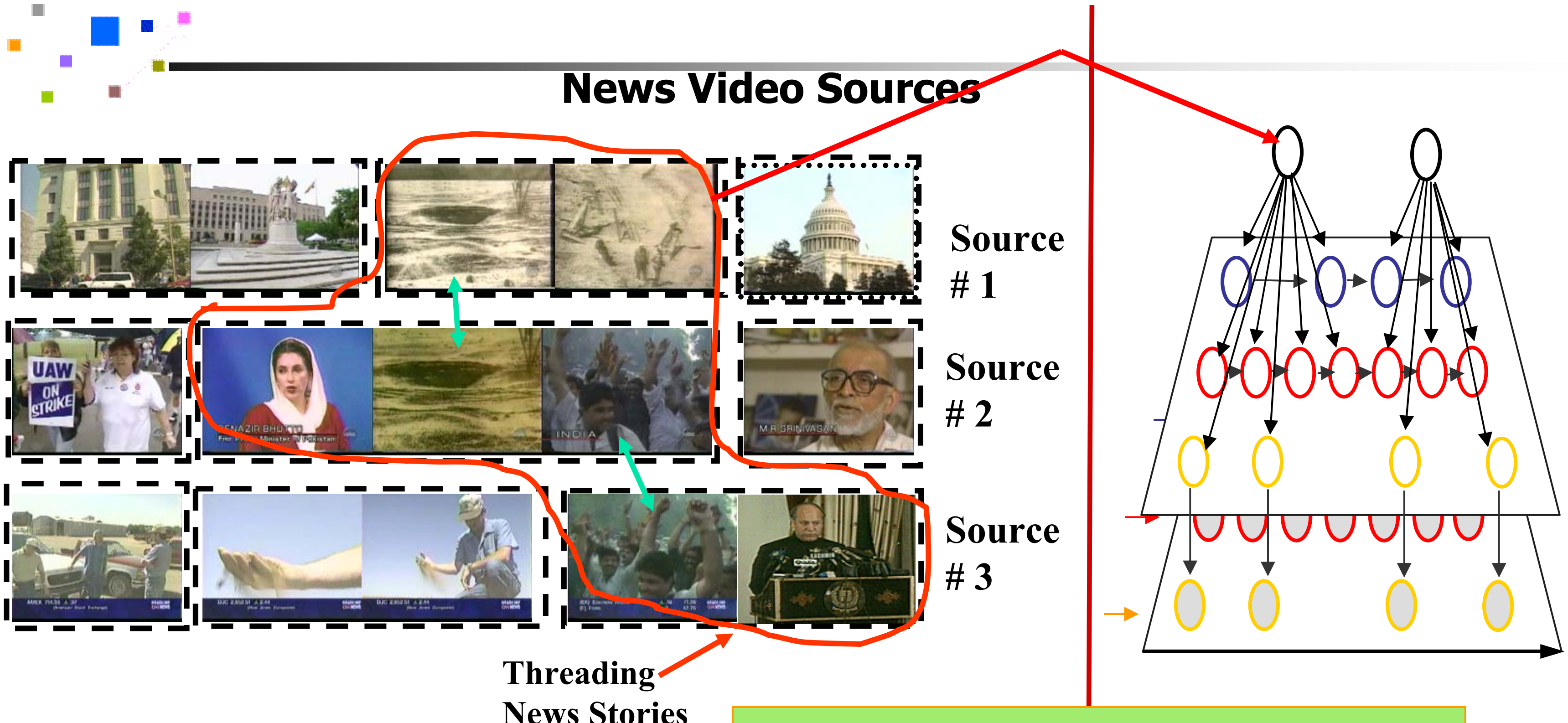
observations X



Example: z ~ "weather"; y ~ … ;

$x_{vconpcet}$ ~"graphics", x_{audio} ~ "speech", x_{text} ~"pressure, temperature …"

Linking Multi-Source Videos, Web Pages



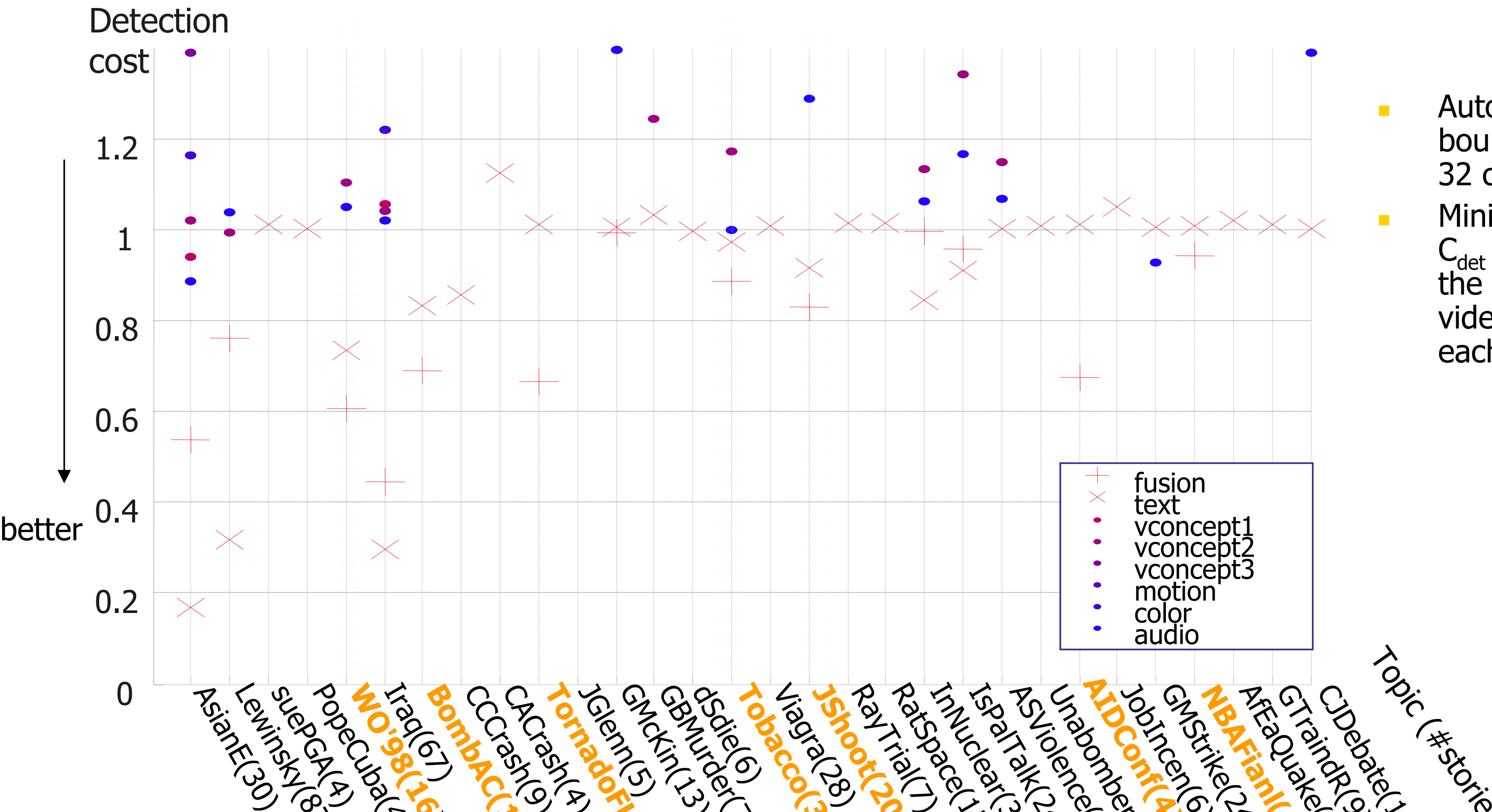
Question: does the discovered thread correspond to distinct semantics?

- **no ground truth**
- **tentatively use TDT topics**

Columbia University in the City of New York

dvmm DIGITAL VIDEO MULTIMEDIA LAB

Evaluate the MM patterns against TDT topics

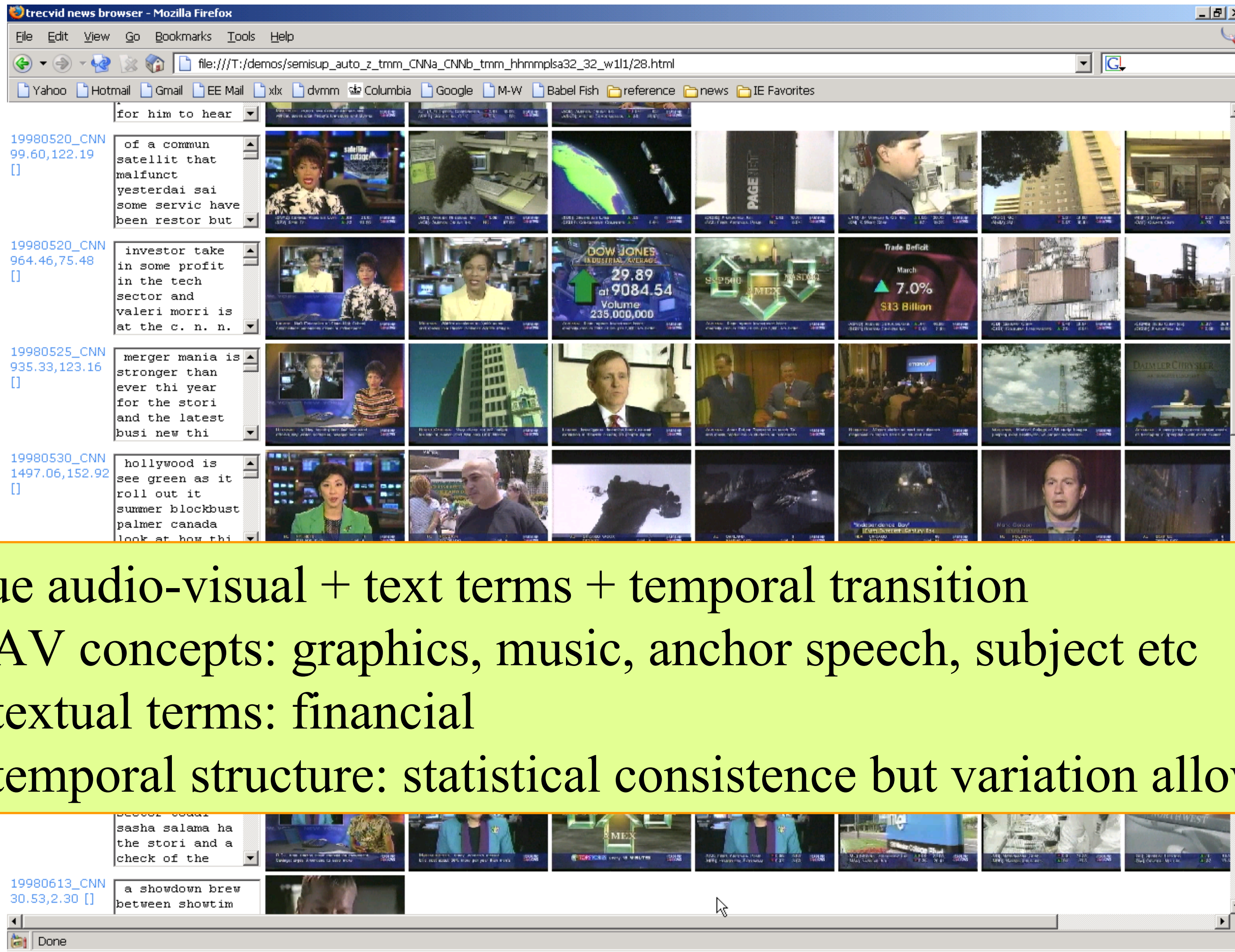


- Auto-story boundary, 32 clusters
- Minimum C_{det} among the CNN videos for each topic

- **unsupervised multi-modal token fusion able to track certain topics more accurately**
- **Such topics tend to be rich in non-textual cues**

Example: Cluster #28

→ Correspond to "Dollars & Sense, CNN" (demo)

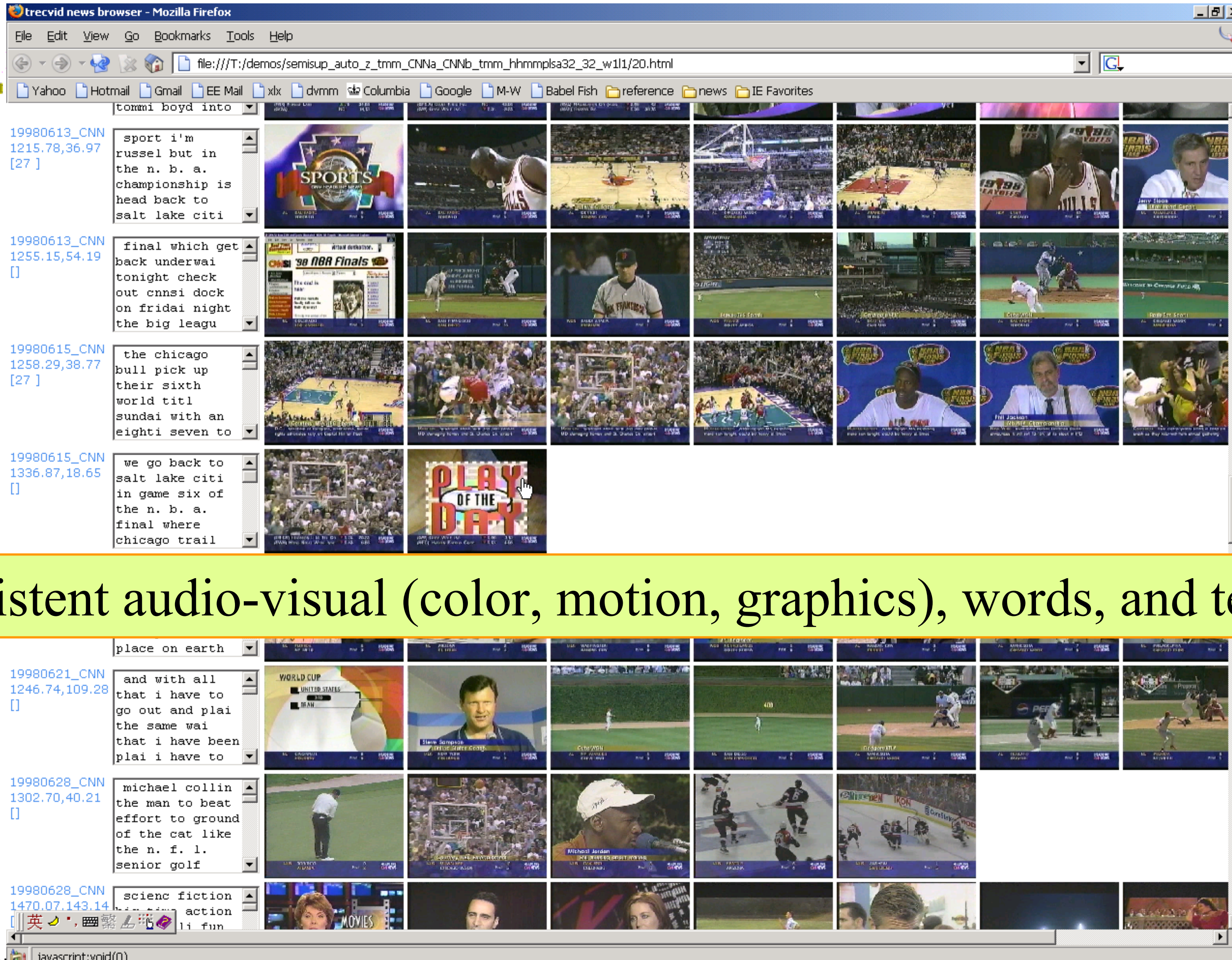


Unique audio-visual + text terms + temporal transition

- AV concepts: graphics, music, anchor speech, subject etc
- textual terms: financial
- temporal structure: statistical consistence but variation allowed

Example: Sports

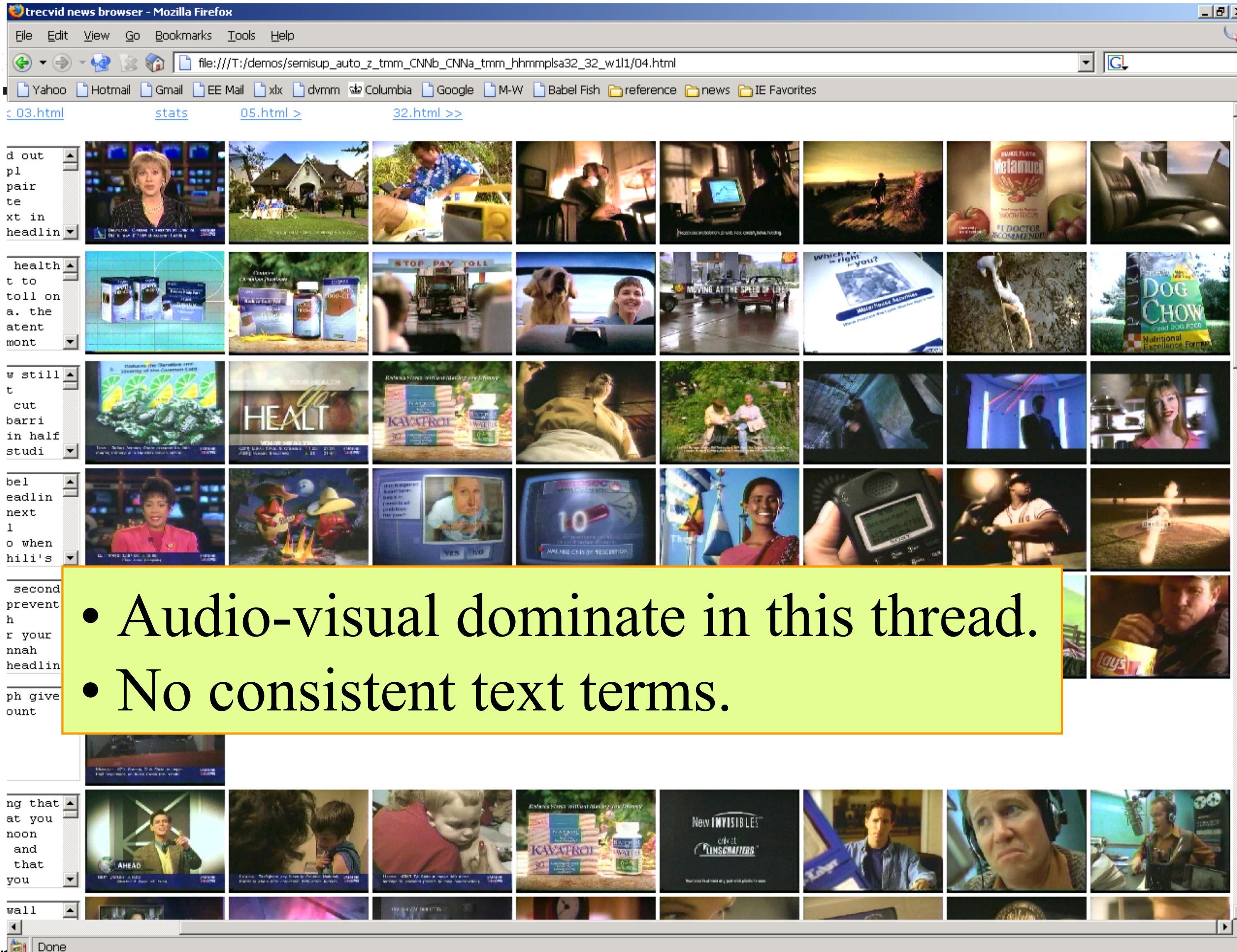
(demo)



Consistent audio-visual (color, motion, graphics), words, and temporal

Example: Advertisements

(demo)



- Audio-visual dominate in this thread.
- No consistent text terms.

Columbia University in the City of New York

DIGITAL VIDEO MULTIMEDIA LAB

Conclusions

Theme: Media Pattern Mining for Semantics Discovery

- Patterns abound in multimedia data
- Mining facilitates auto discovery of salient or novel concepts → scalability
- Current results shown in
  - Multi-level temporal patterns mining through HHMM
  - Fusion between pattern labels and ASR metadata
- Challenging Issues
  - Mining of patterns of different types or complex patterns at higher levels
  - Detection of alerts and novel events
  - Evaluation → Redefine TDT for Multimedia TDT?
  - Visualization

Acknowledgements

- DVMM Group
  http://www.ee.columbia.edu/dvmm
  W. Hsu, H. Sundaram, D-Q Zhang
- IBM TRECVID Team:
  J. R. Smith, C.-Y. Lin, M. Naphade, A. Natsev, B. Tseng, G. Iyengar, H. Nock, et al.

References

[1] L. Xie, S.-F. Chang, A. Divakaran, and H. Sun, "Unsupervised discovery of multilevel statistical video structures using hierarchical hidden Markov models," in *Interational Conference on Multimedia and Expo (ICME)*, (Baltimore, MD), July 2003.

[2] L. Xie, L. Kennedy, S.-F. Chang, C.-Y. Lin, A. Divakaran, and H. Sun, "Discover meaningful multimedia patterns with audio-visual concepts and associated text," in *Interational Conference on Image Processing (ICIP)*, October 2004.

[3] L. Xie, S.-F. Chang, A. Divakaran, and H. Sun, *Unsupervised Mining of Statistical Temporal Structures in Video*, ch. 10. Kluwer Academic Publishers, 2003.

[4] L. Xie, P. Xu, S.-F. Chang, A. Divakaran, and H. Sun, "Structure analysis of soccer video with domain knowledge and hidden markov models," *Pattern Recogn. Lett.*, vol. 25, no. 7, pp. 767–775, 2004.

References

[1] B. Clarkson and A. Pentland, "Unsupervised clustering of ambulatory audio and video," in *International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, 1999.

[2] M. Naphade and T. Huang, "Discovering recurrent events in video using unsupervised methods," in *Proc. Intl. Conf. Image Processing*, (Rochester, NY), 2002.

[3] C. E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton, "Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment," *Science*, vol. 8, no. 262, pp. 208–14, October 1993.

[4] S. Fine, Y. Singer, and N. Tishby, "The hierarchical hidden Markov model: Analysis and applications," *Machine Learning*, vol. 32, no. 1, pp. 41–62, 1998.

[5] A. Iyengar, M. S. Squillante, and L. Zhang, "Analysis and characterization of large-scale web server access patterns and performance," *World Wide Web*, vol. 2, no. 1-2, pp. 85–100, 1999.

[6] P. Duygulu, K. Barnard, N. de Freitas, and D. A. Forsyth, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," in *ECCV*, 2002.

[7] P. Duygulu and H. Wactlar, "Associating video frames with text," in *Multimedia Information Retrieval Workshop, in conjuction with SIGIR 2003*, (Toronto, Canada), August 2003.

[8] N. Oliver, E. Horvitz, and A. Garg, "Layered representations for learning and inferring office activity from multiple sensory channels," in *Proceedings of Int. Conf. on Multimodal Interfaces (ICMI'02)*, (Pittsburgh, PA), October 2002.