Aug. 2005

Multi-Modal Video Search and Pattern Mining

Shih-Fu Chang

Digital Video and Multimedia Lab

Columbia University

Opportunities for video researcher

- A tipping point
  - Prevalence of video content reaching critical point
  - Video as the first class data type
  - Ease and value in using video
- Example applications
  - Consumer media server, DVR, MOD
  - Mobile video, Pod-cast
  - Blog to Video Blog
  - Surveillance, Personal Life-Log (major funding)
  - Video search engines (Goggle, Yahoo)

Usage models

- Search – video Google
  - Search by title
  - Names
  - Content inside video clips (film, news, consumer, surveillance)
- Browsing
  - By topic, genre
- Hyperlink
  - Link video objects to external sites

What do we search in video? (examples)

- TRECVID 2003 **Event**
  - "Find shots of an airplane taking off."

- TRECVID 2004 **Named-person Location**
  - "Find shots of Bill Clinton speaking with a US flag visible behind him."

- IBM Speech Group **Objects**
  - "Find shots containing monkeys or gorillas."

- BBC Logs **Named-location**
  - "Find shots of the Kremlin."

(Images from TRECVID dataset)

Goggle News Threading

- Automatically crawl and track news video topics

- Still mainly relied on text analysis
  - Good opportunity for integrating video analysis

Search and Topic Tracking across Multiple News Channels and Web Pages

A government sponsored IBM-Columbia Joint Project

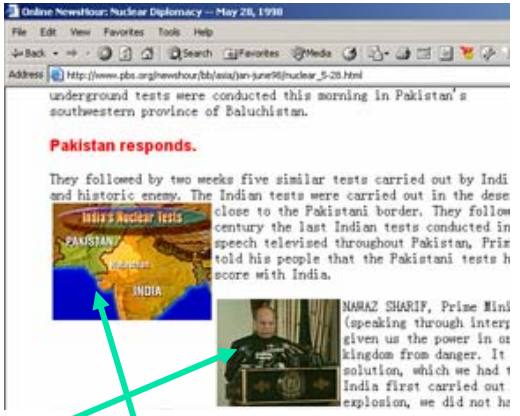
Source # 1

Source # 2

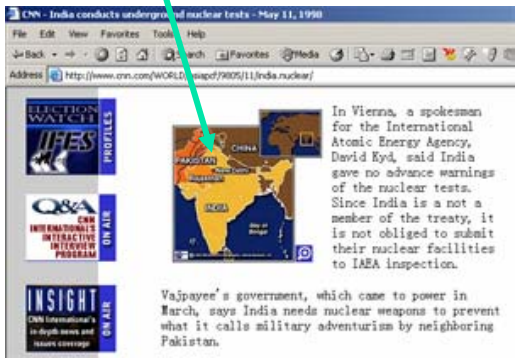
Source # 3

Threading News Stories

Broadcast sources



News Web Site 1



News Web Site 2

A multi-lingual topic example

- Same topic thread across English, Chinese, and Arabic channels

CNN

MSNBC

CCTV

- Imagine an information analyst tracking > 100 channels of broadcast news around the world
- Efficient topic tracking and search tools are important

Video Search/Tracking Calls for Multi-modal search

Query

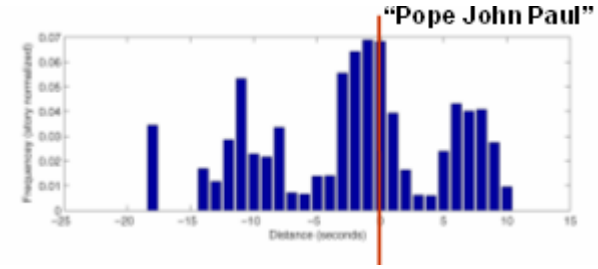
Find
shots of
Pope John
Paul second

Story

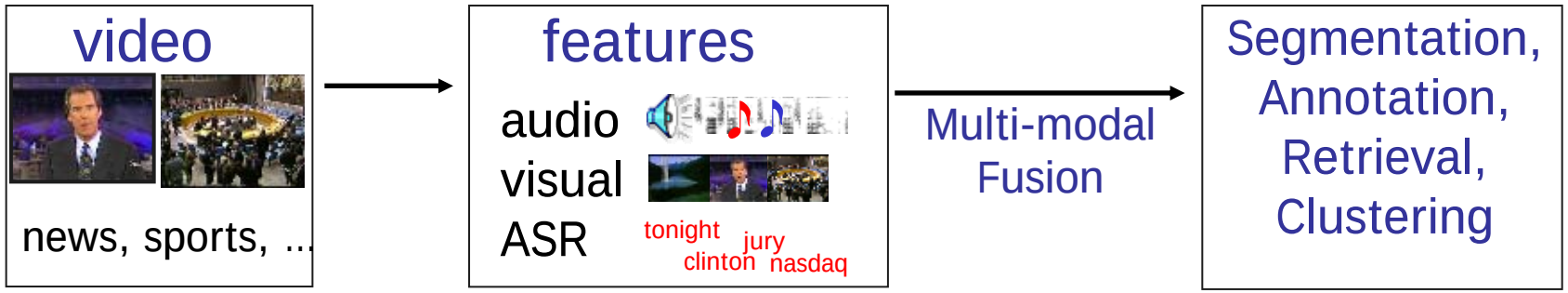
Shot **Shot** **Shot** **Shot** **Shot**

in other news **pope john paul** the **second** will get his first look at the shroud of turin today that's the piece of linen many believe was the burial cloth of jesus the round is on public display for the first time in twenty years it has already drawn up million visitors the pope's visit to northwest italy has also included beatification services for three people the vatican says **john paul** is now the longest serving pope this century he has surpassed **pope** pious the twelfth who served for nineteen years seven months and seven days

- Findings on General Retrieval:
  - Text is effective for recall, but non-text features are essential for precision
- Retrieval on Person-X:
  - A query on person-X, find shots that person-X appears visually
  - Important features:
    named entities in text, face recognizer, and their correlation distributions



A Fruitful Area for Multi-Modal Fusion



- Results of 'Person-X' search suggests
  - Development of various components tools (visual, audio, text)
  - Query-Dependent Model (QDM) for fusing multi-modal features
    → different query strategies for different queries!

Related Activities NIST TRECVID

- Low-level feature detection (motion, shot etc)
- High-level feature detection
  - Image →classifier→ {‘people’, ‘vehicle’, ‘explosion’, etc}
- Story boundary detection
- Search : fully automatic, manual, interactive
- 2005 Data
  - 6 channels in English, Chinese, Arabic
  - >170 hours, 126,000 subshots
  - 39 concepts manually annotated over >80 hours (LSCOM-Lite) : very valuable resource for researchers!
- Participating groups
  - 62 groups (due: high-level feature 8/22, search 9/21/05)

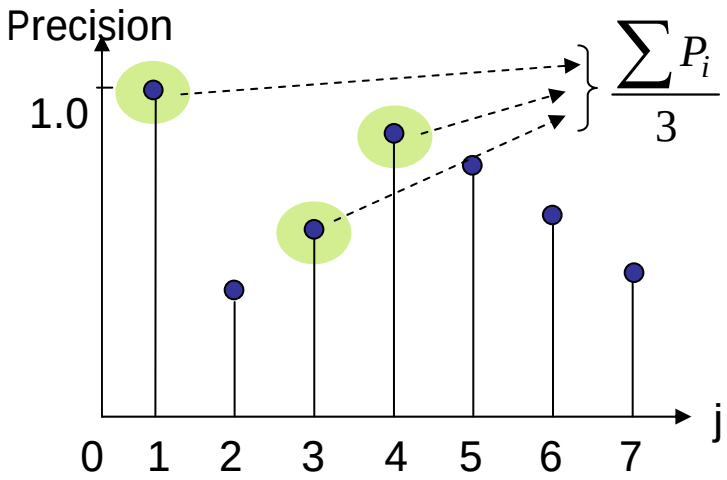
Evaluation Metric: Average Precision

DB → Ranked list of data in response to a query

	D_{15}	D_8	D_{63}	D_{21}	...	...	D_s
Ground truth	*1*	*0*	*1*	*1*	*0*	*0*	*0*
Precision	*1/1*	*1/2*	*2/3*	*3/4*	*3/5*	*3/6*	*3/7*

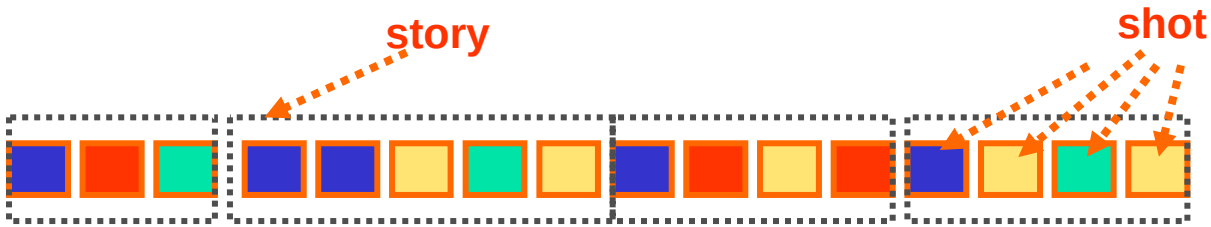
Average precision: $$AP = \frac{1}{R}\sum_{j=1}^{s}\frac{R_j}{j}I_j, \qquad R_j : \#\text{ relevant data at j}$$

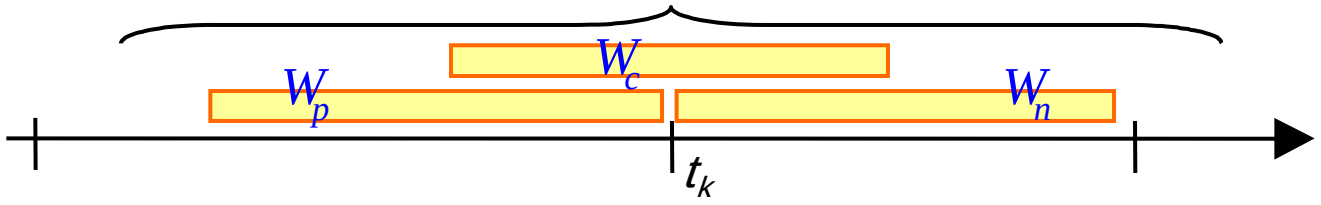
AP measures the average of precision values at R relevant data points



- A Quick Review of Some Building Components for a Video Search System

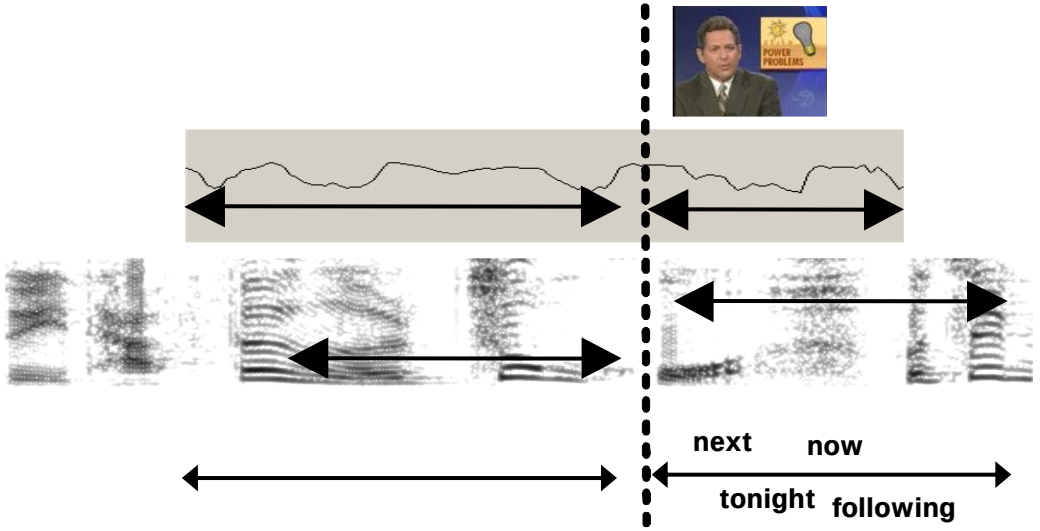
News Story Segmentation





Detect story boundary from multi-modal features

Given observation x_k, estimate probability p(story bnd = YES | x_k)



anchor face?

visual motion?
video caption text?

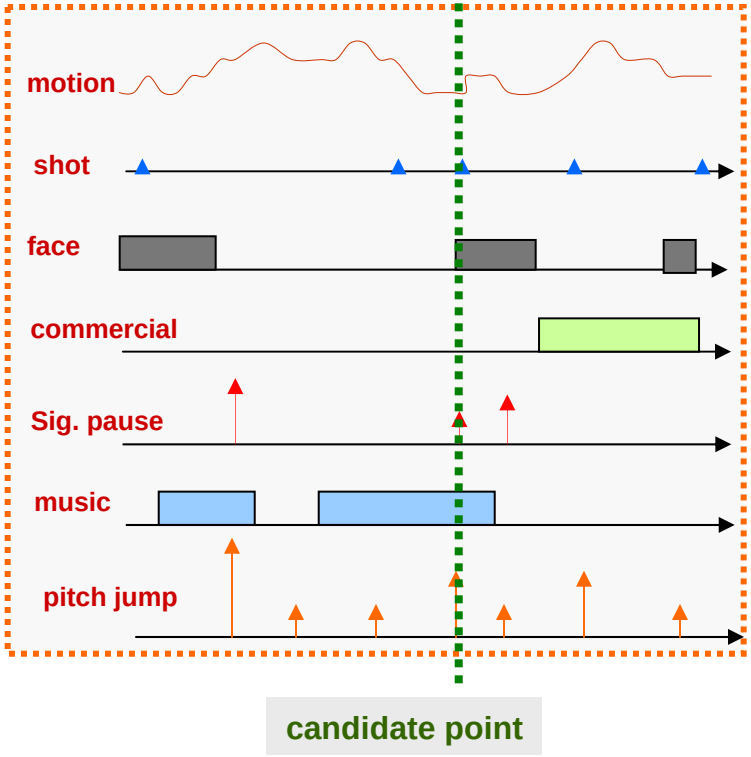
music or speech?
new speech segment?
significant pause?
pitch change?

cue phrases appear?

Anchor face alone: 65%, ASR alone 62%, MM fusion 75% in TRECVID 2003

Issue: diverse feature types and high dimensionality

Modality	Raw Features	Data Type	Value
Video	motion	Point/seg	continuous
	shot boundary	point	binary
	face	segment	continuous
	commercial	segment	binary
Speech /Audio	pause	point	continuous
	pitch jump	point	continuous
	significant pause	point	continuous
	musc./spch. disc.	segment	binary
	spch seg./rapidity	segment	continuous
Text	ASR cue terms	point	binary
	V-OCR cue terms	point	binary
	text seg. score	point	continuous
Misc.	combinatorial	point	binary
	sports	segment	binary



Feature wrappers:

Hsu & Chang
ICASSP '04

ME Model for Feature Fusion & Selection

(195 binary predicates)

Each row represents one predicate f_i

- Anchor after t
- Significant Pause in non-commercial
- Commercial ends/starts
- Speech segment ends after t
- ASR cue term before/after

b	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0
1	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0
2	0	1	1	0	1	0	0	0	0	0	0	1	0	0	0	1	0	0	0
3	0	0	0	0	0	1	0	1	1	0	1	0	1	0	1	1	0	1	0
4	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0	0
5	0	0	1	0	1	0	1	0	0	0	0	0	1	1	0	1	0	0	0
6	0	0	1	0	0	0	1	0	1	0	1	0	1	0	0	1	0	0	0
7	0	0	1	0	0	0	1	0	1	0	0	0	1	0	0	1	0	0	0
8	0	0	1	0	0	0	0	0	0	1	0	1	0	0	0	1	0	0	0
9	0	0	1	0	1	0	1	0	1	0	0	0	1	0	1	0	1	0	0
...	0	0	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	0	0

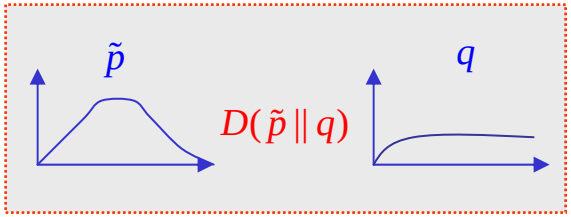
One training case

Maximum entropy model

$$q_\lambda(b \mid x) = \frac{1}{Z_\lambda(x)} e^{\sum_i \lambda_i \cdot f_i(x,b)}$$

$$where \quad f_i(x,b), b \in \{0,1\}$$

- Efficient learning methods
  - match the learned distribution with the empirical distribution
  - estimate the optimal weights
  - select salient feature subset



Discovered features based on Max. Entropy model

*** The first 10 "A+V" features automatically discovered for the CNN channel**

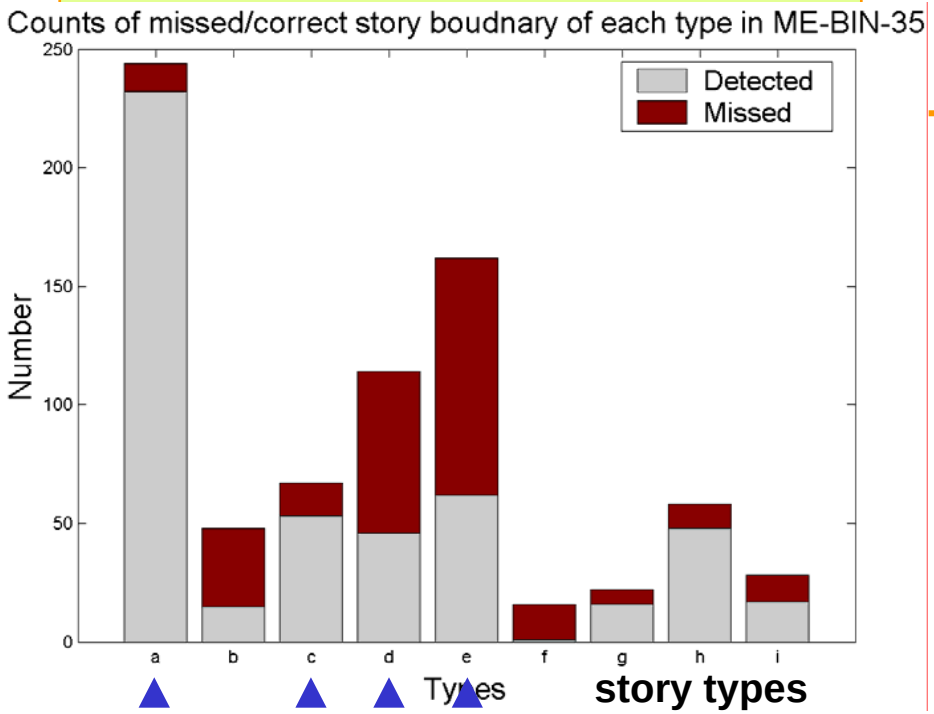
no	raw feature set	gain	λ	interpretation
1	Anchor Face	0.3879	0.4771	An anchor face segment just starts after the candidate point
2	Significant pause & non-commercial	0.0160	0.7471	A significant pause within the non-commercial section appears in the surrounding observation window.
3	Pause	0.0058	0.2434	An audio pause with the duration larger than 2.0 second appears after the boundary point.
4	Significant pause	0.0024	0.7947	The surrounding observation window has a significant pause with the pitch jump intensity larger than the normalized pitch threshold 1.0 and the pause duration larger than 0.5 second.
5	Speech segment	0.0019	-0.3566	A speech segment before the candidate point
6	Speech segment	0.0015	0.3734	A speech segment starts in the surrounding observation window
7	Commercial	0.0015	1.0782	A commercial starts in 15 to 20 seconds after the candidate point.
8	Speech segment	0.0022	-0.4127	A speech segment ends after the candidate point
9	Anchor face	0.0016	0.7251	An anchor face segment occupies at least 10% of next window
10	Pause	0.0008	0.0939	The surrounding observation window has a pause with the duration larger than 0.25 second.

every modality helps : especially anchor face, prosody, and speech segment

Success and failure (TRECVID 2003)

- every modality helps
- important features – anchor face, prosody, speech segment, commercial

- failure cases may need deeper text analysis



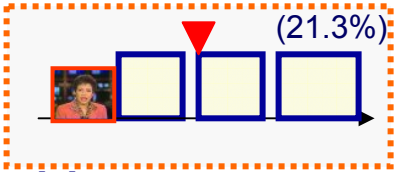
(a): led by an anchor segment

(c): multi-story in an anchor seg.

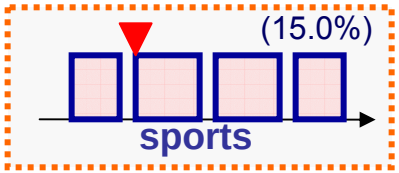
Success Case:

due to anchor & prosody

(demo: sig. pause)



(e): fast short briefings



(d): sports briefings

Failure Case:

- **No significant A-V cues**

(demo: miss)

Basic Search Tools: Find Visually Similar Shots

- Content-based image searches: similarity between query images and search images measured in various spaces.
  - Color
  - Texture
  - Edge

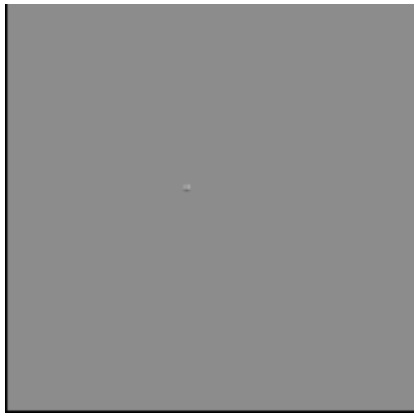
Query Image:

Similar Images:

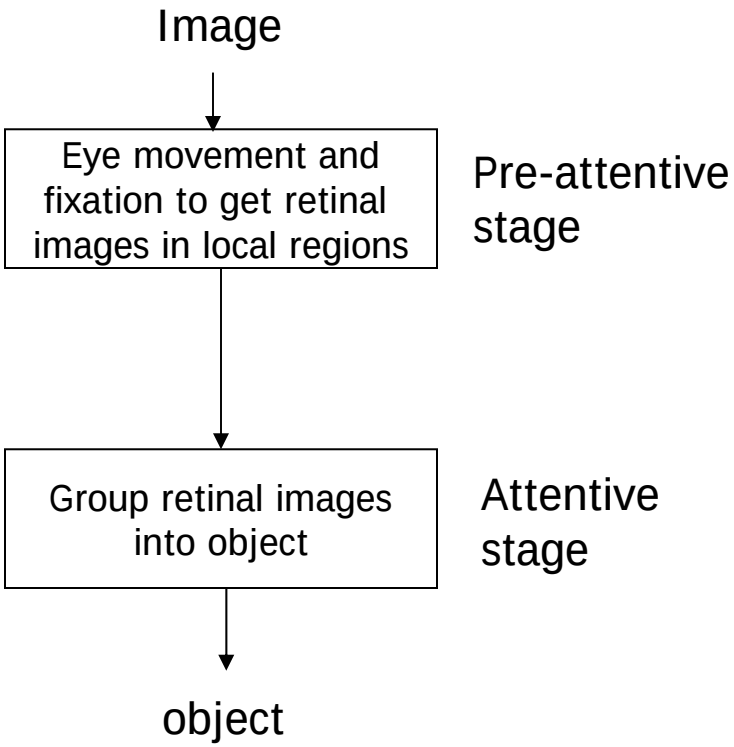
Vision-based Image Understanding: Part-based Object Model

Zhang & Chang, 04

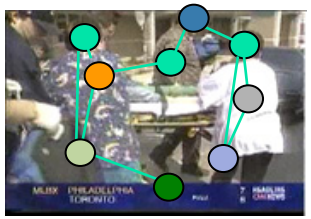
- Human Vision System (HVS) employs Part-based Model in object detection



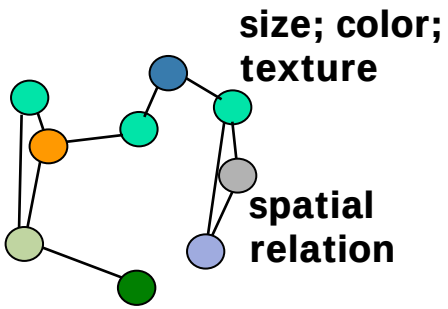
[Rybak et al. 98']



Random Attributed Relational Graph

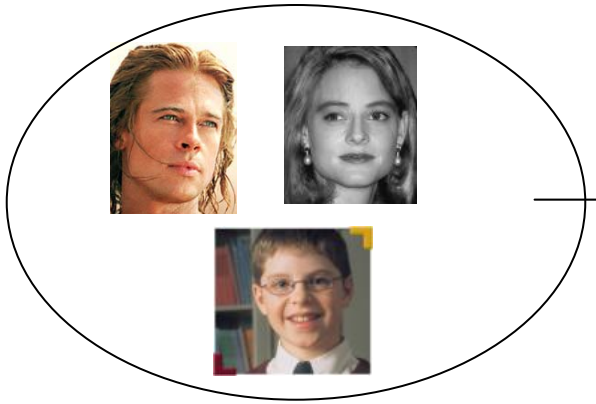


Instance of Image
→ parts, high entropy regions



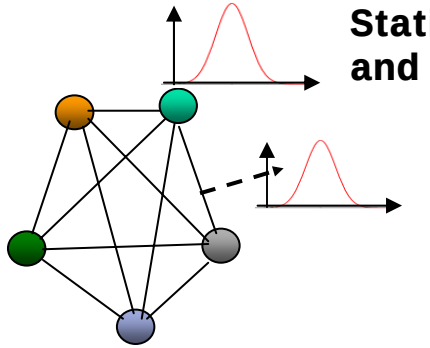
Attributed Relational Graph (ARG)

Graph Representation of Image



collection of training images

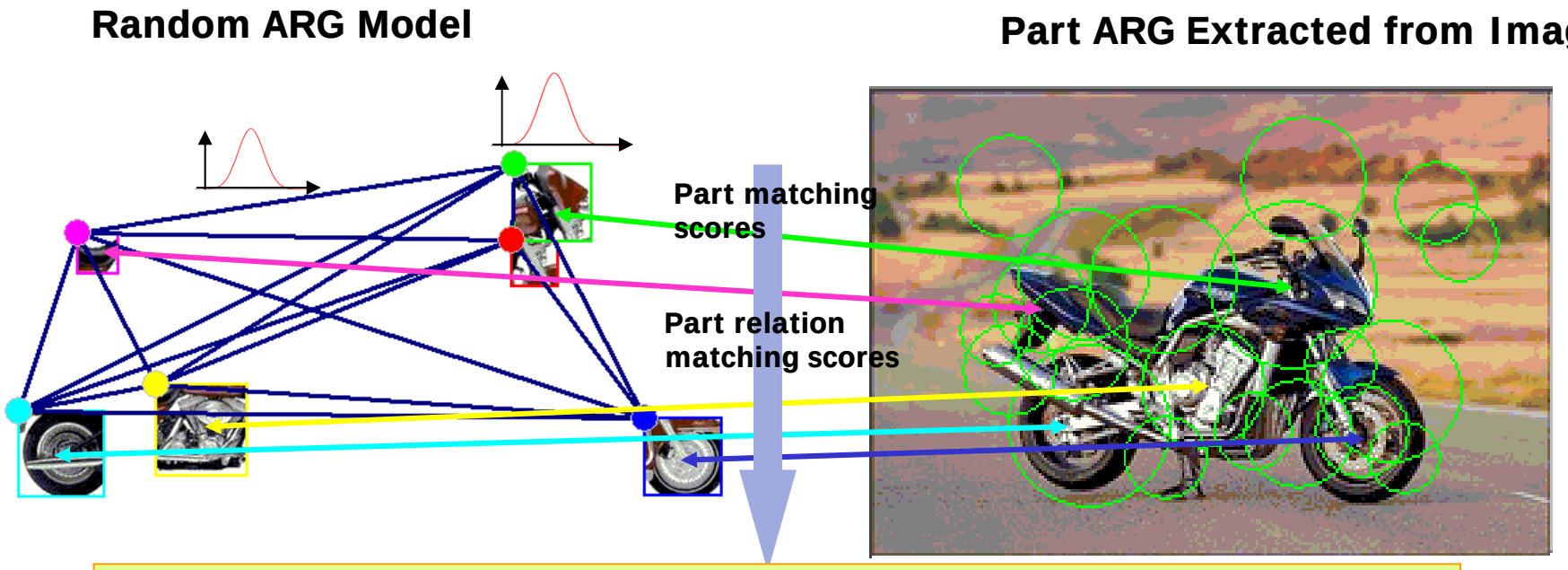
machine learning



Random Attributed Relational Graph (R-ARG)

Statistical Graph Representation of Model

Object Detection: Image-Model Matching



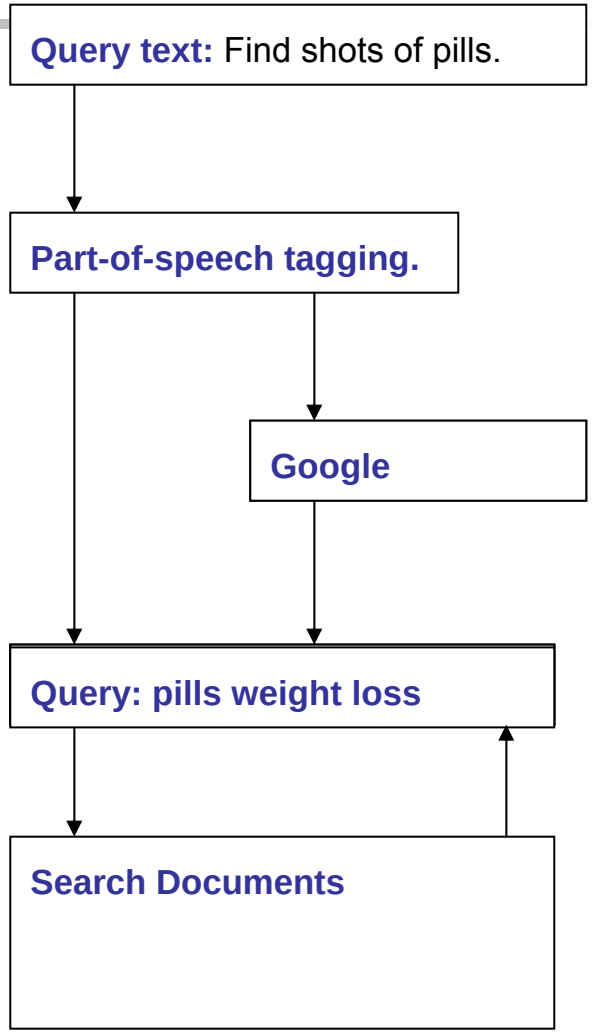
Object Detection: Compute Likelihood Ratio of Model vs. Background

- **Challenge : Finding the correspondence of parts and computing matching probability are NP-complete**
- **Our Solution :**
  - **Apply and develop advanced machine learning techniques – Loopy Belief Propagation (LBP), and Gibbs Sampling plus Belief Optimization (GS+BO)**
  - **Unique feature: compute the probability of each part-node correspondence, instead of overall constellation matching**

demo

Component Search Tools: Text Query Expansion

- Basic text: use key terms from query text.
- Pseudo-relevance feedback: conduct basic search and feedback frequent terms from top relevant documents
- Query expansion: send basic search to WordNet to find synonyms and hypernyms.
- Query expansion: send basic search to Google to find frequent terms in top relevant documents.

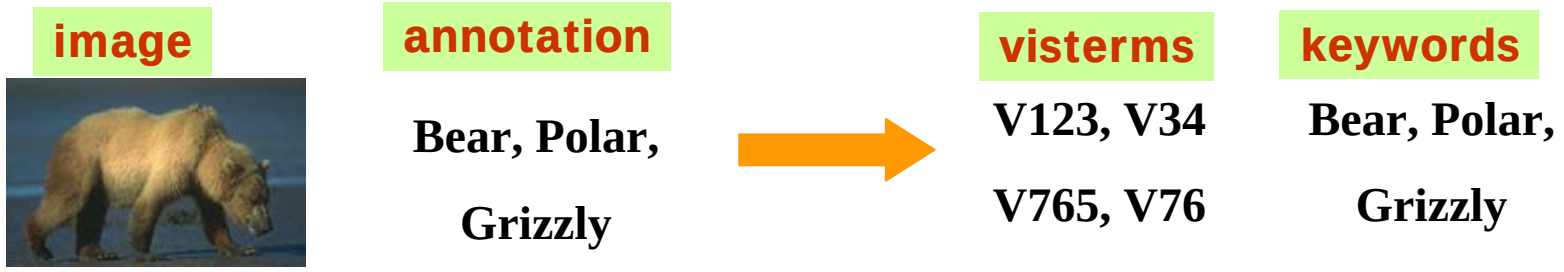


Query Expansion Using Different Sources

Original Query	Simple OKAPI	PRFB	WordNet	Google
Find shots of Osama bin Laden.	osama bin laden	osama bin laden afgahnistan taliban	osama bin laden container bank	osama bin laden usama fbi wanted
Find shots of pills.	pills	pills viagra pfizer	pills lozenge tab dose	pills prescription drug
Find shots with a locomotive (and attached railroad cars if any) approaching the viewer.	locomotive railroad car viewer	locomotive railroad car viewer germany crash wreckage	locomotive railroad car viewer engine railway vehicle track machine spectator	locomitive railroad car viewer steam engine power place locomotion

Image annotation as bilingual analysis

- Annotated images as a bilingual corpus
- Images represented using two vocabularies.
  - Visterms – clustered image features.
  - Keyword Annotations.



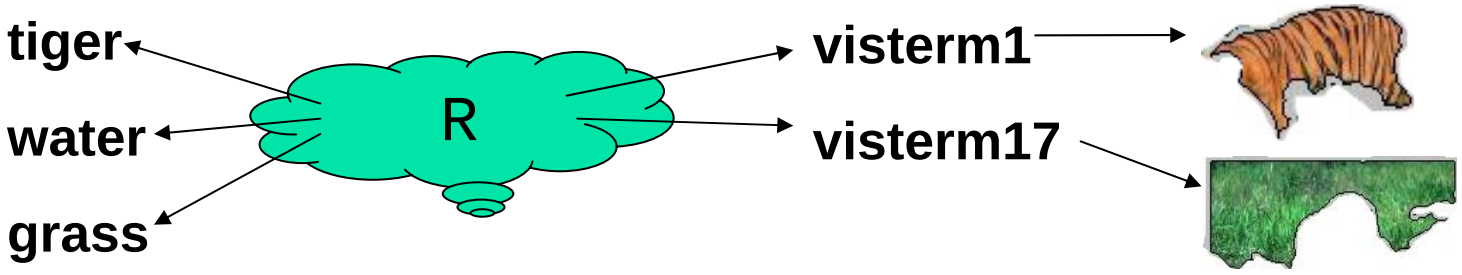
View as

- Machine Translation (*Dugyulu, Barnard, de Freitas and Forsyth*) or
- Cross-Lingual Retrieval Problem (*Jeon, Lavrenko and Manmatha)*

(Courtesy of R. Manmatha U. Mass)

Cross Media Relevance Model - CMRM

- Assumption: Each image-annotation pair is generated from a hidden relevance model.



- Relevance model is a joint distribution of words & visterms.
- Goal: Estimate the relevance model for each test image.
- Annotation for test image I

$$P(w \mid I) \approx P(w \mid v_1 \ldots v_m)$$

$$P(w, v_1 \ldots v_m) = \sum_J P(J) P(w \mid J) \prod_{i=1}^{m} P(v_i \mid J)$$

  - Mixture over all training samples J.

(Courtesy of R. Manmatha U. Mass)

Annotation Examples:

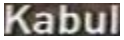
- Compute P(w|I) for different w.
- Probabilistic Annotation:
  - Annotate image with every possible w in the vocabulary with associated probabilities.
  - Useful for retrieval.

NCRM	**graphics_and_text**	**0.9529**
	text_overlay	**0.9413**
	non_studio_setting	**0.5939**
	people_event	**0.5830**
	face	**0.3551**
	male_face	**0.3453**
	...	**...**

(Courtesy of R. Manmatha U. Mass)

Indexing Components: Multi-Modal Analysis for VOCR

- Challenges of general video text detection/recognition
  - Transparency with cluttered background
  - Resolution : variable sizes, as small as 8x10 pixels
  - Different styles : color variation, fonts etc.
  - Some examples:



Text in video

Text with different styles

(Zhang & Chang CVPR 03)

VOCR using knowledge fusion

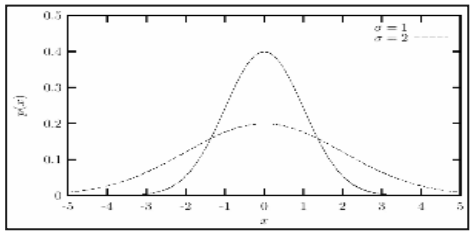
Bayesian Fusion for Video OCR

word

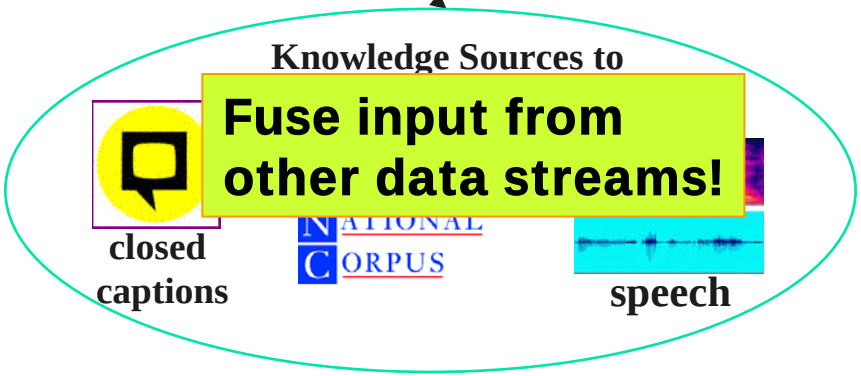
$$\hat{w} = \arg\max_{w} p(w \mid \mathbf{x})$$

image observation

$$= \arg\max_{w} \left[\log p(\mathbf{x} \mid w) + \log p(w) \right]$$



- Model likelihood of features
- Choose discriminative features (e.g., Zernik features for text)



$$p(w) = \alpha_{cc} p(w \mid CC) + \alpha_{BNC} p(w \mid BNC)$$

K2: CC K1: BNC

Automatic Video Highlight Extraction

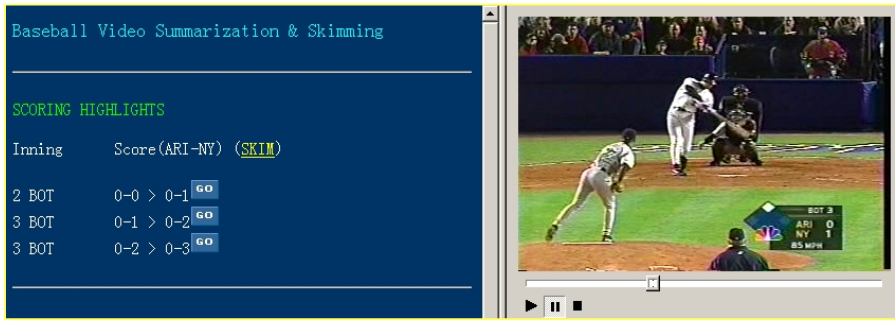
Interactive Event Browsing

Video highlight streaming

- Find semantic events in specific domains e.g., sports, news, surveillance, medical
- Match events to user preferences
- Save tremendous user time, bandwidth, and system power

Personal Sports Highlight System (Demo)

- Columbia's Sports Event Summary System
  - Random access to start of every play
  - Random access to start of every score and other events



Indexing Components: Detecting Image Near Duplicates (IND)

Zhang & Chang, 04

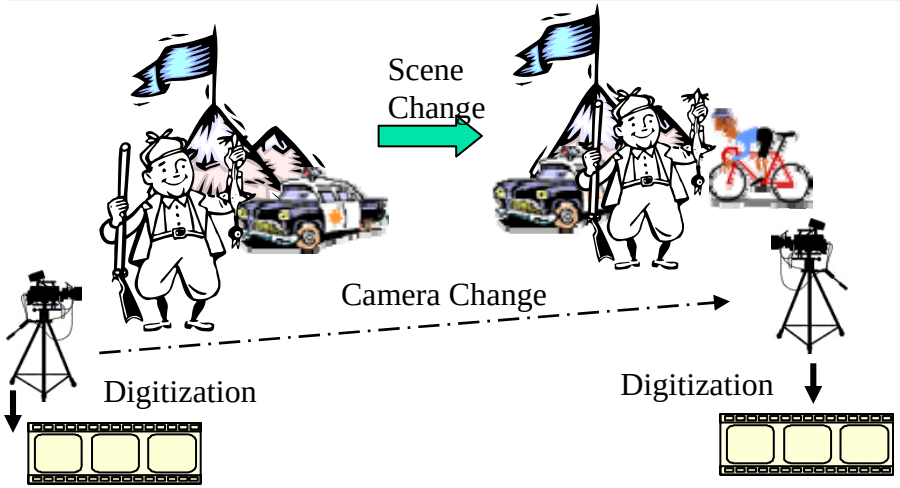
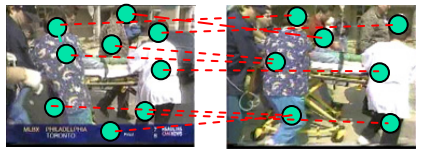


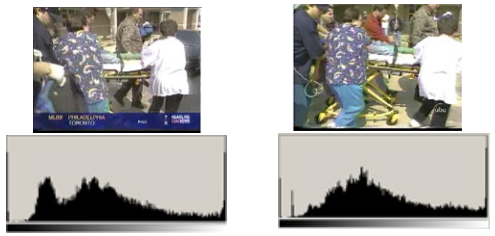
Image Near-Duplicate (IND) variations

- Scene changes: object movement, occlusion etc.
- Camera changes: view point change, panning etc
- Photometric changes: Lighting etc.
- Digitization changes: Resolution, gray scale etc.

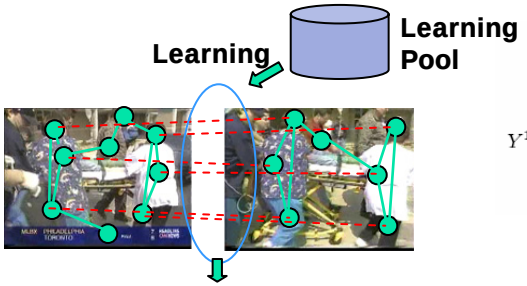
Conventional Approaches

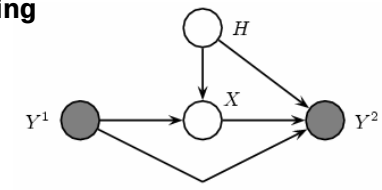
- Image registration or alignment
  - Compute Transform parameters
  - Warping images
- Global image features
  - Color histogram
  - Edge histogram
  - Model vector system





Stochastic Attribute Relational Graph Matching by Learning



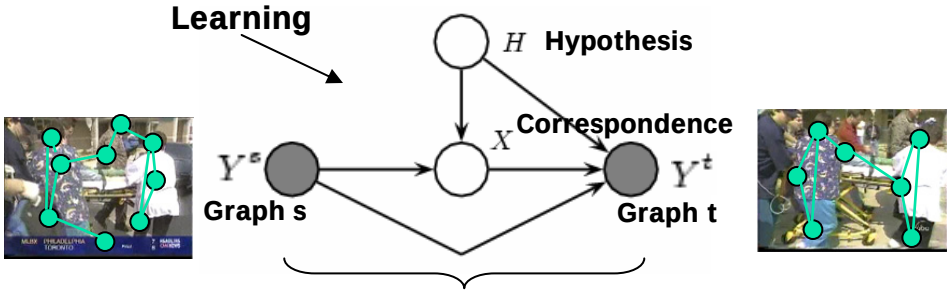


Stochastic Graph Editing Process

(demo)

Statistical Approach to Graph Matching

Stochastic Graph Editing Process



IND vs. non-IND Likelihood Ratio
→ a new similarity measure

Similarity by Likelihood Ratio

$$\text{Similarity} = \frac{\text{P(Graph\_t | Graph\_s, Two\_graph\_is\_IND)}}{\text{P(Graph\_t | Graph\_s, Two\_graph\_is\_not\_IND)}}$$

- **Compute Likelihood**

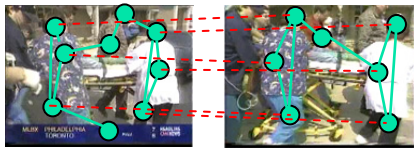
$$p(Y^t|Y^s,H) = \sum_{x\in\chi} p(Y^t|Y^s,x,H)p(x|Y^s,H)\ ,\quad \chi=\{0,1\}^{N\times M}$$

Intractable ! → **So approximate it by using Jensen's lower bound**

$$\ln p(Y^t|Y^s,H=h) \geq \sum_{iu,jv} \hat{q}(x_{iu},x_{jv}) \ln \psi'_{iu,jv}(x_{iu},x_{jv};y^s_{ij},y^t_{uv}) + \sum_{iu} \hat{q}(x_{iu}) \ln \phi'_h(x_{iu},y^s_i,y^t_u) + \mathcal{H}(\hat{q}(x)) + \text{constant}$$

Learning

- **Learning by node-level annotation**



Learning is realized by Parameter Computation

- **Learning by image-level annotation**

Positive Samples

Negative Samples

Learning is realized by Variational E-M

1. **Inference : Compute the approximate distribution $\hat{q}$ by Loopy Belief Propagation**
2. **Learning : Estimate $\psi'_{iu,jv}$ and ϕ'_h by using E-M**

Opportunities Beyond Components

- Statistical Fusion of Multi-Modal Tools

Case Revisited: Multi-modal search

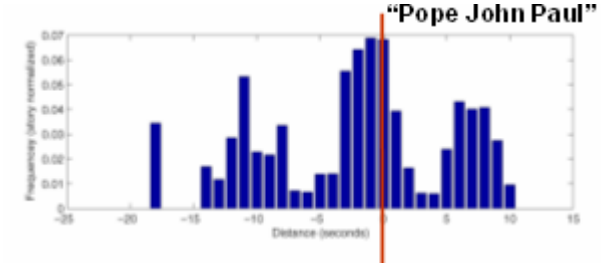
Query

Find
shots of
Pope John
Paul second

Story

Shot **Shot** **Shot** **Shot** **Shot**

in other news **pope john paul** the **second** will get his first look at the shroud of turin today that's the piece of linen many believe was the burial cloth of jesus the round is on public display for the first time in twenty years it has already drawn up million visitors the pope's visit to northwest italy has also included beatification services for three people the vatican says **john paul** is now the longest serving pope this century he has surpassed **pope** pious the twelfth who served for nineteen years seven months and seven days



- Retrieval on Person-X:
  - find shots that person-X appears visually
  - Important features:
    named entities in text, face recognizer, and their correlation distributions
- Results on Person-X retrieval suggest query-specific model (QDM) for fusing multimodal features

Query Dependent Model (QDM) for Retrieval

- Use extensively for question-answering in text
  - Perform query analysis to identify question type and answer target
  - Employ appropriate model for answer selection
- Work by [Yan *et al*, ACM Multimedia 2004]
  - Consider 4 query classes:
    Named Person, Named Object, General Object, Scene
  - Train different QDMs for each class using EM algorithm
  - Search Tools/Features: ASR text, image similarity retrieval (color, texture), anchor, commercial, news subject monologue, and face
  - Tested on TRECVID 2003 corpus with 25 queries

Query Model	MAP (Mean Average Precision)
Text feature only:	0.15
query independent model	0.18
Query dependent model:	0.21

Query Model

Multi-modal

Query Dependent Model for Retrieval -2

- [Chua *et al*, TRECVID 2004] further explore the use of query expansion for news video retrieval
  - Query expansion by including terms from parallel information sources – general Web and Parallel Info Sources (AQUANT corpus)
  - Perform pseudo relevance feedback on text and visual features
  - Consider 6 query classes: Person, Sports, Finance, Weather, Disaster, General
  - Train query specific query-dependent model (QDM) for each class
  - Tested on TRECVID 2004 corpus with 25 queries

	Use General Web for Query Expansion	Use Parallel Info Sources for Query Expansion
Text only w/o QDM	0.047	0.058
Text with QDM	0.071	0.078
Multi-modal with QDM	0.119	0.123
Multi-Modal with QDM + PRF	0.127	0.130

Multi-Modal

Query Exp.

Query Model -- Determine Fusion of Multi-modality Features

$$Final\_Score(S_i) = \sum_{all-\mathrm{mod}alities} \alpha_i^M \bullet Score_i$$

[Chua et al 04]

Class	Wt of NE in Expanded terms	Wt of OCR	Wt of Speaker Identⁿ	Wt of Face Recog-nizer	Weight of Visual Concepts (total of 10 visual concepts used)					
					People	Basket-ball	Hockey	water-body	fire	Etc
PERSON	High	High	High	High	High	Low	Low	Low	Low	.
SPORTS	High	Low	Low	Low	Low	High	High	Low	Low	.
FINANCE	Low	High	Low	High	Low	Low	Low	Low	Low	.
WEATHER	Low	High	Low	High	Low	Low	Low	Low	Low	.
DISASTER	Low	Low	Low	Low	Low	Low	Low	High	High	.
GENERAL	Low	Low	Low	Low	High	Low	Low	Low	Low	.

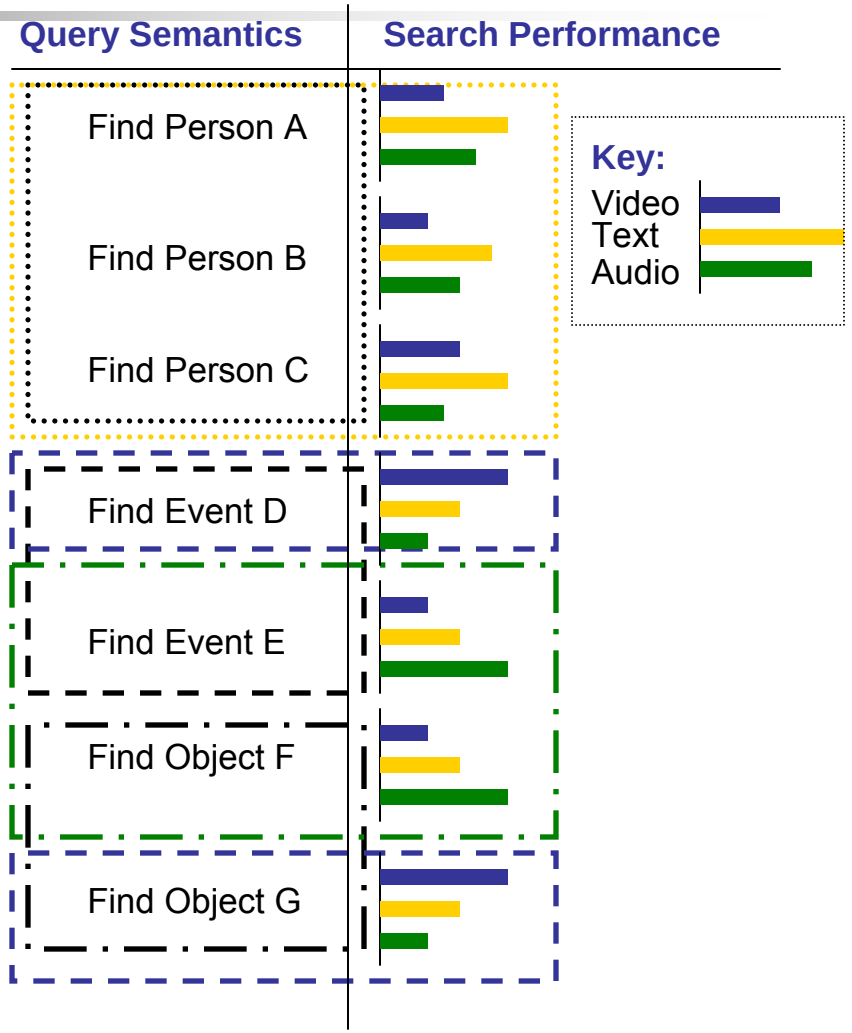
Challenge

- How to automatically discover query classes?
- When and how does each modality help for each query?

Mining of MM Query Classes

Kennedy, Natsev, & Chang, ACMMM 05

- Existing methods: define query classes using human knowledge.
- New method: discover queries according to performance of different searches.



To make query class meaningful: Semantic Space Similarity

- Extract semantic features of each query:
  - counts of nouns, verbs, and named entities (persons, locations, and organizations)
- To map new queries:
  - Compute distance between queries:
  - Wordnet distance
  - cosine distance of semantic features

Query Pool

- 23 from TRECVID 2004
- 25 from TRECVID 2003
- 25 from IBM Speech Group (labeling in progress)
- 130 from BBC logs (labeling in progress)

- Conduct pooled labeling using top results from various searches.
  - Approx. 4000 labeled results per query

Query Examples

- TRECVID 2003
  - "Find shots of an airplane taking off."



- TRECVID 2004
  - "Find shots of Bill Clinton speaking with at least part of a US flag visible behind him."

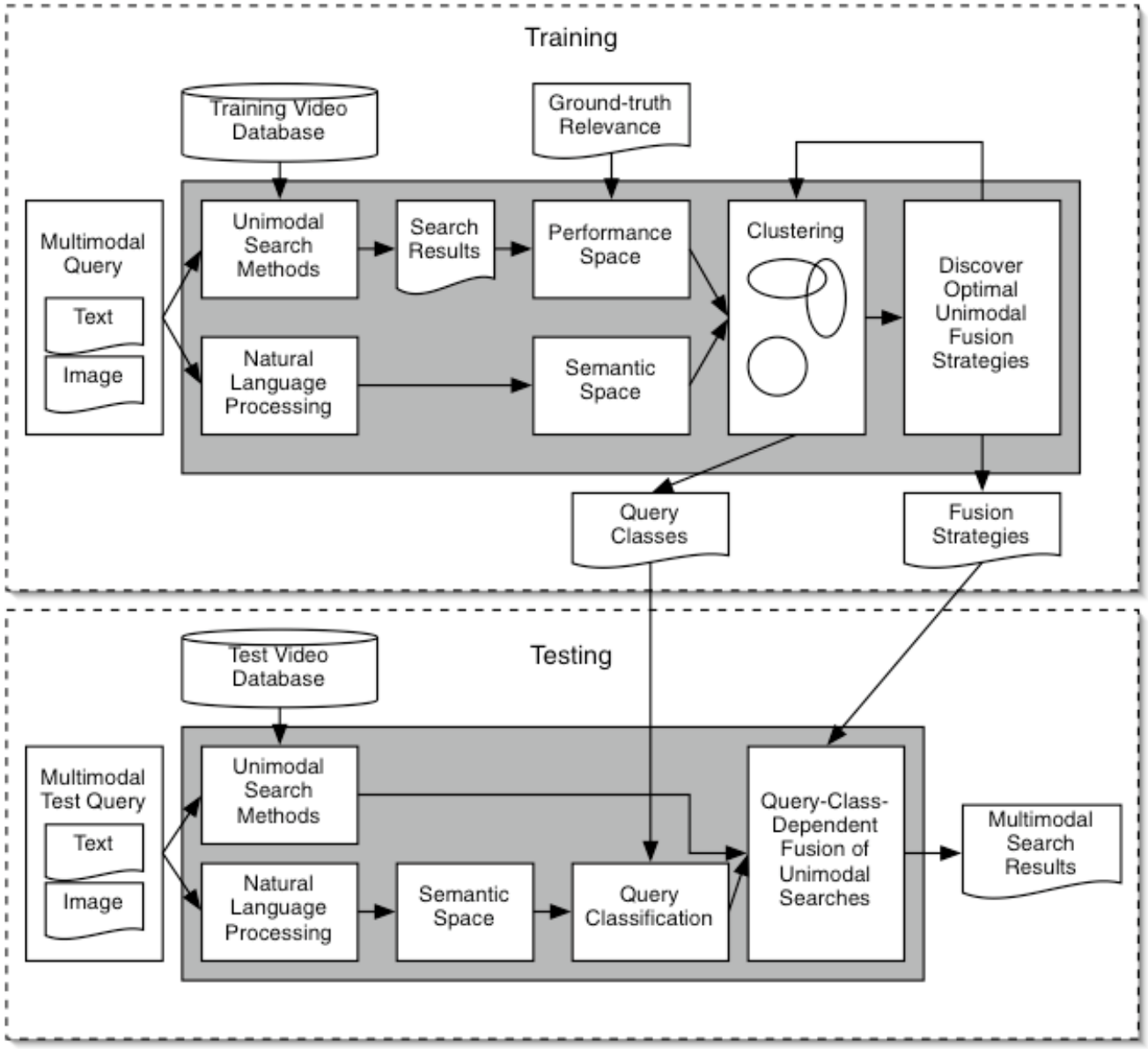


- IBM Speech Group
  - "Find shots containing monkeys or gorillas."


- BBC Logs
  - "Find shots of the Kremlin."



Query Class Mining System



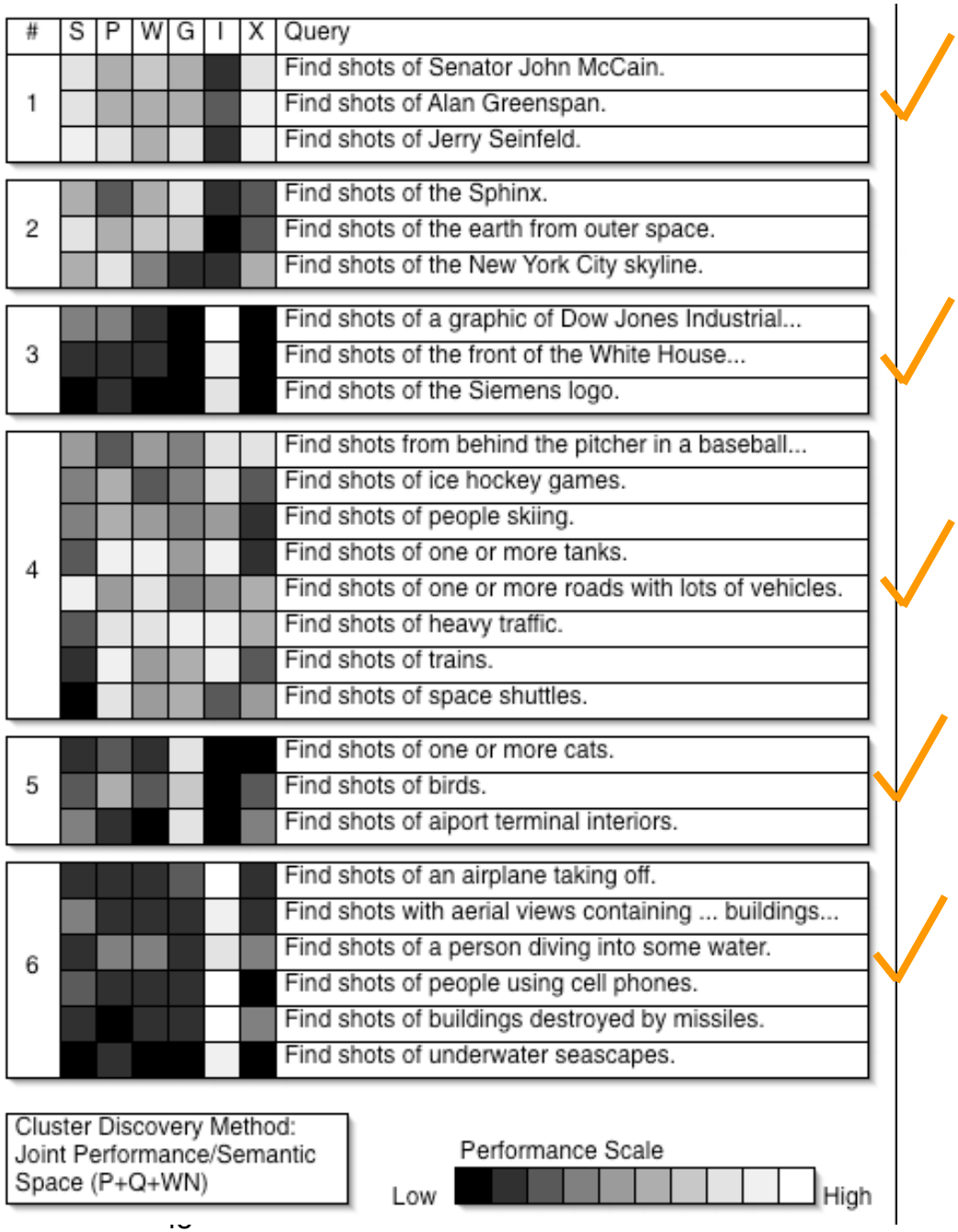
Performance

	Method	Classification		
		O	SD	SVM
Query-Independent	Text	0.0595	-	-
	Multimodal	0.0595	-	-
Hand-defined Classes	CMU	0.0605	-	-
	NUS	0.0598	-	-
Automatically Discovered Classes	P+Q+WN	0.0748	0.0673	0.0359
	P+Q	0.0749	**0.0711**	0.0478
	P+WN	0.0745	0.0645	0.0421
				0320
				0465
				0533
	WN	0.0599	0.0590	0.0544

Confirm best result by
Joint query class mining using performance and semantics

Discovered Query Clusters

- *named persons*: text search and person-X search most useful
- image search benefits *named objects*, *sports*, and *generic scene classes*.
- An interesting *Goggle* class is discovered.



Revisited: Topic Tracking across Multiple News Channels and Web Pages

A government sponsored IBM-Columbia Joint Project

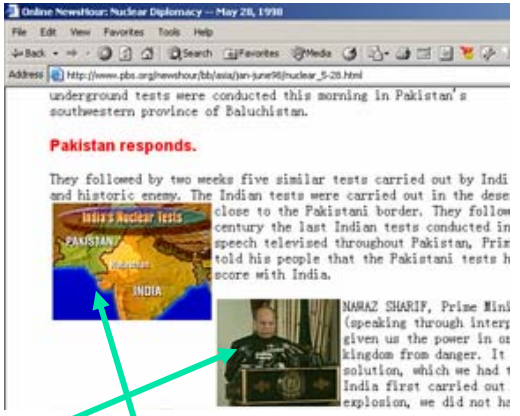
Source # 1

Source # 2

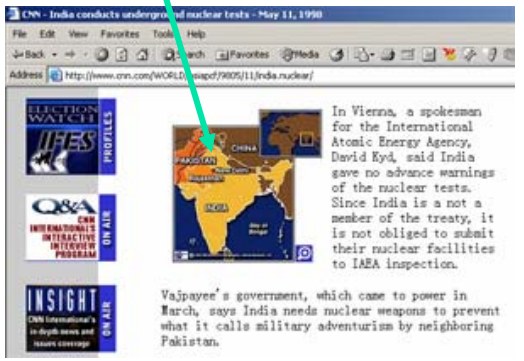
Source # 3

Threading News Stories

Broadcast sources



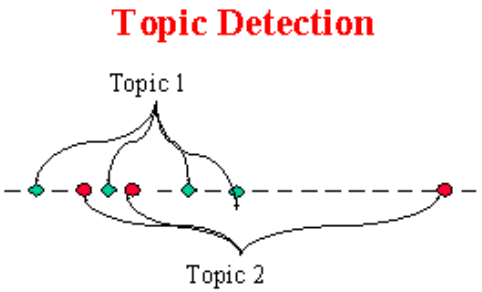
News Web Site 1



News Web Site 2

Text Mining vs. Video Mining

(Topic Detection and Tracking)



Topics: text

Topics: text, scenes, objects

documents

Asian Economic Crisis
Monica Lewinsky Case
War in Iraq
McVeigh's Navy Dismissal
Philippine Elections
Israeli Palestinian Raids
Fossett's Balloon Ride
Casey Martin Sues PGA
Karla
Moun
State
Pope visits Cuba

broadcast video

Hurricane in FL

Disaster

German train derails

- **Addition of A-V information results in needs of sub-topics.**
- **Common and unique visual fists across topics**

Sample topic 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
monica
investigation
president
… …

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

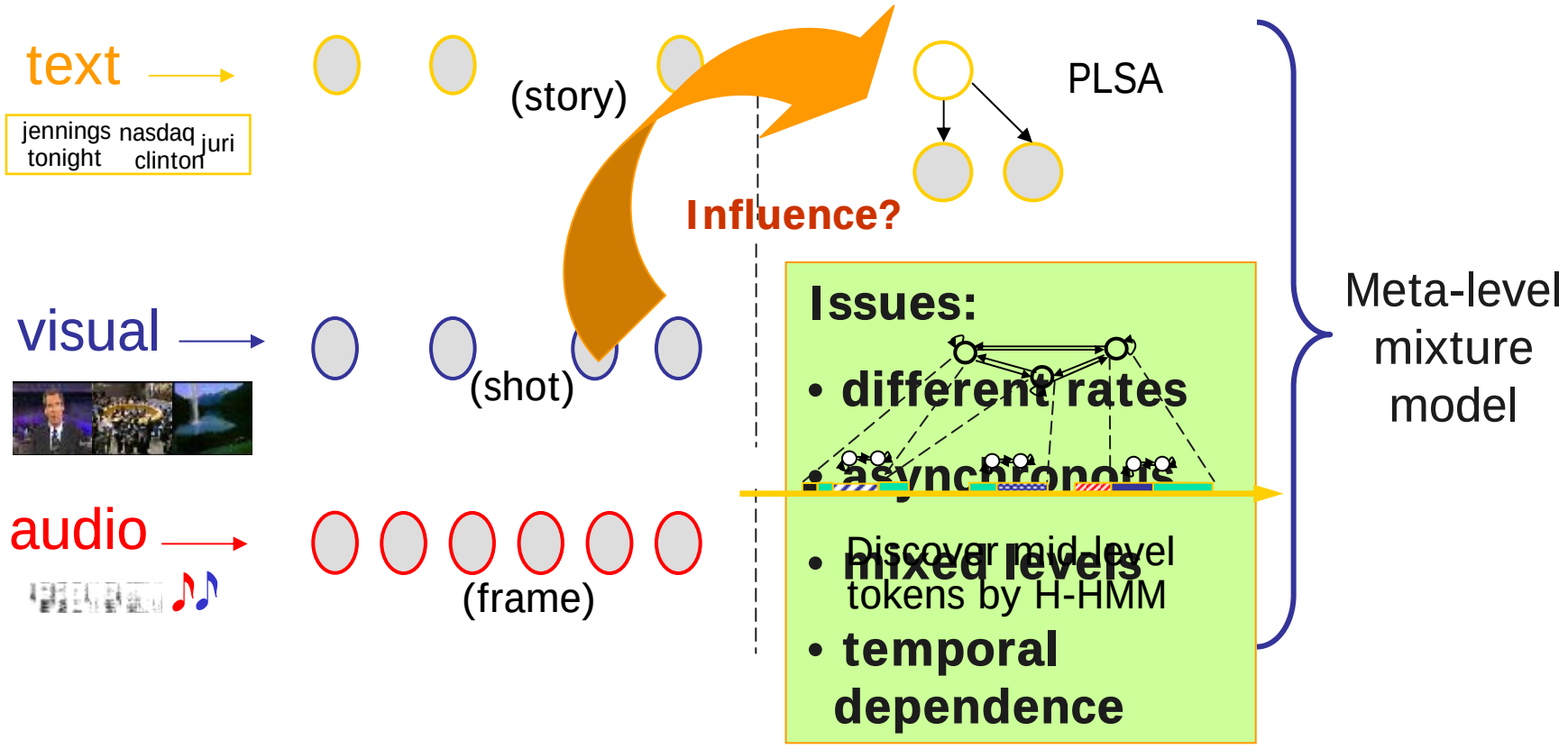
"random" clusters

cancer
increase
secure
temperature
texas
accusation
chance
nasdaq
pressure
center
… …

cancer
africa
temperature
movie
coast
center
rain
strike
… …

- **Text-based clusters reveal semantics, but not AV aspects.**

Use A-V features to refine text clusters

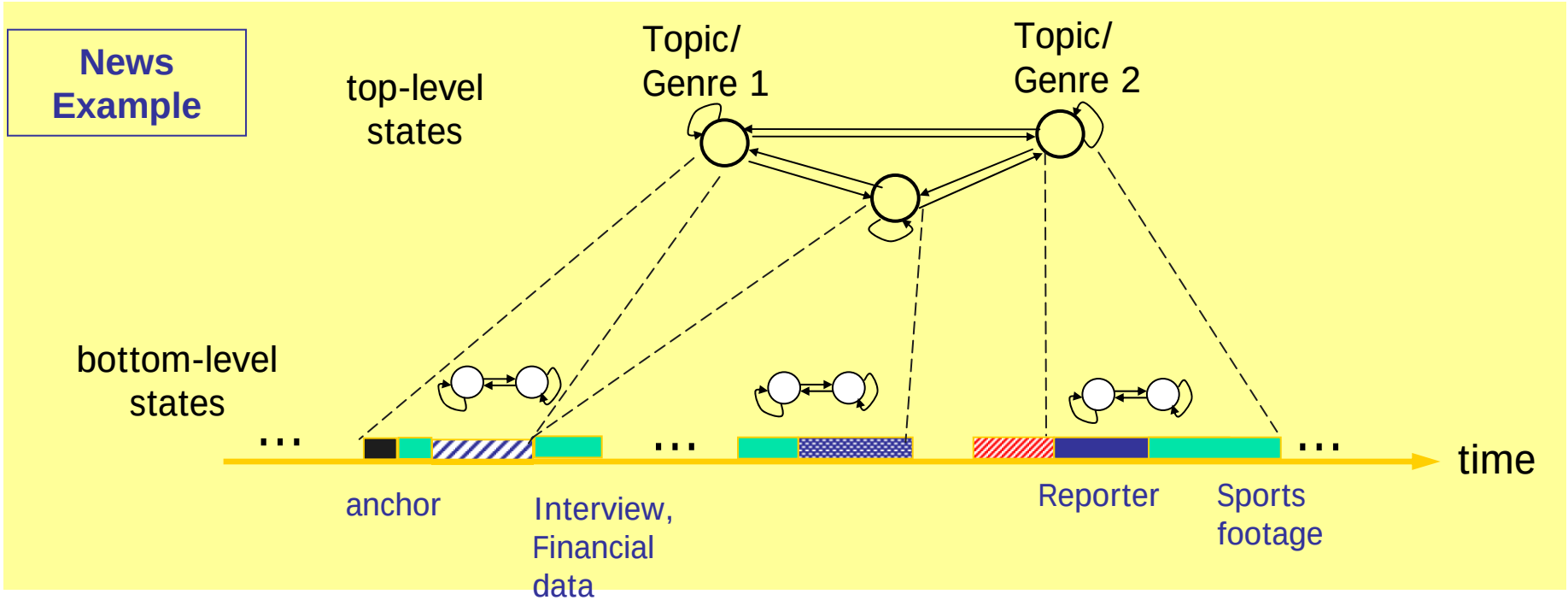


Patterns in Video: Temporal is important

AV Temporal Pattern Mining: A Case for Hierarchical HMM

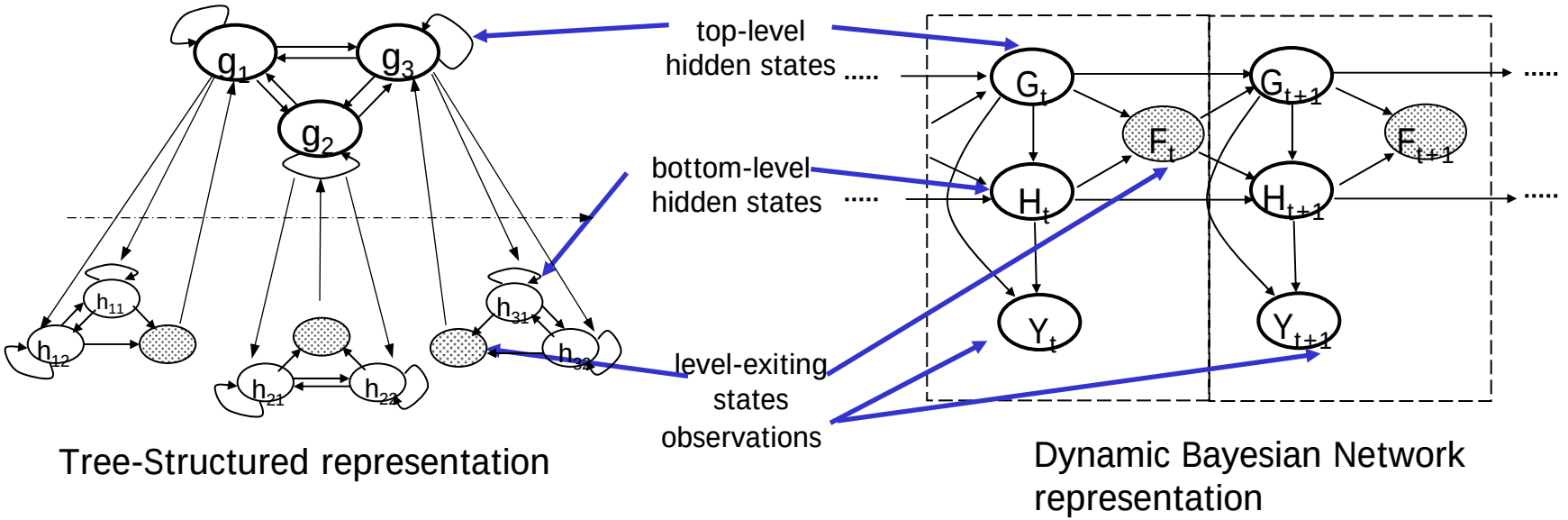
(Xie, Chang, et al '02)

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



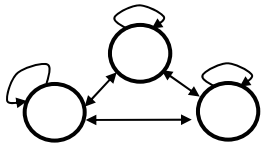
Hierarchical HMM

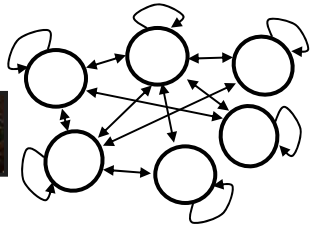
[Fine, Singer, Tishby '98]
[K. Murphy, '01] [Xie et al '02]

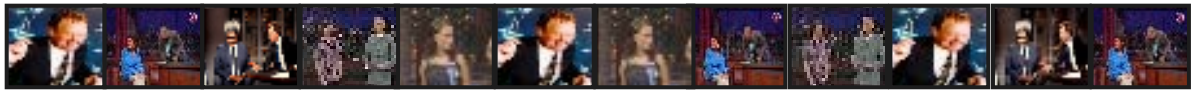
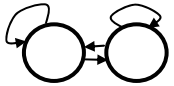


- Flexible control structure (bottom-up control with exit state)
- Extensible to multiple levels and distributions
- Efficient inference technique available
  - Complexity $O(D \cdot T \cdot Q^{\alpha D})$, α=1.5 to 2
- Application in unsupervised discovery has not been explored
  - Questions: how to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

The Need for Model Selection

soccer 

?

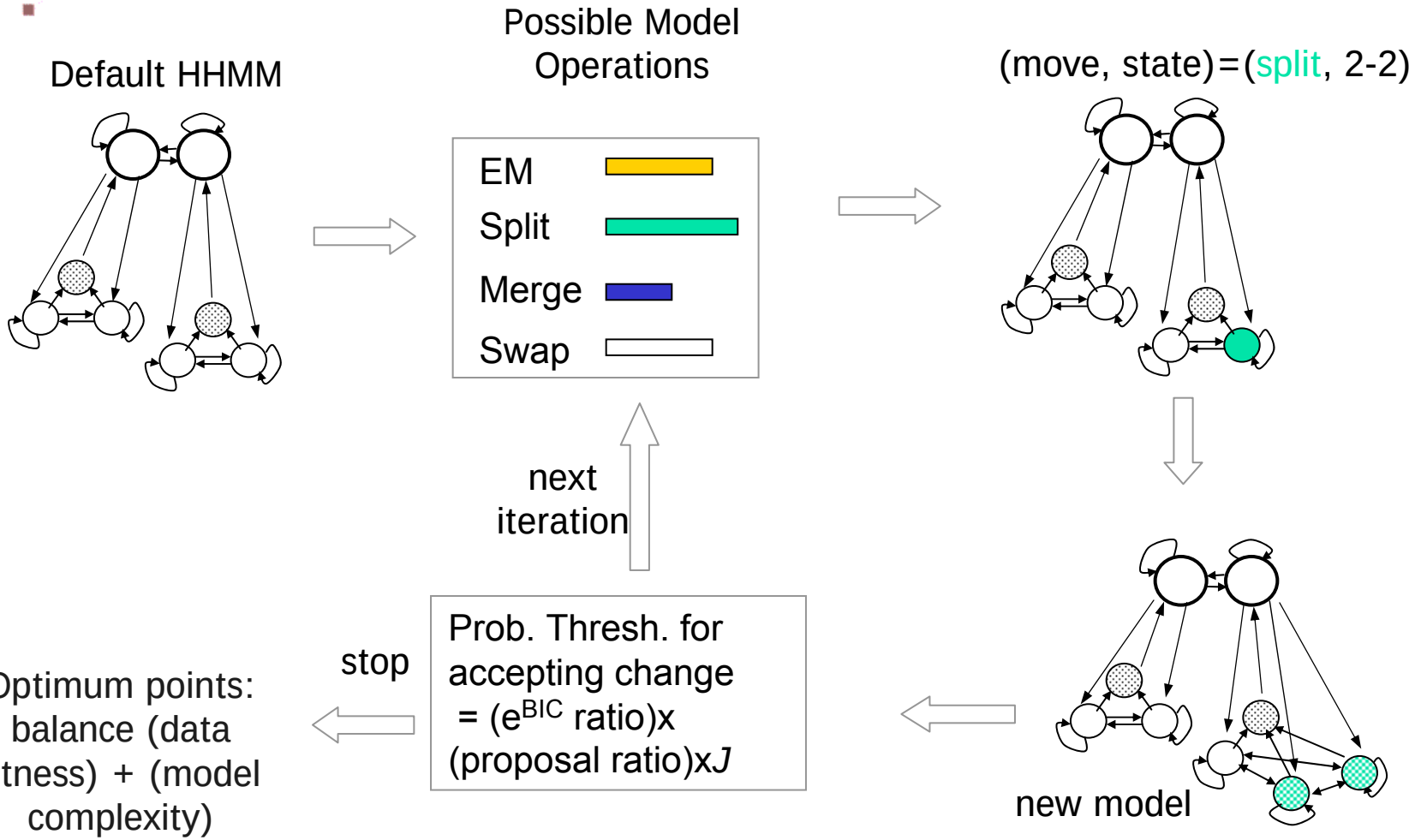
news 

?

talk show 

?

- Different domains have different descriptive complexities.

[Green95]
[Andrieu99]
[Xie ICME03]

Model Selection with RJ-MCMC



Select compact relevant features

MFCC

color histogram

zero-crossing rate

edge histogram

delta energy

zernike moments

LPC coeff.

Spectral rolloff

pitch

Gabor wavelet

motion estimates

keywords

logtf-entropy

tf-idf

people?

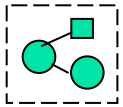
face?

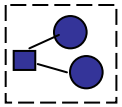
outdoors?

vehicle?

jennings tonight nasdaq clinton juri accus lawyer

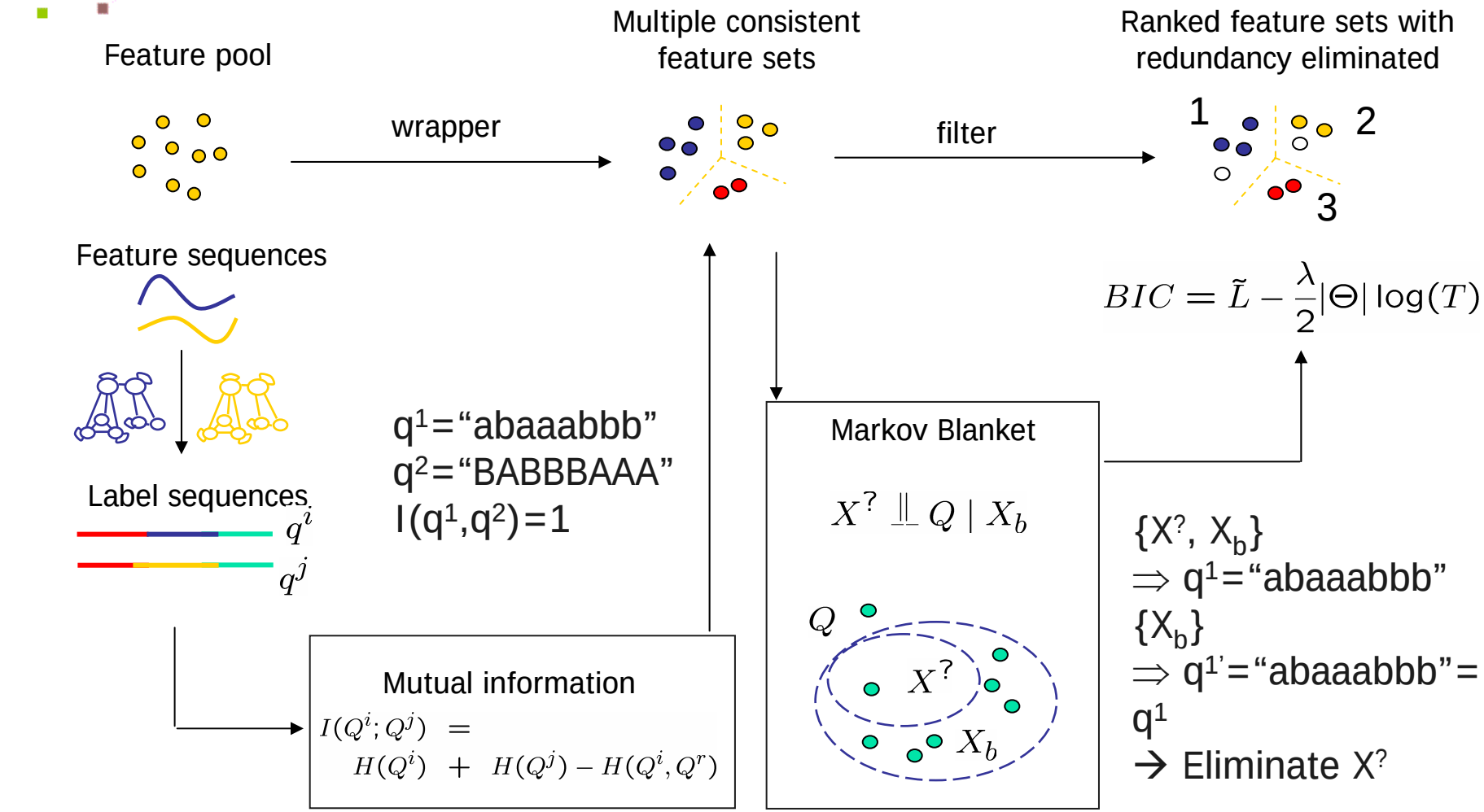
?



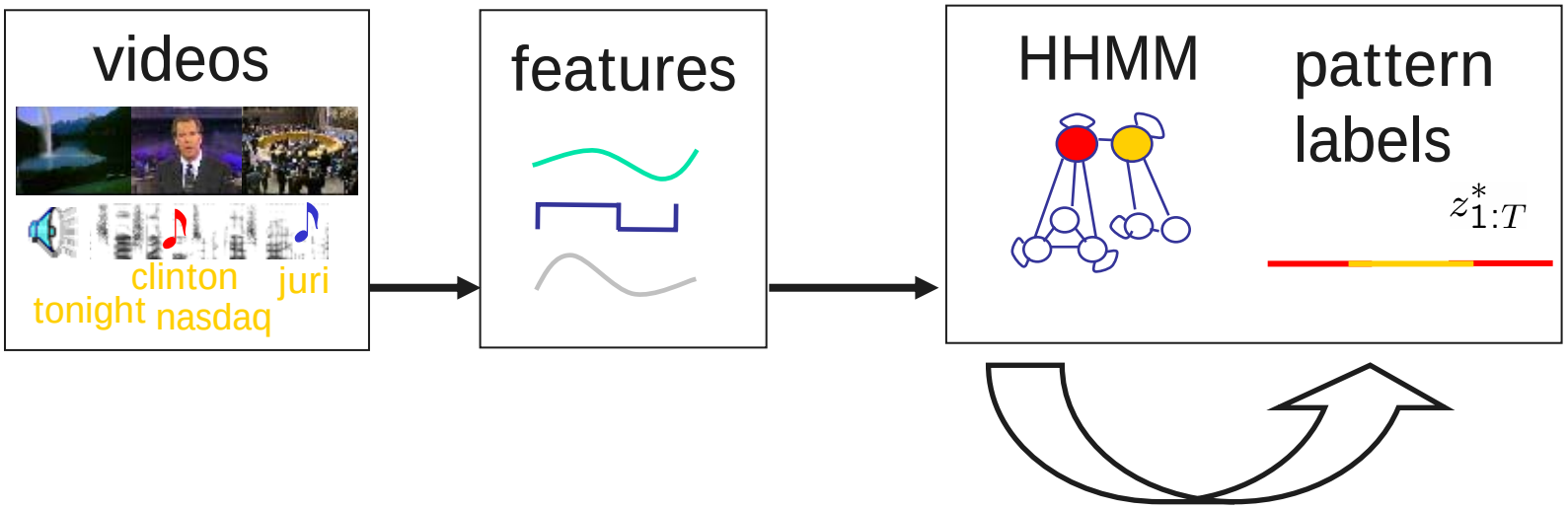


Feature Selection for Temporal Pattern Mining

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



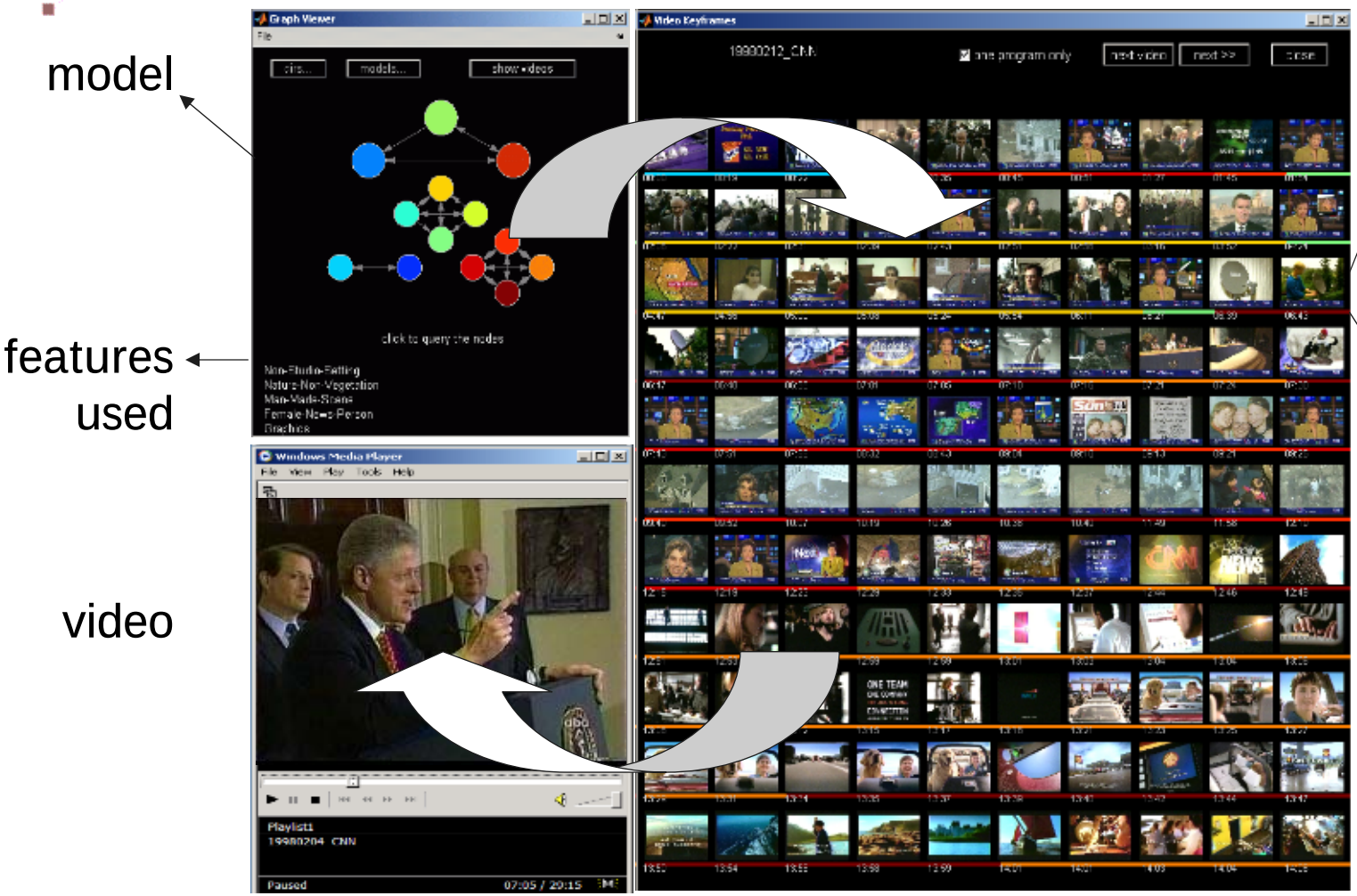
Mapping Videos to Patterns



Maximum-likelihood
state sequence decoding

Xie & Chang '05

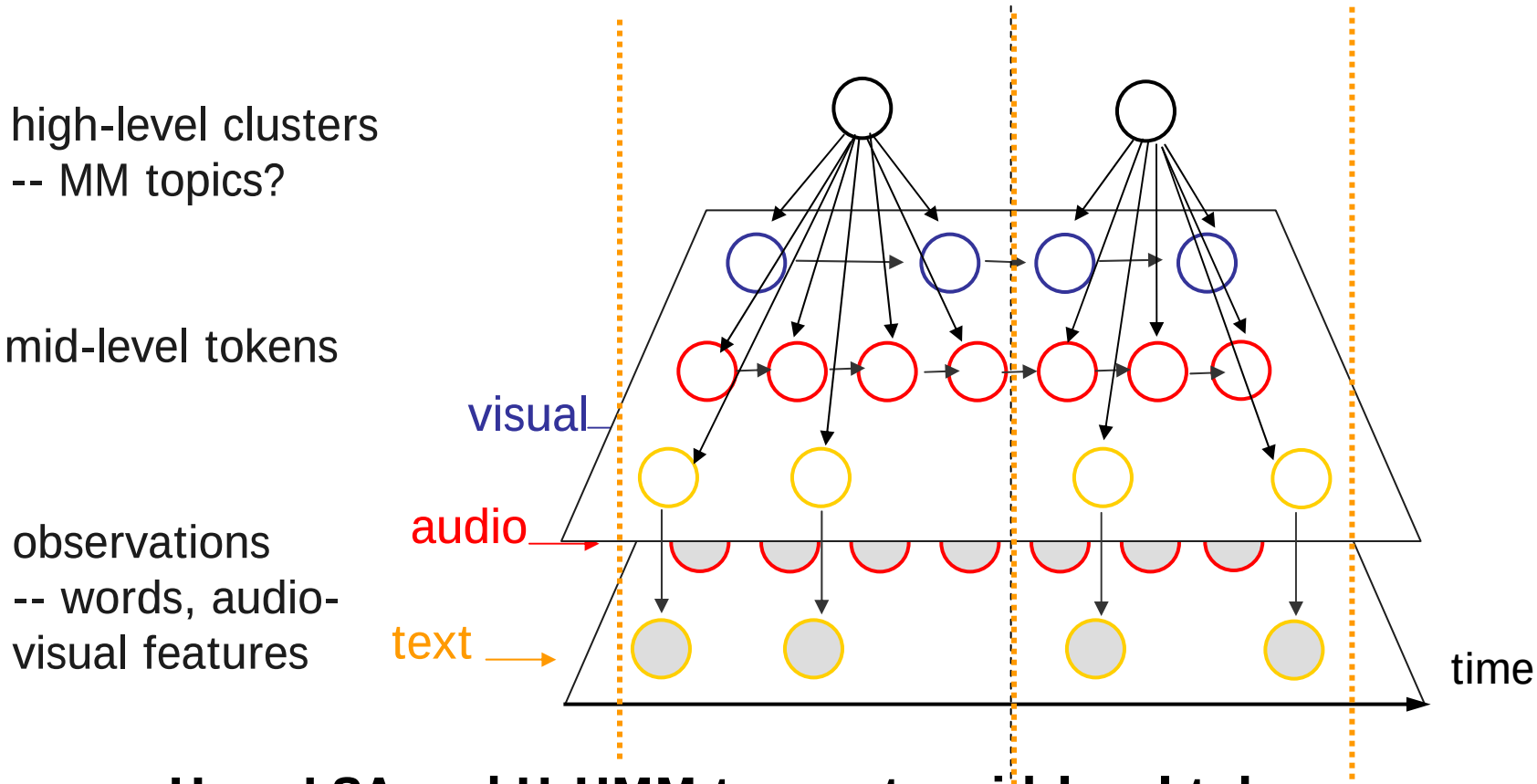
The HHMM Pattern Navigation System



Demo

[Xie *et al*, ICIP'03, ICASSP'05]

Multi-Modal Layered Mixture Model

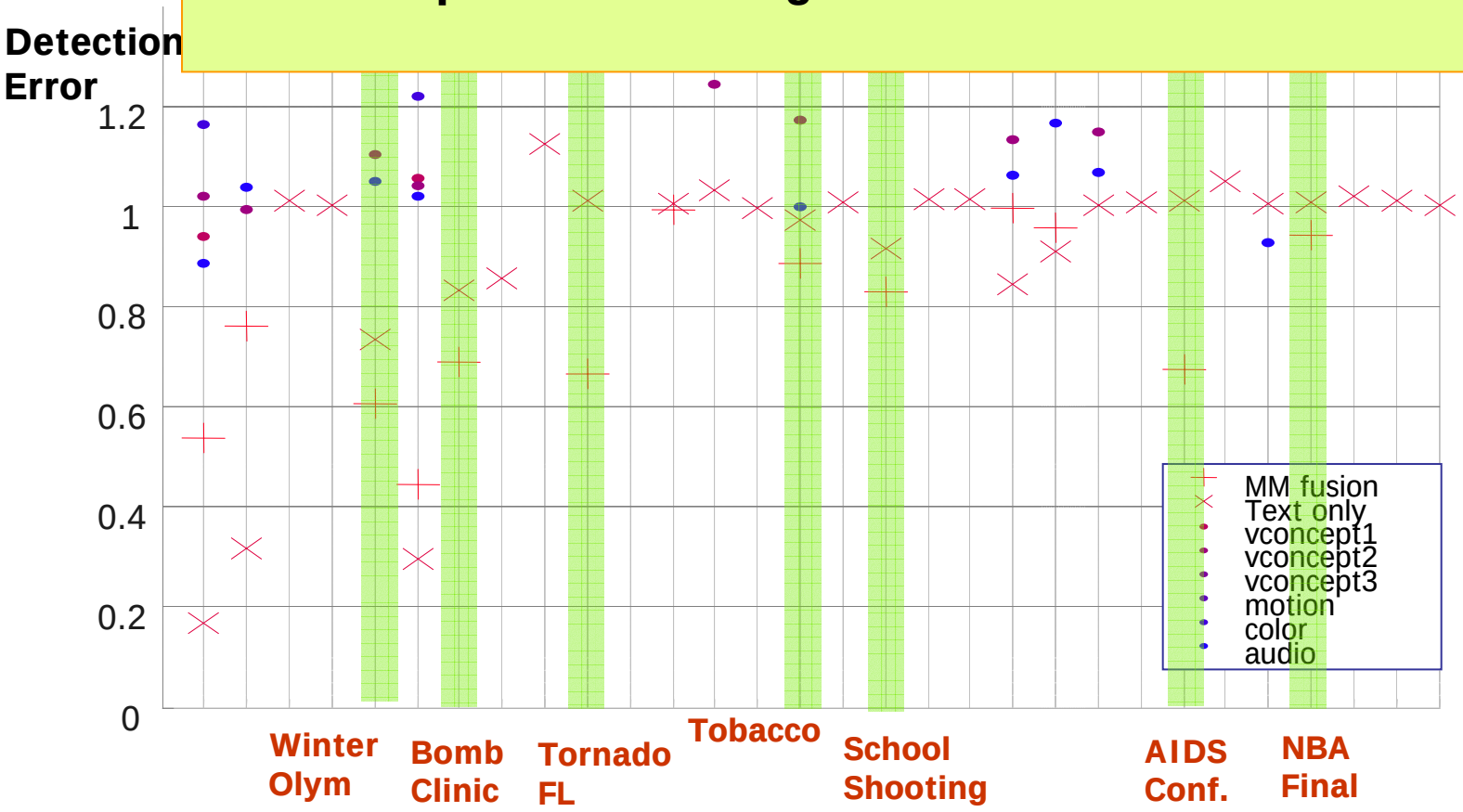


- **Use pLSA and H-HMM to create mid-level tokens**
- **Use story structures to define co-occurrences of tokens**
- **Top-level mixture for capturing the latent semantic aspects**

Topics Improved by MM Fusion

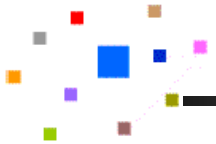
-- Measure overlap of discovered clusters with TDT-2 topic ground truth

- **7 out of 30 topics show improvement by using LMM**
- **Such topics show strong audio-visual cues**



(demo 1)

(demo 2)



Conclusions

- Video Search and Mining offers an exciting field
  - Imminent demands in practical applications
  - Opportunities for advances in image analysis, high-level vision, IR, statistical modeling.
- Benchmark processes and dataset available for checking progress
  - TRECVID, LSCOM (video ontology), Yahoo API
  - Remember the early years of image retrieval research?

Conclusions (2)

- Strategies of multi-modal fusion depend on the query target and user context
  - Determine when and how each search tool is useful
- Video mining promising for discovering query classes and salient events
- Features
  - Low-level similarity matching useful for certain query classes (objects and scenes)
  - High-level concepts (people, location, objects) are useful for filtering and search
  - Use of external information (text query expansion) is promising for retrieval

Open Issues

- Find the right paradigms for mapping
  Tools → User interfaces → User Context
  - For Home, Web, Mobile platforms
- Continued pursuit of effective recognition models
  - content understanding (esp. event!)
  - information retrieval
  - topic tracking and summarization
- Exploit the use of ontology and knowledge sources
- Exploit existing and new dataset and evaluation
  - realistic use scenarios
  - feature/data pool, golden standards, and copyright issues
  - TrecVid benchmark 2002-5

Acknowledgment

- Columbia University
  - W. Hsu, L. Kennedy, Y. Wang, L. Xie, D.Q. Zhang,
- Some topics are joint work with
  - A. Divakaran, M. Franz, G. Iyengar, C. Lin, J.R. Smith, H. Sun
- Additional Slide Sources
  - University of Massachusetts
    - R. Manmatha
  - National University of Singapore
    - T.-S. Chua

Other projects

- Sports highlight summarization
- Generation of low-power light-weight H.264 video streams
- TrustFoto: Image tampering detection
- Echocardiogram Medical Video Indexing

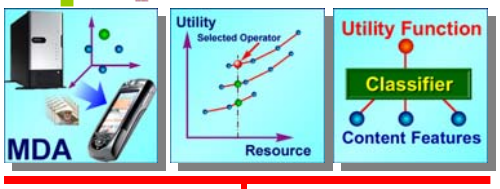
Automatic Video Highlight Extraction

Interactive Event Browsing

Video highlight streaming

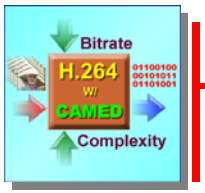
- Find semantic events in specific domains e.g., sports, news, surveillance, medical
- Match events to user preferences
- Save tremendous user time, bandwidth, and system power

Video Coding



Video Adaptation & Scalable Coding

H.264/AVC



Power-Aware Video Coding

Enhancing compression performance & network-friendly representation

Universal Media Access

Mobile Video Application

Traditional Video Coding
H.261,3(+, ++),
MPEG-1, 2, 4, FGS

Light encoder & Heavy decoder

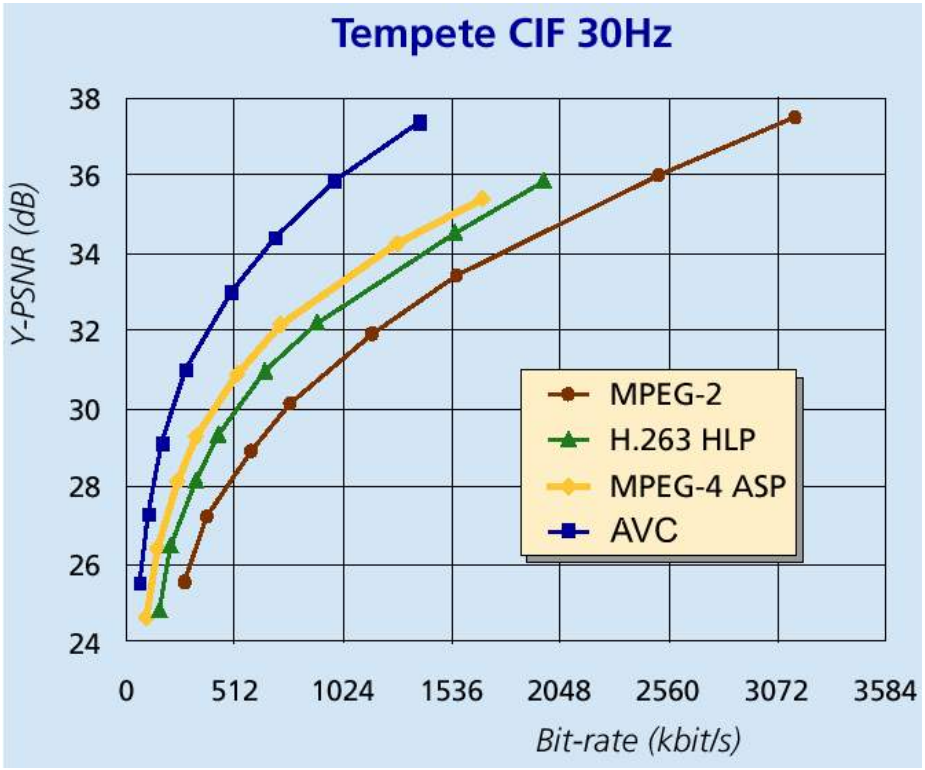
Others

Distributed Video Coding

Multi-view Video Coding

Power efficient video streams

with Yong Wang



- State of the art H.264 doubles the capacity, but consumes much more power.
- We have developed a new technique to reduce the core cost by 60%.
- [demo]

Image Forgery Detection: Columbia TrustFoto project

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

with Tian-Tsong Ng

- 10% of color photos published are retouched or altered [WSJ '89]
- March 2003: A Iraq war news photograph on LA Times front page was found to be a photomontage
- Feb 2004: A photomontage showing John Kerry and Jane Fonda together was circulated on the Internet
- Adobe Photoshop: 5 million registered users
- Image Manipulation Contest: www.worth1000.com, 85,000 work

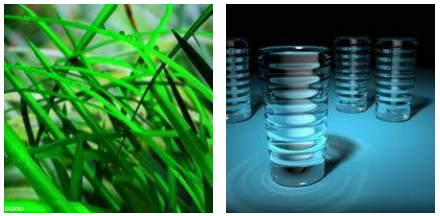
Images downloaded from http://www.camerairaq.com/faked_photos/

Related Problem: Image Source Identification

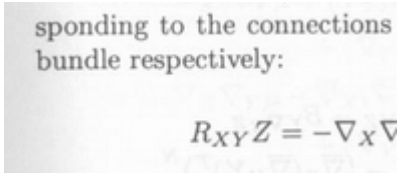
- Identify image production devices: camera, computer graphics, printer, and scanner, etc.

From which camera?

CG Or Photo?



From which printer?



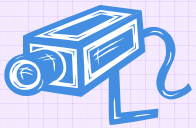
Images from
http://www.alias.com/eng/etc/fakeorfoto/

Users


Forensics Investigation


Criminal Investigation


Insurance Processing


Surveillance video


Intelligence Services

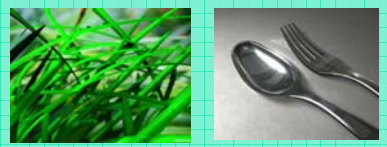

Financial Industry


Journalism

Input Images


Camera Images


Computer Graphics



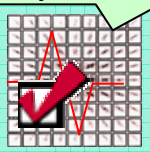
TrustFoto System

With Expert Intervention

Acquisition Device Modeling

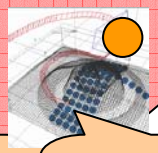
Natural Image Statistics Analysis

Image Manipulation Detection



Signal Processing Approach

Computer Graphics Approach

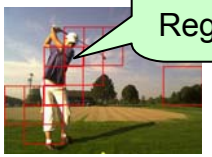


3D Geometry Reconstruction

Inverse Rendering

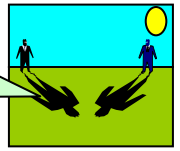
3D Scene Consistency Checking

Diagnostic Output/ Decision Report



Suspicious Regions

Inconsistent Shadows



Report

Smoothing

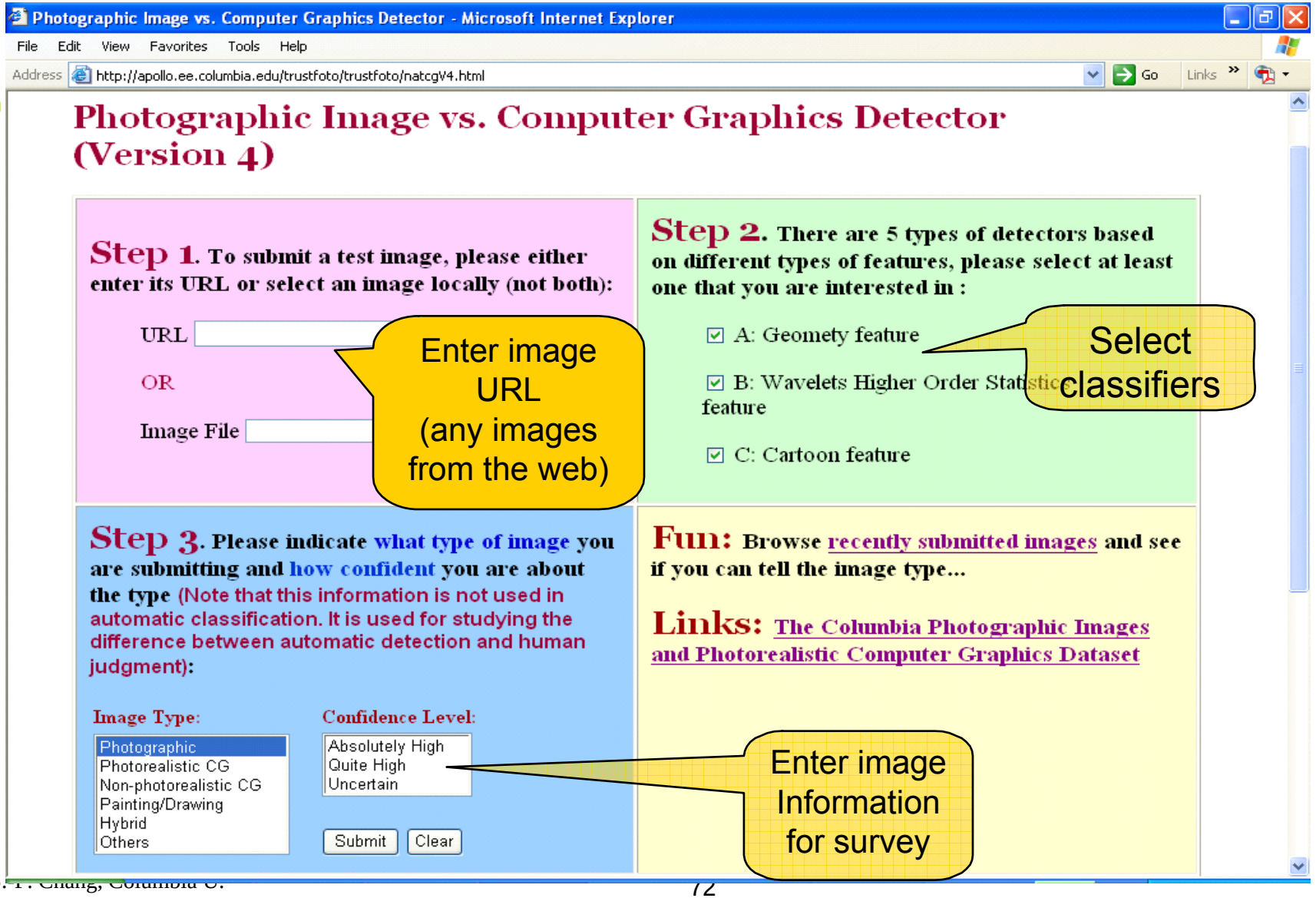
Splicing

Sharpening

Computer-Graphics

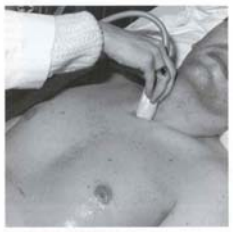
Columbia CG-Photo Online Demo

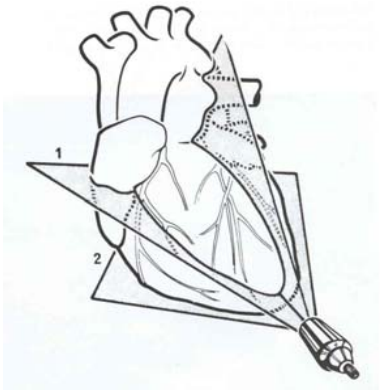
URL: http://www.ee.columbia.edu/trustfoto/demo-photovscg.htm

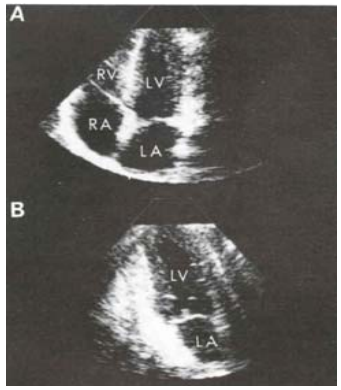


Echocardiogram Video – Digital Library & Remote Medicine

(Ebadollahi, Chang, & Wu '01 '02]



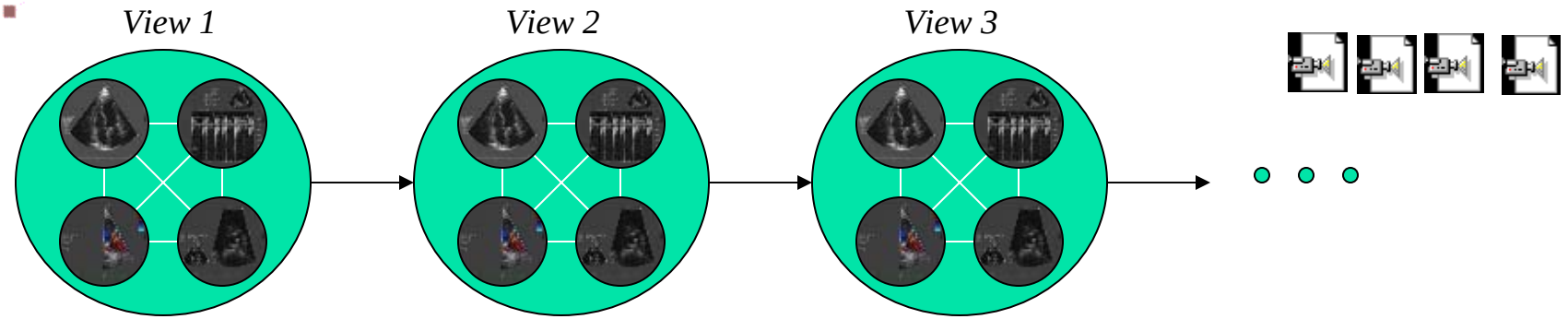


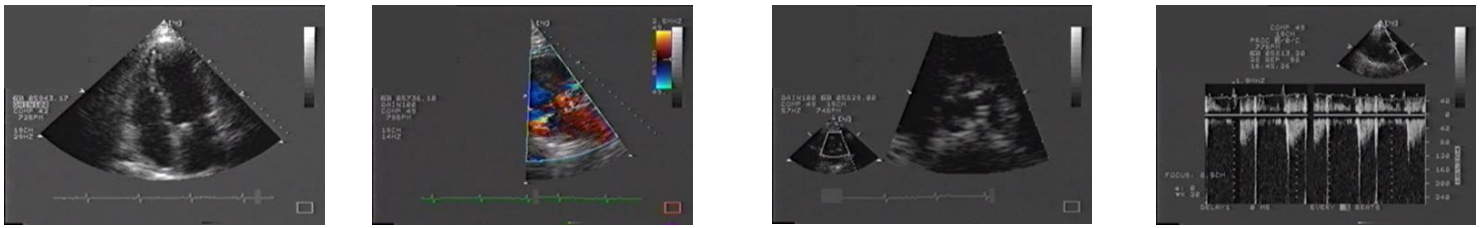


(@1994 from Echocardiography by Harvey Feigenbaum. Reproduced by permission of Lippincot Williams & Wilkins, Inc.)

- Remote patients may not have access to clinical specialists
- Lossy video compression and transmission may not be acceptable
- Semantic/syntactic summary provides an effective solution.

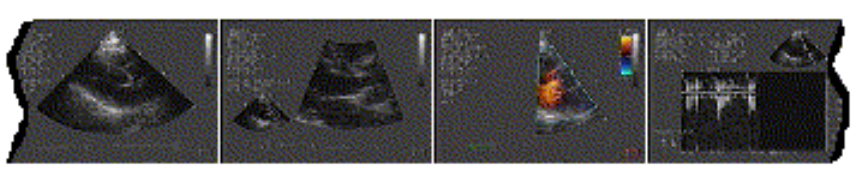
Analyze spatio-temporal structures



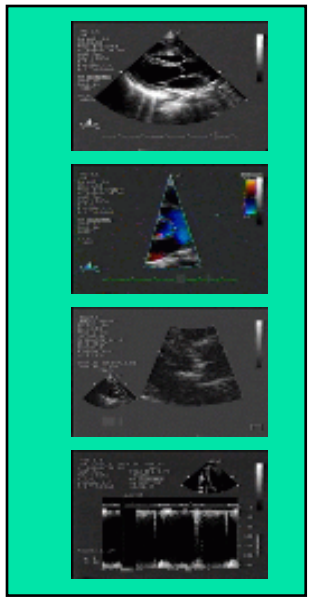


- Deterministic patterns following AAC standard + random orders in actual production → object/scene modeling and detection
- Content-adaptive transmission
  → Transmit selective views/beats/frames only, details on demand

Echo Video Digital Library & Remote Medicine



Echo Video Acquisition



Domain Knowledge

View Recognition

Video Clinic Summary

Diagnosis Reports

Augmented User Interface

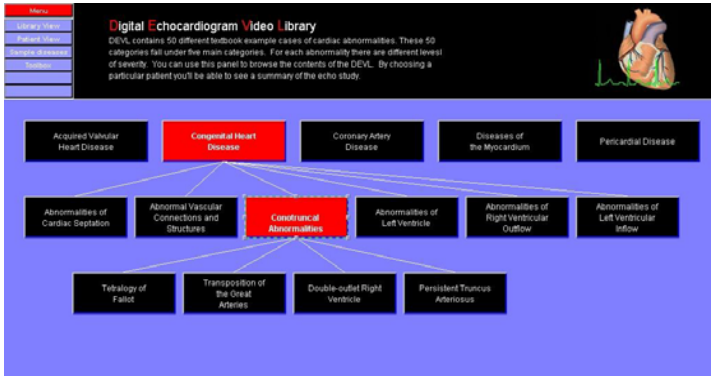
Event /Abnormality Detection

- Database/Teaching
- Selective storage/ transmission

View-based browser

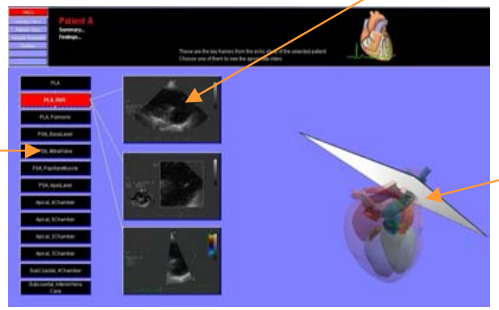
DEVL Medical Echo Library Interfaces (demo)

Disease Taxonomy Interface



View Browsing Interface

Table of Contents showing list of views



Representative frames of modes under selected view

3D model showing transducer angle

3D Heart Model courtesy of New York University School of Medicine