Emerging Multimedia Understanding Technologies

MMSP 2005 Tutorial
Oct. 30, 2005

Ching-Yung Lin[1] and Belle L. Tseng[2]

[1]: Exploratory Stream Processing Systems , IBM T. J. Watson Research Center, Hawthorne, NY 10532 USA

[2]: Adaptive Information and Knowledge Organization, NEC Laboratories America, Cupertino, CA 95014 USA

Outline

- Part I: Techniques for Semantic Video Understanding
- Part II: Standards and State-of-the-Art Systems
- Part III: Emerging Technologies and Applications

Part I: Techniques for Semantic Video Understanding

- Overview and History (8 mins)
- *Frameworks for Learning Generic Visual Concepts (10 mins)*
- *Audio-Visual Feature Extraction and Selection (10 mins)*
- *Machine Learning and Statistical Discriminant Technologies (20 mins)*
- *Multimodality Audio/Visual/Text Fusion (10 mins)*
- Summary (2 mins)

Part II: Standards and State-of-the-Art Systems

- Overview of An Application Scenario: Multimedia Semantic Retrieval Framework (3 mins)
- *MPEG-7 Overview (10 mins)*
- *Systems for Concept Modeling (15 mins)*
- *TREC Video Concept Retrieval and Detection Benchmarking (20 mins)*
- *MPEG-21 and Video Personalization and Summarization System (20 mins)*
- Demo (10 mins)
- Summary (2 mins)

Part III: Emerging Technologies

- Overview of Emerging Technologies (5 mins)
- *Part-Based Object Recognition (15 mins)*
- *Unsupervised Pattern Discovery (15 mins)*
- *Imperfect Learning and Cross-modality Autonomous Learning (15 mins)*
- *Distributed Video Semantic Filtering and Routing (15 mins)*
- Summary and Open Discussion (10 mins)

Speaker Contact Information

- Ching-Yung Lin

  IBM T. J. Watson Research Center

  (also affiliated with Columbia Univ. and Univ. of Washington)

  chingyung@us.ibm.com

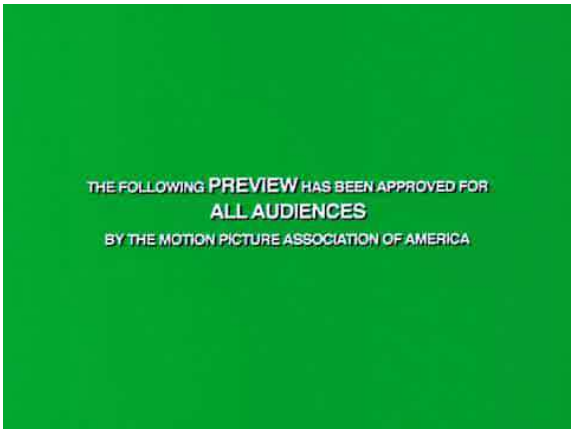
  http://www.research.ibm.com/people/c/cylin
- Belle L. Tseng

  NEC Laboratories America

  belle@sv.nec-labs.com

Part I: Techniques for Semantic Video Understanding

Multimedia Semantic Concept Detection



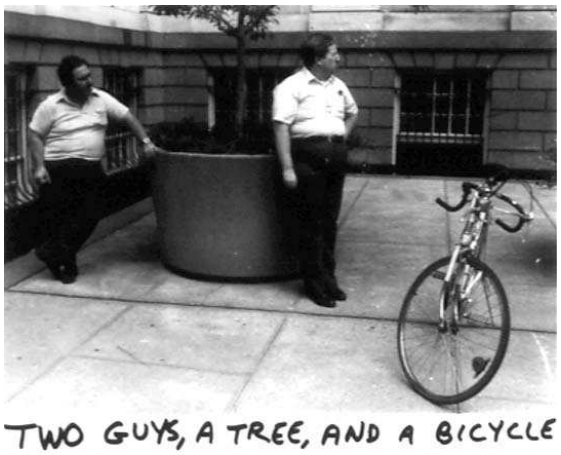
A picture is worth 1000 words !!??

- Fact 1: A film frame = 480 x 260 pixels = 374,400 bytes = 2,995,200 bits = 599,040 alphabets = 124,025 English words
- Fact 2: A 2 hour 10 mins movie = 187,200 frames = 70,087,680,000 bytes = 23,217,480,000 English words = 7,067 years of words a person can talk
- Fact 3: 128 hours of video = 13,648,457 frames = 31,131,584,804,571 bytes = storage space of 1556 20-GB hard disks at my laptop

A picture's worth one thousand words…

A lovely couple hand-in-hand walking together!!

A picture is worth seven words



By Prof. Elliott, Dept. of Communications, Cal State Univ. Fullerton

Words worth one thousand pictures

truth love nature time fate life

education fact knowledge change death eternity

right wrong good bad spirit mind past future

feeling memory intention chance wonder

pain justice courage nobility envy

gratitude respect beauty

Internet

By Prof. Gerald Grow, Florida A & M University

Words that picture can't say:

Confucius says:

"When things are investigated, then true knowledge is achieved;
when true knowledge is achieved, then the will becomes sincere;
when the will becomes sincere, then the heart sees correctly;
when the heart sees correctly, then the personal life is cultivated;
when the personal life is cultivated, then the family life is regulated;
when the family life is regulated, then the national life is orderly;
and when the national life is orderly, then there is peace in this world."

From *Li-Ji* (Record of Rites)

格物 致知 誠意 正心 修身 齊家 治國 平天下

By Prof. Elliott, Dept. of Communications, Cal State Univ. Fullerton

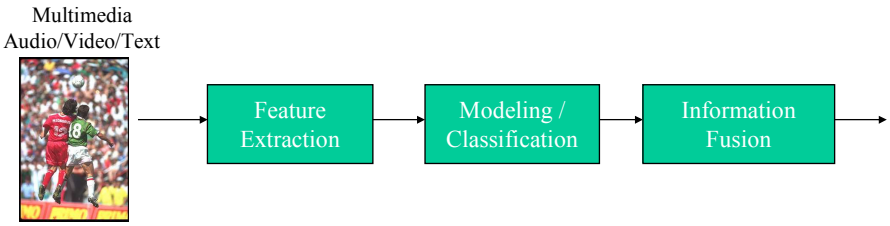
A picture's worth one thousand words…

Therefore, sometimes it can be used for implication!!

A picture's worth one thousand words…

Which One Thousand??

I.1 Framework for Generic Multimedia Concept Detection



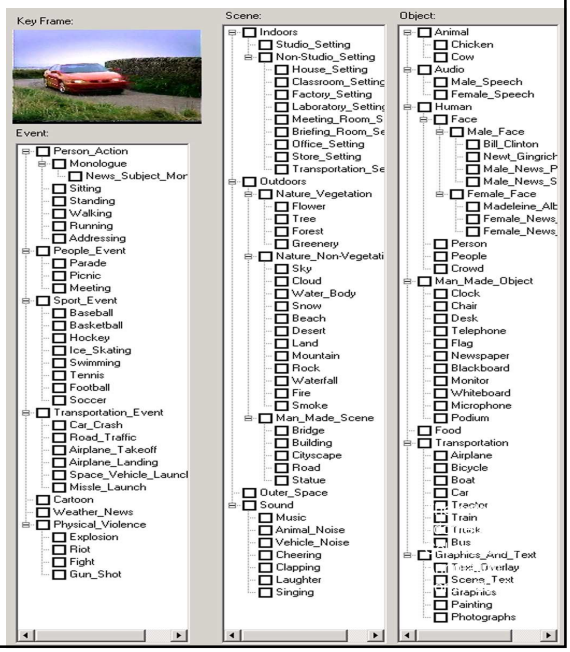
Multimedia Semantic Concept Detection & Mining

– which one thousand?

- Multimedia Concept Detection:
  - Objects:
    - Visual Objects:, Tree, Person, Hands, …
    - Audio Objects: Music, Speech, Sound, …
  - Scenes:
    - Background: Building, Outdoors, Sky
  - Relationships:
    - The (time, spatial) relationships between objects & scenes
  - Activities:
    - Holding Hand in Hand, Looking for Stars

Ontology for Multimedia Concept Description

- Event
- Scene
- Object



Challenges

- Focuses of previous Computer Vision/ Image Processing / Pattern Recognition modeling techniques:
  - Face, People
  - Optical Character Recognition (OCR)
  - Car, or specific objects
  - Specific human actions: walking, running

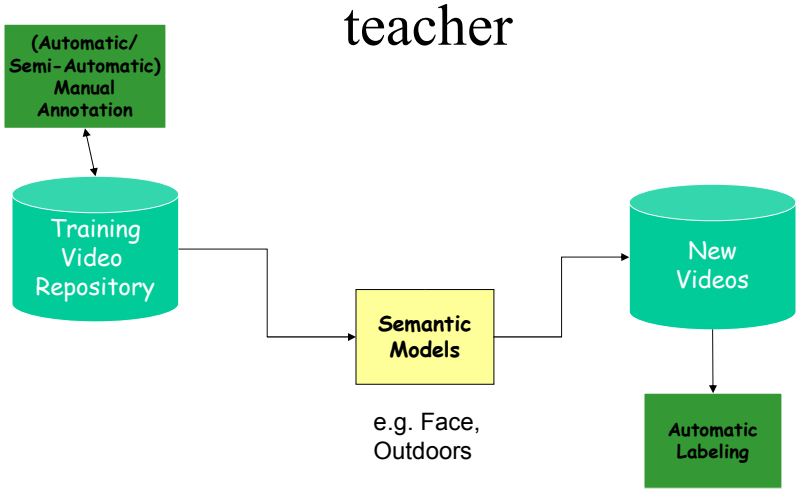
=> Most of previous researches are conducted in controlled environments with known camera settings and/or scenes

Challenge I: multimedia understanding data usually are short shots with unknown settings

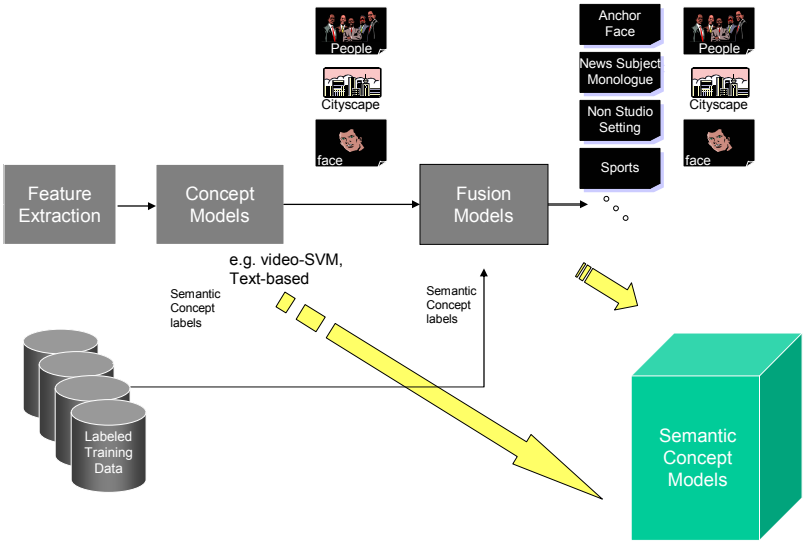
Challenge II: Require large amount of concept models

Challenge III: May need to combine both audio and visual information

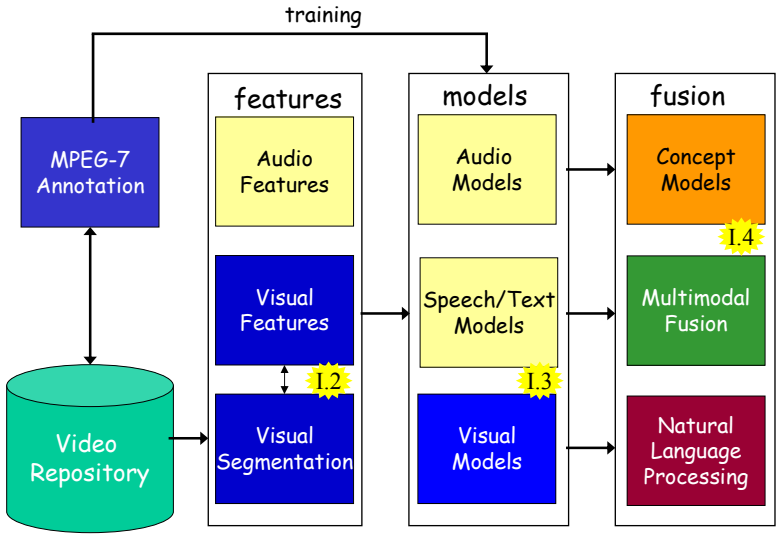
Learning video semantics with a teacher



Training Phase – From Pixels to Semantics

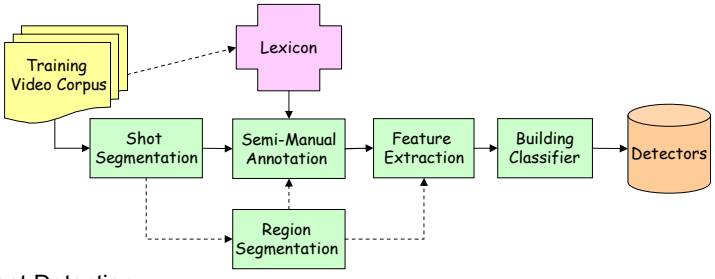


Multimedia Semantic Modeling Framework

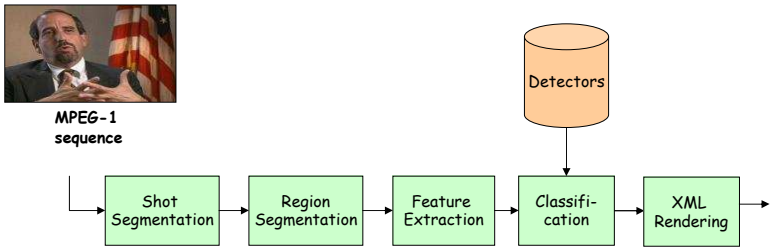


Framework to build generic concept detectors

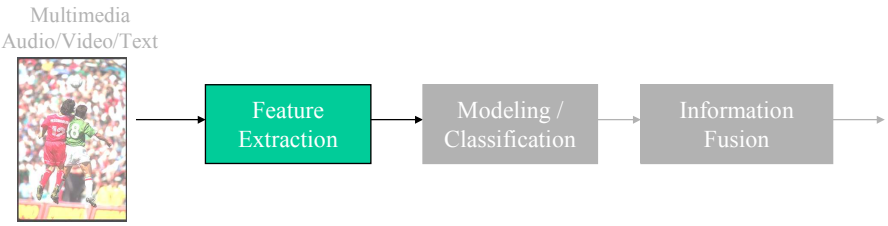
Model Training:



Concept Detection:



I.2 Extract Low-Level Features

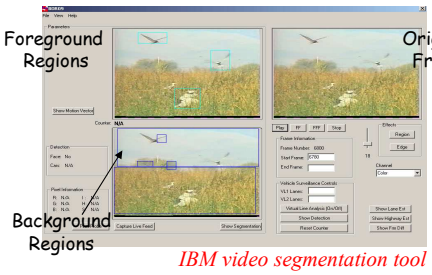


Visual Segmentation and Feature Extraction

I.2

Visual Segmentation

- Shot Segmentation
  - Color visual features
  - Time-scale differencing
- Region Segmentation
  - Object segmentation
  - Background segmentation (5 regions/shot)
  - For 10 primary detectors:
    - Use region segmentation: face, people, landscape, cityscape, monologue
    - Use global features: Outdoors, Indoors



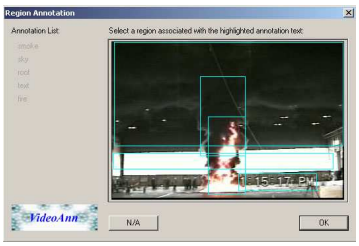
IBM video segmentation tool

Low-Level Visual Features

- Color:
  - Color histograms (YCbCr histograms 8x3x3, RGB 512 dim)
  - Auto-Correlograms (YCbCr, 8x3x3 dim)
  - Moments (7 dim)
- Structure and Shape:
  - Edge orientation histogram (32 dim)
  - Dudani Moment Invariants (6 dim)
  - Normalized width and height of bounding box (2 dim)
- Texture:
  - Co-occurrence texture (48 dim)
  - Coarseness, Contrast, Directionality (3 dim)
  - Wavelet Texture (12 dim)
- Motion:
  - Motion vector histogram (6 dim)

Visual Segmentation

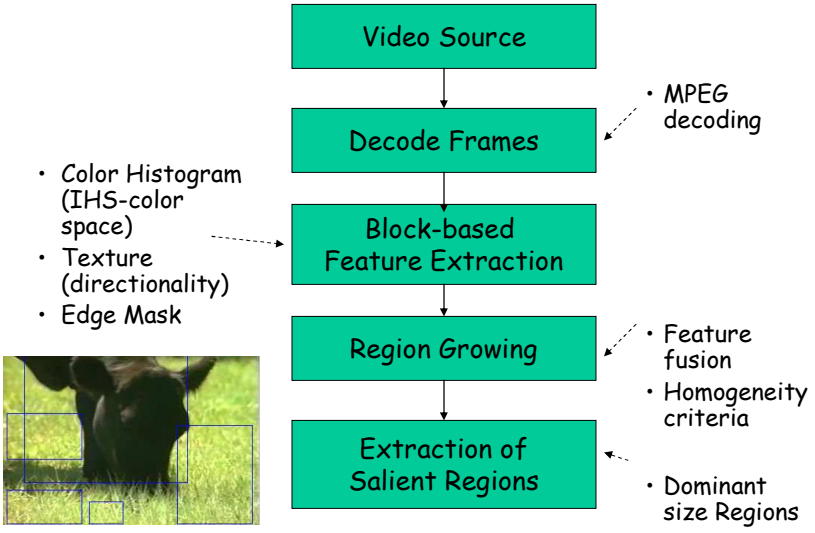
- Visual Segmentation is the No. 1 under- achievement research topic.
  (IEEE SP society 50 years review, 1998)
- **Objective**: Which pixels belong to which objects?
  - Objects vs. Regions
- Segmentation of Unconstrained Video *vs*. Computer Vision:
  - Difficulty for background learning (domain training)
  - No explicit camera information
  - Composite videos (text, graphics,..)
  + MPEG motion vectors (noisy)
- Prior Art:
  - Zhong *et. al., "VideoQ"*
  - Meyer *et. al., " Region Based Tracking using Affine Motion Models"*



Manual Annotated Regions

➔ Challenge: Segment regions corresponding to semantic concepts in short clips

Region of Interest Segmentation



Example: Fusion of Multiple Features

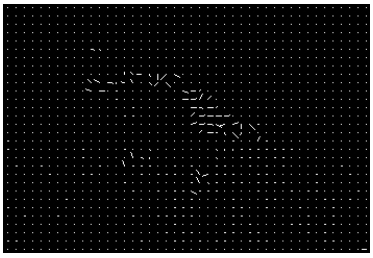
Original Video Frame

Color and Edge

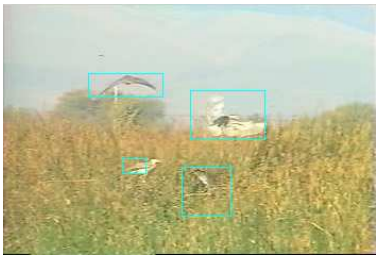
Color, Edge, and Texture

Segmentation Results

Example: Foreground Object Segmentation

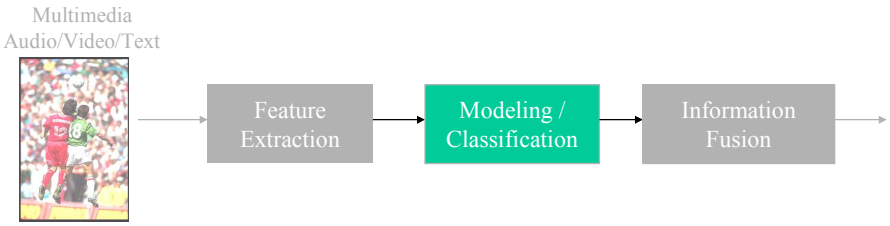


Motion Vectors



Segmentation Results

I.3 Statistical Discriminant Supervised Learning Models



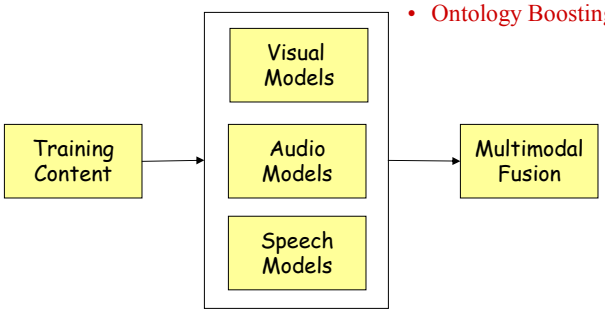
Statistical Learning Techniques

Modeling I.3

- Probabilistic Density Modeling: Gaussian Mixture Models (GMM), Hidden Markov Models (HMM)
- Discriminant Learning: Support Vector Machines (SVM), Neural Networks

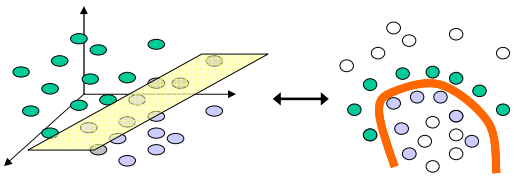
Fusion I.4

- Probabilistic Model: Naïve Bayes
- Graphical Model: Bayesian Networks
- SVM Fusion
- Ensemble Fusion
- Ontology Boosting

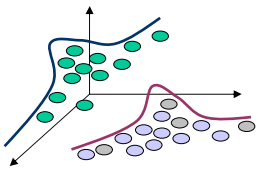


Statistical Modeling Approaches

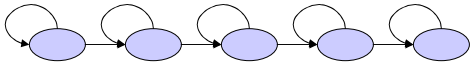
SVM



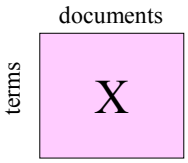
GMM



HMM



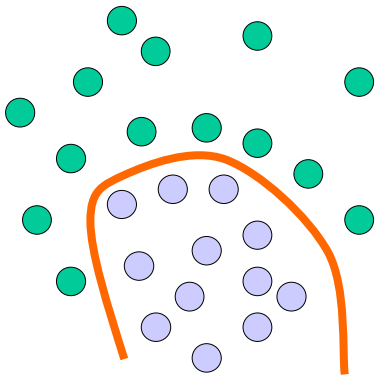
LSA



Discriminant Modeling

Support Vector Machines

- **Objective**: How to separate observations of distinct classes
- **Solution**: Project to higher dimensionality to determine linear separability



+ve

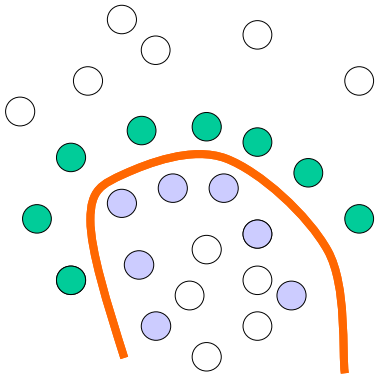
-ve

Slide Source: M. Naphade

Discriminant Modeling

Support Vector Machines

- **Objective**: How to separate observations of distinct classes
- **Solution**: Project to higher dimensionality to determine linear separability



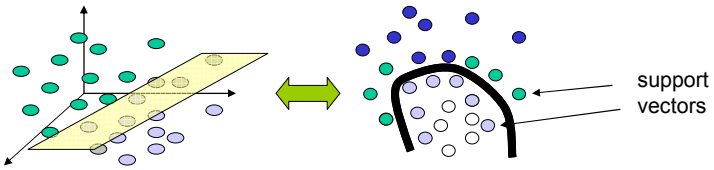
+ support vectors

- support vectors

useless examples

Slide Source: M. Naphade

SVM Classifiers



- Support Vector Machine
  - Largest margin hyperplane in the projected feature space
  - With good kernel choices, all operations can be done in low-dimensional input feature space

- SVM Classifier:

$$f(x) = \sum_{i=1}^{S} a_i \cdot k(x, x_i) + b$$

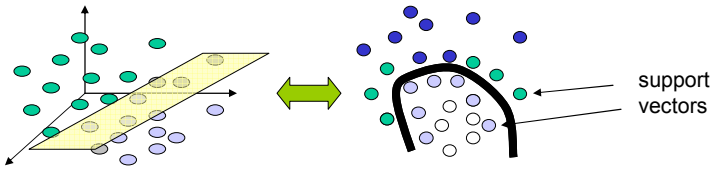
where *S* is the number of support vectors, k(.,.) is a kernel function. E.g., $k(x, x_i) = e^{\frac{\|x - x_i\|}{r}}$

- Complexity *c*: operation (multiplication, addition) required for classification

$$c \propto S \cdot D$$

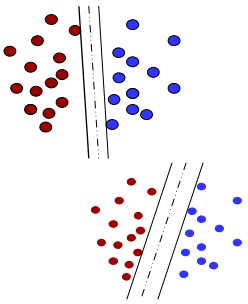
where *D* is the dimensionality of the feature vector

Support Vector Machine



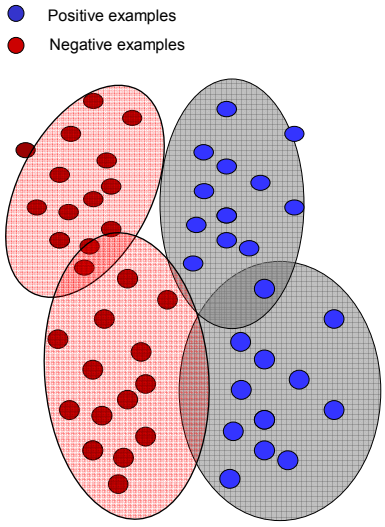
$$f(x) = \sum_{i=0}^{N_s} \alpha_i y_i K(x, s_i) + b$$

$$K(x_i, x_j) = e^{-\|x_i - x_j\|^2 / 2\sigma^2}$$



- Support Vector Machine
  - Largest margin hyperplane in the projected feature space
  - With good kernel choices, all operations can be done in low-dimensional input feature space
  - We use Radial Basis Functions as our kernels
  - Sequential Minimal Optimization = currently, fastest known algorithm for finding hyperplane

Gaussian Mixture Models



- The probability of examples belonging a cluster (e.g., positive set) is modeled as a weighted summation of multiple Gaussians distributed in the feature space

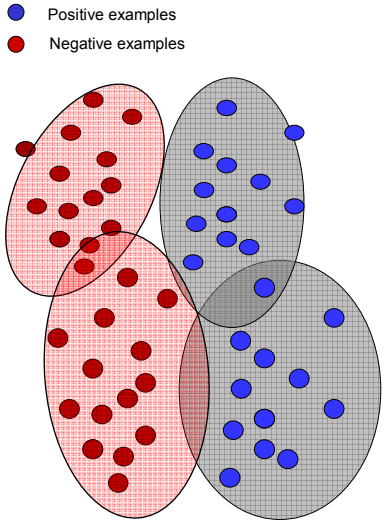
- Given a vector **x** at N-dimensional feature space, the probability that it belongs to a model C is:

$$p(\mathbf{x} \mid C) = \sum_{j=1}^{M} p(\mathbf{x} \mid c_j) p(c_j)$$

where M is the number of Gaussian components. Each component is an N-dimensional Gaussian function which is determined by its mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

$$p(\mathbf{x} \mid c_j) = \frac{1}{2\pi \left|\boldsymbol{\Sigma}_j\right|^{1/2}} \exp^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1} (\mathbf{x}-\boldsymbol{\mu}_j)}$$

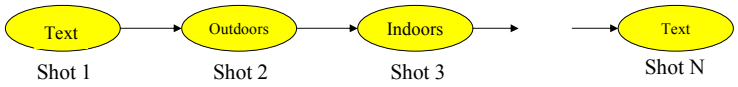
Gaussian Mixture Models



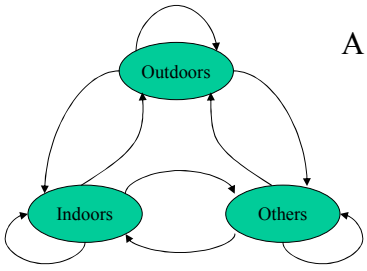
- The probability of examples belonging a cluster (e.g., positive set) is modeled as a weighted summation of multiple Gaussians distributed in the feature space.
- Pros:
  - The size of model parameters can be very small. In a lot of applications, we may assume the Gaussian components at different dimensions are independent. Thus (mean, std) at each feature dim.
- Cons:
  - Usually Estimation-Maximization (EM) Models are used to estimate models. The results may vary depending on the initial condition

Modeling Video Context with Markov Model

- Video Structure



- **Context Modeling Method I : Markov Models**
  - State = Concept, not Union of States = Concept
  - Use Training set to estimate a Transition Probability Matrix
  - Use Fusion methods to combine the results of individual shot-based classifiers and the prediction from the previous shot(s).



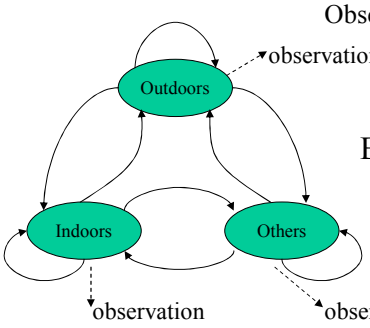
$$A = \begin{pmatrix} P(S_{1,t}|S_{1,t-1}) & P(S_{2,t}|S_{1,t-1}) & P(S_{3,t}|S_{1,t-1}) \\ P(S_{1,t}|S_{2,t-1}) & P(S_{3,t}|S_{2,t-1}) & P(S_{3,t}|S_{2,t-1}) \\ P(S_{1,t}|S_{3,t-1}) & P(S_{2,t}|S_{3,t-1}) & P(S_{3,t}|S_{3,t-1}) \end{pmatrix}$$

Example:

$$A_{FR} = \begin{pmatrix} 0.7684 & 0.0511 & 0.1805 \\ 0.1123 & 0.6763 & 0.2114 \\ 0.2474 & 0.1211 & 0.6311 \end{pmatrix}$$

Modeling Video Context Method with HMM and SVM

- **Context Modeling Method II : Hidden Markov Models + SVM**
  - Use classification results as observations in each state.
  - Various selection of classifiers can be selected:
    - SVM & SVM fusion
    - GMM models
    - HMM models ➔ Hierarchical HMM.
  - User Vertabi algorithm to estimate the states of video sequences. The result will be binary.



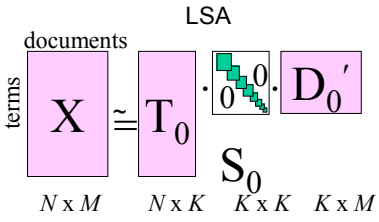
Observation Probability Matrix:

$$B = \begin{pmatrix} P(O_1|S_1) & P(O_2|S_1) & P(O_N|S_1) \\ P(O_1|S_2) & P(O_2|S_2) & P(O_N|S_2) \\ P(O_1|S_3) & P(O_2|S_3) & P(O_N|S_3) \end{pmatrix}$$

O = (O1, O2, .. ON) discrete values

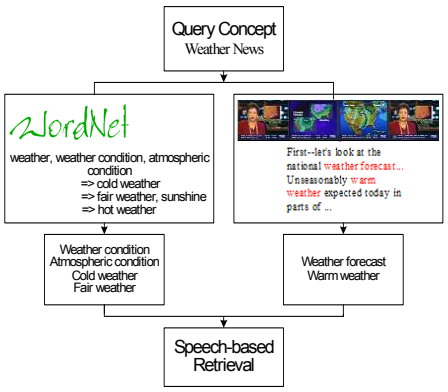
Content Analysis: *Latent Semantic Analysis*

- Latent Semantic Analysis (LSA) [Landauer, Dumais 1997]
  - Capture the semantic concepts of documents by mapping words into the latent semantic space which captures the possible synonym and polysemy of words
  - Training based on different level of documents. Experiments show the synergy of the # of training documents and the psychological studies of students at 4th, 10th, and college level. Used as an alternative to TOEFL test.
  - Based on truncated SVD of document-term matrix: optimal least-square projection to reduce dimensionality



Speech and Text-based Topic Detection

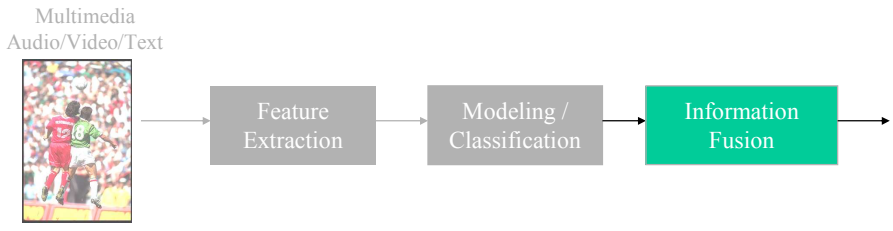
- Unsupervised learning from WordNet:



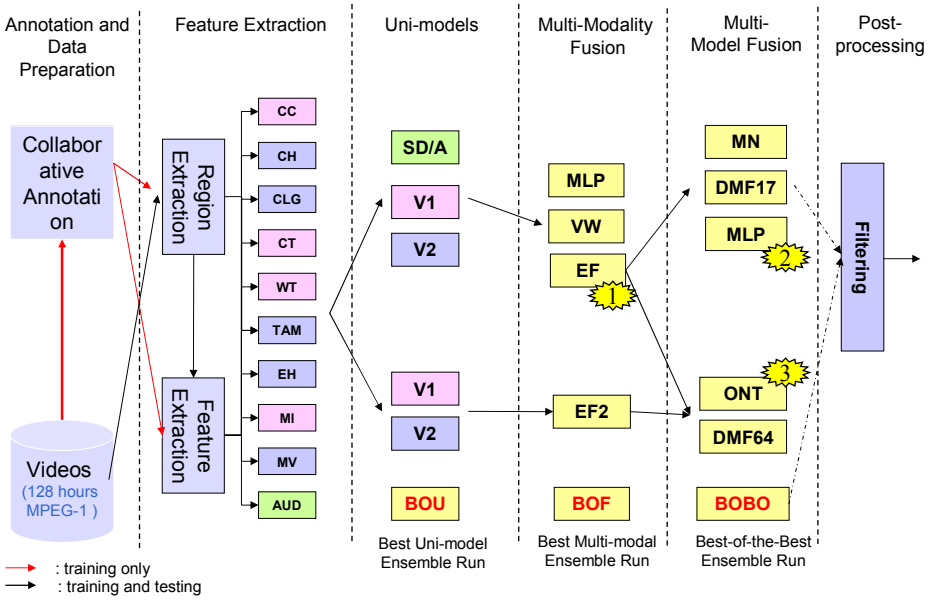
Airplane	Animal	Building	Weather news
airplane	animal	building	Weather
aeroplane	beast	edifice	atmospheric
plane	brute	walk-up	cold weather
airline	critter	butchery	freeze
airbus	darter	apart build	frost
twin-aisle	peeper	tenement	sunshine
airplane	creature	architecture	temper
amphibian	Fauna	call center	scorcher
biplane	microorganism	sanctuary	sultry
passenger	poikilotherm	Bathhouse	thaw
fighter			warm
aircraft			downfall
			hail
			rain

Example of Topic – Related Keywords

I.4 Information Fusion



An Example Video Concept Detection Systems

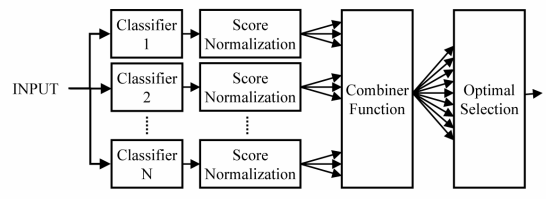


Multi-Modality/ Multi-Concept Fusion Methods

Ensemble Fusion:

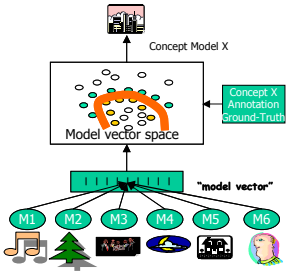
Normalization: rank, Gaussian, linear.

- Combination: average, product, min, max, and grid-weighted average.

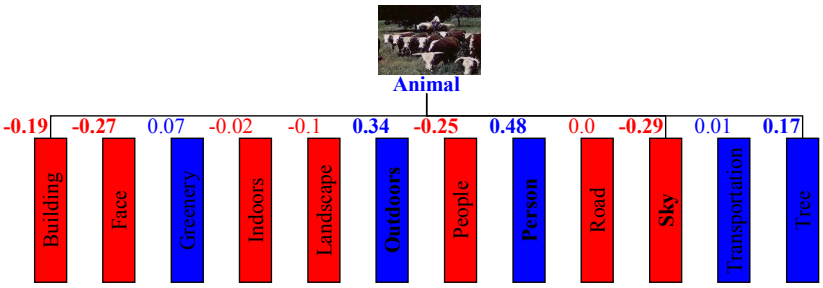


Re-Classification: a.k.a. Discriminative Multiple Fusion (DMF)

- Using SVM to re-classify the output results of Classifier 1-N.
- No normalization required.
- Since the validation set was used for parameter tuning of the models, a k-fold cross validation experiment was set-up for tuning parameters of the re-classifier.



Semantic Modeling Through Multi-Layer Perceptron Neural Networks [Natsev et. al. 2004]

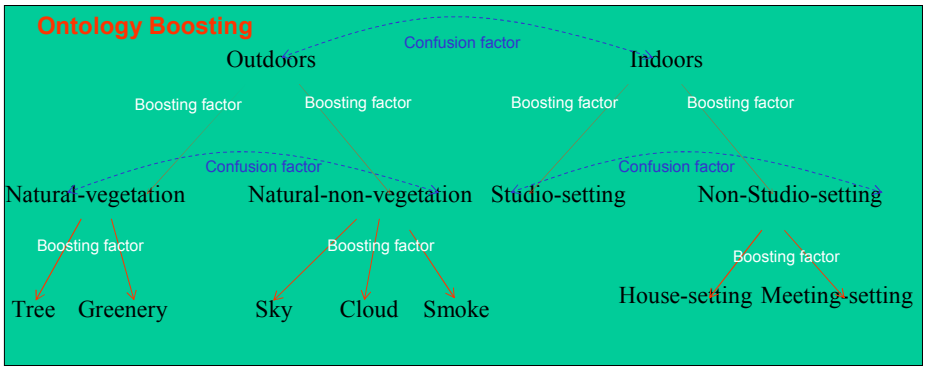


- **Problem:** Given a (small) set of related concept exemplars, learn concept representation
- **Approach:** Learn and exploit semantic correlations and class co-dependencies
  - Build (robust) classifiers for set of basis concepts (e.g., SVM models)
  - Model (rare) concepts in terms of known (frequent) concepts, or anchors
  - Learn weights of separating hyper-plane through regression:
    - Optimal linear regression (through Least Squares fit)
    - Non-linear MLP regression (through Multi-Layer Perceptron neural networks)
    - Can be used to boost performance of basis models or for building additional models

Courtesy By Apostol Natsev

Multimodal Fusion using Ontology Boosting

- Basic Idea
  - Concept hierarchy is created based on semantics ontology
  - Classifiers influence each other in this ontology structure
  - Try best to utilize information from reliable classifiers
- Influence Within Ontology Structure
  - Boosting factor
  - Confusion factor



Summary of Part I

- Finding semantics in video is a challenge issue.
- Generic concept detectors can be developed based on supervised machine learning methods.
- Several statistical classification models – such as Support Vector Machines, Gaussian Mixture Models, Hidden Markov Models, Singular Vector Decomposition are commonly used for multimedia concept detection.
- Information fusion methods can be applied to uni-modality models and multi-modality models.
- Fusion of classifiers from different modality or different features usually improves the classification accuracy.

.

Part I References

- B. Adams, G. Iyengar, Ching-Yung Lin, Milind Naphade, Chalapathy Neti, Herriet Nock and John R. Smith, "**Semantic Indexing of Multimedia Content Using Visual, Audio and Text Cues**," *EURASIP Journal on Applied Signal Processing*, 2003.

- A. Amir, W, Hsu, G. Iyengar, Ching-Yung Lin, M. Naphade, A. Natsev, C. Neti, H. J. Nock, J. R. Smith, Belle L. Tseng, Y. Wu, D. Zhang, "**IBM Research TRECVID-2003 System**," *Proc. NIST Text Retrieval Conf. (TREC),* Gaithersburg, MD, November, 2003.

- Xiaodan Song, Ching-Yung Lin and Ming-Ting Sun, "**Speech-based Video Retrieval Using WordNet**," *IEEE Intl. Conf. on Multimedia & Expo*, Amsterdam, Netherlands, July 2005.

- Belle L. Tseng, Ching-Yung Lin, Milind Npahade, Apostol Natsev and John R. Smith, "**Normalized Classifier Fusion for Semantic Visual Concept Detection**," *IEEE Intl. Conf. on Image Processing*, Barcelona, September 2003.

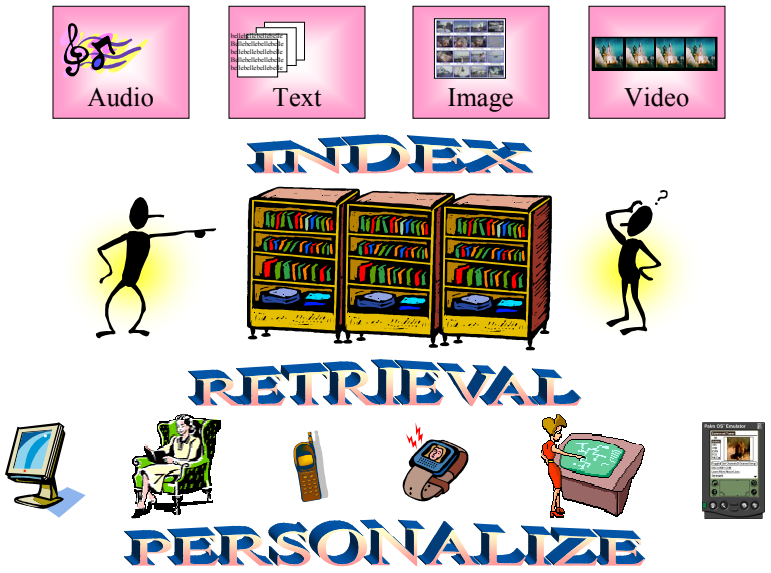
- Y. Wu, B. L. Tseng and J. R. Smith, "**Ontology-based Multi-Classification Learning for Video Concept Detection"** *IEEE International Conference on Multimedia and Expo (ICME),* June 2004.

Part II: Standards and State-of-the Art Systems

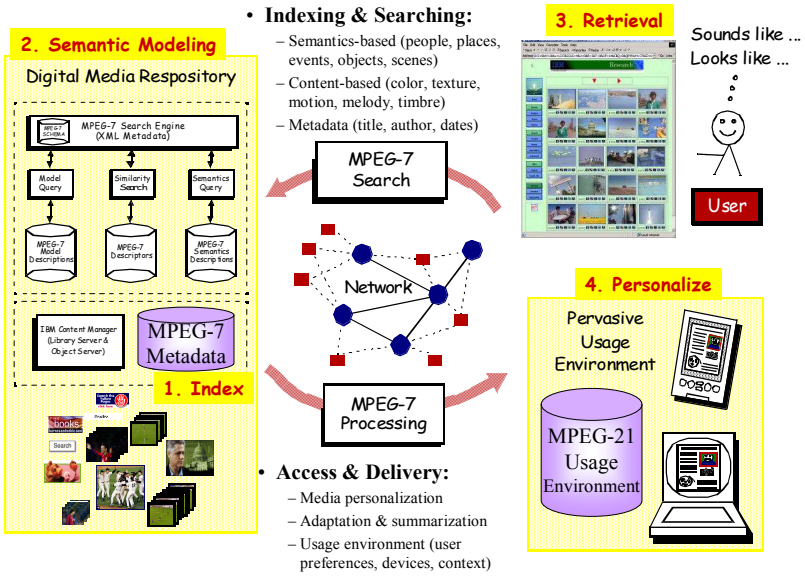
Part II: Standards and State-of-the-Art Systems

- Overview of An Application Scenario: Multimedia Semantic Retrieval Framework (3 mins)
- *MPEG-7 Overview (10 mins)*
- *Systems for Concept Modeling (15 mins)*
- *TREC Video Concept Retrieval and Detection Benchmarking (20 mins)*
- *MPEG-21 and Video Personalization and Summarization System (20 mins)*
- Demo (10 mins)
- Summary (2 mins)

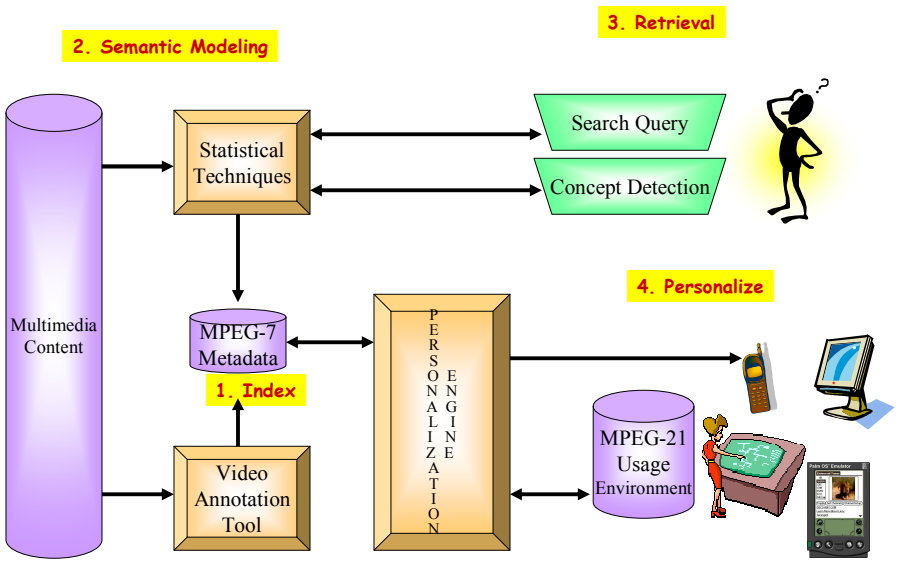
Motivation

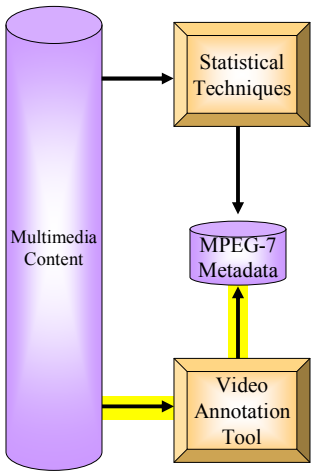


Multimedia Semantic Framework



Multimedia Semantic Framework

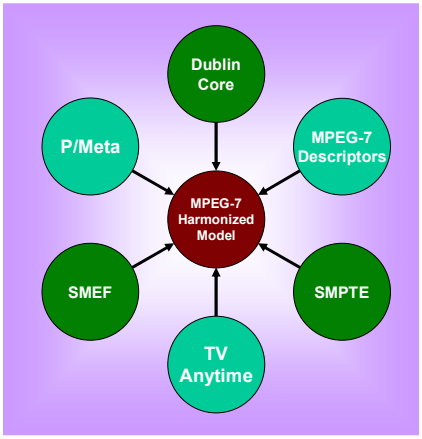




II.1. MPEG-7 Multimedia Description Scheme

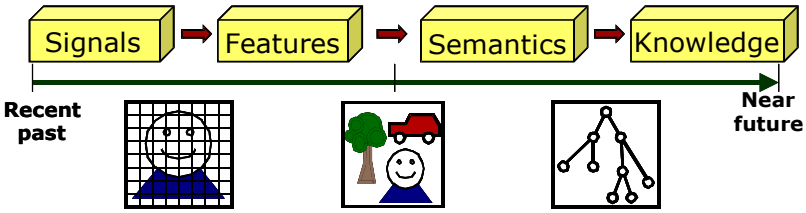
Multimedia Metadata Standards

- MPEG-7: Moving Picture Experts Group
  - Infrastructure standard for Multimedia Metadata
  - Supports interpretation of the information's meaning
  - Supports broad range of applications
- SMEF – Standard Media Exchange Framework:
  - BBC developed data models for information involved in the Production, Development, Use, and Management of media assets
- P/Meta – EBU P/Meta Project:
  - Exchange of program content between high-level business functions of EBU members: Production, Delivery/Broadcast, & Archive
- SMPTE – Metadata dictionary & MXF:
  - Addresses Program Interchange independent of format
- Dublin Core Metadata Initiative:
  - Interoperable online metadata standards supporting broad range of purposes and business models.
- TV-Anytime – TV-Anytime Metadata:
  - Attractors/descriptors used e.g. in Electronic Program Guides (EPG), or in Web pages to describe content.
- Indecs – Indecs Metadata Framework
  - An international initiative of rights owners creating metadata standards for e-commerce.

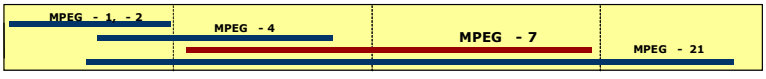


- MPEG-7 Harmonized Model:
  - Harmonized elements with other standards and existing practices
  - Extensible framework
  - Registration authority for classification schemes, controlled terms, ontologies

Towards Knowledge Management and Transaction Enrichment for Digital Media

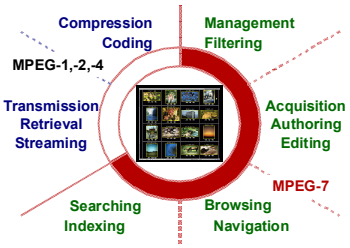


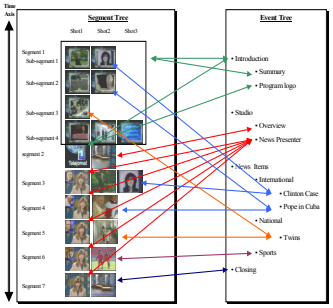
Applications			
MPEG-1,-2,-4 Video storage Broadband Streaming video delivery	**MPEG-4,-7** Content-based retrieval Multimedia filtering Content adaptation	**MPEG-7** Semantic-based retrieval and filtering Enterprise content mgmt.	**MPEG-21** E-commerce of Electronic content Digital items
Problems and Innovations			
Compression Coding Communications	Similarity searching Object- and feature-based coding	Modeling and classification Personalization and summarization	Media mining Decision support IPMP (rights)



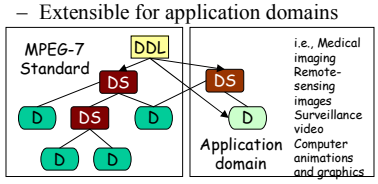
MPEG-7 Overview

XML Metadata for Multimedia Content Description

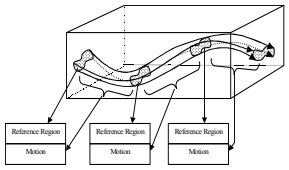




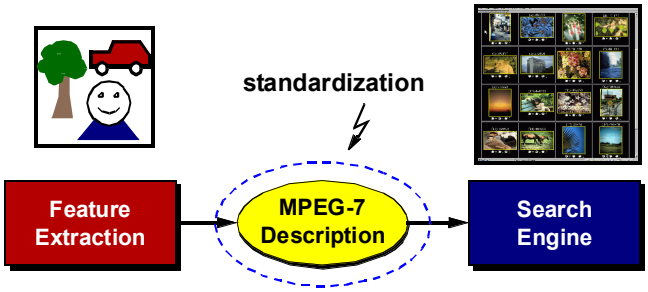
- MPEG-7 Normative elements:
  - Descriptors and Description Schemes
  - DDL for defining Description Schemes
  - Extensible for application domains



- Rich, highly granular descriptions:
  - Video segments, moving regions, shots, frames, …
  - Audio-visual features: color, texture, shape, …
  - Semantics: people, events, objects, scenes, …



MPEG-7 Standard Scope



Industry Competition

Feature Extraction:	MPEG-7 Scope:	Search Engine:
Content analysis (D, DS)	Description Schemes (DSs)	Searching & filtering
Feature extraction (D, DS)	Descriptors (Ds)	Classification
Annotation tools (DS)	Language (DDL)	Complex querying
Authoring (DS)	Coding Schemes (CS)	Indexing
	MPEG-7 Concepts	Personalization

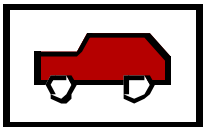
MPEG-7 Multimedia Content Description

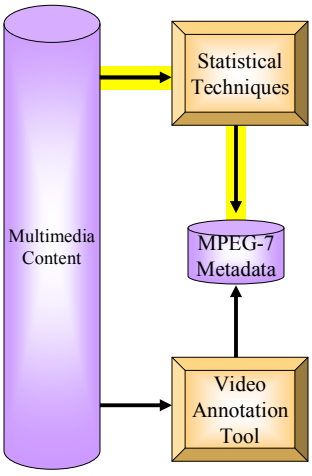
- **What is MPEG-7 about?**
  - Specification of a "Multimedia Content Description Interface"
  - Developed by International Standards Organization (ISO) and International Electrotechnical Commission (IEC)
  - Standardized representation of multimedia metadata in XML (XML Schema Language)
  - Describes audio-visual content at a number of levels (features, structure, semantics, models, collections, immutable metadata)
  - MPEG-7 completed in Oct. 2001
- **MPEG-7 Key Points:**
  - MPEG-7 is not a video coding standard
  - MPEG-7 is an XML metadata standard for describing multimedia content
  - MPEG-7 also provides a binary compression system for MPEG-7 descriptions (called BiM)
  - MPEG-7 descriptions can be embedded in the video streams or stored separately as documents or in databases

MPEG-7 MDS: Free Text Annotation Example

The following example gives an MPEG-7 description of a car that is depicted in an image:

```
<Mpeg7>
  <Description xsi:type="SemanticDescriptionType">
    <Semantics>
      <Label>
        <Name> Car </Name>
      </Label>
      <Definition>
        <FreeTextAnnotation>
          Four wheel motorized vehicle
        </FreeTextAnnotation>
      </Definition>
      <MediaOccurrence>
        <MediaLocator>
          <MediaUri> image.jpg </MediaUri>
        </MediaLocator>
      </MediaOccurrence>
    </Semantics>
  </Description>
</Mpeg7>
```

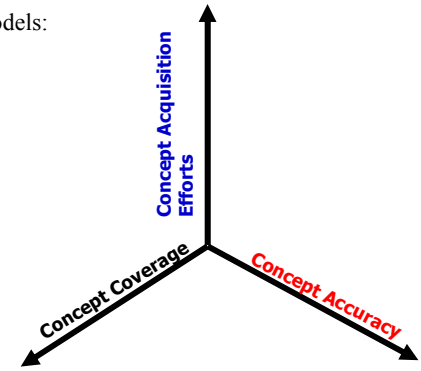




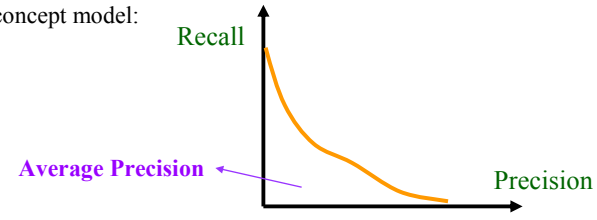
II.2. Multimedia Semantic Modeling

Performance Measurement of Concept Modeling

- From the viewpoints of a set of concept models:

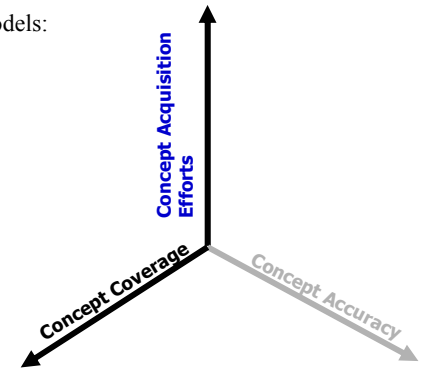


- From the viewpoints of a concept model:

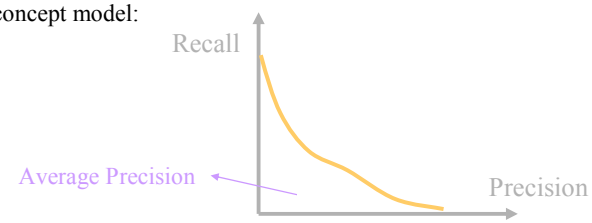


Performance Measurement of Concept Modeling

- From the viewpoints of a set of concept models:



- From the viewpoints of a concept model:



Ontology and Concept Coverage

- Representation of someone's or some group's perception of a real-world domain
- Validation:
  - Accuracy
  - Completeness
  - Conflict-free
  - No redundancy
- Theories of Ontology:
  - Chen (1976):
    - Entity-Relation Modeling Grammer:
      - Entity (and entity set)
      - Relationship
      - Attributes
  - Bunge (1977):
    - Things, properties of things, state of things, laws, events in things, or couplings

Generating Open Semantic Meta-Data Resources for Multimedia (http://mp7.watson.ibm.com/projects/VideoCAforum.html)

- Video Collaborative Annotation Forum 2003:

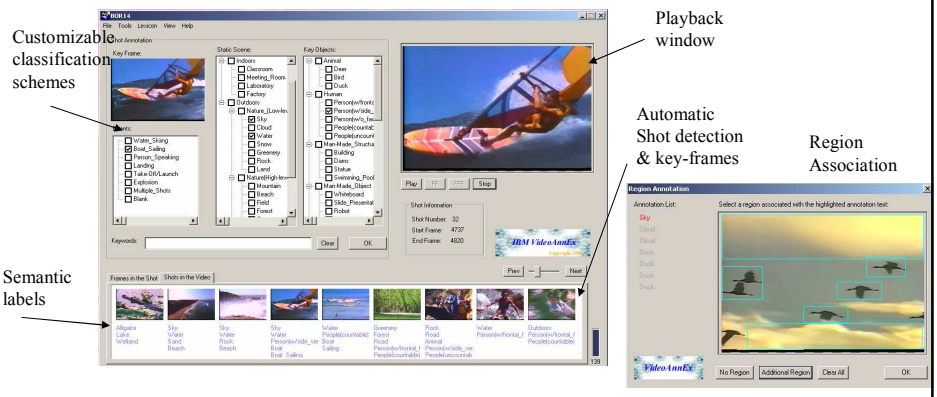
Year	Data	# of Annotators	Lexicon	Source	Example
2003	62.2 hrs	111 @ 23 groups -- Accenture, CMU, CLIPS, Columbia U., CWI, Dublin, EPFL, EURECOM, Fudan U., IBM, Intel, KDDI, Tsing-Hua U., U. Singapore, TUT, UCF, U. Chile, UniDE, U. Geneva, U. Glasgow, U. Mass, UNC, U. Oulu	133 – audio & visual: *35 A&V events, 38 visual scenes, 11 sounds, 49 visual objects*	CNN & ABC news, (1998) C-SPAN (1998, 2000)	

- Initial Steps (03/03 – 05/03): Initialized discussions, proposal, testing environments, groups signed-in, discussed the 1st draft of lexicon, user-feedback for the annotation system.
- Phase I (05/03 – 06/03): Assigned 37 sample videos to groups, client tool improvement, finalized the lexicon, groups get sample videos and set up annotation environment, checked the validity of annotation results.
- Phase II (06/03 – 07/03): Assigned 106 videos to groups. Completed annotating the TRECVID 2003 development set. Cleaned the annotated XMLs.
- User Study (10/03): questionaire – responses from 38 annotators @ 17 groups

➔ 433K ground-truth labels were annotated at 47,364 shots in 62.2 hr of videos

Semi-Automatic Labeling: *VideoAnnEx*

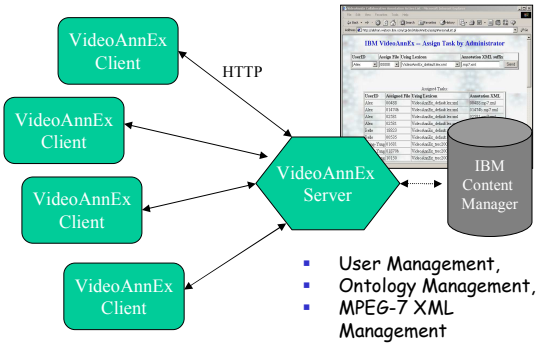
- IBM MPEG-7 Annotation Tool:
  - First Public MPEG-7 semantic annotation tool (Aug. 2002).
  - Creates MPEG-7 semantic description metadata
  - Supports customizable semantic classification schemes
  - Provides learning framework for propagating annotations
  - http://alphaworks.ibm.com/tech/videoannex



Semi-Supervised Manual Indexing and Supervised Learning: *IBM VideoAnnEx Collaborative Multimedia Annotation System*

- July 2001: Version 1.0
- July 2002: Version 1.4: IBM Alphaworks public software – first public available shot-based/user-oriented semi-automatic MPEG-7 Annotation Tool (~ 2000 downloads)
- April 2003: Version 2.0: Collaborative Multimedia Annotation Tool (current version 2.1.2)
- April – July 2003: Collaborative Multimedia Annotation Forum – 23 participating academic/research institutes

- Annotation Learning,
- Template,
- Multi-thread processing



- User Management,
- Ontology Management,
- MPEG-7 XML Management

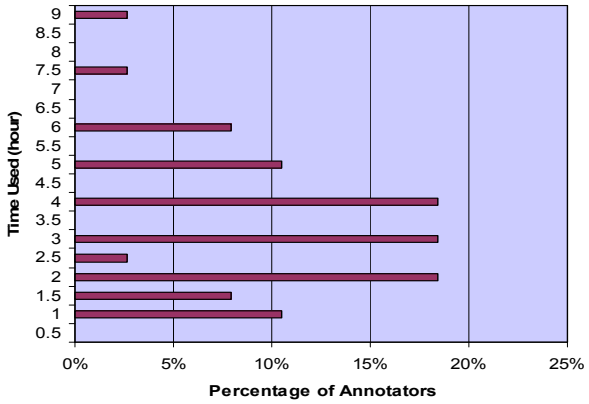
➔ Leading Tool on Multimedia Meta-Data Generation

Other Multimedia Annotation Tools

- Bargeron, et. al., "MRAS: Microsoft Research Annotation System" – a web-based system for annotating multimedia web content. Annotations include comments and audio in the distance learning scenerio.
  - Does not make use of lexicon.
  - Does not make use of shots.
  - Does not include personalized management system.
- Steves, et al, "SMAT: Synchronous Multimedia and Annotation Tool", NIST
  - Annotate images.
  - No granularity on videos.
  - No controlled-term items.
- Nack and Putz, "Semi-automated Annotation of Audio-Visual Media in News", GMD, Germany
  - Stand-alone application
  - Need to specify shots manually.
  - Does not include controlled-term items.
- RICOH – MovieTool MPEG-7 XML labeling Tool – Users need full understanding on MPEG-7, need to specify shots or other granularity.

Concept Acquisition Efforts (1 of 2)

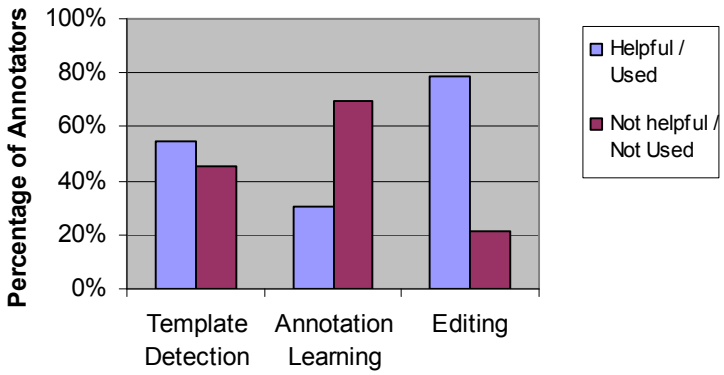
- Time used to annotate audio-visual concepts at the shot/region level on a 30-min video using VideoAnnEx V2.1



Time used for annotation?
➔Average: 3.38 hours, Median: 3 hours

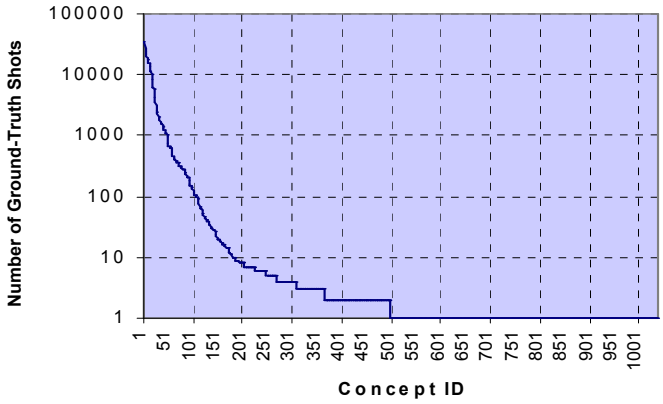
Concept Acquisition Efforts (2 of 2)

- Are these VideoAnnEx features useful in helping the annotation?



Concept Coverage – Annotation Results and User Subjective Evaluations (1 of 4)

- Distribution of the annotated concept labels



➔The number of example decays fast as the concept number increase

Concept Coverage (2 of 4) -- Detailed Distribution of Annotation Results (out of 46374 shots)

35571	Graphics_And_Text
35403	Human
26757	Audio
24224	Sound
22066	Outdoors
21737	Male_Speech
20532	Text_Overlay
18971	Face
18140	Person_Action
18120	Indoors
15546	Music
15293	Male_Face
14043	Non-Studio_Setting
11687	Female_Speech
11350	Monologue
10268	Man_Made_Object
10230	Female_Face
8278	Person
6289	Nature_Non-Vegetation
5991	Graphics
5910	People
5668	Man_Made_Scene
4173	Scene_Text
3450	Transportation
3172	Studio_Setting
2523	Nature_Vegetation
2428	Sport_Event
2400	Sky
2312	News_Subject_Monologue
2062	House_Setting
2035	Building

1124	Road
1074	Food
1069	Office_Setting
1024	Tree
1017	Water_Body
949	Commercial
761	Female_News_Subject
675	Meeting_Room_Setting
655	Basketball
653	Transportation_Event
633	Singing
618	Microphone
608	Desk
605	Cityscape
491	Vehicle_Noise
458	Chair
450	Snow
431	Photographs
416	Cartoon
414	Cloud
412	CNN_Text_Overlay
391	Briefing_Room_Setting
374	Flag
374	Running
364	Graphic/Text_Overlay
360	Mountain
351	Addressing
347	Meeting
326	Bill_Clinton
325	Forest

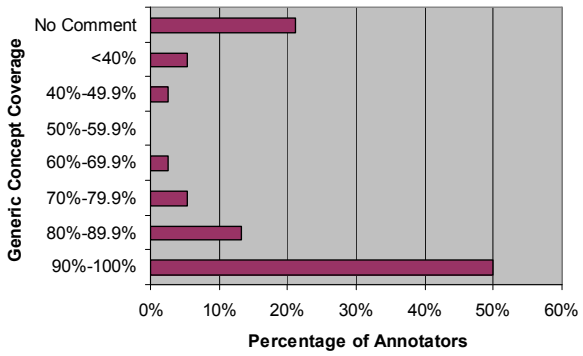
222	Boat
205	Transportation_Setting
203	Hockey
202	Classroom_Setting
200	Rock
176	Baseball
162	Road_Traffic
155	Laboratory_Setting
151	Painting
147	Animal_Noise
145	Clapping
143	Podium
129	Fire
123	Desert
114	Smoke
112	Bridge
106	Newspaper
104	Blank
104	Explosion
102	Hand
101	Horse
96	Laughter
95	Parade
93	Outer_Space
91	Ice_Skating
74	Bus
68	40th_Anniversary_Of_The_Freedom_Rides
67	Tractor
64	Football
63	Car_Crash

38	Gun_Shot
38	Blackboard
37	Interview
35	Madeleine_Albright
34	Chicken
32	Golf
31	Cut
31	Swimming
31	Powerpoint
30	Newt_Gingrich
29	Riot
29	Players
29	Book
28	Statue
27	Noise
26	Tennis
23	Sun
22	Map
21	Space_Vehicle_Launch
20	Picnic
20	Whiteboard
20	Fencing
19	Eating
19	Airplane_Landing
18	Human_Hand
18	Jacques_Nasser
18	Cinema_Setting
17	Gun
17	Skiing
17	Guitar
16	Camera

12	Congress
11	Chocolate
11	CNN_Splash_Page
11	Princess_Diana
11	Olympics
10	LCI
10	Stock_Exchange
10	Aliens
10	Car_Part
10	Bird
10	Porch_Setting
10	Sea
9	Bottle
9	Restaurant_Setting
9	Mop
9	House
9	Children
9	Rocket
8	Bra
8	Computer
8	News_Person_Monologue
8	Cigarette
8	Shoe
8	Lamp
8	Acupuncture
8	Aligator
8	Movie
8	Flowers
8	Wrecked_Car
8	NBA
7	Zoom_Out

Concept Coverage – Annotation Results and User Subjective Evaluations (3 of 4)

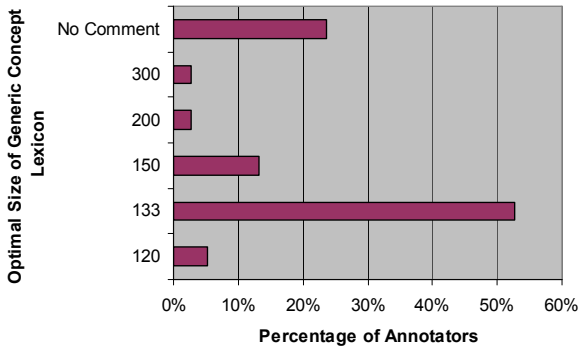
- Subjective Evaluation at the Coverage of Generic Concepts of the Lexicon:



❑ How many generic concepts in this news-domain video set?
➔ About 147 - 164

Concept Coverage – Annotation Results and User Subjective Evaluations (4 of 4)

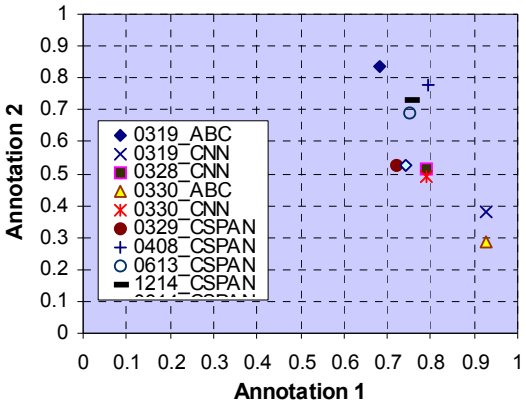
- Subjective practical bound of the size of lexicon:



❑ Major causes of practical limitation: Display window size, Human memory

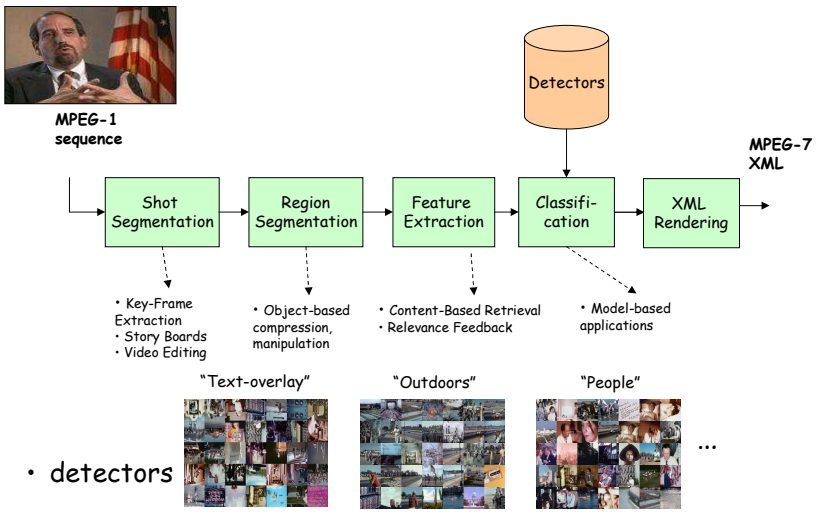
Concept Annotation Accuracy -- Human Factors in Subjective Annotations

- Annotation False Alarm: should be close to 0.
- Experiment of Annotation Miss: 10 development videos are annotated by two annotators.



Annotation Completeness?
➔ Average: 68%; Average of the better annotator: 78.7%

Automatic Labeling: *VideoAl*



• detectors

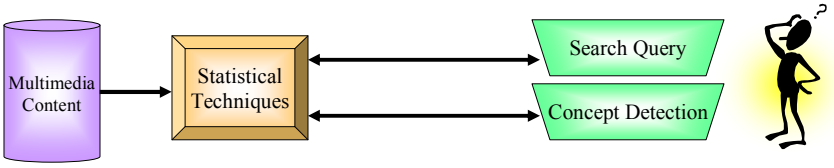
More than 100 Detectors for Automatic Labeling

- Sport_Event
- Transportation_Event
- Cartoon
- Weather_News
- Physical_Violence
- Indoors
- Outdoors
- Outer_Space
- Animal
- Human
- Man_Made_Object
- Food
- Transportation
- Graphics_And_Text
- Sitting
- Standing
- Walking
- Running
- Addressing
- Parade
- Picnic
- Meeting
- Baseball
- Basketball
- Hockey
- Ice_Skating
- Swimming
- Tennis
- Football
- Soccer

- Car_Crash
- Road_Traffic
- Airplane_Takeoff
- Airplane_Landing
- Space_Vehicle_Launch
- Missle_Launch
- Explosion
- Riot
- Fight
- Gun_Shot
- Studio_Setting
- Non-Studio_Setting
- Nature_Vegetation
- Nature_Non-Vegetation
- Man_Made_Scene
- Chicken
- Fire
- Smoke
- Bridge
- Male_Face
- Female_Face
- Bill_Clinton
- Newt_Gingrich
- Male_News_Person
- Male_News_Subject
- Madeleine_Albright
- Female_News_Person
- Female_News_Subject
- Cityscape

- Cow
- Dog
- Fish
- Horse
- Pig
- Face
- Person
- People
- Crowd
- Clock
- Chair
- Desk
- Telephone
- Flag
- Newspaper
- Blackboard
- Monitor
- Whiteboard
- Microphone
- Podium
- Airplane
- Bicycle
- Boat
- Car
- Tractor
- Train
- Truck
- Bus
- Building

- Text_Overlay
- Scene_Text
- Graphics
- Painting
- Photographs
- House_Setting
- Classroom_Setting
- Factory_Setting
- Laboratory_Setting
- Meeting_Room_Setting
- Briefing_Room_Setting
- Office_Setting
- Store_Setting
- Transportation_Setting
- Flower
- Tree
- Forest
- Greenery
- Cloud
- Sky
- Water_Body
- Snow
- Beach
- Desert
- Land
- Mountain
- Rock
- Waterfall
- Road



II.3. Retrieval and TREC Benchmarking

NIST TREC Video Benchmark

- Tasks: Shot Boundary Detection (2001 - 2005)
  Semantic Video Retrieval Query (2001 - 2005)
  Semantic Concept Detection (2002 - 2005)
  Story Boundary (2003 – 2004), Camera Motion (2005), Exploratory BBC(2005)
- Corpus: *2001* – 14 hours from NASA and BBC
  2002 – 74 hours from Internet Movie Archive
  2003, 2004 – 192 hours from CNN, ABC, etc.
  2005 – 170 hours from LBC (Arabic), CCTV, NTDTV (Chinese), CNN, NBC
- Video Retrieval Topic Examples:

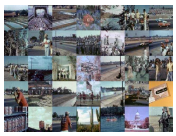
Topic 2. Scenes that depict a lunar vehicle traveling on the moon.

Topic 13. Speaker talking in front of the US flag.

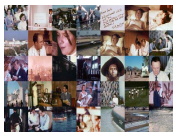
Topic 48. Examples of overhead zooming-in views of canyons in Western US.

- Semantic Concept Detector Examples:

"Text-overlay"



"Outdoors"



"People"

Participating TREC Video Benchmark

- Participation in NIST TREC Video Benchmarking helps everyone:
  - Common Large Video Corpus
  - Common Formats
  - Common Questions
  - Common Deadlines
- Much easier to compare the performance of different algorithms and systems.
- Allows different group to exchange assets – from data preparation to system modules.
- The number of participating groups grows from 12 in 2001 to 60+ in 2005.

TREC 2001 Video Retrieval

- **Goal**: promote progress in content-based retrieval (CBR) from digital video via open, metrics-based evaluation
- **Corpus**: 11 hours MPEG-1/-2 digital video
- **Topics**: 74 query topics developed by participants
- **Tasks** (due Aug. 2001):
  - Identify shot boundaries (automatic)
  - Given statement of information need, return ranked list of shots which best satisfy the need
    - Known item search (automatic comparison to reference)
    - General statements of information need (human assessment per shot of whether the shot meets the need or not)
  - Interactive and fully automatic approaches are allowed

Who Are the Participants:

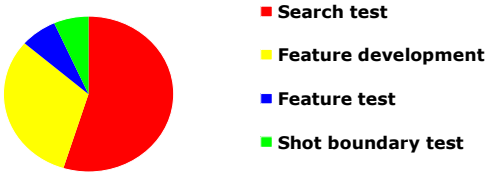
17 Groups and Their Completed Tasks in 2002

	Shot Boundary	Concept Detection										Search	
		1	2	3	4	5	6	7	8	9	10	Int.	Man.
Carnegie Mellon U.		X	X	X	X	X		X	X	X	X	X	X
CLIPS-IMAG (Fr)	X			X	X				X		X		X
CWI Amsterdam (NL)													X
Dublin City University				X					X	X		X	
Fudan Univ. (China)	X	X	X	X	X	X	X	X	X	X	X		X
IBM Research (US)	X	X	X	X	X	X	X	X	X	X	X	X	X
Imperial C. London	X											X	X
Indiana University												X	
Institut Eurecom (Fr)		X	X	X	X	X	X	X					
Mediamill/U Amsterdam		X	X	X	X	X	X	X	X	X	X		
Microsoft Research Asia	X	X	X	X	X	X	X	X	X	X	X	X	X
National U. Singapore	X												
Prous Science (Esp)													X
RMIT University (Aus)	X												
Univ. Bremen (D)	X	X	X										
U. Maryland/INSA/Oulu								X				X	X
Univ. Oulu/VTT (Fin)					X	X	X		X	X		X	X

TREC-2002 Video Data

- Difficult to get video data for use in TREC because ©
- Used mainly Internet Archive
  - –advertising, educational, industrial, amateur films 1930-1970
  - –produced by corporations, non-profit organisations, trade groups, etc.
  - –Noisy, strange color, but real archive data
  - –73.3 hours partitioned as follows:

Sample Video



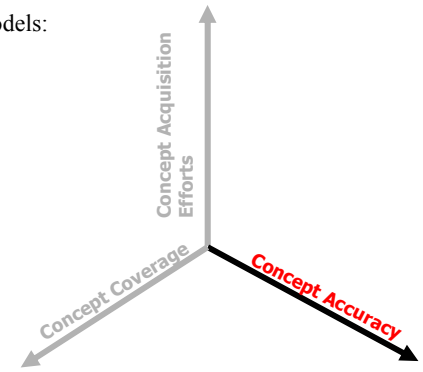
TREC Video 2003 and 2004

- Establish collaborative annotation forum
- Corpus: 192 hours of CNN, ABC, and CSPAN news: 63 hours for development, 65 hours for testing in 2003 and 64 hours for testing in 2004.
- 4 benchmark tasks
  - Shot boundary detection,
  - Story segmentation,
  - Semantic Concept detection,
  - Search task

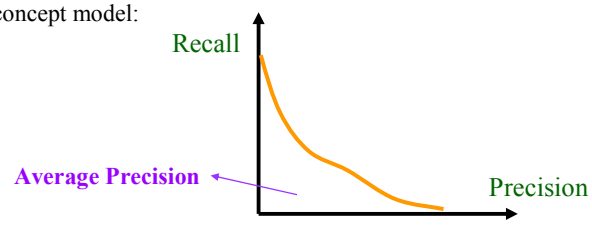
Sample Video

Performance Measurement of Concept Modeling

- From the viewpoints of a set of concept models:



- From the viewpoints of a concept model:



Performance Metric -- Precision

- Example:
  - *"Find shots from behind the pitcher in a baseball game as the batter is visible"*
  - *Precision at the first 10 returns:*
    8 / 10 = 0.8
  - *Precision at the first 20 returns:*
    12 / 20 = 0.6
  - *Precision* (at N) :

$$P@N = \frac{N_{CORRECT}}{N}$$

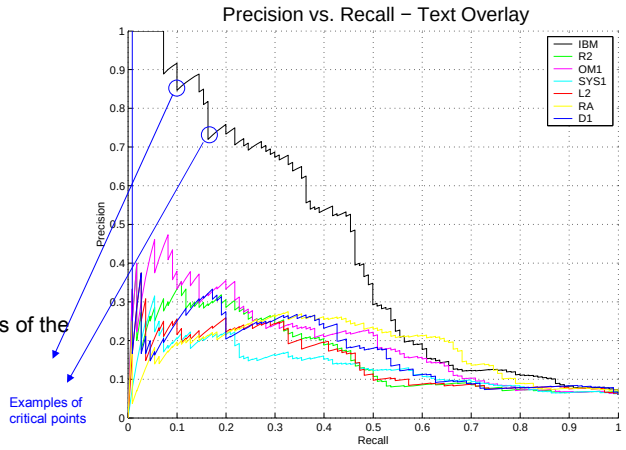
Performance Metric -- Recall

- Example:
  - *"Find shots from behind the pitcher in a baseball game as the batter is visible"*
  - *Assume there are totally 30 "positive" shots in the whole data set.*
  - *Recall at the first 10 returns:*
    8 / 30 = 0.27
  - *Recall at the first 20 returns:*
    12 / 30 = 0.4
  - *Recall* (at N) :

$$R@N = \frac{N_{CORRECT}}{N_{GT}}$$

Performance Metric -- Precision-Recall Curve and Average Precision

- Example:
  - Finding relevant shots that include overlay-text

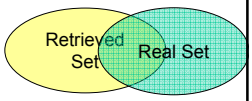


- *Average Precision* :

$$AP = \frac{1}{N_{GT}} \sum_{N_i} P@N_i$$

where N_i are all the ranks of the relevant retrieved shots.

Average Precision is the area under the P-R (ROC) curve



Performance Metric - Average Precision

- Example:
  - *"Find shots from behind the pitcher in a baseball game as the batter is visible"*
  - *Assume there are totally 18 "positive" shots in the whole data set.*
  - *Average Precision:*

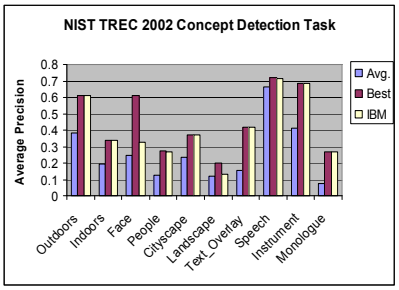
Sum of precision @ Ni =
1+ 1+ 1+ 1+
1+ 1+ 1+ 1+
+ 9/11+
10/13+11/14+
12/18
= 11.05

AP = 11.05 / 18 = 0.61

TREC 2002 Concept Detector Performance

	Outdoors	Indoors	Face	People	Cityscape	Landscape	Text_Overlay	Speech	Instrument	Monologue
Avg.	0.385	0.192	0.247	0.125	0.233	0.12	0.152	0.661	0.412	0.073
Best	0.609	0.335	0.613	0.274	0.374	0.198	0.418	0.721	0.686	0.268
IBM (M-1)	0.609	0.335	0.327	0.271	0.374	0.134	0.418	0.713	0.686	0.268
IBM-2 (M-2)	0.6		0.288				0.181			
r1/2/3	0.573	0.128	0.613	0.274	0.154		0.13	0.642	0.257	0.082
Sys1/2	0.395	0.246	0.111	0.071	0.325	0.198	0.118	0.663	0.564	0.009
RA	0.456	0.205	0.473		0.348		0.151	0.57	0.511	0.009
MSU								0.721		0.149
OM1	0.471		0.15		0.303		0.184			
T1/T2				0.248	0.299	0.193		0.645	0.637	
Best others	0.573	0.246	0.613	0.274	0.348	0.198	0.184	0.721	0.637	0.149
IBM ranks	(1st,2nd) / 11	1st / 8	(3rd, 4th) / 10	2nd / 9	1st / 11	4th / 7	(1st, 3rd) / 10	2nd / 13	1st / 12	1st / 9
Ground Truth	962	351	415	486	521	127	110	1382	1221	38

*: the highest score; *: the second highest score



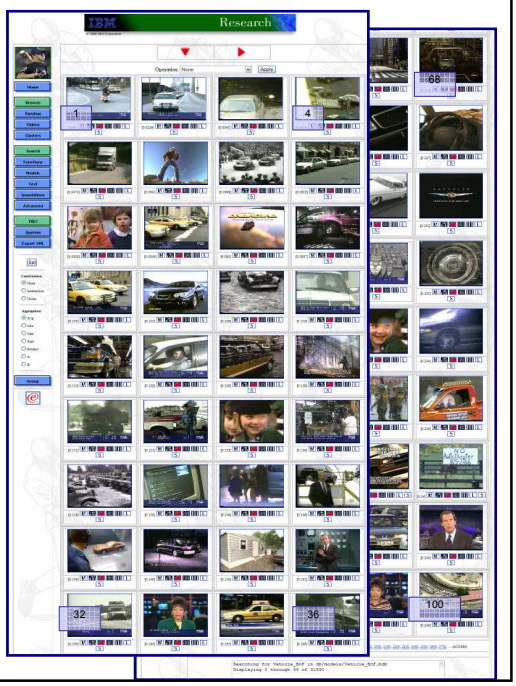
- Mean Average Precision of the best system 0.414 compared to 0.264 for next best system (58% better)
- Strengths of the best system: many features, reasonably good training data, solid training and validation framework

Concept Detection Example: Cars

- ***"Car/truck/bus**: segment contains at least one automobile, truck, or bus exterior*"
- Concept was trained on the annotated training set.
- Results are shown on the test set (see full results online)
  - 83 out of the top 100 are correct
  - TREC Evaluation by NIST

Run	Precision
Best	0.83
Best IBM	0.83
Best non-IBM	0.69
Average non-IBM	0.26

By IBM TRECVID 2003 Team

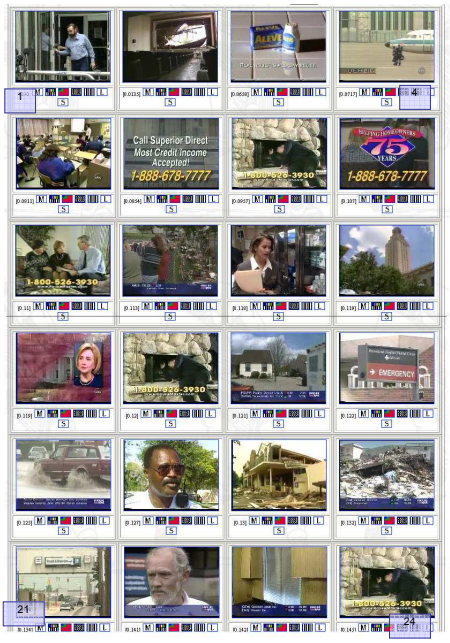


Concept Detection Example: Building

- ***"Building**: segment contains a building. Buildings are walled structures with a roof.*"
- Concept was trained on the annotated training set.
- Results are shown on the test set (see full results online)
  - 71 out of the top 100 are correct
  - TREC Evaluation by NIST

Run	Precision
Best	0.71
Best IBM	0.71
Best non-IBM	0.51
Average non-IBM	0.33

By IBM TRECVID 2003 Team



Concept Detection Example: Ms. Albright

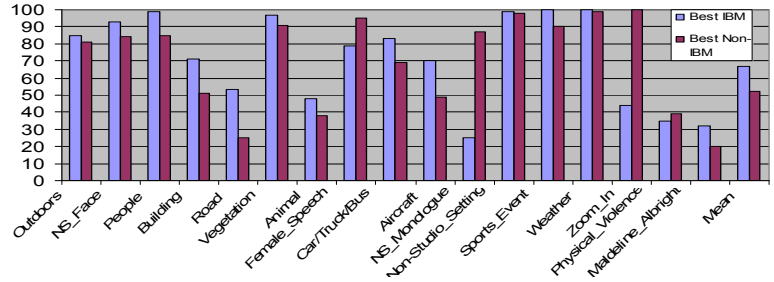
- ***"Person X**: segment contains video of person x (x = Madeleine Albright).*"
- Results at the CF2 (validation set)

Run	Average Precision
Best IBM Audio Models	0.30
Best IBM Visual Models	0.29
Best of Fusion	0.47

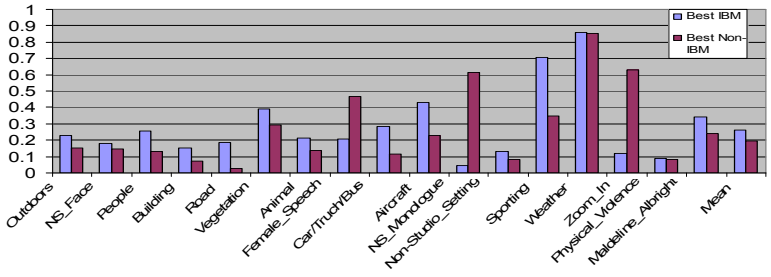
- Results are shown on the test set (see full results online)
  - TREC Evaluation by NIST

Run	Precision
Best	0.32
Best IBM	0.32
Best non-IBM	0.20
Average non-IBM	0.04

Comparison of Precision at Top 100 and Average Precisions (2003)



Precision @ Top 100



Official NIST Average Precision Values

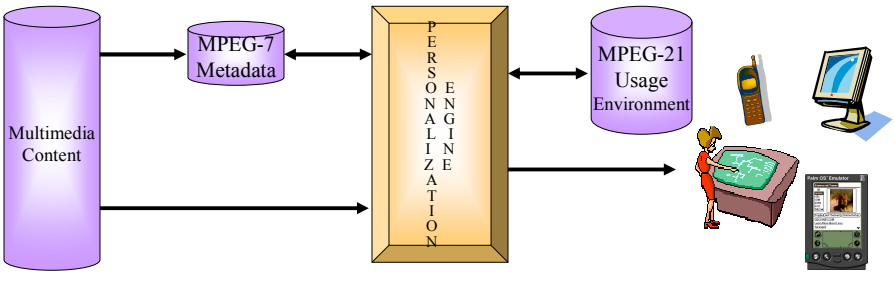
Comparison of the Detection Accuracy with Different Model/Modal Fusion Methods

- Precision on the Top-100 Return:

	Outdoor	NSFace	People	Building	Road	Vege.	Animal	F_Spee	Vehicle	Aircra	Monol.	NonStudio	Sports	Weather	Zoom_In	Violence	Albright	Mean
BOU	81	80	90	53	46	96	10	46	68	38	24	97	81	79	44	33	32	58.706
EF	67	77	95	60	33	97	47	69	80	63	25	96	99	98	44	28	28	65.059
BOF	71	77	97	71	52	93	47	69	80	47	25	96	98	100	44	35	32	66.706
DMF17	82	93	90	54	49	97	45	35	76	70	1	99	98	99	44	9	28	62.882
DMF64	82	73	79	53	41	96	33	79	56	67	0	93	98	99	44	34	4	60.647
MLP_BOR	78	75	97	61	53	94	47	38	70	65	1	95	100	97	44	27	30	63.059
MLP_EFC	73	67	97	41	33	96	48	19	49	60	3	97	99	99	44	27	27	57.588
MN	85	55	99	52	45	97	47	66	81	63	25	96	99	98	44	22	28	64.824
ONT	67	77	95	56	42	97	47	69	83	69	6	94	99	98	44	28	28	64.647
BOBO	85	73	99	56	52	93	10	66	56	63	0	97	98	99	44	22	32	61.471
Maximum:	85	93	99	71	53	97	48	79	83	70	25	99	100	100	44	35	32	66.706
Average:	76.857	73.857	93.429	55.429	45	95.71	44.857	53.571	70.714	63	8.714	95.71429	98.71	98.5714	44	26	25.286	62.908

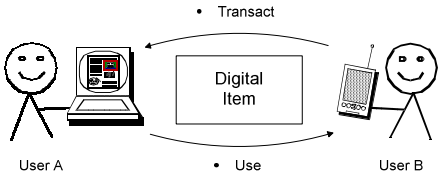
- Some observations from the detection results of TRECVID 2003:
  - Multi-Modality Fusion have better results than individual audio or video models.
  - None of the multi-Concept fusion methods stand out to be the best – performance varies based on concepts.
  - Multi-Concept Fusion did not significantly improve (or worse) the mean precision values from single-model multi-modality fusions.

II.4. Personalization & Summarization System

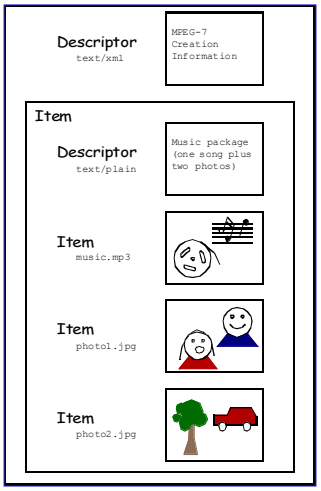


MPEG-21 Multimedia Framework
"Transactions of Digital Items"

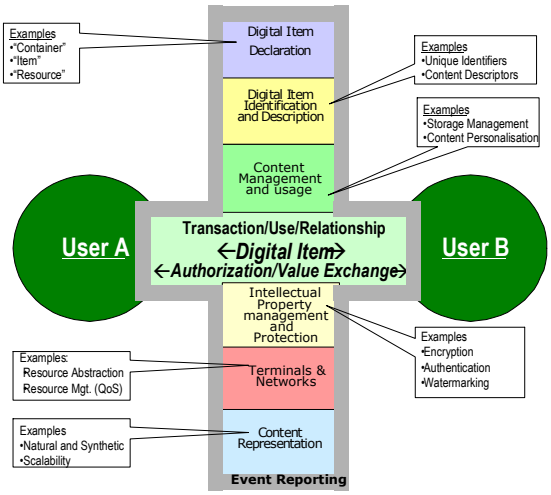
- Users and participants in the content value chain seamlessly exchange content in form of "digital items" across networks and devices
- Framework supporting all forms of electronic content/intellectual property (video, music, learning objects, on-line reports, etc.)
- Digital Item = bundling of:
  - Essence (i.e., media resources)
  - Metadata
  - Rights expressions
  - Identifiers



Example: Digital music package



MPEG-21 Standard Framework



- "Interoperable Multimedia Framework"
- "E-Commerce for E-Content"
- "Digital Audio-Visual Framework"
- **Vision**: "To enable transparent and augmented use of multimedia resources across a wide range of networks and devices."
- **Goal**: Integration of technologies for content *identification* and *consumption*
- **Output**: ISO technical report and technical specification (Int'l Standard in 2003)

MPEG-21 Multimedia Framework Standard

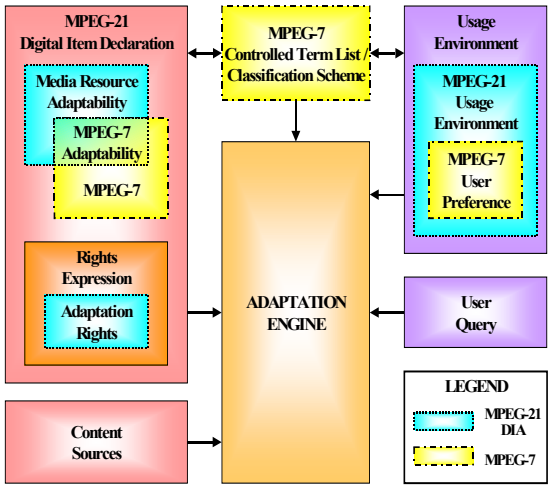
MPEG-21 Part	Current Status	Completion Date
Multimedia Framework (Technical Report)	TR	Nov. 2004
Digital Item Declaration	IS	Mar. 2003
Digital Item Identification	IS	Mar. 2003
Intellectual Property Management & Protection (IPMP)	CD	Mar 2005 (Jan. 2006)
Rights Expression Language	IS	Apr. 2004
Rights Data Dictionary	IS	May 2004
Digital Item Adaptation	IS	Oct. 2004
Digital Item Processing	FCD	Apr. 2005 (July. 2005)

RED: Target ISO Standardization (IS) Date

Digital Item Adaptation

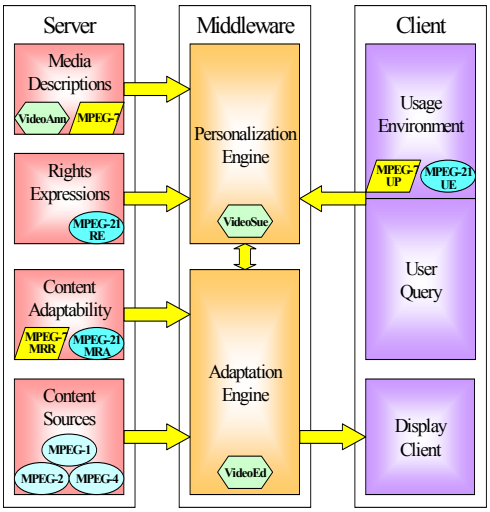
using MPEG-7 and MPEG-21

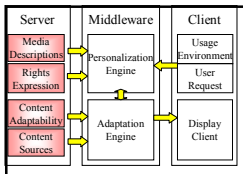
- Media Description
  - MPEG-7 Description Schemes
- Usage Environment
  - MPEG-7 User Preferences
  - MPEG-21 Usage Environment
- Content Adaptability
  - MPEG-7 Media Resource Requirement
  - MPEG-21 Media Resource Adaptability



A Personalization and Summarization System Architecture

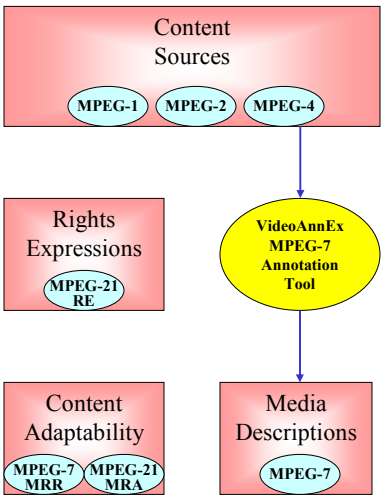
- Database Server
  - Content Sources
  - MPEG-7 Media Descriptions
  - MPEG-21 Rights Expression
  - Content Adaptability
- Media Middleware
  - Select Personalized Contents in Personalization Engine
  - Retrieve and Adapt Contents in Adaptation Engine
- User Client
  - Request for Personalized Content
  - Communicate Usage Environment

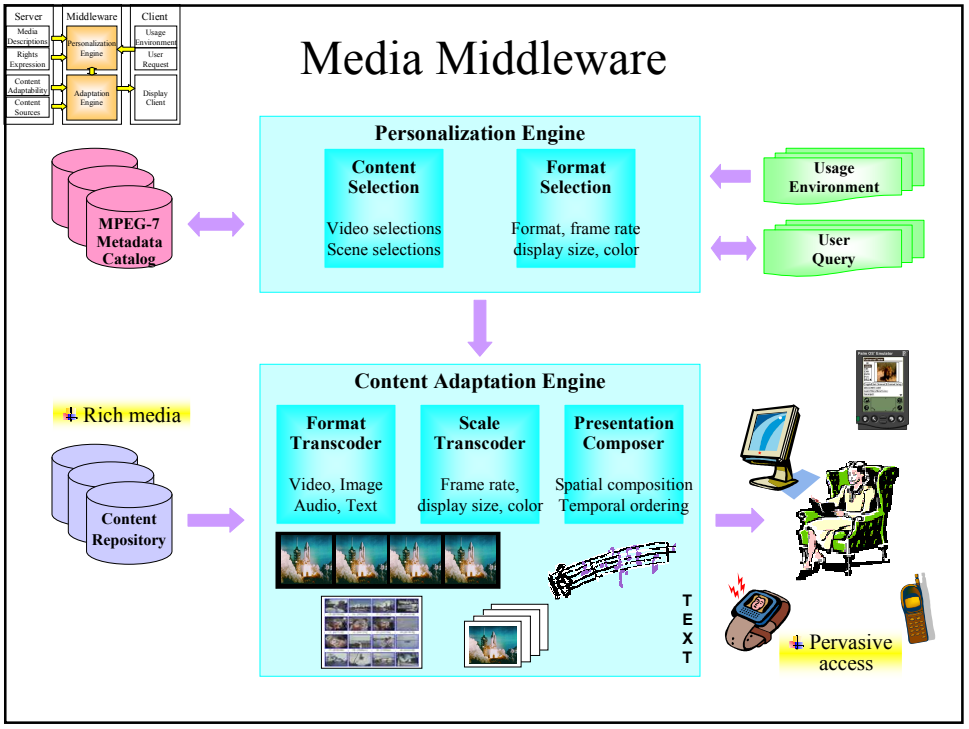


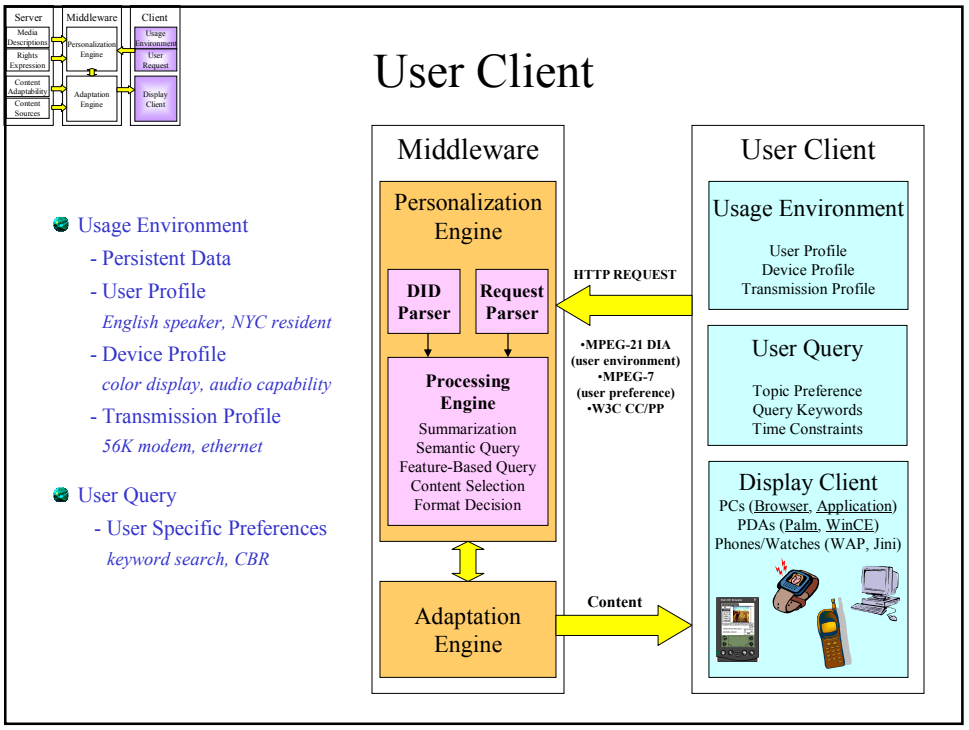


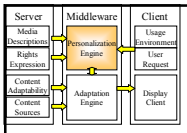
Database Server

- Media Descriptions
  - MPEG-7 DS
- Rights Expression
  - MPEG-21 RE
- Content Adaptability
  - MPEG-7 Media Resource Requirement
  - MPEG-21 Media Resource Adaptability




Server
Middleware
Client
Media Descriptions
Rights Expression
Content Adaptability
Content Sources
Personalization Engine
Adaptation Engine
Usage Environment
User Request
Display Client
Media Middleware
Personalization Engine
Content Selection
Video selections
Scene selections
Format Selection
Format, frame rate
display size, color
Usage Environment
User Query
MPEG-7 Metadata Catalog
Content Adaptation Engine
Format Transcoder
Video, Image
Audio, Text
Scale Transcoder
Frame rate,
display size, color
Presentation Composer
Spatial composition
Temporal ordering
Rich media
Content Repository
TEXT
Pervasive access


Server
Middleware
Client
Media Descriptions
Rights Expression
Content Adaptability
Content Sources
Personalization Engine
Adaptation Engine
Usage Environment
User Request
Display Client
User Client
Usage Environment
- Persistent Data
- User Profile
English speaker, NYC resident
- Device Profile
color display, audio capability
- Transmission Profile
56K modem, ethernet
User Query
- User Specific Preferences
keyword search, CBR
Middleware
Personalization Engine
DID Parser
Request Parser
Processing Engine
Summarization
Semantic Query
Feature-Based Query
Content Selection
Format Decision
Adaptation Engine
HTTP REQUEST
•MPEG-21 DIA (user environment)
•MPEG-7 (user preference)
•W3C CC/PP
Content
User Client
Usage Environment
User Profile
Device Profile
Transmission Profile
User Query
Topic Preference
Query Keywords
Time Constraints
Display Client
PCs (Browser, Application)
PDAs (Palm, WinCE)
Phones/Watches (WAP, Jini)



Media Middleware
Video Summarization on Usage Environment

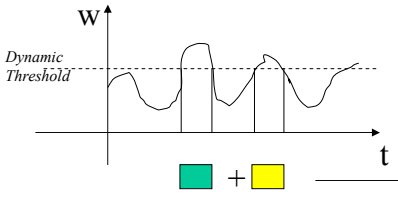
P = user preference vector, where p_i denotes the preference weighting for concept i.
A = attribute matrix, where score $a_{i,j}$ is defined as the relevance of concept i in shot j.
W = weighted importance vector, where w_i is the weighted concept for shot i.

video segment weighting

attribute relevance matrix

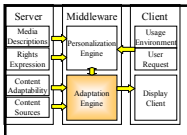
user preference vector

$$\vec{W} = \begin{bmatrix} w_1 \\ w_2 \\ \dots \\ w_M \end{bmatrix} = \begin{bmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,N} \\ a_{2,1} & \dots & \dots & \dots \\ \dots & \dots & a_{i,j} & \dots \\ a_{M,1} & \dots & \dots & a_{M,N} \end{bmatrix} * \begin{bmatrix} p_1 \\ p_2 \\ \dots \\ p_N \end{bmatrix}$$

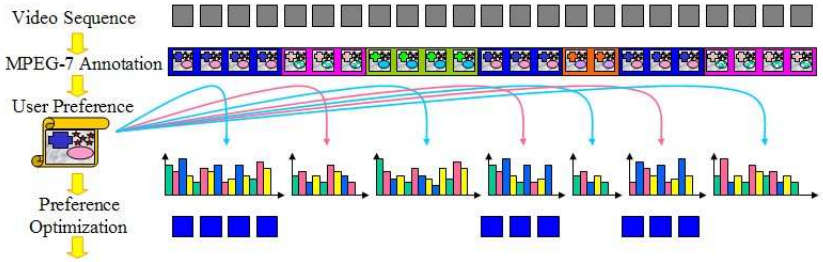


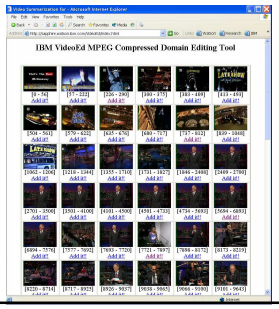
A is generated manually, semi-automatically, or automatically

Summarized video

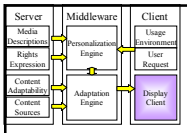


Media Middleware
Compressed-Domain MPEG Editing Tool



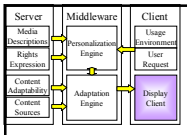


- Run-time user selection of video shots
- Fast creation of personalized summaries
- On-line preview of editing results
- Compatible with MPEG-1 and MPEG-2
- Demo http://sapphire.watson.ibm.com/VideoEd



User Client
PDA Devices

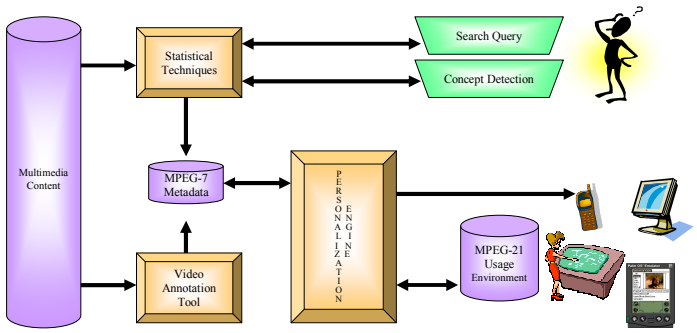
- Browsing
  - Channels
  - Links
- Preferences
  - Video Source
  - Preferences
  - Time Constraint
- Queries
  - Topics
  - Keyword Search
  - Time Constraints



User Client
IBM Websphere Portal Server

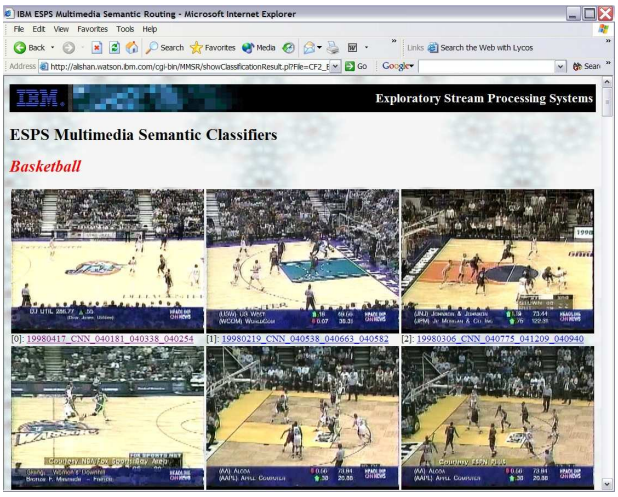
- Usage Environment
  - User Preference Topics
    [news, entertainment, education]
  - Device [terminal, PDA]
  - Network Constraint
- User Query
  - Topic Preferences
  - Keyword Search
  - Time Constraint

Demo and Part II Summary



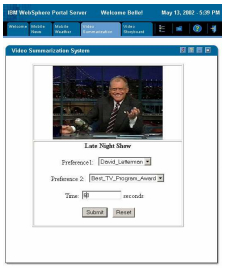
Demo -- Semantic Concept Detection

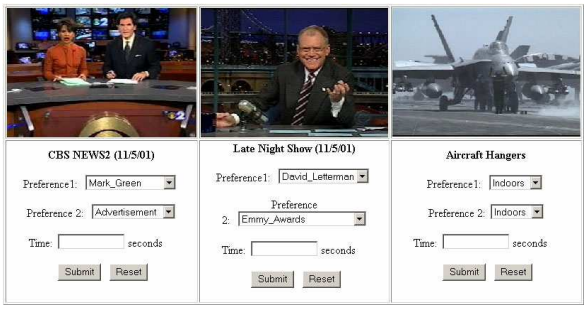
- http://www.research.ibm.com/VideoDIG
- E.g.:



Demo – Personalization and Summarization

- IBM Websphere Portal Server for Personalization
- User Query
  - Topic Preferences 1 = David Letterman
  - Topic Preferences 2 = Guest Introductions
  - Time Constraint = 60 seconds



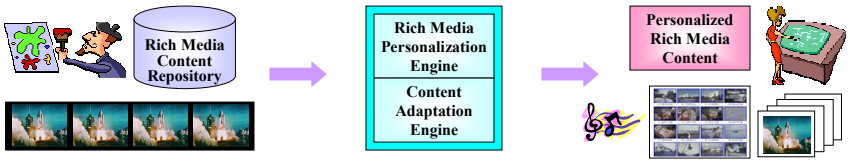


Interoperable Multimedia Content Management:

Relevant Emerging Standards – MPEG-7 and MPEG-21

- Metadata is critical for describing essential aspects of content:
  - Main topics, author, language, publication, etc.
  - Events, scenes, objects, times, places, etc.
  - Rights, packaging, access control, content adaptation, etc.
- Conformity with open metadata standards will be a vital:
  - Allows faster design and implementation
  - Interoperability with broad field of competitive standards-based tools and systems
  - Leveraging of rich set of standards-based technologies for critical functions such as content extraction, advanced search, and personalization
- Relevant critical standards for interoperable multimedia CM:
  - MPEG 7 Multimedia Content Description Interface
    - ISO/IEC standard for multimedia metadata (XML-Schema based)
  - MPEG 21 Multimedia Framework
    - ISO/IEC standard for transactions of digital items, rights management, and content adaptation
- Summary of benefits:
  - MPEG-7 allows interoperability of systems and tools for multimedia content analysis, annotation, indexing, searching, and filtering
  - MPEG-21 allows interoperable transactions of digital multimedia content

Summary of Video Personalization and Summarization



- Video Personalization and Summarization System allows universal access and personalized content of rich media to any user environment at anytime and anywhere
- Innovations
  - Standards-based (MPEG-7 & MPEG-21) interoperable solution
  - Off-line semi-automatic MPEG-7 XML content annotation
  - Digital item adaptation and transactions using MPEG-21
  - Optimized content adaptation to user query and user environment
  - Real-time compressed domain video composer for MPEG contents
- Applications
  - Enterprise rich media (i.e., e-Learning, video conferencing, news, communications)
  - Wireless (i.e., personalized video clips, image slide shows)
- Conclusion
  - Methods for automatic rich media annotation, indexing, retrieval, and optimized personalization and presentation

Part II References

- Belle L. Tseng, Ching-Yung Lin and John R. Smith, "**Video Personalization and Summarization System based on MPEG-7 and MPEG-21**," *IEEE Multimedia Magazine*, Jan.-Mar., 2004.
- Ching-Yung Lin, Belle L. Tseng and John R. Smith, "**Video Collaborative Annotation Forum: Establishing Ground-Truth Labels on Large Multimedia Datasets**," *Proc. of NIST Text Retrieval Conf. (TREC)*, Gaithersburg, MD, November 2003.
- A. Amir, W, Hsu, G. Iyengar, Ching-Yung Lin, M. Naphade, A. Natsev, C. Neti, H. J. Nock, J. R. Smith, Belle L. Tseng, Y. Wu, D. Zhang, "**IBM Research TRECVID-2003 System**," *Proc. NIST Text Retrieval Conf. (TREC)*, Gaithersburg, MD, November, 2003.
- Ching-Yung Lin, Belle L. Tseng, Milind Naphade, Apostol Natsev and John R. Smith, "**VideoAL: A Novel End-to-End MPEG-7 Automatic Labeling System**," *IEEE Intl. Conf. on Image Processing*, Barcelona, September 2003.
- MPEG Committee Documents – Programme of Work http://www.itscj.ipsj.or.jp/sc29/29w42911.htm

Part III: Emerging Technologies

Part III: Emerging Technologies

- Overview of Emerging Technologies (5 mins)
- *Part-Based Object Recognition (15 mins)*
- *Unsupervised Pattern Discovery (15 mins)*
- *Imperfect Learning and Cross-modality Autonomous Learning (15 mins)*
- *Distributed Video Semantic Filtering and Routing (15 mins)*
- Summary and Open Discussion (10 mins)

Part-based Representation for Image

Parts
(corners, regions, Maximum-Entropy-Regions)
[Kadir et al. 04']

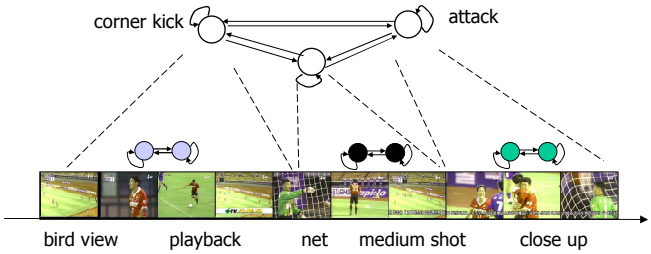
Image := a set of parts,

Part attributes := {Part appearance, Part relations}

Video Pattern Discovery with HHMM

[Xie et. al 2003]

- Supervised: Tracking, speech, DNA sequence recognition [Fine, Singer, Tishby'98] [Zweig 1997], [Ivanov'00]
- Left-right: Video clustering [Clarkson'99][Naphade'02]
- Application in unsupervised discovery has not been explored.

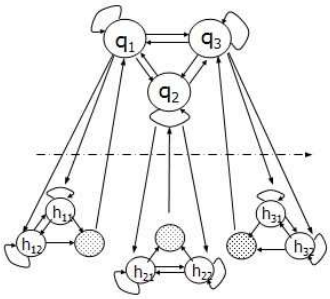


Top-level : events

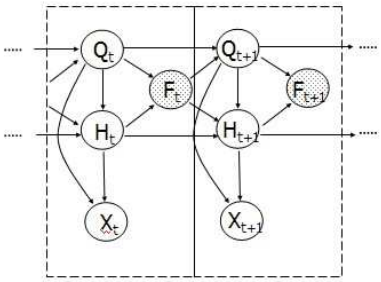
Bottom-level : temporal variations within an event

Hierarchical HMM

[Fine, Singer, Tishby'98]
[Murphy'01][Xie et al. ICME03]



State-space representation



DBN representation, unrolled in time

- Flexible control structure, extensible to multiple levels
- Efficient inference and estimation in O(T)
- Multi-level Viterbi algorithm for sequence labeling

(1) the model ✓ (2) its size? (3) which features?

Autonomous Learning

- Life-long learning from TV, dictionary/encyclopedia ?
- Can a computer learn visual concepts automatically from encyclopedia and media news?

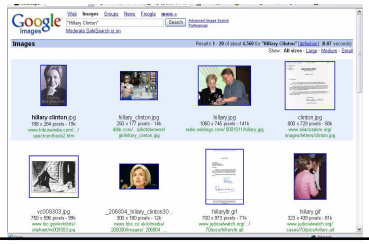
Autonomous Learning

- Correlation between different modalities makes autonomous learning possible:
  - Audio, close captions, and visual data in video sequences

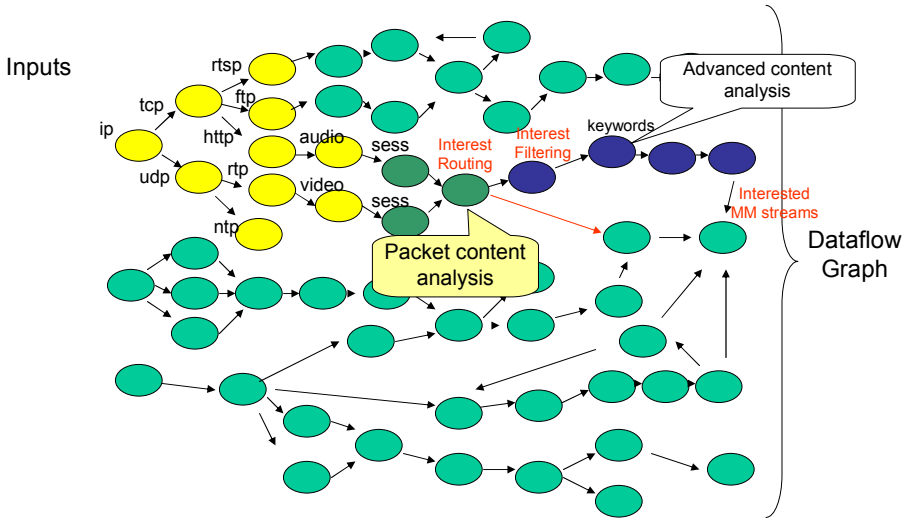
First--let's look at the national weather forecast... Unseasonably warm weather expected today in parts of ...

0.7658 0.7682 0.7746

  - Texts and images from web images, encyclopedia, …



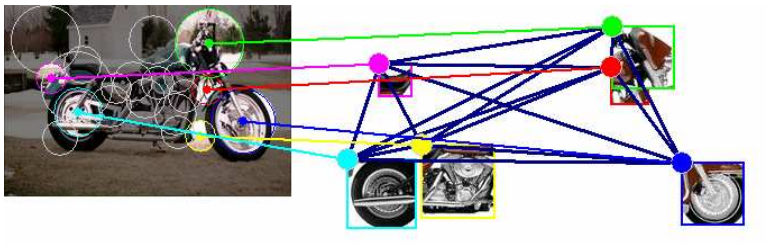
Semantic MM Routing and Filtering



III.1. Part-based Object Detection by Learning Random Attributed Graphs

Courtesy of Dongqing Zhang
Dept. of Electrical Engineering
Columbia University

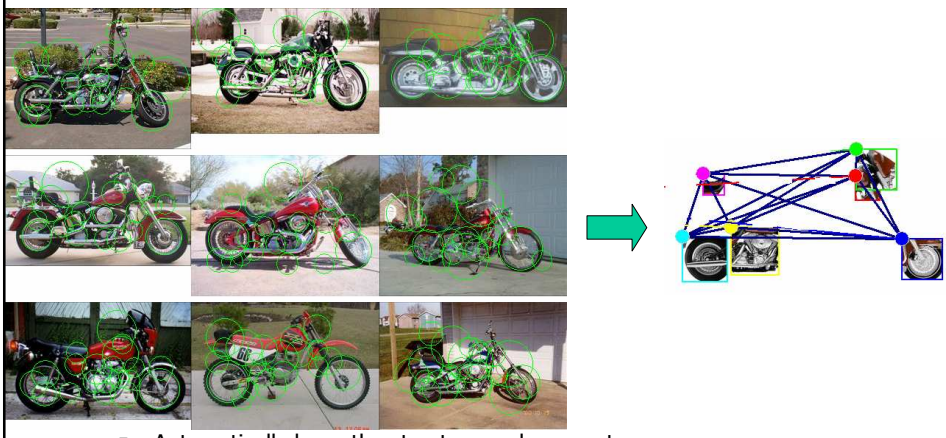
Problem 1 : Object Detection and Part Identification



a. Does the input image contain the specified object ?

b. Where are the object parts ?

Problem 2 : Learning Part-based Object Model



- Automatically learn the structure and parameters
- Minimum supervision : no object location and part location

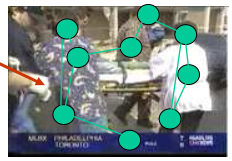
Prior Work on Part-based Object Detection

Model with Hand-built structure	Model without spatial structure	Model with learned structure and part statistics
Pictorial structure, [Felzenszwalb & Huttenlocher 98'] **Elastic Bunch Graph,** [Wiskott et. al 97'] **MRF model,** [Li 94']	**AdaBoost,** [Viola & Jones, 01']	**Constellation Model,** [Burl, Weber, Fergus, Perona, Caltech, Oxford 98'-04']

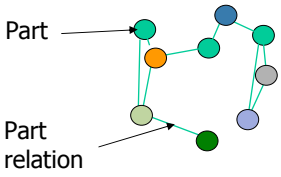
This new model :
Graph-based representation;
Can handle multi-view object detection

Part-based Representation of Visual Scene

ARG

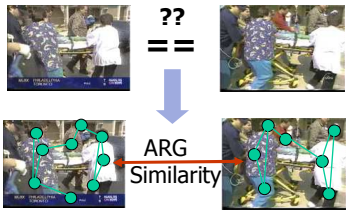


Visual scenes are considered as the composition of the parts with certain spatial/attribute relations, modeled as Attributed Relational Graph (ARG)



Attributed Relational Graph (ARG)

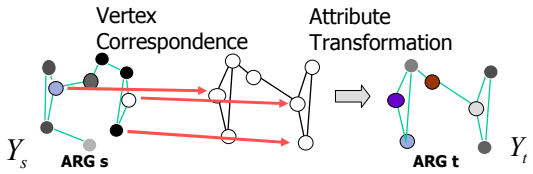
IND Detection
as Computing ARG similarity



ARG based on Interest Point Detection

- Region-based representation had very bad performance !
- Interest point detector: SUSAN (Smallest Univalue Segment Assimilating Nucleus) corner detector
- Local features at vertexes
  - Spatial location, Color, Gabor filter coefficients
- Part relational features at edges
  - Spatial coordinate difference

Stochastic Framework for ARG Similarity



Stochastic Process that Transforms ARG s to ARG t

ARG similarity is the likelihood or likelihood ratio of the stochastic process that transforms source ARG to target ARG

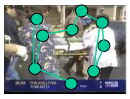
$$S(G^s, G^t) = \frac{p(Y^t|Y^s, H=1)}{p(Y^t|Y^s, H=0)}$$

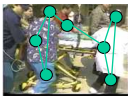
H: Hypotheses:

H = 1, Graph *t* is similar to Graph *s*

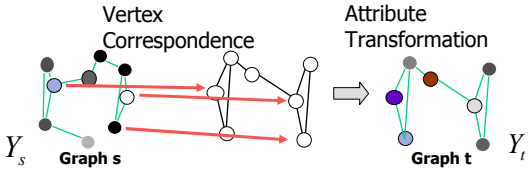
H = 0, Graph *t* is not similar to Graph *s*

Non-linear Scene Transformation





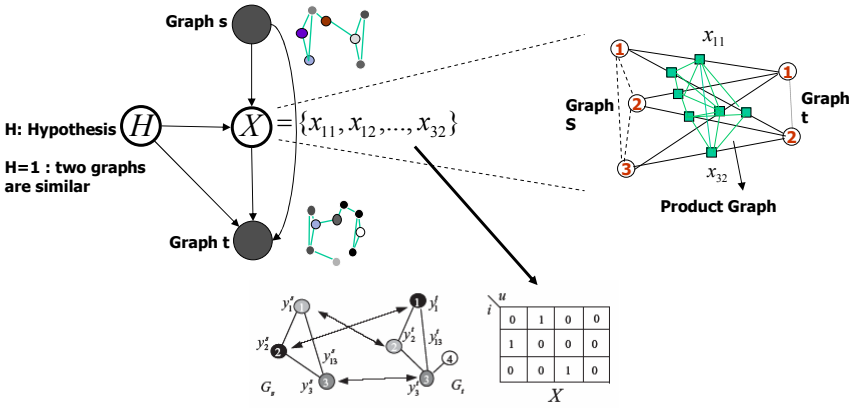
- Scene changes: object movement, occlusion etc.
- Camera changes: view point change, panning etc
- Photometric changes: Lighting etc.
- Digitization changes: Resolution, gray scale etc.



Model
Occlusion of objects
Addition of objects

Model
Object appearance change,
Object move,
Photometric change

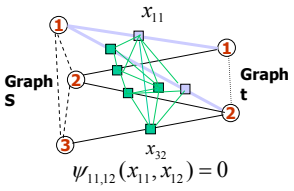
Generative Model of the Stochastic Transformation Process

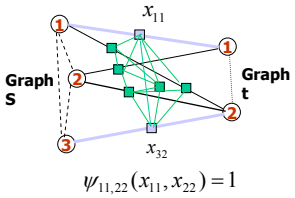


Transformation Likelihood

Transformation Likelihood

$$p(Y^t|Y^s,H) = \sum_{X \in \chi} p(Y^t|Y^s,X,H)p(X|Y^s,H)$$





Prior **MRF** for constraints

$$p(X|Y^s,H=h) = \frac{1}{Z(h)} \prod_{iu,jv} \psi_{iu,jv}(x_{iu},x_{jv}) \prod_{iu} \phi_h(x_{iu})$$

Conditional density for attribute transformation

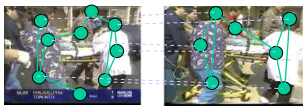
$$p(Y^t|X,Y^s,H) = \prod_{iu,jv} p(y^t_{uv}|x_{iu},x_{jv},y^s_{ij}) \prod_{iu} p(y^t_u|x_{iu},y^s_i)$$

Transformation likelihood is:

$$p(Y^t|Y^s,H=h) = \frac{Z'(Y^t,Y^s,h)}{Z(h)}$$

Learning to Match ARGs

- Feature point level learning: Label every feature point pairs



Vertex-level annotation

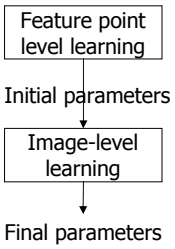
- Image level learning: Label duplicate pairs and non-duplicate pairs
  - Use Variational Expectation-Maximization (E-M)

Positive Samples

Negative Samples

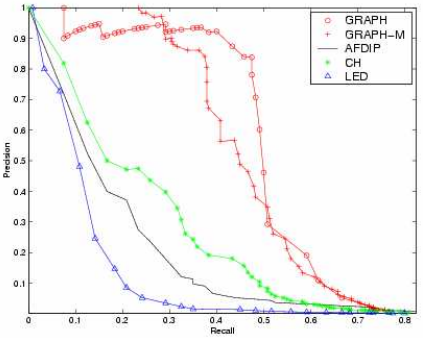
Experiments and Results

- Data set
  - Images are picked up from TREC-VID 2003 video frames (partly based on TDT2 topic detection ground truth)
  - 150 duplicate pairs, 300 non-duplicate images
- Learning
  - Training set: 30 duplicate pairs, 60 non-duplicate images
  - Feature point level learning
    - 5 duplicate pairs, 10 non-duplicate images
  - Image level learning
    - 25 duplicate pairs, 50 non-duplicate images



Compare with other similarity measures

Precision



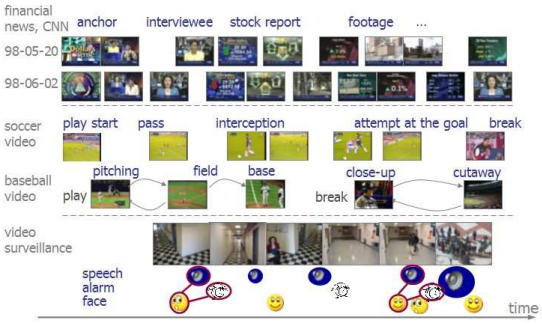
Recall

- (**CH**) HSV color histogram
- (**LED**) Local Edge Descriptor
- (**AFDIP**) Average feature distance of interest points
- (**GRAPH**) ARG matching with learning
- (**GRAPH-M**) ARG matching with manual parameter adjustment

Summary of Part-based ARG Visual Modeling Algorithm

- Statistical part-based similarity measure performs much better than global color histogram and grid-based edge map
- Learning-based ARG matching not only save human cost,but also may give better performance

III.2. Unsupervised Pattern Discovery for Multimedia Sequences



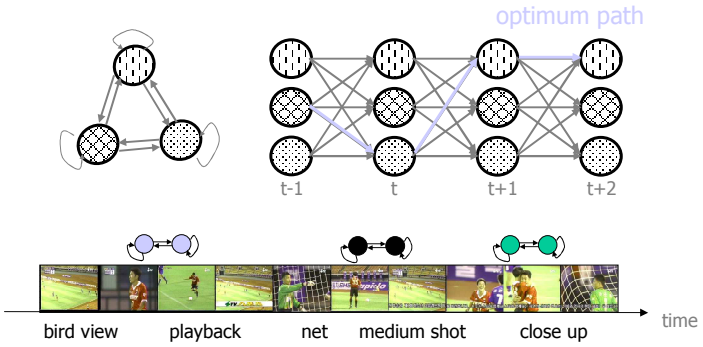
Courtesy of Lexing Xie
Dept. of Electrical Engineering
Columbia University

Unsupervised Pattern Discovery

- The goal in three folds:
  - Build suitable computational models
  - Identify appropriate content features
  - Generalize to different domains
- The benefits
  - Save annotation time
  - Facilitate manual filtering and browsing
  - Mining useful and novel patterns

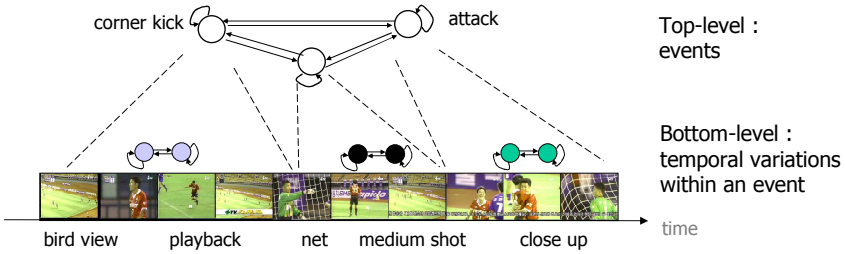
HMM for Temporal Patterns

- Need to capture the appearance and transition
  → hidden Markov model, supervised learning
  - speech recognition (>90%)
  - event recognition in soccer videos (83.5%) [Xie'02]

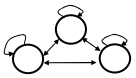


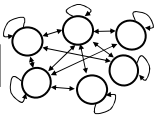
Modeling Hierarchical Video Patterns

- Need to capture the appearance and transition
  → hidden Markov model, supervised learning
  - speech recognition (>90%)
  - event recognition in soccer videos (83.5%) [Xie'02]
- Patterns occur at a number of levels
  - States in each level correspond to different semantic concepts
  - Follow different transition models



The Need for Model Selection

soccer   ?

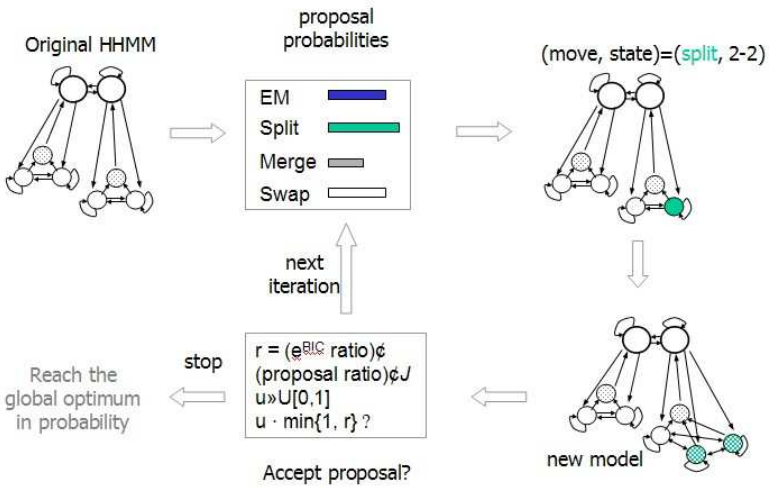
news   ?

talk show   ?

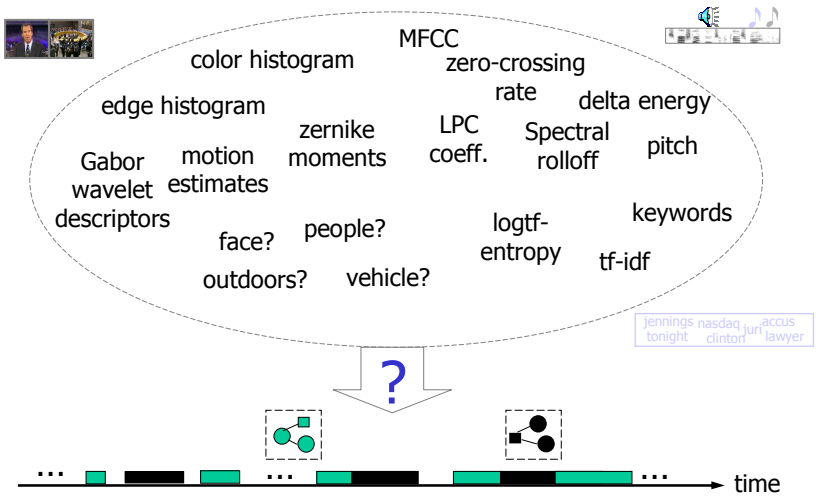
- Different domains have different descriptive complexities.

Model Selection with RJ-MCMC

[Green95]
[Andrieu99]
[Xie et.al 03]



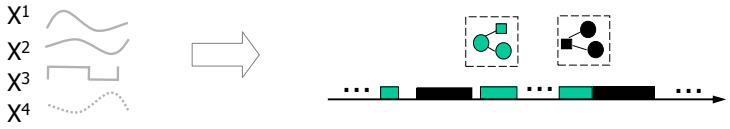
Which Features Shall We Use?

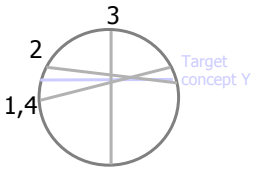


Feature Selection

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

- Goal: to identify a good subset of measurements, so as to improve generalization and reduce computation.
- Criteria: eliminate irrelevance and redundancy.



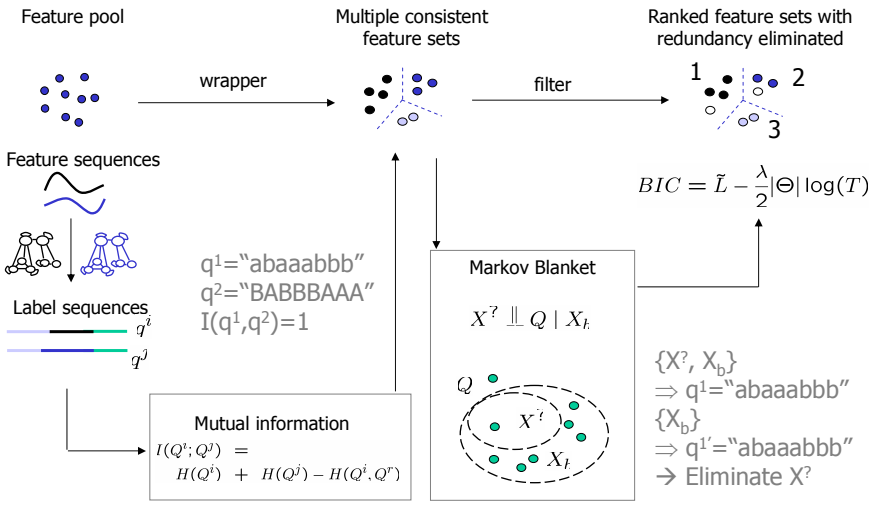


Our problem:

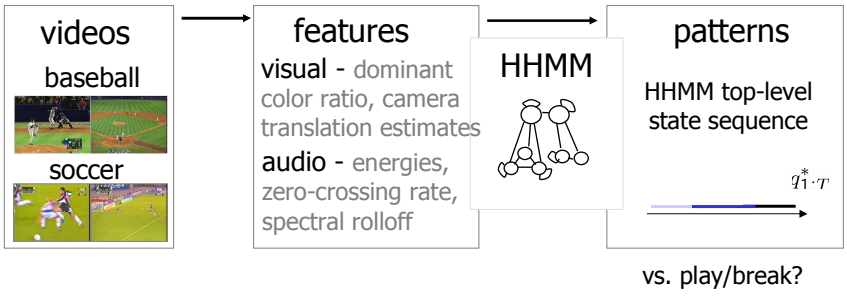
Unsupervised →	No target concept defined *a priori*
Temporal sequence →	Samples not i.i.d.

Feature Selection

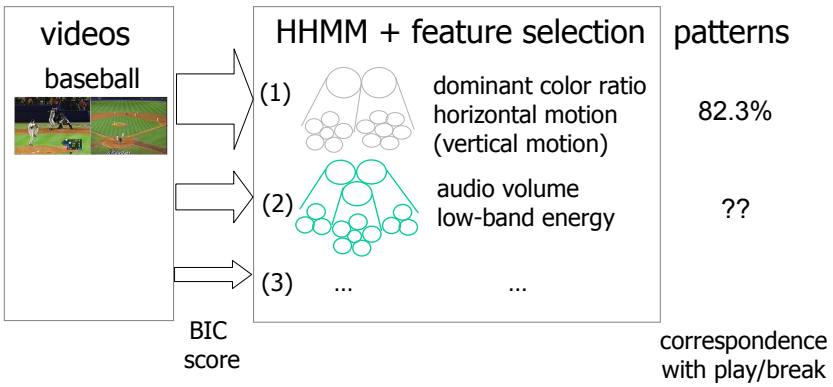
[Koller'96] [Xing'01]
[Xie et al. ICIP'03]



Results: on Sports Videos



Results: on Baseball Videos



Results: Comparison

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%
multi-HMM	Y	N	75.5§1.8%
LR-HHMM	N	N	73.1§1.1%
HMM	N	N	54.0§1.0%
K-Means	N	N	64.0§10.%

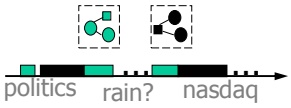
Automatic selection of both model and features

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'

Summary of Unsupervised Clustering

- Multimedia pattern discovery
  - Unsupervised discovery of temporal patterns
  - Finding meaningful patterns across multiple modalities
- Future work
  - Learning and perception
  - Applications to other content collections, generalize to other domains.

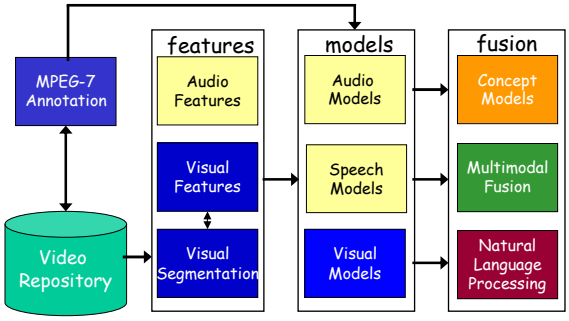


III.3. Imperfect Learning and Autonomous Learning

Courtesy of Xiaodan Song
Dept. of Electrical Engineering
University of Washington

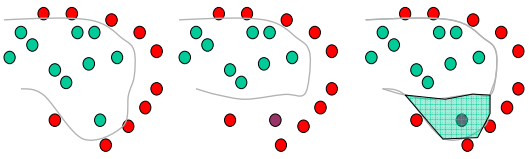
There are scalability problems in the supervised video semantic annotation framework

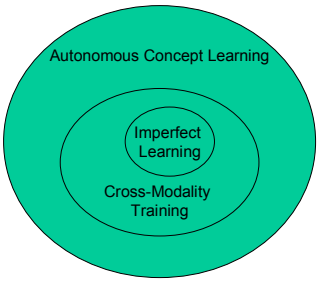
- For training:
  - Tremendous human effort required: we required extensive human labeling effort to have ground truth for training. E.g., 111 researchers from 23 groups annotated 460K semantic labels on 62 hours of videos in 2003.
  - Concept ontology is pre-defined: we won't be able to train more basic concepts if they are not annotated. For instance, 133 concepts in the 2003 ontology.



Imperfect Learning and Autonomous Learning

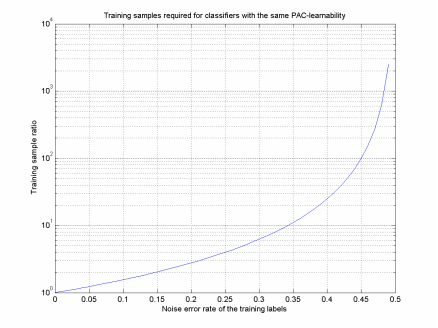
- Autonomous Learning of Video Concepts through Imperfect Training Labels Obtained Through Text Recognition – *Can concept models be automatically learned from cross-modality information?* :
  - Develop theories and algorithms for supervised concept learning from imperfect annotations -- imperfect learning
  - Develop methodologies to obtain imperfect annotation – learning from cross-modality information or web links
  - Develop algorithms and systems to generate concept models – novel generalized Multiple-Instance Learning algorithm with Uncertain Labeling Density





Imperfect Learning: theoretical feasibility

- Imperfect learning can be modeled as the issue of noisy training samples on supervised learning.
- Learnability of concept classifiers can be determined by probably approximation classifier (pac-learnability) theorem.
- Given a set of "fixed type" classifiers, the pac-learnability identifies a minimum bound of the number of training samples required for a fixed performance request.
- If there is noise on the training samples, the above mentioned minimum bound can be modified to reflect this situation.
- The ratio of required sample is independent of the requirement of classifier performance.
- Observations: practical simulations using SVM training and detection also verify this theorem.



A figure of theoretical requirement of the number of sample needed for noisy and perfect training samples

Training samples required when learning from noisy examples

- **Theorem** Let $h < 1/2$ be the rate of classification noise and N the number of rules in the class **C**. Assume $0 < e, h < 1/2$. Then the number of examples, m, required is at least

$$m \geq \max\left[\frac{\ln(2\delta)}{\ln(1-\varepsilon(1-2\eta))}, \log_2 N \cdot (1-2\varepsilon(1-\delta)+2\delta))\right]$$

and at most

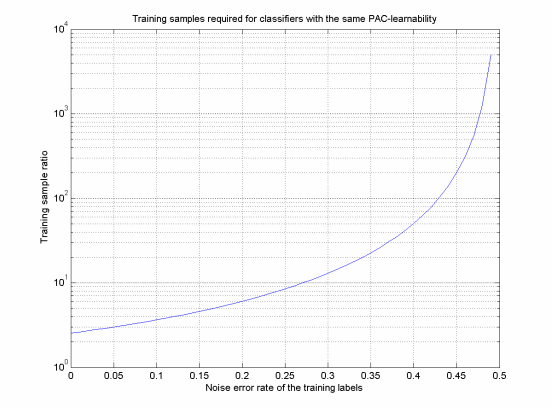
$$\frac{\ln(N/\delta)}{\varepsilon \cdot (1-\exp(-\frac{1}{2}(1-2\eta)^2))}$$

r is the ratio of the required noisy training samples v.s. the noise-free training samples

$$r_\eta = (1-\exp(-\tfrac{1}{2}(1-2\eta)^2))^{-1}$$

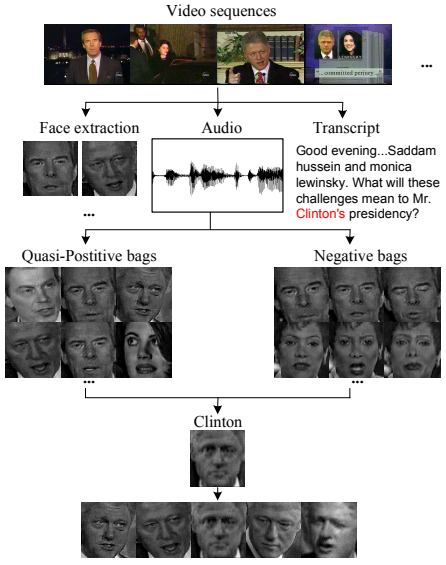
Training samples required when learning from noisy examples

- Ratio of the training samples required to achieve PAC-learnability under the noisy and noise-free sampling environments. This ratio is consistent on different error bounds and VC dimensions of PAC-learnable hypothesis.



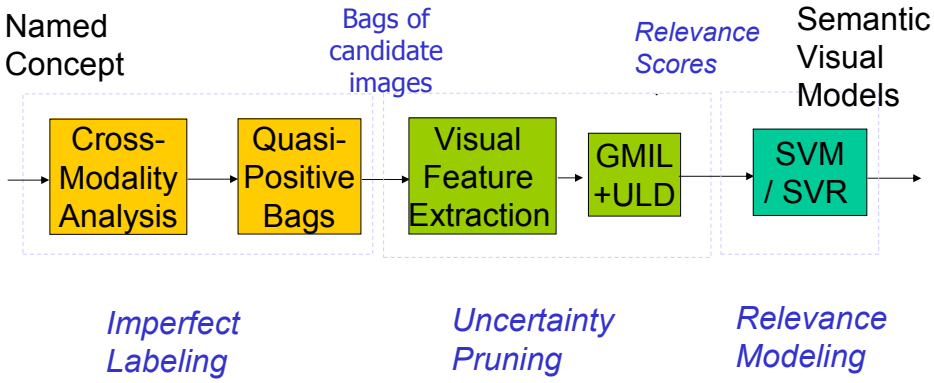
Learning from Cross-Modality Information

- Objective:
  - –Generate Visual Concept Models without Annotation
    - •Cross-Modality Training using Automatic Speech Recognition Transcription
    - •Continuously Learning Concepts from Broadcasting TVs
    - •Large Scale of Visual Concept Learning from Sub-Optimal Annotations such as Google Images
- Methodologies:
  - –Multiple Instance Learning with Quasi-Positive Bags
  - – Support Vector Machine Regression

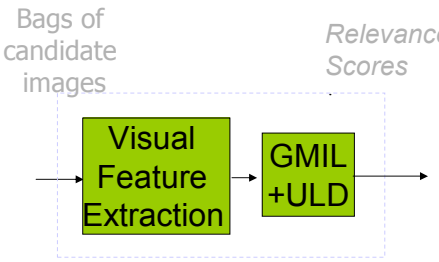


Example: Training Face Models

System Diagram for Autonomous Learning



Uncertainty Pruning



- Face-detection
- Features: Eigenfaces, color histogram, color layout, MPEG-7 features, …
- Generalized MIL (GMIL) allows quasi-positive bags, solved using ULD

Multiple-Instance Learning

- The trainer only labels collections of examples (bags).
- A bag is labeled
  - "negative" if all the examples in it are negative.
  - "positive" if there is at least one positive example in it.

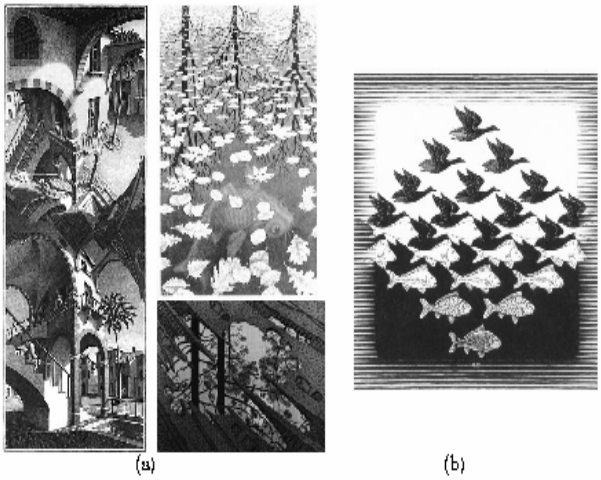


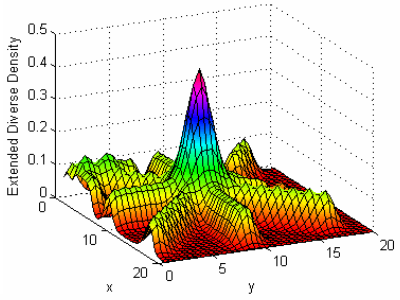
Figure: Some (a) positive and (b) negative training examples

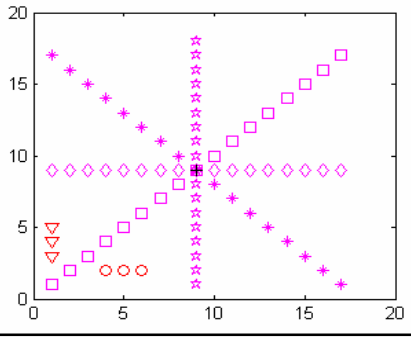
Actual positive instance: Trees [Maron 98]

Uncertain Labeling Density (ULD)

$$\arg\max \frac{\sum_i P\left(t \mid B_i^+\right) \cdot \prod_i \Pr\left(t \mid B_i^-\right)}{Z}$$

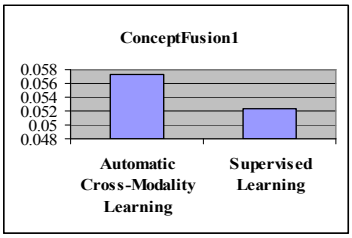
Z is a normalization constant to keep ULD in [0,1].



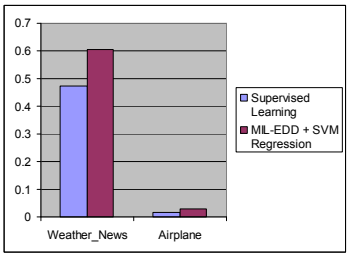


Preliminary Detection Results based on Cross-Modality Autonomous Concept Training

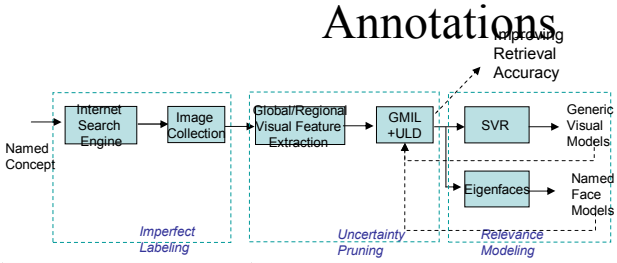
- Experiment1: Face Models were trained using the Feature Train Set
- Detection on the ConceptFusion1 Set

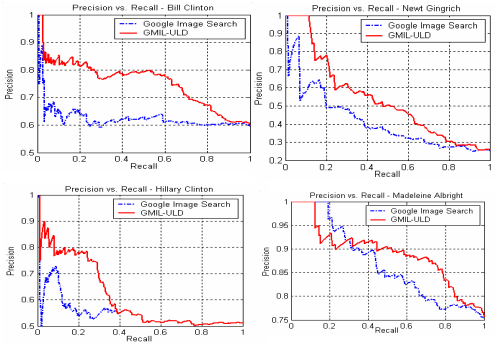


- Experiment 2: Use the ConceptValidate set for training Weather_News and Airplane Concepts
- Apply the TREC 2003 ConceptValidate ASR-based Unimodal Model Rank as one of the input of MIL-EDD
- The novel MIL-EDD+SVM Regression Models of Weather_News and Airplane Models perform better in the Concept Fusion 1 testing set.



Preliminary Detection Results based on Web Annotations





We proposed
- Generalized Multiple-Instance Learning by introducing "Quasi-Positive Bags"
- "Uncertain Labeling Density" (ULD)
- "Bag K-Means" algorithm
- "Bag Fuzzy K-Means" algorithm

We proved:
- When there is no negative bag, the DD algorithm is trying to find the centroid of the cluster by K-Means with "Bag Distance".
- The "Bag K-Means" algorithm converges

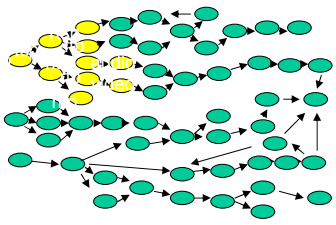
Table 1: Comparison of Average Precision

Average Precision	Bill Clinton	Newt Gingrich	Hillary Clinton	Madelein e Albright
Google Image Search	0.6250	0.4100	0.5467	0.8683
GMIL-ULD	0.7546	0.5339	0.6107	0.8899

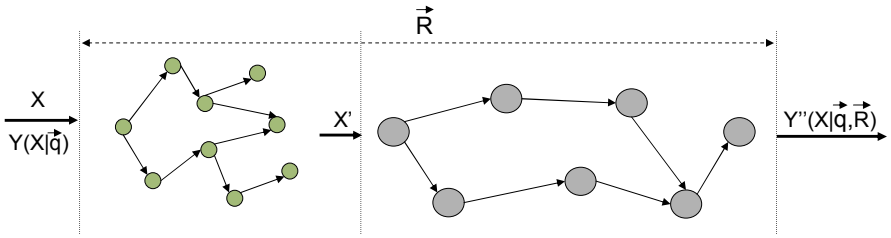
Summary of Autonomous Learning

- A cross-modality automatic learning framework
- The framework consists of Imperfect Labeling, Uncertainty Pruning, and Relevance Modeling.
- We proposed new concepts of Generalized Multiple Instance Learning and Uncertain Labeling Density.
- The learned model using our proposed approach can get good results compared to two base-line algorithms (a supervised learning algorithm based on SVM, and SVR-based confidence ranking).

III.4. Semantic Filtering



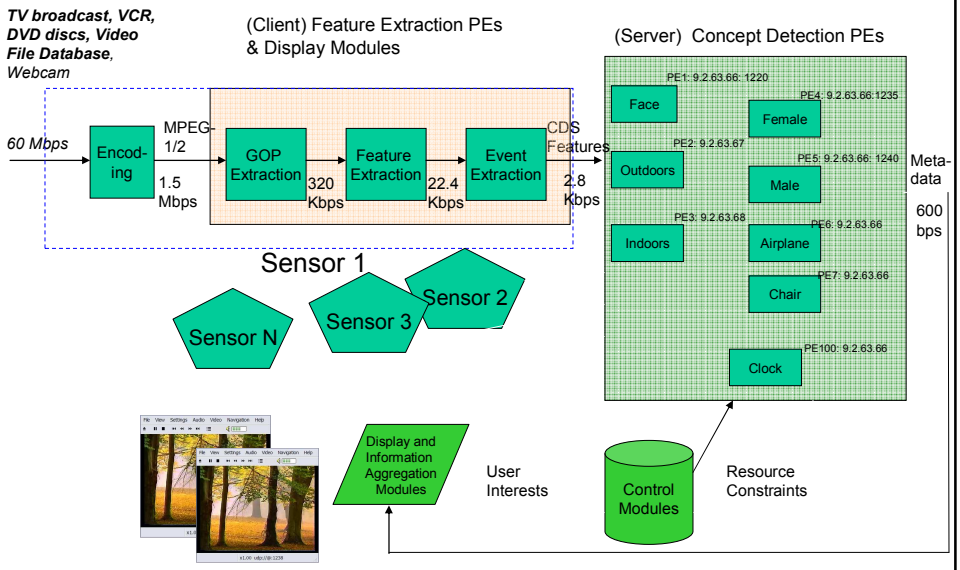
Objective



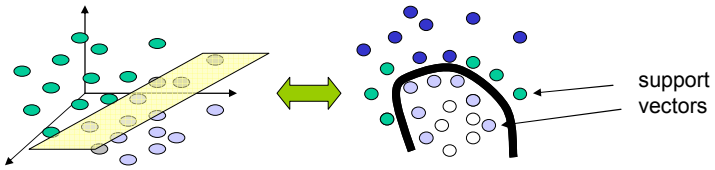
- Input data X – Queries q – Resource R
  - Y(X | q): Relevant information
  - Y'(X | q, R) ∈ Y(X | q): Achievable subset given R
- Configurable Parameters of Processing Elements to maximize relevant information:
  - Y''(X | q, R) > Y'(X | q, R),

  with resource constraint.
- Required resource-efficient algorithms for:
  - Classification, routing and filtering of signal-oriented data: (audio, video and, possibly, sensor data)

An Example of Video Semantic Filtering Framework



Complexity Analysis on SVM classifiers



- Support Vector Machine
  - Largest margin hyperplane in the projected feature space
  - With good kernel choices, all operations can be done in low-dimensional input feature space

- SVM Classifier:

$$f(x) = \sum_{i=1}^{S} a_i \cdot k(x, x_i) + b$$

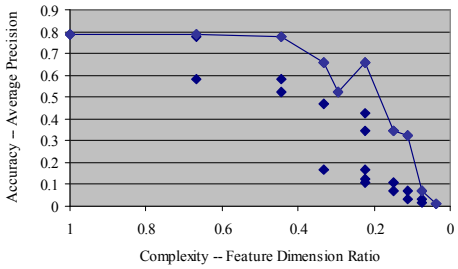
where *S* is the number of support vectors, k(.,.) is a kernel function. E.g., $k(x, x_i) = e^{-\frac{\|x - x_i\|}{r}}$

- Complexity *c*: operation (multiplication, addition) required for classification

$$c \propto S \cdot D$$

where *D* is the dimensionality of the feature vector

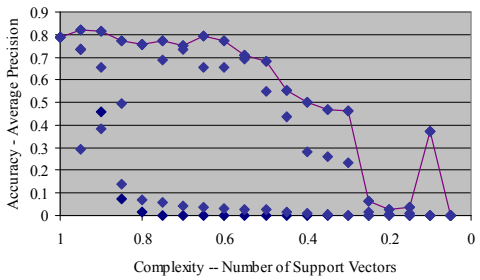
Complexity-Accuracy Curves for Adaptive PE Operations – Feature Dimension Reduction



- Experimental Results for Weather_News Detector
- Model Selection based on the Model Validation Set
- E.g., for Feature Dimension Ratio 0.22, (the best selection of features are: 3 slices, 1 color, 2 texture selections), the accuracy is decreased by 17%.

Slice	Color	Texture	Feature Ratio	AP
3	3	3	1	0.7861
3	3	2	0.666666667	0.7861
3	2	3	0.666666667	0.7757
2	3	3	0.666666667	0.5822
3	2	2	0.444444444	0.7757
2	3	2	0.444444444	0.5822
2	2	3	0.444444444	0.5235
3	3	1	0.333333333	0.4685
3	1	3	0.333333333	0.6581
1	3	3	0.333333333	0.1684
2	2	2	0.296296296	0.5235
3	2	1	0.222222222	0.427
3	1	2	0.222222222	0.6581
2	3	1	0.222222222	0.1241
2	1	3	0.222222222	0.3457
1	3	2	0.222222222	0.1684
1	2	3	0.222222222	0.1065
2	2	1	0.148148148	0.0699
2	1	2	0.148148148	0.3457
1	2	2	0.148148148	0.1065
3	1	1	0.111111111	0.3219
1	3	1	0.111111111	0.0314
1	1	3	0.111111111	0.07
2	1	1	0.074074074	0.0318
1	2	1	0.074074074	0.0173
1	1	2	0.074074074	0.07
1	1	1	0.037037037	0.0123

Complexity-Accuracy Curves for Adaptive Reduction on the Number of Support Vectors



- Proposed Novel Reduction Methods:
  - Ranked Weighting
  - P/N Cost Reduction
  - Random Selection
  - Support Vector Clustering and Centralization
- Experimental Results on Weather_News Detectors show that complexity can be at 50% for the cost of 14% decrease on accuracy

Summary of Semantic Filtering

Novelty:

- Demonstrate filtering and routing algorithms for selected formats of audio/video signals
- Further study inherent tradeoffs {accuracy – rate – complexity} between processing-efficient representations of signal-oriented data and classification techniques
- Create breakthrough technologies for highly scalable, adaptive, self-organizing, and intelligent management of high volume of media data streams

Characteristics:

- On-Demand Video Routing
- Novel processing-efficient visual classifiers:
  - The first system with 100 concept classifiers.
  - Improved accuracy performance over state-of-the-art classifiers.
  - Improved classification efficiency over state-of-the-art classifiers.
- Novel Technologies for adaptive classifiers:
  - Feature space dimension reduction.
  - Support Vector reduction.

Open Issues

- Understanding of composite higher level semantics.
- Reducing human effort in model training.
- Fast and distributed concept detection.
- Multimodality fusion.
- Noise effect in concept modeling and testing.
- Hardware concern – power management, algorithm/system complexity, scalability.

Part III References

- Dongqing Zhang and Shih-Fu Chang, "**Detecting Image Near-Duplicate by Stochastic Attributed Relational Graph Matching with Learning**", ACM Multimedia, New York, Oct. 2004.
- Lexing Xie, L. Kennedy, Shih-Fu Chang, Ajay Divakaran, Huifan Sun and Ching-Yung Lin, "**Layered Dynamic Mixture Model for Pattern Discovery in Asynchronous Multi-Modal Streams**," *IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, Philadelphia, PA, March 2005.
- Ching-Yung Lin, Xiaodan Song and Gang Wu, "**Imperfect Learning for Autonomous Concept Modeling**," *SPIE EI 2005 – Storage and Retrieval for Media Databases,* San Jose, January 2005.
- Xiaodan Song, Ching-Yung Lin and Ming-Ting Sun, "**Automatic Visual Concept Training Using Imperfect Cross-Modality Information**," Y.P. Tan, K.H. Yap and L. Wang, editors, Intelligent Multimedia Processing with Soft Computing, Springer, 2004.
- Ching-Yung Lin, Olivier Verscheure and Lisa Amini, "**Semantic Routing and Filtering for Large-Scale Video Streams Monitoring**," *IEEE Intl. Conf. on Multimedia & Expo*, Amsterdam, Netherlands, July 2005.

Acknowledgment

- IBM T. J. Watson Research Center
  Lisa Amini, Arnon Amir, Giridharan Iyengar, Milind Naphade, Apostol Natsev, Chalapathy Neti, Harriet Nock, John R. Smith, Olivier Verscheure
- Columbia University
  Shih-Fu Chang, Lexing Xie, Lyndon Kennedy, Dongqing Zhang, Winston Hsu
- University of Washington
  Ming-Ting Sun, Xiaodan Song
- University of California, Santa Barbara
  Edward Chang, Yi Wu, Gang Wu