IBM Research

# *Learning and Understanding from Multimodal Signals*

**Ching-Yung Lin**

Exploratory Stream Processing Systems
IBM T. J. Watson Research Center

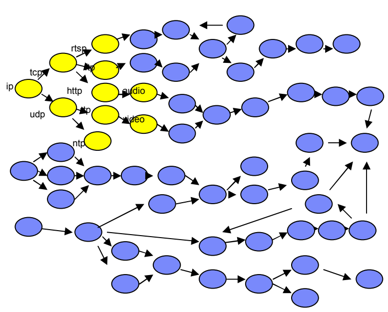
December 23, 2005





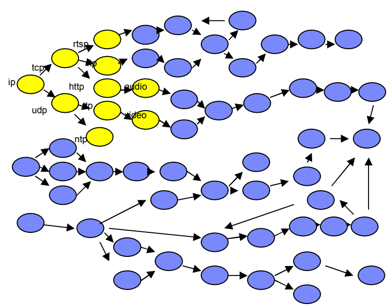
## Outline – Learning and Understanding from Multimodal Signals

- Motivation
- Understanding Multimodality Sensing Signals
- Learning from Multimodality Information
- Mining Large-Scale Multimodality Streams
- Conclusions



# Outline – Learning and Understanding from Multimodal Signals

- Motivation
- Understanding Multimodality Sensing Signals
- Learning from Multimodality Information
- Mining Large-Scale Multimodality Streams
- Conclusions









## 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**

## A picture's worth one thousand words…

Therefore, sometimes it can be used for implication!!

although sometimes it lies…







## Motivation

- **Understanding Multimodality Sensing Signals**
  - Recognize generic visual, audio, text and behavior information
- **Learning from Multimodality Information**
  - Utilize recognition result
    - Cross-Modality Learning
    - Integrated Learning
- **Mining Large-Scale Multimodality Streams**
  - Monitor information streams

# The Angle of My Views -- Research Path

- **In NTU** *(under Prof. Pei) ('91-'93)* **:**
  - Image/Video Pattern Recognition and Compression
- **In Columbia U.** *(under Prof. Anastassiou and Prof. Chang) ('96-'00)* **:**
  - Multimedia Security
- **In IBM Research** *(with Dr. Smith, Dr. Tseng, Dr. Naphade, Dr. Natsev, Dr. Iyengar, Dr. Nock, Dr. Chalapathy, etc) ('00-'04)* **:**
  - Semantic Recognition of Video Content
  - NIST TREC Video Retrieval Benchmarking
- **In IBM Research, U. of Washington and Columbia U.** *('04-'05)* **:**
  - Multimodality Signal Learning, Classification and Filtering
    - Imperfect Learning, Autonomous Learning
    - Large-Scale Stream Information Filtering
    - Modeling Human Behavior and Social Dynamics
    - Developing Smart Mobile/Wearable Multimedia Sensors



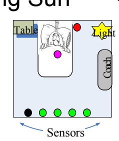




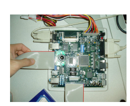
# My Social Network – Current Collaborators

Xiaodan Song
Ya-Ti Peng
Prof. Ming-Ting Sun

Univ. of Washington




Victor Sutan
Jason Cardillo

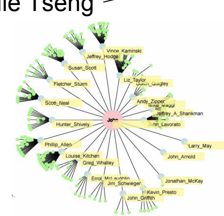


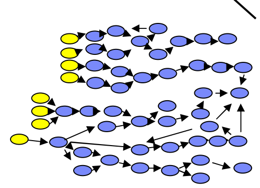
Columbia Univ.

ego

Dr. Belle Tseng

NEC



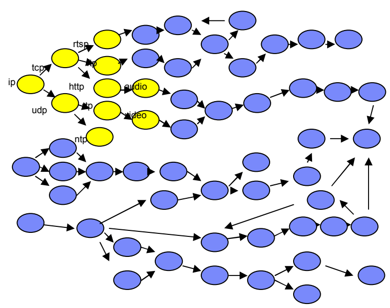


Dr. Lisa Amini
Dr. Oliver Verscheure
Dr. Anshul Sehgal
Dr. Upendra Chaudhari
Dr. Xiaohui Gu
Navneet Panda (UCSB)

IBM

# Outline – Learning and Understanding from Multimodal Signals

- Motivation
- Understanding Multimodality Sensing Signals
- Learning from Multimodality Information
- Mining Large-Scale Multimodality Streams
- Conclusions





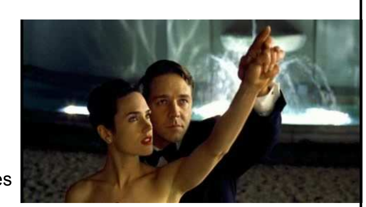




# Video Semantic Concept Detection & Mining

*A picture is worth 1000 words – which one thousand?*

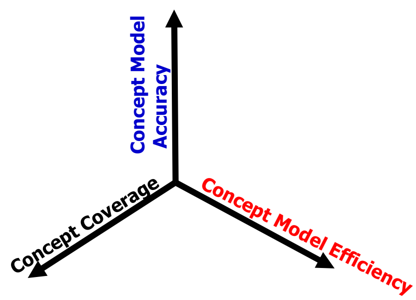
context

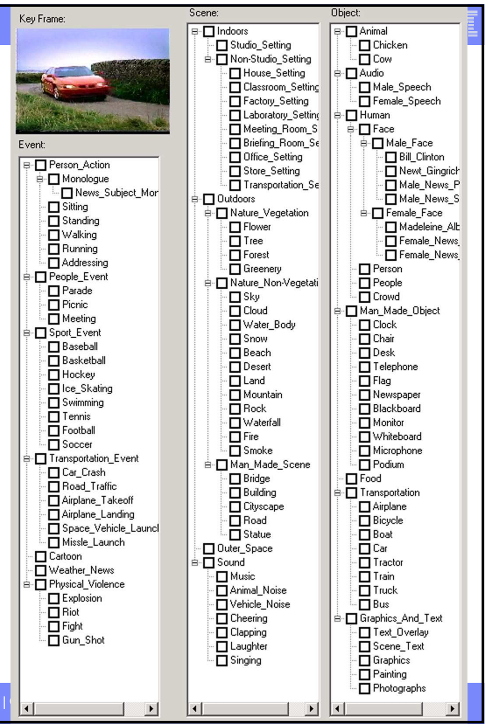


- Multimedia Concept Detection:
  - Objects:
    - Visual Objects:, Tree, Person, Hands, …
    - Audio Objects: Music, Speech, Sound, …
  - Scenes:
    - Background: Building, Outdoors, Sky
  - Relationships:
    - The (time, spatial) relationships between objects & scenes
  - Activities:
    - Holding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

# Video Semantic Concept Detection

- Design Ontology: Video => Text
- Validation Metrics:



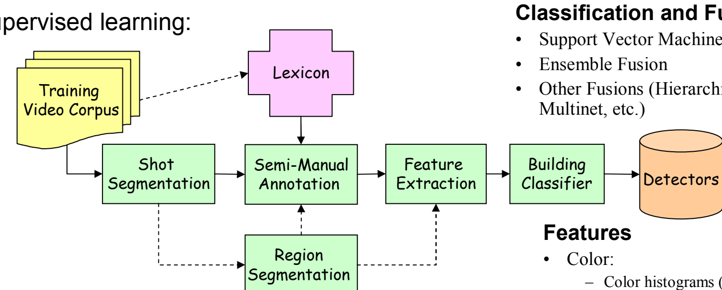



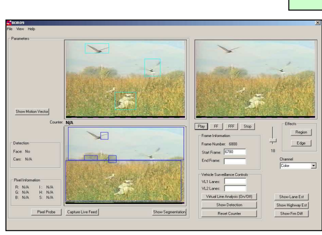





# Supervised Learning for Building Generic Concept Detectors

Supervised learning:



- **Regions**
  – Object (motion, Camera registration)
  – Background (5 lg regions / shot)

### Classification and Fusion:

- Support Vector Machines (SVM)
- Ensemble Fusion
- Other Fusions (Hierarchical, SVM, Multinet, etc.)

### Features

- Color:
  – Color histograms (72 dim, 512 dim), Auto-Correlograms (72 dim)
- Structure & Shape:
  – Edge orientation histogram (32 dim), Dudani Moment Invariants (6 dim), Aspect ratio of bounding box (1dim)
- Texture:
  – Co-occurrence texture (48 dim), Coarseness (1 dim), Contrast (1 dim), Directionality (1 dim), Wavelet (12 dim)
- Motion:
  – Motion vector histogram (6 dim)

## Supervised Learning on Automatic Speech Recognition Results of Video

- **Generating documents by collecting the words occurring symmetrically around the center of a shot ( +-2 surrounding shots)**
- **Mapping the documents from the annotations of the shots for a particular concept**
- **Removing stop-words and stemming**
- **Calculating Information gain (IG) [Yang and Pedersen, ICML 1997] for each term to select the keywords**

$$G(t) = -\sum_{i=1}^{m} P_r(c_i)\log P_r(c_i) + P_r(t)\sum_{i=1}^{m} P_r(c_i|t)\log P_r(c_i|t) + P_r(\bar{t})\sum_{i=1}^{m} P_r(c_i|\bar{t})\log P_r(c_i|\bar{t})$$

- **Ordering the keywords by the decreasing of the information gain**

Example of Topic – Related Keywords

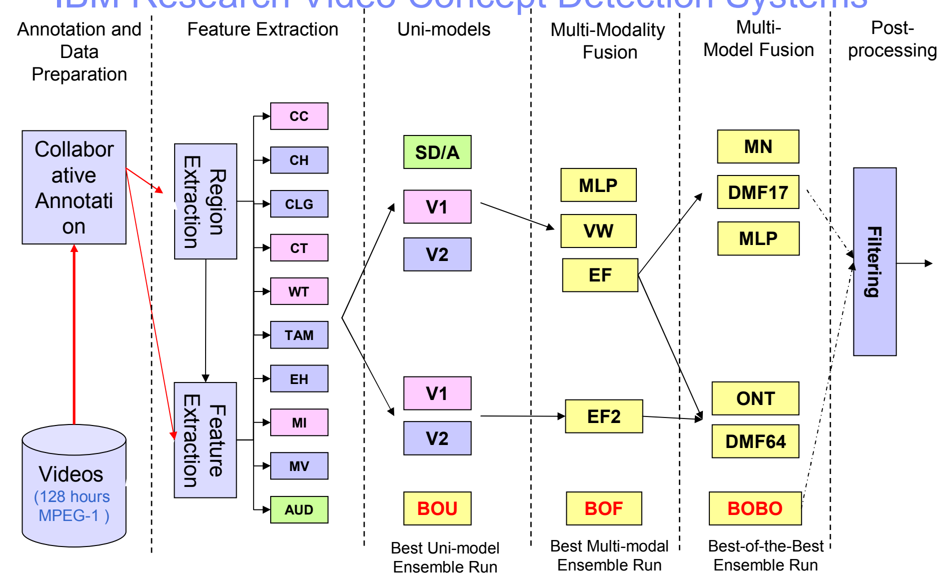
| Airplane | Animal | Building | Weather news |
|---|---|---|---|
| Plane | animal | house | rain |
| fly | Africa | president | temperature |
| ground | park | school | weather |
| military | safari | damage | forecast |
| land | nation | tornado | shower |
| weapon | land | court | storm |
| pilot | wild | city | thunderstorm |
| generator | gorilla | destroy | snow |
| airplane | wildlife | town, | lake |
| hospital | elephant | police | southeast |
| war | extinct | residence | warm |
| Iraq | breed | building | meteorologist |

- **Observations**
  - Part of the words in these two keyword lists are the same
  - Some keywords generated by supervised learning represent the context relationship instead of the lexical or semantic relationship.
    - "Africa" has a high value of information gain for Animal
    - "Airplane" and "Iraq"






## IBM Research Video Concept Detection Systems



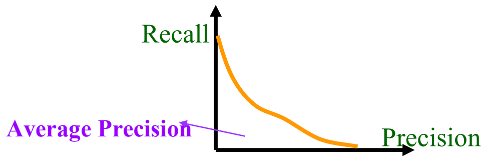
## Performance Metric -- Precision-Recall Curve and Average Precision

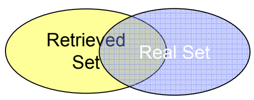
- **Example:**
  - Find shots from behind the pitcher in a baseball game as the batter is visible

*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.









## NIST TREC Video Benchmark

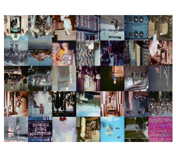
- Tasks: Shot Boundary Detection (2001 - 2005)
  Semantic Video Retrieval Query (2001 - 2005)
  Semantic Concept Detection (2002 - 2005)
  Story Boundary (2003 – 2004), Camera Motion (2005), Exploratory BBC(2005)
- Corpus: *2001* – 14 hours from NASA and BBC
  *2002* – 74 hours from Internet Movie Archive
  *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.*

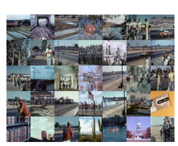
*Topic 13. Speaker talking in front of the US flag.*

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

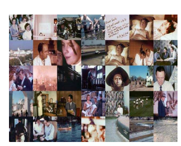
- **Semantic Concept Detector Examples:**



*"Text-overlay"*

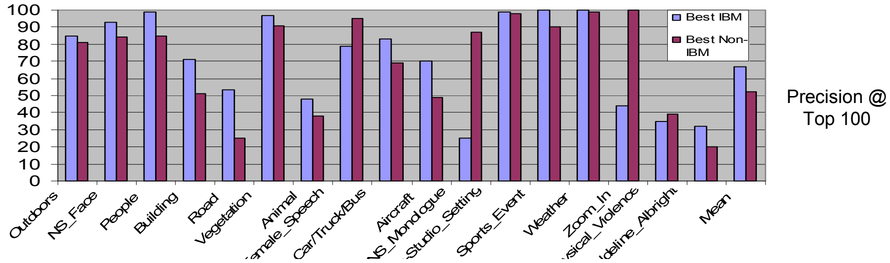


*"Outdoors"*

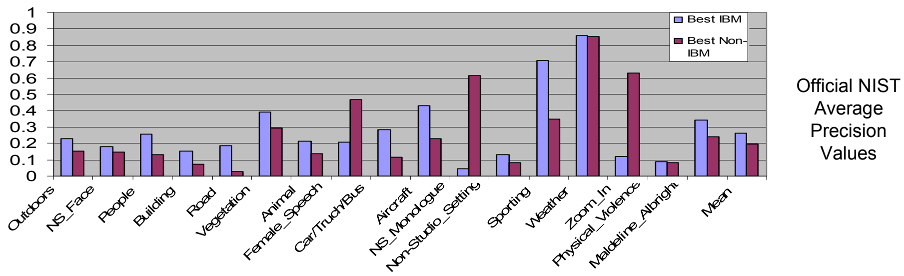


*"People"*

## Reference – Comparison of Precision at Top 100 and Average Precisions




Precision @ Top 100

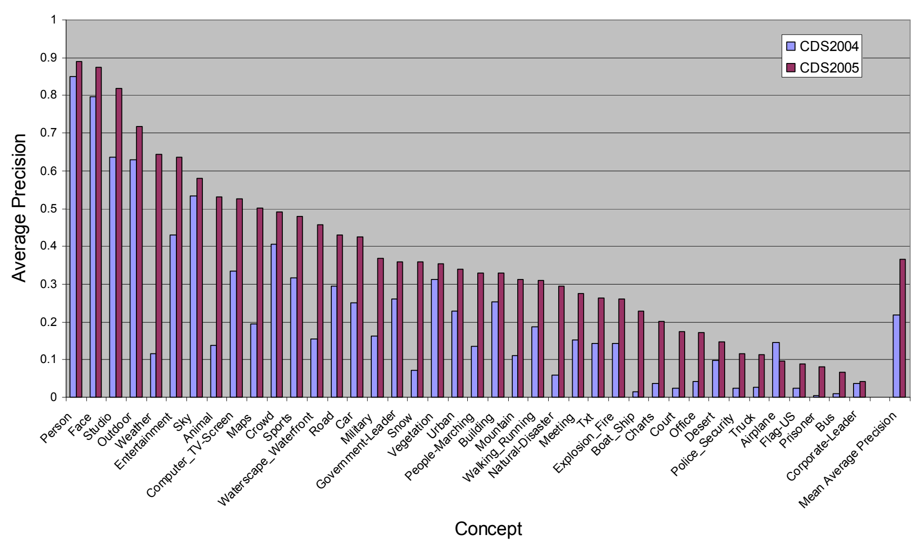



Official NIST Average Precision Values



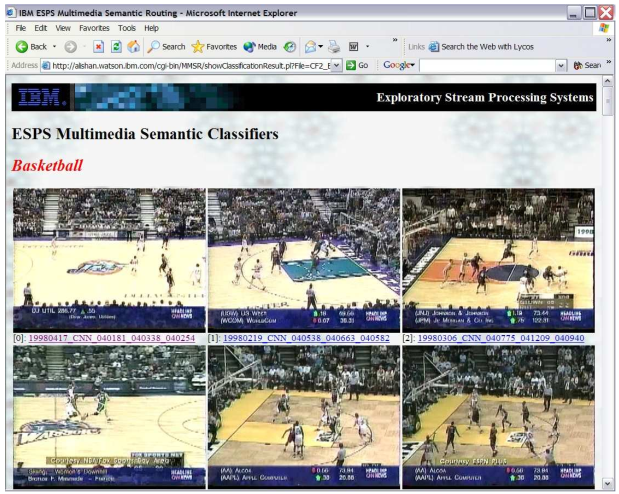



## Evaluation of the TREC 2005 vs. TREC 2004 videos



# Demo -- Novel Semantic Concept Filters

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







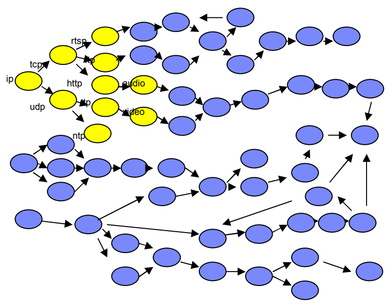


# Some Key Lessons Learnt

- **Visual information and speech information are roughly as important.**
- **The more detectors, the better.**
- **Scalability is a big issue.**

## Outline – Learning and Understanding from Multimodal Signals

- Motivation
- Understanding Multimodality Sensing Signals
- Learning from Multimodality Information
- Mining Large-Scale Multimodality Streams
- Conclusions



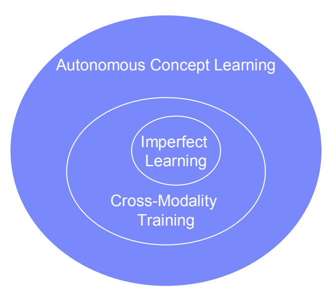






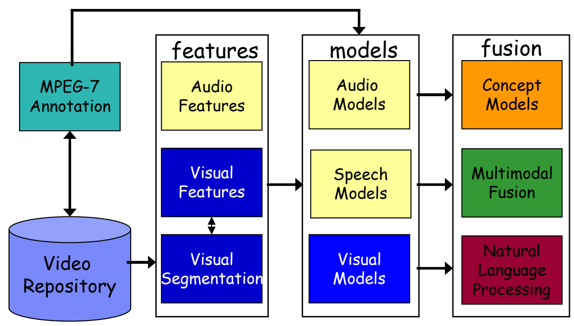
## A solution for the scalability issues at training..

- Autonomous Learning of Video Concepts through Imperfect Training Labels:
  - 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



## 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.



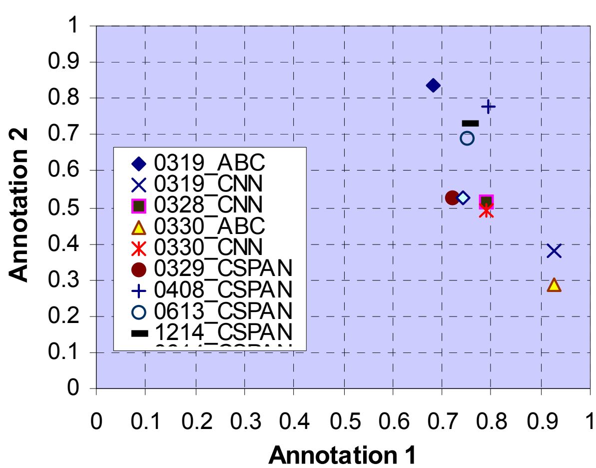




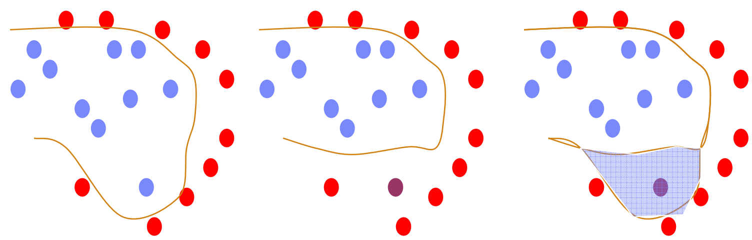


## Also -- To Err is Human

- **10 development videos are annotated by two annotators in Video Concept Annotation Forum 2003.**



## Can concept models be learned from imperfect labeling?



Example: The effect of imperfect labeling on classifiers (left -> right: perfect labeling, imperfect labeling, error classification area)






## 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.

# PAC-identifiable

- PAC-identifiable: PAC stands for *probably approximate correct*. Roughly, it tells us a class of concepts **C** (defined over an input space with examples of size **N**) is PAC learnable by a learning algorithm **L**, if for arbitrary small $\delta$ and $\varepsilon$, and for all concepts *c* in **C**, and for all distributions **D** over the input space, there is a *1-$\delta$* probability that the hypothesis *h* selected from space **H** by learning algorithm **L** is approximately correct (has error less than ε).

$$\Pr_D(\Pr_X(h(x) \neq c(x)) \geq \varepsilon) \leq \delta$$

- Based on the PAC learnability, assume we have *m* independent examples. Then, for a given hypothesis, the probability that m examples have not been misclassified is $(1-e)^m$ which we want to be less than $\delta$. In other words, we want $(1-e)^m <= \delta$. Since for any $0 <= x < 1$, $(1-x) <= e^{-x}$ , we then have:

$$m \geq \frac{1}{\varepsilon}\ln(\frac{1}{\delta})$$






# Sample Size v.s. VC dimension

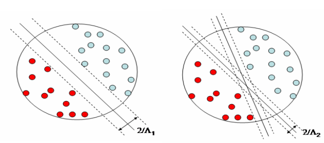
- **Theorem 2** Let **C** be a nontrivial, well-behaved concept class. If the VC dimension of **C** is *d*, where *d* < ∞, then for *0* < *e* < *1* and

$$m \geq \max(\frac{4}{\varepsilon}\log_2\frac{2}{\delta}, \frac{8d}{\varepsilon}\log_2\frac{13}{\varepsilon})$$

any consistent function A: Sc**C** is a learning function for **C,** and, for *0* < *e* < *1/2*, m has to be larger than or equal to a lower bound,

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

For any *m* smaller than the lower bound, there is no function A: Sc**H**, for any hypothesis space **H**, is a learning function for **C**. The sample space of **C**, denoted SC, is the set of all m-samples over all c in **C**.




$$d \leq \min(\Lambda^2 R^2 + 1, n + 1)$$

Example: VC dimension of SVM

# Noisy Samples

- **Theorem 4** 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

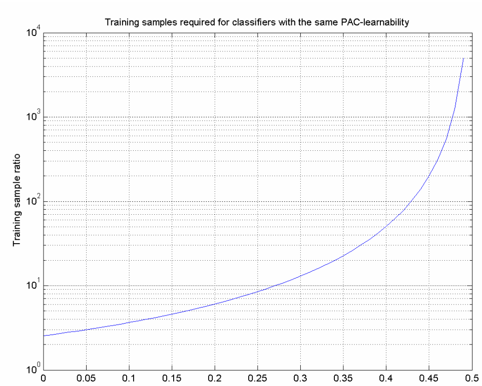
$$r_\eta = (1-\exp(-\tfrac{1}{2}(1-2\eta)^2))^{-1}$$



# 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.

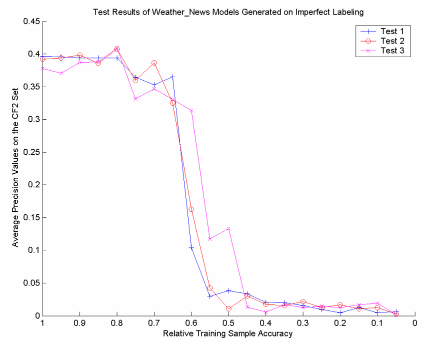


IBM Research

# Experiments -- example:

- We simulated annotation noises by randomly change the positive examples in manual annotations to negatives.
- Because *perfect* annotation is not available, accuracy is shown as a relative ratio to the manual annotations in [10].
- In this figure, we see the model accuracy is not significantly affected for small noises.
- A similar drop on the training examples is observed at around 60% - 70% of annotation accuracy (i.e., 30% - 40% of missing annotations).





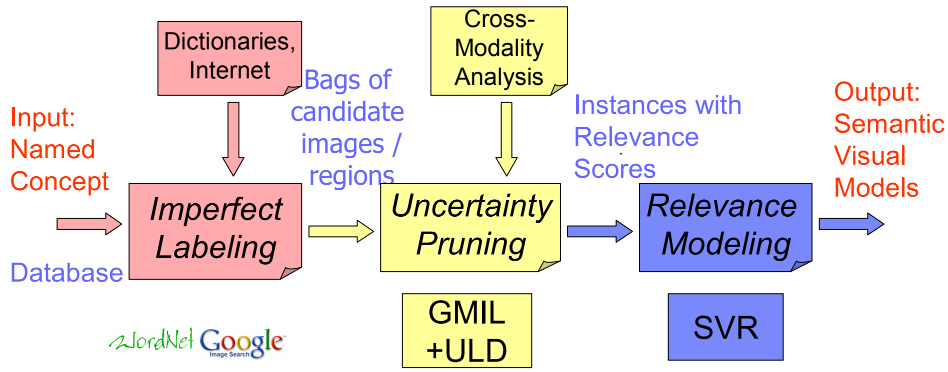


IBM Research

# Challenges to Realize Autonomous Learning

- **When there are no human supervision**
  - Associate the images/videos with semantic labels
  - *Our solution:*
  - Imperfect Labeling by unsupervised text-based search
  - Find the "right" object; Reduce the influence of mislabeling
  - *Our solution:*
  - Uncertainty Pruning by Generalized Multiple Instance Learning (GMIL) + Uncertainty Labeling Density

## Autonomous Learning Scheme




- Uncertainty Pruning:
  - Generalized Multiple Instance Learning (GMIL)
  - Uncertainty Labeling Density (ULD)
- Relevance Modeling:
  - Support Vector Regression (SVR)





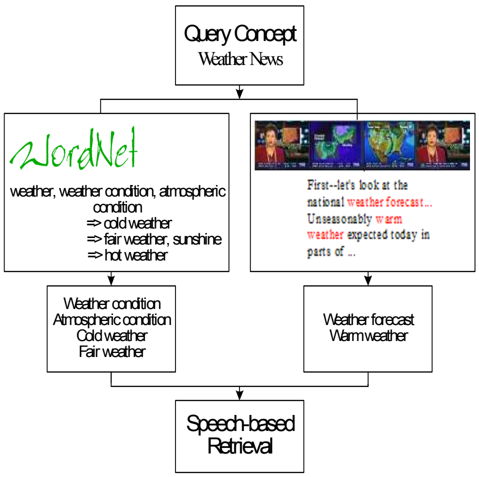

## Imperfect Labeling

*"First – let's look at the national weather forecast… Unseasonably warm weather expected today …"*

**Goal: Leverage cross-modality correlation to associate the image/video with labels**

## Other Issues: 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

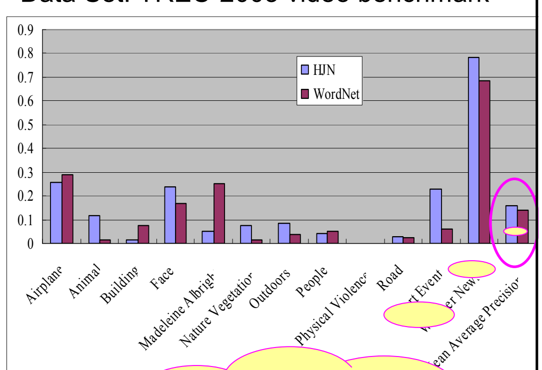





## Imperfect Labeling (cont.)

- **For video sequences:**
  - Query expansion using WordNet
  - Rank video shots based on Okapi algorithm
  - Ranking → imperfect labels
- **For web images**
  - Extract Labels from Existing Search Engine
    - *i.e.*, Google image Search

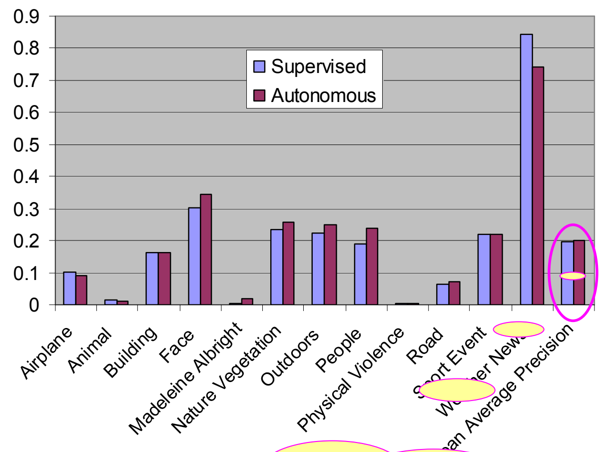
Data Set: TREC-2003 video benchmark




Our Imperfect Labeling is comparable with supervised learning (HJN)

# Experimental Results

- **Training data**
  - "ConceptTraining" in the NIST Video TREC 2003 corpus (78 video sequences with 28,055 keyframes).
- **Testing data**
  - "ConceptFusion1" (13 news videos with 5,037 keyframes)




Our Autonomous Learning is comparable with supervised learning(SVM)





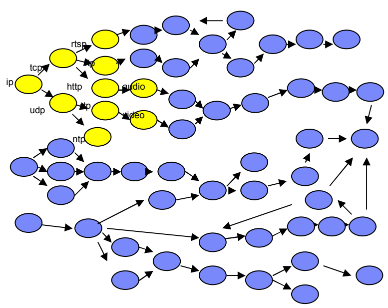
# Experimental results

Dataset: Google Image Search Results

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

## Outline – Learning and Understanding from Multimodal Signals

- Motivation
- Understanding Multimodality Sensing Signals
- Learning from Multimodality Information
- Mining Large-Scale Multimodality Streams
- Conclusions





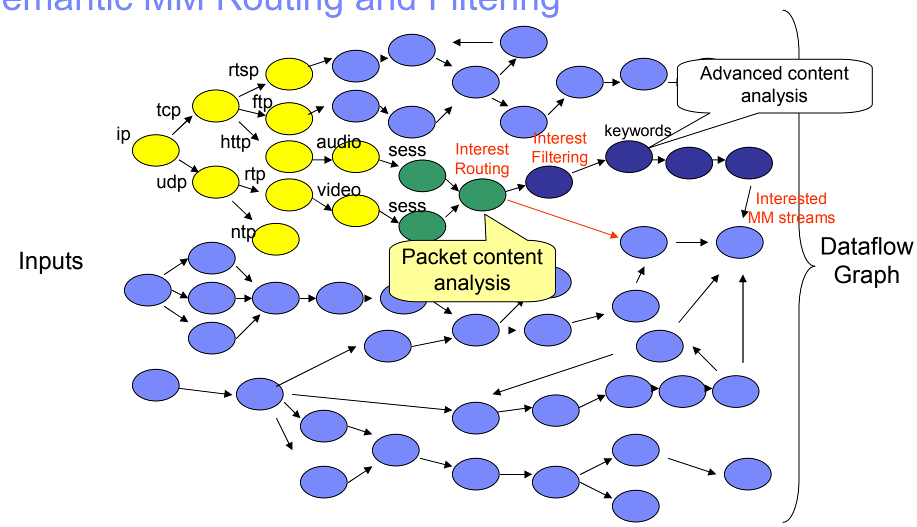




## What is the “large-scale” we are considering?

- 10Gbit/s Continuous Feed Coming into System
- Types of Data
  - Speech, text, moving images, still images, coded application data, machine-to-machine binary communication
- System Mechanisms
  - Telephony: 9.6Gbit/sec (including VoIP)
  - Internet
    - ✓ Email: 250Mbit/sec (about 500 pieces per second)
    - ✓ Dynamic web pages: 50Mbit/sec
    - ✓ Instant Messaging: 200Kbit/sec
    - ✓ Static web pages: 100Kbit/sec
    - ✓ Transactional data: TBD
  - TV: 40Mb/sec (equivalent to about 10 stations)
  - Radio: 2Mb/sec (equivalent to about 20 stations)

## Semantic MM Routing and Filtering



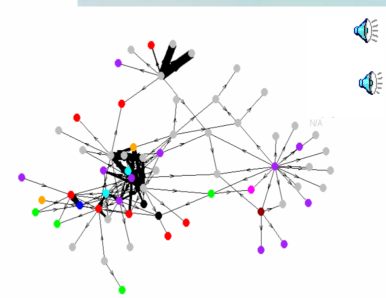

| per PE rates | 200-500MB/s | ~100MB/s | 10 MB/s |
|---|---|---|---|



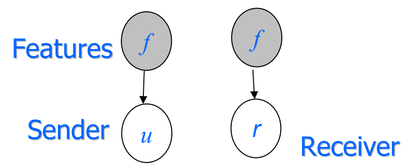



## Activities Monitoring through Public Radio Signals

- **Objective:**
  - Early understanding and monitoring on what's happening
- **Technical Challenges:**
  - Monitor many noisy channels simultaneously Understanding content
  - Understanding speaker
  - Understanding the relationships of speakers
  - How information being flowed and how to organize



## Methods for Speaker Identification

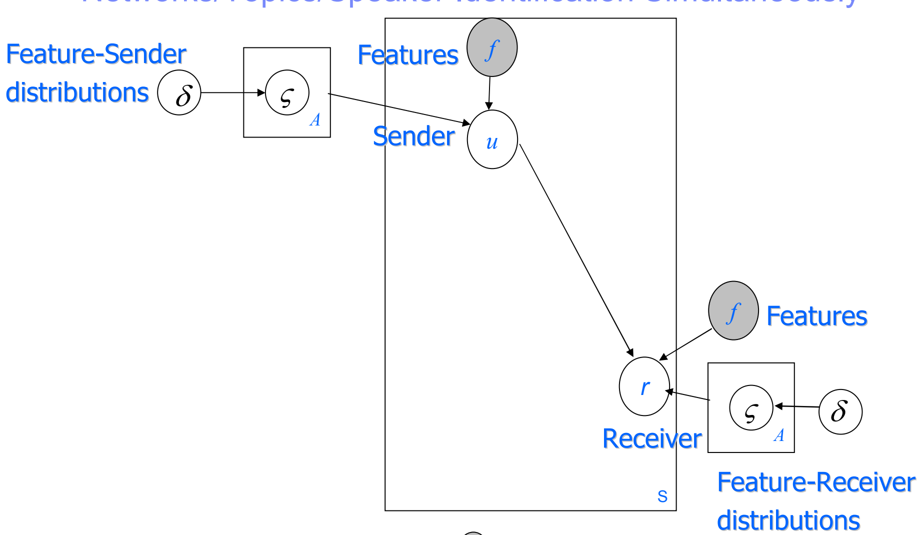



: observations



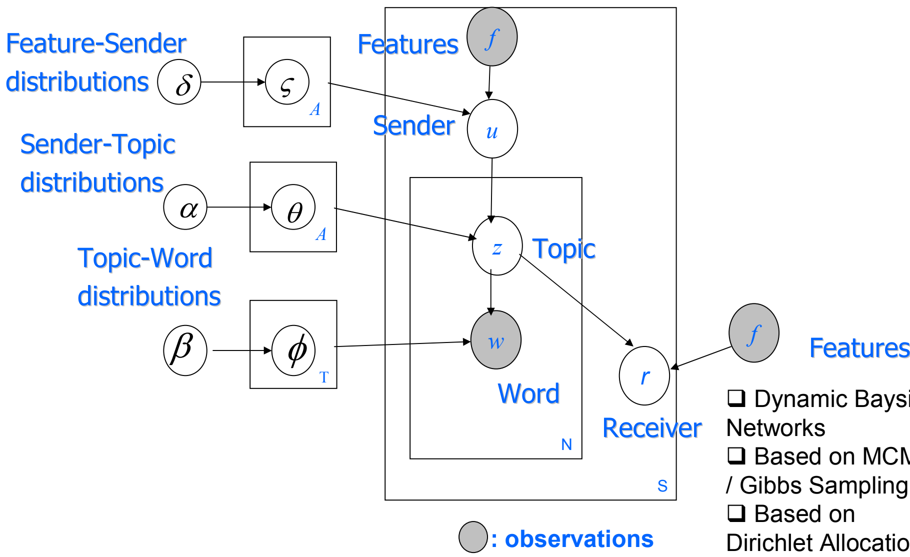



## Novel RadioDIG Algorithm – Building Dynamic Social Networks/Topics/Speaker Identification Simultaneously




: observations

## Novel RadioDIG Algorithm – Building Dynamic Social Networks/Topics/Speaker Identification Simultaneously



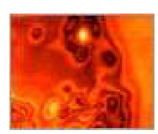

- ❑ Dynamic Baysian Networks
- ❑ Based on MCMC / Gibbs Sampling
- ❑ Based on Dirichlet Allocation



## Human – a complex multimodality subject/object

- ❑ "*Human and Social Dynamics (HSD)*" is identified as one of the five NSF key priorities among:
  - Nanoscale Science and Engineering
  - Biocomplexity in the Environment
  - Human and Social Dynamics
  - Mathematical Sciences
  - Cyberinfrastructure



  - (http://www.nsf.gov/news/priority_areas/)

# Great and Extensive Potential Impacts on Social Computing

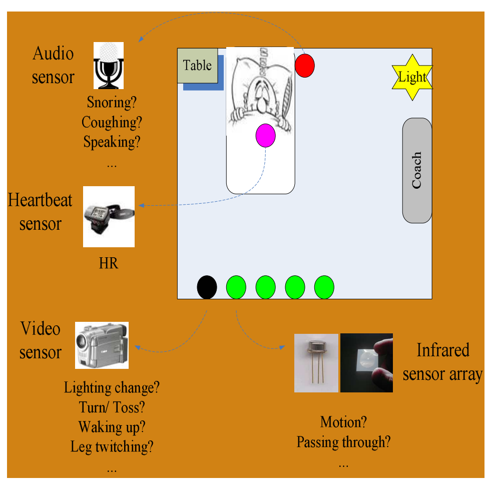
- A deeper understanding of people's routines and interactions will have significant impacts on…
  - Designing public spaces and office environments, developing computer collaboration, recommendation, and assistance tools.
  - Provide personalized and dynamic life pattern log, service/guidance
  - Better anticipation of human and social changes (ex. causes & responses)
  - Improvement of human interactions/ collaborations
  - Prediction of information flow and efficient information spread
  - Help risk assessment and decision-making





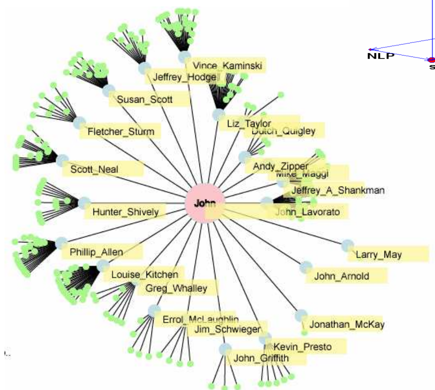

# Low-cost Multimodality Sensors for Sleep Situation Inference / Logging

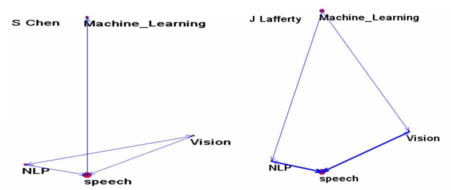
- **Understand human night-time activity – *Sleep***
- **What we have done:**
  - Using visual, audio, heartbeat, infrared sensors to monitor a person's sleep patterns
  - Measurement of sleep quality
  - Logging and retrieval of sleep situation
- **What we are going to do:**
  - Early detection and long term monitoring of sleep related diseases

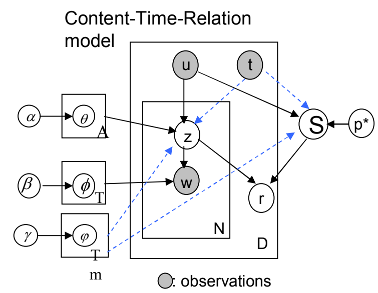


# Modeling Dynamic Human Behavior and Properties

– Establish personal *CommunityNet*
– Establish personal *ExpertiseNet*
– Establish personal *InterestNet*





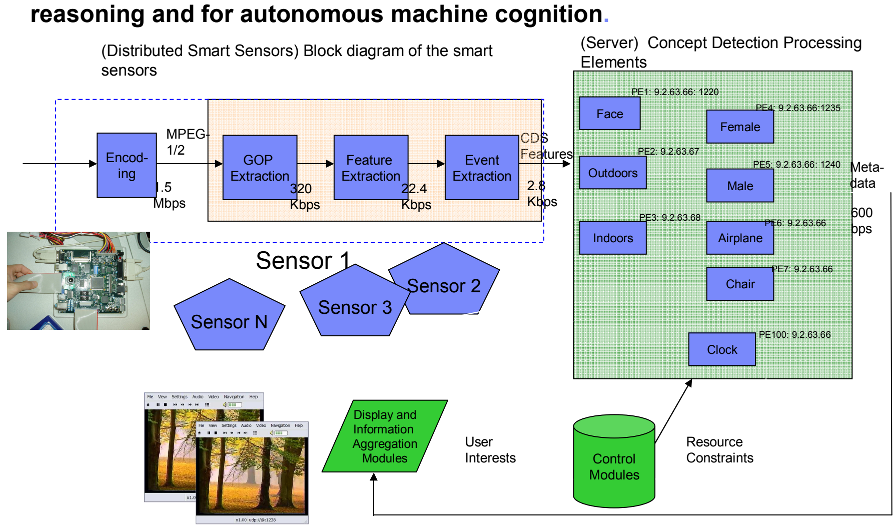






# Smart Semantic Video Cameras

- **Objectives: Build smart video sensors to execute real-time visual reasoning and for autonomous machine cognition.**



## Smart Semantic Video Cameras


*Will this become possible ??*






## Questions?

- **Thanks for attending!!**
- **http://www.research.ibm.com/people/c/cylin**
- **chingyung@us.ibm.com**