Recent Advances and Open Issues of Digital Image/Video Search

Prof. Shih-Fu Chang

Digital Video and Multimedia Lab

Columbia University

www.ee.columbia.edu/dvmm

June 2007

Acknowledgement

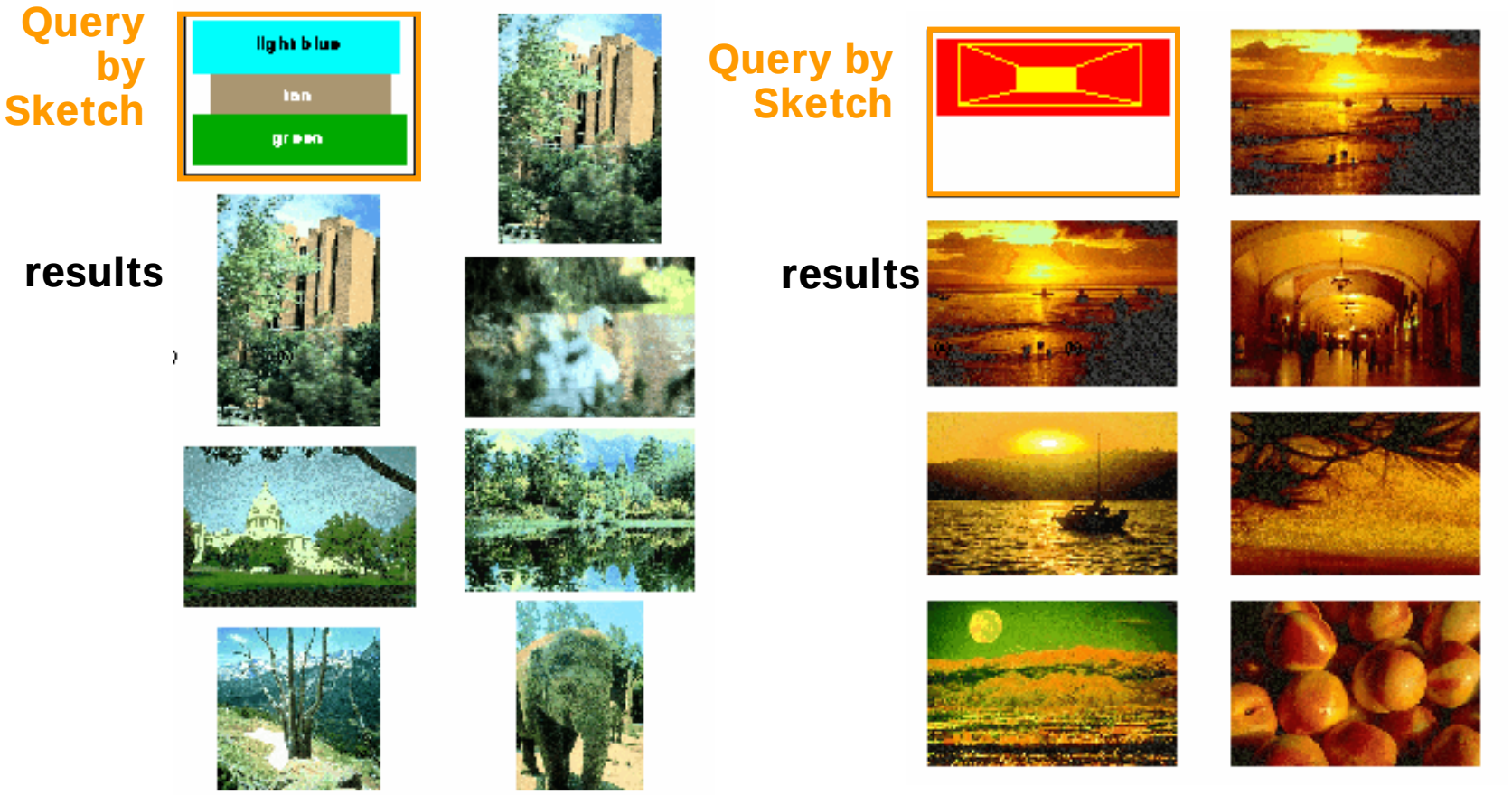
- Columbia University
  - Winston Hsu, Wei Jiang, Lyndon Kennedy, Akira Yanagawa, Eric Zavesky, Dongqing Zhang
- Kodak Research
  - Alex Loui, Jiebo Luo
- IBM Research
  - Shahram Ebadollahi, Milind Naphade, Paul Natsev, John R. Smith, Lexing Xie
- CMU
  - Alex Hauptmann

Need of Image/Video Search

- Explosive growth of online image/video data, personal media, broadcast news videos, etc.
- 5 billion images on the Web, 31 million hours of TV programs each year
- Successful services like Youtube and Flickr
- Image/video search exciting opportunity

Seeking the image search tools
-- an active research area over decade

IBM QBIC '95, Columbia VisualSEEk '96



However... User Expectation in Practice

> "…type in a few words at most, then expect the engine to bring back the perfect results. More than 95 percent of us never use the advanced search features most engines include, …" – *The Search*, J. Battelle, 2003

- Keyword search is the primary search method.
- User input is often limited, thus
  - Difficult to get detailed visual descriptions
  - Difficult to get user feedback for refined search

Google Zeitgeist publishes top keywords monthly

Google Image Queries: April 2007

Newsmakers	Getaways
1. virginia tech	1. hawaii
2. knut	2. dubai
3. yuri gagarin	3. mexico
4. shaha riza	4. chelsea
5. kurt vonnegut	5. london

Greece - Top Gaining Queries: April 2007

1. notis	6. χαρτης θεσσαλονικης	11. chelsea fc
2. holly valance	7. ΜΕΤΑΦΡΑΣΗ	12. manchester united
3. shrek	8. land rover	13. Μύκονος
4. melodia	9. anastasia	14. πλουταρχος
5. erreway	10. mega tv	15. swimming

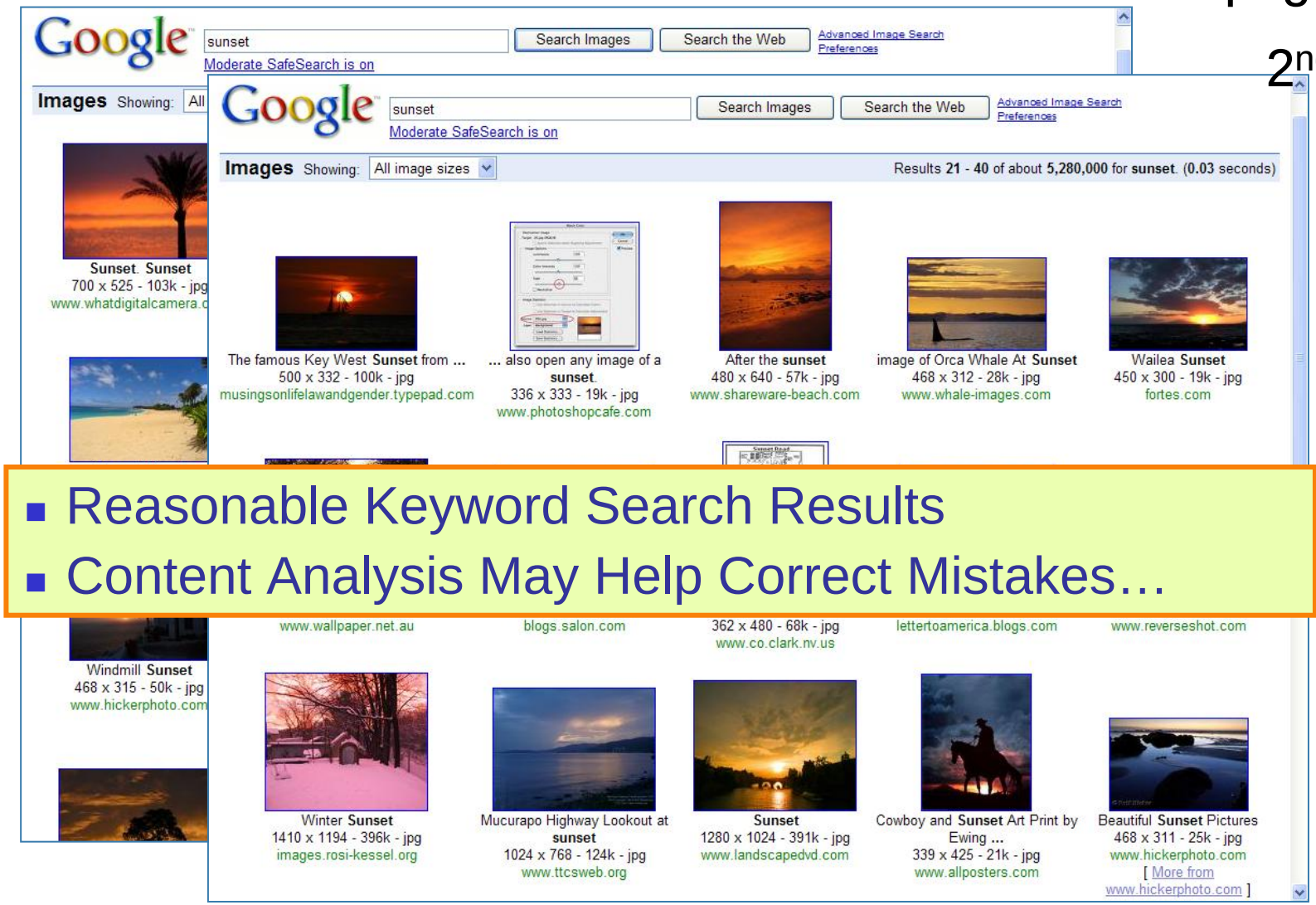
Back to top

Examples of Keyword Image Search

query: “sunset”

1st page

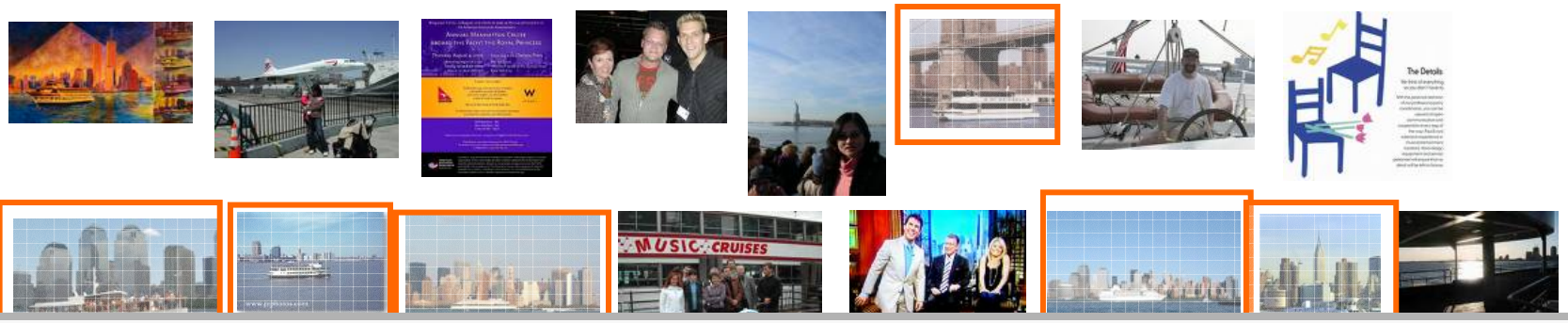
2nd page



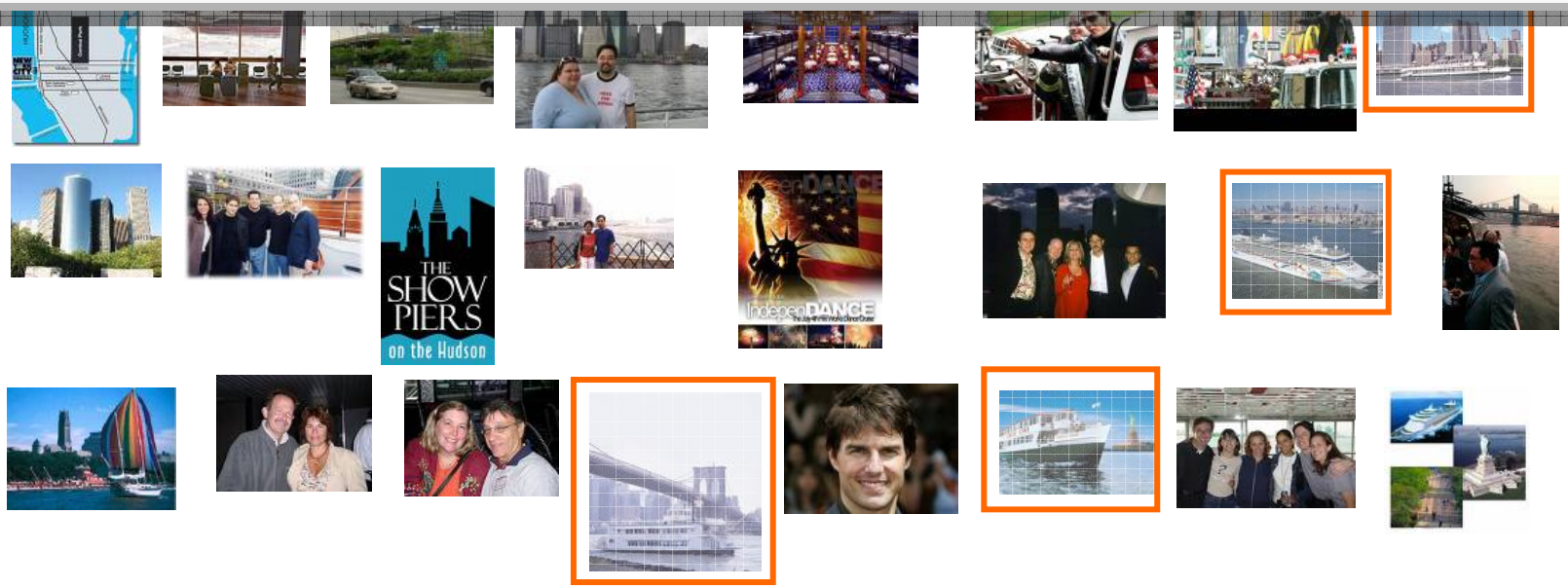
- Reasonable Keyword Search Results
- Content Analysis May Help Correct Mistakes…

Example Search

- Text Query on Google: *“Manhattan Cruise”*



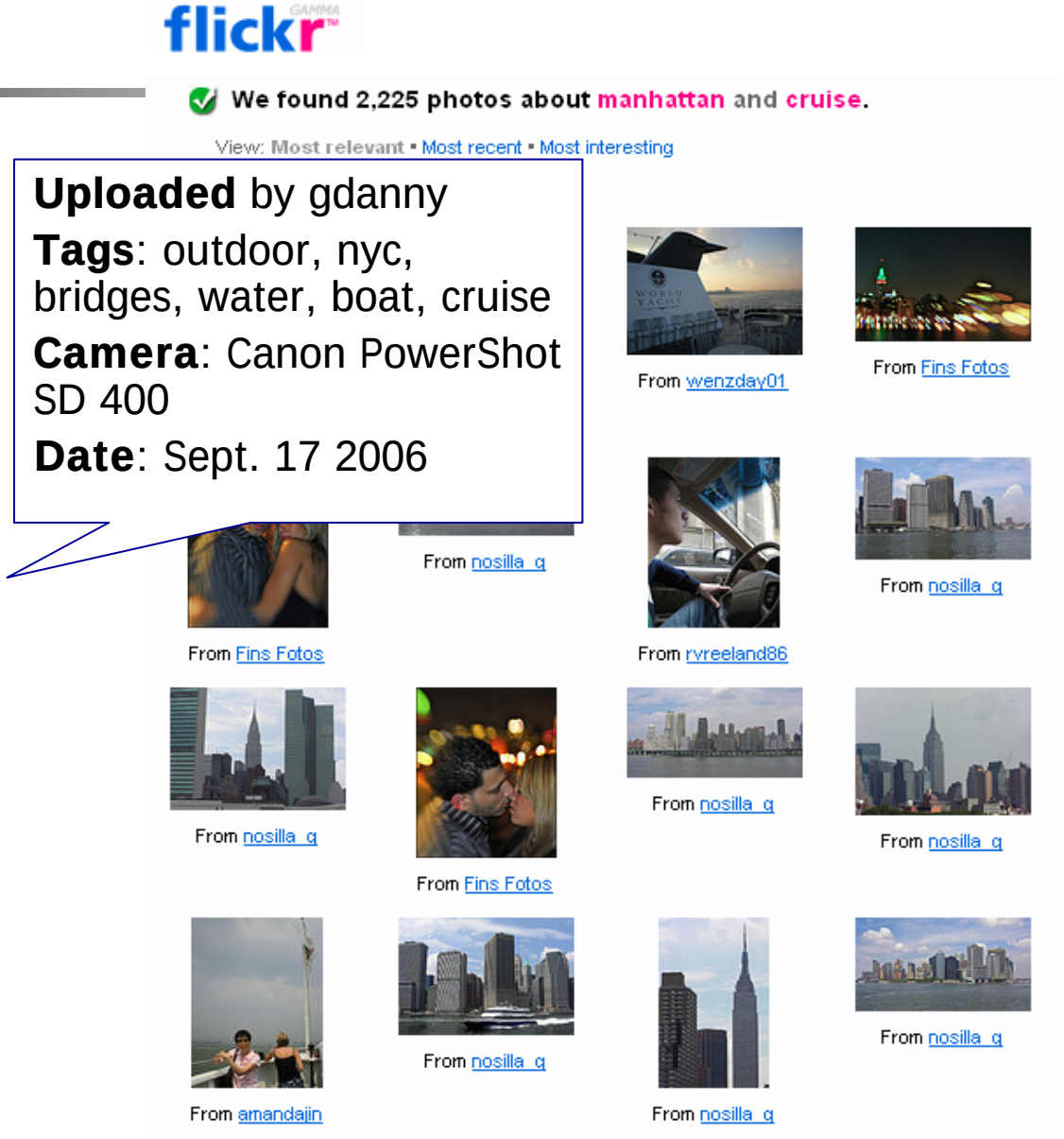
- Image content analysis may help refine results



How about Social-Net Tagging?

- Yahoo-flickr millions of users,

Social tags are often subjective and incomplete.



Insufficient Precision of Social Tags

- Test New York City landmark labels

	precision
Bronx-Whitestone Br.	1.00
Brooklyn Br.	0.38
Chrysler Building	0.65
Columbia University	0.30
Empire State Building	0.18
Flatiron Building	0.70
George Washington Br.	0.48
Grand Central	0.37
Times Square	0.56
Verrazano Narrows Br.	0.66
World Trade Center	0.13

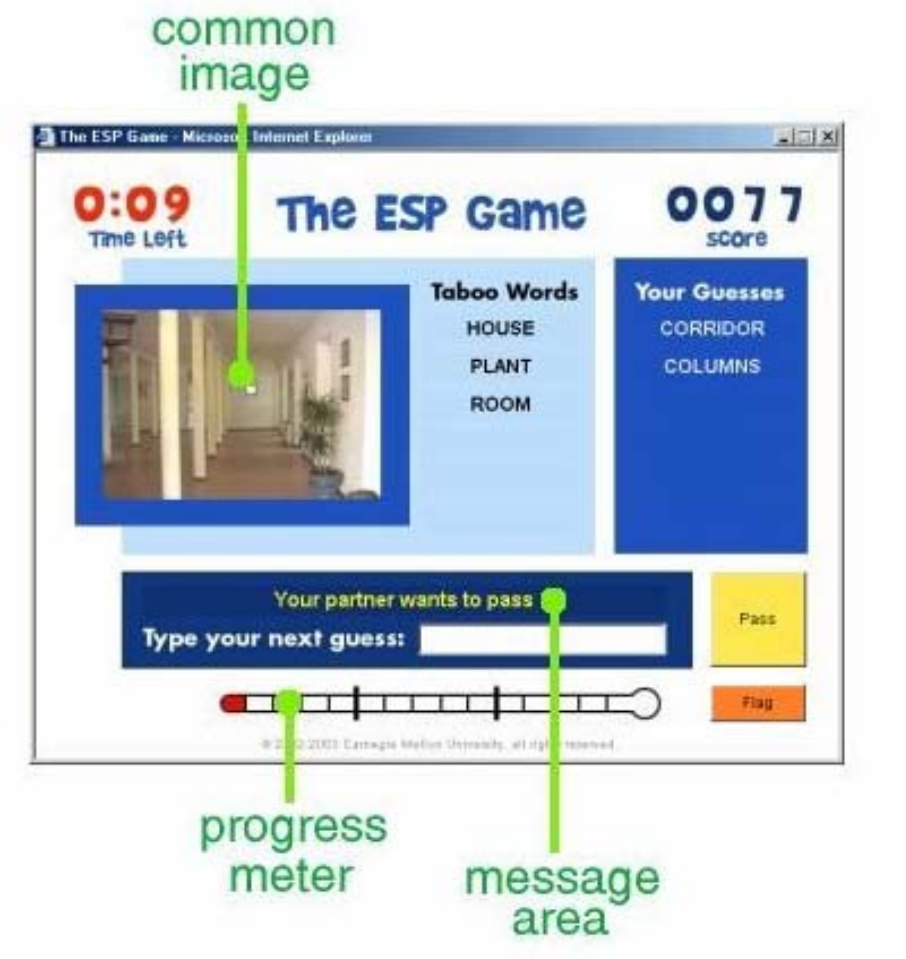
Many tags from social networks are of low precision

(due to batch uploading?)

An Interesting Paradigm:
Image Tagging via Game Playing

(Von Ahn & Dabbish, CHI 04)

- Used in *Goggle Image Labeler*
- Use competitive games to motivate users
- Has attracted many participants for free!
  - Some users spent hours in a day
- Claim the potential of annotating the whole Web in just few months!
  - 5 Billion images



My Personal Experience

512 x 768 pixels
matched: **talk**
www.indaba-southafrica.co.za
Partner's guesses:
speaker, presentation, stage, stand, the dti, conference, person, poster, talk

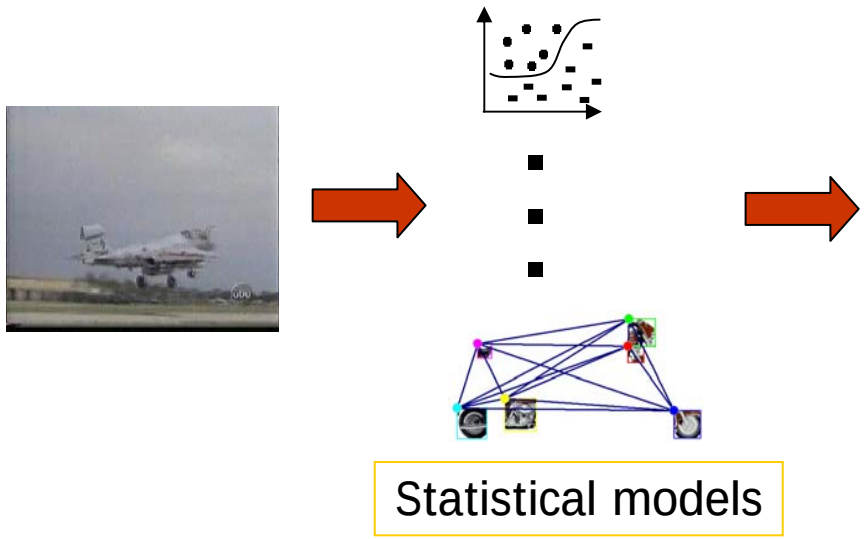
720 x 480 pixels
matched: **space**
www.goodcowfilms.com
Partner's guesses:
earth, blue, video game, satellite, space

1024 x 768 pixels
matched: **heart**
nba-time.skyblog.com
Partner's guesses:
heart, basket, basket ball, sport

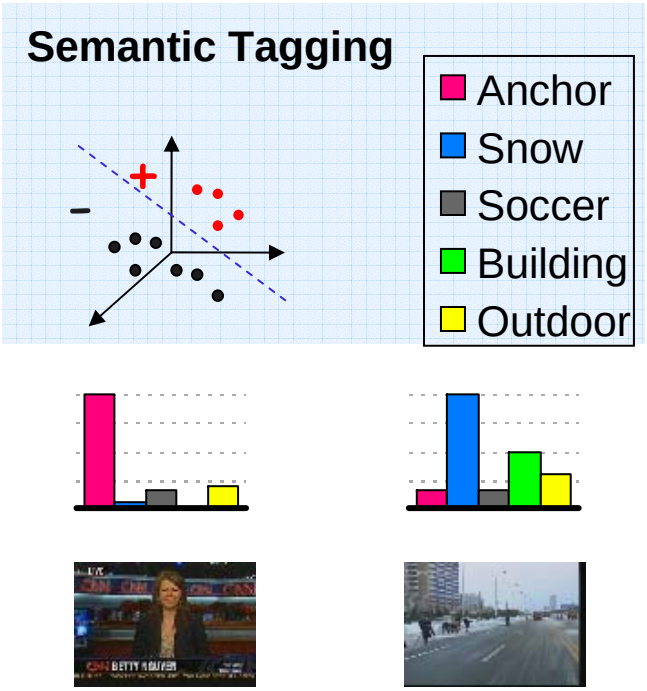
- Affected by personal background and proficiency
- Matched words tend to be obvious ones
  - Scoring system encourages guessing partner's thoughts, rather than indexing image content
  - Players sometimes just perform manual OCR
- An interesting paradigm, but perhaps not a complete solution

Opportunity for Content Analysis: Large-Scale Auto. Image Tagging Framework

- Audio-visual features
- Surrounding text
- SVM or graph models
- Context fusion



- Rich semantic description based on content analysis



Large-Scale Concept Detectors Publicly Available

- Columbia374
  - 374 baseline detectors for LSCOM multimedia ontology
- MediaMill
  - 491 concept detectors for LSCOM and MediaMill 101 Lexicons
- IBM MARVEL Search System
  - Trials with BBC, CNN
  - Real-time standalone detectors from IBM AlphaWorks
- Others ...

What Concept to Detect?

- One effort: Large Scale Concept Ontology for Multimedia (LSCOM)
  - Joint effort by news/intelligence analysts, librarians, researchers
  - Broadcast News Domain
  - Selection Criteria
    - useful, detectable, observable
  - 834 concepts defined, 449 concepts annotated
  - Labeled over 61,000 shots of TRECVID 2005 data set
    - 33 Million judgments collected, 100 person-month labor
  - Download by 170+ groups so far
    - http://www.ee.columbia.edu/dvmm/lscom/

LSCOM data set downloaded by 170+ groups

- Yahoo! Research
- Intel
- AT&T
- FXPAL
- University of Amsterdam
- Oxford University
- Nanyang Technological University, Singapore
- National Taiwan University
- Tsinghua University
- KDDI, Japan
- Dublin City University, Ireland
- University of Central Florida
- University of Texas, Austin
- UC Berkeley
- Others …

Example LSCOM Concepts (449)

- Event/Activity (56 - 13%)
  - Airplane taking off, car crash, explosion, etc
- People (113 - 25%)
  - Person, male/female, firefighter, etc
- Location (89 - 20%)
  - Cityscape, hospital, airfield, etc
- Object (135 - 30%)
  - Vehicle, map, tank, power plant, etc
- Scene (49 - 10%)
  - Vegetation, urban, interview, etc
- Program (7 - 2%)
  - Entertainment, weather, finance, etc

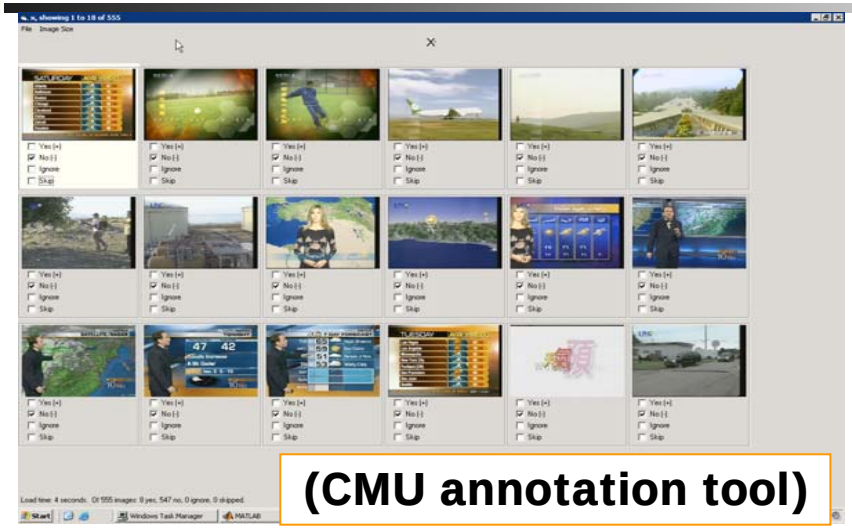
Another Effort : Consumer Video Ontology
(Kodak-Columbia, 2007)

- Activity (6)
- Occasion (16)
- Scene (15)
- Object (25)
- People (11)
- Sound (14)
- Camera Motion (5)
- Object Motion (3)
- Social (4)

Activity:
Occasion :
Scene:
Object:
People:
Sound:
Camera Motion:
Object Motion:
Social:

Lexicon and annotation over 1300 consumer videos planned to be released in ACM MIR 2007

How to Obtain High-Quality Annotations?

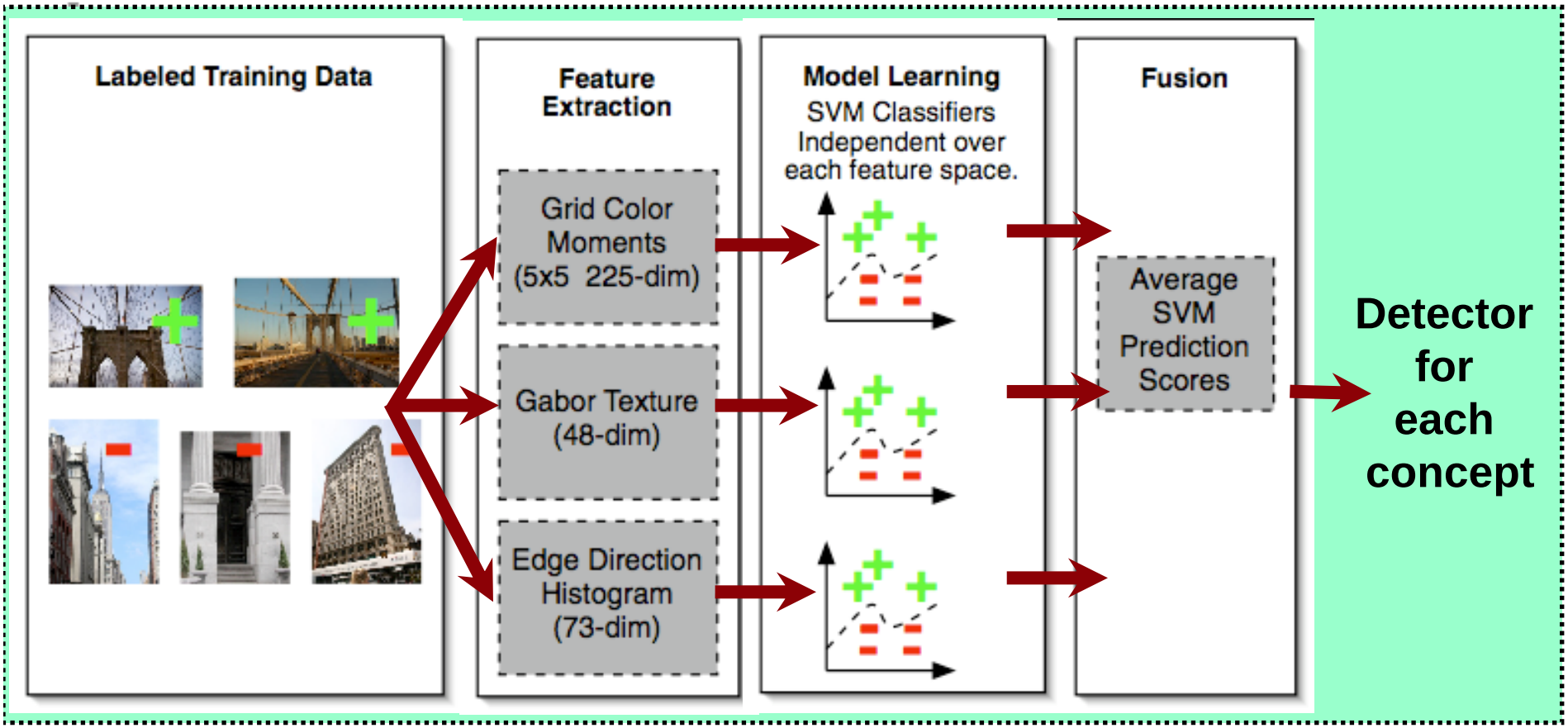


(CMU annotation tool)

IBM EVA tool

- Label only one concept at a time
  - Instead of open text simultaneous labeling
  - Found to be more accurate
  - And faster!
- But hard to scale up
  - Throughput: 1-3 sec/image
  - A laborious process (100 person-month) for 33 M labels
  - No user incentive

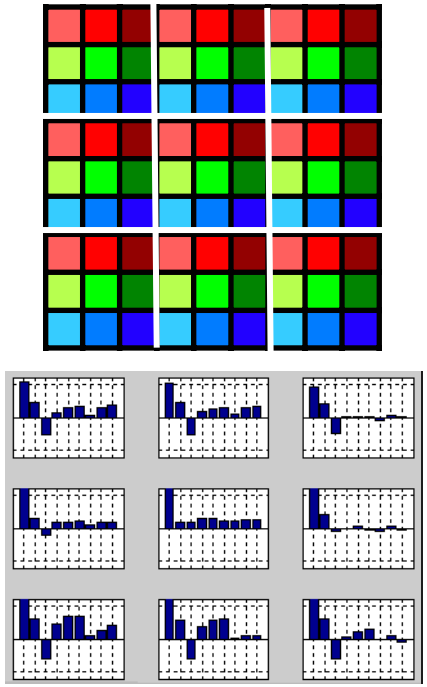
Building Image Classifiers – starting from the basics



- General for all concepts, easy to implement
- Late fusion better than early fusion
- 374 baseline detectors (*Columbia 374*) released

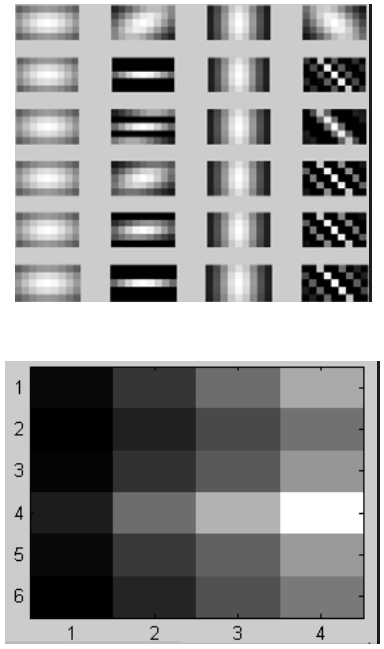
Examples of Basic Image Features

grid layout + color moment



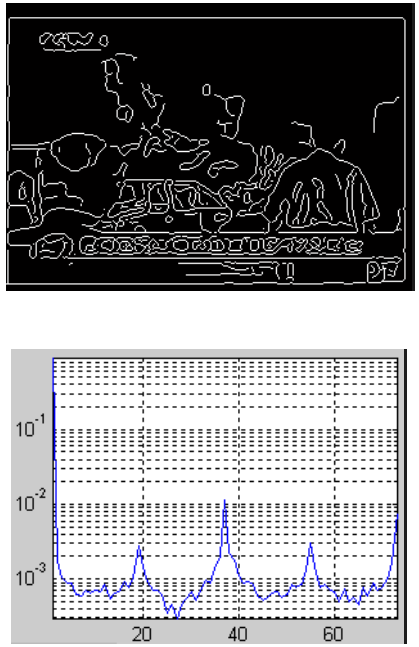
225 dimensions

Gabor texture

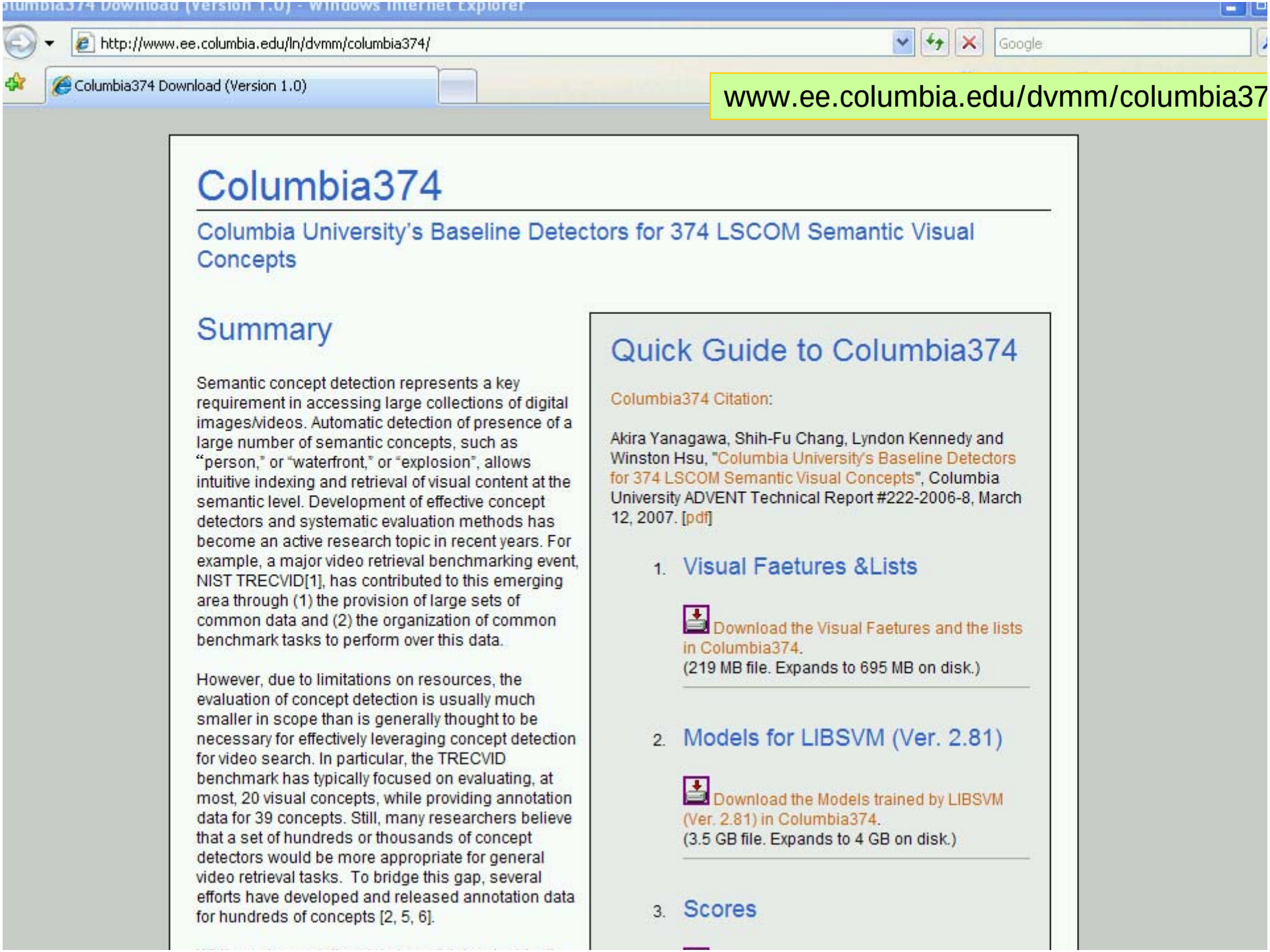


48 dimensions

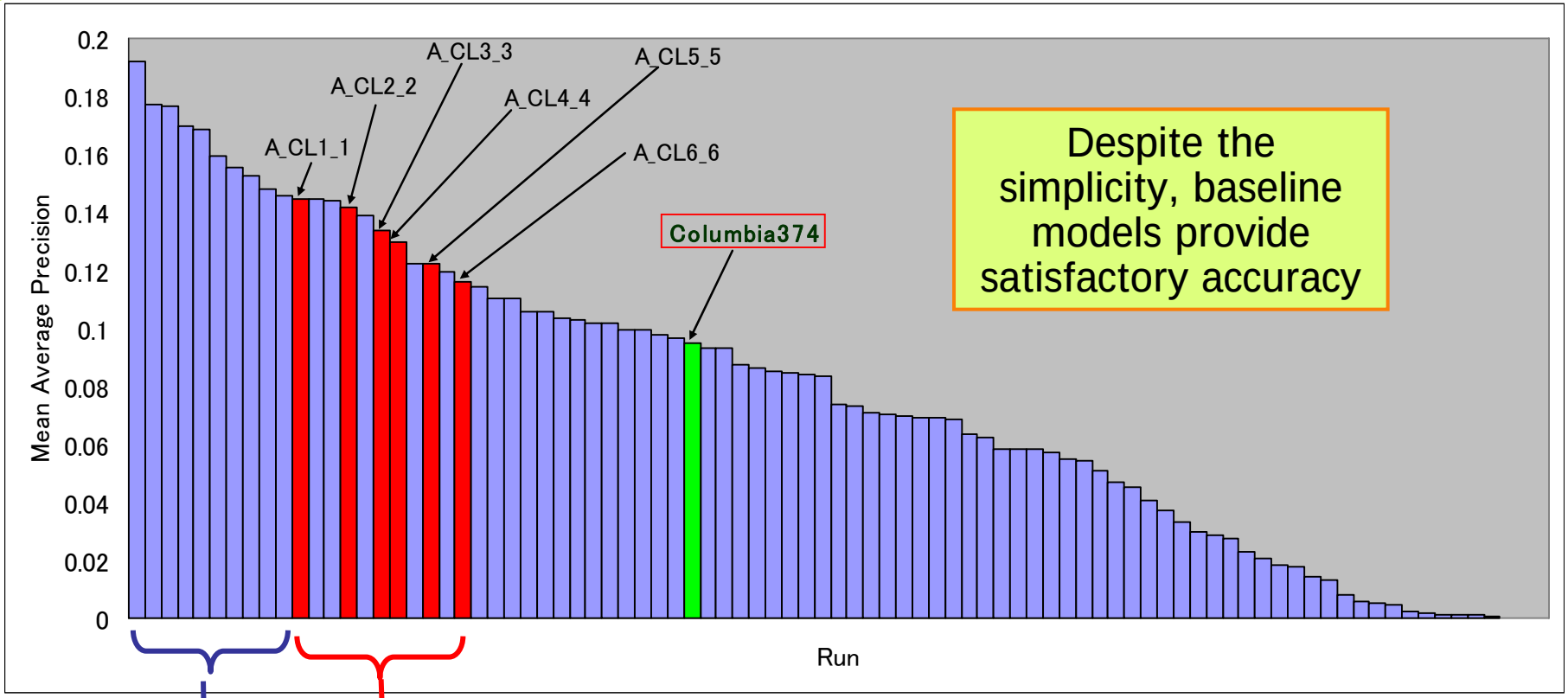
edge direction histogram



73 dimensions


http://www.ee.columbia.edu/ln/dvmm/columbia374/
Google
Columbia374 Download (Version 1.0)
www.ee.columbia.edu/dvmm/columbia37
Columbia374
Columbia University's Baseline Detectors for 374 LSCOM Semantic Visual Concepts
Summary
Semantic concept detection represents a key requirement in accessing large collections of digital images/videos. Automatic detection of presence of a large number of semantic concepts, such as "person," or "waterfront," or "explosion", allows intuitive indexing and retrieval of visual content at the semantic level. Development of effective concept detectors and systematic evaluation methods has become an active research topic in recent years. For example, a major video retrieval benchmarking event, NIST TRECVID[1], has contributed to this emerging area through (1) the provision of large sets of common data and (2) the organization of common benchmark tasks to perform over this data.
However, due to limitations on resources, the evaluation of concept detection is usually much smaller in scope than is generally thought to be necessary for effectively leveraging concept detection for video search. In particular, the TRECVID benchmark has typically focused on evaluating, at most, 20 visual concepts, while providing annotation data for 39 concepts. Still, many researchers believe that a set of hundreds or thousands of concept detectors would be more appropriate for general video retrieval tasks. To bridge this gap, several efforts have developed and released annotation data for hundreds of concepts [2, 5, 6].
Quick Guide to Columbia374
Columbia374 Citation:
Akira Yanagawa, Shih-Fu Chang, Lyndon Kennedy and Winston Hsu, "Columbia University's Baseline Detectors for 374 LSCOM Semantic Visual Concepts", Columbia University ADVENT Technical Report #222-2006-8, March 12, 2007. [pdf]
1. Visual Faetures &Lists
Download the Visual Faetures and the lists in Columbia374.
(219 MB file. Expands to 695 MB on disk.)
2. Models for LIBSVM (Ver. 2.81)
Download the Models trained by LIBSVM (Ver. 2.81) in Columbia374.
(3.5 GB file. Expands to 4 GB on disk.)
3. Scores

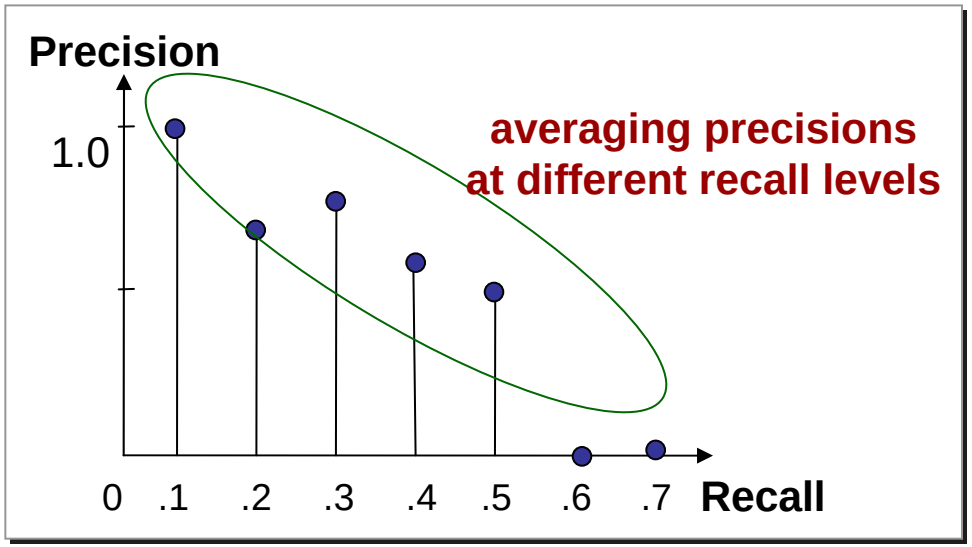
Performance of baseline models in TRECVID 2006



Performance Metric – Average Precision (AP)

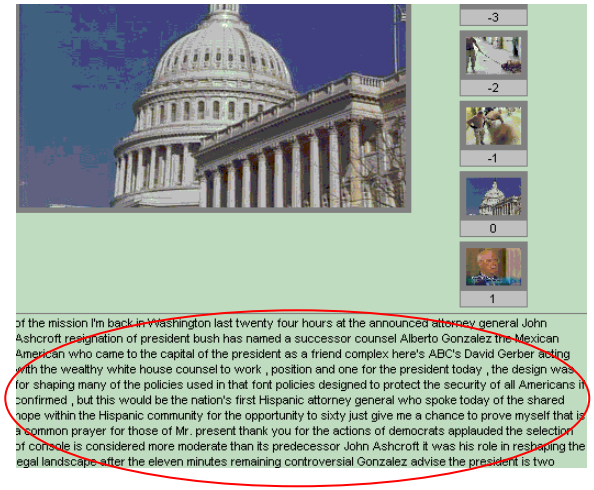
#	hit	R	P
I_1	+	0.10	1.00
I_2	-	-	-
I_3	+	0.20	0.67
I_4	+	0.30	0.75
I_5	-	-	-
I_6	-	-	-
I_7	+	0.40	0.57
I_8	-	-	-
I_9	-	-	-
I_{10}	+	0.50	0.5
I_{11}	-	-	-

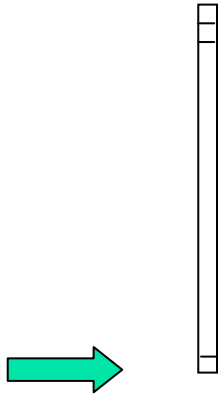
- Rank detection results
- Compute precision/recall values at each hit point
- Average precision at different recall values
- Mean average precision (MAP)
  - Mean of APs across concepts
  - Top TRECVID system MAP ~0.3



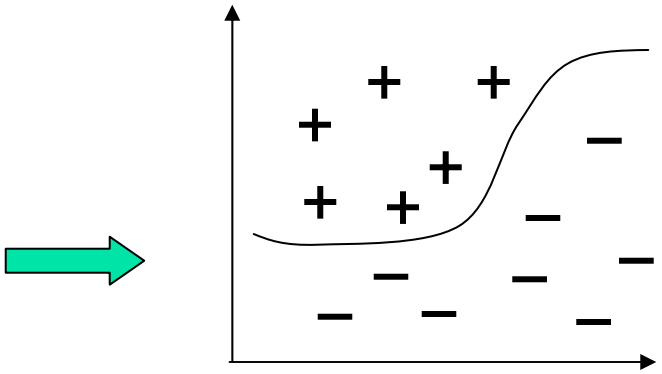
Comparison with Text-based Classifiers

- Extract text features from closed caption, ASR or translated ASR (for foreign videos)
- Apply classifiers : Naïve Bayes, SVM, Max. Entropy …



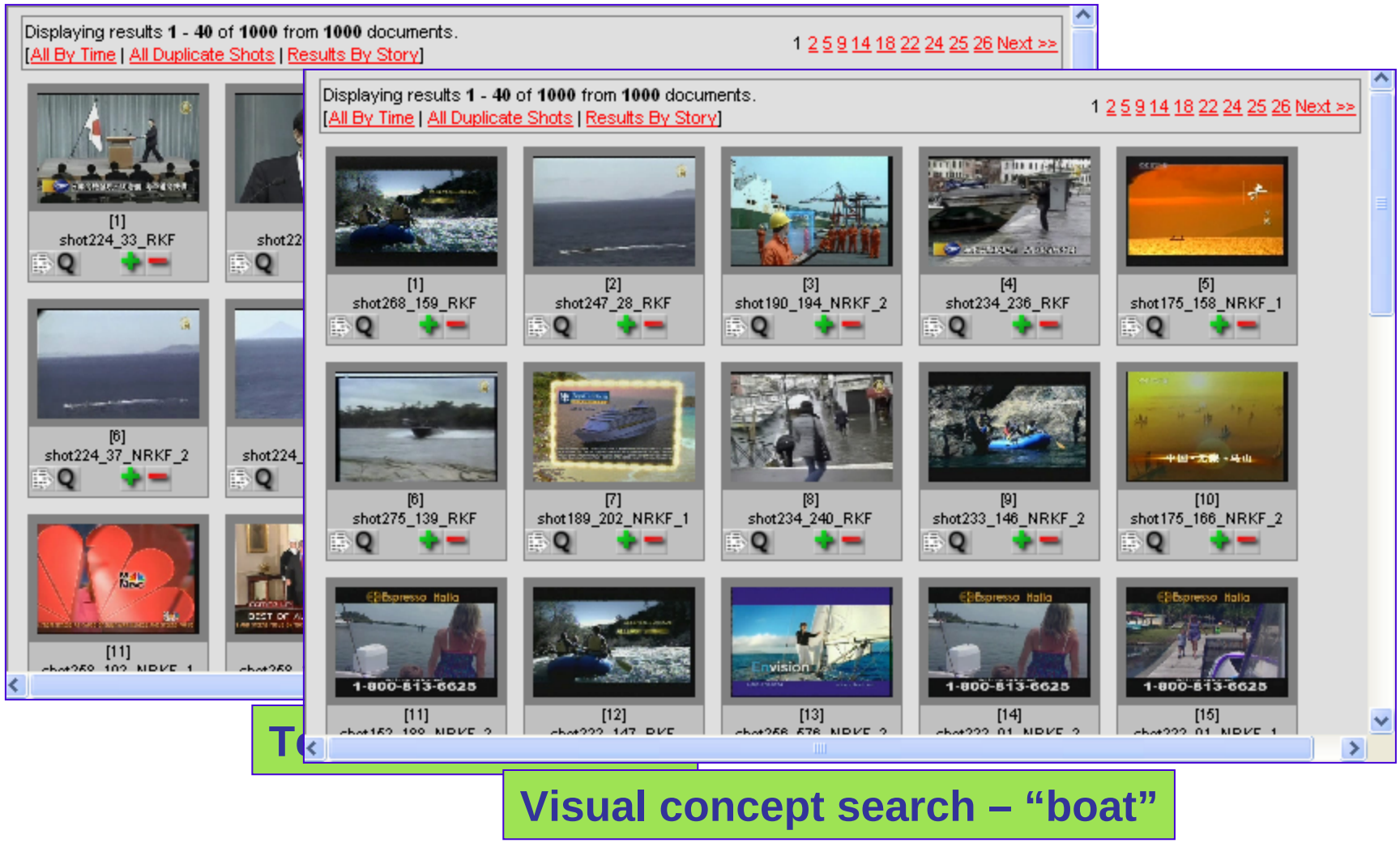


Feature vector



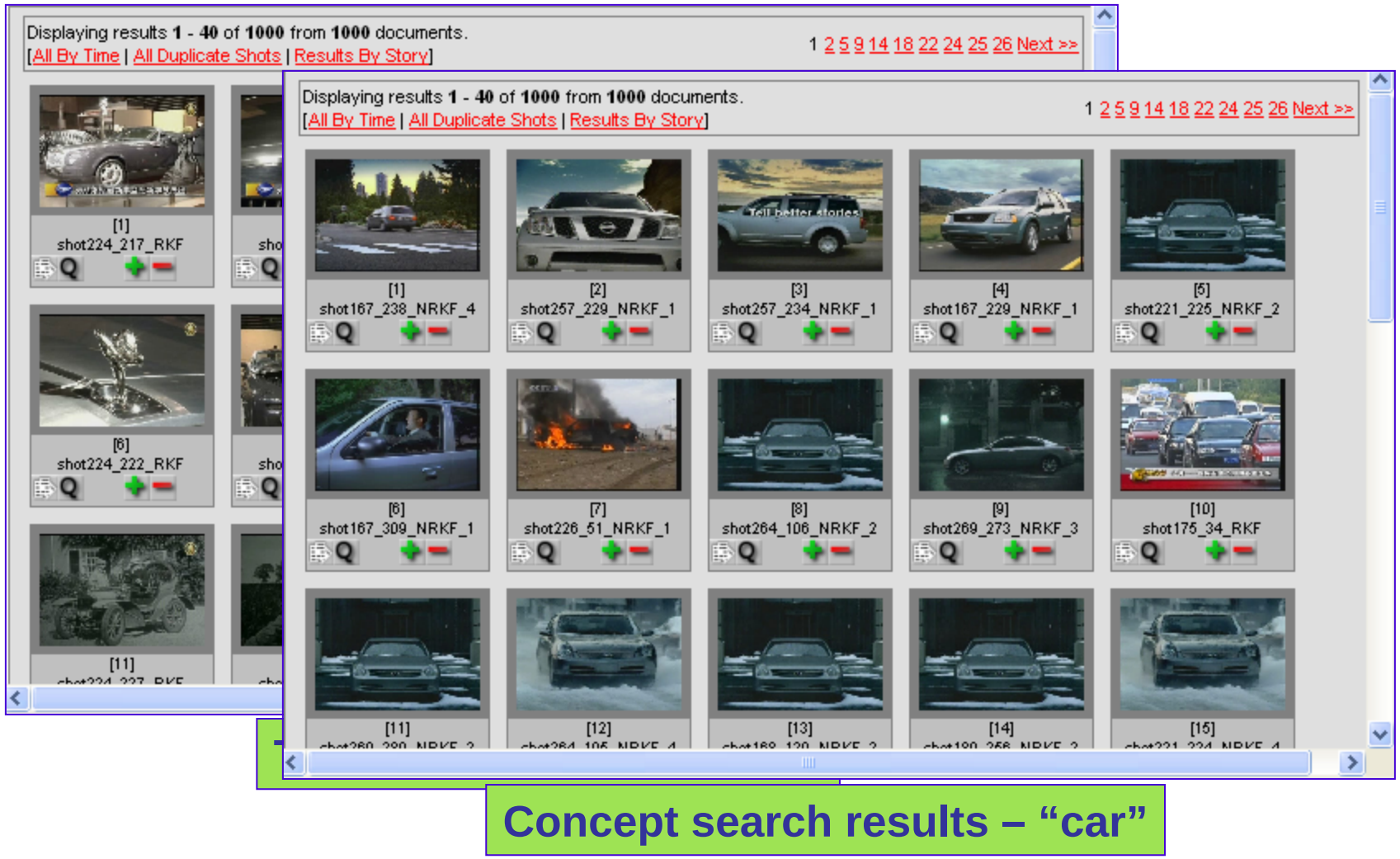
Text classifiers usually lower than visual models by 3-5 times.

Comparing Semantic Classification (text vs. visual) (1)



(images from TRECVID)

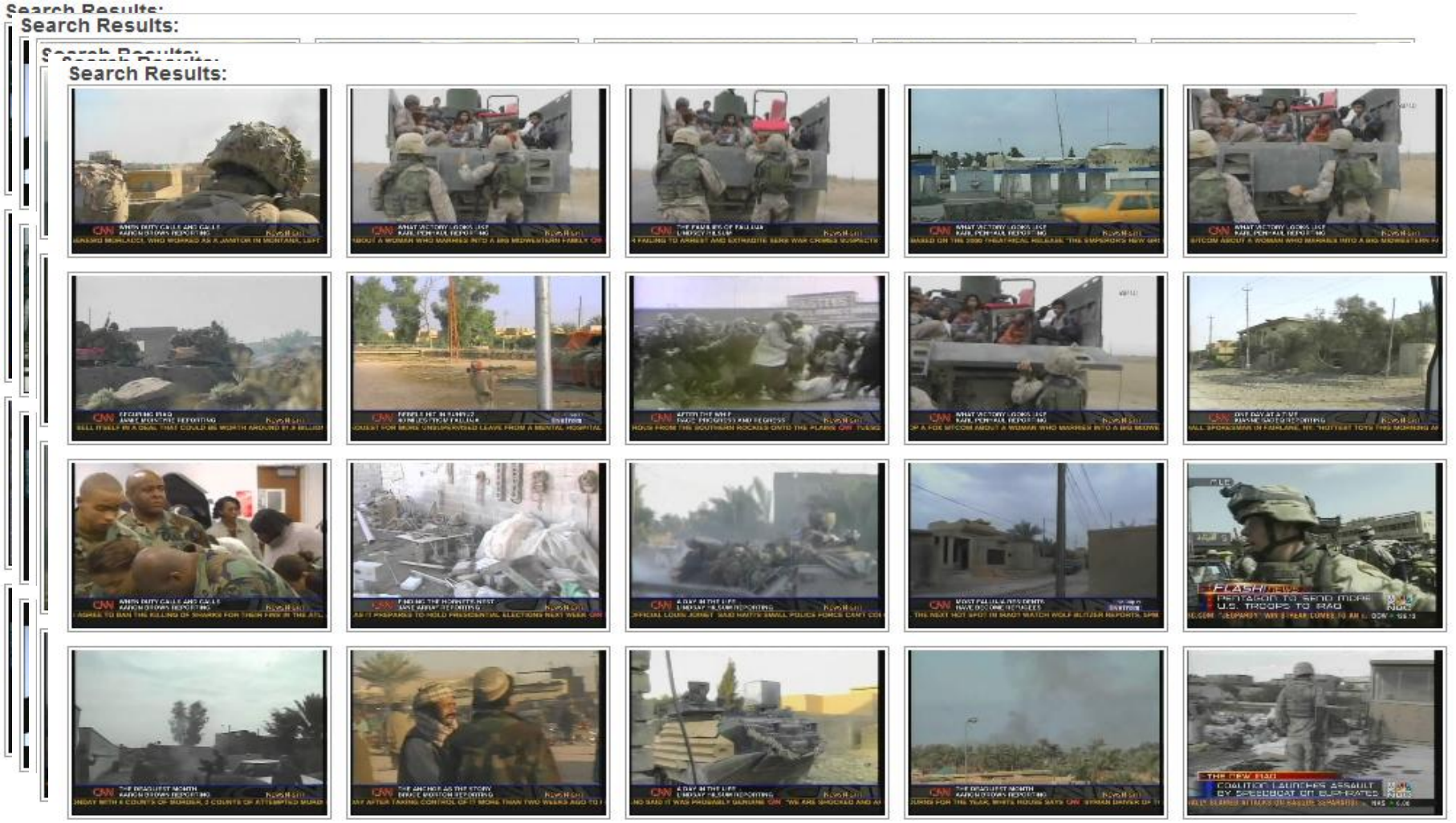
Comparing Semantic Classification (text vs. visual) (2)



Concept search results – “car”

Example: good detectors for LSCOM concept

waterfront | bridge | crowd | explosion fire | US flag | Military personnel

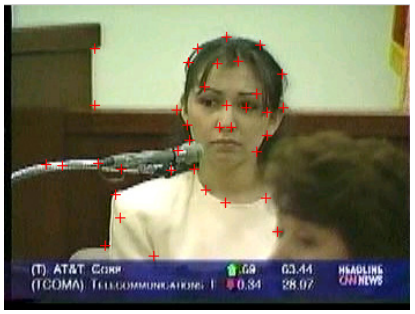


Columbia University in the City of New York

Complexity

- Slow training, fast detection
- Training time over 61K shots per SVM
  - 20 mins (color), 3.5 mins (texture), 1.2 mins (edge)
  - About 1 month over 20 PCs to train *Columbia374*
- Feature extraction
  - 0.4 sec (color), 0.4 sec (texture), 2.8 sec (edge) per image
- Testing
  - < 30 ms per image

Advanced Features and Models



Interest points

Part detection

Local features

SIFT,
Gabor filter,
PCA projection,
Color histogram,
Moments ...

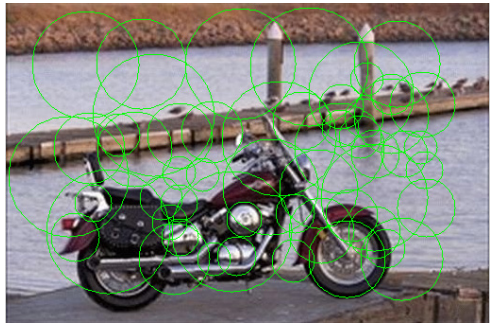
Part-based representation

Bag

Structural Graph

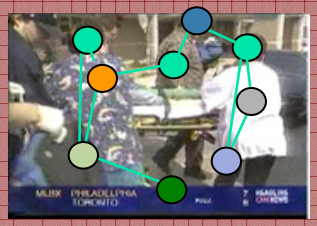
Attributed Relational Graph

Segmented Regions

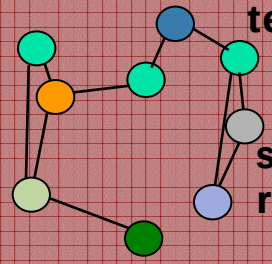


Maximum Entropy Regions

Representation and Learning

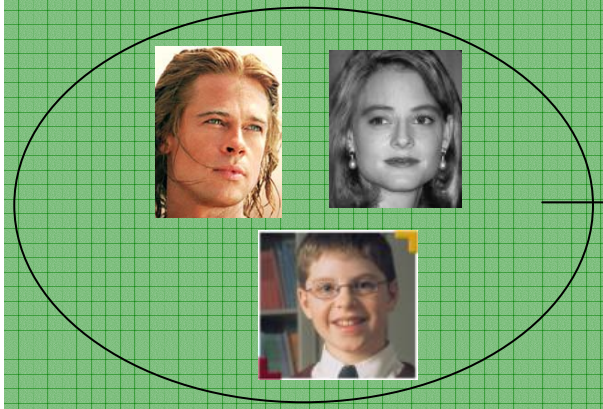


Individual images
→ Salient points, high entropy regions



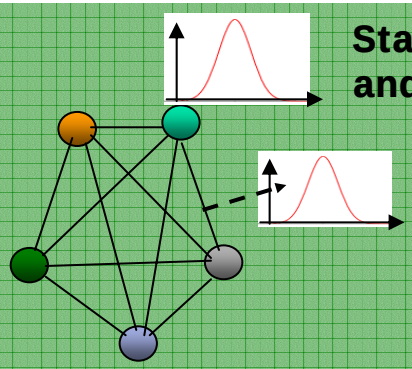
Attributed Relational Graph (ARG)

Graph Representation of an Image



Collection of training images

machine learning

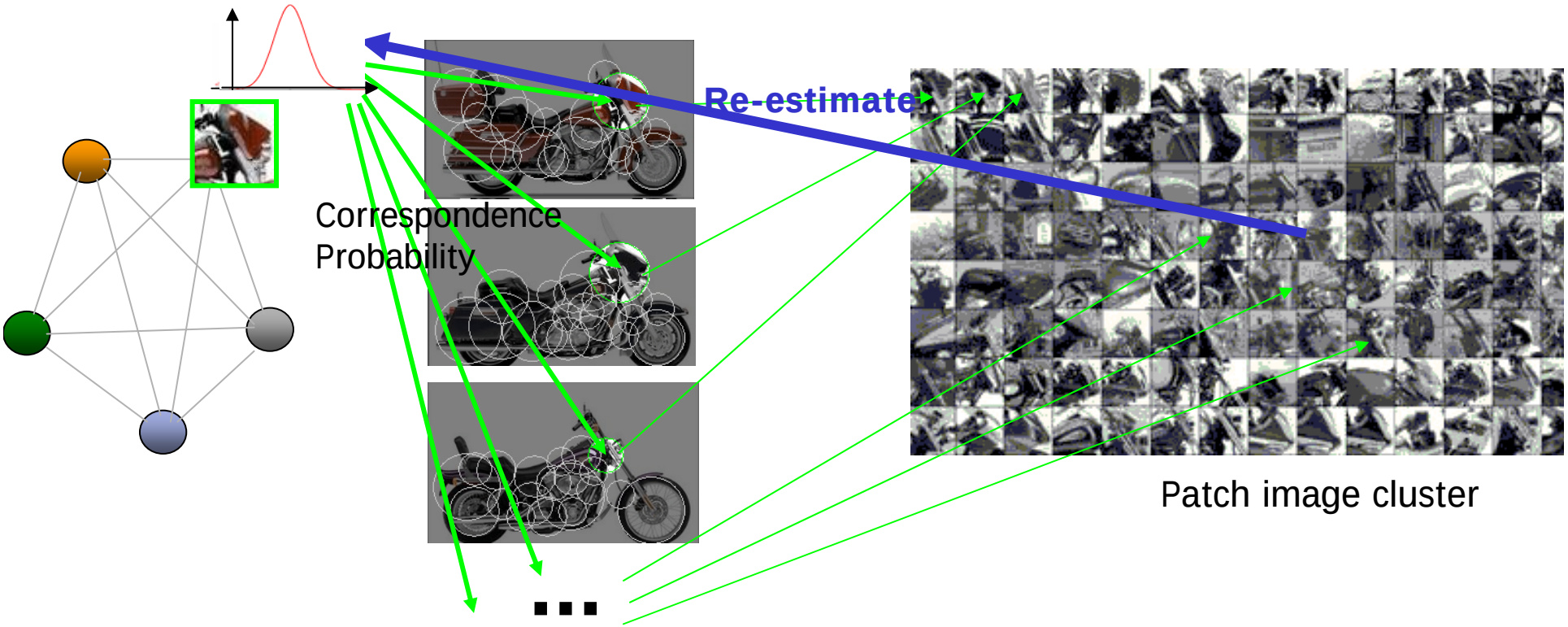


Random Attributed Relational Graph (RARG)

Statistical Graph Representation of Model

(Zhang & Chang CVPR 06)

Learning Graph Object Model



- **BUT**
  - **Finding the correspondence of parts and computing matching probability are NP-complete**
  - **Use advanced machine learning methods: Loopy Belief Propagation, and Gibbs Sampling plus Belief Optimization**

demo

Model Contextual Relations among Concepts

Hard/specific concept

"Government-Leader" Detector

Context-based Model

Government-Leader

\-

\+

Flag, podium

Face

Generic concept

"Face" Detector

Generic concept

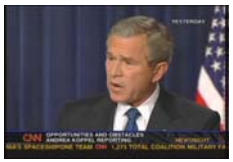
"flag" "podium"

Context Information

dvmm DIGITAL VIDEO • MULTIMEDIA LAB

Boosted Conditional Random Fields

(Jiang, Chang, Louie ICASSP '07)



face detector

Gov. Leader detector

music detector

$P(\text{face}|\text{image})$

$P(\text{gov. leader}|\text{image})$

$P(\text{flag}|\text{image})$

Initial scores

boosting

context-based model

Conditional Random Field

face

Gov. leader

flag

updated scores

$\tilde{P}(\text{face}|\text{image})$

$\tilde{P}(\text{gov. leader}|\text{image})$

$\tilde{P}(\text{flag}|\text{image})$

Predict When will Context Help

Concept Fusion (CF) does not always help:

- Complex inter-conceptual relationships difficult to estimate from limited training samples
- Strong classifiers may suffer by fusion with inaccurate context

Use CF for concept C_i , only if C_i is weak or with strong context

$$E(C_i) > \lambda$$

weak self

or

$$\frac{\sum_{C_j, j \neq i} I(C_j; C_i) E(C_j)}{\sum_{C_j, j \neq i} I(C_j; C_i)} < \beta$$

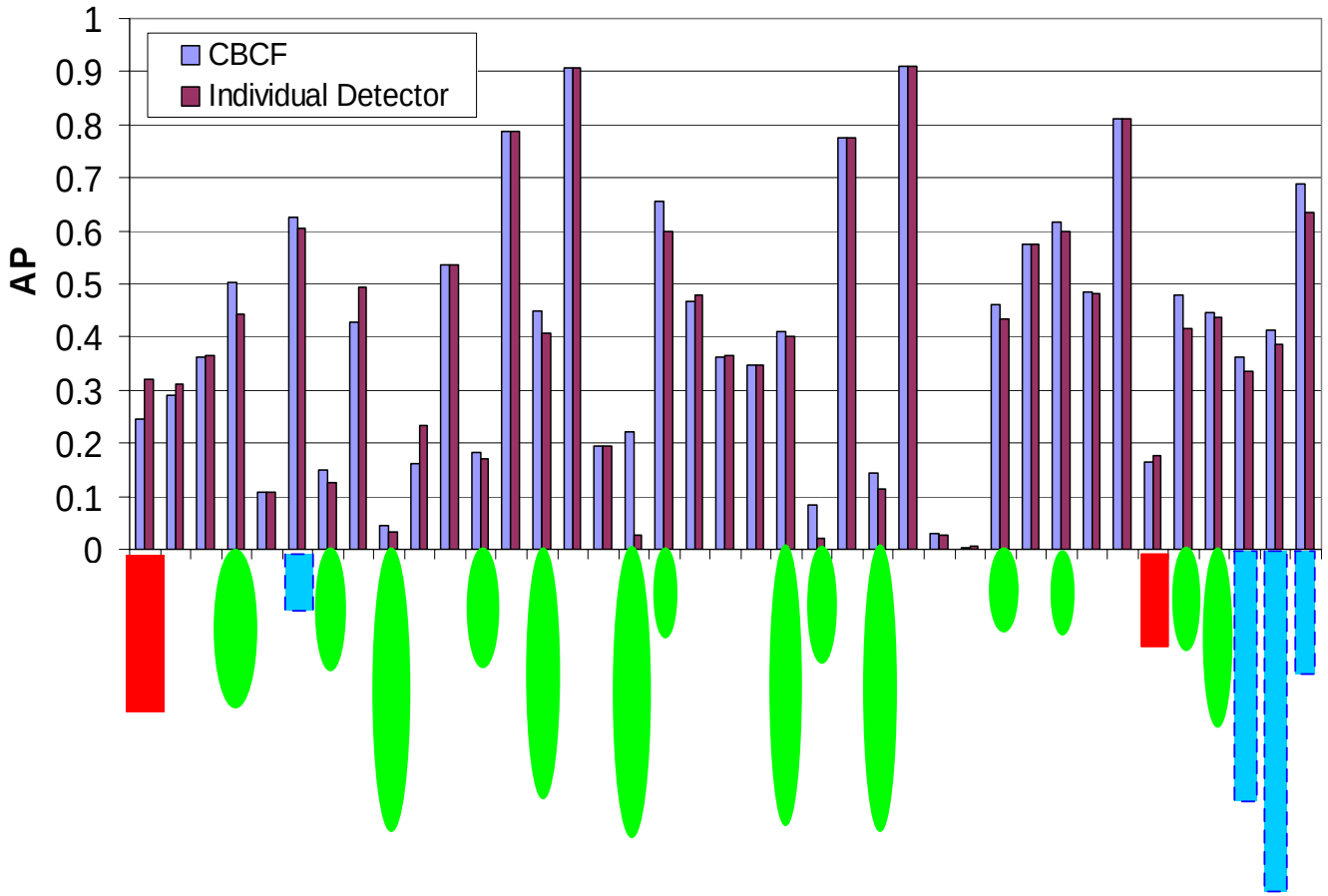
Strong context

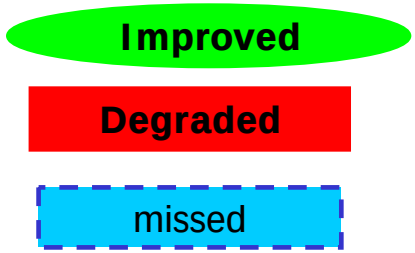
$I(C_i; C_j)$ -- mutual information between C_i and C_j

$E(C_i)$ -- error rate of independent detector for C_i

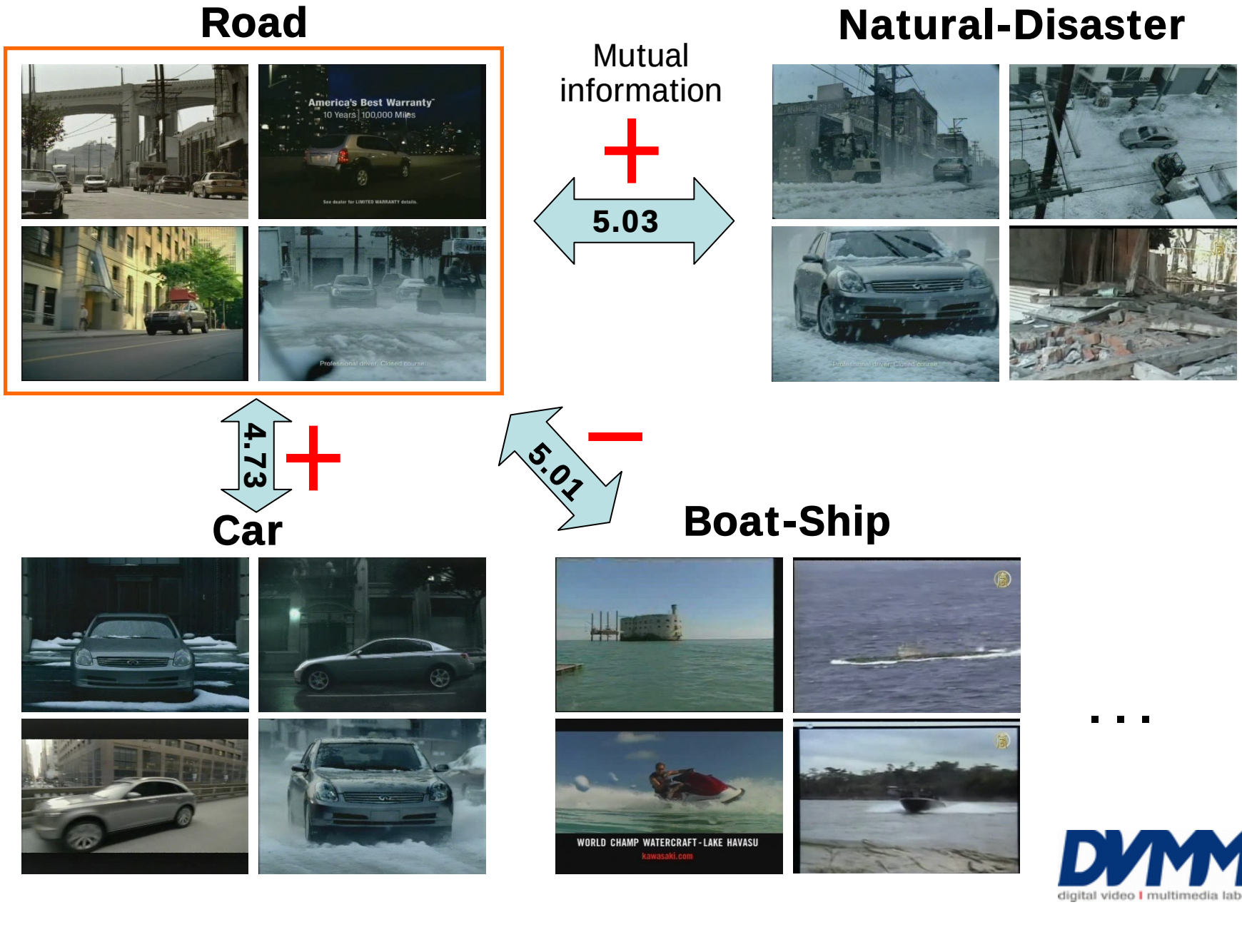
Fuse Context Only When it is Predicted to Help

- Apply context fusion smartly
- 16 concepts are predicted to improve
- 14 actually improve, 4 missed





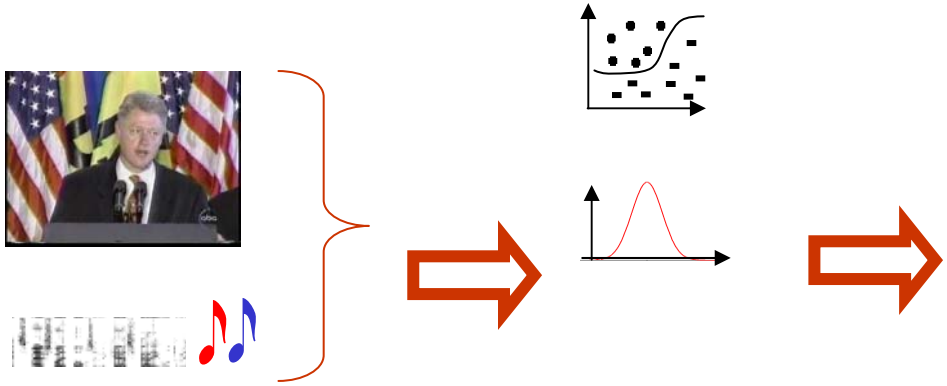
Example: Road Detector



Combine Context-Fusion with User Input

- automatic content recognition
  +
  context fusion
  +
  user input

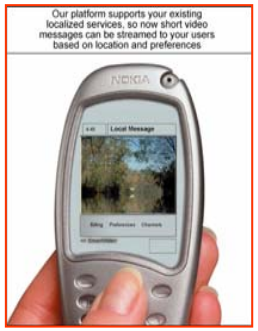
Instead of Automatic Tagging 100% of Concepts



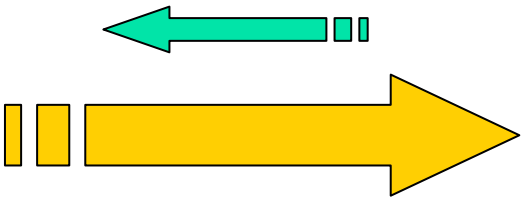
Tags: Person: 79%
Clinton: 75%
US Flag: 80%
Podium: 70%
Give speech: 75%
Press conference: 65%

What if we can ask users to help 1-2 labels?

Active Tagging



User labels: park, picnic



Auto. generated labels:
people, tree, mountain, etc

Which Questions to Ask?

(20 Questions Game)

(Jiang, Chang, Loui, ICIP 06)

User labels: ???

All Concepts of interest --- $S_1, \cdots, S_N$

Questions to ask user --- $S^*_{1:M} = \{S^*_1, \cdots, S^*_M\}$

Selection criterion:

minimize the expected Bayes classification error $\overline{E}$

[Hellman et al., IEEE Trans. Info. Theory, 1970]

$$\overline{E} = \frac{1}{N}\sum_{i=1}^{N}\overline{E}(S_i) \le \frac{1}{2N}\sum_{i=1}^{N}H(S_i) - \frac{1}{2N}\sum_{i=1}^{N}I(S_i; S^*_{1:M})$$

entropy

Mutual info.

Best concepts for active tagging: high uncertainty or high mutual info

Examples of Best Questions

Active Tagging

Best Questions to Ask User?

1. Outdoor?
2. Airplane?

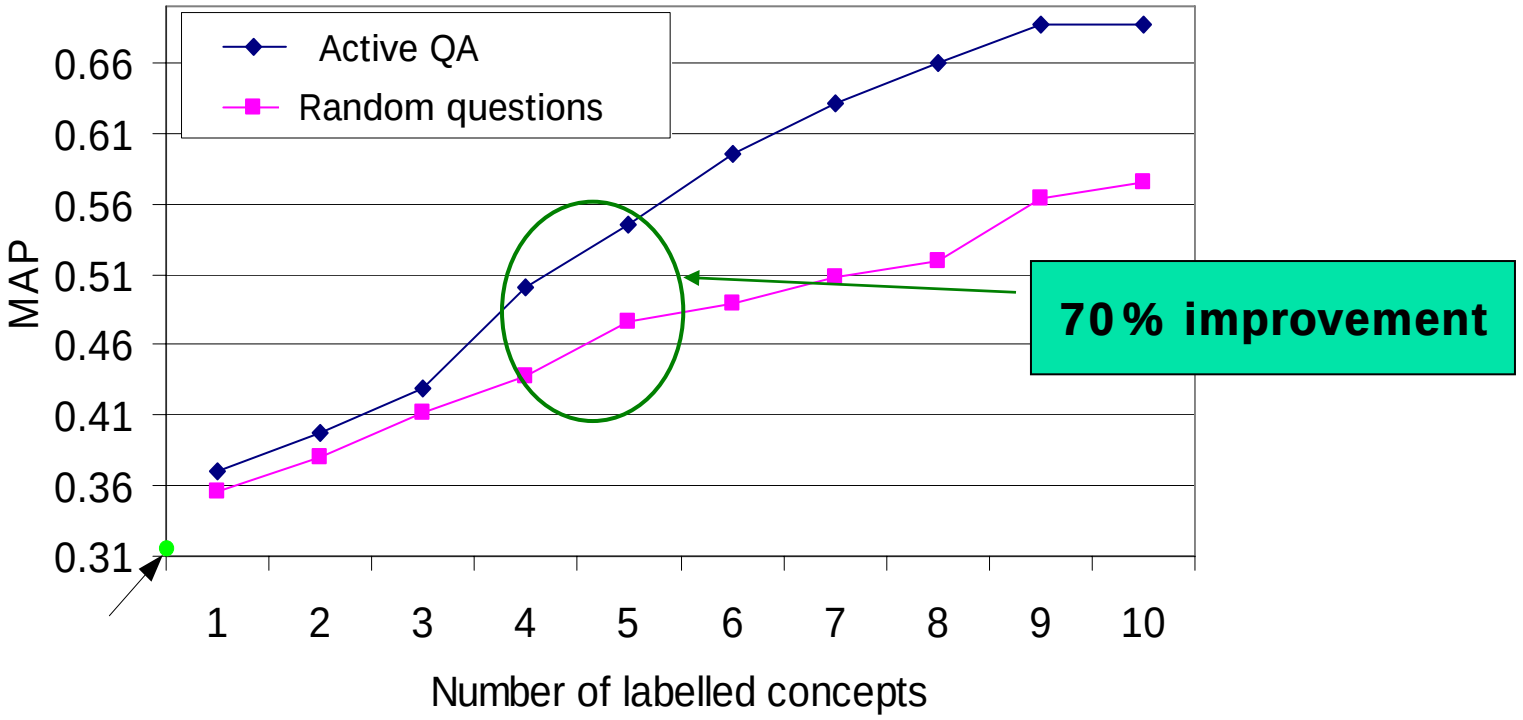
1. Outdoor?
2. Court?

1. Outdoor?
2. Building?

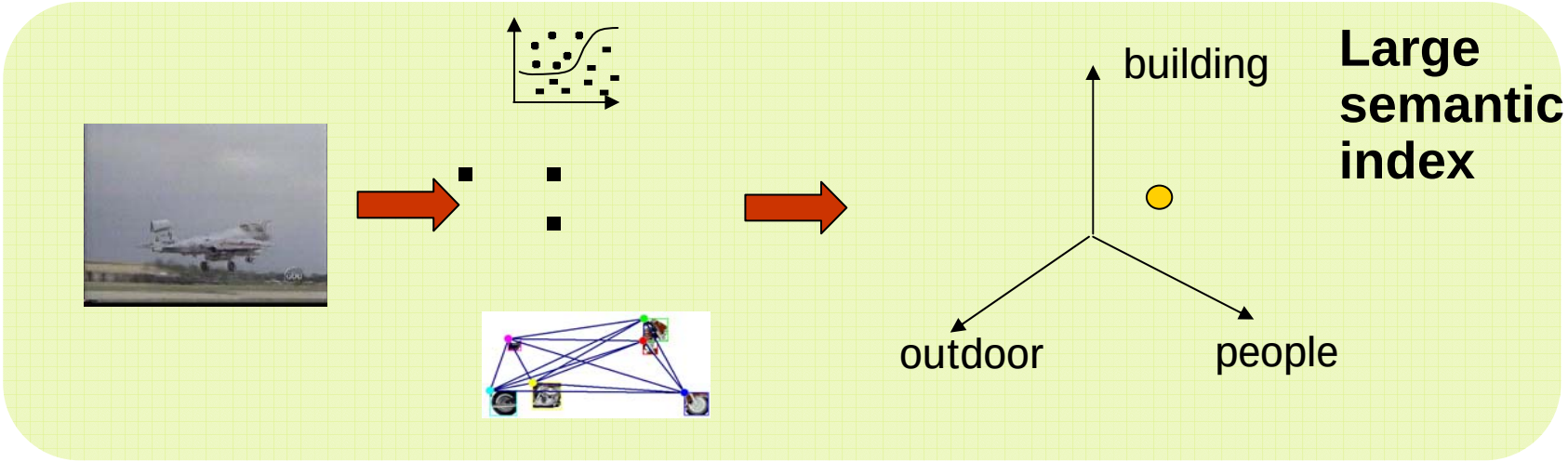
urban, waterscape, etc

Active Tagging Consistently Help

- Performance gain increases with more user input
- Magic number here is 4

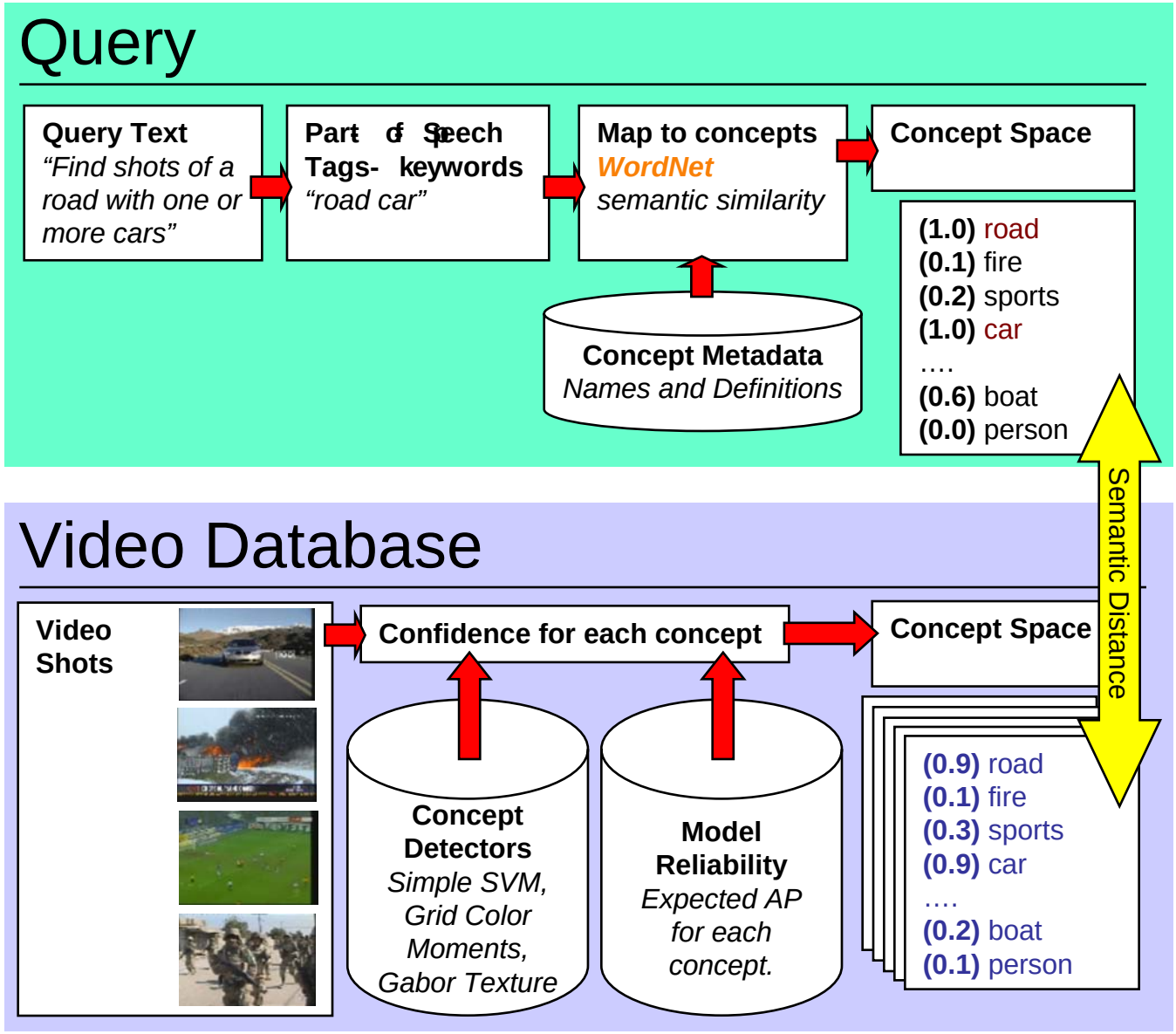


Power of Concept-based Representation



New applications:
Search, Filtering, Pattern Mining

Power of Semantic Indexing: Text-to-Concept Search



- **Map text queries to concept**
- **Use ontology (Wordnet) to find matched concepts**
- **Videos represented by semantic descriptors**
- **Measure semantic distance between query and concept**

Mapping search topics to concepts

TRECVID search topics

Finds shots with one or more emergency vehicles in motion (e.g., ambulance, police car, fire truck, etc.)

Matched Concepts:
Emergency_Room, Vehicle

Find shots with a view of one or more tall buildings (more than 4 stories) and the top story visible.

Matched Concepts: Building

Find shots with
or entering a ve

re soldiers, police,
oner.

Research issue:
what concept to use?
How to fuse multiple concepts?

Matched Concepts: Person, Vehicle

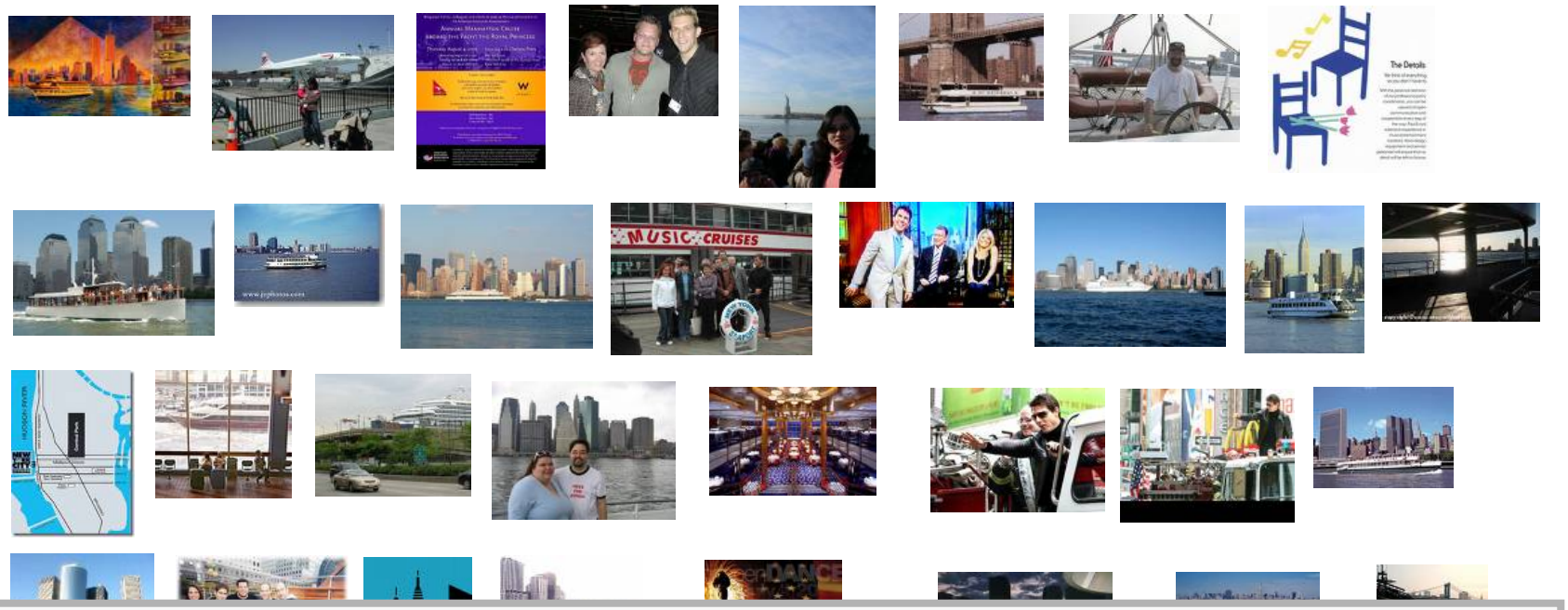
Matched Concepts:
Guard, Police_Security, Prisoner, Soldier

Concept Search Demo

- Concept search case 1 (link)
- Concept search case 2 (link)
- Multimodal search (link)

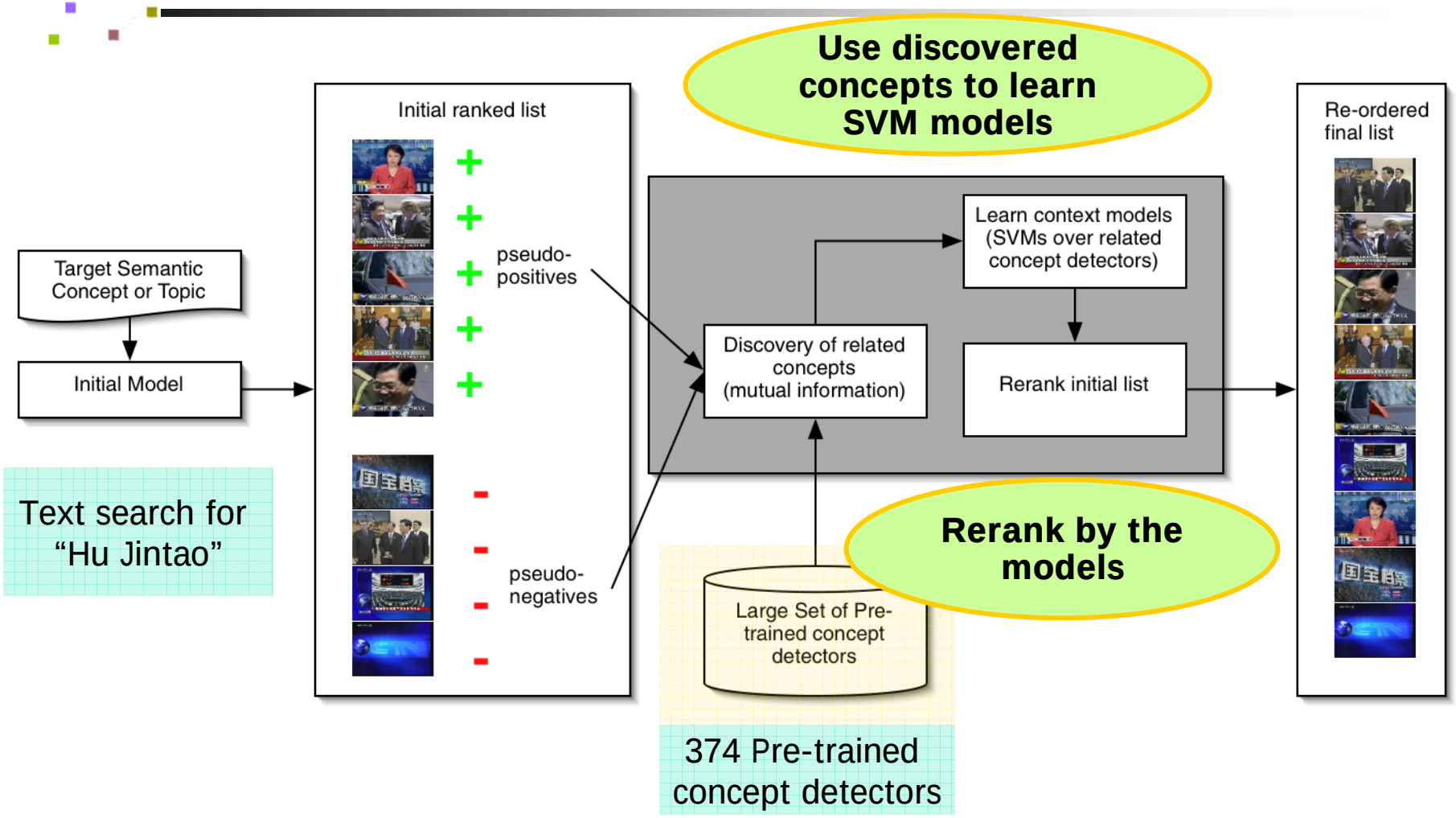
Other Applications: Semantic Mining

Top results from text query: *"Manhattan Cruise"*



- What are the dominant semantic concepts in the initial result set?
  - *Boat, waterfront, sky, buildings*, …

Decipher dominant visual concepts behind each search topic



Related concepts have high mutual information with target labels.

- Mutual information:

$$I(T;C) = \sum_{T}\sum_{C} P(T,C)\log\frac{P(T,C)}{P(T)P(C)}$$

  - T: initial text search score
  - C: concept detection score (quantized)
  - Include both positive and negative correlations

Examples: discovered concepts per query

Query Topic	Positive Concepts	Negative Concepts
(158) Find shots of a helicopter in flight.	Weapons, Explosion Fire, Airplane, Smoke, Sky	Person, Civilian Person, Face, Talking, Male Person, Sitting
(164) Find shots of boats or ships.	Waterscape, Outdoor, Sky, Fire, Exploding Ordnance	Person, Civilian Person, Face, Adult, Suits, Ties
(179) Find shots of Saddam Hussein.	Court, Politics, Government Leader, Suits, Lawyer, Judge	Desert, Protest, Crowd, Mountain, Outdoor, Animal

Three main types of topic-concept relationships:

Generally Present Obvious, expected relationships. Concepts that are always together.

News Story Relationships unique to news story. Reranking uniquely powerful.

Mistaken Relationships due to detection errors. Rare, but still helpful.

Another Application: Event Detection

- Typical approach to event detection: detect object of interest, track over time, model spatio-temporal dynamics

Difficult for events without explicit objects and motions, e.g., "riot", "street battle", "flight combat", "parade"

New Approach: Concept-based Event Representation (1)

Apply large-scale concept detectors

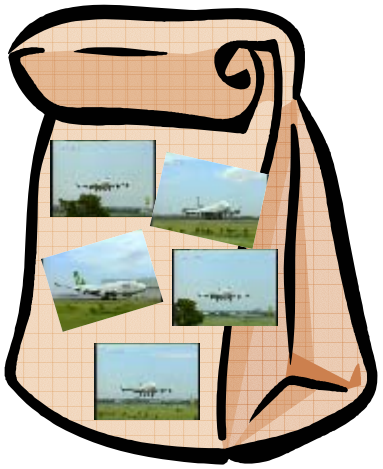
airplan

indoor

crowd

sky

- Map image frames to confidence scores in the semantic concept space
- A video is associated with a bag of concept features



bag of concept features

Concept-based Event Representation (2)

Large-scale concept detectors

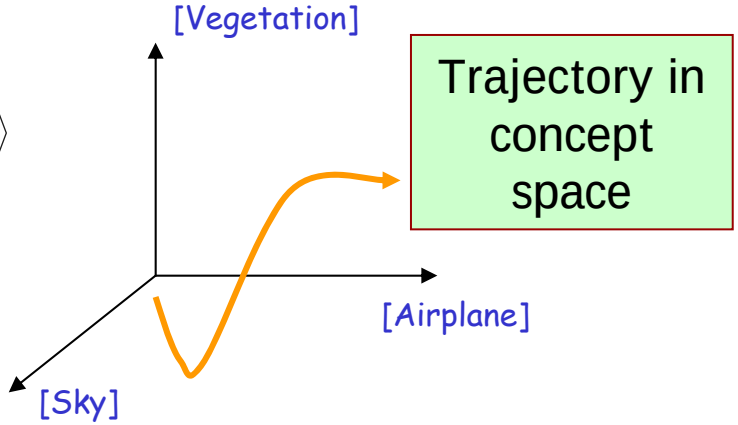
airplan

indoor

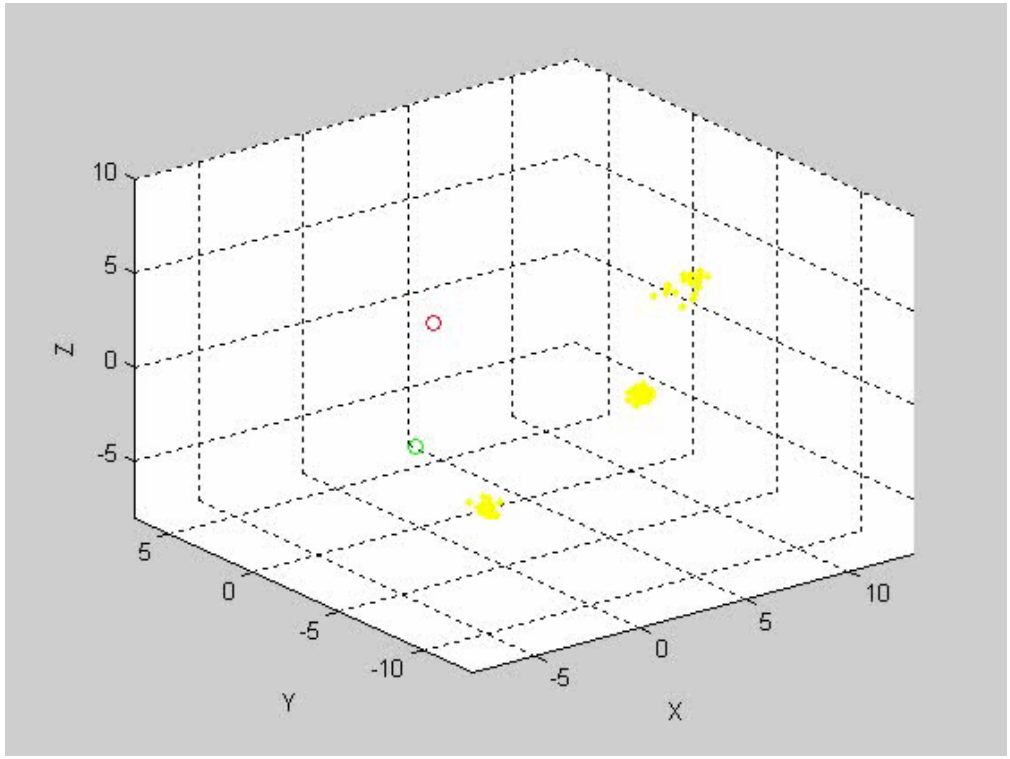
crowd

sky

- Trace the concept scores over time
- A video corresponds to a trajectory in the concept space

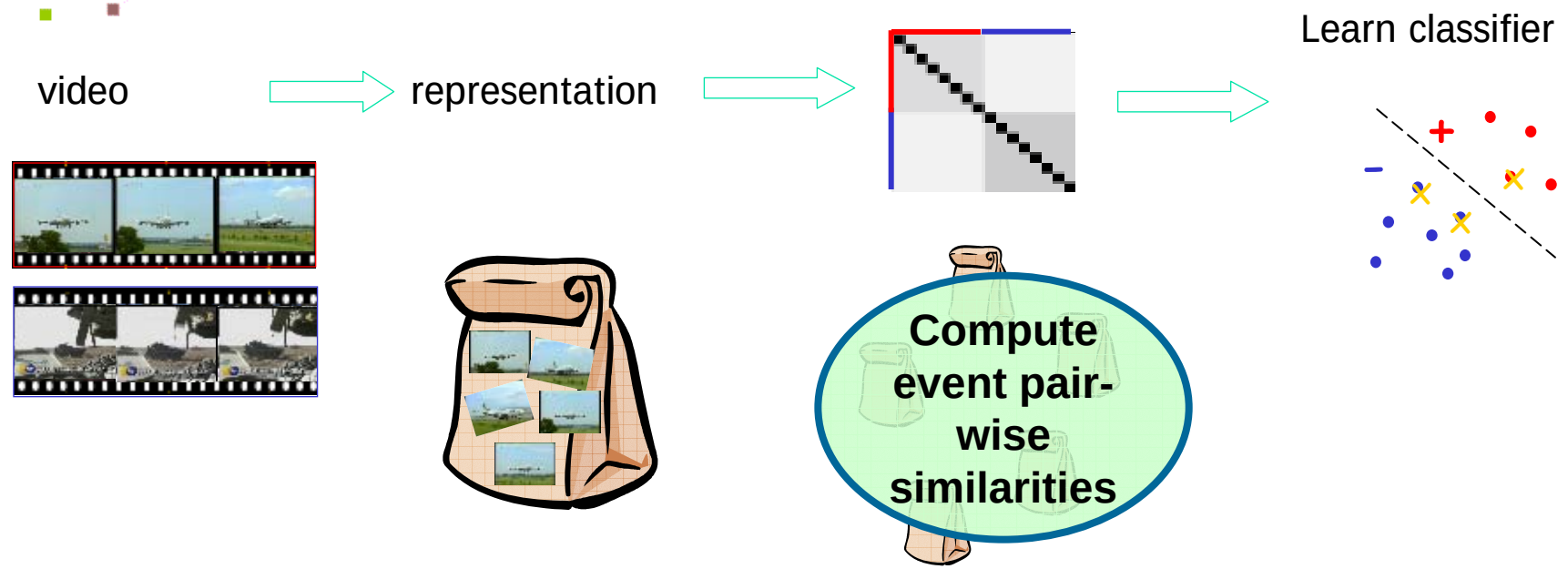


Video Events as Distinct Traces in the Concept Space



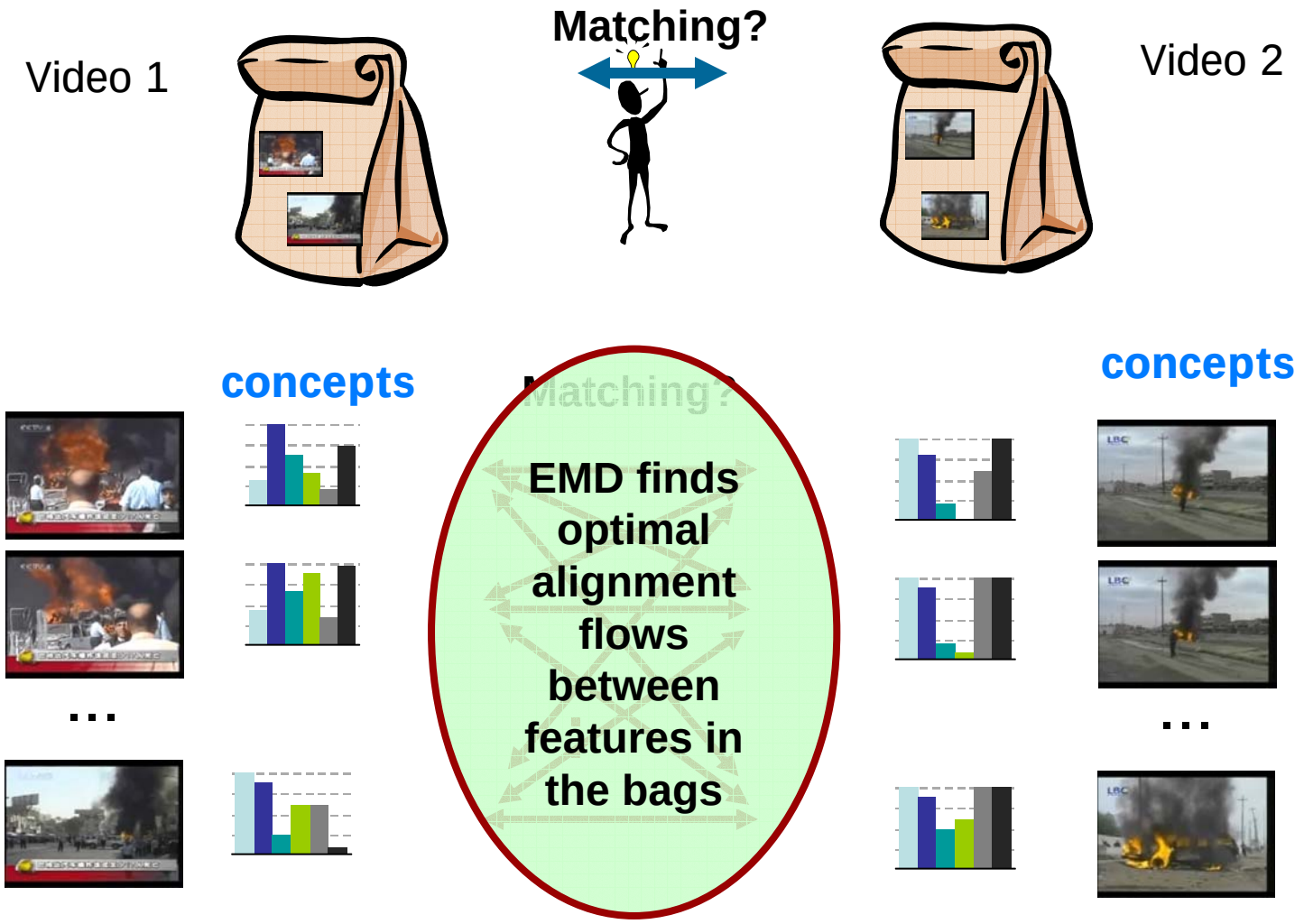
Red: video 1, Green: video 2, Yellow: other videos

From Representation to Event Detection



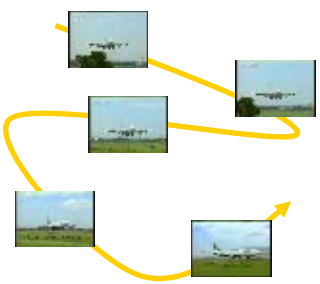
Earth-Mover's Distance (EMD)

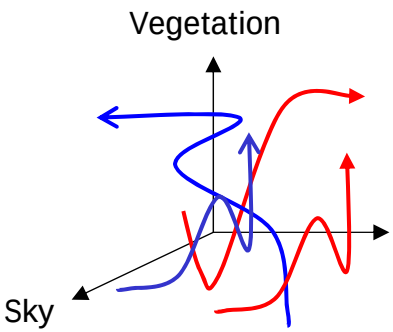
[Xu & Chang, CVPR 07]



HMM+SVM to Model Event Traces in the Concept Space

[Xie et al, ICME 06]

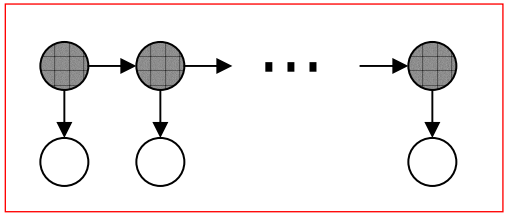


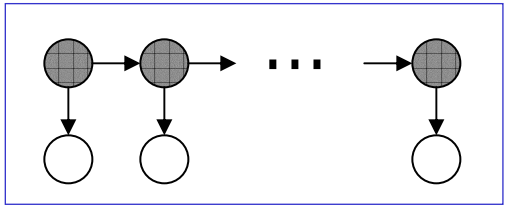


Event Trajectories in high-dimensional space

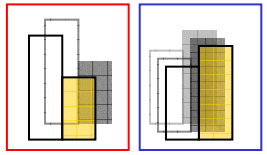
Model dynamics in each dimension

HMM

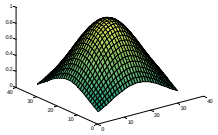




Compute model statistics

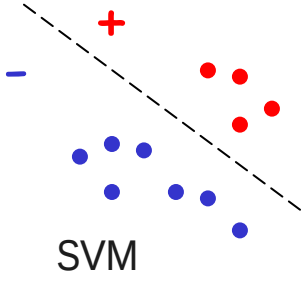


state-histogram



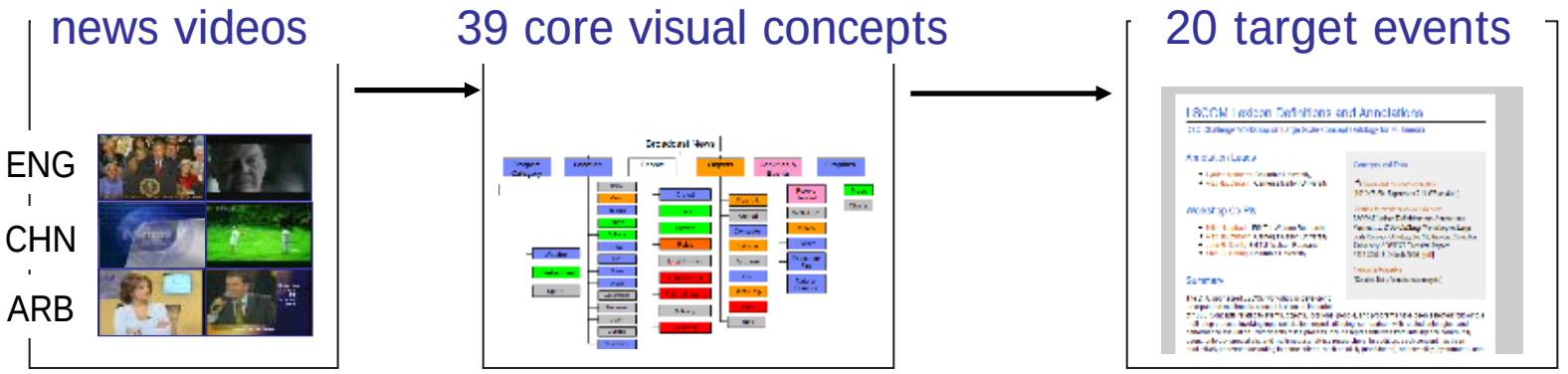
fisher score

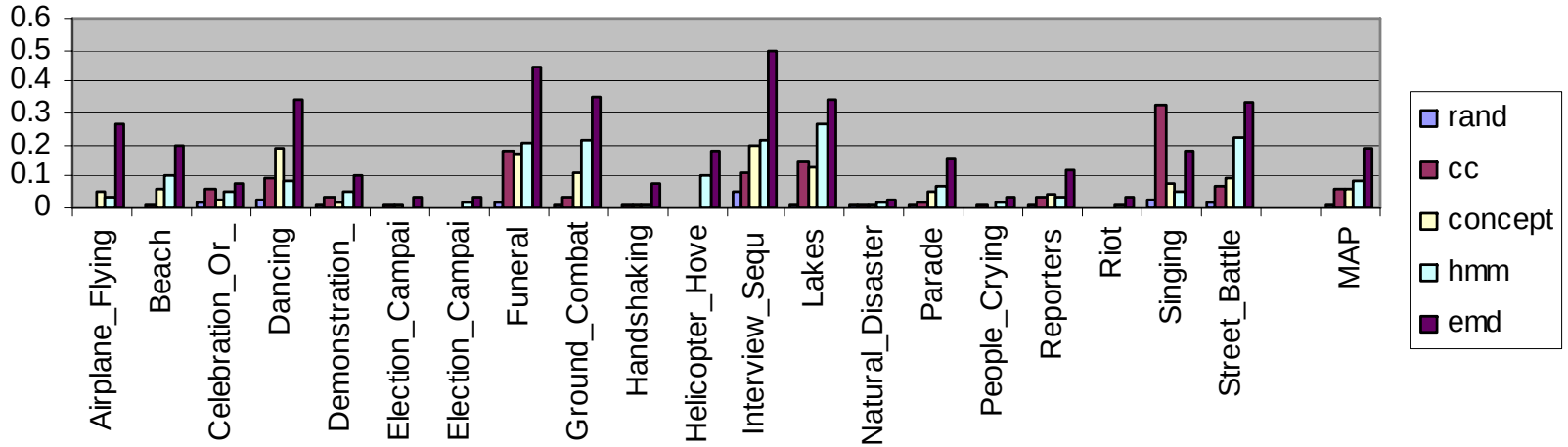
Train Classifiers



SVM

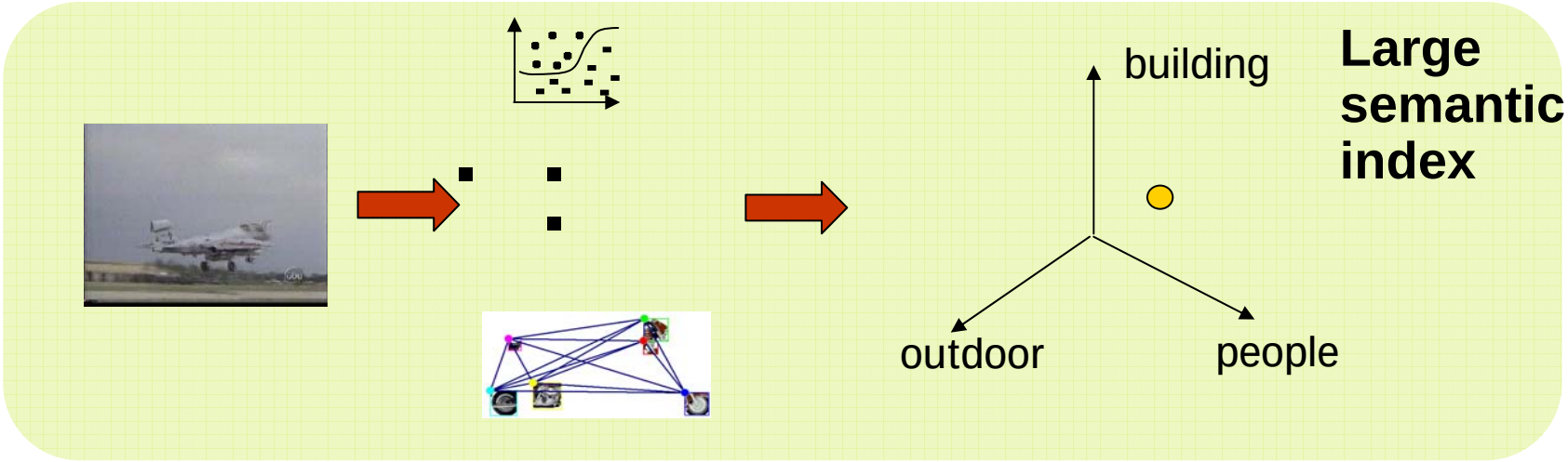
Performance Testing Using TRECVID Data

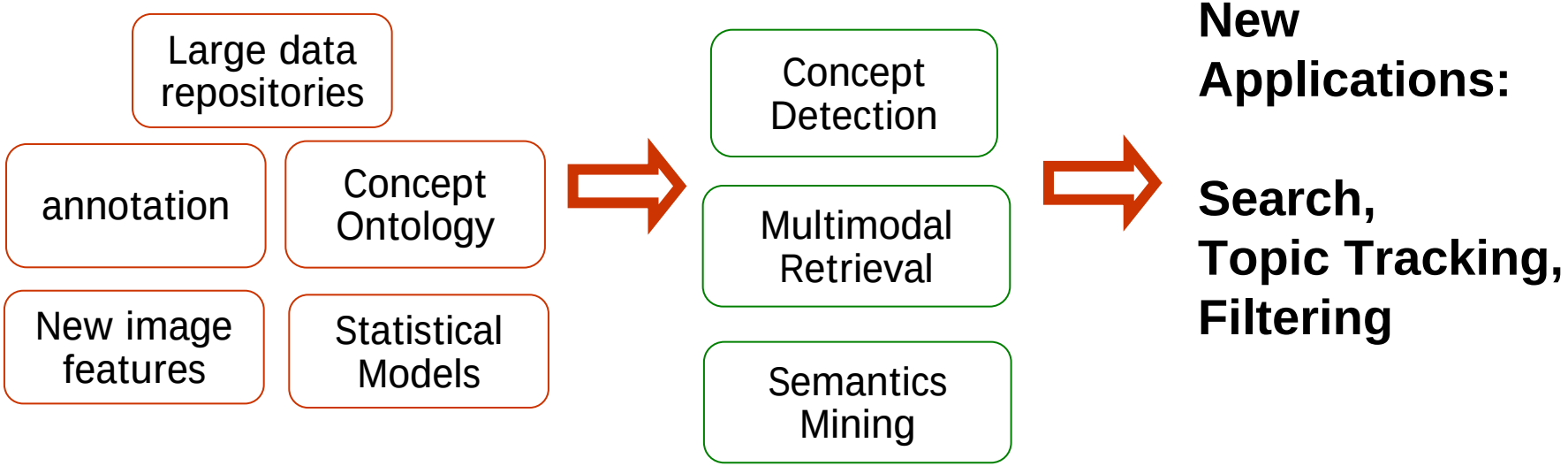




Significant performance gain by applying temporal models (EMD and HMM)

Landscape of Semantic Image/Video Indexing





Open Issues and Opportunities

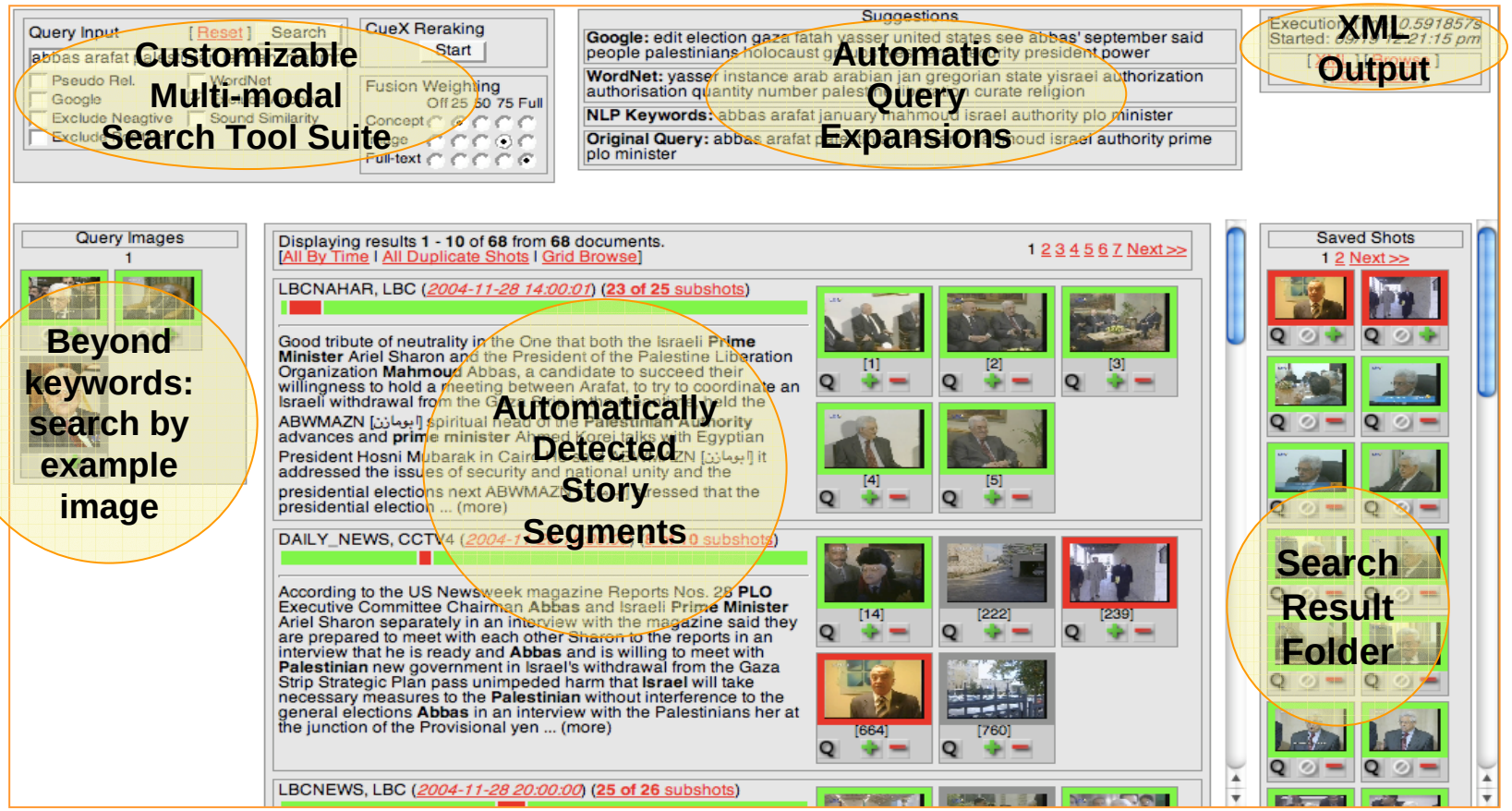
- Systematic Way of Extending Ontology?
- Cross-Domain Model Development:
  - General detectors for News, Consumer, Web?
  - Sharing of data and models?
- Complexity and speed
  - Feature selection, fast training
- Map user queries to visual concepts
  - Wordnet for visual search?

More Information

- Columbia DVMM Lab
  - http://www.ee.columbia.edu/dvmm
  - 374 SVM-based concept detectors available soon
  - Online video search demos
- LSCOM lexicon and annotation
  - http://www.ee.columbia.edu/lscom
- Columbia374 concept detectors
  - http://www.ee.columbia.edu/columbia374

Columbia Video Search System

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



Prototype includes 160 hours, 3 languages (English, Arabic, Chinese), 6 channels