Unleashing Video Search

John R. Smith (with Apostol Natsev, Quoc-Bao Nguyen, Jelena Tesic, Lexing Xie, Rong Yan, Joachim Siedl, Christian Penz, Florian Letz)
Senior Manager, Intelligent Information Management Dept.
IBM T. J. Watson Research Center
Contact: jrsmith@watson.ibm.com

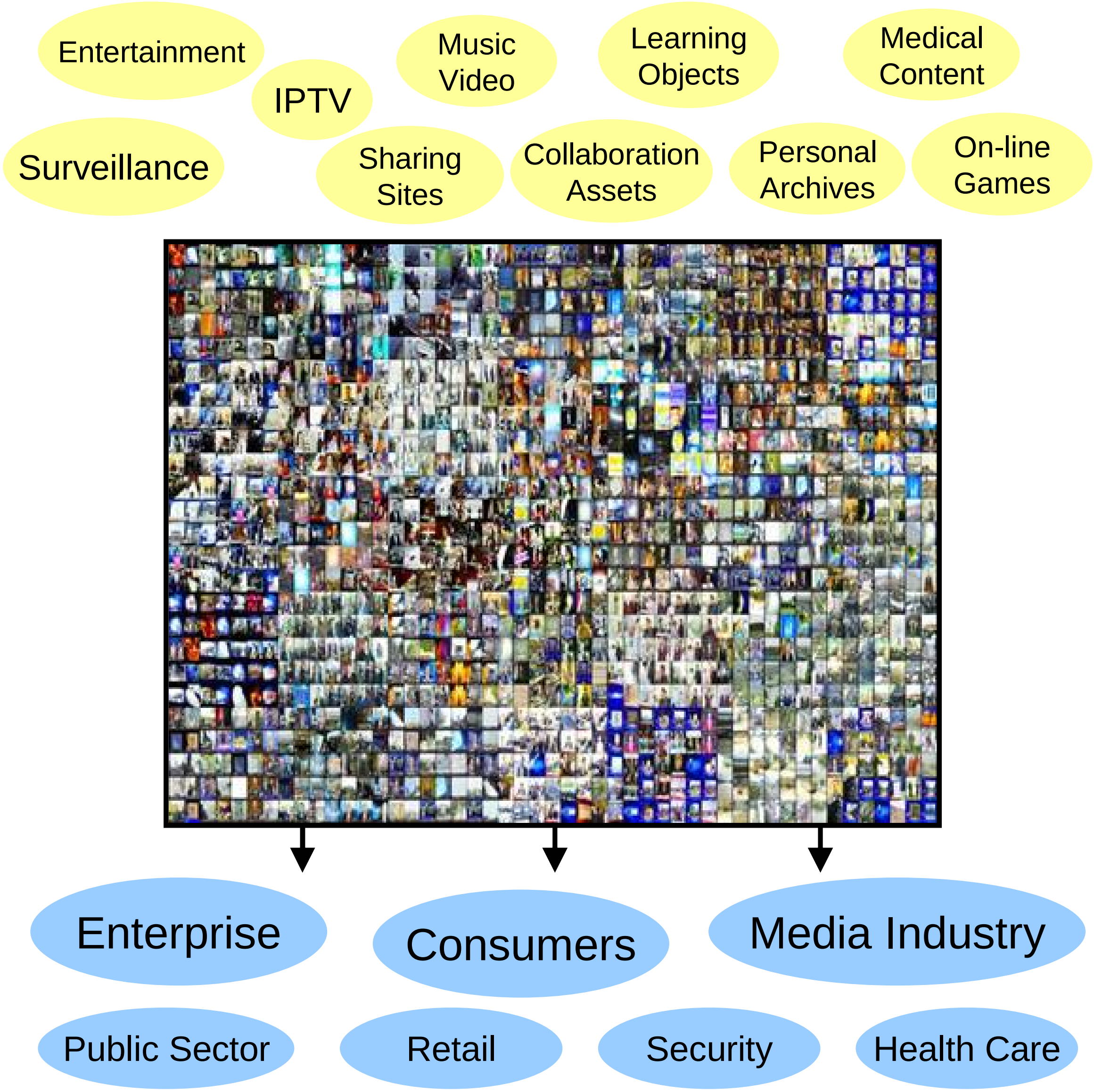
May 2008

Outline

- Challenges of video search
- Making video search better:
  - Visual recognition of content
  - Semantic labeling of visual clusters
  - Multi-modal video search
  - Concept-based video query expansion
- Video retrieval evaluations:
  - TRECVID
  - VideOlympics
- MPEG-7 multimedia content description standard

Video is rapidly becoming a regular part of our digital lives

- Growing deluge requires more effective solutions for organizing, managing & searching video content
- Manual indexing is costly, time-consuming and inadequate
- New technologies are needed to automate processing and unlock value of large repositories
- Metadata standards are needed to support interoperable search



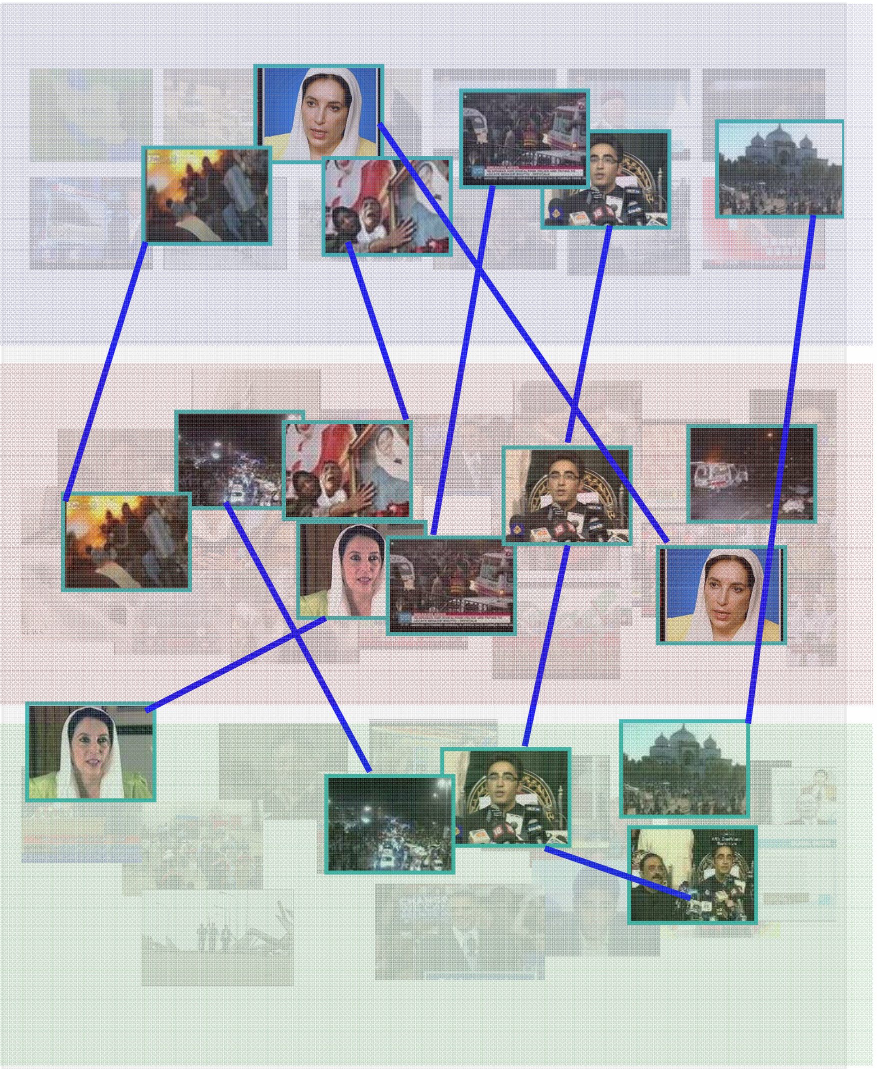
Tremendous growth of video is increasing expectation that it will be as easy to search as text

Insight Extraction Across Diverse Sources of Video and Image Content

Broadcast Video

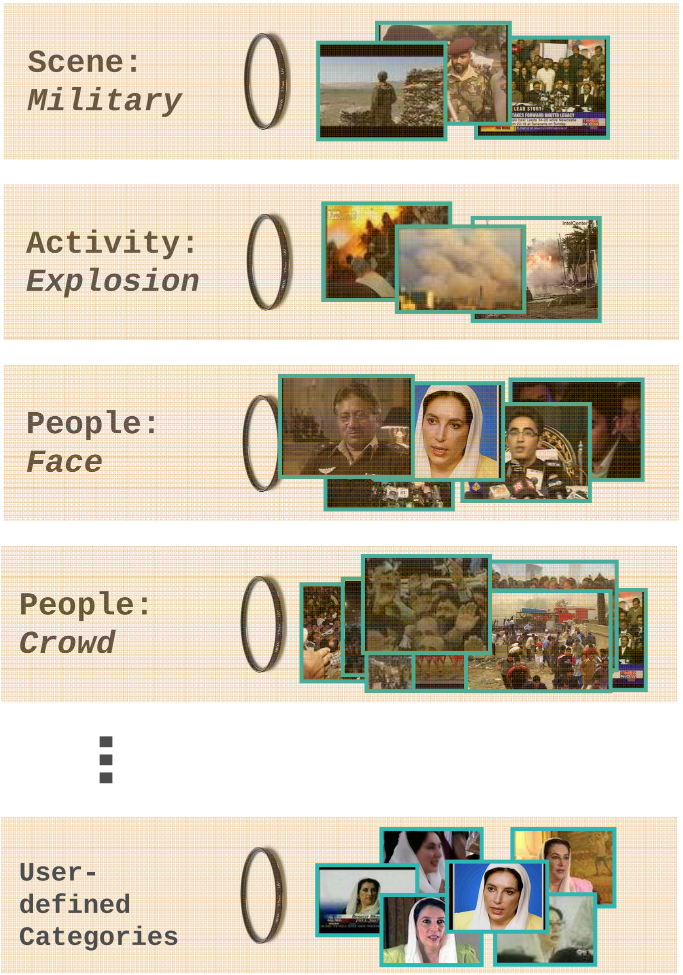
User Generated Content (UGC)

Video Blogs



Duplicate Detection and Linking

Visual Scenes Classification



Scene: *Military*

Activity: *Explosion*

People: *Face*

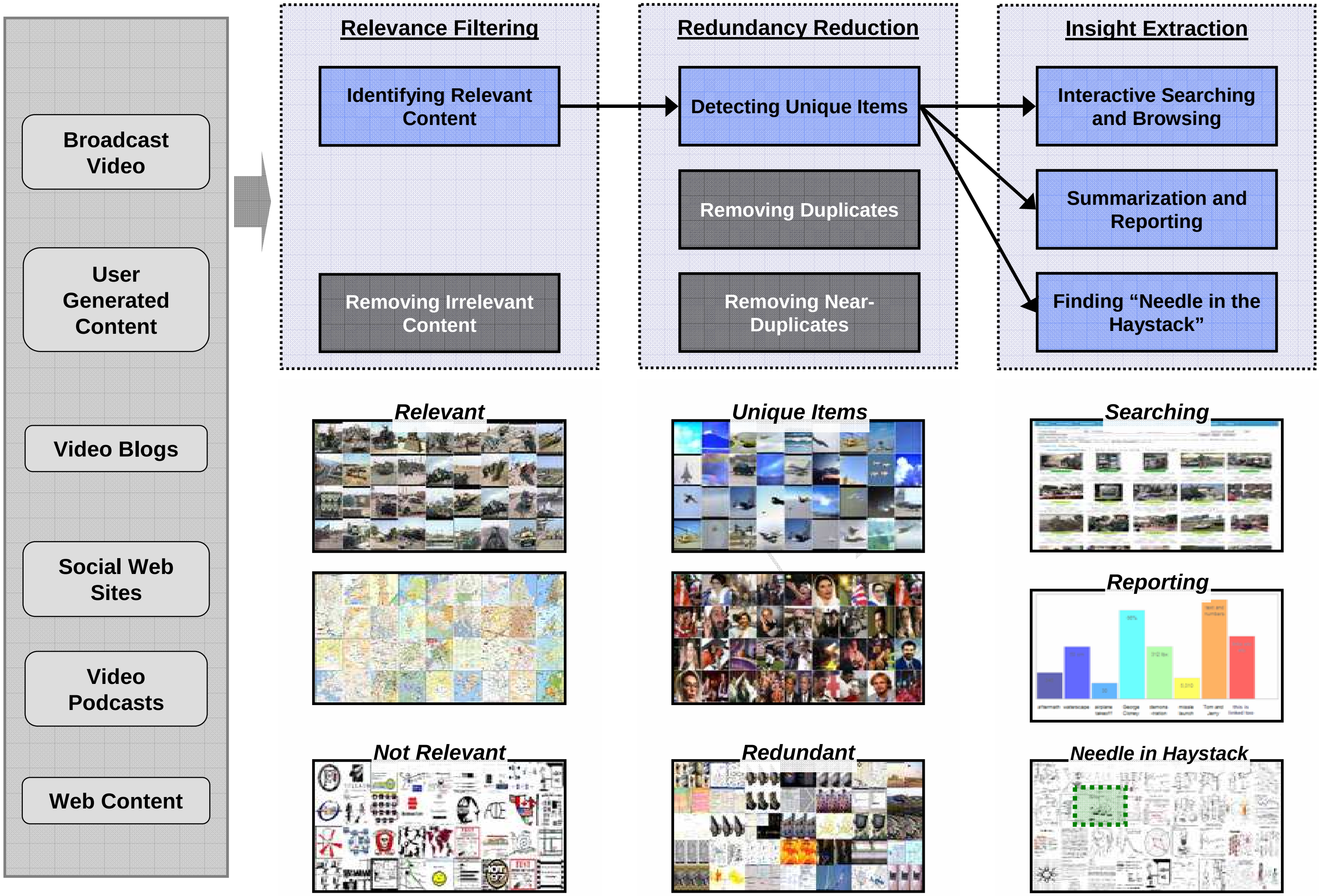
People: *Crowd*

⋮

User-defined Categories

Visual Semantic Classification

Ability to process and recognize visual semantics in video & image data can turn massive amounts of digital content into actionable intelligence

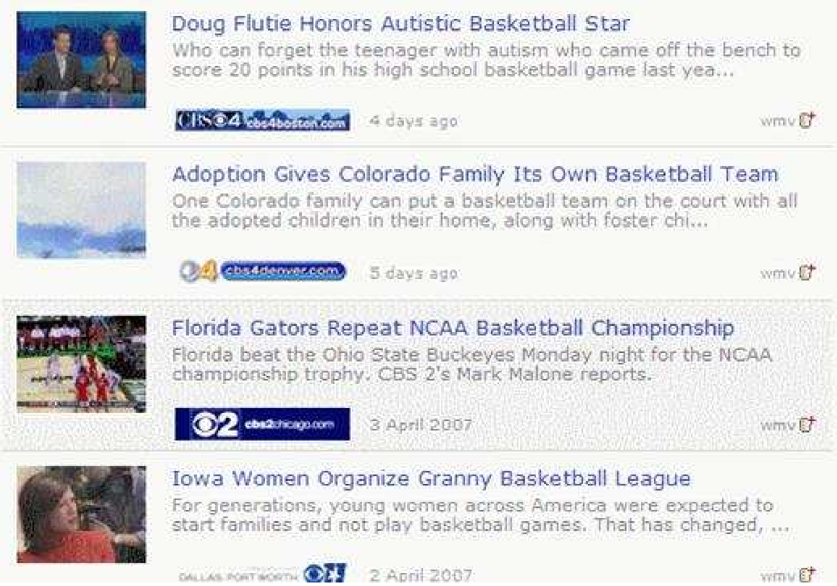


Unfortunately, it is still difficult to find relevant video content

Today's Basic Video Search is not Satisfying for Users

- Frustrating: too many videos to wade through
- Chaotic: hard to find content of interest
- Funky: cannot separate professional from UGC
- Inconsistent: video quality mixed

*www.emarketer.com



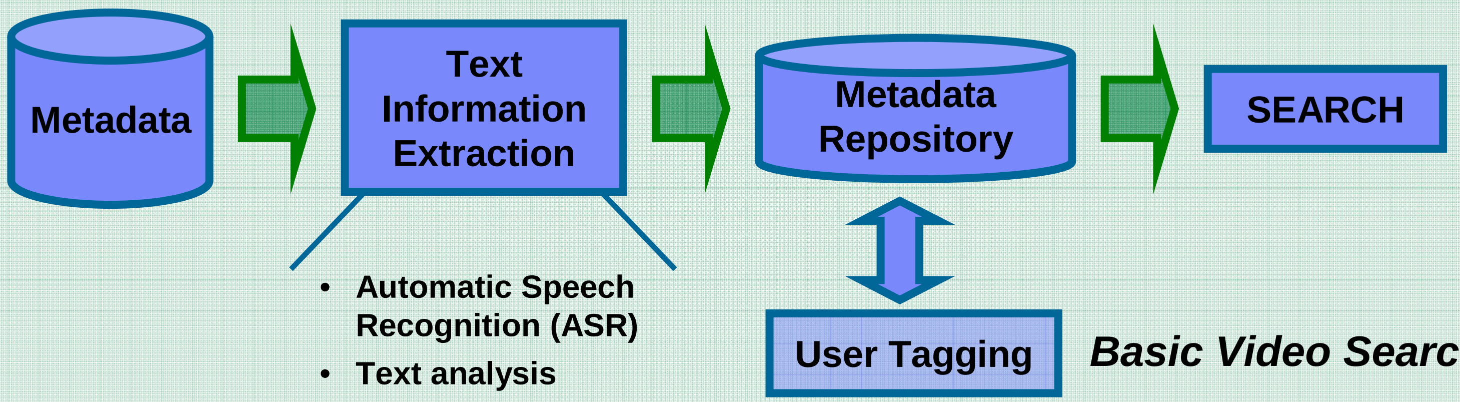
Doug Flutie Honors Autistic Basketball Star
Who can forget the teenager with autism who came off the bench to score 20 points in his high school basketball game last yea...
CBS 4 cbs4boston.com 4 days ago wmv

Adoption Gives Colorado Family Its Own Basketball Team
One Colorado family can put a basketball team on the court with all the adopted children in their home, along with foster chi...
4 cbs4denver.com 5 days ago wmv

Florida Gators Repeat NCAA Basketball Championship
Florida beat the Ohio State Buckeyes Monday night for the NCAA championship trophy. CBS 2's Mark Malone reports.
2 cbs2chicago.com 3 April 2007 wmv

Iowa Women Organize Granny Basketball League
For generations, young women across America were expected to start families and not play basketball games. That has changed, ...
DALLAS FORT WORTH 11 2 April 2007 wmv

- Program guides (EPG)
- Professional metadata
- Web text
- Audio transcripts



Metadata

Text Information Extraction

- Automatic Speech Recognition (ASR)
- Text analysis

Metadata Repository

SEARCH

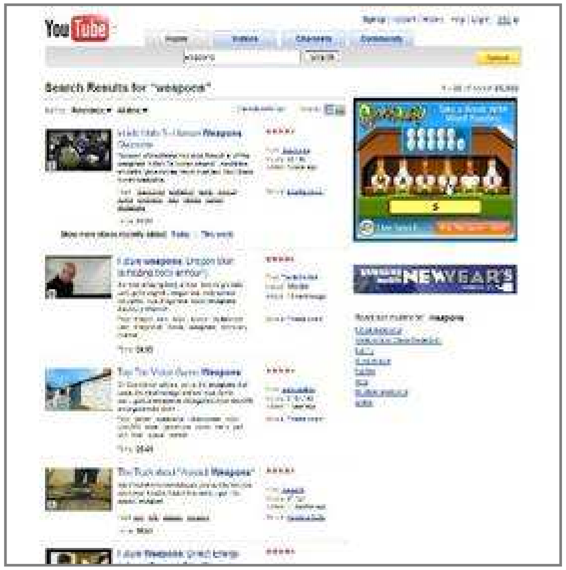
User Tagging

Basic Video Search

Today's Web-based video search is not adequate in either depth or breadth

YouTube — "Weapons"

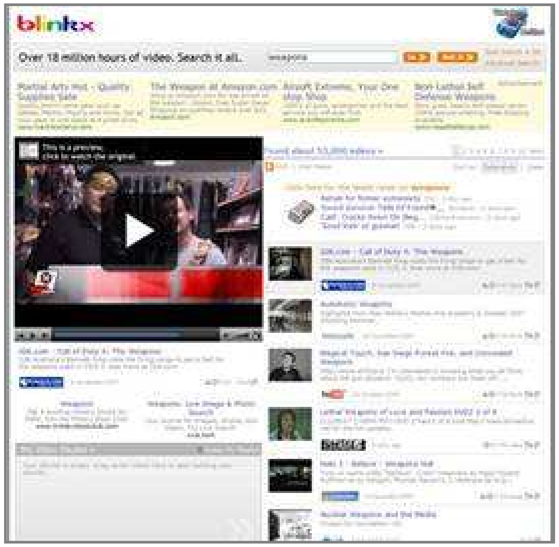
- **Depth:**
  - Cannot distinguish matches showing weapons scenes
- **Breadth:**
  - Does not broadly search the Web
  - User-generated and user provided video



Search Results for "weapons"

Blinkx — "Weapons"

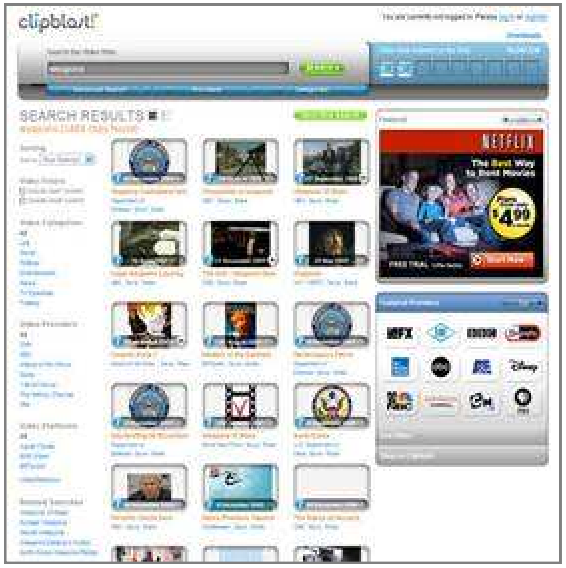
- **Depth:**
  - 53,000 matches related to "weapons"
  - No way to obtain clips showing weapons scenes
- **Breadth:**
  - Results limited to partner content



Over 18 million hours of video. Search it all.

Clipblast — "Weapons"

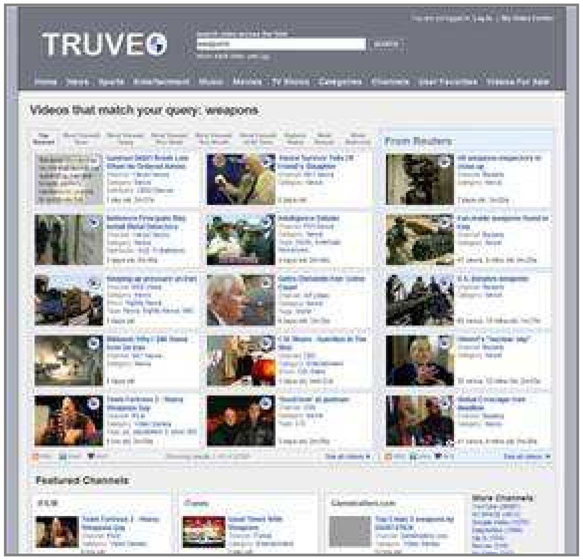
- **Depth:**
  - Search relies on text descriptions
- **Breadth:**
  - Limited to partner content



SEARCH RESULTS

TruVeo (AOL) — "Weapons"

- **Depth:**
  - No ability to refine search based on visual content
  - Search relies on text scraping from Web
- **Breadth:**
  - Preference for AOL and partner content



Videos that match your query: weapons

From Reuters

Featured Channels

All video search relies on metadata (e.g., manually authored, automatically extracted, scraped, etc.) – but, today's metadata is not good enough!!!

Issue	What's wrong
Too sparse	Few video objects have any metadata
Inadequate	Mainly tags or few keywords, program-guide info for broadcast video, speech available in few cases
Coarse-grain	At level of digital objects only
Not visual	Does not describe what is visually depicted
Ambiguous	Taxonomies not widely used; folksonomies creating new problems
Inconsistent	Vocabularies and taxonomies not standardized
Subjective	Limited verification across users
Not trustworthy	Professional metadata mixed-in with noise

Complement and Enhance Professional Cataloging and Social Tagging Approaches

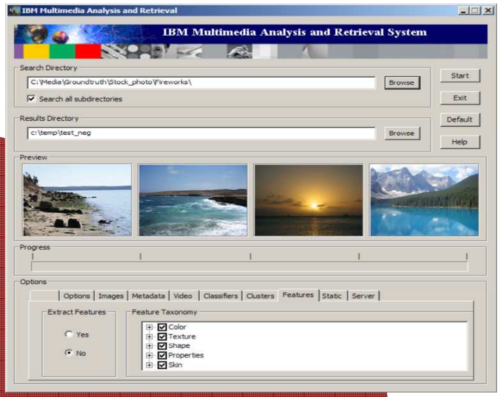
IBM Research

Manual Cataloging – By Professionals

Pros	Cons
▪ Controlled vocabularies & standard taxonomies ▪ Higher quality	▪ Costly ▪ Human resource intensive ▪ Cannot keep up
▪ Example: Fox, CNN, BBC, Broadcast TV	

Automated Tagging – By Machine

Pros	Cons
▪ Lower human cost ▪ Domain & data driven approach to semantics	▪ Requires training of models ▪ Lower quality than manual tagging
▪ Example: Marvel, Informedia, TRECVID concept detection	


IBM Multimedia Analysis and Retrieval

IBM Multimedia Analysis and Retrieval System

Search Directory: C:\Media\Groundtruth\Stock_photo\Fireworks\

☑ Search all subdirectories

Results Directory: c:\temp\test_neg

Preview

Progress

Options

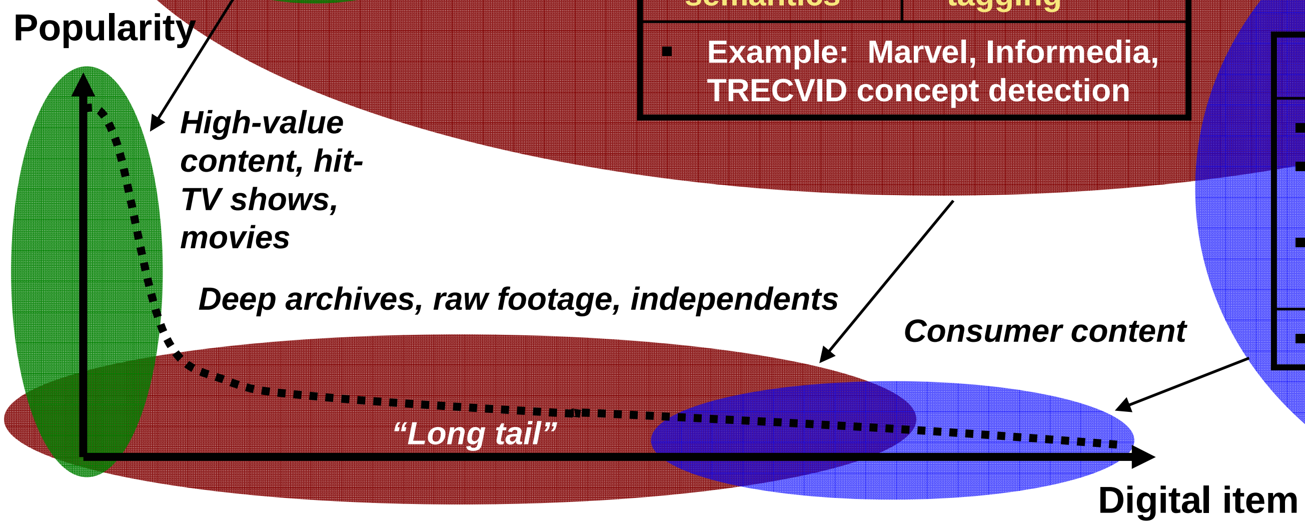
Extract Features: ○ Yes ● No

Feature Taxonomy: ☑ Color ☑ Texture ☑ Shape ☑ Properties ☑ Skin

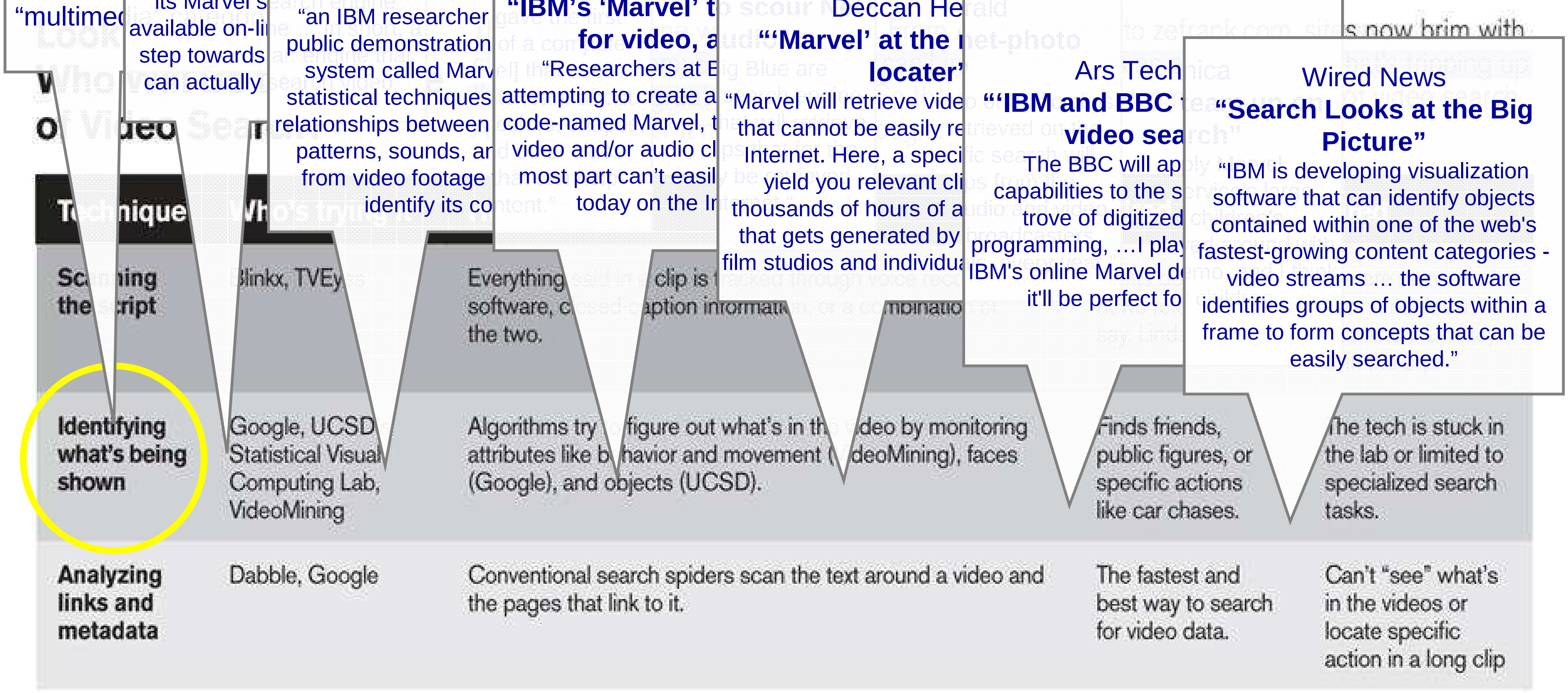
Browse
Browse
Start
Exit
Default
Help
Options | Images | Metadata | Video | Classifiers | Clusters | Features | Static | Server

Social Tagging – By Users

Pros	Cons
▪ User driven ▪ Emergent folksonomies ▪ Serpendipitous browsing	▪ Ambiguity ▪ Uncontrolled vocabulary ▪ Synonyms
▪ Examples: Del.icio.us and Flickr	


Popularity
High-value content, hit-TV shows, movies
Deep archives, raw footage, independents
Consumer content
“Long tail”
Digital item

The market is seeking a new approach for effective video search



Wall Street Journal
IBM system wins Wall Street Journal Innovation Award in "multimedia" category

The Economist
"Google Meet TiVo"
"IBM made a test version of its Marvel search engine available on-line ... step towards an engine that can actually search video."

Information Week
"In Search of Better Video Search"
"an IBM researcher gave the first public demonstration of a computer system called Marvel that uses statistical techniques ... relationships between patterns, sounds, and ... from video footage that ... identify its content."

c/Net
"IBM's 'Marvel' to scour Net for video, audio"
"Researchers at Big Blue are attempting to create a search engine, code-named Marvel, that will retrieve video and/or audio clips that for the most part can't easily be retrieved today on the Internet."

Deccan Herald
"'Marvel' at the net-photo locater'"
"Marvel will retrieve video that cannot be easily retrieved on the Internet. Here, a specific search will yield you relevant clips from the thousands of hours of audio and video that gets generated by broadcasters, film studios and individuals every year."

Ars Technica
"'IBM and BBC team up on video search"
The BBC will apply Marvel capabilities to the services large trove of digitized children's programming, ...I played around with IBM's online Marvel demo, and I think it'll be perfect for children

Wired News
"Search Looks at the Big Picture"
"IBM is developing visualization software that can identify objects contained within one of the web's fastest-growing content categories - video streams … the software identifies groups of objects within a frame to form concepts that can be easily searched."

Technique	Who's trying it			
Scanning the script	Blinkx, TVEyes	Everything said in a clip is tracked through voice recognition software, closed-caption information, or a combination of the two.		
Identifying what's being shown	Google, UCSD's Statistical Visual Computing Lab, VideoMining	Algorithms try to figure out what's in the video by monitoring attributes like behavior and movement (VideoMining), faces (Google), and objects (UCSD).	Finds friends, public figures, or specific actions like car chases.	The tech is stuck in the lab or limited to specialized search tasks.
Analyzing links and metadata	Dabble, Google	Conventional search spiders scan the text around a video and the pages that link to it.	The fastest and best way to search for video data.	Can't "see" what's in the videos or locate specific action in a long clip

IBM Multimedia Analysis and Retrieval System is recognized as leader in research and development of break-through techniques for video content-based analysis and search

Progress in Multimedia Content-Based Retrieval

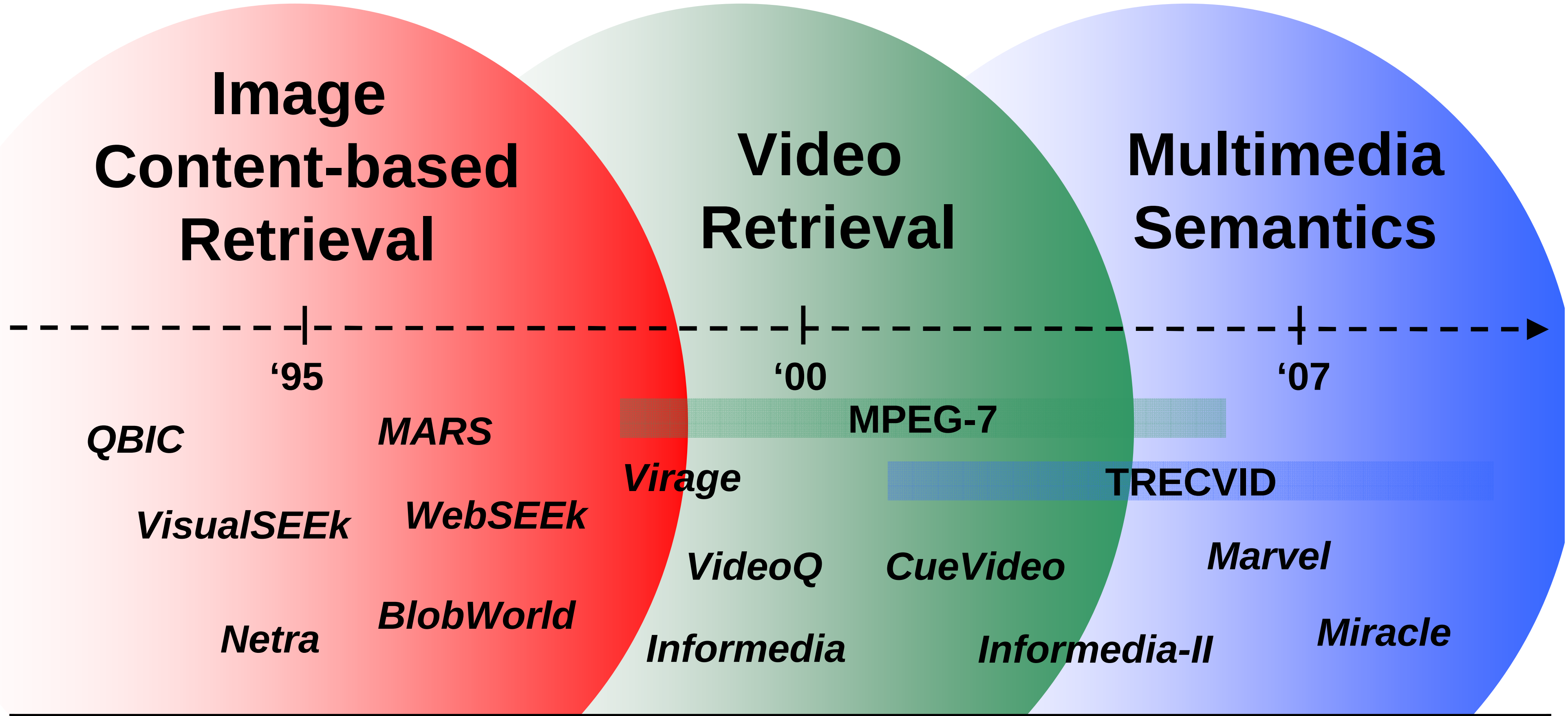


Image Content-based Retrieval

Video Retrieval

Multimedia Semantics

'95

'00

'07

QBIC

MARS

VisualSEEk

WebSEEk

Netra

BlobWorld

MPEG-7

Virage

TRECVID

VideoQ

CueVideo

Informedia

Informedia-II

Marvel

Miracle

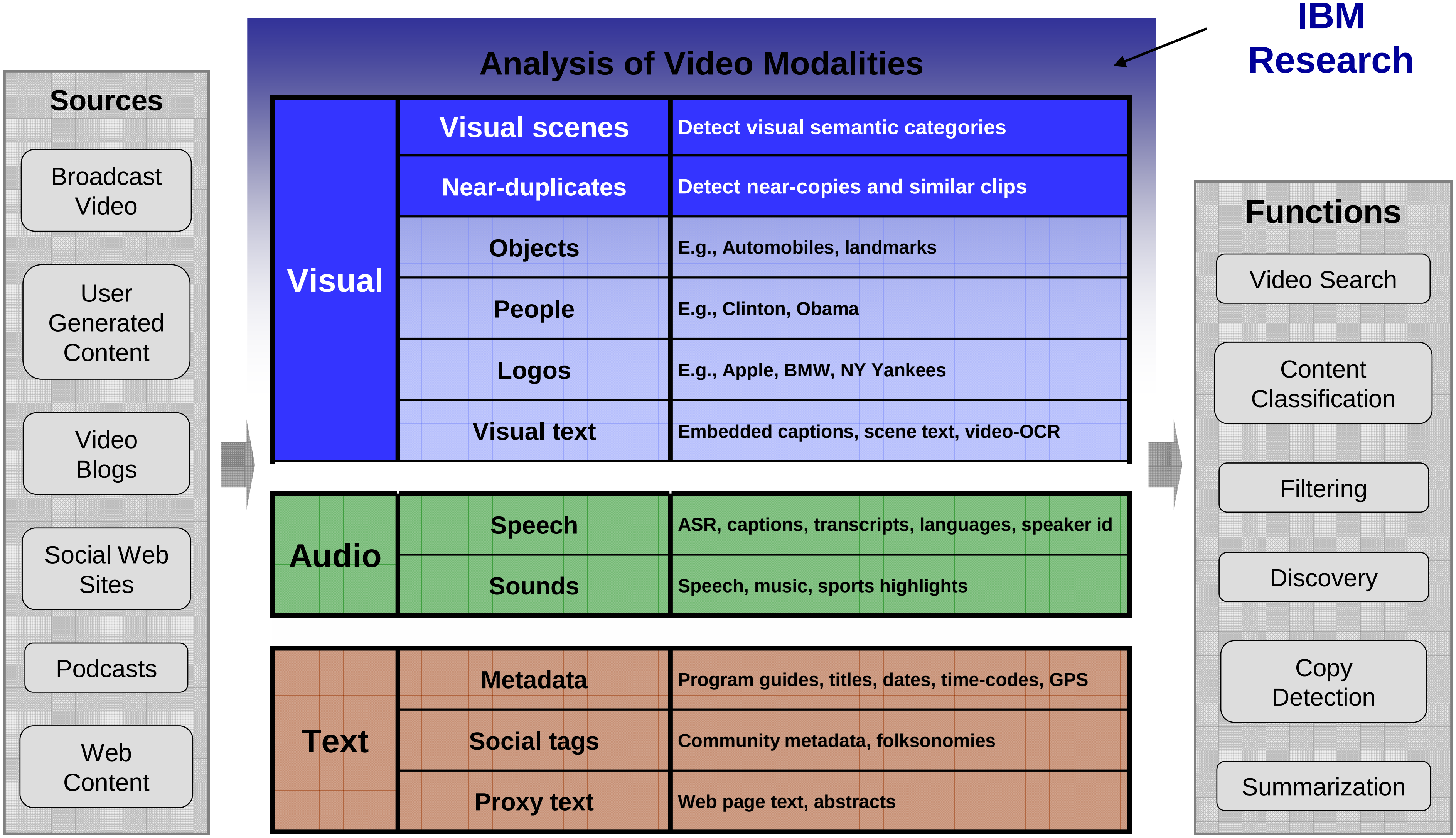
- Visual feature extraction (color, texture, shape)
- Relevance feedback
- Spatial region search

- Shot boundary detection
- Speech indexing
- Motion extraction
- Video summarization

- Semantics learning
- Scene classification
- Semantic clustering
- Search fusion

Bridging the Semantic Gap:
Analyze visual features and apply machine learning techniques to classify video scenes automatically

Making sense of the digital video chaos requires extracting meaningful information across multiple modalities (visual, audio, text, speech)



IBM Research

Sources

- Broadcast Video
- User Generated Content
- Video Blogs
- Social Web Sites
- Podcasts
- Web Content

Analysis of Video Modalities

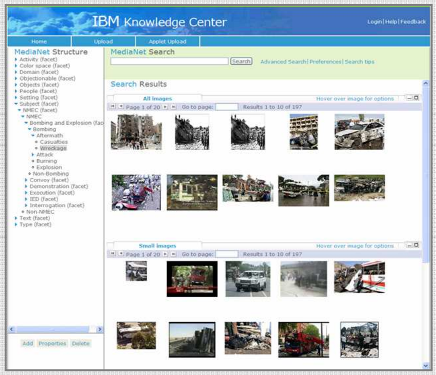
Visual	Visual scenes	Detect visual semantic categories
	Near-duplicates	Detect near-copies and similar clips
	Objects	E.g., Automobiles, landmarks
	People	E.g., Clinton, Obama
	Logos	E.g., Apple, BMW, NY Yankees
	Visual text	Embedded captions, scene text, video-OCR
Audio	Speech	ASR, captions, transcripts, languages, speaker id
	Sounds	Speech, music, sports highlights
Text	Metadata	Program guides, titles, dates, time-codes, GPS
	Social tags	Community metadata, folksonomies
	Proxy text	Web page text, abstracts

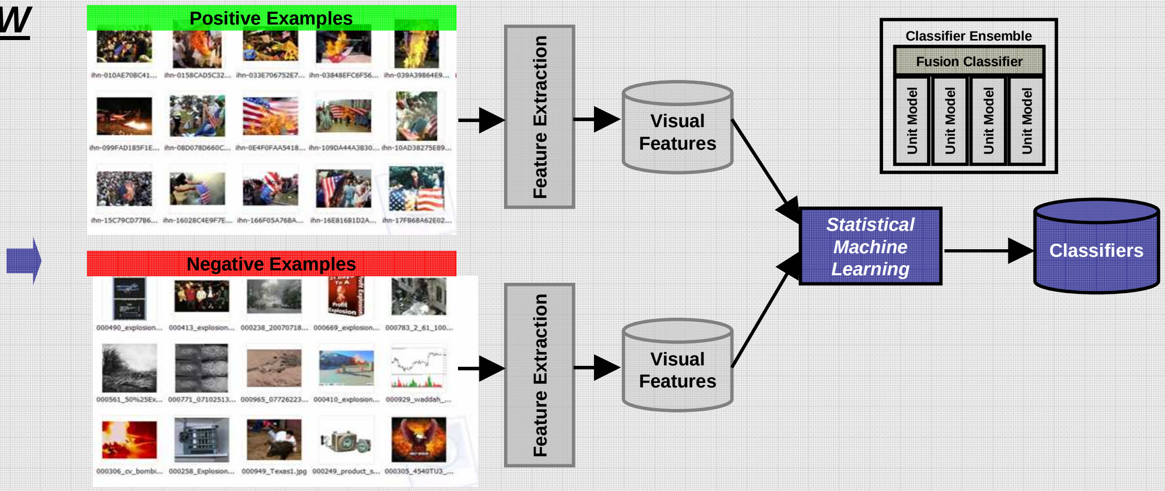
Functions

- Video Search
- Content Classification
- Filtering
- Discovery
- Copy Detection
- Summarization

Marvel – Software for learning visual categories and classifying and recognizing image/video content

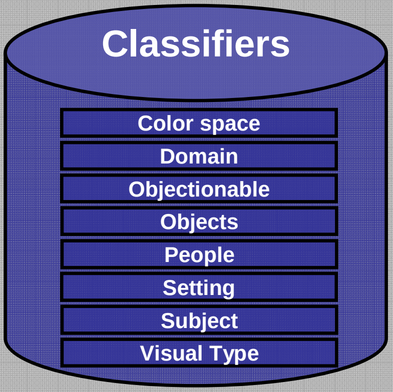
Image/Video Content Learning SW



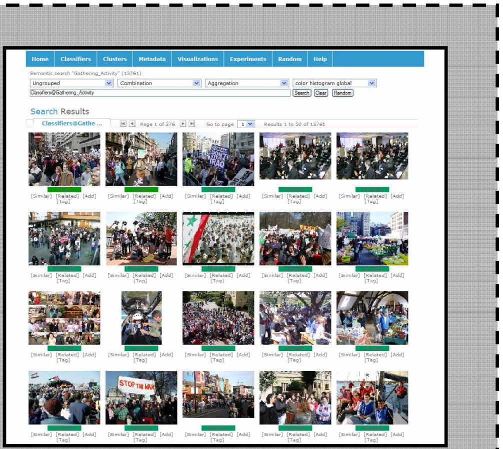


Image/Video Classification Software

Unknown images/video

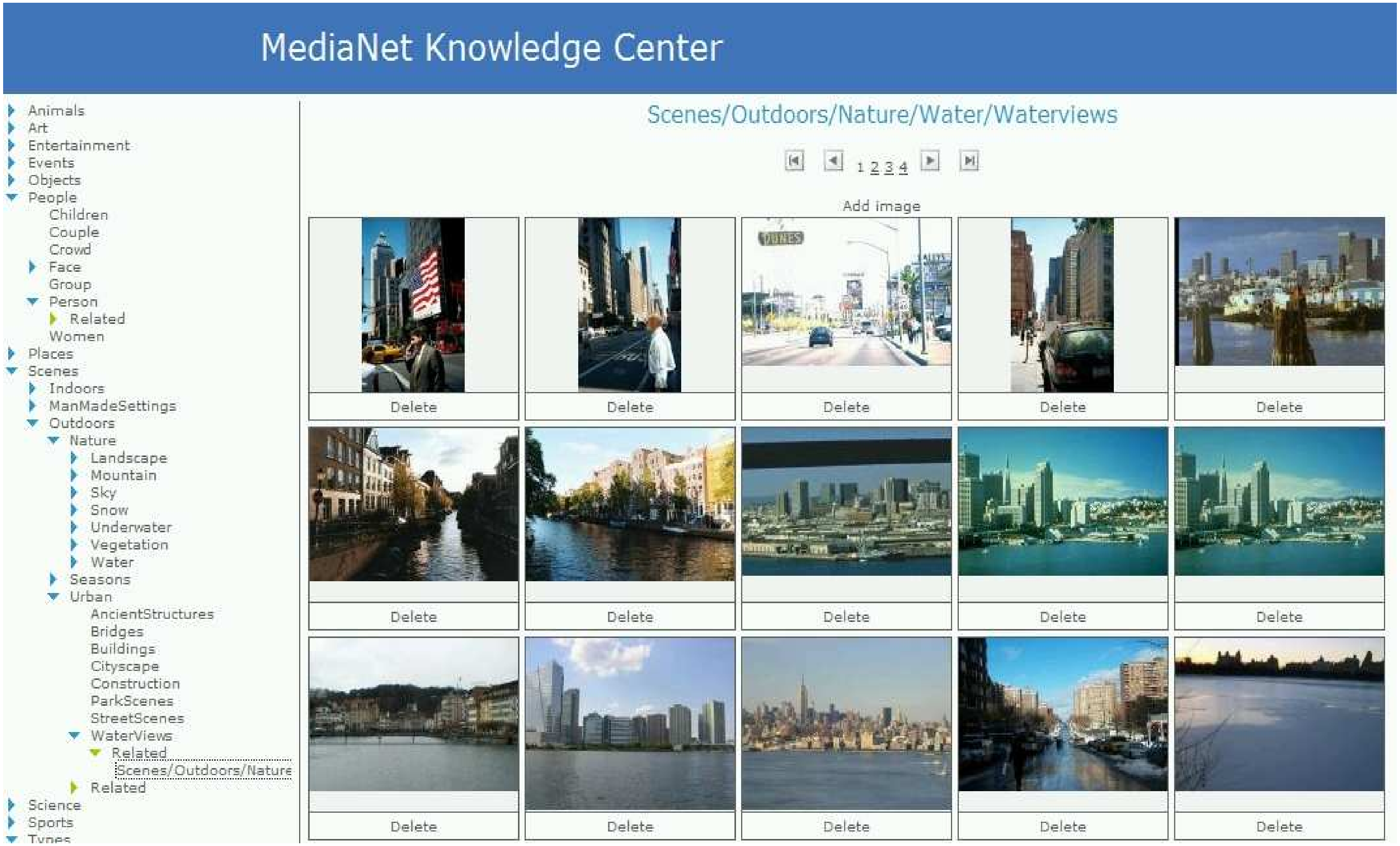


Taxonomy



Classified images/video

Semantic models are created from training examples that are managed using multimedia taxonomies



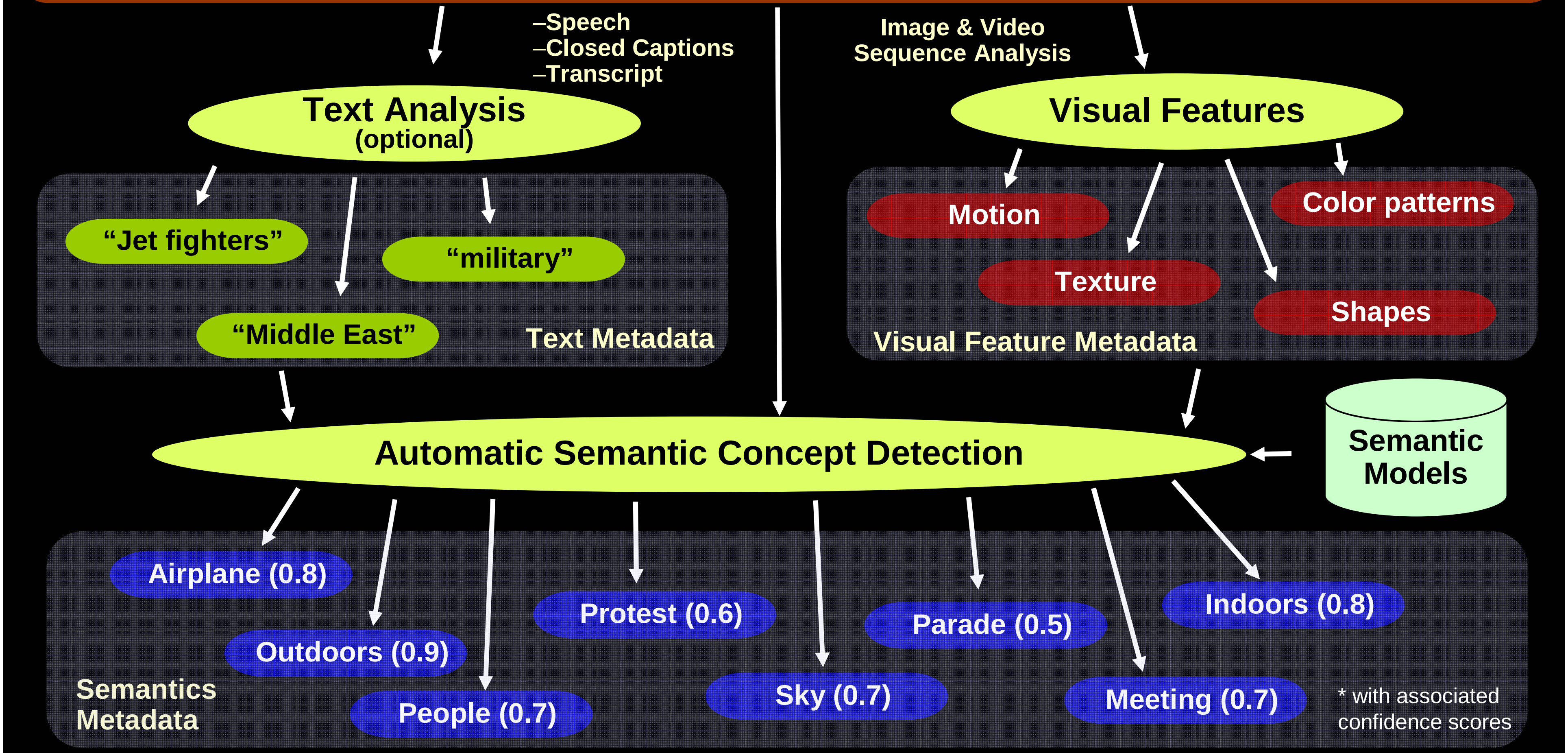
MediaNet Knowledge Center

Scenes/Outdoors/Nature/Water/Waterviews

Semantic Tagging of Multimedia Content

Video:

Associated speech: Today, **jet fighters** practiced maneuvers and forces increased **military** preparations as tensions in **Middle East** reached …



–Speech
–Closed Captions
–Transcript

Image & Video Sequence Analysis

Text Analysis (optional)

“Jet fighters”

“military”

“Middle East”

Text Metadata

Visual Features

Motion

Color patterns

Texture

Shapes

Visual Feature Metadata

Automatic Semantic Concept Detection

Semantic Models

Airplane (0.8)

Outdoors (0.9)

People (0.7)

Protest (0.6)

Sky (0.7)

Parade (0.5)

Meeting (0.7)

Indoors (0.8)

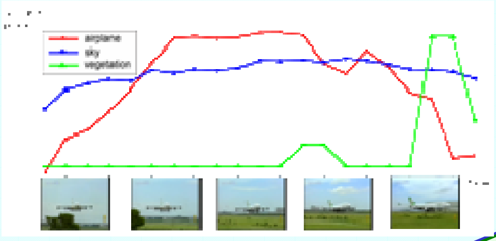
Semantics Metadata

* with associated confidence scores

Multimedia Semantic Analysis and Search

Semantics Modeling:

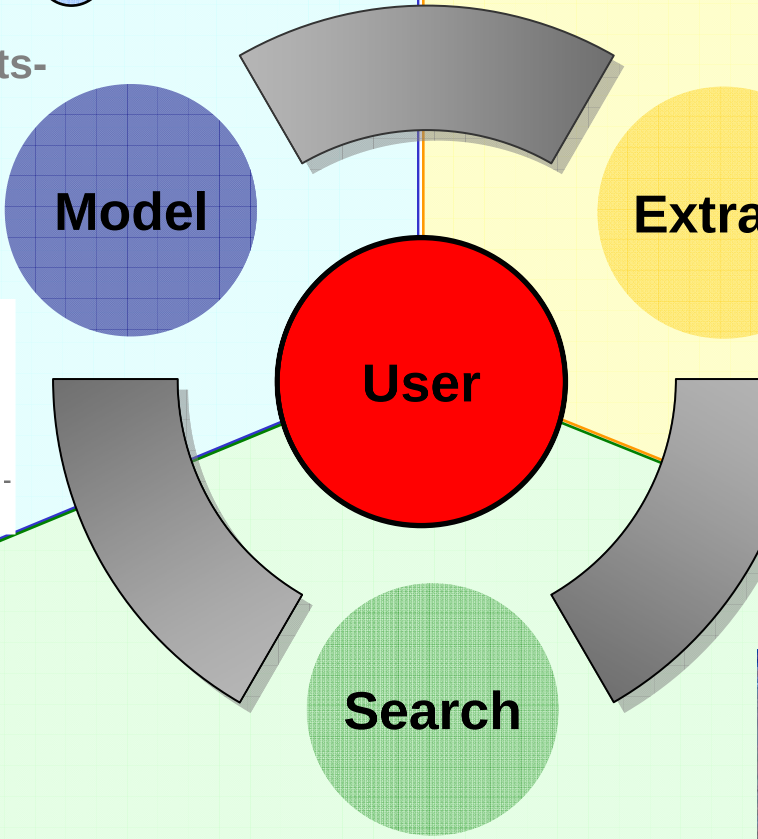
- Modeling Large-scale Semantic Spaces 1
- Multi-Granular & Parts-based Modeling
- Temporal Event Semantic Modeling
- Sequence Rhythm



Content Extraction:

- 2 Semantic Labeling of Visual Clusters
- Temporal Pattern Mining
- Cross-Channel Topic Tracking

Nature, Day, Outdoors



Searching, Browsing & Interaction:

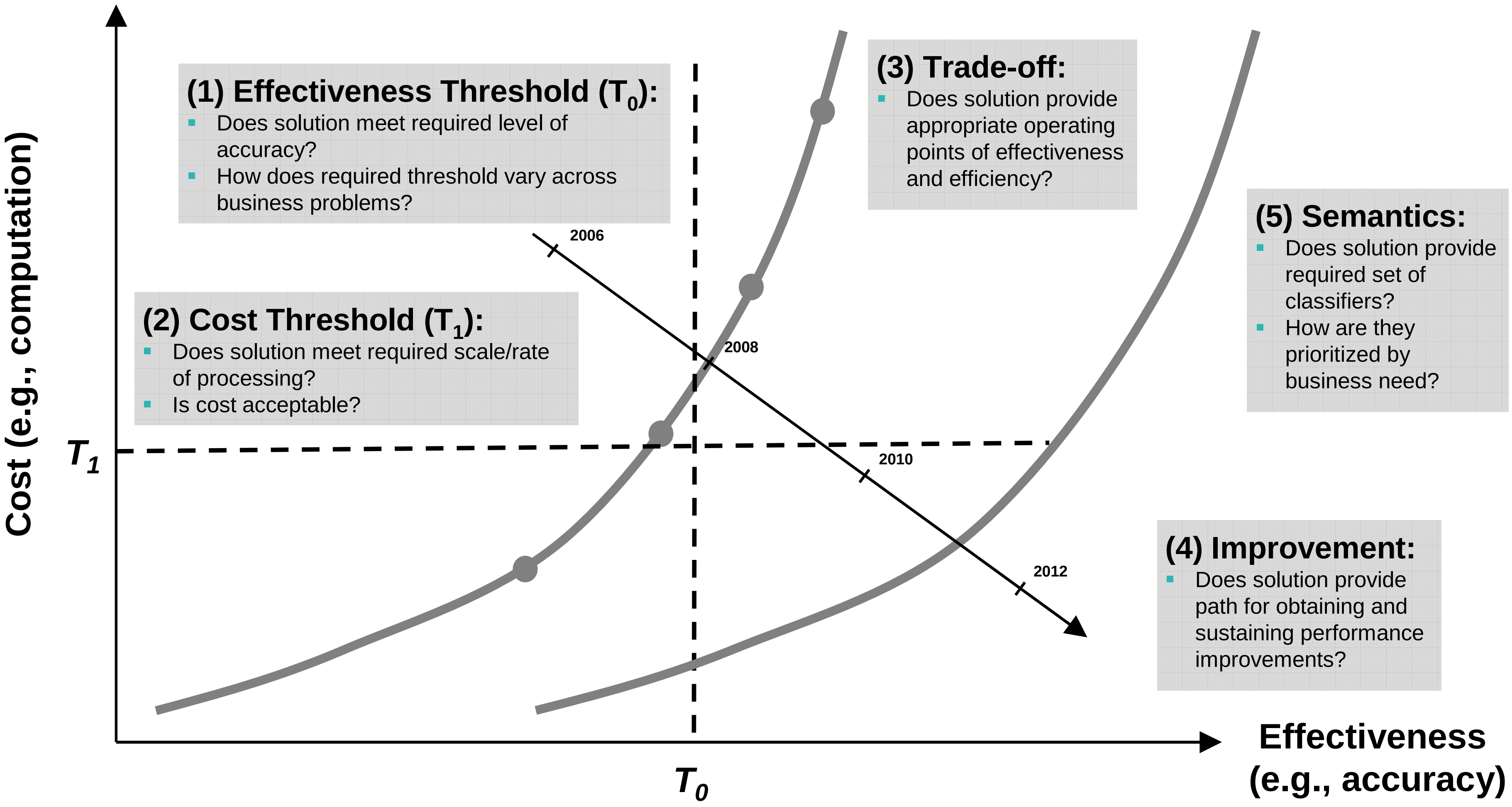
- Multi-Modal Search and Retrieval 3
- Query Expansion for Multimodal Video Retrieval 4
- Query-Class Dependent Video Search

MARVel, MPEG-7 Video Search Engine (IBM T. J. Watson Research Cent...
MARVel

(1) Modeling Large-scale Semantic Spaces

Challenge with any nascent technology is to fit to suitable problem set

- **Five Dimensions of Visual Recognition Performance**



(1) Effectiveness Threshold (T_0):

- Does solution meet required level of accuracy?
- How does required threshold vary across business problems?

(2) Cost Threshold (T_1):

- Does solution meet required scale/rate of processing?
- Is cost acceptable?

(3) Trade-off:

- Does solution provide appropriate operating points of effectiveness and efficiency?

(5) Semantics:

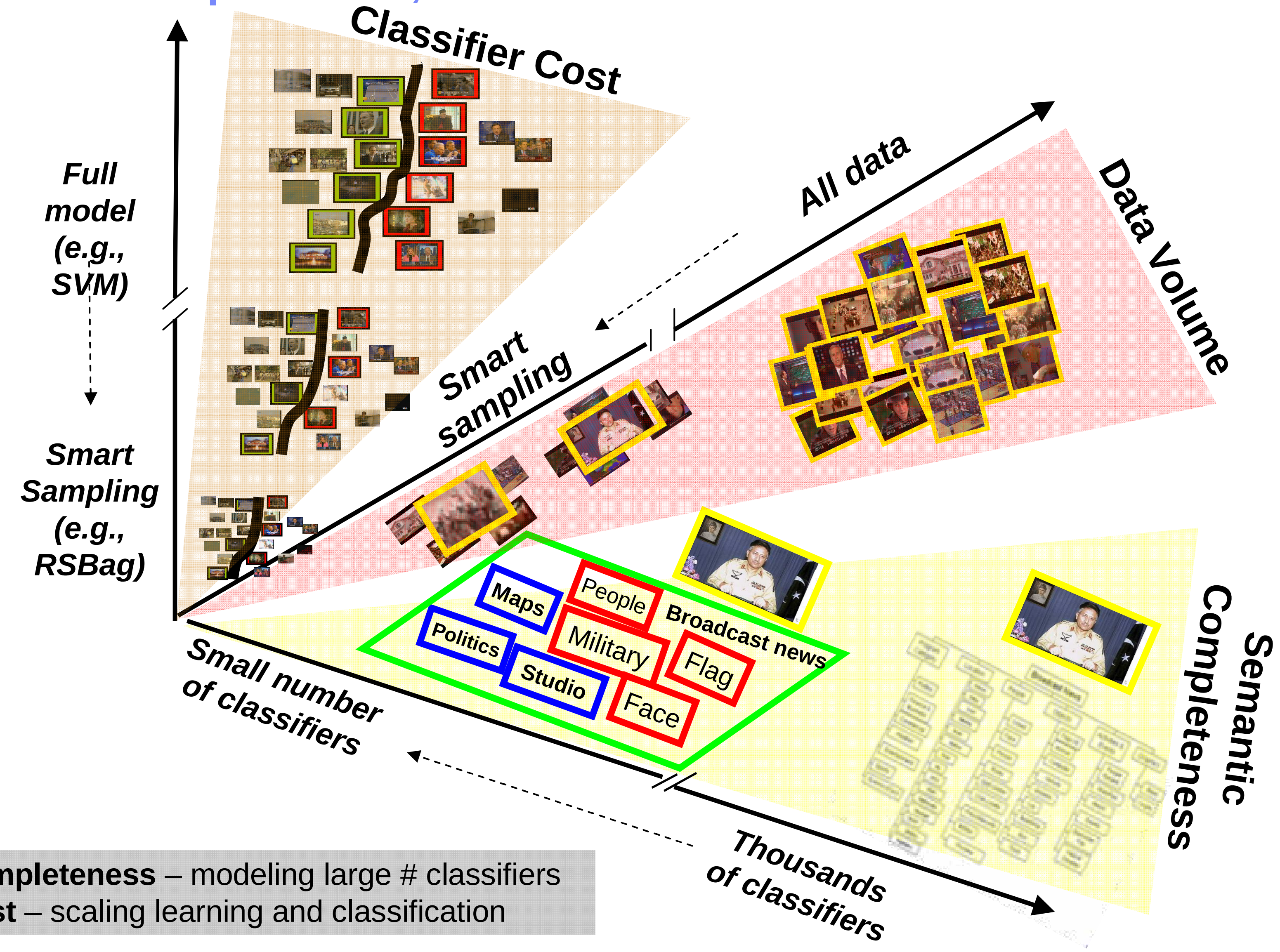
- Does solution provide required set of classifiers?
- How are they prioritized by business need?

(4) Improvement:

- Does solution provide path for obtaining and sustaining performance improvements?

Scalability in Visual Semantic Classification is Achieved by Trading-off Semantic Completeness, Classifier Costs and Data Volumes

Smart feature sampling attains 50-100x speed-up in learning and classification

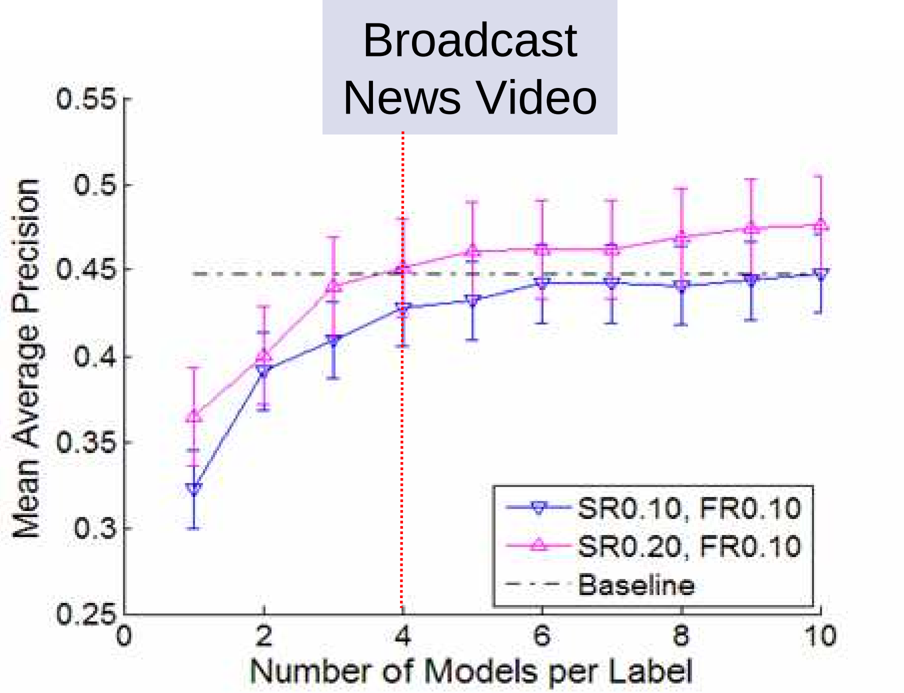


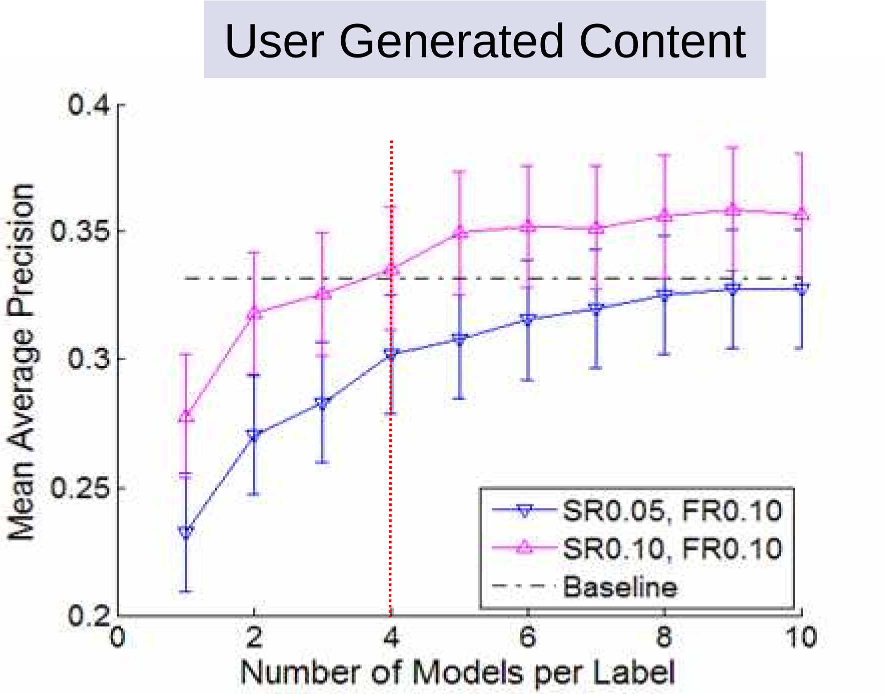
- **Semantic Completeness** – modeling large # classifiers
- **Classifier Cost** – scaling learning and classification

Smart feature sampling during learning of visual semantic classifiers allows efficient scaling to large number of video semantic classifiers

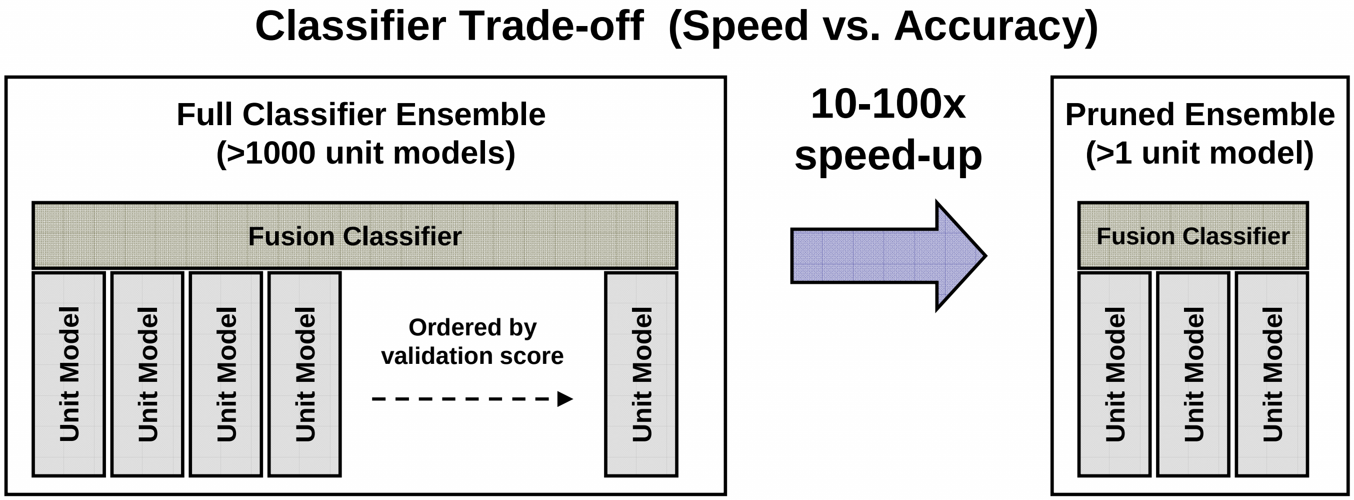
- Smart feature sampling of features greatly speeds-up learning and classification
- Easy-to-use trade-off of classification accuracy and computation
- Unit models can be leveraged across multiple semantic concepts for greater efficiency

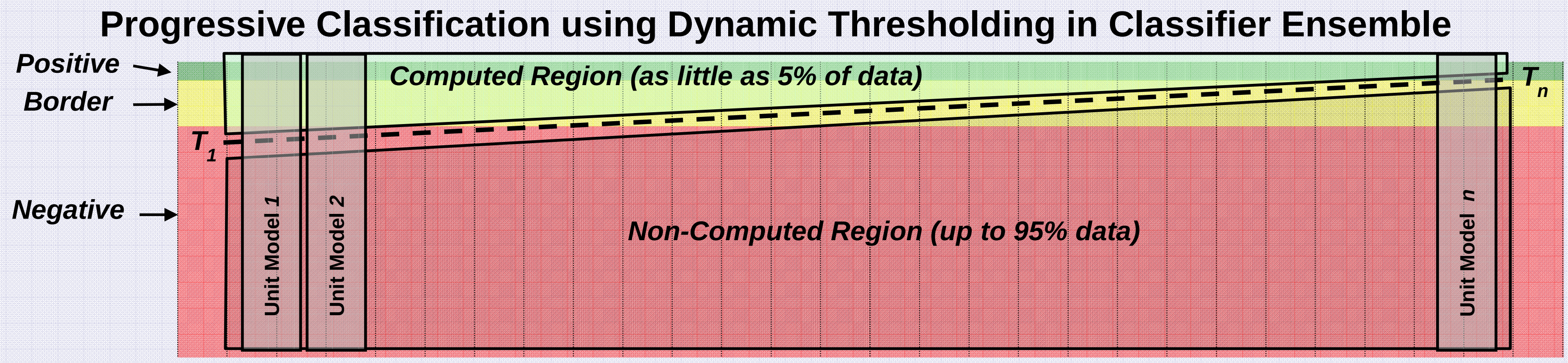
*** Classification Accuracy reaches high value using small number of unit models**


Broadcast News Video


User Generated Content

IBM’s solution uses a highly granular ensemble classifier approach built on 140 visual descriptors that supports large-scale processing through progressive classification and run-time trade-off in accuracy and speed

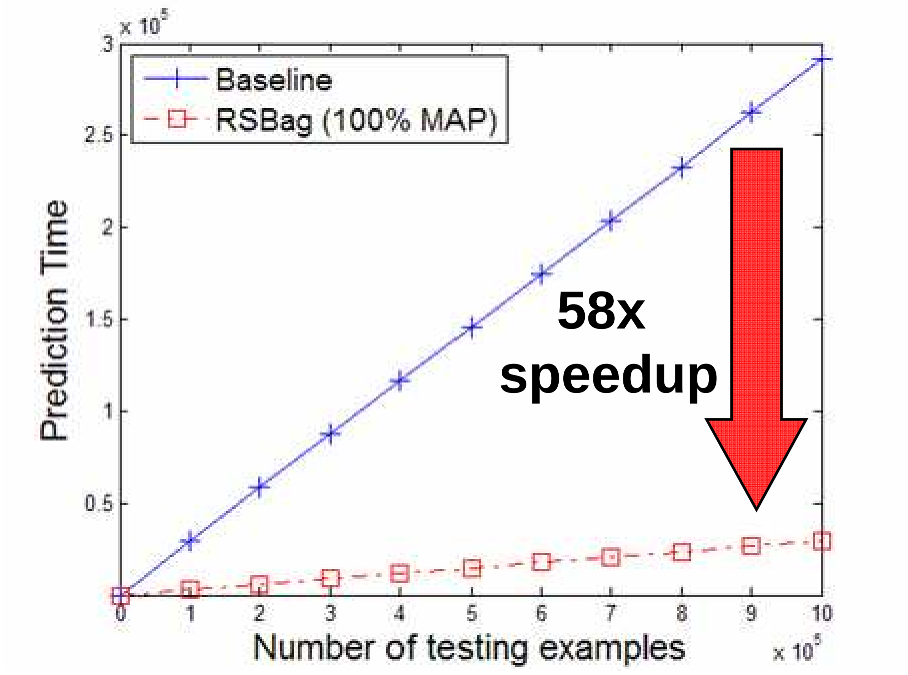



Progressive Classification using Dynamic Thresholding in Classifier Ensemble
Positive
Border
Negative
Computed Region (as little as 5% of data)
Non-Computed Region (up to 95% data)
T_1
T_n
Unit Model 1
Unit Model 2
Unit Model n

Significant speed-up in learning makes it possible to learn new visual semantic models in near real-time as needed

Broadcast News Video

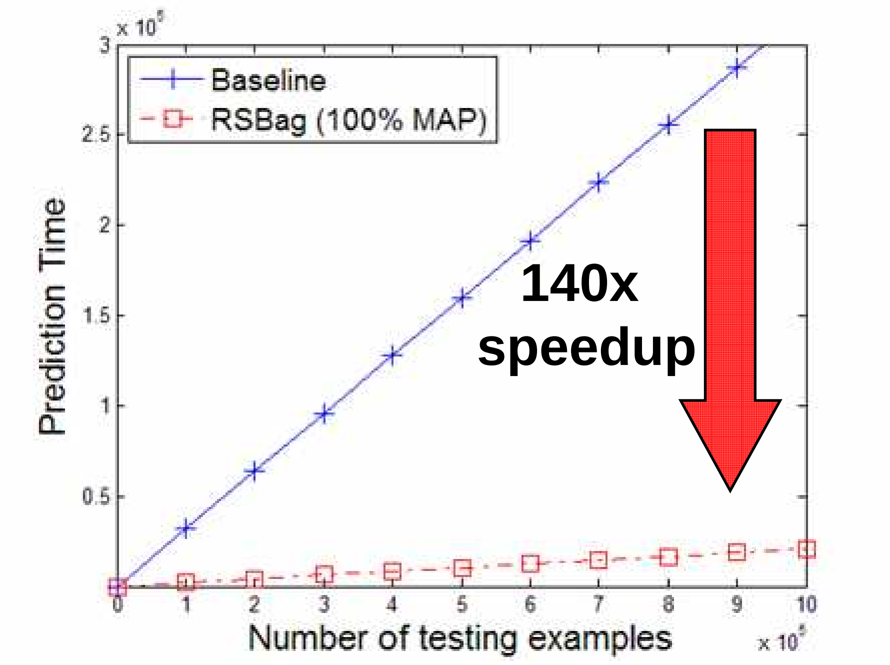
	All data	Smart Sampling
Training Time	*98 min*	*101 sec*



58x speedup

User Generated Content

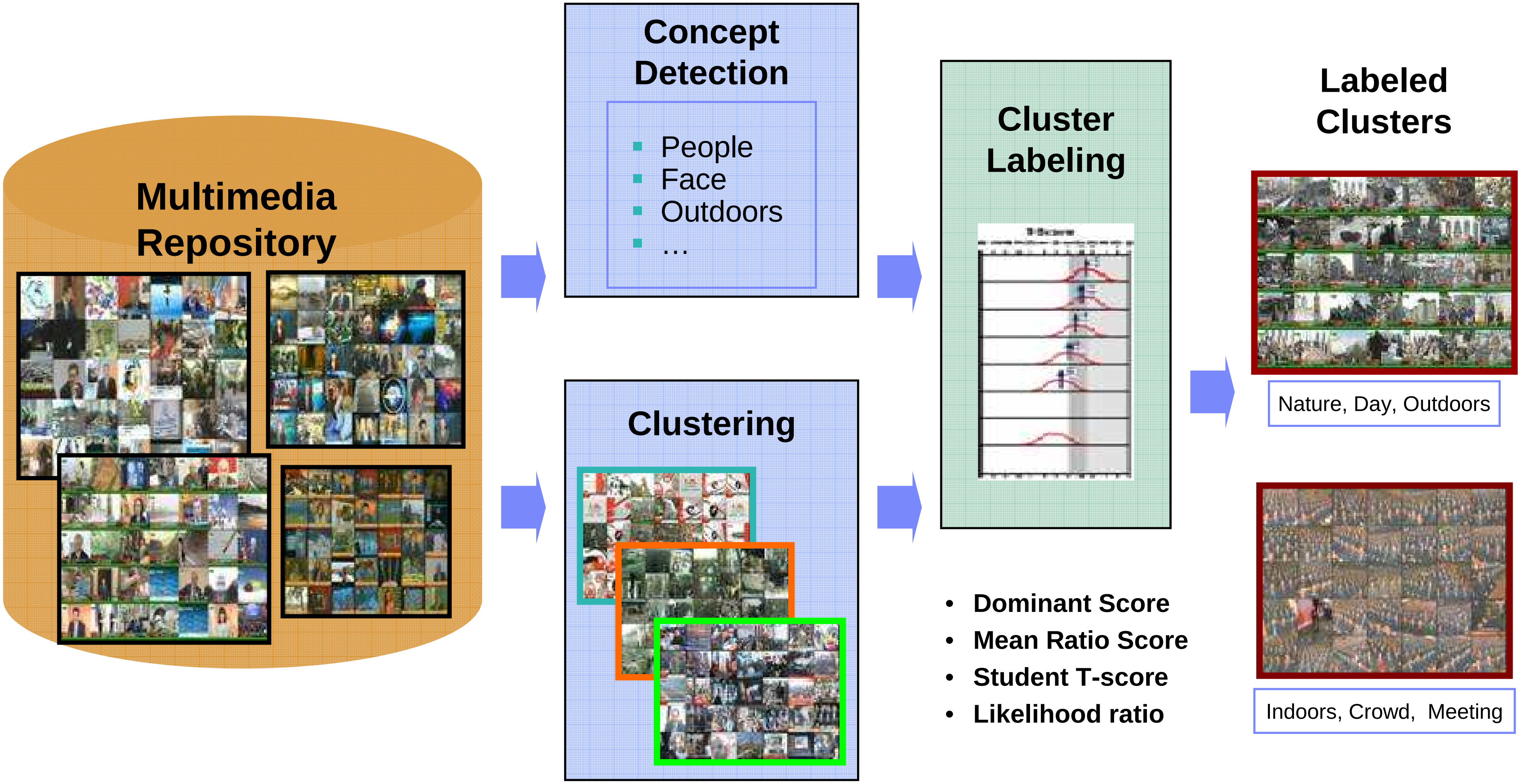
	All data	Smart Sampling
Training Time	*85 min*	*36 sec*



140x speedup

(2) Semantic Labeling of Visual Clusters

Semantic Labeling of Visual Clusters – Discovering Descriptive & Discriminative Semantics (*ICME-2006*)



Multimedia Repository

Concept Detection

- People
- Face
- Outdoors
- …

Clustering

Cluster Labeling

- **Dominant Score**
- **Mean Ratio Score**
- **Student T-score**
- **Likelihood ratio**

Labeled Clusters

Nature, Day, Outdoors

Indoors, Crowd, Meeting

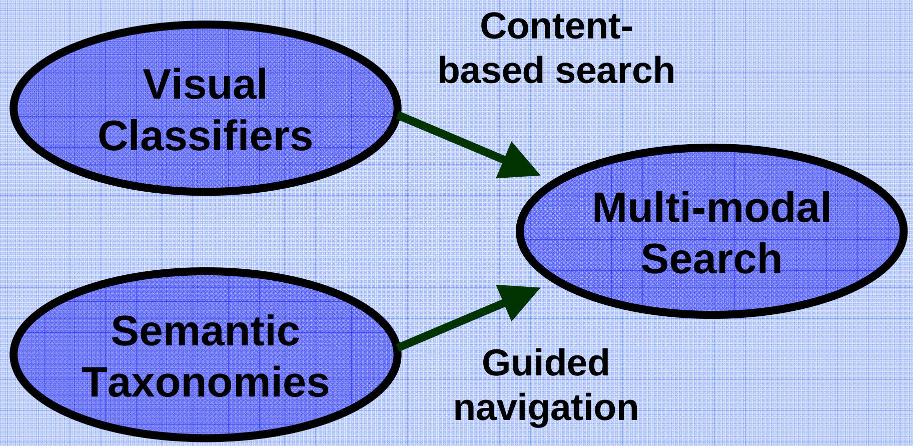
(3) Multi-Modal Search and Retrieval

IBM's "content-based" approach improves video analysis by classifying scenes visually and allows multi-modal search of video content

Sources

- Broadcast Video
- Video Blogs
- User Generated Content
- Social Web Sites
- Podcasts
- Web Content

Advanced multi-modal search


Visual Classifiers

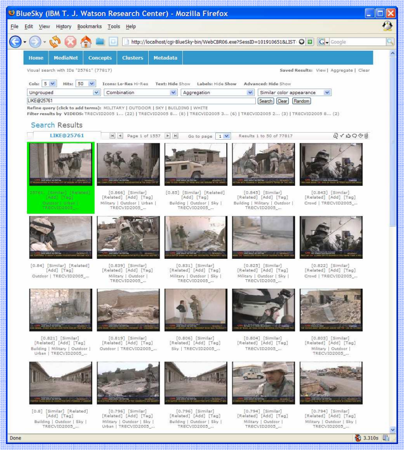
Semantic Taxonomies

Content-based search

Guided navigation

Multi-modal Search

- *More accurate (>2.5x)*
- *Faster (>250x)*
- *Precise matches within clips*


BlueSky (IBM T. J. Watson Research Center) - Mozilla Firefox

http://localhost/cgi-BlueSky-bin/WebCBR06.exe?SessID=101910651&LIST

Visual search with IDs "25761" (77817)

Saved Results: View | Aggregate | Clear

Cols: 5 Hits: 50 Icons: Lo-Res Hi-Res Text: Hide Show Labels: Hide Show Advanced: Hide Show

Ungrouped | Combination | Aggregation | Similar color appearance

LIKE@25761

Refine query (click to add terms): MILITARY | OUTDOOR | SKY | BUILDING | WHITE

Filter results by VIDEOS: TRECVID2005 1... (22) | TRECVID2005 8... (8) | TRECVID2005 3... (6) | TRECVID2005 2... (3) | TRECVID2005 8... (2)

Search Results

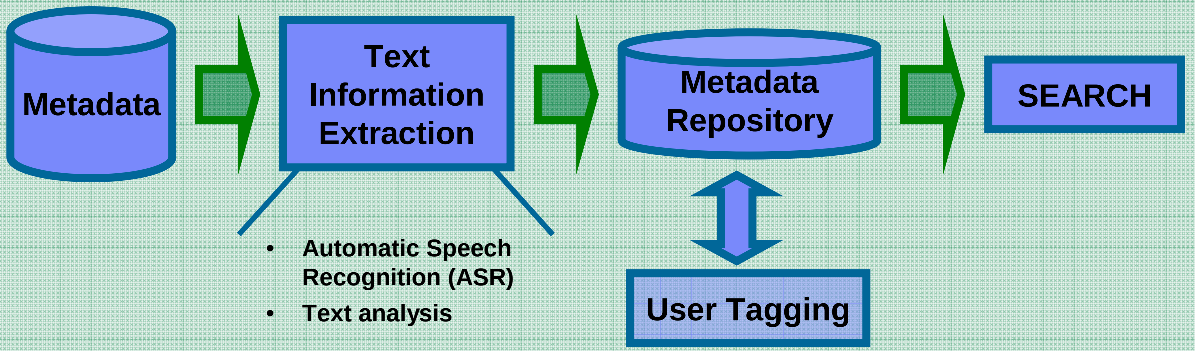
LIKE@25761 Page 1 of 1557 Go to page 1 Results 1 to 50 of 77817

25761. [Similar] [Related] [Add] [Tag] Outdoor | Urban | TRECVID2005_...	[0.866] [Similar] [Related] [Add] [Tag] Military | Outdoor | Urban | TRECVID2005_...	[0.85] [Similar] [Related] [Add] [Tag] Building | Outdoor | Sky | TRECVID2005_...	[0.845] [Similar] [Related] [Add] [Tag] Building | Military | Outdoor | TRECVID2005_...	[0.843] [Similar] [Related] [Add] [Tag] Crowd | TRECVID2005_...
[0.84] [Similar] [Related] [Add] [Tag] Outdoor | TRECVID2005_...	[0.839] [Similar] [Related] [Add] [Tag] Military | Outdoor | TRECVID2005_...	[0.831] [Similar] [Related] [Add] [Tag] Military | Outdoor | Sky | TRECVID2005_...	[0.825] [Similar] [Related] [Add] [Tag] Military | Outdoor | Sky | TRECVID2005_...	[0.822] [Similar] [Related] [Add] [Tag] Crowd | TRECVID2005_...
[0.821] [Similar] [Related] [Add] [Tag] Building | Military | Outdoor | Urban | TRECVID2005_...	[0.819] [Similar] [Related] [Add] [Tag] Outdoor | TRECVID2005_...	[0.806] [Similar] [Related] [Add] [Tag] Sky | TRECVID2005_...	[0.804] [Similar] [Related] [Add] [Tag] TRECVID2005_...	[0.803] [Similar] [Related] [Add] [Tag] Military | Outdoor | TRECVID2005_...
[0.8] [Similar] [Related] [Add] [Tag] Building | Outdoor | Sky | TRECVID2005_...	[0.796] [Similar] [Related] [Add] [Tag] Military | Outdoor | Sky | Urban | TRECVID2005_...	[0.796] [Similar] [Related] [Add] [Tag] Building | Outdoor | Sky | TRECVID2005_...	[0.794] [Similar] [Related] [Add] [Tag] Military | Outdoor | TRECVID2005_...	[0.794] [Similar] [Related] [Add] [Tag] Military | Outdoor | Sky | TRECVID2005_...

Done 3.310s

Basic Text-based Video Search

- **Program guides (EPG)**
- **Professional metadata**
- **Web text**
- **Audio transcripts**


Metadata

Text Information Extraction

- Automatic Speech Recognition (ASR)
- Text analysis

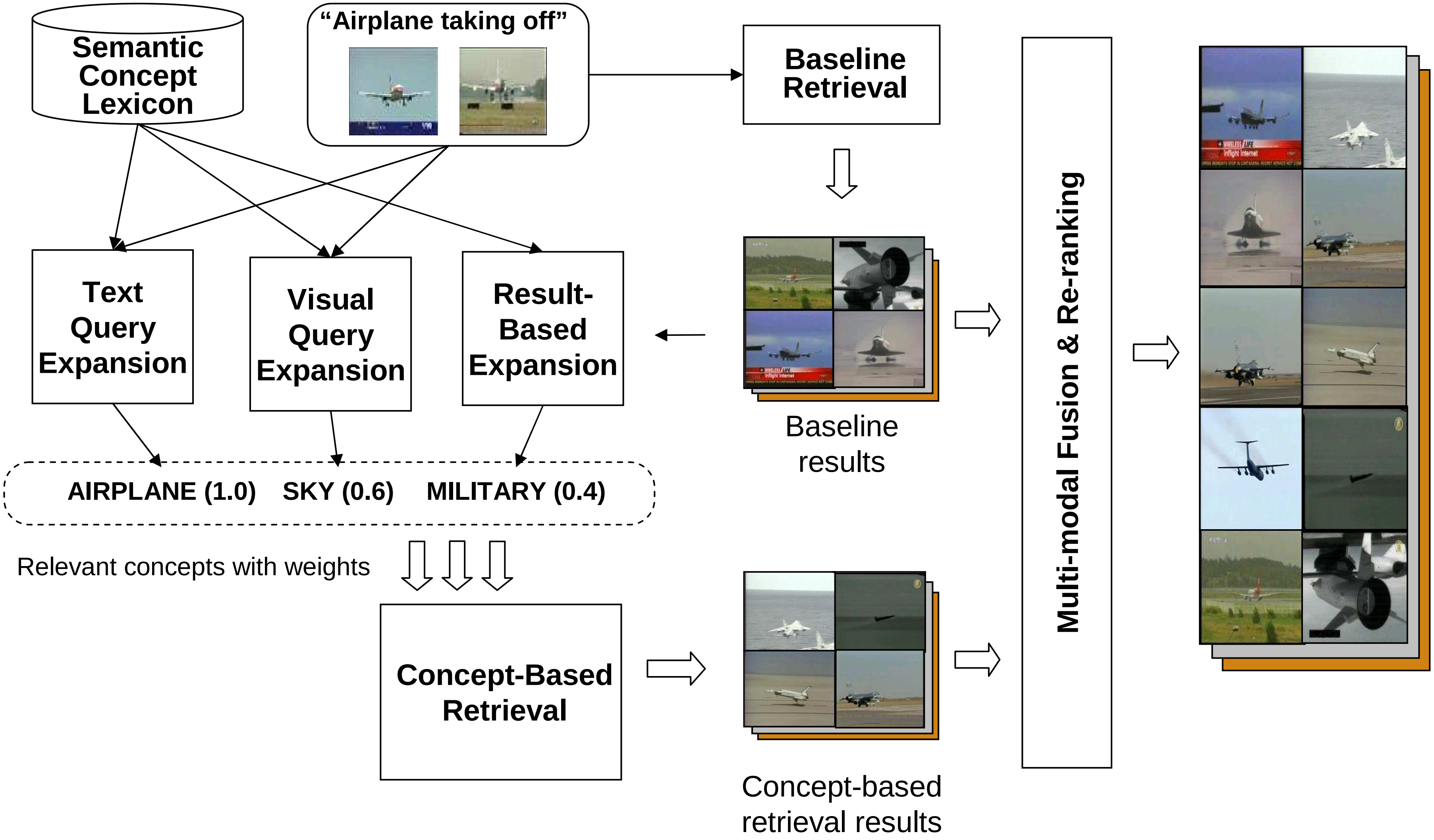
Metadata Repository

User Tagging

SEARCH

(4) Query Expansion for Multi-modal Video Retrieval

Query Expansion for Multi-modal Video Retrieval (*ACM Multimedia, Sept. 2007*)



Semantic Concept Lexicon

"Airplane taking off"

Baseline Retrieval

Text Query Expansion

Visual Query Expansion

Result-Based Expansion

Baseline results

AIRPLANE (1.0) SKY (0.6) MILITARY (0.4)

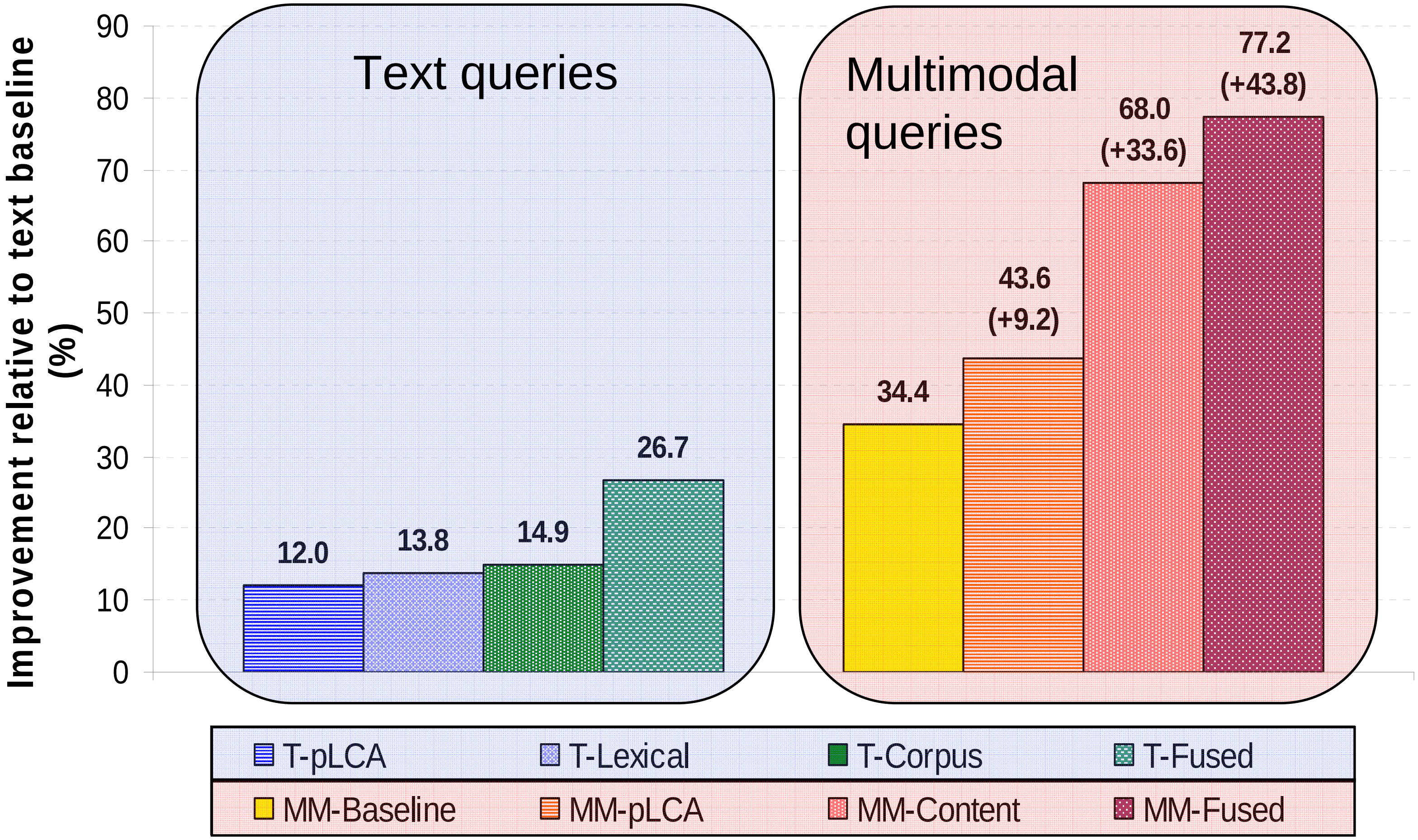
Relevant concepts with weights

Concept-Based Retrieval

Concept-based retrieval results

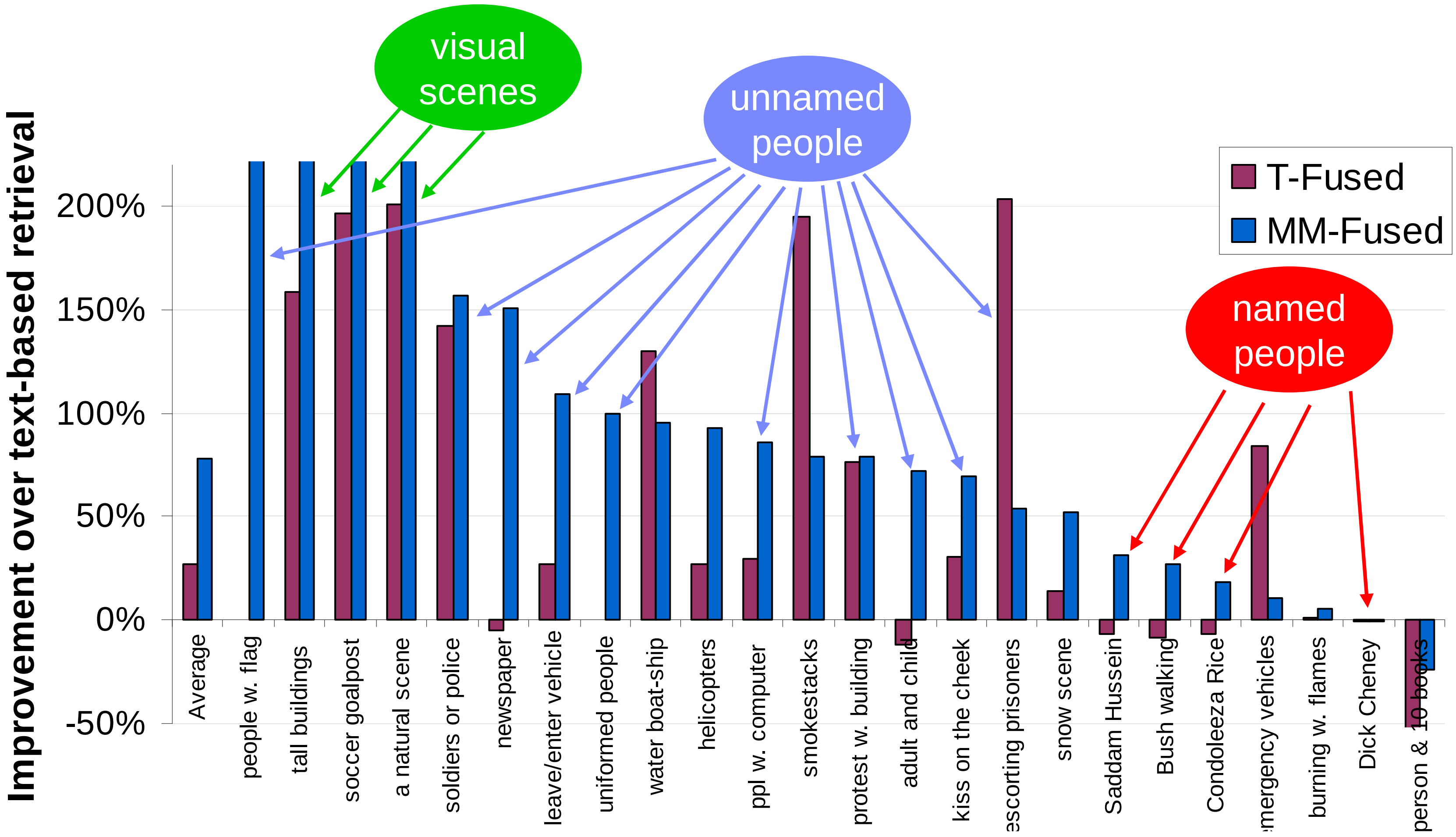
Multi-modal Fusion & Re-ranking

Empirical Evaluation & Comparison (TRECVID 2006 data)



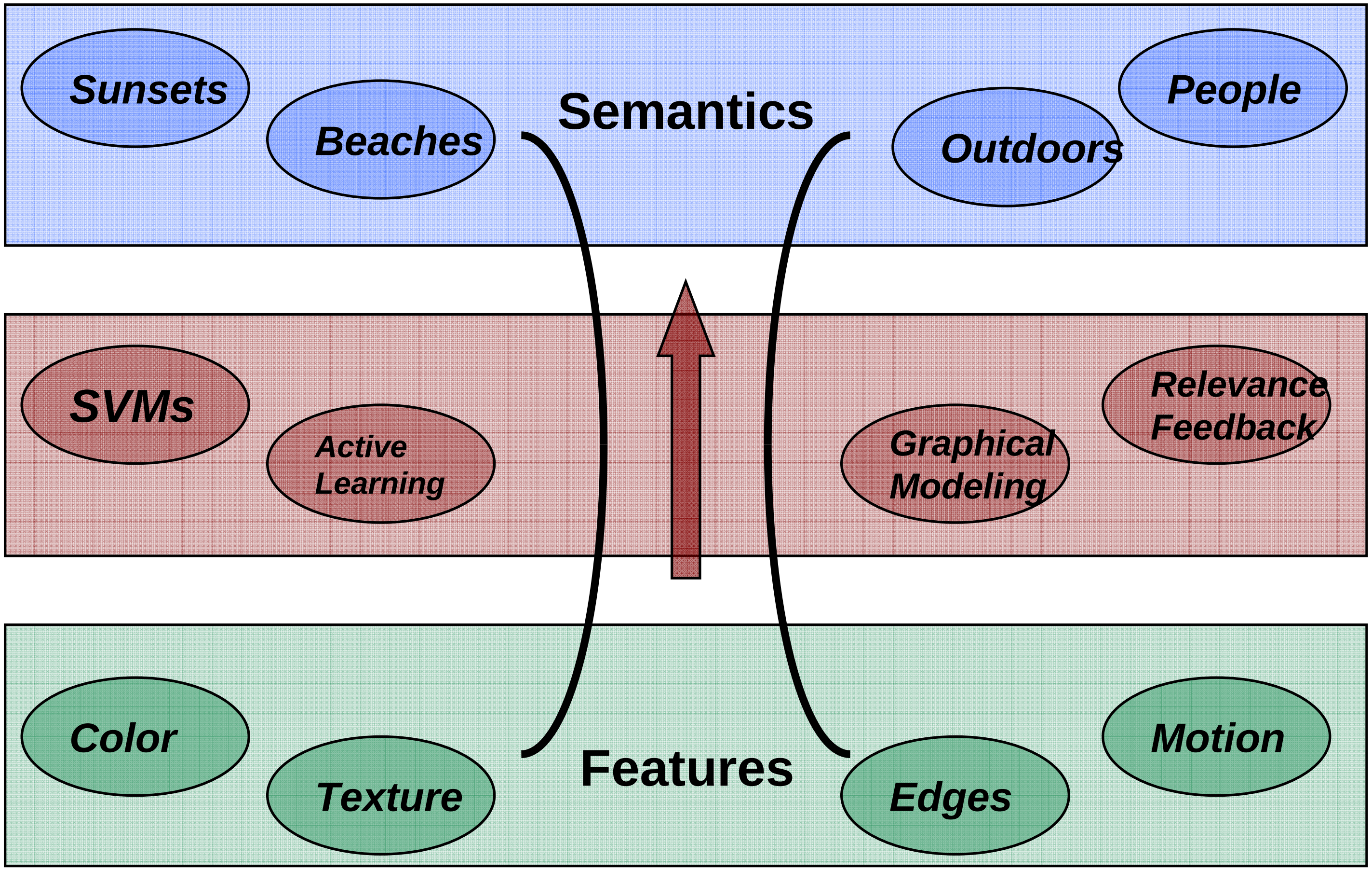
- Text-based expansion approaches perform comparably but are complementary
- Content-based approaches bring significant further improvements

Empirical Evaluation & Comparison (Cont’d)



Related efforts on the modeling of large video semantic spaces

Bridging the Multimedia Semantic Gap – What's the Destination?

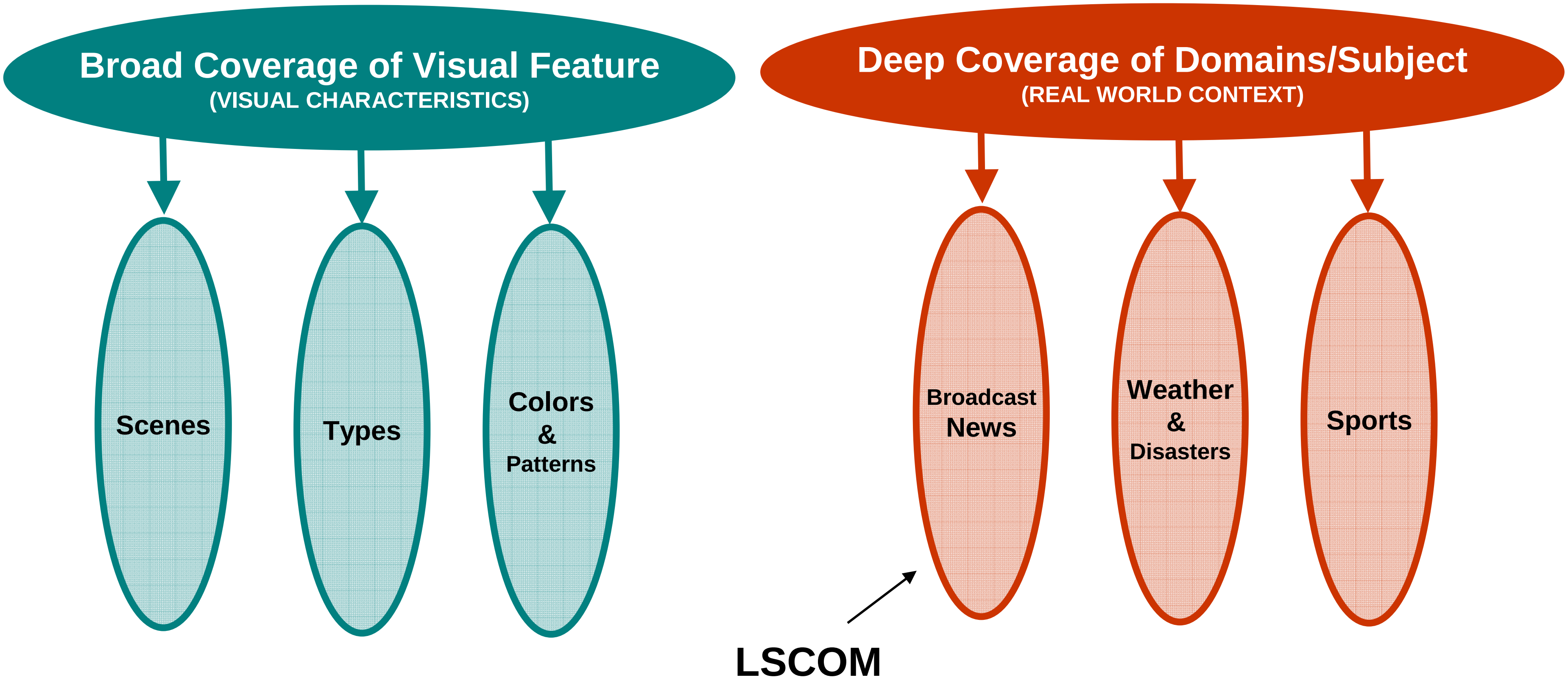


- How do we fully develop the semantic space itself?
- Research is producing powerful learning tools
- Foundation established (e.g., MPEG-7)

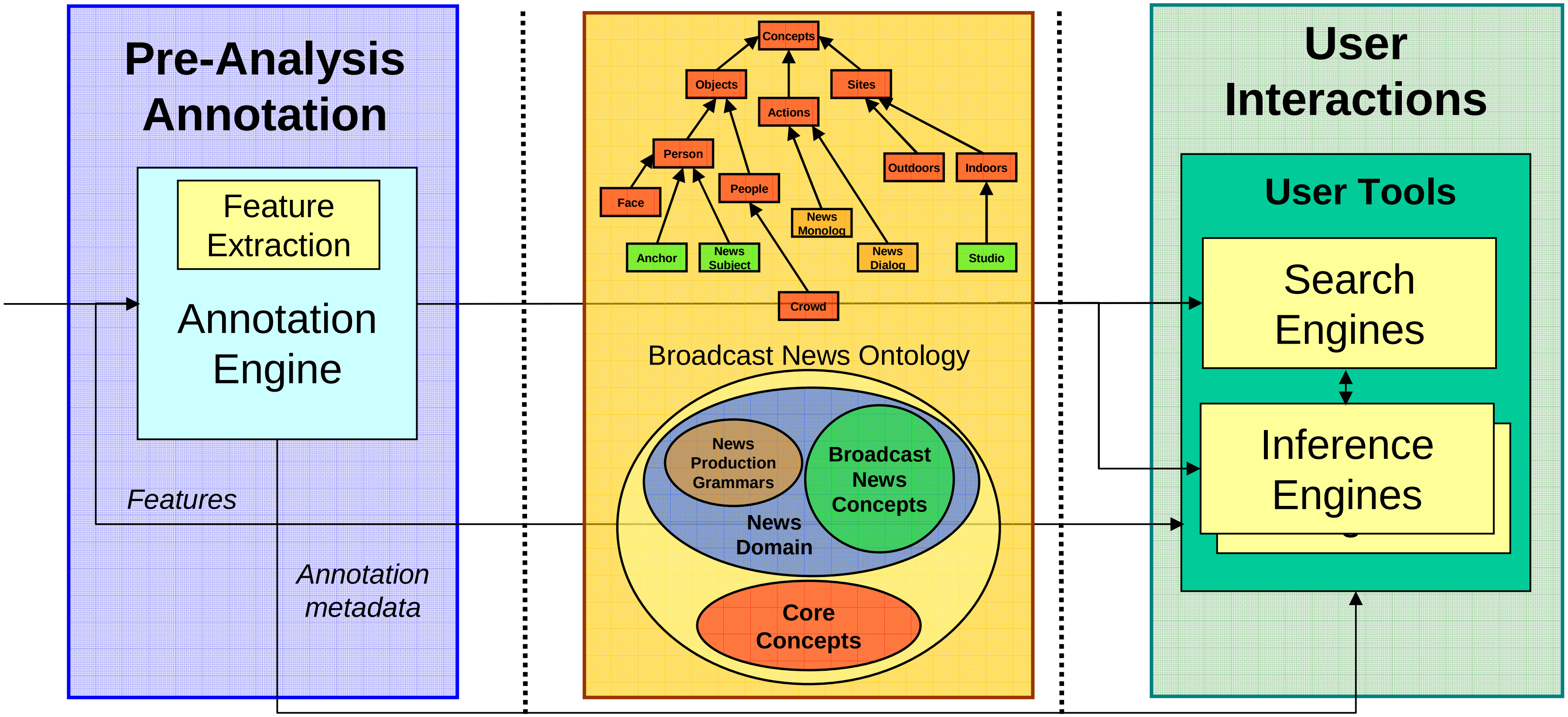
- Working on the foundation and the bridge, but what is the ultimate destination?
- Don't want to build a bridge to nowhere !!!

Structuring Multimedia Semantic Spaces

- Multimedia ontologies resemble faceted taxonomies but use richer semantic relationships among nodes that contain multimedia signifiers
- Can be developed to support different perspectives on multimedia content (i.e. visual characteristics vs. subject hierarchy



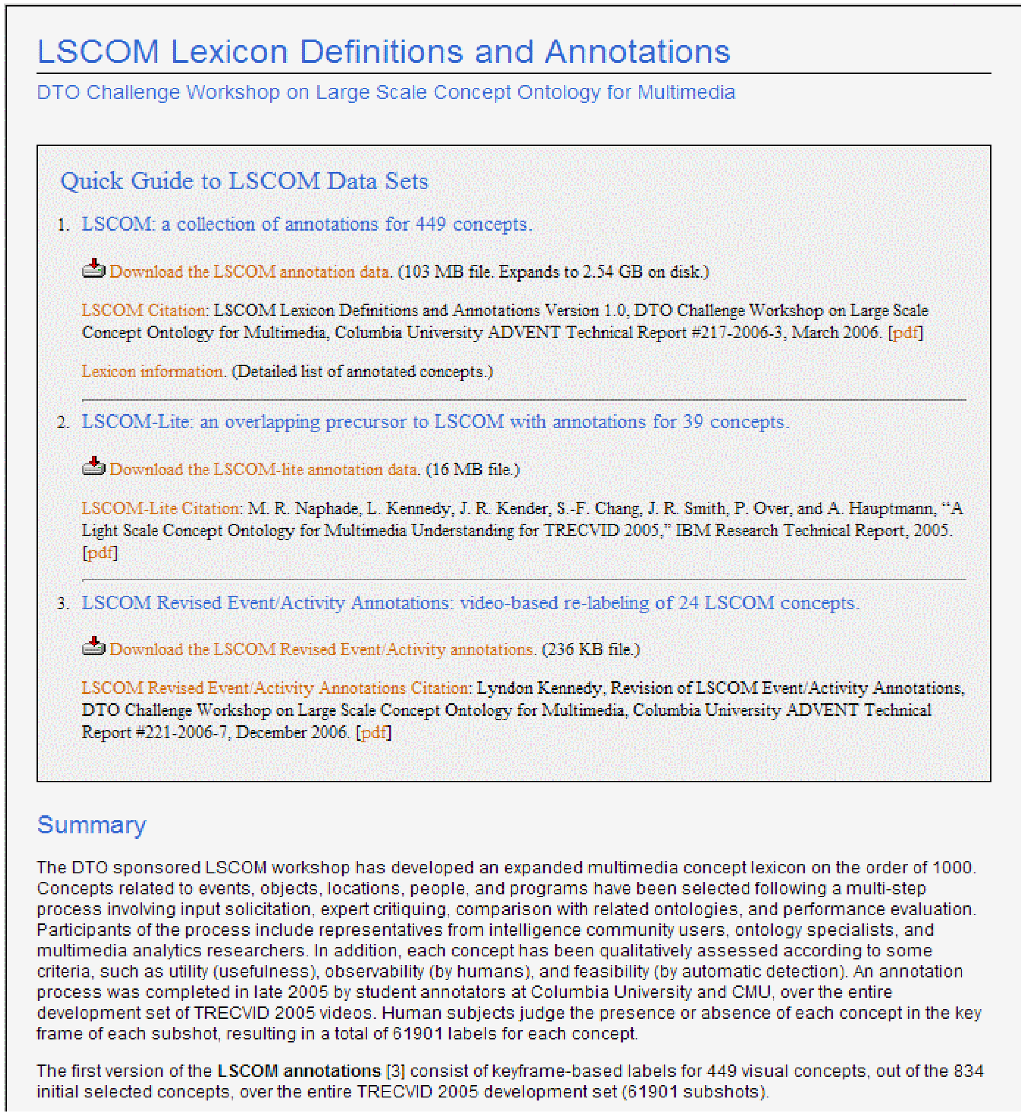
Large Scale Concept Ontology for Multimedia Understanding (LSCOM*) – 1,000 Semantic Concepts



- **LSCOM is collaborative effort to develop a large standardized taxonomy for describing multimedia broadcast news video**
- **<u>Designed to optimize</u>: (1) utility for facilitating end-user access, (2) coverage of large semantic space, (3) feasibility for automated extraction, (4) observability in diverse multimedia broadcast news data sets**

*** M. Naphade, J.R. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, J. Curtis, *IEEE MultiMedia*, July 2006**

Large Scale Concept Ontology for Multimedia (LSCOM)



LSCOM Lexicon Definitions and Annotations

DTO Challenge Workshop on Large Scale Concept Ontology for Multimedia

Quick Guide to LSCOM Data Sets

1. LSCOM: a collection of annotations for 449 concepts.

   Download the LSCOM annotation data. (103 MB file. Expands to 2.54 GB on disk.)

   LSCOM Citation: LSCOM Lexicon Definitions and Annotations Version 1.0, DTO Challenge Workshop on Large Scale Concept Ontology for Multimedia, Columbia University ADVENT Technical Report #217-2006-3, March 2006. [pdf]

   Lexicon information. (Detailed list of annotated concepts.)

2. LSCOM-Lite: an overlapping precursor to LSCOM with annotations for 39 concepts.

   Download the LSCOM-lite annotation data. (16 MB file.)

   LSCOM-Lite Citation: M. R. Naphade, L. Kennedy, J. R. Kender, S.-F. Chang, J. R. Smith, P. Over, and A. Hauptmann, "A Light Scale Concept Ontology for Multimedia Understanding for TRECVID 2005," IBM Research Technical Report, 2005. [pdf]

3. LSCOM Revised Event/Activity Annotations: video-based re-labeling of 24 LSCOM concepts.

   Download the LSCOM Revised Event/Activity annotations. (236 KB file.)

   LSCOM Revised Event/Activity Annotations Citation: Lyndon Kennedy, Revision of LSCOM Event/Activity Annotations, DTO Challenge Workshop on Large Scale Concept Ontology for Multimedia, Columbia University ADVENT Technical Report #221-2006-7, December 2006. [pdf]

Summary

The DTO sponsored LSCOM workshop has developed an expanded multimedia concept lexicon on the order of 1000. Concepts related to events, objects, locations, people, and programs have been selected following a multi-step process involving input solicitation, expert critiquing, comparison with related ontologies, and performance evaluation. Participants of the process include representatives from intelligence community users, ontology specialists, and multimedia analytics researchers. In addition, each concept has been qualitatively assessed according to some criteria, such as utility (usefulness), observability (by humans), and feasibility (by automatic detection). An annotation process was completed in late 2005 by student annotators at Columbia University and CMU, over the entire development set of TRECVID 2005 videos. Human subjects judge the presence or absence of each concept in the key frame of each subshot, resulting in a total of 61901 labels for each concept.

The first version of the **LSCOM annotations** [3] consist of keyframe-based labels for 449 visual concepts, out of the 834 initial selected concepts, over the entire TRECVID 2005 development set (61901 subshots).

www.ee.columbia.edu/dvmm/lscom

- **What is it?** – lexicon covering large semantic space for broadcast news analysis from IC perspective
  - >1,000 concepts
  - Large annotated video data set (449 visual concepts, 24 temporal activities)
- **Impact to-date:**
  - LSCOM-lite used in TRECVID
  - Downloaded by >170 groups
- **Available for download:**
  - LSCOM lexicon
  - LSCOM annotations
  - "Columbia374" SVM models

Sample of 170+ institutions downloading LSCOM

- 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 ...

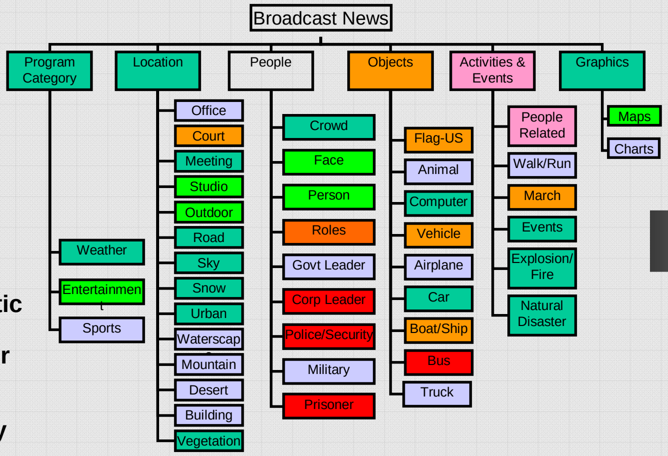
Link to download log

Large Scale Concept Ontology for Multimedia Understanding (LSCOM*) – 1,000 Semantic Concepts

Design

Taxonomy Design:

- *LSCOM 1,000 Semantics Concepts
- Designed to optimize:
  1. Utility for facilitating user access
  2. Coverage of large semantic space
  3. Feasibility for automated extraction
  4. Observability in diverse broadcast news data sets

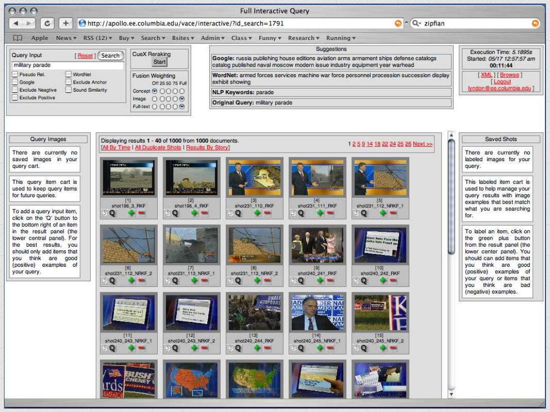


* *IEEE MultiMedia,* Summer 2006

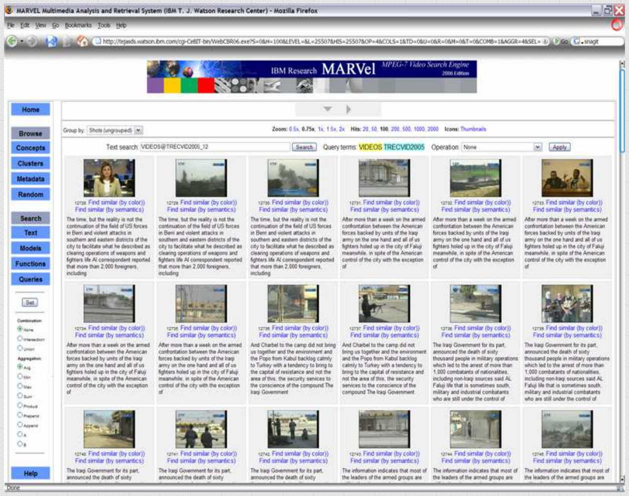
Annotate

Annotated Concepts:

- Event/Activity (56 - 13%) - *Airplane taking off, car crash, shaking hands*
- People (113 - 25%) - *Female person, firefighter, judge*
- Location (89 - 20%) - *Hospital, airfield, cityscape*
- Object (135 - 30%) - *Power plant, tent, vehicle*
- Scene (49 - 10%) - *Vegetation, interview, urban*
- Program (7 - 2%) - *Entertainment, weather, finance*



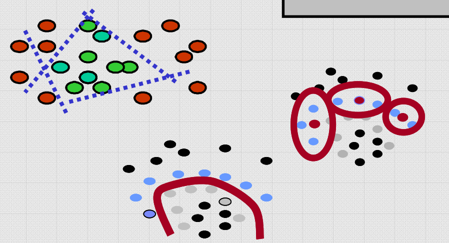
Use

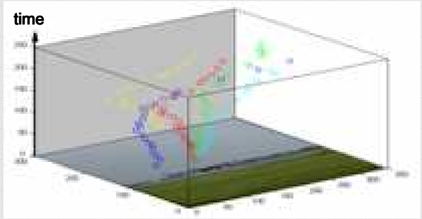

MARVEL Multimedia Analysis and Retrieval System (IBM T. J. Watson Research Center) - Mozilla Firefox
IBM Research MARVel
MPEG-7 Video Search Engine
Home
Browse
Concepts
Clusters
Metadata
Random
Search
Text
Models
Functions
Queries
Help

Search & Retrieval:

- Use-case driven assessment for searching & topic threading
- Multiple search engine implementation and evaluation
- Support for automatic, manual and interactive search

Model





Semantics Modeling:

- Appearance-based feature extraction (color, texture, shape, edges, motion)
- Machine learning and statistical modeling (SVMs, GMMs, Nearest Neighbor)
- Multi-feature and multi-model fusion
- Scalable modeling using a massive distributed computing infrastructure

Public evaluations such as TRECVID

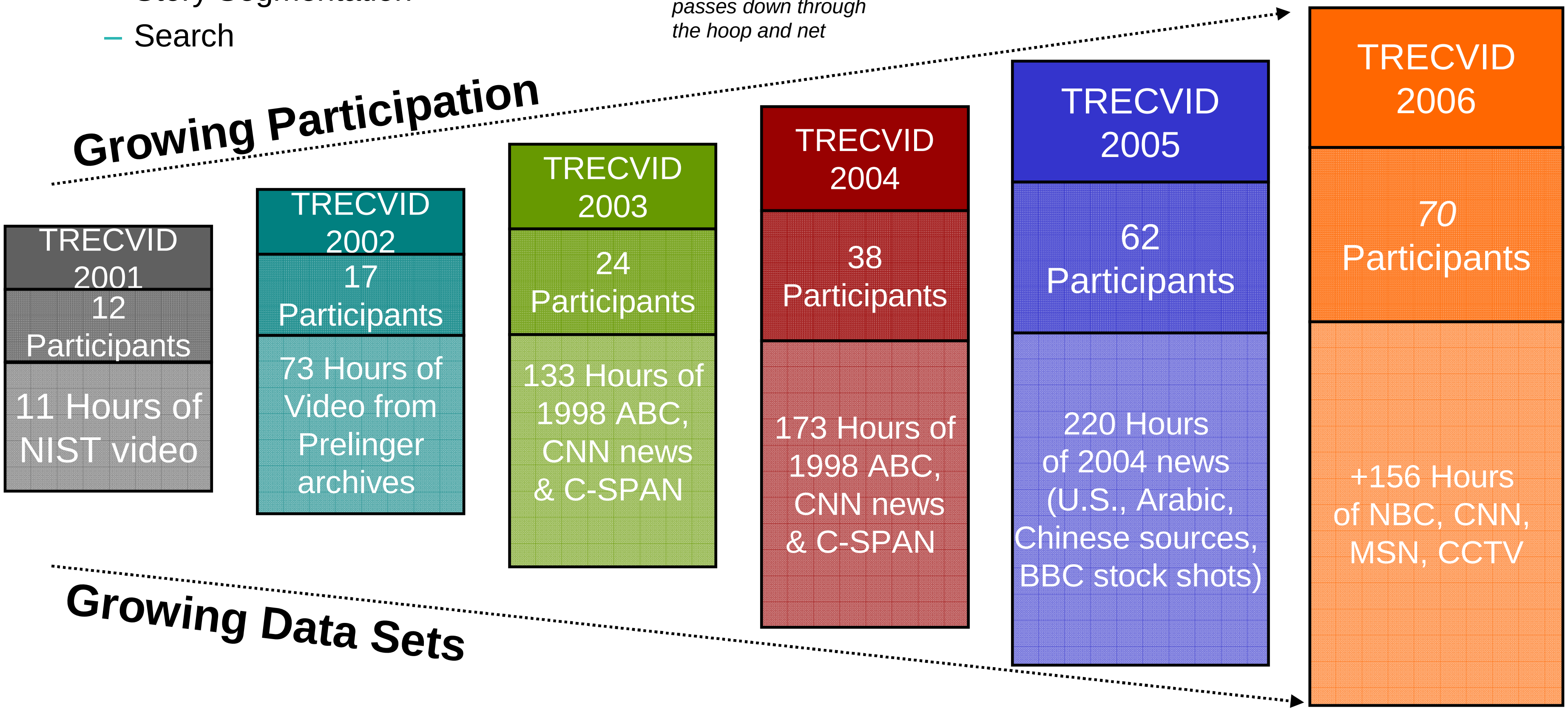
NIST TRECVID Video Retrieval Benchmark at a Glance

- **TRECVID:**
  - NIST benchmark for evaluating state of the art in video retrieval
- **Benchmark tasks:**
  - Shot Boundary Determination
  - Semantic Concept Detection
  - Story Segmentation
  - Search

Topic 101: *Find shots of a basket being made - the basketball passes down through the hoop and net*

Topic 129: *Find shots zooming in on the US Capitol dome.*

Topic 104 and 167: *Find shots of an airplane taking off*



Growing Participation

TRECVID 2001
12 Participants
11 Hours of NIST video

TRECVID 2002
17 Participants
73 Hours of Video from Prelinger archives

TRECVID 2003
24 Participants
133 Hours of 1998 ABC, CNN news & C-SPAN

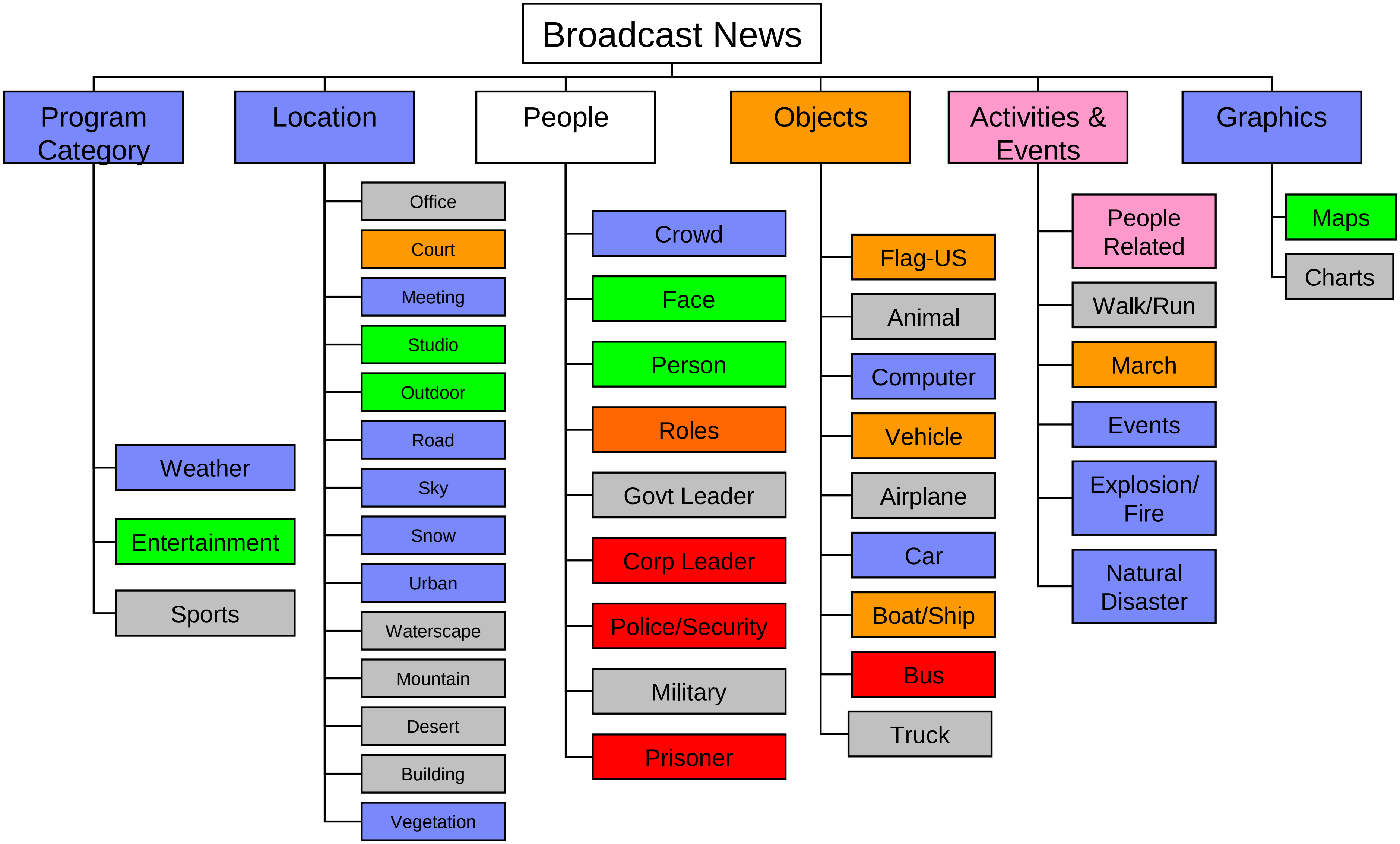
TRECVID 2004
38 Participants
173 Hours of 1998 ABC, CNN news & C-SPAN

TRECVID 2005
62 Participants
220 Hours of 2004 news (U.S., Arabic, Chinese sources, BBC stock shots)

TRECVID 2006
70 Participants
+156 Hours of NBC, CNN, MSN, CCTV

Growing Data Sets

“LSCOM-lite” for TRECVID High-Level Feature Detection



- Broadcast News
  - Program Category
    - Weather
    - Entertainment
    - Sports
  - Location
    - Office
    - Court
    - Meeting
    - Studio
    - Outdoor
    - Road
    - Sky
    - Snow
    - Urban
    - Waterscape
    - Mountain
    - Desert
    - Building
    - Vegetation
  - People
    - Crowd
    - Face
    - Person
    - Roles
    - Govt Leader
    - Corp Leader
    - Police/Security
    - Military
    - Prisoner
  - Objects
    - Flag-US
    - Animal
    - Computer
    - Vehicle
    - Airplane
    - Car
    - Boat/Ship
    - Bus
    - Truck
  - Activities & Events
    - People Related
    - Walk/Run
    - March
    - Events
    - Explosion/ Fire
    - Natural Disaster
  - Graphics
    - Maps
    - Charts

TRECVID

Video corpus

- Broadcast news from U.S., Arabic, and Chinese sources
  - TRECVID 2005: 160 hrs
  - TRECVID 2006: 240 hrs
- Speech transcripts based on
  - Speech Recognition
  - Machine Translation

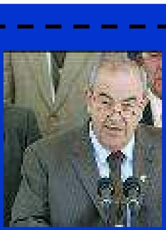
Query topics

- Brief description of topic
- 5-10 visual examples/topic
- 24-25 topics each year
- Typical topic classes:
  - Named people (Person-X)
  - Generic people interactions
  - Sports
  - Objects/Events
  - Scenes/settings

Search types

- Automatic, manual, interactive

***Topic 149**: Find shots of Condoleeza Rice*



***Topic 150**: Find shots of Iyad Allawi, the former prime minister of Iraq*

***Topic 151**: Find shots of Omar Karami, the former prime minister of Lebannon*

***Topic 152**: Find shots of Hu Jintao, president of the People's Republic of China*

***Topic 153**: Find shots of Tony Blair.*

***Topic 159**: Find shots of George W. Bush entering or leaving a vehicle (e.g., car, van, airplane, helicopter, etc) (he and vehicle both visible at same time)*

***Topic 157**: Find shots of people shaking hands*

***Topic 161**: Find shots of people with banners or signs*

***Topic 163**: Find shots of a meeting with a large table and more than two people*

***Topic 165**: Find shots of basketball players on the court*

***Topic 171**: Find shots of a goal being made in a soccer match*

***Topic 156**: Find shots of tennis players on the court both players visible at same time*

***Topic 158**: Find shots of a helicopter in flight*

***Topic 164**: Find shots of a ship or boat*

***Topic 167**: Find shots of an airplane taking off*

***Topic 160**: Find shots of something (e.g., vehicle, aircraft, building) on fire with flames & smoke visible*

***Topic 168**: Find shots of a road with one or more cars*

***Topic 170**: Find shots of a tall building (with more than 5 floors above the ground)*

ACM CIVR'07 VideOlympics Showcase (July, 2007)

- A video search showcase that goes beyond the regular demo session and a small size of TRECVID participants
  - The showcase participants will simultaneously do an **interactive search task** during the VideOlympics showcase event.
  - Paul Over from NIST will provide **text-only** search topics onsite
  - Unlike TRECVID, results are submitted immediately after they are found.
  - Fun to do for the participants and fun to watch for the conference audience
- The first VideOlympics event is a great success
  - 9 retrieval systems submitted from worldwide participants and great interest from the audience in the conference
- Video: http://videolympics.org/
- Next year: CIVR'08, Niagara Falls, Canada

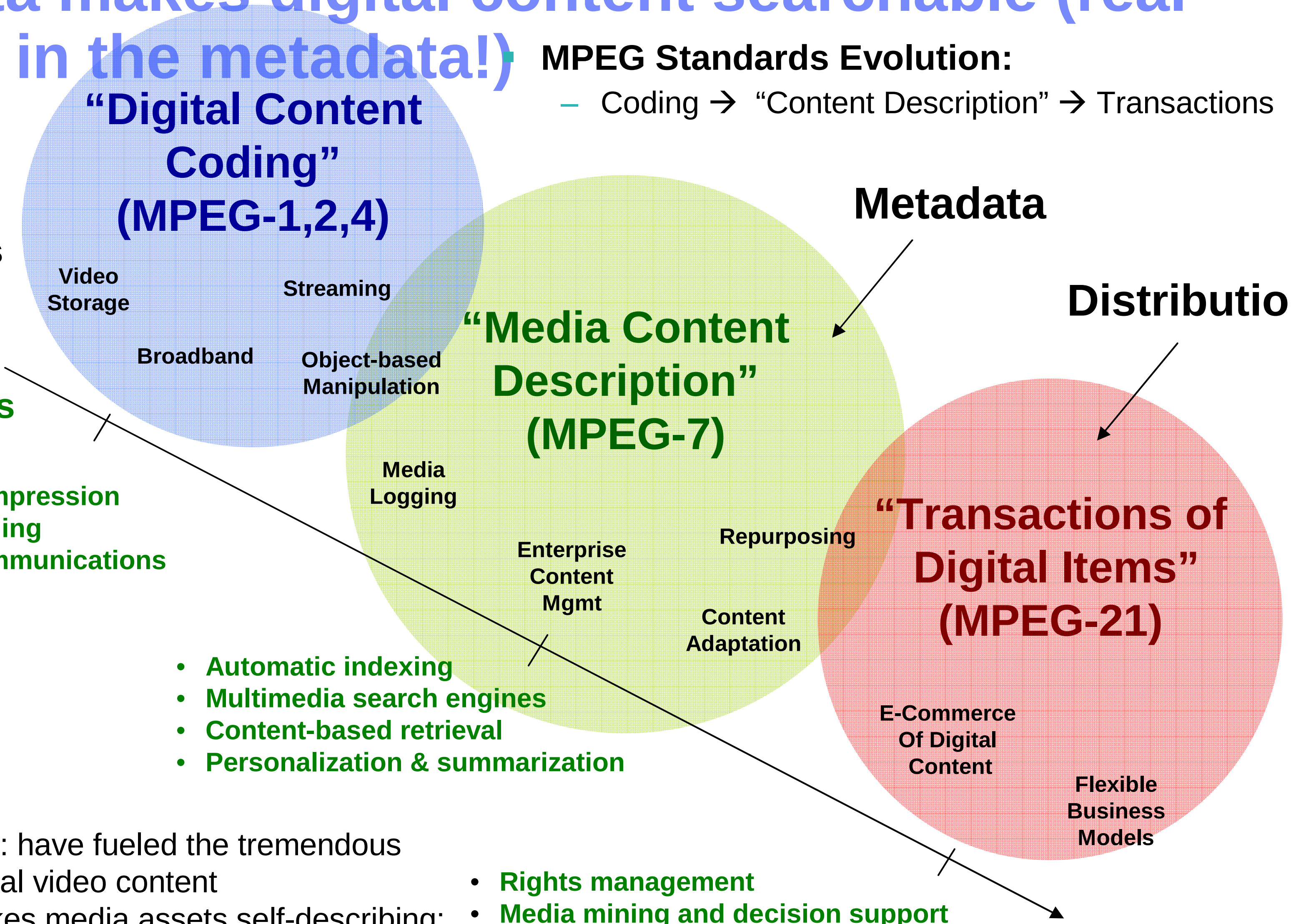
Role of MPEG-7 as a way to store metadata generated for video in a fully standards-based searchable representation.

Metadata makes digital content searchable (real value is in the metadata!)

- **MPEG Standards Evolution:**
  - Coding → "Content Description" → Transactions

- **Applications**

- **Technologies**


"Digital Content Coding" (MPEG-1,2,4)
Video Storage
Streaming
Broadband
Object-based Manipulation
"Media Content Description" (MPEG-7)
Media Logging
Enterprise Content Mgmt
Repurposing
Content Adaptation
Metadata
Distribution
"Transactions of Digital Items" (MPEG-21)
E-Commerce Of Digital Content
Flexible Business Models

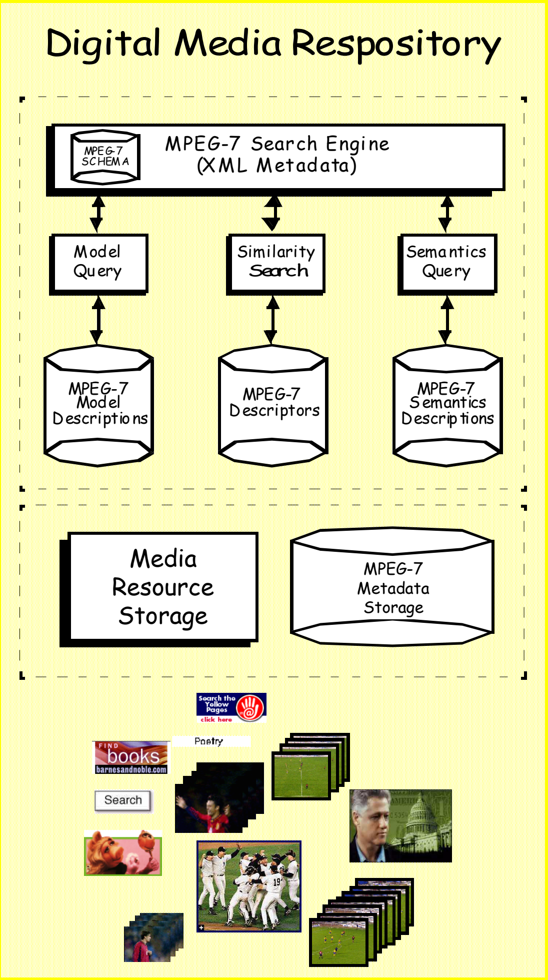
- Compression
- Coding
- Communications

- Automatic indexing
- Multimedia search engines
- Content-based retrieval
- Personalization & summarization

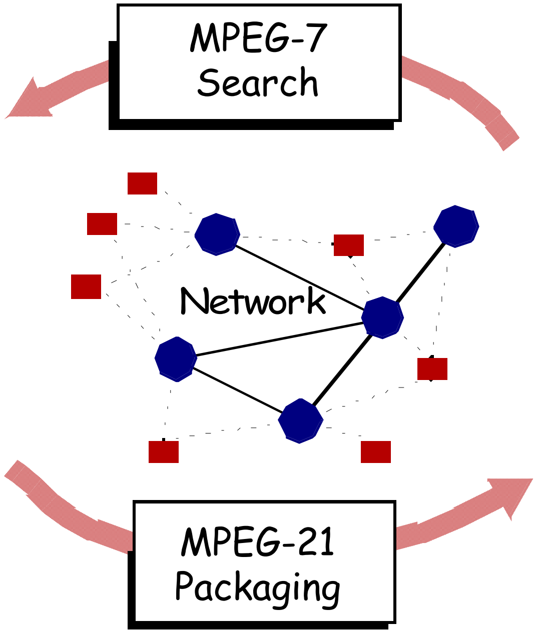
- Rights management
- Media mining and decision support

- **MPEG-1,-2,-4**: have fueled the tremendous growth in digital video content
- **MPEG-7**: makes media assets self-describing; allows content-based access

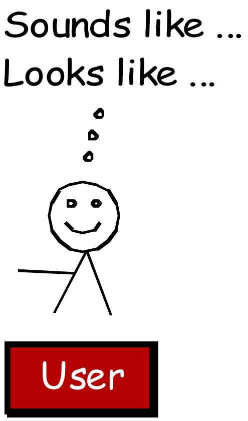
MPEG-7/-21 Multimedia Indexing, Searching and Delivery

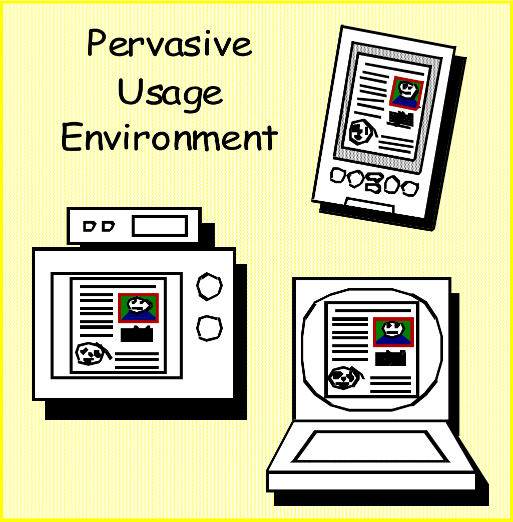


- **Multimedia Indexing & Searching:**
  - Semantics-based (people, places, events, objects, scenes, speech)
  - Immutable metadata (titles, dates)
  - Content-based (color, texture, motion, melody, timbre)



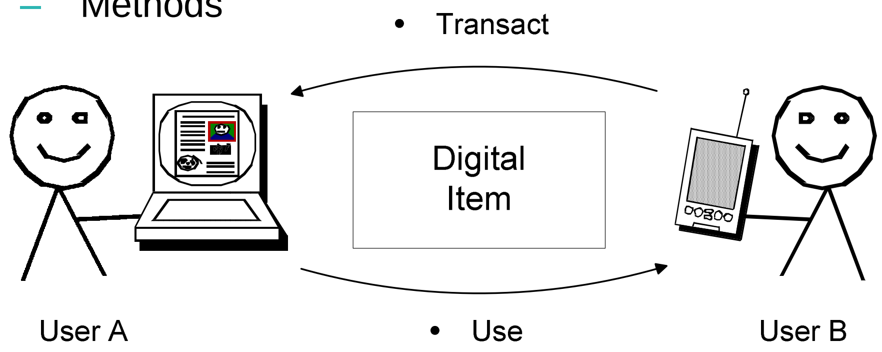
- **Multimedia Access & Delivery:**
  - Media content personalization
  - Adaptation & summarization
  - Usage environment (context, devices, user preferences)





MPEG-21 Multimedia Framework: “Transactions of Digital Items”

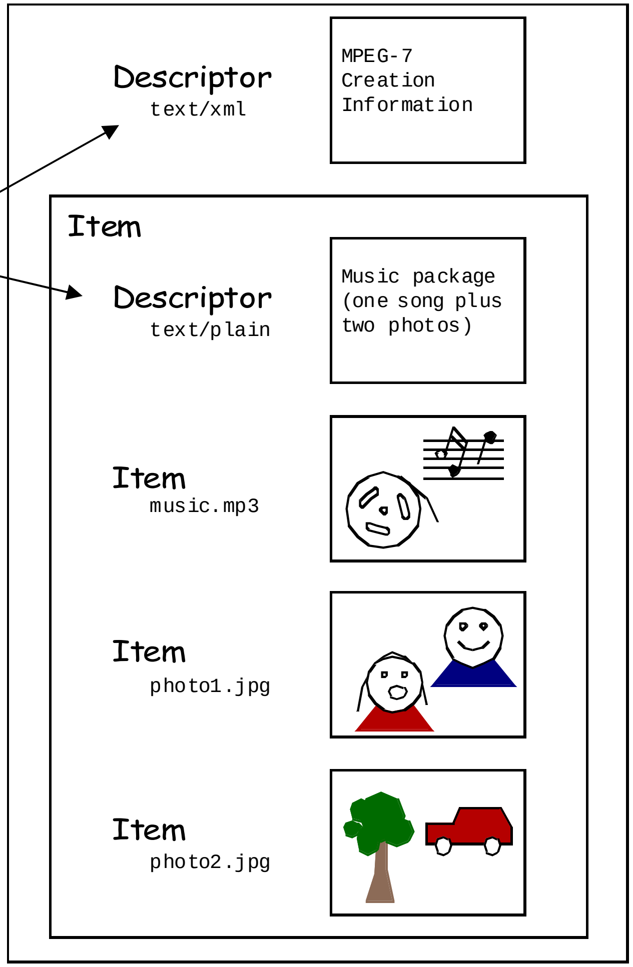
- Users and participants in the content value network seamlessly exchange content in form of “digital items” across networks and devices
- Framework supporting all forms of electronic content/intellectual property (video, music, learning objects, on-line reports, etc.)
- Digital Item = bundling of:
  - Media resource
  - Metadata (eg., MPEG-7)
  - Rights expressions
  - Identifiers
  - Methods



- Example: Digital music package

MPEG-7 Metadata

MPEG-21 Digital Item



These approaches together go a long way to truly unleash video search.

References

- **Demos and Tools:**
  - IBM Research Marvel “lite”
    - http://www.alphaworks.ibm.com/tech/imars
- **Links:**
  - IBM Research Intelligent Information Management Department:
    - http://www.research.ibm.com/iim