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
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*
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

**Clipblast** — “Weapons”
- **Depth:**
  - Search relies on text descriptions
- **Breadth:**
  - Limited to partner content

**Blinkx** — “Weapons”
- **Depth:**
  - 53,000 matches related to “weapons”
  - No way to obtain clips showing weapons scenes
- **Breadth:**
  - Results limited to partner content

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

<table>
<thead>
<tr>
<th>Issue</th>
<th>What’s wrong</th>
</tr>
</thead>
<tbody>
<tr>
<td>Too sparse</td>
<td>Few video objects have any metadata</td>
</tr>
<tr>
<td>Inadequate</td>
<td>Mainly tags or few keywords, program-guide info for broadcast video, speech available in few cases</td>
</tr>
<tr>
<td>Coarse-grain</td>
<td>At level of digital objects only</td>
</tr>
<tr>
<td>Not visual</td>
<td>Does not describe what is visually depicted</td>
</tr>
<tr>
<td>Ambiguous</td>
<td>Taxonomies not widely used; folksonomies creating new problems</td>
</tr>
<tr>
<td>Inconsistent</td>
<td>Vocabularies and taxonomies not standardized</td>
</tr>
<tr>
<td>Subjective</td>
<td>Limited verification across users</td>
</tr>
<tr>
<td>Not trustworthy</td>
<td>Professional metadata mixed-in with noise</td>
</tr>
</tbody>
</table>
Automatic Machine Tagging can Complement and Enhance Professional Cataloging and Social Tagging Approaches

**Manual Cataloging – By Professionals**

<table>
<thead>
<tr>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Controlled vocabularies &amp; standard taxonomies</td>
<td>Costly</td>
</tr>
<tr>
<td>Higher quality</td>
<td>Human resource intensive</td>
</tr>
<tr>
<td>Cannot keep up</td>
<td></td>
</tr>
<tr>
<td>Example: Fox, CNN, BBC, Broadcast TV</td>
<td></td>
</tr>
</tbody>
</table>

**Automated Tagging – By Machine**

<table>
<thead>
<tr>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower human cost</td>
<td>Requires training of models</td>
</tr>
<tr>
<td>Domain &amp; data driven approach to semantics</td>
<td>Lower quality than manual tagging</td>
</tr>
<tr>
<td>Example: Marvel, Informedia, TRECVID concept detection</td>
<td></td>
</tr>
</tbody>
</table>

**Social Tagging – By Users**

<table>
<thead>
<tr>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>User driven</td>
<td>Ambiguity</td>
</tr>
<tr>
<td>Emergent folksonomies</td>
<td>Uncontrolled vocabulary</td>
</tr>
<tr>
<td>Serendipitous browsing</td>
<td>Synonyms</td>
</tr>
<tr>
<td>Example: Del.icio.us and Flickr</td>
<td></td>
</tr>
</tbody>
</table>

Popularity
- High-value content, hit-TV shows, movies
- Deep archives, raw footage, independents

“Long tail”
- Consumer content
- Digital item
The market is seeking a new approach for effective video search

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**
  - QBIC
  - VisualSEEk
  - Netra
  - BlobWorld
  - Visual feature extraction (color, texture, shape)
  - Relevance feedback
  - Spatial region search

- **Video Retrieval**
  - MARS
  - WebSEEk
  - VideoQ
  - Shot boundary detection
  - Speech indexing
  - Motion extraction
  - Video summarization

- **Multimedia Semantics**
  - MPEG-7
  - Virage
  - CueVideo
  - Informedia
  - Semantics learning
  - Scene classification
  - Semantic clustering
  - Search fusion

- **Timeline**
  - '95: QBIC, VisualSEEk, Netra
  - '00: MARS, WebSEEk, VideoQ
  - '07: MPEG-7, Virage, CueVideo, Informedia
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)

### Analysis of Video Modalities

<table>
<thead>
<tr>
<th>Sources</th>
<th>Functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Broadcast Video</td>
<td>Video Search</td>
</tr>
<tr>
<td>User Generated Content</td>
<td>Content Classification</td>
</tr>
<tr>
<td>Video Blogs</td>
<td>Filtering</td>
</tr>
<tr>
<td>Social Web Sites</td>
<td>Discovery</td>
</tr>
<tr>
<td>Podcasts</td>
<td>Copy Detection</td>
</tr>
<tr>
<td>Web Content</td>
<td>Summarization</td>
</tr>
</tbody>
</table>

**Visual**
- **Visual scenes**: Detect visual semantic categories
  - 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
Marvel – Software for learning visual categories and classifying and recognizing image/video content

**Image/Video Content Learning SW**

- Feature Extraction → Visual Features → Positive Examples
- Feature Extraction → Visual Features → Negative Examples
- Statistical Machine Learning
- Classifier Ensemble

**Image/Video Classification Software**

- Classifier Ensemble
- Fusion Classifier
- Classifiers
- Taxonomy
- Classified images/video

- Unknown images/video
- Classifiers
- Color space, Domain, Objectionable, Objects, People, Setting, Subject, Visual Type
- Classified images/video

© 2008 IBM Corporation
Semantic models are created from training examples that are managed using multimedia taxonomies.
Semantic Tagging of Multimedia Content

Video:

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

Text Analysis (optional)

“Jet fighters”

“military”

“Middle East”

Visual Features

Motion

Texture

Color patterns

Shapes

Automatic Semantic Concept Detection

Airplane (0.8)

Outdoors (0.9)

People (0.7)

Protest (0.6)

Parade (0.5)

Sky (0.7)

Indoors (0.8)

Meeting (0.7)

Semantics Metadata

People (0.7)

Sky (0.7)

Meeting (0.7)

* with associated confidence scores

Text Metadata

Speech

Closed Captions

Transcript

Presentation

Image & Video Sequence Analysis

Visual Feature Metadata

Semantic Models

IBM Research

Unleashing Video Search | John R. Smith | 2008
Multimedia Semantic Analysis and Search

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

Content Extraction:
- Semantic Labeling of Visual Clusters
- Temporal Pattern Mining
- Cross-Channel Topic Tracking

Searching, Browsing & Interaction:
- Multi-Modal Search and Retrieval
- Query Expansion for Multimodal Video Retrieval
- Query-Class Dependent Video Search

Model

Extract

User

Search
(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?

  4. **Improvement:**
     - Does solution provide path for obtaining and sustaining performance improvements?

  5. **Semantics:**
     - Does solution provide required set of classifiers?
     - How are they prioritized by business need?
Scalability in Visual Semantic Classification is Achieved by Trading-off Semantic Completeness, Classifier Costs and Data Volumes

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

Smart feature sampling attains 50-100x speed-up in 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
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.

**Classifier Trade-off (Speed vs. Accuracy)**

- **Full Classifier Ensemble** (>1000 unit models)
  - Fusion Classifier
  - Ordered by validation score
- **Pruned Ensemble** (>1 unit model)
  - 10-100x speed-up

**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)**
Significant speed-up in learning makes it possible to learn new visual semantic models in near real-time as needed.
(2) Semantic Labeling of Visual Clusters
Semantic Labeling of Visual Clusters – Discovering Descriptive & Discriminative Semantics (ICME-2006)

Multimedia Repository

Concept Detection
- People
- Face
- Outdoors
- ...

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.

**Advanced multi-modal search**

- **Visual Classifiers**
- **Semantic Taxonomies**

**Multi-modal Search**

- Guided navigation

**Sources**

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

**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**

**SEARCH**

- More accurate (>2.5x)
- Faster (>250x)
- Precise matches within clips
(4) Query Expansion for Multi-modal Video Retrieval
Query Expansion for Multi-modal Video Retrieval (ACM Multimedia, Sept. 2007)

- Text Query Expansion
- Visual Query Expansion
- Result-Based Expansion

- Relevant concepts with weights: AIRPLANE (1.0), SKY (0.6), MILITARY (0.4)
- Concept-Based Retrieval
- “Airplane taking off”
- Baseline Retrieval
- Multi-modal Fusion & Re-ranking

MILITARY (0.4)  SKY (0.6)  AIRPLANE (1.0)
Empirical Evaluation & Comparison (TRECVID 2006 data)

Text queries

- T-pLCA
- T-Lexical
- T-Corpus
- T-Fused

Multimodal queries

- MM-Baseline
- MM-pLCA
- MM-Content
- MM-Fused

- Text-based expansion approaches perform comparably but are complementary
- Content-based approaches bring significant further improvements
Empirical Evaluation & Comparison (Cont’d)

Improvement over text-based retrieval

-50% 0% 50% 100% 150% 200%

Average people w. flag
tall buildings soccer goalpost a natural scene soldiers or police newspaper leave/enter vehicle uniformed people water boat ship helicopters ppl w. computer smokestacks protest w. building adult and child kiss on the cheek escoing prisoners snow scene Saddam Hussein Bush walking Condoleeza Rice emergency vehicles burning w. flames Dick Cheney person & 10 books

visual scenes
unnamed people

T-Fused
MM-Fused

named people
Related efforts on the modeling of large video semantic spaces
Bridging the Multimedia Semantic Gap – What’s the Destination?

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

- How do we fully develop the semantic space itself?

- Research is producing powerful learning tools

- Foundation established (e.g., MPEG-7)
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).

![Diagram showing Broad Coverage of Visual Feature (VISUAL CHARACTERISTICS) and Deep Coverage of Domains/Subject (REAL WORLD CONTEXT)]
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
- **Designed to optimize**: (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

Large Scale Concept Ontology for Multimedia (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

www.ee.columbia.edu/dvmm/lscom
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

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

Annotate 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 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

* IEEE MultiMedia, Summer 2006
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

---

**Growing Participation**

<table>
<thead>
<tr>
<th>Year</th>
<th>TRECVID</th>
<th>Participants</th>
<th>Data Sets</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001</td>
<td>12</td>
<td>11 Hours of NIST video</td>
<td></td>
</tr>
<tr>
<td>2002</td>
<td>17</td>
<td>73 Hours of Video from Prelinger archives</td>
<td></td>
</tr>
<tr>
<td>2003</td>
<td>24</td>
<td>133 Hours of 1998 ABC, CNN news &amp; C-SPAN</td>
<td></td>
</tr>
<tr>
<td>2004</td>
<td>38</td>
<td>173 Hours of 1998 ABC, CNN news &amp; C-SPAN</td>
<td></td>
</tr>
<tr>
<td>2005</td>
<td>62</td>
<td>220 Hours of 2004 news (U.S., Arabic, Chinese sources, BBC stock shots)</td>
<td></td>
</tr>
<tr>
<td>2006</td>
<td>70</td>
<td>+156 Hours of NBC, CNN, MSN, CCTV</td>
<td></td>
</tr>
</tbody>
</table>

---

**Growing Data Sets**

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

Broadcast News

Program Category
- Entertainment
- Sports
- Weather

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
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/](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**
  - Video Storage
  - Streaming
  - Broadband
  - Object-based Manipulation
  - Media Logging
  - Enterprise Content Mgmt
  - Repurposing
  - Content Adaptation

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

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

- **Transactions of Digital Items** (MPEG-21)
  - E-Commerce Of Digital Content
  - Flexible Business Models

- **“Media Content Description”** (MPEG-7)
  - Rights management
  - Media mining and decision support

- **“Digital Content Coding”** (MPEG-1,2,4)
  - Compression
  - Coding
  - Communications
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)

- 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 (e.g., 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”

- **Links:**
  - IBM Research Intelligent Information Management Department: