Dynamic Multimodal Fusion in Video Search

Lexing Xie
IBM T J Watson Research Center
joint work with Apostol Natsev, Jelena Tesic, Rong Yan, John R Smith

WOCC-2007, April 28th, 2007
The Multimedia Search Problem

multi-modal query

“Find shots of flying through snow capped mountains”

Impact

- Widely applicable: consumer, media, enterprise, web, science …
- Bridging Traditional Search and Multimedia Knowledge Extraction

Multiple Tipping Points: multimedia semantics learning, multimedia ontology, training and evaluation, …

NIST TRECVID benchmark

- Validation for emerging technologies
- Scaling video retrieval technologies to large-scale applications
### Multimedia Search: Query Topics overview

<table>
<thead>
<tr>
<th>Query Types</th>
<th>Query topic</th>
<th>Query Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Find objects</td>
<td><strong>Topic 13. Speaker talking in front of the US flag</strong></td>
<td>Palm trees, tanks, ship/boat</td>
</tr>
<tr>
<td></td>
<td><strong>Topic 4. Scenes of snow capped mountains</strong></td>
<td>People shaking hands, people with banners, people in a meeting, people entering or leaving a building</td>
</tr>
<tr>
<td>Find events</td>
<td><strong>Topic 48. Other examples of overhead zooming-in views of canyons in Western United States</strong></td>
<td>Soccer score, basketball, tennis, airplane take-off, helicopter in flight</td>
</tr>
<tr>
<td>Find sites</td>
<td><strong>Topic 48. Other examples of overhead zooming-in views of canyons in Western United States</strong></td>
<td>Tall buildings, office setting, road with cars, fire</td>
</tr>
</tbody>
</table>
Outline

- The multimedia search challenge
- A birds-eye’s view of IBM Multimedia Search System
- Query-dependent multimodal fusion
- Evaluation on TRECVID benchmark
- Summary
IBM Multimedia Search System Overview

Approaches:

1. **Text-based**: story-based retrieval with automatic query refinement/re-ranking
2. **Model-based**: automatic query-to-model mapping based on query topic text
3. **Semantic-based**: cluster-based semantic space modeling and data selection
4. **Visual-based**: light-weight learning (discriminative and nearest neighbor modeling) with smart sampling
5. **Fusion**:
   - Query independent
   - Query-class-dependent with soft, hard or dynamic class membership
1 IBM Text Retrieval System

- **Corpus Indexing**
  - Shot-level ASR/MT documents
  - Story-level ASR/MT documents
  - Both aligned at phrase level

- **Query analysis:**
  - Tokenization, phrasing, stemming
  - Part-of-speech tagging & filtering

- **Query refinement:**
  - Pseudo-Relevance Feedback

- **Query execution and fusion**
  - IBM Semantic Search Engine (Juru)
  - Using TF*IDF-based retrieval
  - Fusion of shot- and story-based results

**Performance (MAP):**
- **2005:** 0.09873
- **2006:** 0.05169

*T. Volkmer et al, ICME 2006*
IBM Model-based Retrieval System

Lexical (WordNet) Approach
- Query analysis
  - Tokens, stems, phrases
  - Stop word removal
- Query refinement
  - Automatic mapping of query text to concept models & weights
  - Lexical approach based on WordNet Lesk similarity
- Query execution
  - Concept-based retrieval using statistical concept models
  - Weighted averaging model fusion

Performance (MAP)
- 2006: 0.029

Textual query topic
“Find shots of an airplane taking off”

Query

Query Analysis

“airplane”, “take off”

Query Refinement
(Query-to-Concept Mapping & Weighting)

Outdoors (1.0), Sky (0.8), Military (0.6)

Query Execution
(Concept-Based Retrieval)

Model-based ranking

Tokenizer
WordNet
Concept Lexicon
Concept Models
Repository
Haubold et al. (ICME 2006)
- Visual concepts can help refine query topics
- Data modeling for discriminative learning:
  - Content-based over-sampling of visual query examples to create positive examples
  - Content-based pseudo-negative sampling in the development and test set.
IBM Visual-Semantic Retrieval

- Observations
  - Visual context can disambiguate word senses

- Idea
  - Leverage automatic visual concept detectors for semantic model-based query refinement

- Semantic Concept Lexicon
  - Hierarchy of 39 LSCOM-lite concepts
  - Statistical models based on visual features and machine learning
  - Dimensions correspond to LSCOM-lite

- We model query topic examples in semantic space
  - Any descriptor space
Multi-modal Fusion

- Each modality is good in finding certain things
  - Text: named people, other named entity
  - Visual: semantic scenes consistent in color or layout, e.g., sports, weather
  - Concept Models: non-specific queries, e.g. protest, boat, fire
- Averaging the retrieval models helps in any case
- Query-class or query-cluster dependent fusion helps more [CMU, NUS, Columbia]
- Query soft/hard/dynamic class approach
  - Semantic query analysis for query matching
  - Matching based on PIQUANT-II Q&A text features
  - Weighted linear combination for fusion
  - Training Queries: TRECVID2005

**Performance MAP**
- 2006: 0.0756 -> 0.087 -> 0.0937
Extract Semantic Query Features

Input Query Text

“people with” computer display

Semantic Tagging

Person:CATEGORY
BodyPart:UNKNOWN
Furniture:UNKNOWN

Semantic Categories

Query Feature Vector

person
scene

[IBM UIMA and PIQUANT Analysis Engine]
Three Query-dependent Fusion Strategies

Qclass

Qcomp

Qdyn

new query

training queries

learned weight vectors

subsets of the training query from which weights are obtained.
Outline

- The multimedia search challenge
- A birds-eye’s view of IBM Multimedia Search System
- Query-dependent multimodal fusion
- Evaluation and demo on TRECVID benchmark
- Summary
NIST TRECVID Benchmark at a Glance

- NIST benchmark for evaluating state-of-the-art in video retrieval
- Benchmark tasks:
  - Semantic Concept Detection; Search (2003-); Rushes (2005-)
  - Shot detection and story segmentation

Growing Participation

- TRECVID 2001: 12* Participants
- TRECVID 2002: 17 Participants
- TRECVID 2003: 24 Participants
- TRECVID 2004: 38 Participants
- TRECVID 2005: 42* Participants
- TRECVID 2006: 54* Participants

Growing Data Sets

- TRECVID 2001: 11 Hours NIST video
- TRECVID 2002: 73 Hours Video from Prelinger archives
- TRECVID 2003: 133 Hours 1998 ABC, CNN news & C-SPAN
- TRECVID 2004: 173 Hours 1998 ABC, CNN news & C-SPAN
- TRECVID 2005: 220 Hours of 2004 news from U.S., Arabic, Chinese sources
  - 50 hours BBC stock shots (vacation spots)
- TRECVID 2006: 380 Hours of 2004 and 2005 news from U.S., Arabic, Chinese source
  - 100 hours BBC rushes (interviews)

* Number of participants that completed at least one task
Automatic/Manual Search Overall Performance (Mean AP)

- Multi-modal fusion doubles baseline performance!
- Visual and semantic retrieval have a significant impact over a text baseline

IBM Official Runs:
- Text (baseline): 0.041
- Text (story-based): 0.052
- Multimodal Fusion:
  - Query independent: 0.076
  - Query classes (soft): 0.086
  - Query classes (hard): 0.087

IBM Modal Runs:
- Semantic-based Run: 0.037
- Visual-based Run: 0.0696
TRECVID’06 Fusion Results (Relative Performance)

- Relative improvement (%)
  - query-class fusion vs. query-independent fusion

- Observations Qind → Qclass
  - Concept-related queries improved the most:
    “tall building”, “prisoner”, “helicopters in flight”, “soccer”
  - Named-entity queries improved slightly:
  - Generic people category deteriorated the most:
    “people in formation”, “at computer display”, “w/ newspaper”, “w/ books”
Improvement over TREC06’ and TREC’05

\[ Q_{dyn} \text{ and } Q_{comp} > Q_{class} >> Q_{ind} \]

<table>
<thead>
<tr>
<th>dataset</th>
<th>method</th>
<th>Qind</th>
<th>Oracle</th>
<th>Qclass</th>
<th>Qcomp</th>
<th>Qdyn</th>
</tr>
</thead>
<tbody>
<tr>
<td>trec06</td>
<td>all-queries</td>
<td>0.0756</td>
<td>0.1089</td>
<td>0.0897</td>
<td>0.0929</td>
<td>0.0937</td>
</tr>
<tr>
<td></td>
<td>non-sports</td>
<td>0.0695</td>
<td>0.0835</td>
<td>0.0686</td>
<td>0.0720</td>
<td>0.0728</td>
</tr>
<tr>
<td>trec05</td>
<td>all-queries</td>
<td>0.1681</td>
<td>0.2184</td>
<td>0.1591</td>
<td>0.1621</td>
<td>0.1604</td>
</tr>
<tr>
<td></td>
<td>non-sports</td>
<td>0.1263</td>
<td>0.1681</td>
<td>0.1291</td>
<td>0.1326</td>
<td>0.1307</td>
</tr>
</tbody>
</table>

New improvements come from the increased weights on concepts for object queries (next slide).
Results: Improved upon TRECVID’06 on generic people queries.
Demo

Welcome to the IBM Research Multimedia Analysis and Retrieval System External Site

<table>
<thead>
<tr>
<th>Search Engine System</th>
<th>Video Data Set</th>
<th>Overview</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM Multimedia Analysis and Retrieval System (D08a)</td>
<td>TRECVid 2005 Videos</td>
<td>![Overview Image]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Search Engine System</th>
<th>Video Data Set</th>
<th>Overview</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM Multimedia Analysis and Retrieval System (D08a)</td>
<td>BBC Rushes 2005 Videos</td>
<td>![Overview Image]</td>
</tr>
</tbody>
</table>

IBM Multimedia Analysis and Retrieval System
Contact: John R. Smith, IBM T. J. Watson Research Center

© 2007 IBM Corporation
Thank You

For more information:

- IBM Multimedia Retrieval Demo
  http://mp7.watson.ibm.com/
