Dynamic Multimodal Fusion in Video Search

Lexing Xie, Apostol Natsev, Jelena Tesic
IBM T. J. Watson Research Center

July 5th, ICME 2007, Beijing
The Multimedia Search Problem

- **Wide impact**
  - consumer, media, enterprise, web, science …

- **Multiple Tipping Points:**
  - Multimedia semantics learning,
  - Multimedia ontology,
  - Systematic benchmark and evaluations
## The Challenge of Choices in Multimodal Search

<table>
<thead>
<tr>
<th>Text</th>
<th>Image similarity</th>
<th>Semantic concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Find shots of Condoleezza Rice</td>
<td>Find shots of soccer game with goalpost visible</td>
<td>Find scenes of snow capped mountains</td>
</tr>
</tbody>
</table>

- Text: "...which he was accompanied by his wife are trying to light and Foreign Minister who resigned as successor Condoleezza Rice, Bush was greeted at the airport Governor-General of"
Outline

- The problem
- Dynamic multimodal fusion
  - Query feature extraction
  - Query matching
- Experiments and evaluation
- Conclusion
Prior Work on Query Dependent Fusion

- **Query-classes: manually defined**
  - [Chua et. al., NUS]
  - [Yan et. al., CMU]

- Joint clustering of queries in feature + performance space [Kennedy et. al.]

- **Our approach**
  - Can we fuse without pre-defined “query classes” or “query clusters”? 
  - Deep text analysis for query features and query matching
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

The associated press reported today that, Ms Brown, CEO of Textract Corp, had been recently spotted at the Summit meeting in Zurich. At 42, Ms. Brown, is the youngest CEO at the Summit...

[IBM UIMA and PIQUANT Analysis Engine]
Weighted linear fusion

Query

“Snow capped mountains”

Multimodal retrieval experts

Text  Visual  Concepts

Broadcast Video

\[ r = w_t r_t + w_v r_v + w_c r_c \]

\( r, r_t, r_v, r_c \in \mathcal{R} \)
Three Query-dependent Fusion Strategies

Qclass

new query

Qcomp

Qdyn

training queries

learned weight vectors

subsets of the training query from which weights are obtained.
Experimental setup

- **TRECVID benchmark corpora**
  - Multi-lingual news from 6 channels
  - Training/test split:
    - 2005 24 queries, 110 hours
    - 2006 24 queries, 160 hours

- Measure AP (average precision) on each query

*Average precision graph*

*News videos*
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
Automatic/Manual Search Overall Performance (Mean AP)

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

- Multi-modal fusion doubles baseline performance.
- Text-based query matching has the best performance among TRECVID runs
- Improves upon query-independent fusion by 14%
Query-dependent Fusion Breakdown

- **Observation:**
  - 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”

Relative improvement (%): Qcomp vs. Qind
Dynamic Query Fusion Results

- Qdyn further improves MAP of Qclass by 8%, to 0.094
- The improvement is mainly in generic people queries → where Qclass failed previously
Demo: multimodal interactive search
Conclusion

- Presented a solution to query-dependent multimodal fusion problem for video search
  - Query features based on deep semantic parsing
  - Dynamic query matching
  - Improvement on TRECVID queries

- Future work
  - More retrieval modality, more diverse queries
Thank You

For more information:

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

- “Data modeling strategies for imbalanced learning in video search”, J. Tesic, A. Natsev, L. Xie, J. Smith, ICME’07  → next poster session