CMU-Informedia @ TRECVID 2010
Known-item Search

Lei Bao\textsuperscript{1,2}, Arnold Overwijk\textsuperscript{1}, Alexander Hauptmann\textsuperscript{1}

\textsuperscript{1}School of Computer Science, Carnegie Mellon University
\textsuperscript{2}Institute of Computing Technology, Chinese Academy of Science
Outline

- System overview
- Three retrieval systems
  - Text-based retrieval with Lemur
  - Visual-based retrieval with Bipartite Graph Propagation Model
  - LDA-based multi-modal retrieval
- Multiple query-class dependent fusion
- Conclusions and future work
System overview

0185 Query: Find the video with three black horses eating from a pile of hay with a small red building behind them
0185 Key Visual Cues: horses, hay, red building

Query Reinforcement and Expansion

Text Query
- Keywords
- Visual keywords
- Filter keywords by Flickr API
- Expand keywords by Flickr API

Image Examples from Google Images

Retrieval Systems

Text-based Retrieval with Lemur

Visual-based Retrieval with Bipartite Graph Propagation Model

query-by-keyword  query-by-example

LDA-based Multi-modal Retrieval

Multiple Query-Class Dependent Fusion

Final Ranked Video List
Text-based Retrieval with Lemur

- six query types
  - keywords query
  - keywords filtered by Flickr tags
  - expand keywords by Flickr tags
  - visual cues query
  - visual cues filtered by Flickr tags
  - expand visual cues by Flickr tags

- six fields
  - 3 fields out of 74 in metadata:
    - description
    - title
    - keywords
  - Automatic Speech Recognition (ASR)
    - Microsoft Speech SDK 5.1
    - speech transcription from LIMSI
  - Optical Character Recognition (OCR)
    - all metadata fields, ASR and OCR are combined into 1 field

- fusion: give different weights for fields and query types.
Text-based Retrieval with Lemur

- six query types in six fields, tested on 122 sample topics

<table>
<thead>
<tr>
<th></th>
<th>all</th>
<th>description</th>
<th>title</th>
<th>keywords</th>
<th>ASR</th>
<th>OCR</th>
</tr>
</thead>
<tbody>
<tr>
<td>keywords</td>
<td>0.2549</td>
<td>0.1787</td>
<td>0.0863</td>
<td>0</td>
<td>0.0636</td>
<td>0.0328</td>
</tr>
<tr>
<td>keywords.filtered</td>
<td>0.2911</td>
<td>0.1688</td>
<td>0.0862</td>
<td>0</td>
<td>0.0661</td>
<td>0.0362</td>
</tr>
<tr>
<td>keywords.expand</td>
<td>0.0680</td>
<td>0.0024</td>
<td>0.0082</td>
<td>0</td>
<td>0.0021</td>
<td>0</td>
</tr>
<tr>
<td>visual cues</td>
<td>0.2640</td>
<td>0.1476</td>
<td>0.0842</td>
<td>0</td>
<td>0.0494</td>
<td>0.0351</td>
</tr>
<tr>
<td>visual cues.filtered</td>
<td>0.2785</td>
<td>0.1497</td>
<td>0.0998</td>
<td>0.0027</td>
<td>0.0709</td>
<td>0.0292</td>
</tr>
<tr>
<td>visual cues.expand</td>
<td>0.0569</td>
<td>0.0020</td>
<td>0.0171</td>
<td>0.0006</td>
<td>0.0007</td>
<td>0.0007</td>
</tr>
</tbody>
</table>
Visual-based Retrieval with Bipartite Graph Propagation Model

- **Explicit concepts**
  - pre-defined from human perspective
  - 130 concepts for semantic indexing task
  - 12 color concepts

- **Implicit concepts (latent topics)**
  - discovered from computer perspective
  - 200 implicit concepts: discovered by Latent Dirichlet Allocation (LDA)

- **Bipartite Graph Propagation Model-based Retrieval**
  - the relationship between query and explicit and implicit concepts can be described in a bipartite graph
  - after propagation stability, concept nodes with stronger connections with query nodes will win. The score of each concept node indicates its relevance to the queries
Visual-based Retrieval with Bipartite Graph Propagation Model

- Are query examples helpful?
- Are 12 color concepts helpful?
- Are implicit concepts helpful?
- Is the visual-based retrieval helpful?
  - 36 queries out of 420 have over 0.01 performance
  - in these 36 queries, 16 of them have zero performance in text-based retrieval.

<table>
<thead>
<tr>
<th></th>
<th>explicit (130)</th>
<th>explicit (130 +12 colors)</th>
<th>implicit (200)</th>
<th>explicit + implicit (342)</th>
</tr>
</thead>
<tbody>
<tr>
<td>query-by-keywords</td>
<td>0.0054</td>
<td>0.0064</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>query-by-examples</td>
<td>0.0070</td>
<td>0.0075</td>
<td>0.0047</td>
<td>0.0078</td>
</tr>
<tr>
<td>keywords+examples</td>
<td>0.0079</td>
<td>0.0094</td>
<td>----</td>
<td>0.0099</td>
</tr>
</tbody>
</table>
some reasons for the poor performance

- concept detectors
  - 304 topics out of 420 contain at least one of the predefined concept
  - only 27 topics out of these 304 have over 0.01 performance

- shot-based retrieval vs. video-based retrieval
  - 0185: find the video with three black horses eating from a pile of hay with tress and a small red building behind them

Figure 1. keyframes of the answer video for topic 0185.

- image examples vs. video examples
LDA-based Multi-modal Retrieval

- A generative topic model to describe the joint distribution of textual and visual features
  - the generative process of a video with $N^t$ text words and $N^v$ SIFT visual words
  - draw a topic proportion $\theta|\alpha \sim \text{Dir}(\alpha)$
  - for each text word $w^t$
    - choose a topic $z \sim \text{multinomial}(\theta)$
    - choose a word $w^t$ from $p(w^t|z, \beta^t)$, a multinomial probability conditioned on the topic $z$
  - for each visual word $w^v$
    - choose a topic $z \sim \text{multinomial}(\theta)$
    - choose a word $w^v$ from $p(w^v|z, \beta^v)$, a multinomial probability conditioned on the topic $z$
Multiple Query-class Dependent Fusion

- Ranking features
  - for each query, its ranking features is a $N \times K$ matrix. $N$ is the number of videos in collection. $K$ is the number of experts.
  - assumption: assign the queries with similar ranking features into one class helps to optimize weights for the class-dependent fusion.

- Present query based on ranking features
  - train “ranking words” by clustering, where each word is a $K$-dimensional vector
  - present each query as a bag of “ranking words”

- Cluster queries into several classes
- Optimize fusion weights for each class by exhaustive search
Multiple Query-class Dependent Fusion

- Fuse the results from six fields with keywords query
  - best run out of six
  - single query class dependent fusion
  - 5 query classes dependent fusion
Conclusions & Future Work

- Conclusions
  - textual information contributed the most
  - visual-based retrieval is promising

- Future Work
  - find a better formulation of the query
  - extend the visual-based retrieval from shot-based to video-based
  - re-rank the text-based result with visual feature
  - use multiple query-class dependent fusion to combine the text-based and visual-based retrieval