Multimedia Indexing and Retrieval in the Large Scale

A. Delopoulos
Target users: multimedia archives

- **Professional Audio Visual Content**
  - content which is professionally produced, has an associated industry for commercial exploitation

- **Scientific Content**
  - content where special expertise is required to interpret the content, usually much more specific and often used for research purposes as well as commercial.

- **Shared Public Content**
  - generally multimedia content created by people who are independent of a content-generating company, typically made available at no cost on public web sites

- **Personal Collections**
  - created by people who are mostly amateurs and retained in local storage media
The (large) scale

- Professional Audio Visual Content
  - Video centric
    - TV broadcasters
    - Large public archives associated with broadcasters
    - Estimated volume in Europe:
      - 10 million hours of film
      - 20 million hours of video
      - 20 million hours of audio
    - New production rate:
      - $O(24\text{hours} \times 116 \text{EBU members}) = \text{thousands} \text{ hours per day}$
The (large) scale

- **Professional Audio Visual Content**
  - **Video centric examples**
    - INA (FR): 1.2 Mhours video (+30 Khours/year), 1.6 Mhours audio
    - BBC (GB): 1.2 Mhours video
    - ITN source (GB): 800 Khours of footage
    - RAI (IT): 400 Khours video, 400 Khours audio
    - National Audio Visual Conservation Centre – Library of Congress (US): 1.1 million video items, 2 millions audio items
The (large) scale

- **Professional Audio Visual Content**
  - **Image centric**
    - Picture agencies
      - Subsidiaries of press agencies
  
- **Examples:**
  - Getty images (US): 25 M images
  - Belga images (BE): 3 M images + 6000 images/day
  - European Press Photo Agency: +750 images/day
Manual annotation is almost impossible

- **Detailed annotation**
  - Characterization/categorization of each media unit
    - Image level
    - Shot level (shot = a few seconds of video)

- **Distribution of the effort**
  - Collection of metadata from media producers
    - e.g., photographers
    - Teams of professional archivists
  - Mostly describe the context and not the content
  - Extremely difficult at concept level
    - $M$ concepts $\times N$ items
Automatic annotation

- High throughput
  - Thousands of images/video shots per day
- Less subjective
- Complimentary to available metadata
- At a semantic level
  - Concepts / High level features
Concepts in multimedia retrieval: Why?

- **The only way to index multimedia collections at the semantic level without context information/metadata.**
  - E.g., a subset of Belga images without annotations.
- **Currently the most promising method for efficient video indexing.**
  - Enables video search at shot level.
- **Both content – based and context – based search.**
  - Can be extracted using existing metadata, audiovisual content, as well as speech transcriptions.
  - Keywords rely solely on context (e.g., text in an html page).
- **Resolve query ambiguity.**
  - Similar to Wikipedia disambiguation pages.
User point of view

Two main options

- **Functionality transparent to the user.**
  - The user enters keywords as usual.
  - Keywords are mapped to concepts and the query is automatically expanded.

- **User specifies concepts explicitly.**
  - Concepts are entered using extra search forms or facets (e.g. concept:sky).
  - E.g., “Nicolas Sarkozy concept:crowd” to search for images/shots of Nicolas Sarkozy also depicting a crowd.
System point of view

Input: Multimedia document
- Content + metadata

Output: Concept probability scores
- To be indexed by the search system
Key aspects

- **Selection of Concepts**
  - Useful and Feasible
  - Disambiguation

- **Feature extraction**
  - Informative features
  - Complexity

- **Fusion**
  - Cross media
  - Cross type

- **Supervised learning**
  - Classifier choice
  - Need for ground truth = training data
    - Manual annotation tools
    - Automatic generation
Relevant research achievements in MUG

- **Procedures/protocols for**
  - Concept selection, description, and selection of annotation set.
  - Manual Annotation
    - Tools

- **Concept extraction components.**
  - Feature extraction, fusion and prediction.
  - Tools for development of concept predictors and selection of fusion strategies.

- **Automatic ground truth generation**
  - From clickthrough data
  - From tags
Concept definition protocol
Concept definition protocol (1)

- **Concept selection**
  - Characteristic terms selected using statistical analysis of past query logs
  - Terms filtered by both user and technical parties
  - Approximately 530 concepts

- **Concept disambiguation**
  - Concept name, definition, examples and related keywords.
  - [http://concepts.ee.auth.gr](http://concepts.ee.auth.gr)
Concept definition protocol (2)

Vitalis concept id: 539
Concept short name: chair
Concept description:
A chair is a kind of furniture for sitting, consisting of a back, and sometimes arm rests, commonly for use by one person. Chairs also often have four legs to support the seat raised above the floor. A positive picture contains at least one or more chairs clearly present in the picture. a bench where more people can sit on is a negative example

Relevant keywords:
chair, furniture, sit, sitting

Examples of positive images (optional):

Examples of negative images (optional):

Disambiguation
- Organization:
  Belga
- Name of disambiguator:
  Tom

Annotation
- Organization:
- Name of annotator:
- List of images to be annotated:
  chair.080724-230456.zip
Training set generation protocol
Training set generation protocol

- Manual annotation
- 2 annotation tools
- total 530 concepts
- 458 with > 5 positives

This process leads to fast generation of small training sets!
Cross-domain concept fusion

Concept detection module

**Low-level features**
- WBL, DCOLOR, CSD, HOUGH, TEXT

**Fusion**
- Early fusion
- Cross-domain concept fusion

**Classifiers**
- Array of SVMs (one per concept)
Cross-domain concept fusion

- Models from foreign domains improve effectiveness in a concept fusion scheme.
  - MAP improvement 21.5% for Belga, 14.66% for TRECVID-2005 (compared to early fusion)

![Diagram of concept fusion process]

Features
- Foreign domain base concept classifiers (detect “properties”)
- One score vector per low-level feature!

Score vector to final target concept classification layer
Cross-domain concept fusion (2)

- **Foreign model selection criteria** = “which concepts are appropriate to play the role of base concepts?”
- **Those with maximum information transfer** = 
  \[(\text{Entropy}) - (\text{Mutual Information})\]
  \[
  \arg\max_{s^m} \left[ H(s^m) - \frac{1}{m-1} \sum_{i=1}^{m-1} I(s^i, s^m) \right]
  \]

- **Two alternative “base concept” selection criteria**
  - **MaxInfo**: Maximizes information transfer for all low-level features, minimizes redundancy.
  - **Top-k**: Maximizes information transfer for each low-level feature separately. All low-level features are of equal importance.
- **Common list of base concepts for all target concepts**
Concept fusion vs early fusion

Results on the Belga Images

- Early Fusion (Exp. 1)
- MaxInfo (Exp. 2)

Visual only
Concept fusion vs early fusion

Results on the Belga Images

Visual+Text

A. Delopoulos, CERTH-ITI seminars, Oct. 20, 2010
Cross-domain concept fusion (3)

- Cross-domain concept fusion improves scalability to the number of concepts.

Need of less training data = can afford more concepts

12% vs 100%
Cross-domain concept fusion (3)

- Cross-domain concept fusion improves scalability to the number of concepts.

Need of less training data = can afford more concepts

20%-30% vs 95%

Train base concepts on Belga 100K
Train target concepts on TV2005
Predict on TV2005

Early fusion
Top-K
MaxInfo

Sample percentage

AP
Cross-domain concept fusion improves scalability to the number of concepts.

Need of less training data = can afford more concepts

Train base concepts on Belga 100K
Train target concepts on TV2005
Predict on TV2005

Better for few training samples

Early fusion
Top-K
MaxInfo
Computational complexity

• Rough estimate: Depends on architecture, memory size, etc

• Capable to predict 500 concepts on 20K images on an 8 core server within 1 day

• Comparable to the acquisition rate of professional image archives
Automatic training set generation using clickthrough data

Click through data

- Log files including past user-archive interaction
  - Query terms
  - Returned image set
  - User selection
- Selected image is relevant to query terms with high probability
Click through data: basic method

- In order to collect positive images as ground truth
  - For some concept C, define a set of relevant keywords
  - Identify (logged) queries including these keywords
  - Pick the images that were selected as a result of these queries
- Negative images are randomly chosen
  - Pretty accurate for low prior concepts
Click through data: need for expansion

- *This exact matching procedure may result to too few positive images*
- Need for expansion
  - *at the cost of accuracy loss*
Click through data: expansion approaches

- **Exact matching**
  - select images clicked for queries exactly matching the concept name

- **“Textual similarity”** (based on IR language models)
  - annotate each image with all queries for which it has been clicked
  - select images retrieved
    - for query: (i) concept name (ii) concept keywords
    - using retrieval model: (i) language model (LM) (ii) smoothed LM (LMS)

- **Clickgraph**
  - images clicked for the same query are likely to be relevant to each other
Click through data: selected images

- **Search logs provided by Belga news agency**
  - 101 days (June – October 2007)
  - professional users
  - 9,605 unique queries
  - 35,894 clicked images (out of the 97,628)

- **Selected 25 concepts for experiments**

<table>
<thead>
<tr>
<th>number of clicked images per method</th>
<th>exact</th>
<th>LM</th>
<th>LMS</th>
<th>LMS&lt;sub&gt;key&lt;/sub&gt;</th>
<th>Lm&lt;sub&gt;stem&lt;/sub&gt;</th>
<th>LMS&lt;sub&gt;stem&lt;/sub&gt;</th>
<th>LMS&lt;sub&gt;stem_key&lt;/sub&gt;</th>
<th>clickgraph</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>59.34</td>
<td>104.75</td>
<td>116.18</td>
<td>251.67</td>
<td>102.24</td>
<td>116.06</td>
<td>256.79</td>
<td>1217.18</td>
</tr>
</tbody>
</table>
Click through data: results

- All trained classifiers are better than random
- For visual features (Weibull distribution of edges):
  - combination of manual and searchlog-based training samples performs best consistently over all methods
- For text features (bag of words):
  - searchlog-based training samples produced by less noisy methods perform best
- Text features outperform visual features
- Clickgraph approach is too noisy
  - Needs post filtering
Click through data: results

Text features:
- Searchlog-based
- Manual
- Manual + Searchlog-based

Visual features:
- Searchlog-based
- Manual

Graph showing the performance of different methods with varying features.
Click through data: conclusion

- Few manually annotated images + images selected from clickthrough analysis
  - Effective
  - Scale up to large numbers of concepts
  - Easy to re-implement
    - In new archives
    - New domains
  - A wise choice to keep logs
Automatic training set generation using Flickr tags
Ground truth from Flickr tags

- Use estimated similarity to rank images wrt each concept
  - Positive samples: $N^+$ most similar
  - Negative samples: $N^-$ less similar

<table>
<thead>
<tr>
<th>Keywords</th>
<th>Concept</th>
<th>Image</th>
</tr>
</thead>
<tbody>
<tr>
<td>w1</td>
<td>k1</td>
<td>t1</td>
</tr>
<tr>
<td>w2</td>
<td>k1</td>
<td>t2</td>
</tr>
<tr>
<td>w3</td>
<td>k2</td>
<td>t3</td>
</tr>
<tr>
<td>w4</td>
<td>k2</td>
<td>t4</td>
</tr>
<tr>
<td>w5</td>
<td>k2</td>
<td>t5</td>
</tr>
<tr>
<td>w6</td>
<td>k3</td>
<td>t6</td>
</tr>
<tr>
<td>w7</td>
<td>k3</td>
<td></td>
</tr>
<tr>
<td>w8</td>
<td>k3</td>
<td></td>
</tr>
<tr>
<td>w9</td>
<td>k3</td>
<td></td>
</tr>
</tbody>
</table>

Wordnet based expansion
Corpus based similarity

A. Delopoulos, CERTH-ITI seminars, Oct. 20, 2010
Ground truth from Flickr tags

- Pretty good accuracy of the resulting concept classifiers
- Outperform classifiers based on NUS-WIDE [1] annotation of the same dataset
  - Even when this NUS-WIDE annotation is used for evaluation
- In progress

[1] NUS-WIDE: a real-world web image database from National University of Singapore
by: Tat S. Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, Yantao Zheng
VITALAS:
Video & image Indexing and reTrievAI in the LArge Scale
IP FP6 – 045389, 2006-2010

- This research was part of our contribution to VITALAS
- Integrated platform + new tools for indexing, browsing, searching professional archives of video and images
- Consortium:
  ERCIM (FR), EADS (FR), CWI (NL), FhG (GE), INRIA (FR), Fundación Robotiker (ES), INA (FR), University of Sunderland (UK), CERTH-ITI (GR), Codeworks (UK), Belga (BE), Institut für Rundfunktechnik GmbH (GE)
VITALAS:
Video & image Indexing and reTrievAI in the LArge Scale
TRECVID-2009

- Significantly improved performance compared to the median
Future

- **Larger scale**
  - internet

- **Live retraining based on user feedback**
  - Capture user response
  - Even non verbal

- **Improve classifiers**
  - Beyond SVM