Video & Image Indexing and Retrieval in the Large Scale

FP6 project VITALAS
Arjen P. de Vries
arjen@acm.org
• Three characteristics:
  – Use-case driven
    • Professional users
  – Larg-*ish* collections
    • Target: 10M images, 10k hours of video
  – Cross-media
    • Target: 1K – 3K ‘cross-media’
**Integration Project**

**New Functionalities**
- Visual Relevance Feedback
- Concept and Keyword Suggestion
- Interactive Cartographic Exploration
- Video search by audio (german)
- Video search by text & metadata
- Video detailed view

**New GUI**

**Robustification of services**
- Scalability of services and chains

---

**V1**
- Image Indexation chain
- Search by Image Upload
- 50 cross media concepts

**V1+**

**Early V2 (now)**

**V2 trials**
- Image search by local part
- Global similarity search
- Image search in video
- Concept and keyword suggestion

**New Functionalities**
- Video Indexation chain
- Logo search in audio and video
- Video summarisation
- Audio search in french and English
- Improvement of existing services

**V3 trials**

---

**New Functionalities**
- Video Indexation chain
- Logo search in audio and video
- Video summarisation
- Audio search in french and English
- Improvement of existing services

---

**Final Chorus Conference, Brussels, May 26th-27th 2009**
Personalization, access rights, ...

**User** sets:
- Personal Data
- Preferences

**Administrator** establishes:
- Role (rights to perform system operations)
- Access Profile (rights to access data sources)
Disclaimer

\[
\frac{(15\text{min} - 4\text{min})}{11 \text{ partners}} \approx 1 \text{ min / partner}
\]
Use-case driven?!

Unwanted results.... Just because ‘KIA’ is part of the title/caption
e.g. SLOVAKIA
QUERY: ‘KIA’ AND ‘LOGO’

Results seem correct but..
Only 10 pictures found

Final Chorus C:
Final Chorus Conference, Brussels, May 26th-27th 2009
Scalable Sub-image Search

• Improved access structure PMH V2
  – Random projections to reduce dimensionality, and compressed signatures to further reduce memory resource consumption
  – Real-time Global similarity search on 20 M images and more …
  – Real-time Local search in 100K images (150M local feature descriptors)
Scalable Sub-image Search

- Logos ground truth generation
  - Creation of a new challenging dataset (INRIA)
  - 10,000 BELGA images
  - Annotated with 26 logos (> 10 pixels)
  - 55 internal queries and 24 external queries
  - Made available to third parties for research usage

<table>
<thead>
<tr>
<th>Logo name</th>
<th>Illustration</th>
<th>nb of images</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adidas</td>
<td>adidas</td>
<td>114</td>
</tr>
<tr>
<td>Adidas-text</td>
<td></td>
<td>47</td>
</tr>
<tr>
<td>Base</td>
<td>BASE</td>
<td>408</td>
</tr>
<tr>
<td>Bouguen</td>
<td></td>
<td>11</td>
</tr>
<tr>
<td>Citroen</td>
<td>CITROEN</td>
<td>46</td>
</tr>
<tr>
<td>Citroen-text</td>
<td></td>
<td>97</td>
</tr>
<tr>
<td>CocaCola</td>
<td>CocaCola</td>
<td>32</td>
</tr>
<tr>
<td>CoFidis</td>
<td></td>
<td>17</td>
</tr>
<tr>
<td>Dexia</td>
<td>DEEXIA</td>
<td>494</td>
</tr>
<tr>
<td>Standard Liege</td>
<td></td>
<td>372</td>
</tr>
<tr>
<td>Eleclerc</td>
<td></td>
<td>8</td>
</tr>
<tr>
<td>Ferrari</td>
<td></td>
<td>76</td>
</tr>
<tr>
<td>Gucci</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Kia</td>
<td>KIA</td>
<td>82</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Logo name</th>
<th>Illustration</th>
<th>nb of images</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mercedes</td>
<td></td>
<td>73</td>
</tr>
<tr>
<td>Nike</td>
<td></td>
<td>116</td>
</tr>
<tr>
<td>Peugeot</td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>US President</td>
<td></td>
<td>14</td>
</tr>
<tr>
<td>Puma</td>
<td></td>
<td>54</td>
</tr>
<tr>
<td>Puma-text</td>
<td></td>
<td>22</td>
</tr>
<tr>
<td>Quick</td>
<td></td>
<td>56</td>
</tr>
<tr>
<td>Rochie</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>SNCF</td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>StellaArtois</td>
<td></td>
<td>15</td>
</tr>
<tr>
<td>TNT</td>
<td></td>
<td>66</td>
</tr>
<tr>
<td>VRT</td>
<td></td>
<td>9</td>
</tr>
</tbody>
</table>

Table 1: Annotated logos in BelgaLogos dataset
Cross-media Concepts

- Scalability issues for 1K – 3K concepts:
  - How to define the lexicon?
  - How to use the lexicon?
  - How to obtain training data for supervised ML?
Cross-media Concepts

- **Ground Truth Generation – Phase One**
  - Selection of 650 concepts with content owners
  - Selection of ~1000 positive and negative example/concept
  - Manual binary labelling
  - Disambiguation process

Final Chorus Conference, Brussels, May 26th-27th 2009
How to define the lexicon?

- Candidate concepts are selected semi-automatically
  - Derived from Belga captions and/or logs
  - Statistical test with Europarl as reference corpus
Automatically-selected concept vocabulary

- soccer (soccer, soccers, soccere): 27278.65
- pictur (picture, pictures, pictured, picturing, pictur, pictures, pictureds): 22565.06
- minist (minister, ministers, ministe, minist, ministeer, ministes, ministered): 21035.98
- team (team, teams, teamed, teaming): 18273.36
- cup (cup, cups, cupping, cupped): 18187.04
- citi (city, cities, citi, citys, citis, citie): 17369.72
- leagu (league, leagues, leagu, leagu): 16693.42
- celebr (celebrates, celebrate, celebrations, celebration, celebrated, celebrating, celebrities, celebrity, celebrants, celebrer, celebre, celebrators, celebrateing, celebreated, celebreates, celebres, celebrational, celebral, celebratings, celebrant): 16545.52
- won (won, wons): 16436.22
- championship (championships, championship): 16256.48
How to use the lexicon?

• Wikipedia-based Concept Suggestion:
  – Each Vitalas concept is associated with a Wikipedia article
  – Rank concept-related articles by keyword query
  – Suggest top-k concepts
How to obtain training data?

• Can we use click-through data instead of manually labelled samples?
  – Advantages:
    • Large quantities, no user intervention, collective assessments
  – Disadvantages:
    • Noisy & sparse
    • Queries not based on strict visual criteria
Manual annotations for the VITALAS 596 concepts

<table>
<thead>
<tr>
<th>annotations per concept</th>
<th>positive samples</th>
<th>negative samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEAN</td>
<td>1020.02</td>
<td>930.57</td>
</tr>
<tr>
<td>MEDIAN</td>
<td>998</td>
<td>970</td>
</tr>
<tr>
<td>STDEV</td>
<td>164.64</td>
<td>186.21</td>
</tr>
</tbody>
</table>
VITALAS manual annotations vs. search logs

Search logs and manual annotations for the VITALAS 128 concepts

BELGA NEWS AGENCY
~ 2,000,000 images

VITALAS
(current working set)
100,000 images

manually annotated
(with at least one of 128 concepts)
1925 images

Search logs (SL1)
4585 clicked images

1084 images
• Research questions:
  – How to annotate images with concepts using click-through data?
  – How reliable are click-through data based annotations?
  – What is the effectiveness of these annotations as training samples for concept classifiers?
How to annotate?

• Use queries for which images were clicked
  – Inherent noise: gap between queries/captions and concepts
    • queries describe the content+context of images to be retrieved
    • clicked images retrieved using their captions: content+context
    • concept-based annotations: based on visual content-only criteria
  – Sparsity: only cover part of the collection
    – Mismatch between terms in captions and queries
Effectiveness

• Experiment 1 (visual features):
  – training: search-log based annotations
  – test set for each concept: manual annotations (~1000 images)
  – feasibility study: in most cases, AP considerably higher than the prior

<table>
<thead>
<tr>
<th>Concept c</th>
<th>$T_{c,\text{exact}}$</th>
<th>$T_{c,\text{LM}}$</th>
<th>$T_{c,\text{LMS}}$</th>
<th>$T_{c,\text{LMS}_{\text{key}}}$</th>
<th>$T_{c,\text{LMstem}}$</th>
<th>$T_{c,\text{LMS}_{\text{stem}}}$</th>
<th>$T_{c,\text{LMS}_{\text{stem,\text{key}}}}$</th>
<th>$T_{c,\text{clickgraph}}$</th>
<th>prior</th>
</tr>
</thead>
<tbody>
<tr>
<td>airplane_flying</td>
<td>0.2269</td>
<td>0.3736</td>
<td>0.3736</td>
<td>0.0589</td>
<td>0.3736</td>
<td>0.3736</td>
<td>0.3736</td>
<td>0.0589</td>
<td>0.0262</td>
</tr>
<tr>
<td>airport</td>
<td>0.5920</td>
<td>0.6003</td>
<td>0.6003</td>
<td>0.5496</td>
<td>0.6003</td>
<td>0.6003</td>
<td>0.5501</td>
<td>0.4859</td>
<td>0.3968</td>
</tr>
<tr>
<td>anderlecht</td>
<td>0.5172</td>
<td>0.5499</td>
<td>0.5499</td>
<td>0.5499</td>
<td>0.4547</td>
<td>0.4547</td>
<td>0.5499</td>
<td>0.5172</td>
<td>0.3855</td>
</tr>
<tr>
<td>athlete</td>
<td>0.2166</td>
<td>0.2166</td>
<td>0.2166</td>
<td>0.2166</td>
<td>0.2166</td>
<td>0.2166</td>
<td>0.2166</td>
<td>0.0779</td>
<td>0.1034</td>
</tr>
<tr>
<td>building</td>
<td>0.5353</td>
<td>0.5766</td>
<td>0.5224</td>
<td>0.6082</td>
<td>0.5786</td>
<td>0.5224</td>
<td>0.6056</td>
<td>0.5030</td>
<td>0.4080</td>
</tr>
<tr>
<td>club.brugge</td>
<td>0.3854</td>
<td>0.0623</td>
<td>0.7227</td>
<td>0.7227</td>
<td>0.7227</td>
<td>0.7116</td>
<td>0.7116</td>
<td>0.0972</td>
<td>0.0972</td>
</tr>
<tr>
<td>crowd</td>
<td>0.5159</td>
<td>0.5794</td>
<td>0.5794</td>
<td>0.5333</td>
<td>0.5948</td>
<td>0.5948</td>
<td>0.4511</td>
<td>0.4261</td>
<td>0.3627</td>
</tr>
<tr>
<td>farms</td>
<td>0.4322</td>
<td>0.5478</td>
<td>0.5890</td>
<td>0.7242</td>
<td>0.5813</td>
<td>0.5813</td>
<td>0.7242</td>
<td>0.4208</td>
<td>0.1623</td>
</tr>
<tr>
<td>fashion_model</td>
<td>0.3062</td>
<td>0.0162</td>
<td>0.5453</td>
<td>0.5453</td>
<td>0.5453</td>
<td>0.5453</td>
<td>0.5453</td>
<td>0.4322</td>
<td>0.4322</td>
</tr>
<tr>
<td>fire</td>
<td>0.3868</td>
<td>0.4523</td>
<td>0.4523</td>
<td>0.3371</td>
<td>0.4395</td>
<td>0.4395</td>
<td>0.3424</td>
<td>0.3620</td>
<td>0.4333</td>
</tr>
<tr>
<td>flood</td>
<td>0.0162</td>
<td>0.0162</td>
<td>0.0162</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>logo</td>
<td>0.0010</td>
<td>0.0010</td>
<td>0.0010</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>meadow</td>
<td>0.0010</td>
<td>0.0010</td>
<td>0.0010</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rally_motorsport</td>
<td>0.7540</td>
<td>0.7837</td>
<td>0.8046</td>
<td>0.7197</td>
<td>0.8046</td>
<td>0.7197</td>
<td>0.8046</td>
<td>0.7197</td>
<td>0.4624</td>
</tr>
<tr>
<td>red_devils</td>
<td>0.5694</td>
<td>0.6435</td>
<td>0.6435</td>
<td>0.6435</td>
<td>0.6435</td>
<td>0.6435</td>
<td>0.6435</td>
<td>0.6435</td>
<td>0.4297</td>
</tr>
<tr>
<td>sky</td>
<td>0.2662</td>
<td>0.2662</td>
<td>0.2662</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>soccer</td>
<td>0.3954</td>
<td>0.3954</td>
<td>0.3954</td>
<td>0.3954</td>
<td>0.3954</td>
<td>0.3954</td>
<td>0.3954</td>
<td>0.3954</td>
<td>0.1091</td>
</tr>
<tr>
<td>stadium</td>
<td>0.2169</td>
<td>0.2169</td>
<td>0.2169</td>
<td>0.2169</td>
<td>0.2169</td>
<td>0.2169</td>
<td>0.2169</td>
<td>0.2169</td>
<td>0.1091</td>
</tr>
<tr>
<td>team</td>
<td>0.2153</td>
<td>0.2153</td>
<td>0.2153</td>
<td>0.2153</td>
<td>0.2153</td>
<td>0.2153</td>
<td>0.2153</td>
<td>0.2153</td>
<td>0.1091</td>
</tr>
<tr>
<td>tennis</td>
<td>0.5119</td>
<td>0.5119</td>
<td>0.5119</td>
<td>0.5119</td>
<td>0.5119</td>
<td>0.5119</td>
<td>0.5119</td>
<td>0.5119</td>
<td>0.3403</td>
</tr>
<tr>
<td>volleyball</td>
<td>0.1737</td>
<td>0.1737</td>
<td>0.1737</td>
<td>0.1737</td>
<td>0.1737</td>
<td>0.1737</td>
<td>0.1737</td>
<td>0.1737</td>
<td>0.2076</td>
</tr>
<tr>
<td>war</td>
<td>0.5030</td>
<td>0.4920</td>
<td>0.5059</td>
<td>0.4050</td>
<td>0.4928</td>
<td>0.5022</td>
<td>0.4067</td>
<td>0.4402</td>
<td>0.2523</td>
</tr>
<tr>
<td>MAP</td>
<td></td>
<td>0.5030</td>
<td>0.4920</td>
<td>0.5059</td>
<td>0.4050</td>
<td>0.4928</td>
<td>0.5022</td>
<td>0.4067</td>
<td>0.4402</td>
</tr>
</tbody>
</table>
Example: soccer

manually annotated positive samples

search log based annotated positive samples

test set results

More? See paper published at CIVR 2009!

Final Chorus Conference, Brussels, May 26th-27th 2009
**Scalable Speech Search**

- **Speech Search**
  
  Near-Realtime ASR using Multipass Decoding
  
  - SoA models for German
  - Baseline setups for French and English

  **Hybrid Speech Search Approach**
  
  - Parallel word and syllable ASR decoding
  - Hybrid retrieval from word and subword lattices
  - Advantage: no fixed vocabulary

  **Focus on both Accuracy and Efficiency in Retrieval**
  
  - High precision/recall without need of fixed lexicon
  - Enable vocabulary independent search on 10,000 hours of speech data
Scalable Speech Search

- **Speech Search Results - Hybrid Spoken Term Detection**

  - Data set: Broadcast News + Conversation Shows (German)
  - Exact hybrid on a simulated 10,000h syllable index: 0.06 seconds
  - Enables vocabulary independent speech search on large archive
  - Several retrieval algorithms available
    ⇒ Choice depends on precision / recall / efficiency requirements

<table>
<thead>
<tr>
<th>System Setup</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exact syllable transcript search</td>
<td>0.96</td>
<td>0.68</td>
</tr>
<tr>
<td>Syllable lattice retrieval</td>
<td>0.91</td>
<td>0.71</td>
</tr>
<tr>
<td>Syllable lattice retrieval ( + fuzzy search)</td>
<td>0.80</td>
<td>0.70</td>
</tr>
<tr>
<td>Exact syllable-word-hybrid</td>
<td>0.97</td>
<td>0.82</td>
</tr>
</tbody>
</table>
Scalable Speech Search

<table>
<thead>
<tr>
<th>Treffer 1</th>
<th>Start: 14:44</th>
<th>Match: 100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>... die weltweite <strong>finanzkrise</strong> hat die schwächen des kapitalistischen wirtschaftssystems ans nicht gebracht ...</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Treffer 2</th>
<th>Start: 26:23</th>
<th>Match: 100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>... doch noch retten zu können unter anderem eine firma heidelberg zement war durch die <strong>finanzkrise</strong> erheblichen bedrängnis geraten: außerdem hatte merkel bei fehlt spekulationen ...</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Treffer 3</th>
<th>Start: 27:41</th>
<th>Match: 100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>... die den tod des ihren siebziger jahrige bestätigte in den schreiben heißt es durch die <strong>finanzkrise</strong> verursachte wirtschaftlichen notlage seiner firmen zu wie die ohnmacht nicht behandeln ...</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Alle Informationen anzeigen
Audio Structuring

- Separation into homogeneous audio segments
- Detection of pre-defined acoustic scenarios
  - *Speech parts, speaker gender*
  - *Re-appearing speaker clusters, acoustic concepts*
Scalable Audio Search
Automatic Detection of Repeated Objects

100 hours of TV news

Structuring Object: Generic

Informative Object: Figures

Semantic Object: Maps

Final Chorus
Interactive Exploration

Local Interactive Cartographic Explorer

Global Interactive Cartographic Explorer

Presented at CIVR’08
Browse 100K Belga documents

Global map: the colored clusters are built from textual similarities. Documents selected by search functionalities are highlighted.

- Global map and zoom area
- Text or image search
- Image and text of the document pointed by the mouse
- Labels of the Cluster pointed by the mouse
Local map: local expansion based on image, textual or hybrid proximity. Documents close to each other are connected.

Modality setting.
(First) User Evaluations

- **Search with VITALAS integrated system - Corpus: 100K images (BELGA)**
  - Users were very enthusiastic about the possibilities offered by the advanced functionalities
  - More practice needed on new functionalities to perform better
  - Response time is not sufficient

- **Logo search, evaluation - Corpus 10000 images (BELGA)**
  - Very positive end user feedback
  - Our expert would like to extend the use beyond logos

- **Audio search, evaluation - Corpus: 1 year from ZDF news**
  - Results: very positive end user feedback
  - Timeline representation pertinent
International Benchmarks

• Organize
  – Wikipedia MM task (from INEX → ImageCLEF)
    • More Vitalas spin-off... ImageCLEF 2009 uses a 500k Belga collection!

• Participate
    • Ambition for 2009: Interactive search with professional archivists
  – INEX Efficiency, Entity Ranking
Conclusion

• Scalability:
  – Local and global visual search: 20M global features, 100K locals features
  – Vocabulary independent speech search on 10k hours of video

• Competitive cross-media concept detectors

• Overcome training data bottleneck through log file analysis
Final year’s work

- Main scientific objectives:
  - Scalability
  - Closing the loop between cross-media retrieval and indexing
- Main emphasis in terms of effort:
  - Improve intuitiveness and user control
- Main deliverables:
  - Operational system at user sites
• Many search problems in Vitalas have been approached as XML Retrieval:
  – Concept / term suggestion, concept / meta-data search, query log analysis
  – Features: XQuery integration, incremental indexing, working with arbitrary