Television Linked To The Web

Enrichment of News Show Videos with Multimodal Semi-Automatic Analysis

Daniel Stein\textsuperscript{1}, Evlampios Apostolidis\textsuperscript{2}, Vasileios Mezaris\textsuperscript{2}, Nicolas de Abreu Pereira\textsuperscript{3}, Jennifer Müller\textsuperscript{3}, Mathilde Sahuguet\textsuperscript{4}, Benoit Huet\textsuperscript{4}, Ivo Lašek\textsuperscript{5}

1 Fraunhofer Institute IAIS, Schloss Birlinghoven, Germany
2 Informatics and Telematics Institute, CERTH, Greece
3 rbb - Rundfunk Berlin-Brandenburg, 14482 Potsdam, Germany
4 Eurecom, Sophia Antpolis, France
5 Czech Technical University and University of Economics, Prague, Czech Republic

NEM Summit, Istanbul, October 2012

www.linkedtv.eu
Synopsis

- Introduction: LinkedTV Project
- News Show
- Intelligent Video Analysis
- Results
- Conclusions
LinkedTV — Television Linked To the Web

Vision:
- ubiquitously online cloud of Networked Audio-Visual Content
- decoupled from place, device or source

Aim:
- provide interactive multimedia service for non-professional end-users
- focus television broadcast content as seed videos

Web: http://www.linkedtv.eu

12 Excellent Partners

Fraunhofer STI GMBH
CERTH
UEP
UMONS
CWI
Eurecom
Condat
BEELD EN GELUID
Noterik
U. ST GALLEN
RBB
LinkedTV Workflow

Overall Architecture

- Use Case Scenarios
- Intelligent Video Analysis
- Linking Hypervideo to Web Content
- Interface and Presentation Engine
- Contextualization and Personalization
LinkedTV Workflow

Overall Architecture

Use Case Scenarios

Intelligent Video Analysis

Linking Hypervideo to Web Content

Interface and Presentation Engine

Contextualization and Personalization
Two Use Case Scenarios in LinkedTV

Scenario 1 (this talk): Interactive News Show
- Professional news content produced by RBB
- Seed content: local news show "rbb Aktuell"

- Due to legal constraints: whitelist
- Detailed scenario archetype description

Scenario 2 (not covered here): Hyperlinked Documentary
- Cultural content from S&V (1700 hours of cultural heritage AV-content under CCL)
- Seed content: "Antique Roadshow"
Intelligent Video Analysis

- Video
  - Shot Segmentation
  - Spatio-Temporal Segmentation
  - Moving Objects/Re-detection
  - Speaker Clustering/Recognition
  - Automatic Speech Recognition
  - Forced Alignment

- Visual track
  - Visual Concept Detection
  - Scene/Story Segmentation
  - Face Analysis

- Audio track
  - Named Entity Recognition
  - Keyword Extraction

- Subtitles
  - Tier 2, Tier 1

- XML
Segmentation

- Shot segmentation technique
  - [Tsamoura et. al., 2008]
- News show video performance: “remarkably well”
  - Out of 269 shots detected:
    - 2 had wrong starting points
    - 4 contained multiple shots
    - 11 were too short to evaluate properly
  - Spatio-temporal Segmentation
    - [Mezaris et. al., 2004]
- News show performance: Good
  - False positives due to:
    - Camera movement or zoom in/out (~ 55 %)
    - Gradual transition between frames (~ 10 %)
    - Erroneous motion vectors (~ 35 %)
  - Unwanted effect: false recognition of moving banners which not yield additional information


Concept Detection

- Method was described in [Moumtzidou et. al., 2011]
- 346 concepts from TRECVID 2011 SIN task
- Overall performance:
  - Correctly detected concepts > 64%
  - About 25% of them are characterized as particularly useful mostly related to detecting persons (e.g., person, face, adult)
  - Erroneous concepts vary between 22% - 42% and in many cases achieve high scores (e.g., outdoor, amateur video)

Visit: http://mklab.iti.gr/eventdetection-linkedtv/

Automatic Speech Recognition

Automatic speech recognition for German (using [Schneider08]):

<table>
<thead>
<tr>
<th>segment of one news show</th>
<th>WER</th>
<th>notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>new airport</td>
<td>36.2</td>
<td>outdoor, spontaneous</td>
</tr>
<tr>
<td>soccer riot</td>
<td>44.2</td>
<td>tavern, dialect, background noise</td>
</tr>
<tr>
<td>various news I</td>
<td>9.5</td>
<td></td>
</tr>
<tr>
<td>murder case</td>
<td>24.0</td>
<td></td>
</tr>
<tr>
<td>boxing</td>
<td>50.6</td>
<td>dialect, very spontaneous</td>
</tr>
<tr>
<td>various news II</td>
<td>20.9</td>
<td></td>
</tr>
<tr>
<td>rbb game</td>
<td>39.1</td>
<td></td>
</tr>
<tr>
<td>weather report</td>
<td>46.7</td>
<td>spontaneous, casual</td>
</tr>
</tbody>
</table>

Main obstacles: local dialect, spontaneous speech, background noise

Person Detection

- Face clustering using the face.com api
- Result: generally very good, some erroneous clusters due to side-view

- Speaker Identification using a GMM-HMM model, with 253 German parliament speakers
- Result: 8.0% Equal Error Rate
Conclusions

- We have established:
  - all the different video analysis techniques
  - their exact functionality
  - the connections among them

- Preliminary results work as a solid ground for further improvements

- Many challenges have been addressed but several aspects of the analysis techniques show much room for improvement, e.g.,
  - over-sensitivity of spatiotemporal segmentation algorithm to gradual transitions and camera’s movement
  - adaptation of several TRECVID concepts to the needs of each specific multimedia content (news show, documentary, art show)
  - over-sensitivity of speech recognizer to localized dialects and background noise
Future Plans

- **Incorporate new methods:**
  - Automatic Content Recognition
    - Goal: find parts that are already watched
  - Optical Character Recognition
    - Goal: exploit banner information to obtain a database for face and speaker recognition
  - Topic Segmentation
    - Goal: improve scene segmentation

- **Find synergies between methods:**
  - ASR + Speaker Recognition + Face Detection
    - Person Detection
  - ASR + Topic Classification + Shot Segmentation
    - Story Segmentation
  - Concept Detection + Keywords Extraction + Topic Segmentation
    - Video Similarity/Clustering
Questions?

More information:
http://www.iti.gr/~bmezaris
bmezaris@iti.gr