Multimedia Indexing

Prof. Bernard Mérialdo
EURECOM
Journées du LABRI - 17 Juin 2011

Outline

- Introduction
- TRECVID Video Indexing Benchmarks
- Video and MM Summarization
- Evaluation (BLEU, ROUGE, VERT)
- Conclusions
EURECOM

- Higher education and Research GIE
  - Subsidiary of Institut Télécom
  - Academic members (5 european universities)
  - Industrial members (10 international companies)

- International:
  - 80 Master students (15+ nationalities)
  - 70 PhD students (20+ nationalities)
  - 23 professors (11 nationalities)

- Research activity:
  - 3 departments: mobile, multimedia, security
  - 96 research contracts, 4,6M€
  - 200 publications / year
Multimedia Indexing

- **Information Overload:**
  - Google 40 G web pages
  - **flickr** >5G pictures
  - **YouTube** >15M videos

- **Audio-visual data:**
  - Sound and speech recognition
  - Natural Language Processing
  - Video Analysis

TRECVID Evaluation

- **International Evaluation Benchmark**
  - Organized by NIST since 2001
  - Schedule:
    - Data is distributed to participants
    - Participants run their algorithms, send results to NIST
    - NIST evaluates
    - Results are compared during workshop
  - Several tasks:
    - Semantic Indexing
    - Topic Search
    - Copy Detection
    - Event Detection
    - Summarization (2006-2008)
## TRECVID in numbers

**Participants:**

<table>
<thead>
<tr>
<th>Year</th>
<th>2001</th>
<th>2002</th>
<th>2003</th>
<th>2004</th>
<th>2005</th>
<th>2006</th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
<th>2010</th>
<th>2011</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hours</td>
<td>12</td>
<td>17</td>
<td>24</td>
<td>33</td>
<td>41</td>
<td>54</td>
<td>54</td>
<td>77</td>
<td>63</td>
<td>73</td>
<td>110</td>
</tr>
</tbody>
</table>

**Video data:**

<table>
<thead>
<tr>
<th>Year</th>
<th>Hours of video (training/test)</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001</td>
<td>11 NIST videos</td>
<td></td>
</tr>
<tr>
<td>2002</td>
<td>73 Internet Open Archive</td>
<td></td>
</tr>
<tr>
<td>2003</td>
<td>66/67 TV News (ABC, CNN, CSPAN)</td>
<td></td>
</tr>
<tr>
<td>2004</td>
<td>130/70 TV News (ABC, CNN, CSPAN)</td>
<td></td>
</tr>
<tr>
<td>2005</td>
<td>85/85 TV News (+arabic, chinese)</td>
<td></td>
</tr>
<tr>
<td>2006</td>
<td>170/158 TV News (+arabic, chinese)</td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td>50 BBC Rushes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>50/50 Sound and Vision (dutch)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>18/17 BBC Rushes</td>
<td></td>
</tr>
</tbody>
</table>

---

## TRECVID Video Data

<table>
<thead>
<tr>
<th>Year</th>
<th>Hours of video (training/test)</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008</td>
<td>100/100 Sound and Vision (dutch)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>35/18 BBC Rushes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>200 Surveillance (Gatwick airport)</td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>100/280 Sound and Vision (dutch)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>53/20 BBC Rushes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>150 Surveillance (Gatwick airport)</td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td>200/200 Internet Archive</td>
<td></td>
</tr>
<tr>
<td></td>
<td>180 Sound and Vision</td>
<td></td>
</tr>
<tr>
<td></td>
<td>45 Surveillance (Gatwick airport)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>50 BBC Rushes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>100 HAVIC</td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td>400/200 Internet Archive</td>
<td></td>
</tr>
<tr>
<td></td>
<td>150 Surveillance (Gatwick airport)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>50 BBC Rushes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>100 HAVIC</td>
<td></td>
</tr>
</tbody>
</table>
TRECVID 2011 Tasks

- Known-item search task (automatic, manual, interactive)
  - Search for a known video from text description
    - 0001 KEY VISUAL CUES: man, clutter, headphone
      - QUERY: Find the video of bald, shirtless man showing pictures of his home full of clutter and wearing headphone
    - 0002 KEY VISUAL CUES: Sega advertisement, tanks, walking weapons, Hounds
      - QUERY: Find the video of a Sega video game advertisement that shows tanks and futuristic walking weapons called Hounds.
    - 0003 KEY VISUAL CUES: Two girls, pink T shirt, blue T shirt, swirling lights background
      - QUERY: Find the video of two girls in a pink T shirt and blue T shirt, swirling lights in the background.
    - 0004 KEY VISUAL CUES: George W. Bush, man, kitchen table, glasses, Canada
      - QUERY: Find the video about the cost of drugs, featuring a man in glasses at a kitchen table, a video of Bush, and a sign saying Canada.
    - 0005 KEY VISUAL CUES: village, thatch huts, girls in white shirts, woman in red shorts, man with black hair
      - QUERY: Find the video of an Asian family visiting a village of thatch roof huts showing two girls with white shirts and a woman in red shorts entering several huts with a man with black hair doing the commentary.
TRECVID 2011 Tasks

- Content-based multimedia copy detection
  - Find transformed copies of a video segment

TRECVID 2011 Tasks

- Event detection in airport surveillance video
  - PersonRuns, Pointing, CellToEar, ObjectPut, Embrace, PeopleMeet, PeopleSplitUp
TRECVID 2011 Tasks

- **Instance search** (interactive, automatic)
  Searching a visual occurrence of a target given a few examples (for example logo detection, product and landmark recognition)

Person | Object | Location
--- | --- | ---

TRECVID 2011 Tasks

- **Semantic indexing (SIN)**
  - Find **shots** containing a given semantic concept
  - (previously called « High Level Feature Detection »)

- **Objective: build generic concept detectors**

- **Method:**
  - Assume concept presence is binary (contained or not)
  - System ranks shots by confidence score of presence
  - Best 2000 shots are returned for each feature
  - Manual assessment by NIST (yes/no for each shot)
  - Compute Mean Average Precision (MAP)
TRECVID Semantic Concepts

<table>
<thead>
<tr>
<th>2002</th>
<th>2003</th>
<th>2004</th>
<th>2005</th>
<th>2006</th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
</tr>
</thead>
<tbody>
<tr>
<td>outdoors</td>
<td>indoors</td>
<td>face</td>
<td>people</td>
<td>cityscape</td>
<td>landscape</td>
<td>text overlay</td>
<td>speech</td>
</tr>
<tr>
<td>news subject</td>
<td>news subject</td>
<td>face</td>
<td>people</td>
<td>road</td>
<td>vegetation</td>
<td>animal</td>
<td>female speech</td>
</tr>
<tr>
<td>car/truck/bus</td>
<td>aircraft</td>
<td>news subject</td>
<td>monologue</td>
<td>non-studio setting</td>
<td>sporting event</td>
<td>weather news</td>
<td>zoomin</td>
</tr>
<tr>
<td>Madeleine Albright</td>
<td>Bill Clinton</td>
<td>trains or railroad</td>
<td>cars</td>
<td>beach</td>
<td>basket score</td>
<td>airplane taking off</td>
<td>people walking</td>
</tr>
</tbody>
</table>

- Development Data: 130 concepts (+ ontology relations)
  - 200 hours: 200 hours
  - 130 Kshots: 130 Kshots

- Development data is manually annotated

- Test Data: 200 hours
  - 130 Kshots: 130 Kshots

- 2010: 130 concepts (+ ontology relations)

- 2011: 500 concepts (+ ontology relations)
  - 400 hours: 200 hours
  - 260 Kshots: 130 Kshots
TRECVID Concept Annotation

- Collaborative Annotation
- Active Learning

TRECVID 2011 Collaborative Annotation

Generic Concept Detector

Training Data

Airplane

Not Airplane

Feature Extraction

Supervised Classification

P(X=Airplane)

Feature Extraction

Classifier
Image/Video Features

- **Global:**
  - Color Histogram
  - Wavelets, Gabor filters, Edges
  - spatial arrangements

- **Local:**
  - SIFT, SURF

- **Motion:**
  - Optical flow, activity

- **Audio/Text:**
  - MFCC, speech/music/noise
  - Keywords from metadata or ASR

- **Specific:**
  - Face detection
  - Video-OCR

Local features: Bag Of Words

**Bag-Of-Words histogram**

Keypoint extraction → Clustering → Keypoint feature space → Visual word vocabulary

(Columbia U)
Supervised Classifiers

- **Classifiers:**
  - SVM Support Vector Machines
  - K-NN Nearest Neighbours
  - NN Neural Networks
  - Boosting
  - Ensemble
  - ...

- **Fusion:**
  - Different features / classifiers for a concept
  - Classifiers for different concepts

GDR-ISIS IRIM Action

- **Coordinated action LIG/MRIM, ETIS, LIF, LISTIC, LABRI, LIF, GIPSA, EURECOM**
  - LIG/hNdN : normalized RGB Histogram
  - LIG/gabN : normalized Gabor transform
  - LIG/opp_sift_har : bag of word, opponent sift
  - ETIS/global_X: histogram, quaternionic wavelets
  - LISTIC_Stip_X : nb of Spatio-Temporal Interest Points
  - LaBRI/faces : OpenCV+median temporal filtering
  - LaBRI/residualMotion_nPX : residual motion vectors
  - LIF/percepts : mid-level concepts on several grid blocks
  - GIPSA/AudioSpectro : spectral profile
  - GIPSA/AudioHamonicity
  - EUR/EUR-sm462 : Salency color moments
TRECVID Generic Architecture

For each concept

Cross-concept correlation

Video Data

Feature 1

Classifier 1

Feature 2

Classifier 2

Feature 3

Classifier 3

Feature N

Classifier N

Fusion

TRECVID 2010 SIN Performance

Top 10 InfAP scores by feature (Full runs)

10 life common features

InfAP

0.45

0.4

0.35

0.3

0.25

0.2

0.15

0.1

0.05

0.0

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

Adrenal

Adam_People

Bicycling

Cat_Racing

Casting

Dancing

Disinfection_People

Doorway

Extraction_File

Face_Recognition

Fusion

Greeting_Vehicles

Hitting

Jumping

Kick

Knocking

Landing

Passing

Running

Shaking

Sitting

Standing

Throwing

Vehicles

Walking

17 June 2011

LABRI Invited Presentation
TRECVID SIN Task

- Conclusions on SIN task:
  - System performance improve
    - More data or better model?
  - Cross-domain models are to be improved
    - How to quickly adapt to a new domain?
  - Computation requirements are increasing

Multimedia Summarization

- Overload of Multimedia information, specially videos
  - Lots of TV channels
  - Lots of recording devices

- Summarization is a useful tool:
  - Quickly grasp the main content
  - Decide to watch entire video or not
  - Allows to quickly compare several videos
  - Sometimes find relevant information

- Specific problems:
  - Multi-Media
  - Multi-Video
  - Evaluation
Video Summarization

Extract and assemble
Static keyframes
Dynamic VideoSkim

Video Summarization is difficult

Efficient selection requires:
- Analysis
- Modeling
- “Understanding”
- Evaluation of importance
Video Summarization is easy

- Lots of possible approaches for selection
  - From random choice
  - To numerical optimization
- How to prove that a summary is good (or bad)?
- A major problem is Evaluation

Video Summary Evaluation

- Many proposals, two basic approaches:
  - Objective metrics (quantitative)
    - SVD over feature frame matrix [Gong 2000]
    - Shot Reconstruction Degree [Liu 2004]
    - Shot importance [Uchihashi 1999]
  - User studies (qualitative)
    - Keyframe Counting [Dufaux 2000]
    - User satisfaction [Ngo 2003]
    - Content identification [Smith 1998, Lu 2004]
- Dilema: automatic vs realistic
TRECVID BBC Rushes summarization task

- Rushes from BBC archive
  - Unedited material from dramatic series

Rushes Video Structure

- A rushes video contains:
  - Junk frames
    - Test bar patterns
    - Junk recordings, irrelevant shots
  - Scenes
    - Recordings of a prepared action
    - A scene contains several takes
    - Each take is a tentative recording for the action
    - A take generally starts with a clapboard

<table>
<thead>
<tr>
<th>Scene 1</th>
<th>Scene 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Junk</td>
<td>Junk</td>
</tr>
<tr>
<td>Take 1</td>
<td>Take 1</td>
</tr>
<tr>
<td>Take 2</td>
<td>Take 2</td>
</tr>
<tr>
<td>Take 3</td>
<td></td>
</tr>
</tbody>
</table>
Rushes Video Structure

- Several takes of the same scene

TRECVID BBC Rushes summarization task

- 2006: organize, no evaluation
- 2007: summarize, evaluate
  - List of topics and events built for ground truth
  - 4% summary is built for each video
  - Evaluator watches summary and counts topics present
- 2008: summarize, evaluate
  - 2% summary is built for each video
  - Evaluator watches summary and counts topics present
- 2009: discontinued 😞
**Summarization Evaluation**

- **Ground truth**: human annotation of visible topics
- **Sample for MRS044500**:
  - 2 men in dark suits walk past Ford truck to building entrance
  - 2 men in dark suits enter building
  - person in brown coat opens rear end car and removes wheelchair (seen from front of car)
  - woman walks around car to passenger window (seen from rear end of car)
  - close up of man in passenger seat (seen from front of car)
  - woman in brown coat removes wheelchair and brings it round to the passenger door (seen from front of car)
  - man in beige suit appears (seen from front of car)
  - man in beige suit opens car door (seen from front of car)
  - woman in brown jacket undoes man in car’s seatbelt (seen from front of car)
  - woman in brown jacket helps passenger into wheelchair (seen from front of car)
  - …

---

**Video**

**Summary**

Randomly Selected Topics

Ground Truth

Y/N

Y/N

Y/N

Assessors
EURECOM 2008 Summarization system

COST 292 Summarization System

- Based on Spectral Clustering

- LABRI contribution: mid-level features:
  - Face
  - Camera movement
TRECVID Summarization Evaluation

Internet Multi-Video Summarization

- Collaboration with wikio.fr
  - News aggregator
  - Collect news information from various sites
  - Categorize/gather
  - Contains text, video

- Objective:
  - Multimedia summaries of multi-documents articles
Summarization Evaluation

- **Problem:**
  - No ground truth
  - No perfect summary
  - Different people will create different summaries
  - Different people will generally agree on summary ranking

- **Solution:**
  - Same problem appeared in:
    - Machine translation → BLEU evaluation
    - Text summarization → ROUGE evaluation
  - Idea: compare candidate with a set of references

Machine Translation Evaluation

- **Machine Translation BLEU [Papineni 2002]**
  - “BiLingual Evaluation Understudy”
  - N-gram precision based metric:

  \[
  BLEU_n = \frac{\sum_{n \in \text{C}} \sum_{n \in \text{clip}} \text{count}(n)}{\sum_{n \in \text{C}} \text{count}(n)}
  \]

  - Used in NIST evaluations

  *Cand 1:* It is a guide to action which ensures that the military always obeys the commands of the party

  *Cand 2:* It is to insure the troops forever hearing the activity guidebook that party direct

  *Ref 1:* It is a guide to action that ensures that the military will forever heed Party commands

  *Ref 2:* It is the guiding principle which guarantees the military forces always being under the command of the Party

  *Ref 3:* It is the practical guide for the army always to heed the directions of the party
Text Summarization Evaluation

- **ROUGE [Lin 2004]**
  - “Recall- Oriented Understudy for Gisting Evaluation”
  - N-gram recall based metric
  
  \[
  ROUGE_n = \frac{\sum_{\text{References}} \sum_{n\text{-gram} \in S} \text{count}_{match}(n\text{-gram})}{\sum_{\text{References}} \sum_{n\text{-gram} \in S} \text{count}(n\text{-gram})}
  \]

  - **Cand:** pulse series may ease schizophrenic voices
  - **Ref1:** magnetic pulse series sent through brain may ease imaginary schizophrenic sounds
  - **Ref2:** yale finds magnetic stimulation pulses may provide some relief to schizophrenic voices

Summarization Evaluation

- **VERT [Eurecom 2010]**
  - “Video Excerpt Relevance Threshold”
  - Recall based measure
  - Compares candidate keyframes with reference summaries
  
  \[
  VERT_n = \frac{\sum_{\text{ReferenceSummaries}} \sum_{\text{gram} \in S} w_c(\text{gram}_n)}{\sum_{\text{ReferenceSummaries}} \sum_{\text{gram} \in S} w_s(\text{gram}_n)}
  \]

  - **wC and wS:** weight of gramn
    - Weight = keyframe rank in selection
  - Several variants explored
Summarization Evaluation

- **First Experiments:**
  - Keyframe automatic selection
  - User keyframe selection and ranking
    - Reference summaries
  - Statistical comparison

VERT Evaluation

- **On going experiment:**
  - Large scale experimentation on Wikio laboratory site
  - 30 groups of 6 videos each, for each group, selection of (max) 60 keyframes
  - For each user:
    - Ask for keyframe selection and ranking
      - Reference summaries $R(u,v)$
      - Used to compute VERT
    - Ask for summary evaluation
      - By pair
    - Used to evaluate VERT
Conclusions

- MM Indexing is a hard problem
- Evaluation is required to measure progress
- Machine Learning is effective

But

- Comparative benchmarks tend to limit innovation
- Is more data better than smarter models?
- Which are the right criteria for evaluation?

Thank you

Merci