TRECVID-2013 Semantic Indexing task: Overview

Georges Quénot
Laboratoire d'Informatique de Grenoble

George Awad
Dakota Consulting, Inc
Outline

- Task summary
- Evaluation details
  - Inferred average precision
  - Participants
- Evaluation results
  - Pool analysis
  - Results per category
  - Results per concept
  - Significance tests per category
- Global Observations
- Issues
Semantic Indexing task

- **Goal**: Automatic assignment of semantic tags to video segments (shots)

- **Secondary goals**:
  - Encourage generic (scalable) methods for detector development.
  - Semantic annotation is important for filtering, categorization, searching and browsing.

- **Participants submitted four types of runs**:
  - **Main run** Includes results for 60 concepts, from which NIST and Quaero evaluated 38
    - **Localization run** includes results for 10 pixel-wise localized concepts from the 60 evaluated concepts in main runs. *NEW*
  - **Progress run** Includes results for 60 concept for 3 non-overlapping datasets, from which 2 datasets will be evaluated the next 2 years. *NEW*
  - **Pair run** Includes results for 10 concept pairs, all evaluated.
Semantic Indexing task (data)

- **SIN testing dataset**
  - Main test set (IACC.2.A): 200 hrs, with durations between 10 seconds and 6 minutes.
  - Progress test set (IACC.2.B, IACC.2.C): each 200 hrs and non overlapping from IACC.2

- **SIN development dataset**

- **Total shots:**
  - Much more than in previous TRECVID years, no composite shots
  - Development: 549,434
  - Test: IACC.2.A (112,677), IACC.2.B (107,806), IACC.2.C (113,467)

- Common annotation for 346 concepts coordinated by LIG/LIF/Quaero from 2007-2013 made available.
Semantic Indexing task (Concepts)

- Selection of the 60 target concepts
  - Were drawn from 500 concepts chosen from the TRECVID “high level features” from 2005 to 2010 to favor cross-collection experiments. Plus a selection of LSCOM concepts so that:
    - we end up with a number of generic-specific relations among them for promoting research on methods for indexing many concepts and using ontology relations between them
    - we cover a number of potential subtasks, e.g. “persons” or “actions” (not really formalized)
    - It is also expected that these concepts will be useful for the content-based (instance) search task.
  - Set of relations provided:
    - 427 “implies” relations, e.g. “Actor implies Person”
    - 559 “excludes” relations, e.g. “Daytime_Outdoor excludes Nighttime”
Semantic Indexing task (training types)

- Six training types were allowed:
  - A - used only IACC training data (110 runs)
  - B - used only non-IACC training data (0 runs)
  - C - used both IACC and non-IACC TRECVID (S&V and/or Broadcast news) training data (0 runs)
  - D - used both IACC and non-IACC non-TRECVID training data (0 runs)
  - E – used only training data collected automatically using only the concepts’ name and definition (6 runs)
  - F – used only training data collected automatically using a query built manually from the concepts’ name and definition (3 runs)
- E & F results inconclusive
  - E & F hardly represented - 9 runs
  - only 1 team system provided an E vs F pair
  - no clear difference.
38 concepts evaluated(1)

Single Concepts

3 Airplane*
5 Anchorperson
6 Animal
10 Beach
15 Boat_Ship*
16 Boy*
17 Bridges*
19 Bus
25 Chair*
31 Computers*
38 Dancing
49 Explosion_Fire
52 Female-Human-Face-Closeup
53 Flowers
54 Girl*
56 Government_Leader*
59 Hand
71 Instrumental_Musician*
72 Kitchen*
80 Motorcycle*
83 News_Studio
86 Old_People
89 People_Marching
100 Running
105 Singing*
107 Sitting_down*
117 Telephones
120 Throwing*
163 Baby*
227 Door_Opening
254 Fields*
261 Flags
267 Forest*
274 George_Bush*
342 Military_Airplane*
392 Quadruped
431 Skating
454 Studio_With_Anchor_person

-The 19 marked with “*” are a subset of those tested in 2012
Concepts evaluated (2)

- **Concept pairs**
  - [911] Telephones + Girl
  - [912] Kitchen + Boy
  - [913] Flags + Boat_Ship
  - [914] Boat_Ship + Bridges
  - [915] Quadruped + Hand
  - [916] Motorcycle + Bus
  - [917] Chair + George_W_Bush
  - [918] Flowers + Animal
  - [919] Explosion_Fire + Dancing
  - [920] Government-Leader + Flags

- **Localization concepts**
  - [15] Boat_ship
  - [17] Bridges
  - [19] Bus
  - [25] Chair
  - [59] Hand
  - [80] Motorcycle
  - [117] Telephones
  - [261] Flags
  - [392] Quadruped
NIST evaluated 15 concepts + 5 concept pairs and Quaero evaluated 23 concepts + 5 concept pairs.

- Each feature assumed to be binary: absent or present for each master reference shot
- Task: Find shots that contain a certain feature, rank them according to confidence measure, submit the top 2000
- NIST sampled ranked pools and judged top results from all submissions
- Metrics: inferred average precision per concept
- Compared runs in terms of mean inferred average precision across the:
  - 38 feature results for main runs
  - 10 feature results for concept-pairs runs
Inferred average precision (infAP)

- Developed* by Emine Yilmaz and Javed A. Aslam at Northeastern University
- Estimates average precision surprisingly well using a surprisingly small sample of judgments from the usual submission pools
- More features can be judged with same effort
- Increased sensitivity to lower ranks
- Experiments on previous TRECVID years feature submissions confirmed quality of the estimate in terms of actual scores and system ranking

2013: mean extended Inferred average precision (xinfAP)

- 2 pools were created for each concept and sampled as:
  - Top pool (ranks 1-200) sampled at 100%
  - Bottom pool (ranks 201-2000) sampled at 6.7%

<table>
<thead>
<tr>
<th>48 concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>336,683 total judgments</td>
</tr>
<tr>
<td>12006 total hits</td>
</tr>
<tr>
<td>8012 Hits at ranks (1-100)</td>
</tr>
<tr>
<td>3239 Hits at ranks (101-200)</td>
</tr>
<tr>
<td>755 Hits at ranks (201-2000)</td>
</tr>
</tbody>
</table>

- Judgment process: one assessor per concept, watched complete shot while listening to the audio.
- infAP was calculated using the judged and unjudged pool by sample_eval
<table>
<thead>
<tr>
<th>2013 : 26 Finishers</th>
</tr>
</thead>
<tbody>
<tr>
<td>PicSOM</td>
</tr>
<tr>
<td>INF</td>
</tr>
<tr>
<td>IRIM</td>
</tr>
<tr>
<td>VIREO</td>
</tr>
<tr>
<td>Dcu_savasa</td>
</tr>
<tr>
<td>EURECOM</td>
</tr>
<tr>
<td>VIDEOSENSE</td>
</tr>
<tr>
<td>TOSCA</td>
</tr>
<tr>
<td>FIU_Um</td>
</tr>
<tr>
<td>FHFLIHI</td>
</tr>
<tr>
<td>HFUT</td>
</tr>
<tr>
<td>IBM</td>
</tr>
<tr>
<td>ITI_CERTH</td>
</tr>
<tr>
<td>Quaero</td>
</tr>
<tr>
<td>JRS</td>
</tr>
<tr>
<td>AXES</td>
</tr>
<tr>
<td>NII</td>
</tr>
<tr>
<td>NHKSTRL</td>
</tr>
<tr>
<td>NTT</td>
</tr>
<tr>
<td>FTRDBJ</td>
</tr>
<tr>
<td>SRI_AURORA</td>
</tr>
<tr>
<td>TokyoTechCanon</td>
</tr>
<tr>
<td>Sheffield</td>
</tr>
<tr>
<td>MindLAB</td>
</tr>
<tr>
<td>MediaMill</td>
</tr>
<tr>
<td>UEC</td>
</tr>
</tbody>
</table>
Inferred frequency of hits varies by concept

**from total test shots

- Dancing
- Chair
- Female Human Face Close-up
- Hand
- Instrumental_Musician
- News_studio
- Old_people
- Singing
- anchorperson
- Boy
- Girl
- Old_people
- News_studio
### Total true shots contributed uniquely by team

#### Main runs

<table>
<thead>
<tr>
<th>Team</th>
<th>No. of Shots</th>
<th>Team</th>
<th>No. of shots</th>
</tr>
</thead>
<tbody>
<tr>
<td>NTT</td>
<td>65</td>
<td>FIU</td>
<td>10</td>
</tr>
<tr>
<td>Min</td>
<td>51</td>
<td>Kit</td>
<td>10</td>
</tr>
<tr>
<td>sri</td>
<td>49</td>
<td>FTR</td>
<td>8</td>
</tr>
<tr>
<td>EUR</td>
<td>38</td>
<td>ITI</td>
<td>8</td>
</tr>
<tr>
<td>FHH</td>
<td>32</td>
<td>Dcu</td>
<td>7</td>
</tr>
<tr>
<td>UEC</td>
<td>30</td>
<td>TOS</td>
<td>6</td>
</tr>
<tr>
<td>UvA</td>
<td>25</td>
<td>IBM</td>
<td>2</td>
</tr>
<tr>
<td>JRS</td>
<td>22</td>
<td>She</td>
<td>1</td>
</tr>
<tr>
<td>CMU</td>
<td>18</td>
<td>Tok</td>
<td>1</td>
</tr>
<tr>
<td>HFU</td>
<td>14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>vir</td>
<td>14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NHK</td>
<td>13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pic</td>
<td>11</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Pair runs

<table>
<thead>
<tr>
<th>Team</th>
<th>No. of Shots</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sri</td>
<td>3</td>
</tr>
<tr>
<td>CMU</td>
<td>2</td>
</tr>
<tr>
<td>HFU</td>
<td>1</td>
</tr>
</tbody>
</table>

Fewer unique shots compared to TV2012
Category A results (Main runs)

Median = 0.128

NIST baseline run
Top 10 InfAP scores by concept (Main runs)

<table>
<thead>
<tr>
<th>Rank</th>
<th>Concept</th>
<th>InfAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>3*</td>
<td>Airplane</td>
<td>0.38</td>
</tr>
<tr>
<td>5</td>
<td>Anchorperson</td>
<td>0.53</td>
</tr>
<tr>
<td>6</td>
<td>Animal</td>
<td>0.69</td>
</tr>
<tr>
<td>10</td>
<td>Beach</td>
<td>0.69</td>
</tr>
<tr>
<td>15*</td>
<td>Boat_ship</td>
<td>0.69</td>
</tr>
<tr>
<td>16*</td>
<td>Boy</td>
<td>0.66</td>
</tr>
<tr>
<td>17*</td>
<td>Bridges</td>
<td>0.72</td>
</tr>
<tr>
<td>19</td>
<td>Bus</td>
<td>0.71</td>
</tr>
<tr>
<td>25*</td>
<td>Chair</td>
<td>0.72</td>
</tr>
<tr>
<td>31*</td>
<td>Computers</td>
<td>0.72</td>
</tr>
<tr>
<td>38</td>
<td>Dancing</td>
<td>0.69</td>
</tr>
<tr>
<td>49</td>
<td>Explosion_Fire</td>
<td>0.64</td>
</tr>
<tr>
<td>52</td>
<td>Female_human_face_closeup</td>
<td>0.69</td>
</tr>
</tbody>
</table>

* Common concept in TV2012

Legend:
- Series1
- Series2
- Series3
- Series4
- Series5
- Series6
- Series7
- Series8
- Series9
Statistical significant differences among top 10 A-category runs (using randomization test, p < 0.05)

<table>
<thead>
<tr>
<th>Run name</th>
<th>mean infAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>UvA-Robb_1</td>
<td>0.321</td>
</tr>
<tr>
<td>UvA-Arya_2</td>
<td>0.300</td>
</tr>
<tr>
<td>UvA-Bran_3</td>
<td>0.296</td>
</tr>
<tr>
<td>UvA-Jon_4</td>
<td>0.286</td>
</tr>
<tr>
<td>Quaero-2013-3_3</td>
<td>0.285</td>
</tr>
<tr>
<td>Quaero-2013-2_2</td>
<td>0.285</td>
</tr>
<tr>
<td>TokyoTechCanon_2</td>
<td>0.284</td>
</tr>
<tr>
<td>TokyoTechCanon_1</td>
<td>0.284</td>
</tr>
<tr>
<td>TokyoTechCanon_3</td>
<td>0.283</td>
</tr>
<tr>
<td>Quaero-2013-4_4</td>
<td>0.283</td>
</tr>
</tbody>
</table>
Inferred frequency of hits for concept pairs

<table>
<thead>
<tr>
<th>Concept Pair</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>911 Telephones + Girl</td>
<td>50</td>
</tr>
<tr>
<td>912 Kitchen + Boy</td>
<td>30</td>
</tr>
<tr>
<td>913 Flags + Boat_ship</td>
<td>80</td>
</tr>
<tr>
<td>914 Boat_ship + Bridges</td>
<td>40</td>
</tr>
<tr>
<td>915 Quadruped + Hand</td>
<td>20</td>
</tr>
<tr>
<td>916 Motorcycle + Bus</td>
<td>10</td>
</tr>
<tr>
<td>917 Chair + George_W_Bush</td>
<td>90</td>
</tr>
<tr>
<td>918 Flowers + Animal</td>
<td>10</td>
</tr>
<tr>
<td>919 Explosion_Fire + Dancing</td>
<td>20</td>
</tr>
<tr>
<td>920 Government_leader + Flags</td>
<td>80</td>
</tr>
</tbody>
</table>
Category A results (Concept Pairs)

Mean InfAP:

Only 1 ‘E’ run submitted with score 0!
Why baseline runs sometimes are better than the regular runs?!
Statistical significant differences among top 10 A-category Concept Pairs runs (using randomization test, p < 0.05)

<table>
<thead>
<tr>
<th>Run name</th>
<th>(mean infAP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A_UvA-Shaggydog_8</td>
<td>0.162</td>
</tr>
<tr>
<td>A_UvA-Rickon_7</td>
<td>0.161</td>
</tr>
<tr>
<td>A_TokyoTechCanon_6</td>
<td>0.148</td>
</tr>
<tr>
<td>A_CMU_Todd_and_Rod_3</td>
<td>0.142</td>
</tr>
<tr>
<td>A_TokyoTechCanon_5</td>
<td>0.138</td>
</tr>
<tr>
<td>A_Quaero-2013-P5_5</td>
<td>0.127</td>
</tr>
<tr>
<td>A_Quaero-2013-P7_7</td>
<td>0.120</td>
</tr>
<tr>
<td>A_Quaero-2013-P6_6</td>
<td>0.120</td>
</tr>
<tr>
<td>A_CMU_Sherri_and_Terri_2</td>
<td>0.116</td>
</tr>
<tr>
<td>A_PicSOM_P_6_6</td>
<td>0.113</td>
</tr>
</tbody>
</table>

- A_TokyoTechCanon_6
  - A_Quaero-2013-P6_6
  - A_Quaero-2013-P7_7
  - A_TokyoTechCanon_5
- A_CMU_Todd_and_Rod_3
  - A_CMU_Sherri_and_Terri_2
  - A_PicSOM_P_6_6
Concept localization subtask

• Goal
  • Make concept detection more precise in time and space than current shot-level evaluation.

• Task
  • For each of the 10 concepts
    • For each of the top 1000 main run shots
      • For each I-Frame within the shot that contains the target, return
        • the x,y coordinates of the (UL,LR) vertices of a bounding rectangle containing all of the target concept and as little more as possible.
  • Systems were allowed to submit more than 1 bounding box per I-frame but only one with maximum fscore were judged.
NIST Evaluation framework

- Concept exists in shot (TP)
- Concept not in shot (FP)
- Sampling (random set of sequential I-frames)
  - 271k I-frames
- Concept exists in I-frame (TP)
- Concept not in I-frame (FP)
- Localization human assessors
- Draw bounding box

SIN human assessors
Evaluation metrics

- **Temporal localization**: precision, recall and f-score based on the judged I-frames.

- **Spatial localization**: precision, recall and f-score based on the located pixels representing the concept.

- An average of precision, recall and f-score for temporal and spatial localization across all I-frames for each concept and for each run.
Participants (Finishers)

• 4 teams submitted 9 runs
  • UvA (University Of Amsterdam)
  • SRI AURORA (SRI, Sarnoff, Central Fl.U., U. Mass., Cycorp, ICSI, Berkeley)
  • FTRDBJ (Orange Labs International Centers China)
  • QUAERO (INRIA, LIG, KIT)
Temporal localization results by run
Finding the best bounding box is much harder than finding just the I-frame.
TP vs FP submitted I-frames by run

How can systems find the right balance between TP vs FP I-frames?
Temporal localization results per concept

![Graph showing temporal localization results per concept with F-score on the y-axis and various concepts on the x-axis. The red line indicates the median performance for each concept.]
Spatial localization results per concept

Mean F-score

Airplane  | Boat_ship | Bridges | Bus | Chair | Hand | Motorcycle | Telephones | Flags | quadruped

- Median
Samples of good localization
Samples of less good localization
Results per concept across all teams

Majority of systems submitted a lot of non-target I-frames. While few found a balance.

Most systems submitted bounding boxes ~= G.T boxes AND overlaps. Systems are good in finding the real box sizes 😊
2013 Observations

• No submissions for training types B, C, & D
• Training types E & F still very few
• Fewer unique shots found vs TV2012
• No teams submitted any results for feature sequence in concept pairs!! Why?
• Concept-pairs baseline submissions are better than regular runs! (why? How to improve learning concept pairs?)
• For most localization systems, finding the correct I-frame is much easier than finding the bounding box
Site experiments include (not exhaustive):
- focus on robustness, merging many different representations
- use of spatial pyramids
- improved bag of word approaches
- Fisher/super-vectors, VLADs, VLATs
- audio analysis
- consideration of scalability issues
- improved rescoring methods
- use of semantic features
- work on the kernel size parameter of the SVM-RBF kernel
- work on the “no annotation” conditions: use of socially tagged videos or images and develop strategies for positive example selection
- deep convolutional neural networks (deep learning)
Announcements

• The full set of the 60 single concepts judgments are now available
• New qrels will be made available on the website
• No significant change in systems ranking are observed
SIN 2014

- Globally keep the task similar and of similar scale
- Further explore the “concept pair” and “no annotation” and “localization” variants
- Common training data for the “no annotation” variant is likely will be delivered LIG (F type)
- Sharing of data still proposed by IRIM
- Method for measuring progress over years
- Collaborative annotation unchanged
- Feedback welcome
Sharing of data for TRECVID SIN

- Organized by the IRIM groups of CNRS GRD ISIS.
- IRIM proposes its data sharing organization for the TRECVID SIN task. This comprises:
  - a wiki with read-write access for all
  - a data repository with read access for all and currently a write access only via one of the organizers
  - a small set of simple file formats
  - a (quite) simple directory structure

- Shared data mostly consist in descriptors and classification scores.

- Rewarding principle (same as for other contributions)
  - share and be cited and evaluated
  - use freely and cite
Sharing of data for TRECVID SIN

- Wiki (access with tv13 active participant login/password):
  - http://mrim.imag.fr/trecvid/wiki

- Associated data for SIN 2012 (access with IACC collection login/password):
  - http://mrim.imag.fr/trecvid/sin12

- Related actions:
  - Sharing of low-level descriptors by CMU for TRECVID 2003-2004
  - Mediamill challenge (101 concepts) using TRECVID 2005 data
  - Sharing of detection scores by CU-Vireo on TRECVID 2008-2010 data

- Possible extension to other TRECVID tasks, e.g. MED.