TRECVID-2005 High-Level Feature task: Overview

Wessel Kraaij
TNO
&
Paul Over
NIST
High-level feature task

- Goal: Build benchmark collection for detection methods
- Secondary goal: feature-indexing could help search/browsing
- Feature set selected from feature set used for annotation of development data (LSCOM-lite)
- Examples of thing/activity/person/location
- Collaborative development data annotation effort
  - Tools from CMU and IBM (new tool)
  - 39 features and about 100 annotators
  - Multiple annotations of each feature for a given shot
- Range of frequencies in the common development data annotation
True examples in the common training data

- Feature 38: 5300
- Feature 39: 748
- Feature 40: 910
- Feature 41: 480
- Feature 42: 6084
- Feature 43: 1081
- Feature 44: 626
- Feature 45: 106
- Feature 46: 1570
- Feature 47: 3390

13% of the data consists of true examples, with Feature 42 having the highest number of true examples (6084).
High-level feature evaluation

- 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 pooled submissions to depth 250
- Evaluate performance quality by measuring the average precision etc. of each feature detection method
10 Features

38. People walking/running: segment contains video of more than one person walking or running (tv4: 35)
39. Explosion or fire: segment contains video of an explosion or fire
40. Map: segment contains video of a map
41. US flag: segment contains video of a US flag
42. Building exterior: segment contains video of the exterior of a building (tv3: 14)
43. Waterscape/waterfront: segment contains video of a waterscape or waterfront
44. Mountain: segment contains video of a mountain or mountain range with slope(s) visible
45. Prisoner: segment contains video of a captive person, e.g., imprisoned, behind bars, in jail, in handcuffs, etc.
46. Sports: segment contains video of any sport in action (tv3: 23)
Participants (22/42)  *(up from 12/33 in 2004)*

<table>
<thead>
<tr>
<th>Institution</th>
<th>Country</th>
<th>--</th>
<th>LL</th>
<th>HL</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bilkent University</td>
<td>Turkey</td>
<td>--</td>
<td>LL</td>
<td>HL</td>
<td>SE</td>
</tr>
<tr>
<td>Carnegie Mellon University</td>
<td>USA</td>
<td>--</td>
<td>--</td>
<td>HL</td>
<td>SE</td>
</tr>
<tr>
<td>CLIPS-IMAG, LSR-IMAG, Laboratoire LIS</td>
<td>France</td>
<td>SB</td>
<td>--</td>
<td>HL</td>
<td>--</td>
</tr>
<tr>
<td>Columbia University</td>
<td>USA</td>
<td>--</td>
<td>--</td>
<td>HL</td>
<td>SE</td>
</tr>
<tr>
<td>Fudan University</td>
<td>China</td>
<td>SB</td>
<td>LL</td>
<td>HL</td>
<td>SE</td>
</tr>
<tr>
<td>FX Palo Alto Laboratory</td>
<td>USA</td>
<td>SB</td>
<td>--</td>
<td>HL</td>
<td>SE</td>
</tr>
<tr>
<td>Helsinki University of Technology</td>
<td>Finland</td>
<td>--</td>
<td>--</td>
<td>HL</td>
<td>SE</td>
</tr>
<tr>
<td>IBM</td>
<td>USA</td>
<td>SB</td>
<td>--</td>
<td>HL</td>
<td>SE</td>
</tr>
<tr>
<td>Imperial College London</td>
<td>UK</td>
<td>SB</td>
<td>--</td>
<td>HL</td>
<td>SE</td>
</tr>
<tr>
<td>Institut Eurecom</td>
<td>France</td>
<td>--</td>
<td>--</td>
<td>HL</td>
<td>--</td>
</tr>
<tr>
<td>Johns Hopkins University</td>
<td>USA</td>
<td>--</td>
<td>--</td>
<td>HL</td>
<td>--</td>
</tr>
<tr>
<td>Language Computer Corporation (LCC)</td>
<td>USA</td>
<td>--</td>
<td>--</td>
<td>HL</td>
<td>SE</td>
</tr>
<tr>
<td>LIP6-Laboratoire d'Informatique de Paris 6</td>
<td>France</td>
<td>--</td>
<td>--</td>
<td>HL</td>
<td>--</td>
</tr>
<tr>
<td>Lowlands Team (CWI, Twente, U. of Amsterdam)</td>
<td>Netherlands</td>
<td>--</td>
<td>--</td>
<td>HL</td>
<td>SE</td>
</tr>
<tr>
<td>Mediamill Team (Univ. of Amsterdam)</td>
<td>Netherlands</td>
<td>--</td>
<td>LL</td>
<td>HL</td>
<td>SE</td>
</tr>
<tr>
<td>National ICT Australia</td>
<td>Australia</td>
<td>SB</td>
<td>LL</td>
<td>HL</td>
<td>--</td>
</tr>
<tr>
<td>National University of Singapore (NUS)</td>
<td>Singapore</td>
<td>--</td>
<td>--</td>
<td>HL</td>
<td>SE</td>
</tr>
<tr>
<td>SCHEMA-Univ. Bremen Team</td>
<td>EU</td>
<td>--</td>
<td>--</td>
<td>HL</td>
<td>SE</td>
</tr>
<tr>
<td>Tsinghua University</td>
<td>China</td>
<td>SB</td>
<td>LL</td>
<td>HL</td>
<td>SE</td>
</tr>
<tr>
<td>University of Central Florida / University of Modena</td>
<td>USA, Italy</td>
<td>SB</td>
<td>LL</td>
<td>HL</td>
<td>SE</td>
</tr>
<tr>
<td>University of Electro-Communications</td>
<td>Japan</td>
<td>--</td>
<td>--</td>
<td>HL</td>
<td>--</td>
</tr>
<tr>
<td>University of Washington</td>
<td>USA</td>
<td>--</td>
<td>--</td>
<td>HL</td>
<td>--</td>
</tr>
<tr>
<td>Institution</td>
<td>Features</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------</td>
<td>----------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bilkent University</td>
<td>38</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Carnegie Mellon University</td>
<td>38 39 40 41 42 43 44 45 46 47</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CLIPS-IMAG, LSR-IMAG, Laboratoire LIS</td>
<td>38 39 40 41 42 43 44 45 46 47</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Columbia University</td>
<td>38 39 40 41 42 43 44 45 46 47</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fudan University</td>
<td>38 39 40 41 42 43 44 45 46 47</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FX Palo Alto Laboratory</td>
<td>38 39 40 41 42 43 44 45 46 47</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Helsinki University of Technology</td>
<td>38 39 40 41 42 43 44 45 46 47</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IBM</td>
<td>38 39 40 41 42 43 44 45 46 47</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Imperial College London</td>
<td>38 39 40 41 42 43 44 45 46 47</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Institut Eurecom</td>
<td>38 39 40 41 42 43 44 45 46 47</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Johns Hopkins University</td>
<td>38 39 40 41 42 43 44 45 46 47</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Language Computer Corporation (LCC)</td>
<td>38 39 40 41 42 43 44 45 46 47</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LIP6-Laboratoire d'Informatique de Paris 6</td>
<td>40</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lowlands Team (CWI, Twente, U. of Amsterdam)</td>
<td>38 39 40 41 42 43 44 45 46 47</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mediamill Team (Univ. of Amsterdam)</td>
<td>38 39 40 41 42 43 44 45 46 47</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>National ICT Australia</td>
<td>38 39 40 41 42 43 44 45 46 47</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>National University of Singapore (NUS)</td>
<td>38 39 40 41 42 43 44 45 46 47</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCHEMA-Univ. Bremen Team</td>
<td>40 41 43</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tsinghua University</td>
<td>38 39 40 41 42 43 44 45 46 47</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>University of Central Florida / Univ. of Modena</td>
<td>39 40 41 42 43 44 45 46 47</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>University of Electro-Communications</td>
<td>43 44</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>University of Washington</td>
<td>38 39 40 41 42 43 44 45 46 47</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Number of runs each training type

<table>
<thead>
<tr>
<th>Tr-Type</th>
<th>2005</th>
<th>2004</th>
<th>2003</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>79 (71.8%)</td>
<td>45 (54.2%)</td>
<td>22 (36.7%)</td>
</tr>
<tr>
<td>B</td>
<td>24 (21.8%)</td>
<td>27 (32.5%)</td>
<td>20 (33.3%)</td>
</tr>
<tr>
<td>C</td>
<td>7 (6.3%)</td>
<td>11 (13.3%)</td>
<td>18 (30.0%)</td>
</tr>
<tr>
<td>Total</td>
<td>110</td>
<td>83</td>
<td>60</td>
</tr>
</tbody>
</table>

**System training type:**

- **A** - Only on common dev. collection and the common annotation
- **B** - Only on common dev. collection but not on (just) the common annotation
- **C** - not of type A or B
AvgP by feature (all runs)
2005: AvgP by feature (top 10 runs)

38. People walking/running
39. Explosion/fire
40. Map
41. US flag
42. Building exterior
43. Waterscape/waterfront
44. Mountain
45. Prisoner
46. Sports
47. Car

Previous best result on CNN/ABC
2004: AvgP by feature (top 10 runs)
2003: AvgP by feature (top 10 runs)
AvgP by feature (top 3 runs by per feature)
Max AvgP by number of annotated training examples
Median AvgP by number of annotated training examples
Max AvgP by number true shots found
Median AvgP by number true shots found

![Graph showing median AvgP by number true shots found]
True shots contributed uniquely by team for each feature
Observations

- Participation almost doubled over 2004 (12 -> 22)
- Focus on category A runs (increased comparability)
- Scores are generally higher than in 2004 despite
  - new sources
  - errorful text from speech (via MT)
  - What does it mean?
- Did anybody run last year’s system on this year’s task?
- Features were generally found in all language sources
- Top scores come from fewer systems/groups
To follow: overview of the systems with map > 0.16 (median)

- Only systems that were tested on all 10 features
- Only category A
- Runs were compared on map across 10 features
Overview of approaches

- HLF systems draw from a very wide range of signal processing and machine learning techniques
  - Generic vs feature specific
  - How to do feature selection for visual modalities such as color and texture
  - Visual representation: grid or salient feature clusters
  - Various fusion methods, normalization methods
  - Range of classifiers
Carnegie Mellon University

- **Approach**
  - unimodal / multimodal (as in 2004)
  - learn dependencies between semantic features (by using various graphical model representations): inconclusive
  - global fusion < local fusion
  - multilingual > monolingual
  - multiple text sources > single text source
  - Best run: local fusion

- **Results:**

![Graph showing results](image-url)

TRECVID 2005
Approach
- Parts based object representation (ARG)
- Captures:
  - topological structure (spatial relationships among parts)
  - Local attributes of parts
- Model learns the parameter distribution properties due to differences in photometric conditions and geometry
- Runs vary across classifier combination schemes (fusion/selection)

Results:
- Significantly better than global (i.e. grid based) approach
- Esp. good for visual concepts where topology and local attributes are important (e.g. US flag)
- Text features play only a marginal role (contrastive experiment)
Fudan University

- **Approach:**
  - Several runs
    - Specific feature detectors
    - ASR based
    - Fusion of several unimodal SVM classifiers
    - Contrastive experiments with different dimension reduction techniques (PCA, locality preserving projection)

- **Results:**
  - Best run: 0.19
FXPAL

- **Approach**
  - SVM trained on low level features donated by CMU
  - Classifier combination schemes based on various forms of regression
  - 1st time participation

- **Results**
  - Best result: map=0.18
Helsinki University of Technology

- **Approach:**
  - Self Organizing maps trained on multimodal features and LSCOM lite annotations

- **Result:**
  - 1 run : map 0.2
IBM

- **Approach**
  - **Features:**
    - Visual: Extensive experiments for selecting best feature type and granularity for individual modalities (color, texture etc.)
    - Motion, Text, LSCOM LITE concepts
    - Features also included meta-information such as time of broadcast, channel etc.
  - SVM > (ME, KNN, GMM )
  - Flat and hierarchical feature fusion
  - Variations in classifier fusion methods
  - Feature specific approaches (selection based on held-out data)

- **Results:**

![Graph showing results](image-url)
Approach

1. “Naïve model”:
   - locate salient clusters in feature space
   - Learn HLF<-> clusters models

2. Nonparametric Density estimation (kernel smoothing)

Results:

- Naïve model: performance problems
- NPDE >> Naïve model
Mediamill team (Univ. of Amsterdam)
- presentation follows -

- Approach
  - Authoring metaphor
  - Feature specific combination of content, style and context analysis
  - 101 concept lexicon

- Results:
  - Textual features contribute only a small performance gain
National University of Singapore (NUS)

- **Approach**
  1. Ranked maximal figure of merit: ASR only, texture only, 2 fused runs
  2. HMM for visual dependency (4X4 grid): ASR only, +visual, +audio,genre,OCR . RankBoost fusion

- **Results:**
  - 2\textsuperscript{nd} approach $>>$ 1\textsuperscript{st} approach
University of Washington

- Approach: ?
  - (notebook paper not available yet)

- Results: