TRECVID-2016
Concept Localization : Overview

George Awad
National Institute of Standards and Technology
Dakota Consulting, Inc
• **Goal**
  - Make concept detection more precise in **time** and **space** than current shot-level evaluation.
  - Encourage context independent concepts design to increase their reusability.

• **Task set up**
  - For each of the 10 new test concepts, NIST provided set of ≈1000 shots.
  - Any shot **may or may not** contain the target concept.

• **Task**
  - 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 the ones with maximum f-score were scored.
## 10 New evaluated concepts

<table>
<thead>
<tr>
<th>Non action concepts</th>
<th>New action concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Animal</td>
<td>Bicycling</td>
</tr>
<tr>
<td>Boy</td>
<td>Dancing</td>
</tr>
<tr>
<td>Baby</td>
<td>Instrumental_musician</td>
</tr>
<tr>
<td></td>
<td>Running</td>
</tr>
<tr>
<td></td>
<td>Sitting_down</td>
</tr>
<tr>
<td></td>
<td>Skier</td>
</tr>
<tr>
<td></td>
<td>Explosion_fire</td>
</tr>
</tbody>
</table>
NIST Evaluation framework

• Testing data
  • IACC.2.A-C (600 h, used between 2013 to 2015 in semantic indexing task).

  • About 1000 shots per concept were sampled from the ground truth (with true positive (TP) clips of max = 300, avg = 178, min = 12).

  • Total of 9587 shots and 2205140 i-frames were distributed to systems.

  • Human assessors were given all the i-frames (total of 55789 images) of all TP shots to create the ground truth (drawing bounding box around the concept if it exists).

  • Human assessors had to watch the video clips of the images to verify the concepts.
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: 3 out of 21)

- 3 teams submitted 11 runs
  - TokyoTech (4 runs)
    - Tokyo Institute of Technology
  - NII_Hitachi_UIT (3 runs)
    - National Institute of Informatics; Hitachi, Ltd; University of Information Technology
  - UTS_CMU_D2DCRC (4 runs)
    - University of Technology, Sydney; Carnegie Mellon University; D2DCRC
Temporal localization results by run (sorted by F-score)

Mean per run across all concepts

I-frame F-score

I-frame Precision

I-frame Recall
Mean per run across all concepts

**2013**

- A_UniversityOfAmsterdam
- A_UniversityOfAmsterdam
- A_UniversityOfAmsterdam
- A_Quero-2013-2013
- A_FTRDBJ-M1.local_4
- A_FTRDBJ-M2.local_3
- A_FTRDBJ-M3.local_1
- A_sfiaurora

**2014**

- A_UniversityOfAmsterdam
- D_UniversityOfAmsterdam
- D_UniversityOfAmsterdam
- D_UniversityOfAmsterdam

**2015**

- CCNY_sub1.result.txt
- CCNY_sub2.result.txt
- CCNY_sub3.result.txt
- CCNY_sub4.result.txt
- insightdcu.DCU_Loc
- MediaMill_Qualcomm
- MediaMill_Qualcomm
- MediaMill_Qualcomm
- PicSOM.PicSOM_LO
- PicSOM.PicSOM_LO
- PicSOM.PicSOM_LO
- PicSOM.PicSOM_LO
- TokyoTech.run_tokyo
- TokyoTech.run_tokyo
- TokyoTech.run_tokyo
- TokyoTech.run_tokyo
- TokyoTech.run_tokyo
- Trimp_1.txt
- Trimp_2NEG_04.txt
- Trimp_3NEG_NOC
- Trimp_3NEG_015

2016 (mainly action) >> 2013 & 2014 (mainly objects)

ONLY TP shots were given to systems to localize.

Temporal Localization results
Spatial Localization results by run (sorted by F-score)

- Harder than temporal localization

- Mean Pixel F-score
- Mean Pixel Precision
- Mean Pixel Recall
Spatial Localization results

2016 (actions) > 2013 (objects)
2016 (actions) ~ 2014 (objects)

ONLY TP shots were given to systems to localize.
Most concepts perform better in temporal compared to spatial localization
A lot of resemblance between same concepts
Results per concept across all runs

Many systems submitted a lot of non-target I-frames, while few found a good balance.

submitted bounding boxes approximate the size of ground truth boxes and overlap with them. Many systems are good in finding the real box sizes.
General Observations

• Consistent observations in the last 4 years
  ✓ Temporal localization is easier than spatial localization.
  ✓ Systems report approximate G.T box sizes.

• Performance of action/dynamic concepts are higher than object concepts tested in 2013 to 2014.

• Assessment of action/dynamic concepts proved to be challenging in many cases to the human assessors.

• Lower finishing% of teams compared to signups.
Next team talks

- TokyoTech

- UTS_CMU_D2DCRC