TRECVID 2019

Ad-hoc Video Search Task : Overview

Georges Quénot
Laboratoire d'Informatique de Grenoble

George Awad
Georgetown University;
National Institute of Standards and Technology

Disclaimer
The identification of any commercial product or trade name does not imply endorsement or recommendation by the National Institute of Standards and Technology.
Outline

• Task Definition
• Video Data
• Topics (Queries)
• Participating teams
• Evaluation & results
• General observation
Task Definition

- **Goal:** promote progress in content-based retrieval based on end user **ad-hoc (generic) textual queries** that include searching for persons, objects, locations, actions and their combinations.

- **Task:** Given a test collection, a query, and a master shot boundary reference, return a ranked list of at most 1000 shots (out of 1,082,657) which best satisfy the need.

- **Testing data:** 7475 Vimeo Creative Commons Videos (V3C1), 1000 total hours with mean video durations of 8 min. Reflects a wide variety of content, style and source device.

- **Development data:** ≈2000 hours of previous IACC.1-3 data used between 2010-2018 with concept and ad-hoc query annotations.
Query Development

• Test videos were viewed by 10 human assessors hired by the National Institute of Standards and Technology (NIST).

• 4 facet descriptions of different scenes were used (if applicable):
  – Who: concrete objects and beings (kind of persons, animals, things)
  – What: are the objects and/or beings doing? (generic actions, conditions/state)
  – Where: locale, site, place, geographic, architectural
  – When: time of day, season

• In total assessors watched random selection of ≈1% (12000 videos) of the V3C1 segmented shots.

• All random shots were selected to cover all original 7475 videos.

• 90 candidate queries chosen from human written descriptions to be used between 2019 to 2021 including 20 progress topics (10 shared with the Video Browser Showdown (VBS)).
TV2019 Queries by complexity

- **Person + Action + Object + Location (most complex)**
  - Find shots of a woman riding or holding a bike outdoors
  - Find shots of a person smoking a cigarette outdoors
  - Find shots of a woman wearing a red dress outside in the daytime

- **Person + Action + Location**
  - Find shots of a man and a woman dancing together indoors
  - Find shots of a person running in the woods
  - Find shots of a group of people walking on the beach

- **Person + Action/state + Object**
  - Find shots of a person wearing a backpack
  - Find shots of a race car driver racing a car
  - Find shots of a person holding a tool and cutting something
TV2019 Queries by complexity

- **Person + Object + Location**
  - Find shots of a person wearing shorts outdoors
  - Find shots of a person in front of a curtain indoors

- **Person + Object**
  - Find shots of a person with a painted face or mask
  - Find shots of person in front of a graffiti painted on a wall
  - Find shots of a person in a tent

- **Object + Location**
  - Find shots of one or more picnic tables outdoors
  - Find shots of coral reef underwater
  - Find shots of one or more art pieces on a wall
TV2019 Queries by complexity

• **Object + Action**
  - Find shots of a drone flying
  - Find shots of a truck being driven in the daytime
  - Find shots of a door being opened by someone
  - Find shots of a small airplane flying from the inside

• **Person + Action**
  - Find shots of a man and a woman holding hands
  - Find shots of a black man singing
  - Find shots of a man and a woman hugging each other

• **Person/being + Location**
  - Find shots of a shirtless man standing up or walking outdoors
  - Find shots of one or more birds in a tree
TV2019 Queries by complexity

- **Object**
  Find shots of a red hat or cap

- **Person**
  Find shots of a woman and a little boy both visible during daytime
  Find shots of a bald man
  Find shots of a man and a baby both visible
Training and run types

• Three run submission types:
  ✓ Fully automatic (F): System uses official query directly (37 runs)
  ✓ Manually-assisted (M): Query built manually (10 runs)
  ✓ Relevance Feedback (R): Allow judging top-5 once (0 runs)

• Four training data types:
  ✓ A – used only IACC training data (7 runs)
  ✓ D – used any other training data (33 runs)
  ✓ E – used only training data collected automatically using only the query text (7 run)
  ✓ F – used only training data collected automatically using a query built manually from the given query text (0 runs)

• New novelty run was introduced to encourage retrieving non-common relevant shots easily found across runs.
# Main Task Finishers: 10 out of 19

<table>
<thead>
<tr>
<th>Team</th>
<th>Organization</th>
<th>Runs</th>
</tr>
</thead>
<tbody>
<tr>
<td>INF</td>
<td>Carnegie Mellon University (USA); Monash University (Australia); Renmin University (China); Shandong University (China)</td>
<td>M: 4</td>
</tr>
<tr>
<td>Kindai_kobe</td>
<td>Department of Informatics, Kindai University; Graduate School of System Informatics, Kobe University</td>
<td>M: 4</td>
</tr>
<tr>
<td>EURECOM</td>
<td>EURECOM</td>
<td>M: 3</td>
</tr>
<tr>
<td>IMFD_IMPRESEE</td>
<td>Millennium Institute Foundational Research on Data (IMFD) Chile; Impresee Inc ORAND S.A. Chile</td>
<td>M: 4</td>
</tr>
<tr>
<td>ATL</td>
<td>Alibaba group; ZheJiang University</td>
<td>M: 4</td>
</tr>
<tr>
<td>WasedaMeiseiSoft bank</td>
<td>Waseda University; Meisei University; SoftBank Corporation</td>
<td>M: 4</td>
</tr>
<tr>
<td>VIREO</td>
<td>City University of Hong Kong</td>
<td>M: 2</td>
</tr>
<tr>
<td>FIU_UM</td>
<td>Florida International University; University of Miami</td>
<td>M: 6</td>
</tr>
<tr>
<td>RUCMM</td>
<td>Renmin University of China; Zhejiang Gongshang University</td>
<td>M: 4</td>
</tr>
<tr>
<td>SIRET</td>
<td>Charles University</td>
<td>M: 4</td>
</tr>
</tbody>
</table>

**M**: Manually-assisted, **F**: Fully automatic, **R**: Relevance feedback, **N**: Novelty run
# Progress Task Submitters: 9 out of 10

<table>
<thead>
<tr>
<th>Team</th>
<th>Organization</th>
<th>Runs</th>
</tr>
</thead>
<tbody>
<tr>
<td>INF</td>
<td>Carnegie Mellon University (USA); Monash University (Australia) Renmin University (China) Shandong University (China)</td>
<td>-</td>
</tr>
<tr>
<td>Kindai_kobe</td>
<td>Department of Informatics, Kindai University; Graduate School of System Informatics, Kobe University</td>
<td>-</td>
</tr>
<tr>
<td>EURECOM</td>
<td>EURECOM</td>
<td>-</td>
</tr>
<tr>
<td>ATL</td>
<td>Alibaba group; ZheJiang University</td>
<td>-</td>
</tr>
<tr>
<td>WasedaMeiseiSoft bank</td>
<td>Waseda University; Meisei University; SoftBank Corporation</td>
<td>4</td>
</tr>
<tr>
<td>VIREO</td>
<td>City University of Hong Kong</td>
<td>2</td>
</tr>
<tr>
<td>FIU_UM</td>
<td>Florida International University; University of Miami</td>
<td>-</td>
</tr>
<tr>
<td>RUCMM</td>
<td>Renmin University of China; Zhejiang Gongshang University</td>
<td>-</td>
</tr>
<tr>
<td>SIRET</td>
<td>Charles University</td>
<td>4</td>
</tr>
</tbody>
</table>

M: Manually-assisted, F: Fully automatic, R: Relevance feedback, N: Novelty run
Evaluation

Each query assumed to be binary: absent or present for each master reference shot.

NIST judged top ranked pooled results from all submissions 100% and sampled the rest of pooled results.

**Metrics:** *Extended inferred average precision per query.*

Compared runs in terms of **mean extended inferred average precision** across the 30 queries.
Mean Extended Inferred Average Precision (XInfAP)

2 pools were created for each query and sampled as:

- Top pool (ranks 1 to 250) sampled at 100 %
- Bottom pool (ranks 251 to 1000) sampled at 11.1 %
- % of sampled and judged clips from rank 251 to 1000 across all runs and topics (min= 10.8 %, max = 86.4 %, mean = 47.6 %)

Judgment process: one assessor per query, watched complete shot while listening to the audio. infAP was calculated using the judged and unjudged pool by sample_eval tool

<table>
<thead>
<tr>
<th>30 queries</th>
</tr>
</thead>
<tbody>
<tr>
<td>181649 total judgments</td>
</tr>
<tr>
<td>23549 total hits</td>
</tr>
<tr>
<td>10910 hits at ranks (1 to 100)</td>
</tr>
<tr>
<td>8428 hits at ranks (101 to 250)</td>
</tr>
<tr>
<td>4211 hits at ranks (251 to 1000)</td>
</tr>
</tbody>
</table>

# Hits >> IACC data (2016-2018)
Inferred frequency of hits varies by query

Inf. Hits / query

Queries

Inf. hits

0.00
1,000.00
2,000.00
3,000.00
4,000.00
5,000.00

0.5% of test shots

person holding a tool and cutting something

person in front of a curtain indoors

shirtless man standing up or walking outdoors

woman wearing a red dress outside in the daytime

truck being driven in the daytime

0.5% of test shots
Total **unique relevant** shots contributed by team across all runs

Top scoring teams not necessary contributing a lot of unique true shots
Novelty Metric

• **Goal**
  Novelty runs are supposed to retrieve more unique relevant shots as opposed to more common relevant shots easily found by most runs.

• **Metric**
  1- A weight is given to each topic and shot pairs in the ground truth such that highest weight is given to unique shots:

  \[
  \text{TopicX}_\text{ShotY}\_\text{weight} = 1 - \left(\frac{N}{M}\right)
  \]

  Where \(N\) : Number of times Shot Y was retrieved for topic X by any run submission.
  \(M\) : Number of total runs submitted by all teams

  E.g. A unique relevant shot weight = 0.978 (given 47 runs in 2019), a shot submitted by all runs = 0.

  2- For Run R and for all topics, we calculate the summation S of all *unique* shot weights ONLY.

  Final novelty score = \(S/30\) (the mean across all evaluated 30 topics)
Novelty scores

- **Novelty runs**
- **Common runs**
Sorted overall scores
(37 Fully automatic runs, 9 teams)

Mean Inf. AP

Median = 0.08

Runs
Sorted scores

(10 Manually-assisted runs, 3 teams)

Mean Inf. AP

Runs

Median = 0.09
Top 10 runs by query (Fully Automatic)

- person in front of a graffiti painted on a wall
- person wearing a backpack
- inside views of a small airplane flying
- Person holding tool and cutting something
Top 10 runs by query (Manually-Assisted)
Unique vs Common relevant shots

13% of all hits are unique
## Performance in the last 4 years?

<table>
<thead>
<tr>
<th></th>
<th>IACC.3 Dataset</th>
<th></th>
<th>V3C1 Dataset</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2016</td>
<td>2017</td>
<td>2018</td>
<td>2019</td>
</tr>
<tr>
<td><strong>Automatic</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teams</td>
<td>9</td>
<td>8</td>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td>Runs</td>
<td>30</td>
<td>33</td>
<td>33</td>
<td>37</td>
</tr>
<tr>
<td>Min $x\text{InfAP}$</td>
<td>0.000</td>
<td>0.026</td>
<td>0.003</td>
<td>0.014</td>
</tr>
<tr>
<td>Max $x\text{InfAP}$</td>
<td>0.054</td>
<td>0.206</td>
<td>0.121</td>
<td>0.163</td>
</tr>
<tr>
<td>Median $x\text{InfAP}$</td>
<td>0.024</td>
<td>0.092</td>
<td>0.058</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Manually-Assisted</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teams</td>
<td>8</td>
<td>5</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>Runs</td>
<td>22</td>
<td>19</td>
<td>16</td>
<td>10</td>
</tr>
<tr>
<td>Min $x\text{InfAP}$</td>
<td>0.005</td>
<td>0.048</td>
<td>0.012</td>
<td>0.033</td>
</tr>
<tr>
<td>Max $x\text{InfAP}$</td>
<td>0.169</td>
<td>0.207</td>
<td>0.106</td>
<td>0.152</td>
</tr>
<tr>
<td>Median $x\text{InfAP}$</td>
<td>0.043</td>
<td>0.111</td>
<td>0.072</td>
<td>0.09</td>
</tr>
</tbody>
</table>
### Easy vs difficult topics overall (2019)

<table>
<thead>
<tr>
<th>Easiness</th>
<th>Hardness</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Top 10 Easy</strong> sorted by count of runs with $\text{InfAP} \geq 0.3$</td>
<td><strong>Top 10 Hard</strong> sorted by count of runs with $\text{InfAP} &lt; 0.3$</td>
</tr>
<tr>
<td>person in front of a graffiti painted on a wall</td>
<td>one or more picnic tables outdoors</td>
</tr>
<tr>
<td>coral reef underwater</td>
<td>inside views of a small airplane flying</td>
</tr>
<tr>
<td>person in front of a curtain indoors</td>
<td>person holding a tool and cutting something</td>
</tr>
<tr>
<td>person wearing shorts outdoors</td>
<td>door being opened by someone</td>
</tr>
<tr>
<td>person wearing a backpack</td>
<td>woman wearing a red dress outside in the daytime</td>
</tr>
<tr>
<td>bald man</td>
<td>a black man singing</td>
</tr>
<tr>
<td>person with a painted face or mask</td>
<td>truck being driven in the daytime</td>
</tr>
<tr>
<td>shirtless man standing up or walking outdoors</td>
<td>man and a woman holding hands</td>
</tr>
<tr>
<td>man and a baby both visible</td>
<td>man and a woman hugging each other</td>
</tr>
<tr>
<td>drone flying</td>
<td>woman riding or holding a bike outdoors</td>
</tr>
</tbody>
</table>
Statistical significant differences among top 10 “F” runs (using randomization test, p < 0.05)

<table>
<thead>
<tr>
<th>Run</th>
<th>Mean Inf. AP score</th>
</tr>
</thead>
<tbody>
<tr>
<td>C_D_ATL.19_2</td>
<td>0.163 #</td>
</tr>
<tr>
<td>C_D_ATL.19_1</td>
<td>0.161 #</td>
</tr>
<tr>
<td>C_D_RUCMM.19_1</td>
<td>0.160 #</td>
</tr>
<tr>
<td>C_D_ATL.19_4</td>
<td>0.157 #</td>
</tr>
<tr>
<td>C_D_RUCMM.19_2</td>
<td>0.152 #</td>
</tr>
<tr>
<td>C_D_RUCMM.19_4</td>
<td>0.127 *</td>
</tr>
<tr>
<td>C_D_RUCMM.19_3</td>
<td>0.124 *</td>
</tr>
<tr>
<td>C_D_WasedaMeiseiSoftbank.19_1</td>
<td>0.123 *</td>
</tr>
<tr>
<td>C_D_Inf.19_3</td>
<td>0.118 *</td>
</tr>
<tr>
<td>C_D_Inf.19_2</td>
<td>0.118 *</td>
</tr>
</tbody>
</table>

#* : no significant difference among each set of runs

- Runs higher in the hierarchy are significantly better than runs more indented.
Statistical significant differences among top 10 "M" runs (using randomization test, p < 0.05)

<table>
<thead>
<tr>
<th>Run</th>
<th>Mean Inf. AP score</th>
</tr>
</thead>
<tbody>
<tr>
<td>C_D_WasedaMeiseiSoftbank.19_2</td>
<td>0.152</td>
</tr>
<tr>
<td>C_D_WasedaMeiseiSoftbank.19_3</td>
<td>0.136 #</td>
</tr>
<tr>
<td>C_D_WasedaMeiseiSoftbank.19_1</td>
<td>0.133 #</td>
</tr>
<tr>
<td>C_D_VIREO.19_2</td>
<td>0.118 #</td>
</tr>
<tr>
<td>C_D_WasedaMeiseiSoftbank.19_4</td>
<td>0.114</td>
</tr>
<tr>
<td>C_D_VIREO.19_1</td>
<td>0.066 *</td>
</tr>
<tr>
<td>C_A_SIRET.19_3</td>
<td>0.035 *</td>
</tr>
<tr>
<td>C_A_SIRET.19_2</td>
<td>0.035 *</td>
</tr>
<tr>
<td>C_A_SIRET.19_1</td>
<td>0.034 !</td>
</tr>
<tr>
<td>C_A_SIRET.19_4</td>
<td>0.033 !</td>
</tr>
</tbody>
</table>

!#* : no significant difference among each set of runs

- Runs higher in the hierarchy are significantly better than runs more indented.
Processing time vs Inf. AP (“F” runs)
Across all topics and runs

Few topics with fast response and high score
Processing time vs Inf. AP ("M" runs) 
Across all topics and runs

Consuming longer time processing queries didn’t help many teams/topics
Samples of (tricky/failed) results

- Truck driven in the daytime
- Drone flying
- Person in a tent
- Person wearing shorts
- Man and a woman holding hands
- Black man singing
- Birds in a tree
- Red hat or a cap

TRECVID 2019
2019 Main approaches

• Two main competing approaches: “concept banks” and “(visual-textual) embedding spaces”

• Currently: significant advantage for “embedding space” approaches, especially for fully automatic search and even overall

• Training data for semantic spaces: MSR and TRECVID VTT tasks, TGIF, IACC.3, Flickr8k, Flickr30k, MS COCO], and Conceptual Captions
2019 Main approaches

- **Alibaba Group** (presentation to follow):
  - Fully automatic (0.163): mapping video embedding and language embedding into a learned semantic space with graph sequence and aggregated modeling, and gated CNNs

- **Renmin University of China and Zhejiang Gongshang University** (presentation to follow):
  - Fully automatic (0.160): Word to Visual Word (W2VV++) similar to TRECvid 2018 plus “dual encoding network” and BERT as text encoder

- **Waseda University; Meisei University; SoftBank Corporation** (presentation to follow):
  - Manually assisted (0.152): concept-based retrieval similar to previous years’ concept bank approach
  - Fully automatic (0.123): visual-semantic embedding (VSE++)
2019 Main approaches

• Shandong Normal University; Carnegie Mellon University; Monash University:
  – Fully automatic (0.118): submitted fully automatic runs but notebook paper currently only about their INS task participation.
• City University of Hong Kong (VIRE0) and Eurecom:
  – Manually assisted runs (0.118): concept based approach with manual query parsing and manual concept filtering
  – Fully automatic (0.075): concept based approach
• Kindai University and Kobe University:
  – Fully automatic (0.087): embedding that maps visual and textual information into a common space
• Florida International University; University of Miami (presentation to follow)
  – Fully automatic (0.082): weighted concept fusion and W2VV
2019 Task observations

- New dataset: Vimeo Creative Commons Collection (V3C1) is being used for testing.
- Development of 90 queries to be used between 2019-2021 including progress subtask.
- Run training types are dominated by “D” runs. No relevance feedback submissions received.
- New “novelty” run type (and metric). Novelty runs proved to submit unique true shots compared to common run types.
- Stable team participation and task completion rate. Manually-assisted runs decreasing.
- High participation in the progress subtask.
- Absolute number of hits are higher than previous years.
- We can’t compare performance with IACC.3 (2016-2018) : New dataset + New queries.
- Fully automatic and Manually-assisted performance are almost similar.
- Among high scoring topics, there is more room for improvement among systems.
- Among low scoring topics, most systems scores are collapsed in small narrow range.
- Dynamic topics (actions, interactions, multi-facets ..etc) are the hardest topics.
- Most systems are slow. Few topics scored high in fast time.
- Task is still challenging!
RUCMM 2019 system on previous years

<table>
<thead>
<tr>
<th>Previous best run</th>
<th>TRECVID edition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2016</td>
</tr>
<tr>
<td><strong>Ours:</strong></td>
<td></td>
</tr>
<tr>
<td>Run 4</td>
<td>0.163</td>
</tr>
<tr>
<td>Run 3</td>
<td>0.161</td>
</tr>
<tr>
<td>Run 2</td>
<td>0.165</td>
</tr>
<tr>
<td>Run 1</td>
<td>0.169</td>
</tr>
<tr>
<td>Dual Encoding*</td>
<td>0.162</td>
</tr>
</tbody>
</table>
Interactive Video Retrieval subtask will be held as part of the Video Browser Showdown (VBS)

At MMM 2020
26th International Conference on Multimedia Modeling, January 5-8, 2020 Daejeon, Korea

- 10 Ad-Hoc Video Search (AVS) topics: Each AVS topic has several/many target shots that should be found.
- 10 Known-Item Search (KIS) tasks, which are selected completely random on site. Each KIS task has only one single 20 s long target segment.
- Registration for the task is now closed
9:10 – 12:20 : Ad-hoc Video Search

9:10 - 9:40 am Ad-hoc Video Search Task Overview

9:40 - 10:10 am Learn to Represent Queries and Videos for Ad-hoc Video Search, RUCMM Team - Renmin University of China; Zhejiang Gongshang University

10:10 - 10:40 am Zero-shot Video Retrieval for Ad-hoc Video Search Task, WasedaMeiseiSoftbank Team – Waseda University; Meisei University; SoftBank Corporation

10:40 - 11:00 am Break with refreshments

11:00 - 11:30 am Query-Based Concept Tree for Score Fusion in Ad-hoc Video Search Task, FIU_UM Team – Florida Intl. University; University of Miami

11:30 - 12:00 pm Hybrid Sequence Encoder for Text Based Video Retrieval, ATL Team – Alibaba Group

12:00 - 12:20 pm AVS Task discussion
2019 Questions and 2020 plans

• Was the task/queries realistic enough?!
• How teams feel the difference between IACC data vs V3C?
• Do we need to change/add/remove anything to the task in 2020?
• Is there any specific reason for the low submissions in “E” & “F” training type runs? (training data collected automatically from the given query text)
• Do we need the relevance feedback run type? 0 submissions this year.
• Did any team run their 2019 system on IACC.3 (2016-2018) topics? (Yes)
• Any feedback about the new novelty metric (runs)?
• Engineering versus research efforts?
• Shared “consolidated” concept banks?
  • How to encourage teams to share resources/concept models,... etc.
• Current plan is to continue the task V3C1 for main and progress subtask.
• Please continue participating in the “progress subtask” to measure accurate performance difference
• What about an explainability subtask (related to embedding approaches)?
## AVS Progress subtask

<table>
<thead>
<tr>
<th>Submission year</th>
<th>Evaluation year</th>
</tr>
</thead>
<tbody>
<tr>
<td>2019</td>
<td>2019</td>
</tr>
<tr>
<td>2019</td>
<td>Submit 50 queries (30 new + 20 common) Eval 30 new Queries</td>
</tr>
<tr>
<td>2019</td>
<td>2020</td>
</tr>
<tr>
<td>2019</td>
<td>2020</td>
</tr>
<tr>
<td>2019</td>
<td>2021</td>
</tr>
<tr>
<td>2019</td>
<td>2020</td>
</tr>
<tr>
<td>2019</td>
<td>2021</td>
</tr>
<tr>
<td>2020</td>
<td>2020</td>
</tr>
<tr>
<td>2020</td>
<td>Submit 40 queries (20 new + 20 common) Eval 30 (20 new + 10 common)</td>
</tr>
<tr>
<td>2020</td>
<td>2021</td>
</tr>
<tr>
<td>2021</td>
<td>2020</td>
</tr>
<tr>
<td>2021</td>
<td>2021</td>
</tr>
<tr>
<td>2021</td>
<td>Submit 40 queries (20 new + 20 common) Eval 30 (20 New + 10 common)</td>
</tr>
</tbody>
</table>

**Goals:**
- Evaluate 10 (set A) common queries submitted in 2 years (2019, 2020)
- Evaluate 10 (set B) common queries submitted in 3 years (2019, 2020, 2021)
- Evaluate 20 common queries submitted in 3 years (2019, 2020, 2021)
- Ground truth for 20 common queries can be released only in 2021