# TRECVID 2016 INSTANCE RETRIEVAL

## INTRODUCTION AND TASK OVERVIEW

Wessel Kraaij
The Netherlands Organisation for
Applied Scientific Research TNO; Leiden University

George Awad
Dakota Consulting ; National Institute of Standards and Technology



National Institute of Standards and Technology

# Table of contents

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

# Task

## From 2013 – 2015

- The task asked systems to find a specific object, person or location in any context using a small set of image and video examples.

## In 2016

- A new query type was used: *find a specific person in a specific location.*

## System task:

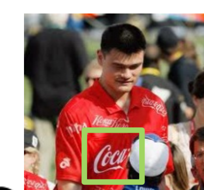
- Given a topic with:
  - 4 example images of the target person
  - 4 Region of Interest (ROI)-masked images of the target person
  - 4 shots from which the target person example images came
  - 6 to12 image and video examples of a known location
- Return a list of up to 1000 shots ranked by likelihood that they contain the topic target person in the target location
- **Automatic** or **interactive** runs are accepted

NIST
National Institute of Standards and Technology

# Background

- The many dimensions of searching and indexing video collections
  - crossing the semantic gap: search task, semantic indexing task
  - visual domain: shot boundary detection, copy detection, INS
  - machine learning vs. high dimensional search given spatio temporal constraints
- Instance search:
  - searching with a visual example (image or video) of a target person/ location/object
  - hypothesis: systems will focus more on the target, less on the visual/ semantic context
  - Investigating region of interest approaches, image segmentation.
- Existing commercial applications using visual similarity
  - logo detection (sports video)
  - product / landmark recognition (images)





# Data …

The British Broadcasting Corporation (BBC) and the Access to Audiovisual Archives (AXES) project made **464 h** of the BBC soap opera EastEnders available for research

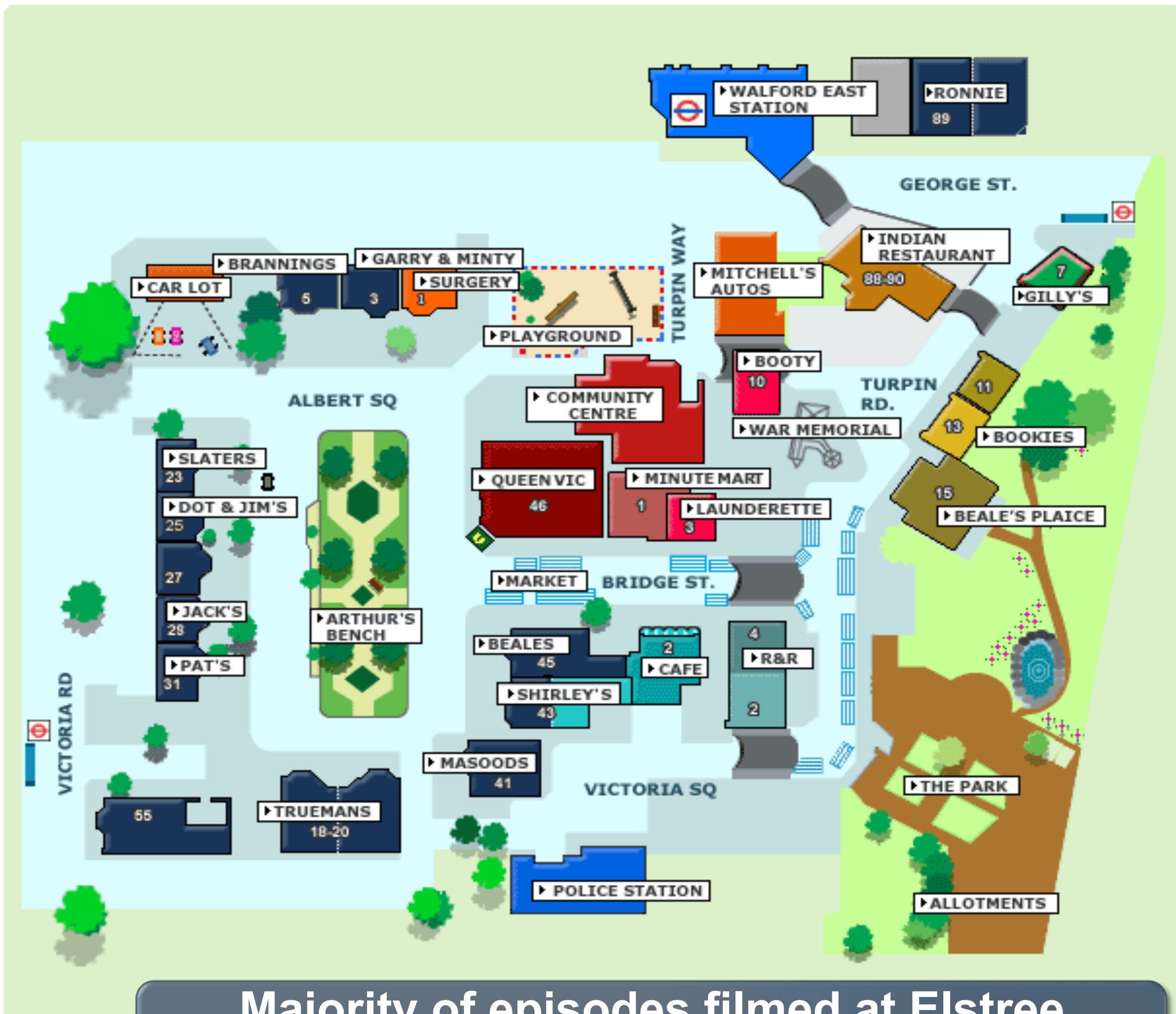
- 244 weekly “omnibus” files (MPEG-4) from 5 years of broadcasts
- 471527 shots
- Average shot length: 3.5 seconds
- Transcripts from BBC
- Per-file metadata

Represents a “small world” with a slowly changing set of:

- People (several dozen)
- Locales: homes, workplaces, pubs, cafes, open-air market, clubs
- Objects: clothes, cars, household goods, personal possessions, pets, etc
- Views: various camera positions, times of year, times of day,

Use of fan community metadata allowed, if documented

# EastEnders’ world




**Majority of episodes filmed at Elstree studios. Sometimes filmed on ‘location’.**

# Topic creation procedure @ NIST

- Viewed several test videos to develop a list of recurring people, locations and their overlapping.
- Chose 10 master locations and identified 6 to 12 image and video examples to each depending on location type (private: kitchen, room, etc; public: pub, café, market, etc)
- Created ≈90 topics targeting recurring specific persons in specific locations.
- Chose representative sample of 30 topics. Each topic includes images for target persons from test videos, many from the sample video (ID 0) and a named location.
- Filtered example shots from the submissions if it satisfies the topic.

# Global test condition: type of training data

Effect of examples – 2 conditions:

- A – one or more provided images – no video
- E - video examples (+ optionally image examples)

# Topics – segmented "person" example images

Brad

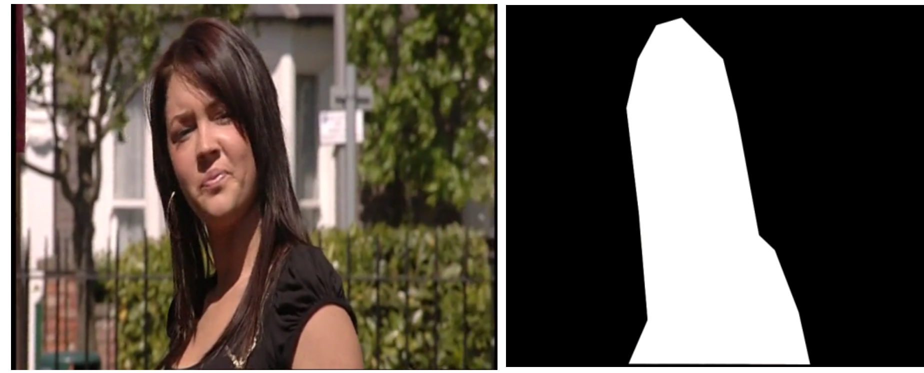
Dot

Fatboy

Jim

# Topics – segmented example images

**Pat**



**Stacey**

**Patrick**

# Topics – 10 Master locations

**Foyer**

**Kitchen1**

**Kitchen2**

**LR1**

**LR2**

**Cafe1**

**Cafe2**

**Laundrette**

**market**

**Pub**

# Topics – 2016

| | Jim | Dot | Brad | Stacey | Pat | Patrick | Fatboy |
|---|---|---|---|---|---|---|---|
| Pub | x | x | x | x | x | x | x |
| Foyer | x | x | x | x | x | | |
| LR1 | x | x | x | x | x | | x |
| Kitchen1 | x | x | x | x | x | x | |
| Laundrette | x | | x | x | x | x | x |

30 x topics : find {jim, Dot, Brad, Stacey, Pat, Patrick, Fatboy} in {Pub,Foyer,LR1,Kitchen1,Laundrette}

NIST National Institute of Standards and Technology

# INS 2016: 13 Finishers (out of 30)

| | |
|---|---|
| U_TK | University of Tokushima |
| UQMG | University of Queensland - DKE Group of ITEE |
| **insightdcu** | **Dublin City University; Polytechnic University of Catalonia** |
| **ITI_CERTH** | **Centre for Research and Technology Hellas** |
| IRIM | EURECOM; LABRI; LIG;LIP6; LISTIC |
| JRS | JOANNEUM RESEARCH |
| **BUPT_MCPRL** | **Beijing University of Posts and Telecommunications** |
| NII_Hitachi_UIT | National Institute of Informatics; Hitachi, Ltd; U. of Inf. Tech. |
| WHU_NERCMS | Wuhan University |
| **PKU-ICST** | **Peking University** |
| SIAT_MMLAB | Shenzhen Institutes of Advanced Technology;Chinese Academy of Sciences |
| TRIMPS_SARI | Third Research Inst. of the Ministry of Public Security; Chinese Academy of Sciences |
| **TUC** | **Technische Universitaet Chemnitz** |

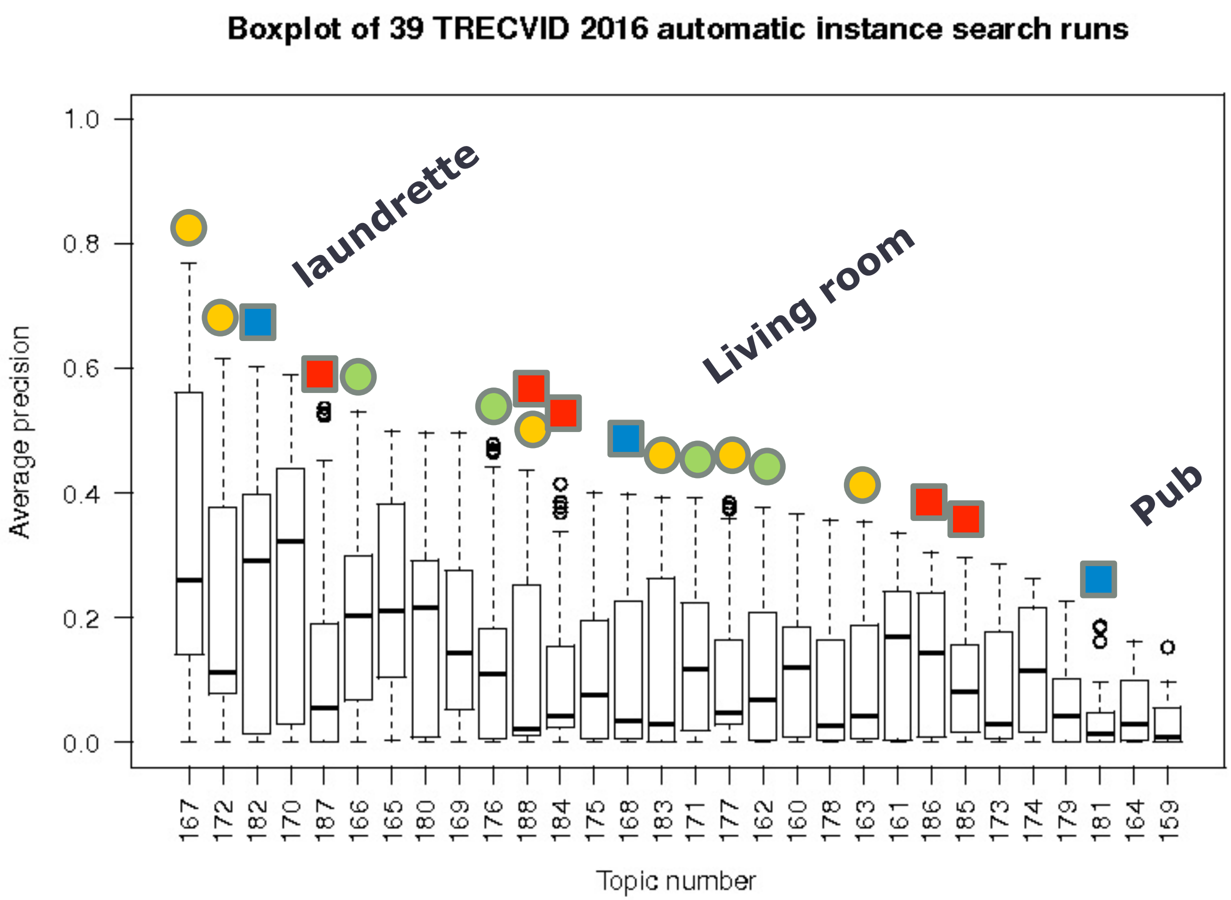
**BLUE indicates team submitted interactive runs**

# Evaluation

For each topic the submissions were pooled and judged down to at least rank 120 (on average to rank 288, max 520), resulting in 136744 judged shots (≈ 600 person-h).

- 10 NIST assessors played the clips and determined if they contained the topic target or not.
- 13800 clips (avg. 460 / topic) contained the topic target (10 %)
- True positives per topic: min 13 med 276 max 1614
- The task is treated as a form of search and thus the trec_eval_video tool was used to calculate average precision, recall, precision, etc.
- To measure efficiency, speed was also measured.

# Results by topic - automatic




| # | Query |
|---|---|
| 167 | Find **Dot** in this **Living Room** |
| 172 | Find **Brad** in this **Living room** |
| 182 | Find Fatboy in this Laundrette |
| 170 | Find **Brad** in this Laundrette |
| 187 | Find **Pat** at this **Foyer** |
| 166 | Find **Dot** at this **Foyer** |
| 165 | Find **Dot** in this Kitchen |
| 180 | Find Patrick in this Laundrette |
| 169 | Find **Brad** in this Kitchen |
| 176 | Find Stacey at this **Foyer** |
| 188 | Find **Pat** in this **Living Room** |
| 184 | Find **Pat** in this Pub |
| 175 | Find Stacey in this Laundrette |
| 168 | Find Brad in this Pub |
| 183 | Find Fatboy in this **Living room** |
| 171 | Find Brad at this **Foyer** |
| 177 | Find Stacey in this **Living room** |
| 162 | Find Jim at this **Foyer** |
| 160 | Find **Jim** in this Kitchen |
| 178 | Find Patrick in this Pub |
| 163 | Find **Jim** in this **Living Room** |
| 161 | Find **Jim** in this Laundrette |
| 186 | Find **Pat** in this Laundrette |
| 185 | Find **Pat** in this Kitchen |
| 173 | Find Stacey in this Pub |
| 174 | Find Stacey in this Kitchen |
| 179 | Find Patrick in this Kitchen |
| 181 | Find Fatboy in this Pub |
| 164 | Find **Dot** in this Pub |
| 159 | Find **Jim** in this Pub |

**What is the effect of person vs location on the performance ?**

# Automatic Run results + Randomization testing

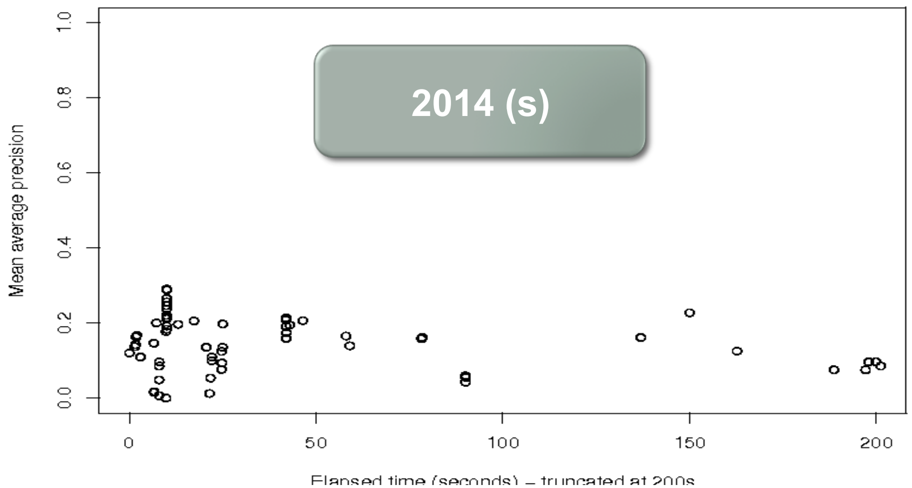
## Top 10 runs across all teams (automatic)

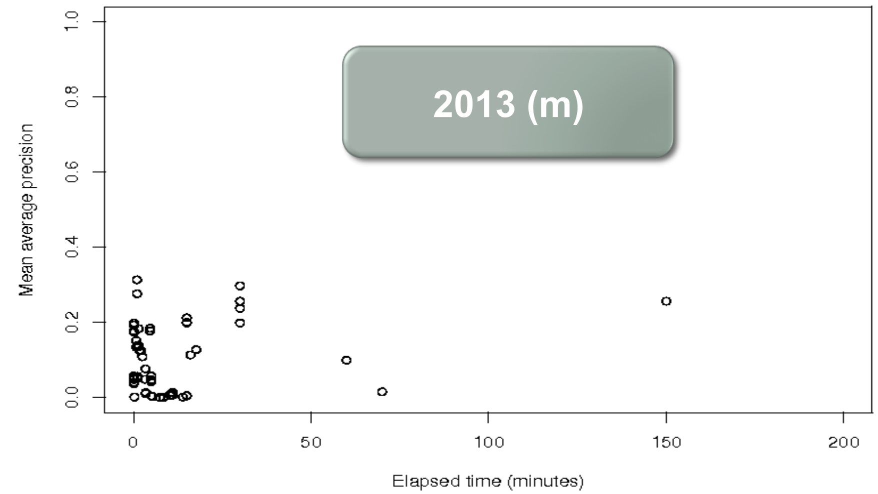
| MAP | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.370 | F_E_PKU_ICST_1 | = | > | > | > | > | > | > | > | > | > |
| 0.364 | F_E_PKU_ICST_3 | | = | | > | > | > | > | > | > | > |
| 0.349 | F_E_PKU_ICST_5 | | | = | | > | > | > | > | > | > |
| 0.335 | F_A_PKU_ICST_4 | | | | = | > | > | > | > | > | > |
| 0.328 | F_A_PKU_ICST_6 | | | | | = | | > | > | > | > |
| 0.317 | F_A_PKU_ICST_7 | | | | | | = | > | > | > | > |
| 0.244 | F_A_NII_Hitachi_UIT_1 | | | | | | | = | | | > |
| 0.230 | F_A_NII_Hitachi_UIT_4 | | | | | | | | = | | |
| 0.230 | F_A_BUPT_MCPRL_3 | | | | | | | | | = | |
| 0.229 | F_A_NII_Hitachi_UIT_2 | | | | | | | | | | = |

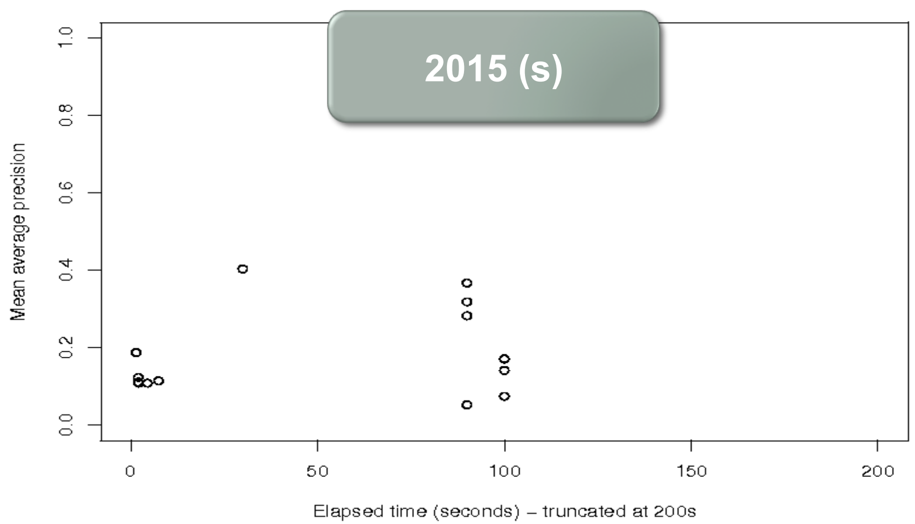
**p = probability the row run scored better than the column run due to chance**

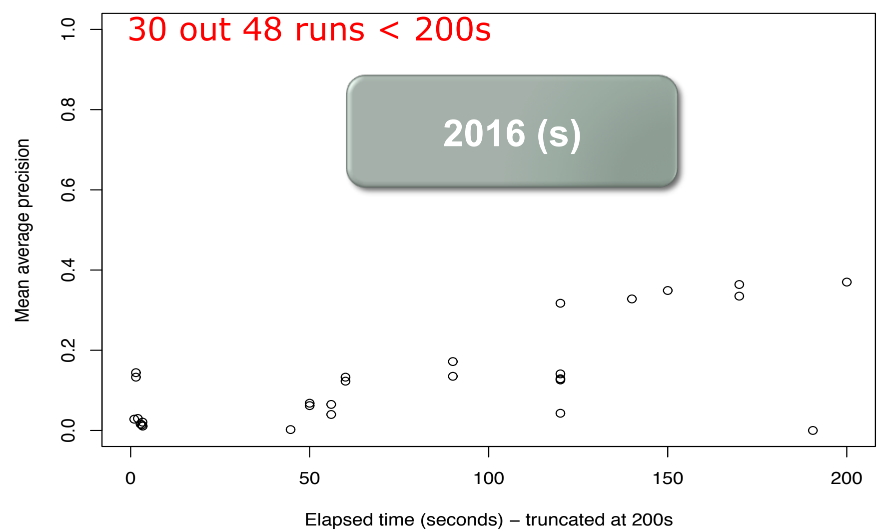
> $p < 0.05$

# Mean Average Precision (MAP) vs. per query clock processing time (automatic)



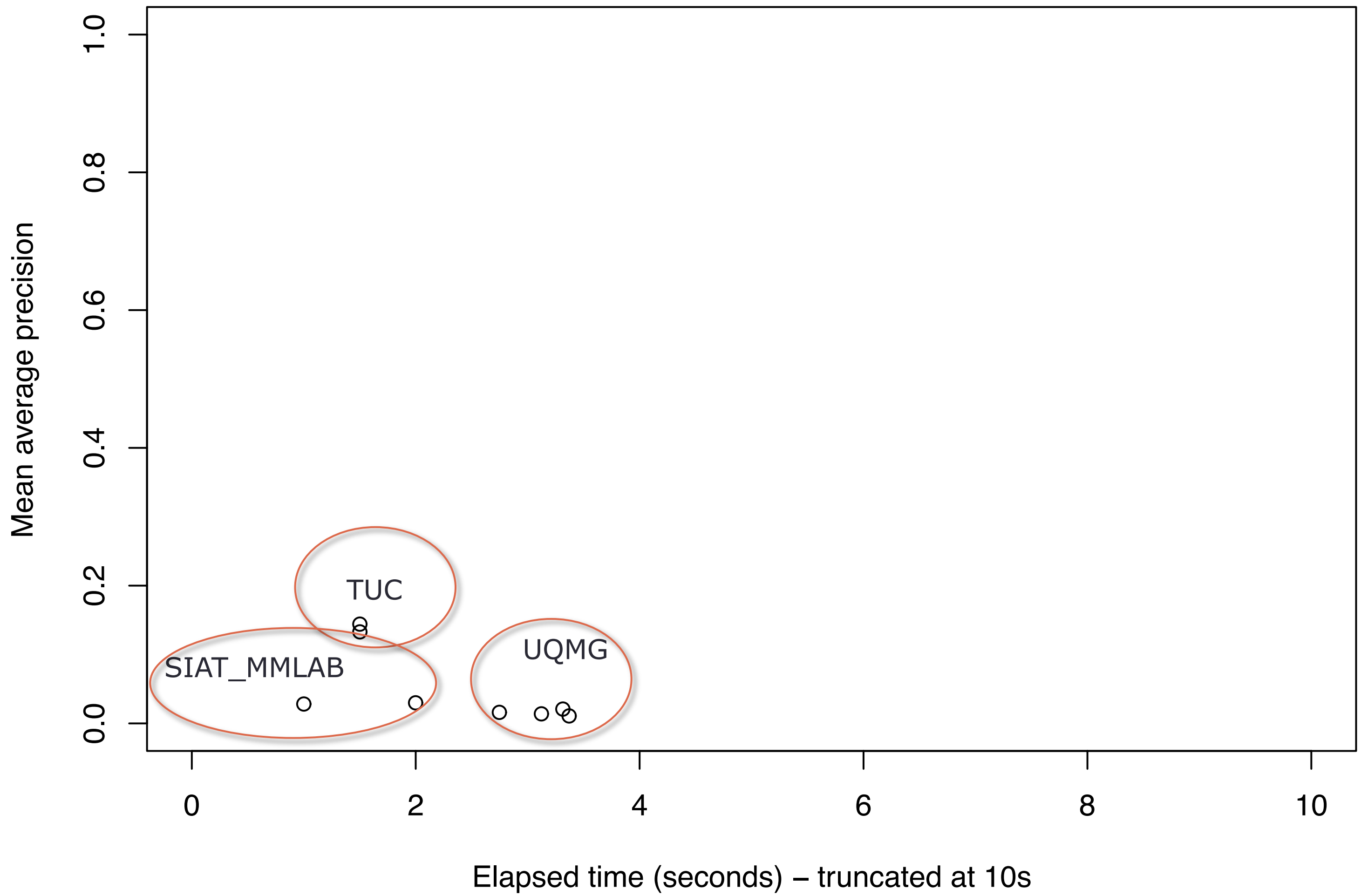



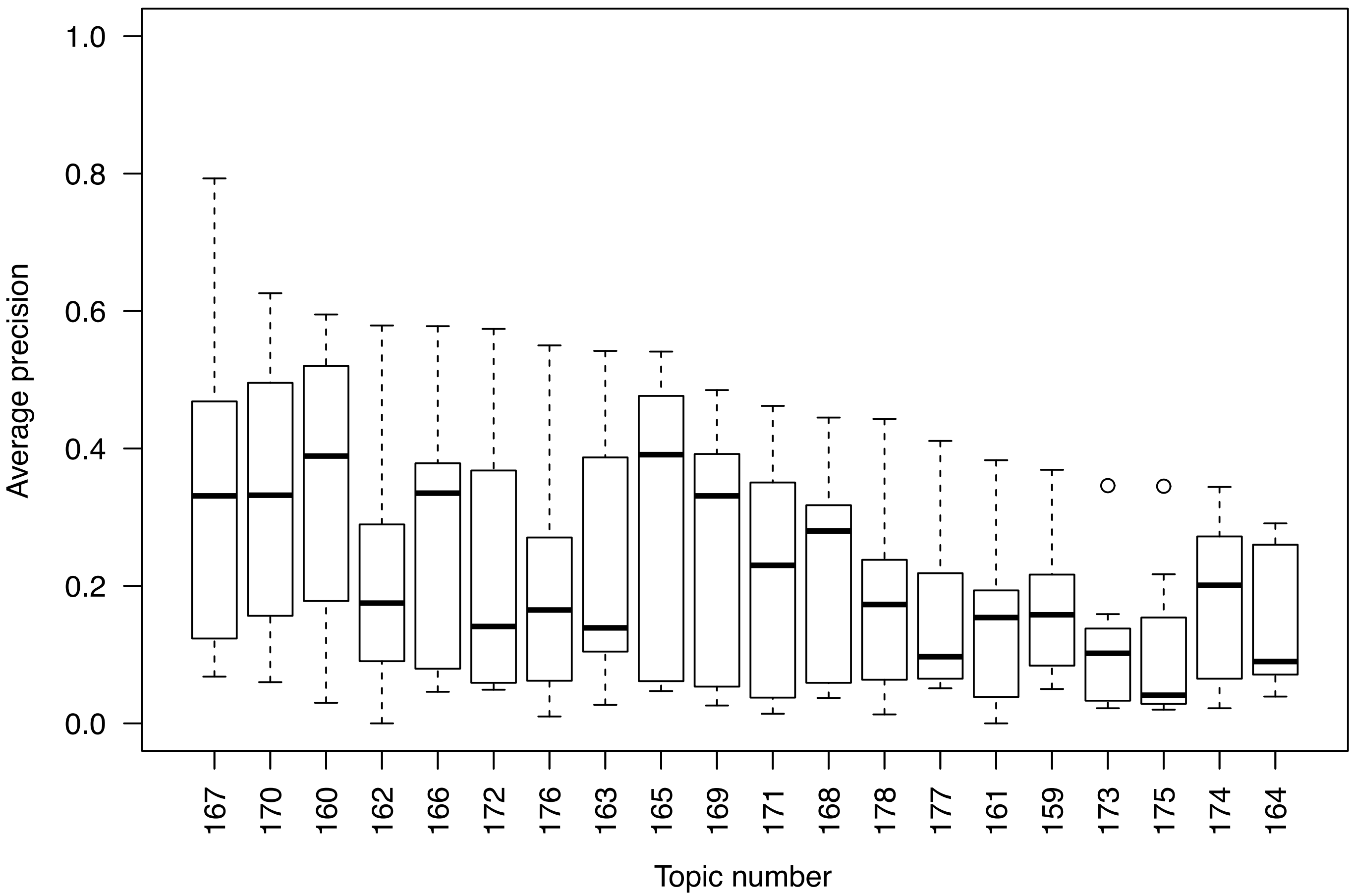

National Institute of Standards and Technology

# MAP vs. fastest query processing time
## (<=10 s, automatic)



# Results by topic - interactive

**Boxplot of 7 TRECVID 2016 interactive instance search runs**




| # | Query |
|---|---|
| 167 | Find **Dot** in this Living Room |
| 170 | Find Brad in this Laundrette |
| 160 | Find Jim in this Kitchen |
| 162 | Find Jim at this **Foyer** |
| 166 | Find **Dot** at this **Foyer** |
| 172 | Find Brad in this Living room |
| 176 | Find **Stacey** at this **Foyer** |
| 163 | Find Jim in this Living Room |
| 165 | Find **Dot** in this Kitchen |
| 169 | Find Brad in this Kitchen |
| 171 | Find Brad at this **Foyer** |
| 168 | Find Brad in this **Pub** |
| 178 | Find Patrick in this **Pub** |
| 177 | Find **Stacey** in this Living room |
| 161 | Find Jim in this Laundrette |
| 159 | Find Jim in this **Pub** |
| 173 | Find **Stacey** in this **Pub** |
| 175 | Find **Stacey** in this Laundrette |
| 174 | Find **Stacey** in this Kitchen |
| 164 | Find **Dot** in this **Pub** |

# Interactive Run Results, Randomization testing

## Top 10 runs across all teams (interactive)

| MAP | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 0.484 | I_E_PKU_ICST_2 | = | > | > | > | > | > | > |
| 0.318 | I_A_TUC_1 | | = | | > | > | > | > |
| 0.285 | I_A_BUPT_MCPRL_4 | | | = | > | > | > | > |
| 0.224 | I_A_TUC_2 | | | | = | > | > | > |
| 0.114 | I_A_ITI_CERTH_1 | | | | | = | > | > |
| 0.059 | I_A_insightdcu_3 | | | | | | = | > |
| 0.036 | I_E_insightdcu_1 | | | | | | | = |
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 |

**p = probability the row run scored better than the column run due to chance**

**> $p < 0.05$**

# Automatic vs interactive topics
## (ranked by max performance on the topic)

### Automatic

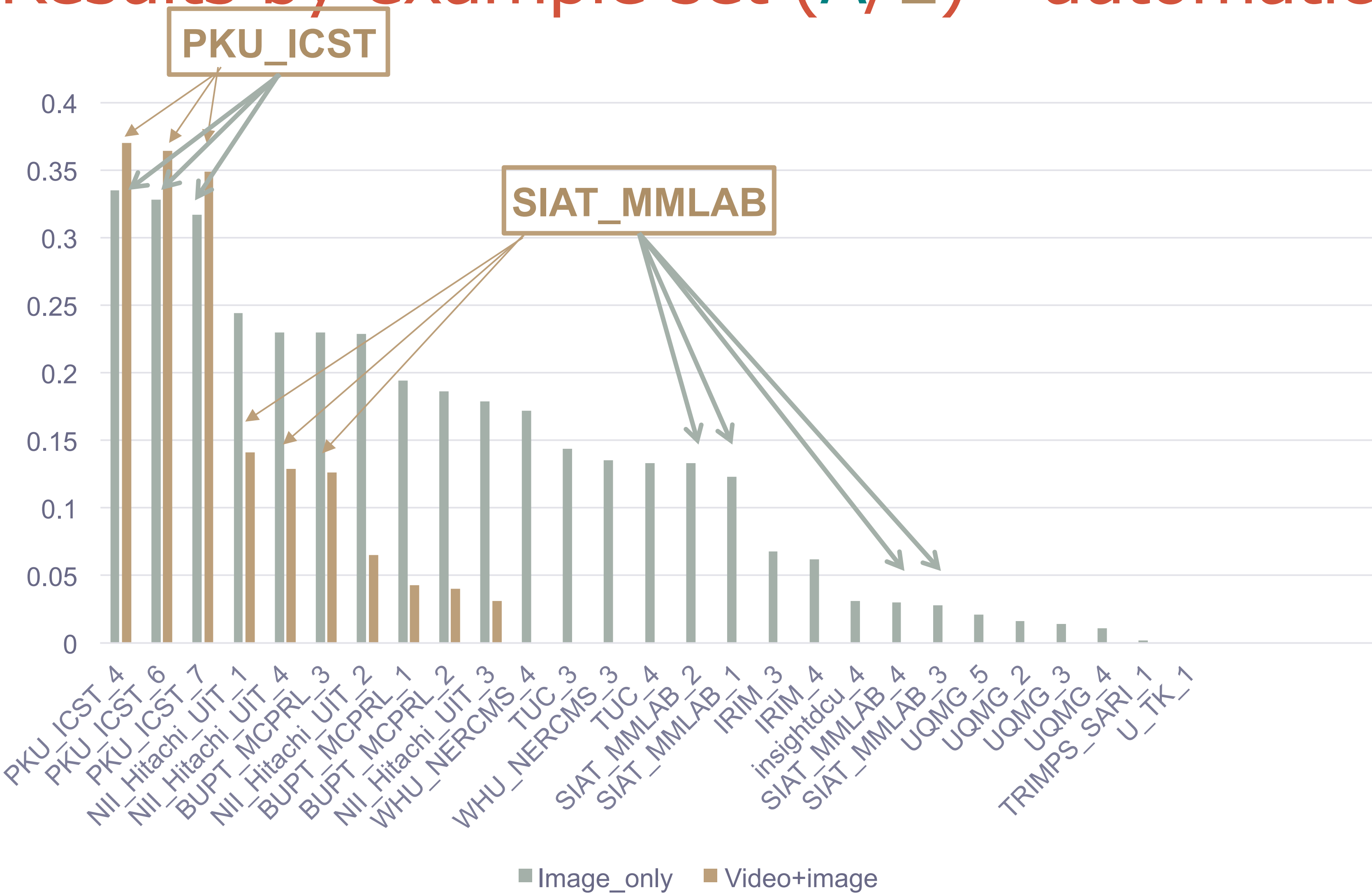
~~186 Find Pat in this Laundrette~~
~~185 Find Pat in this Kitchen~~
160 Find Jim in this Kitchen
~~187 Find Pat at this Foyer~~
177 Find Stacey in this **Living room**
163 Find Jim in this **Living Room**
175 Find Stacey in this Laundrette
167 Find Dot in this **Living Room**
~~188 Find Pat in this Living Room~~
170 Find Brad in this Laundrette
159 Find Jim in this Pub
~~180 Find Patrick in this Laundrette~~
165 Find Dot in this Kitchen
~~179 Find Patrick in this Kitchen~~
~~184 Find Pat in this Pub~~
~~183 Find Fatboy in this Living room~~
172 Find Brad in this **Living room**
~~181 Find Fatboy in this Pub~~
174 Find Stacey in this Kitchen
169 Find Brad in this Kitchen
164 Find Dot in this Pub
171 Find Brad at this **Foyer**
~~182 Find Fatboy in this Laundrette~~
168 Find Brad in this Pub
166 Find Dot at this **Foyer**
178 Find Patrick in this Pub
173 Find Stacey in this Pub
176 Find Stacey at this **Foyer**
162 Find Jim at this **Foyer**
161 Find Jim in this Laundrette

### Interactive

167 Find **Dot** in this Living Room
170 Find Brad in this Laundrette
160 Find Jim in this Kitchen
162 Find Jim at this **Foyer**
166 Find **Dot** at this **Foyer**
172 Find Brad in this Living room
176 Find **Stacey** at this **Foyer**
163 Find Jim in this Living Room
165 Find **Dot** in this Kitchen
169 Find Brad in this Kitchen
171 Find Brad at this **Foyer**
168 Find Brad in this **Pub**
178 Find Patrick in this **Pub**
177 Find **Stacey** in this Living room
161 Find Jim in this Laundrette
159 Find Jim in this **Pub**
173 Find **Stacey** in this **Pub**
175 Find **Stacey** in this Laundrette
174 Find **Stacey** in this Kitchen
164 Find **Dot** in this **Pub**

**Some hard topics were boosted by the interactive users.**

# Results by example set (A/E) - automatic



# WHU-NERCMS team runs

- MAP Results:
  - F NO NERCMS 1 0.758
  - F NO NERCMS 2 0.632
  - F NO NERCMS 3 0.135
  - F NO NERCMS 4 0.172

- Officially evaluated by NIST
- Do not fit the pre-specified ‘automatic’ or ‘interactive’ task categories.
- Talk follows.

# INS 2016: 13 Finishers (out of 30)

No papers

| | |
|---|---|
| U_TK | University of Tokushima |
| UQMG | University of Queensland - DKE Group of ITEE |
| **PKU-ICST** | **Peking University** |
| TRIMPS_SARI | Third Research Inst. of the Ministry of Public Security; Chinese Academy of Sciences |

**BLUE indicates team submitted interactive runs**

# Some general observations about the task

- New task on the Eastenders dataset:
  - Increase in number of participants and stable #of finishers
  - BBC does not permit giving out data to new teams.
  - … spawned some really interesting new architectures
- Task guidelines should become more clear about what is allowed for task categories
  - Add categories for additional data which is used?
  - Add manual query type?
- E condition shows that tracking characters pays off
- Interactive search task:
  - Limited participation, just a few teams perform relevance feedback, mostly cleaning up result lists

# Some general observations about the task

- First year: no development data.
- Specific methods for faces do help significantly
  - Mostly CNN based
- Detecting / Learning location is difficult since they are occluded by people.
  - Best location strategies combine CNN and BOVW using traditional SIFT features
- More and more work on scene threading (linking related shots).

# Overview of submissions (1)

- 10 out of 13 teams described INS runs for the TV notebook
- 4 teams will present their INS experiments

**2:00 - 2:20**, WHU_NERCMS (Wuhan University - Natl. Eng. Research Center for Multimedia Software)
**2:20 - 2:40**, NII_HITACHI-UIT (National Institute of Informatics; Hitachi; U. of Inf. Tech.)
**2:40 - 3:00**, BUPT_MCPRL (Beijing University of Posts and Telecommunications)

**3:00 - 3:20**, **Break** with refreshments

**3:20 - 3:40**, TUC (TU Chemnitz - Junior Professorship Media Computing - Chair Media Informatics)
**3:40 - 4:00**, INS Discussion

# Overview of submissions (2)

- Almost all systems have dedicated pipelines for persons and locations
  - Person recognition relies mostly on CNN models
  - Location often based on traditional BOVW accompanied with CNN features
- Ranking is based on fusion (several experiments) followed by postfiltering strategies
- Exploring external data such as closed captions, fan resources for additional evidence,

# Finding an optimal representation

- **BUPT:**
  - **Locations:** tv15 system (SIFT, VCG19)
  - **Persons:** DLIB detection, VCG-Face, Openface Use CNN for both local and global features + 3 local features
- **InsightDCU:**
  - **Locations:** VCG-places-205
  - **Persons:** VCG16-faces
- **PKU-ICST:**
  - **Locations:** fuse CNN, SIFT BOW
  - **Persons:** VCG-face, relevance feedback, person-re-identification (tracking on clothing), ASR search
- **IRIM**
  - **Locations:** LIMSI SIFT (cleaning up scene) CNN places205
  - **Persons:** face tracking, openface embedding

# Finding an optimal representation (2)

- **ITI-CERTH:**
  - **Locations:** Convolutional neural networks (CNN) imagenet
  - **Persons:** CNN based face detector (Sun et al.)
- **JRS:**
  - MPEG compact video descriptors (no person specific pipeline)
- **NII-Hitachi**
  - **Locations: BOVW,** remove human regions, top K - reranking
  - **Persons:** VCG Face
- **TU_CHEMNITZ**:
  - **Locations:** CNN based  (annotated first episode)
  - **Persons:** CNN based (trained on first episode)

# Finding an optimal representation (3)

- **SIAT:**
  - **Locations:** SIFT based
  - **Persons:** VCG faces, CNN based person re-identification (bounding boxes for persons)
- **WHU:**
  - **Locations:** BOVW + CNN features || Strategy is to manually choose particular objects in a location to serve as 'clean' discriminating query objects
  - **Persons:** Scale-Adaptive Deconvolutional Regression (SADR) Network, VCG 16 for features , speaker identification and captions

# Dealing with query images

- How to exploit the mask (focus vs background)
  - **JRS:** blurring area outside the mask
  - **Wuhan:** manual selection of ROI on different query images: helped significantly for locations
  - **InsightDCU:** only face part of masked ROI is used
- Combining sample images
  - Usually late fusion
  - **PKU:** transformations on samples (for CNN)
  - **WUHAN:** extra images from the web for characters and locations
- Exploiting the full query video clip (for query expansion)
  - Successfully applied by **IRIM, PKU, SIAT**
  - Full clips are also mined for interactive runs (Chemnitz)

# Matching & Ranking

- Typically: fusing or intersecting location and character search results

- Experiments with similarity function:
  - **BUPT** Query adaptive late fusion (like 2015)
  - **Wuhan:** Asymmetrical query adaptive matching
  - **SIAT,WHU:** Hamming embedding
  - **TUC:** linear weighted fusion 2/3 person 1/3 location

- Pseudo relevance feedback, query expansion:
  - BUPT, INSIGHT
  - PKU: Semi supervised learning for discarding noisy videos (linear algebra method on similarity matrix)

# Postprocessing the ranked list (1)

- **IRIM:**
  - Credits filtering / remove ads and opening / end credits
  - Shot threads clustering
- **NII-HITACHI:**
  - Geometric verifcation
  - CNN filtering
- **Wuhan university:**
  - Extensive filtering step:
  - Outdoor (vehicle, hippopotamus, indian elephant, castle) QUESTION: did the team look at the test data to construct the filter??
  - Remove shots without target persons
  - Groundtruth shots of previous years orthogonal topics can be omitted

NIST National Institute of Standards and Technology

# Postprocessing the ranked list (2)

- **SIAT:**
  - spatial verification for locations
- **TU Chemnitz:**
  - Improved version of semantic sequence clustering, effect of semantic sequence clustering is mixed (like pseudo relevance feedback, the technique really depends on the precision at 10 for the initial run. If the precision is low, the MAP will decrease because of drift.

# Interactive experiments

- **TU_CHEMNITZ:** 2 runs; system with sequence clustering <u>increases</u> on baseline interactive
- **BUPT:** 1 run (significantly better than automatic)
- **INSIGHTDCU:** 2 runs, designed to clean up 2 alternative automatic runs
- **ITI_CERTH:** 1 interactive run (no automatic), their first system with CNN
- **PKU_ICST:** 1 run Label max 10 positive examples, use as additional query images, Discard negative examples; big increase in MAP

# INS 2017 discussion

- What do people think of the new task?
- Do we need additional categories (e.g., manual)?
- Do we need additional run categories (for type of external training data)?
- Which data can be used as development data for next year (e.g., for characters and locations)?
- How do we keep the ‘ad hoc’ element in the task in a closed world? Should we move to new data in the future?

# INS 2017 plans

✓Continue with same test data and new set of 30 topics

✓Continue the same topics type: location + person

- Use same training video for a small set of named locations
- Topics will contain
  - reference by name to one of known locations
  - ad hoc person target with 4 image examples and source video shots
- Task: search for shots containing the target person in the target location