TRECVID 2015 INSTANCE RETRIEVAL

INTRODUCTION AND TASK OVERVIEW

Wessel Kraaij
TNO; Radboud University Nijmegen

TNO innovation for life

Paul Over
NIST

George Awad
Dakota Consulting ; NIST

NIST
National Institute of Standards and Technology

Task

Example use case: *browsing a video archive, you find a video of a person, place, or thing of interest to you, known or unknown, and want to find more video containing the same target, but not necessarily in the same context.*

System task:

- Given a topic with:
  - 4 example images of the target
  - 4 ROI-masked images
  - 4 shots from which the example images came
  - a target type (OBJECT/LOGO, PERSON, LOCATION)
  - Attribute Multi <Yes/No> : single vs multiple instances (‘the’ vs ‘a’)
  - <topic title>
- Return a list of up to 1000 shots ranked by likelihood that they contain the topic target
- **Automatic** or **interactive** runs are accepted

Data …

The BBC and the AXES project made **464 hours** 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

Topic creation procedure @ NIST

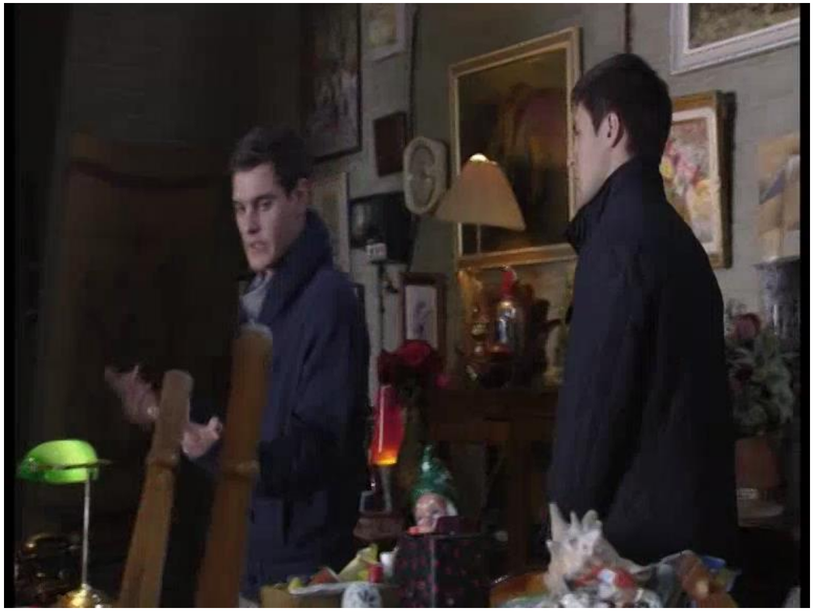
- Viewed every tenth video
- Created ~90 topics targeting recurring specific objects or persons
  - Emphasized objects over people
  - People: mixture of unnamed extras, named characters
  - Objects: most clearly bounded, various sizes, most rigid, some mobile (e.g. varying contexts)
  - All: various camera angles/distances, some variation in lighting
- Chose representative sample of 30 topics, then example images from test videos, many from the sample video (ID 0)
- Filtered example shots from the submissions

Global test condition: type of training data

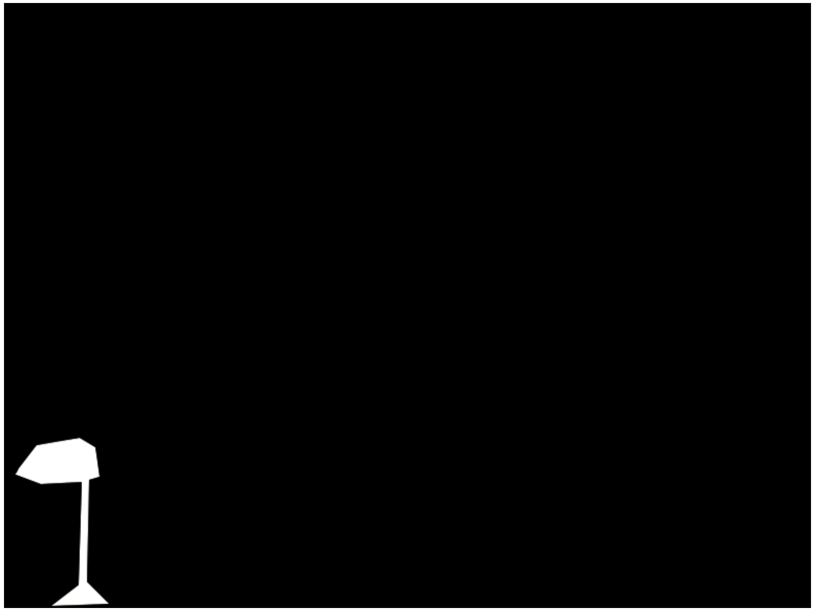
Effect of examples – 2 conditions:

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

Topics – segmented example images



Source



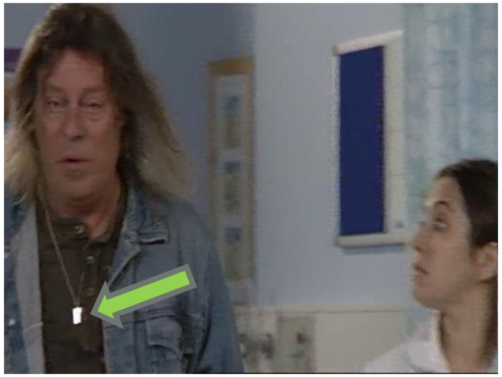
Region of interest mask

“this brass piano lamp with green shade”

Topics – 26 Objects

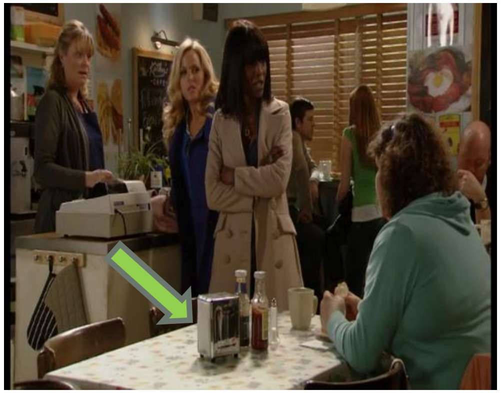
Topic: **True positives:**

129 265



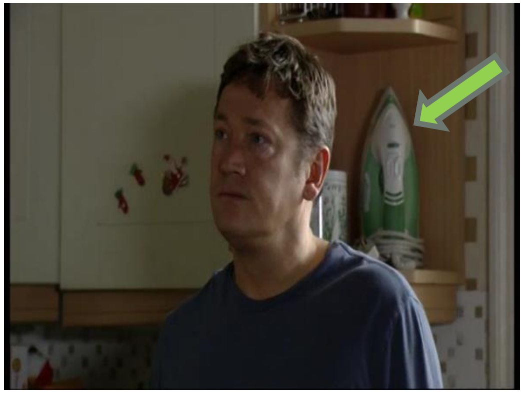
this silver necklace …

130 1735



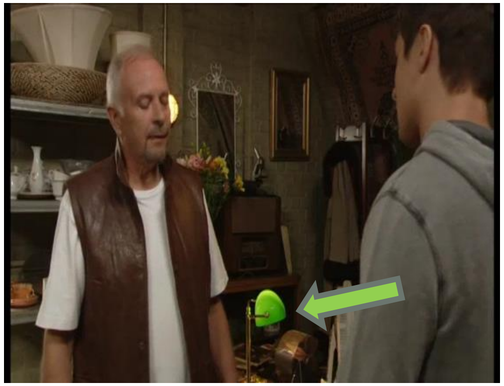
a chrome napkin holder

131 402



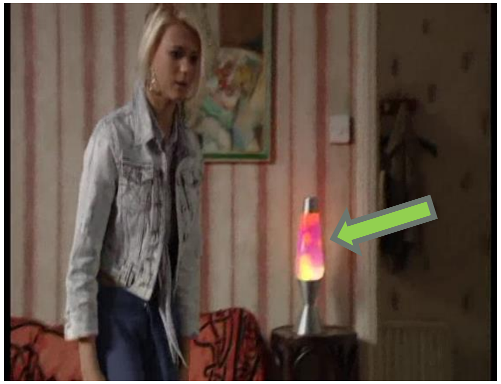
a green and white iron

132 68



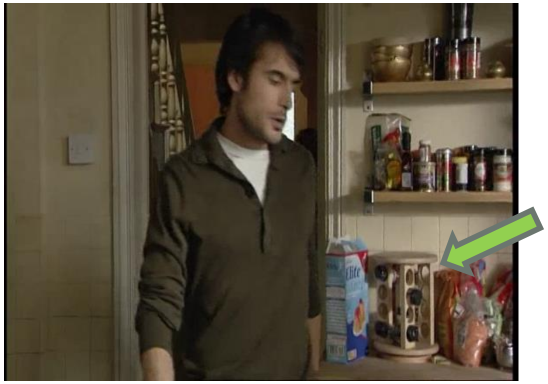
this brass piano lamp

133 112



this lava lamp

134 472

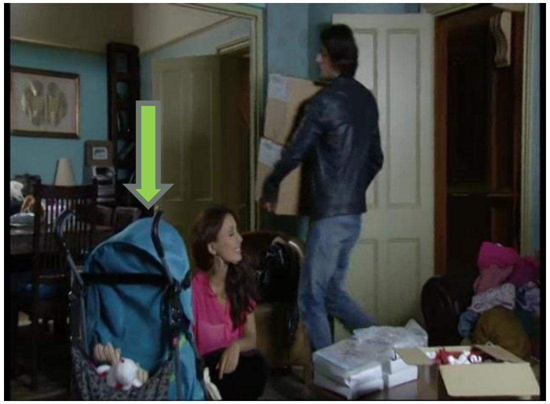


this cylindrical spice rack

Topics – 26 Objects (cont.)

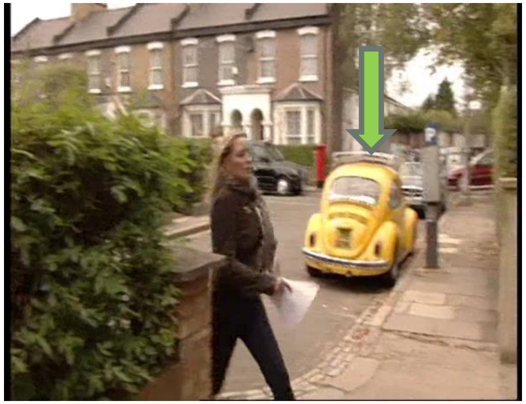
Topic: **True positives:**

135 60



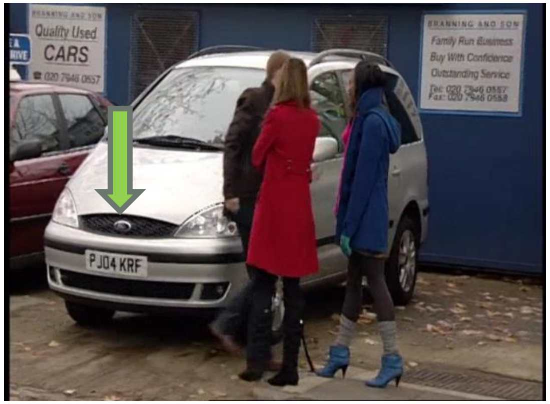
this turquoise stroller

136 83



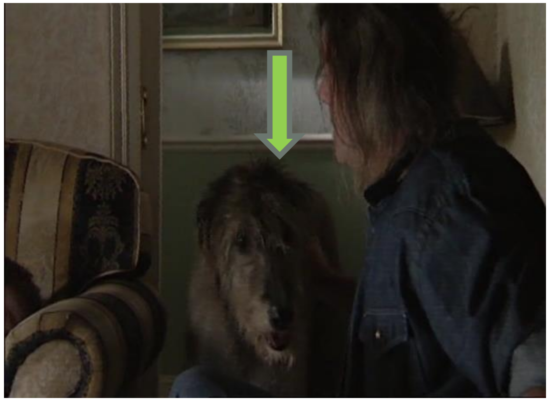
this yellow VW beetle

137 134



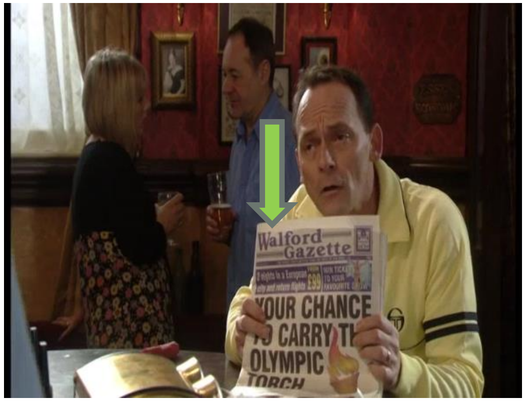
a Ford script logo

139 33



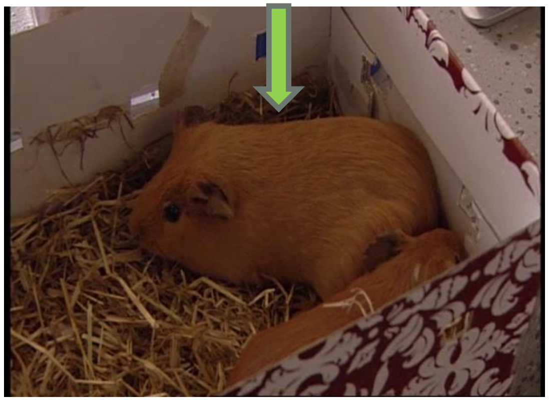
this shaggy dog

140 95



a Walford Gazette banner

141 52

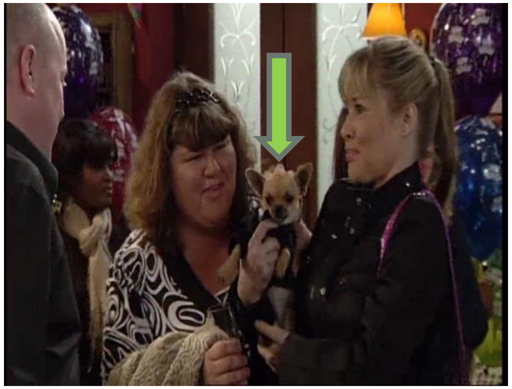


this guinea pig

Topics – 26 Objects (cont.)

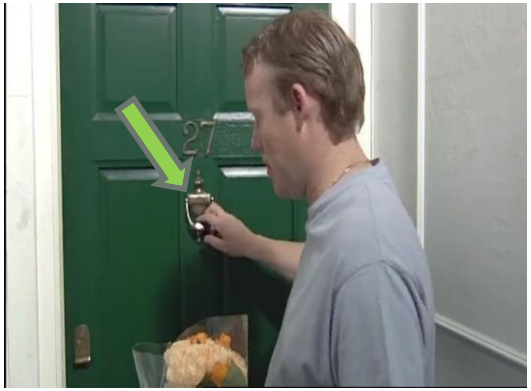
Topic: **True positives:**

142 44



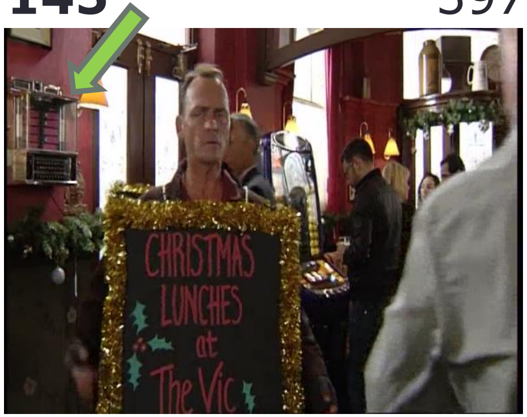
this chihuahua (Prince)

144 256



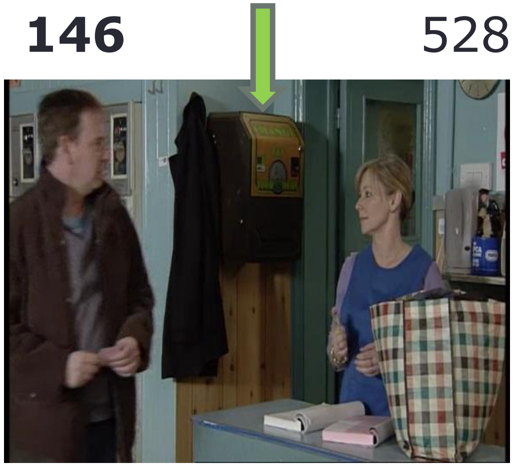
this doorknocker on #27

145 397



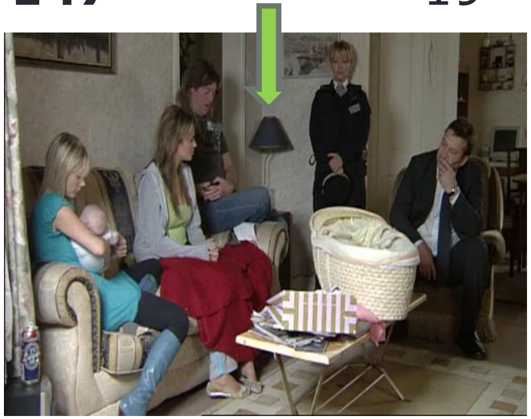
this jukebox wall unit

146 528



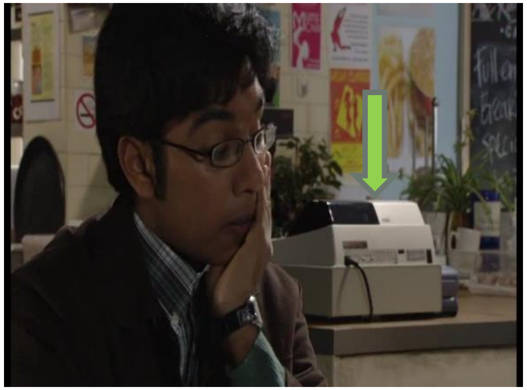
this change machine

147 19



this table lamp

148 1308

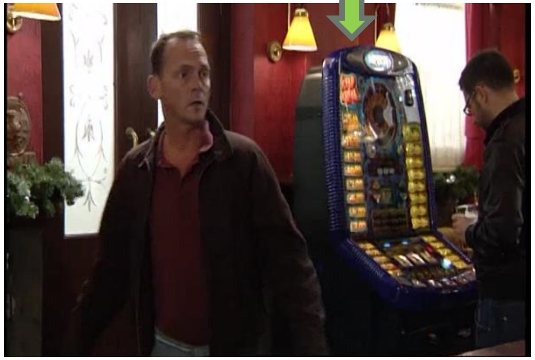


this cash register

Topics – 26 Objects (cont.)

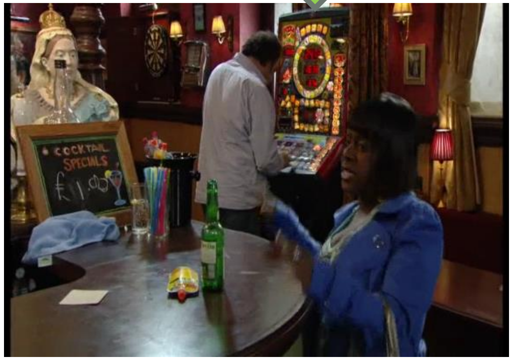
Topic: **True positives:**

150 1103



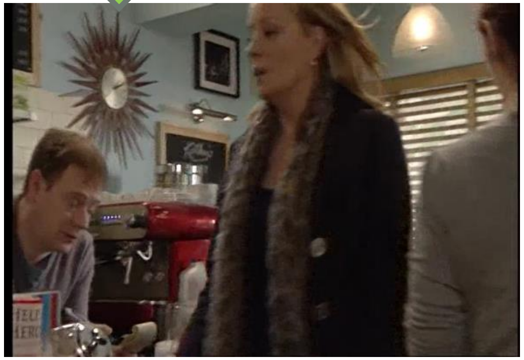
this IMPULSE game

152 638



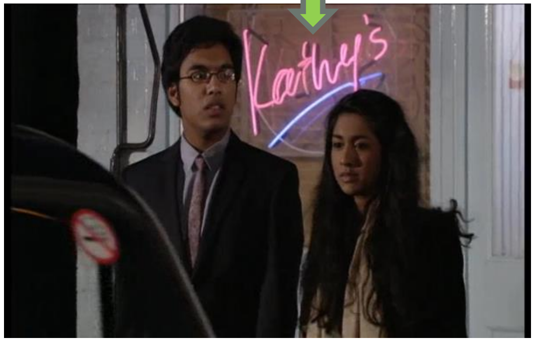
this PIZZA game

153 874



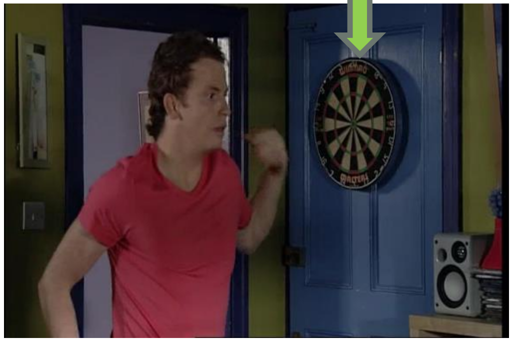
this starburst wall clock

154 747



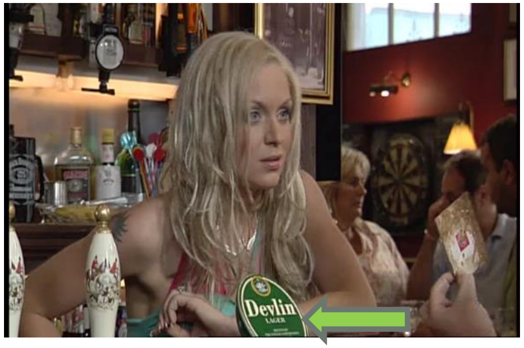
this neon Kathy's sign

155 127



this dart board

156 661

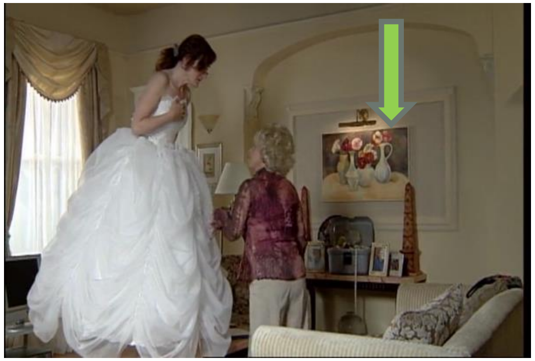


a 'DEVLIN' lager logo

Topics – 26 Objects (cont.)

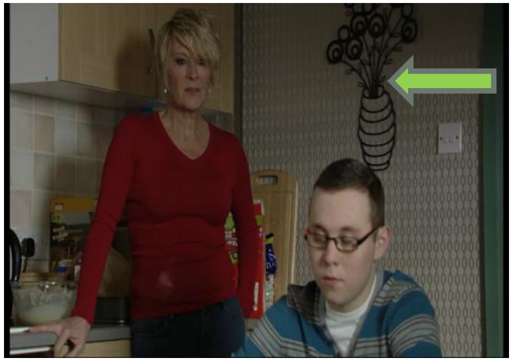
Topic: **True positives:**

157 682



this picture of flowers

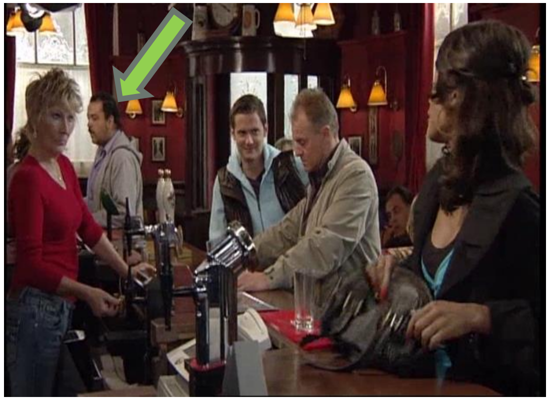
158 437



this flat wire vase with flowers

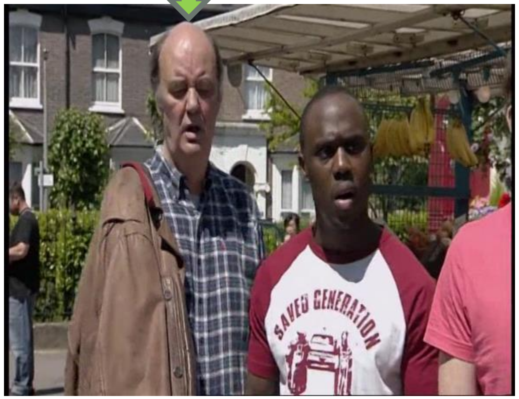
Topics – 2 Persons

138 448



this man with moustache

143 105

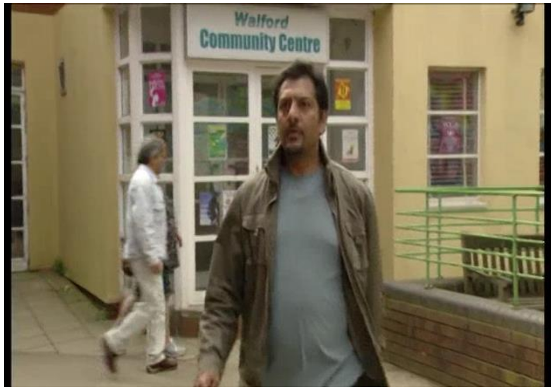


this bald man

this man

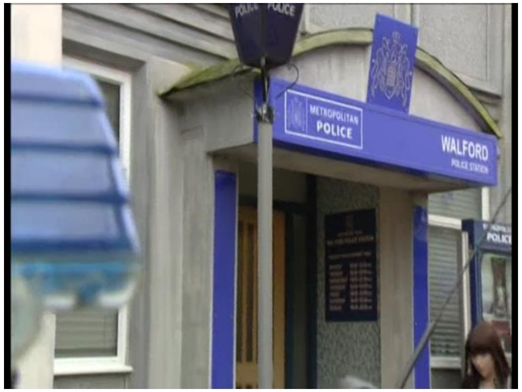
Topics – 2 Locations

149 286



this Walford Community Center entrance from street

151 94



this Walford Police Station entrance from street

INS 2015: 14 Finishers (2014:23, 2013:22, 2012:24)

BUPT_MCPRL	**Beijing University of Posts and Telecommunications**
ITI_CERTH	**Centre for Research and Technology Hellas**
insightdcu	**Dublin City University; University Polytechnica Barcelona**
NII_Hitachi_UIT	National Institute of Informatics; Hitachi, Ltd; U. of Inf. Tech.
NTT	NTT Communication Science Laboratories
ORAND	ORAND S.A. Chile
PKU-ICST	**Peking University ICST**
TUC	**Technische Universitaet Chemnitz**
Trimps	Third Research Institute of the Ministry of Public Security,China
Tsinghua_IMMG	Tsinghua University
Sheffield_UETLahore	University of Sheffield, Lahore U. of Engineering and Technology
UQMG	University of Queensland - DKE Group of ITEE
U_TK	University of Tokushima
NERCMS	Wuhan University

BLUE indicates team submitted interactive runs

Evaluation

For each topic the submissions were pooled and judged down to at least rank 100 (on average to rank 350, max 460), resulting in 205527 judged shots (~ 600 person-hrs).

10 NIST assessors played the clips and determined if they contained the topic target or not.

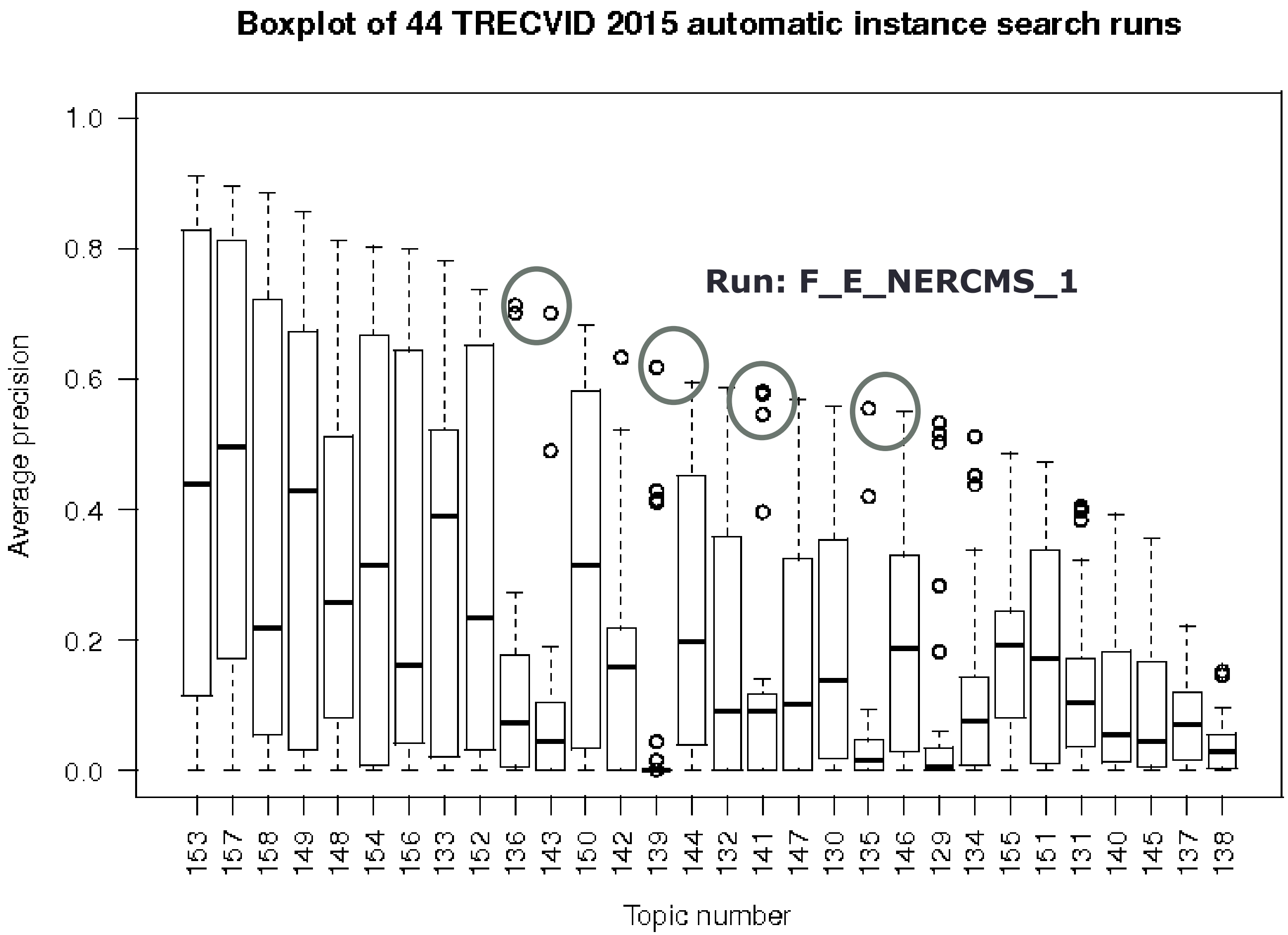
12265 clips (avg. 408.8 / topic) contained the topic target (6%)

True positives per topic: min 19 med 275.5 max 1735

Table lamp Napkin holder

trec_eval_video was used to calculate average precision, recall, precision, etc.

Results by topic - automatic



Targets with single location in BLUE

#	Text
153	this starburst wall clock
157	this picture of flowers
158	this flat wire vase with flowers
*149	this Walford Community Cntr…
148	this cash register
154	this neon Kathy's sign
156	a 'DEVLIN' lager logo
133	this lava lamp
152	this PIZZA game
136	this yellow VW beetle…
+143	this bald man
150	this IMPULSE game
142	this Chihuahua dog
139	this shaggy dog
144	this doorknocker on #27
132	this brass piano lamp…
141	this guinea pig
147	this table lamp…
130	a chrome napkin holder
135	this turquoise stroller
146	this change machine
129	this silver necklace
134	this cylindrical spice rack
155	this dart board
*151	this Walford Police Station…
131	a green and white iron
140	a Walford Gazette banner
145	this jukebox wall unit
137	a Ford script logo
+138	this man with moustache

***: location**
+: person

National Institute of Standards and Technology

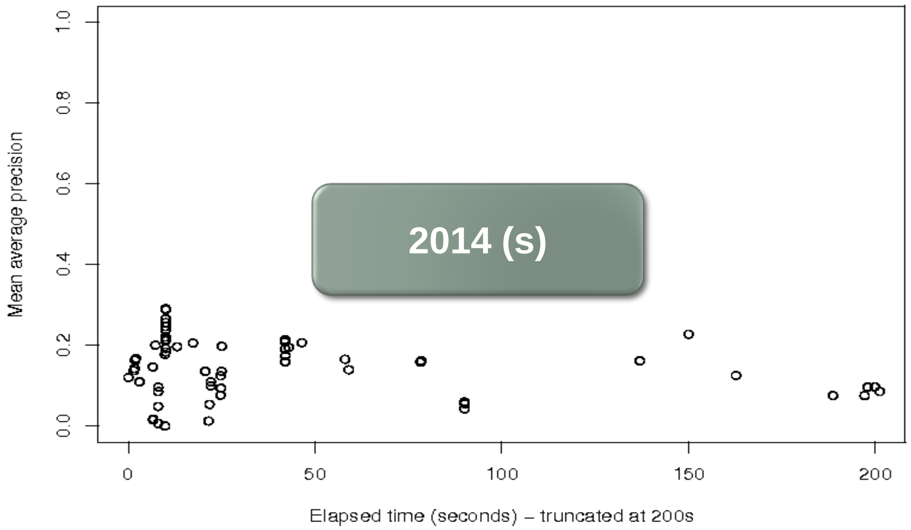
Run results + Randomization testing

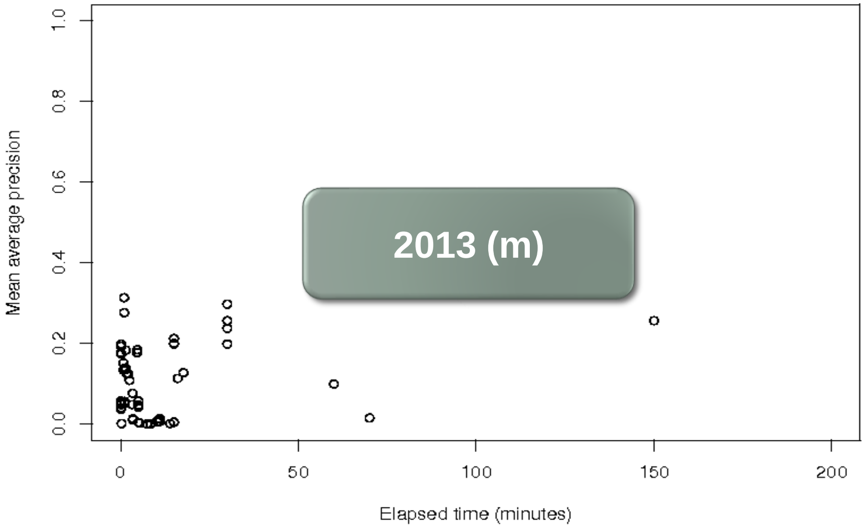
MAP	Top 10 runs across all teams (automatic)										
0.453	F_E_PKU_ICST_1	=		>				>		>	
0.443	F_E_PKU_ICST_3		=								
0.424	F_A_PKU_ICST_4			=							
0.424	F_A_NII_Hitachi_UIT_3				=						
0.418	F_A_NII_Hitachi_UIT_4					=					>
0.415	F_A_NII_Hitachi_UIT_2						=				>
0.403	F_A_BUPT_MCPRL_4							=			
0.403	F_A_BUPT_MCPRL_3								=		
0.403	F_A_BUPT_MCPRL_1									=	
0.401	F_A_NII_Hitachi_UIT_1										=
		1	2	3	4	5	6	7	8	9	10

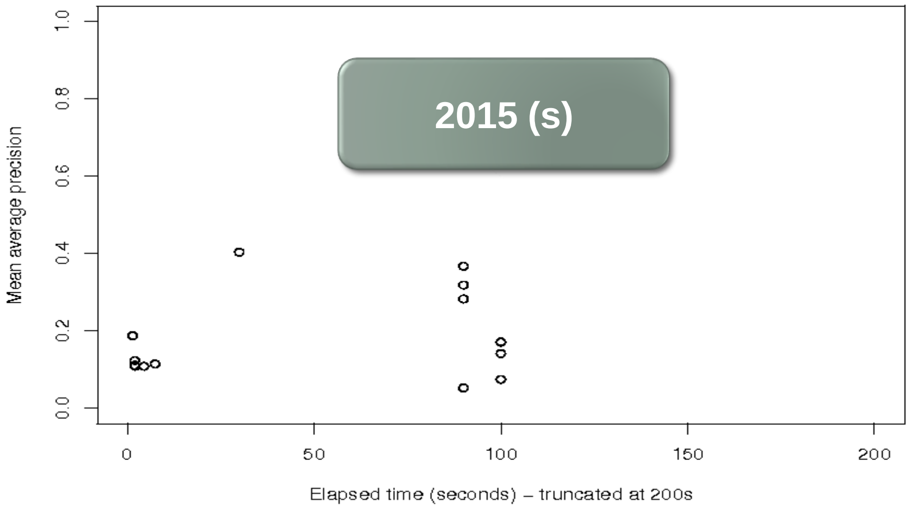
p = probability the row run scored better than the column run due to chance

> $p < 0.05$

MAP vs. per query clock processing time (automatic)



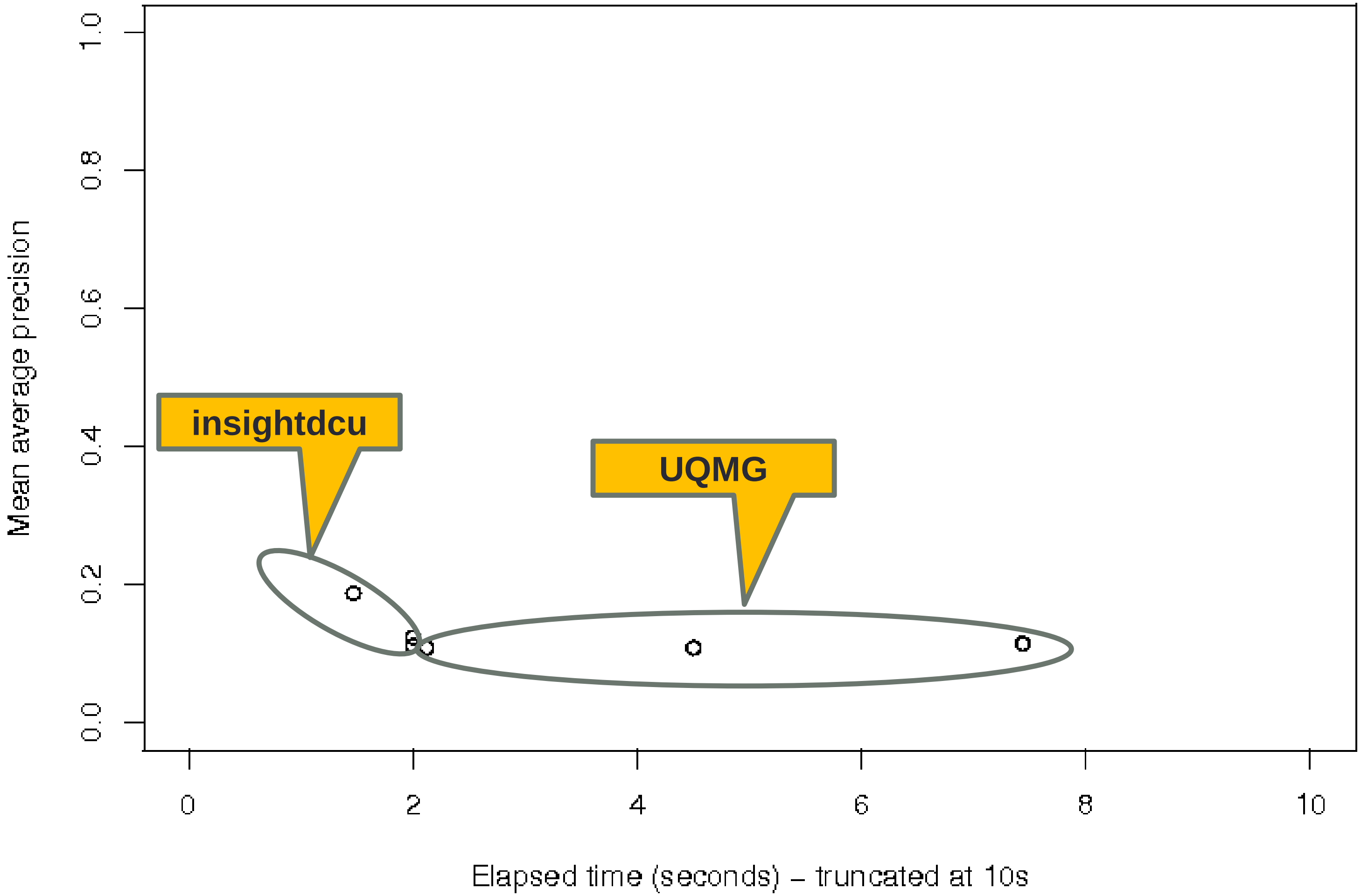




17 out 50 runs
< 200s

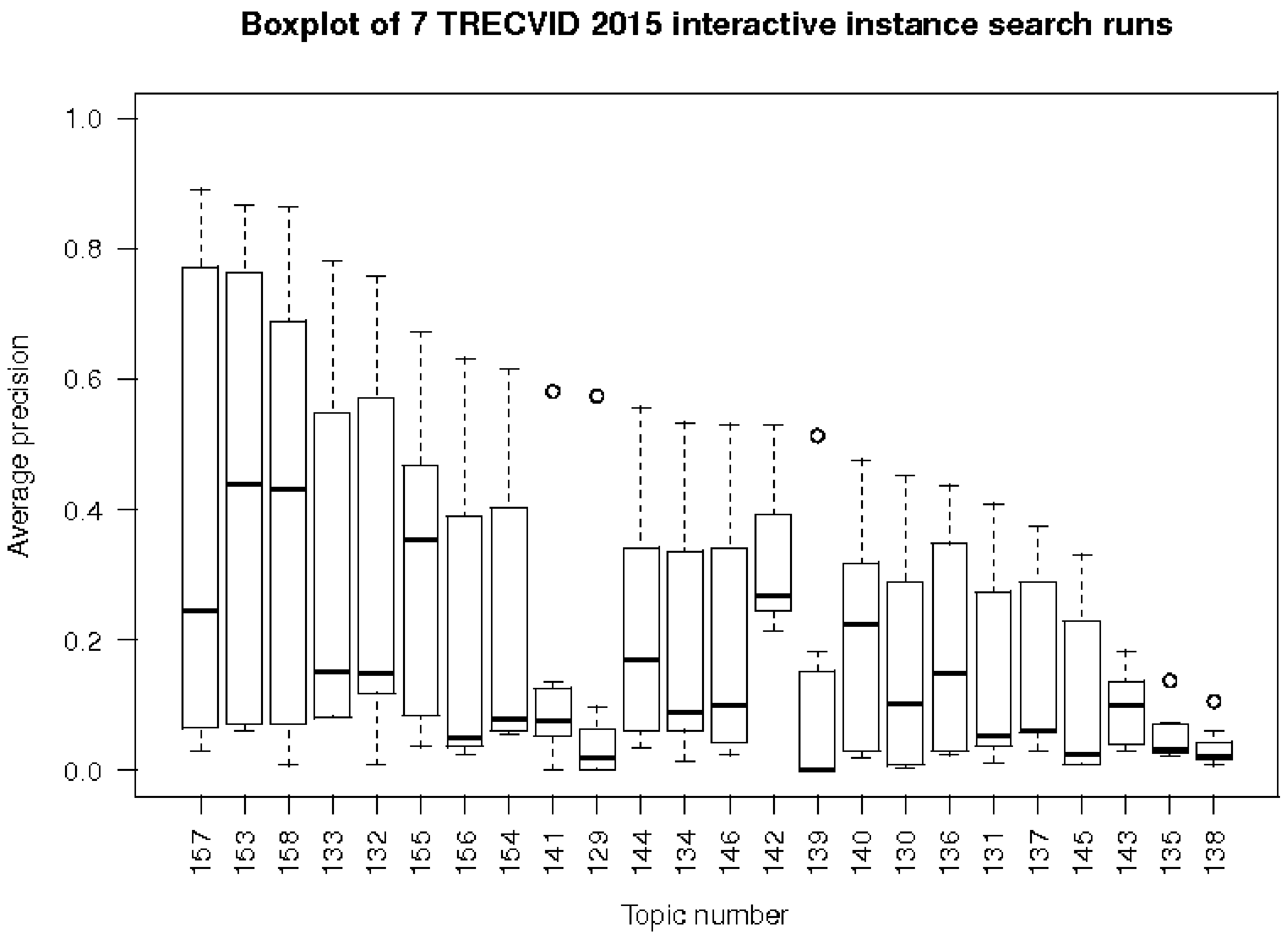
MAP vs. fastest query processing time

(<=10 s, automatic)



Results by topic - interactive

Targets with single location in BLUE



#	Text
157	this picture of flowers
153	this starburst wall clock
158	this flat wire vase with flowers
133	this lava lamp
132	this brass piano lamp…
155	this dart board
156	a 'DEVLIN' lager logo
154	this neon Kathy's sign
141	this guinea pig
129	this silver necklace
144	this doorknocker on #27
134	this cylindrical spice rack
146	this change machine
142	this Chihuahua dog
139	this shaggy dog
140	a Walford Gazette banner
130	a chrome napkin holder
136	this yellow VW beetle…
131	a green and white iron
137	a Ford script logo
145	this jukebox wall unit
+143	this bald man
135	this turquoise stroller
+138	this man with moustache

Run Results, Randomization testing

Top 10 runs across all teams (interactive)

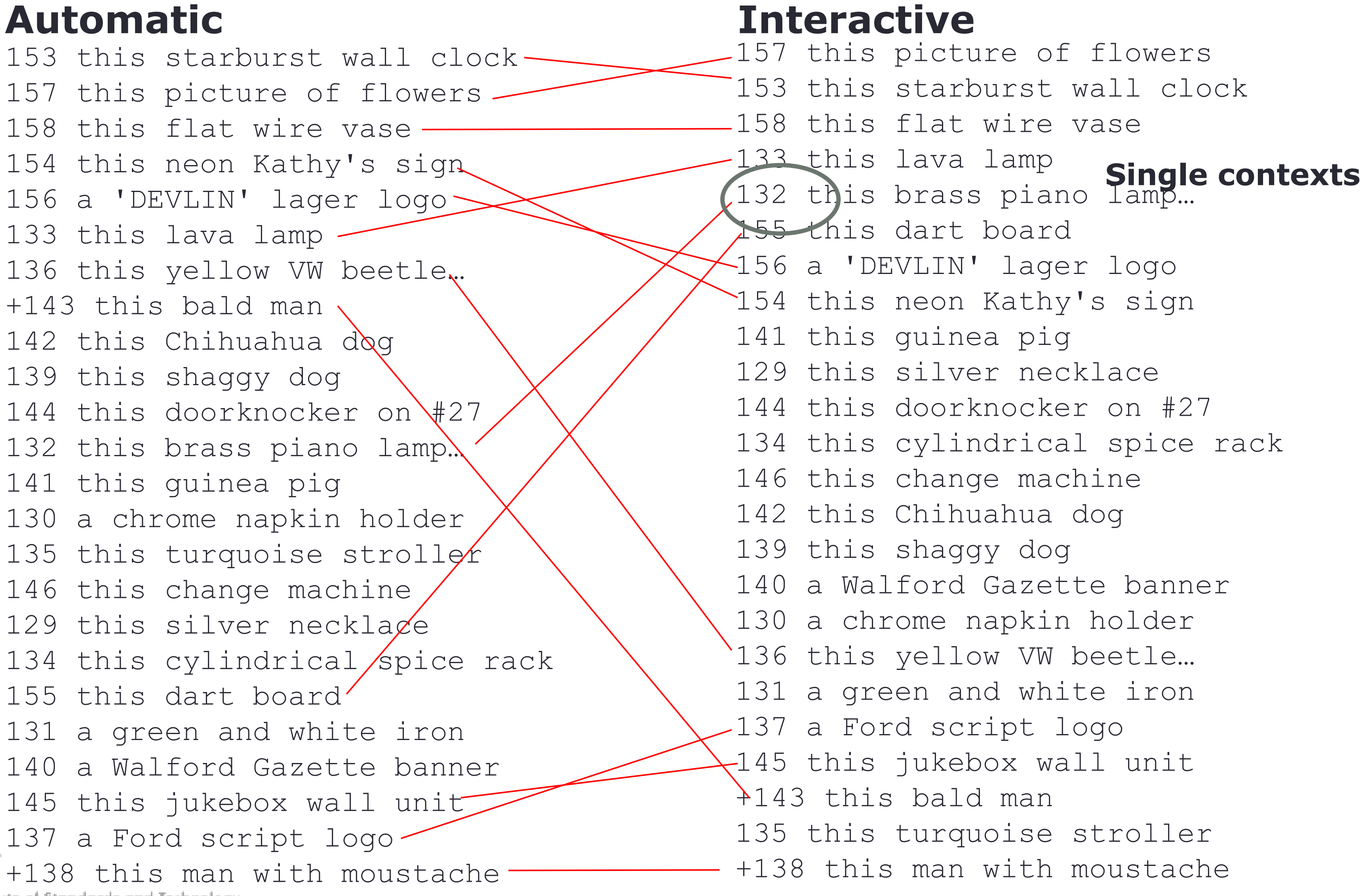
MAP								
0.517	I_E_PKU_ICST_2	=	>	>	>	>	>	>
0.388	I_A_BUPT_MCPRL_2		=	>	>	>	>	>
0.269	I_A_insightdcu_3			=	>	>	>	>
0.171	I_E_TUC_1				=	>	>	>
0.064	I_A_ITI_CERTH_1					=		>
0.053	I_A_ITI_CERTH_2						=	
0.046	I_A_ITI_CERTH_3							=
		1	2	3	4	5	6	7

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

$>$ $p < 0.05$

Automatic vs interactive topics

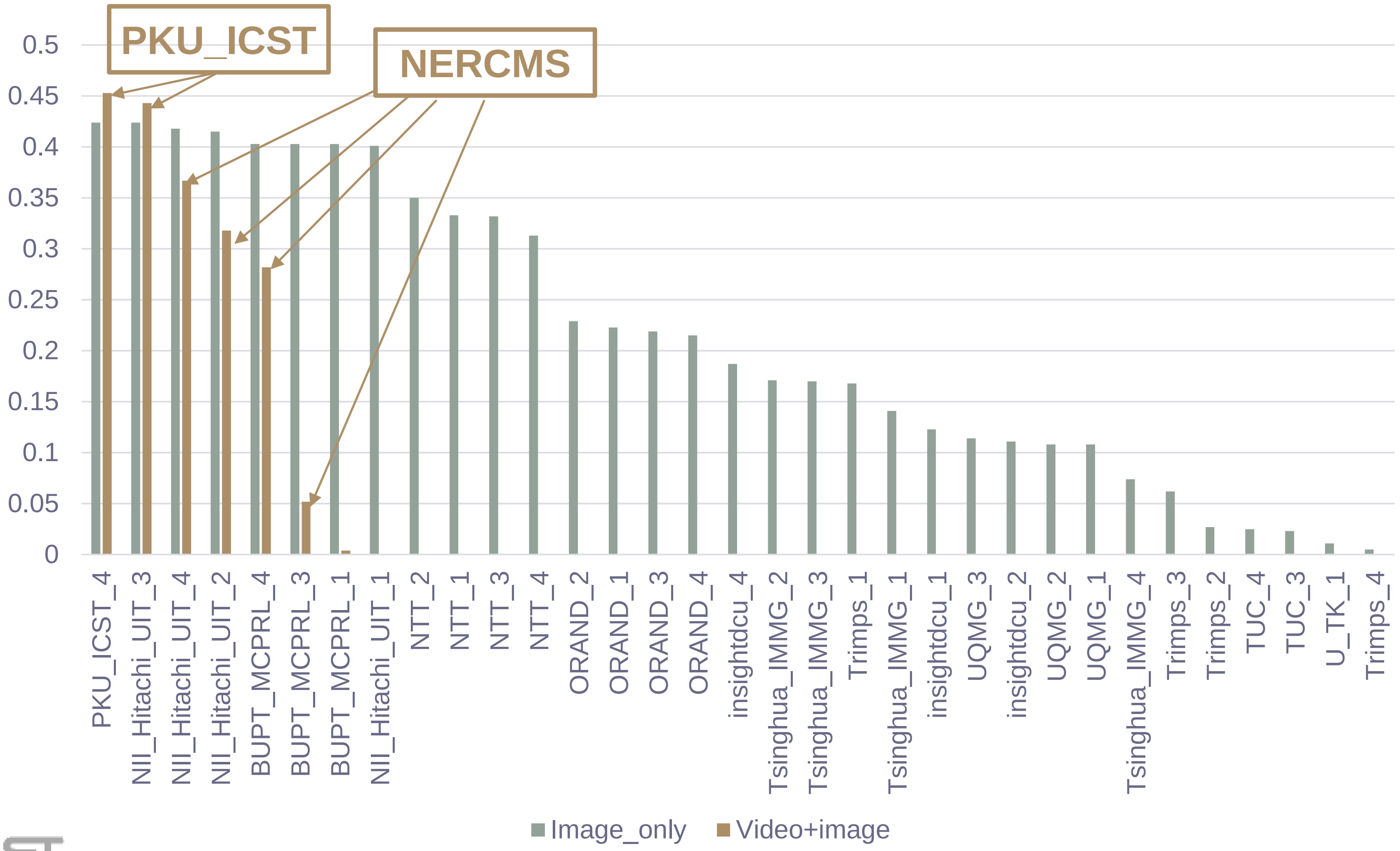
(ranked by max performance on the topic)



Automatic	Interactive
153 this starburst wall clock	157 this picture of flowers
157 this picture of flowers	153 this starburst wall clock
158 this flat wire vase	158 this flat wire vase
154 this neon Kathy's sign	133 this lava lamp
156 a 'DEVLIN' lager logo	132 this brass piano lamp…
133 this lava lamp	155 this dart board
136 this yellow VW beetle…	156 a 'DEVLIN' lager logo
+143 this bald man	154 this neon Kathy's sign
142 this Chihuahua dog	141 this guinea pig
139 this shaggy dog	129 this silver necklace
144 this doorknocker on #27	144 this doorknocker on #27
132 this brass piano lamp…	134 this cylindrical spice rack
141 this guinea pig	146 this change machine
130 a chrome napkin holder	142 this Chihuahua dog
135 this turquoise stroller	139 this shaggy dog
146 this change machine	140 a Walford Gazette banner
129 this silver necklace	130 a chrome napkin holder
134 this cylindrical spice rack	136 this yellow VW beetle…
155 this dart board	131 a green and white iron
131 a green and white iron	137 a Ford script logo
140 a Walford Gazette banner	145 this jukebox wall unit
145 this jukebox wall unit	+143 this bald man
137 a Ford script logo	135 this turquoise stroller
+138 this man with moustache	+138 this man with moustache

Single contexts

Results by example set (A/E) - automatic



Some general observations about the task

- 3rd iteration on the Eastenders dataset:
  - Drop in number of participants
  - MAP has increased, not clear if this means progress
    - But: participants report a bit of progress (compared to last year systems)
  - Persons are still the most difficult category
  - progress smaller, perhaps needs new challenge
- E condition was used by just a few teams
  - But the E (video) condition was used for top runs
- Interactive search task
  - Helps improving MAP of instances with varying backgrounds

Overview of submissions (1)

- 11 out of 14 teams described INS runs for the TV notebook
- 4 teams will present their INS experiments

- **2:30 - 2:50**, NTT (NTT Comm. Science Lab.; NTT Media Intelligence Lab.)
- **2:50 - 3:10**, NERCMS (Wuhan University - Natl. Eng. Res. Center for MM Software)
- **3:10 - 3:30**, BUPT_MCPRL (Beijing University of Posts and Telecommunications)

- **3:30 - 3:50**, **Break** with refreshments

- **3:50 - 4:10**, NII_HITACHI-UIT (National Inst. of Informatics; Hitachi; U. of Inf. Tech.)
- **4:10 - 4:30**, Discussion

Overview of submissions (2)

- Nearly all systems use some form of SIFT local descriptors
  - Large variety of experiments adressing representation, fusion or efficiency challenges
- Most systems also include a CNN component
  - Better understanding when CNN can help
- Many experiments with post-processing (spatial verification, feedback)
- Exploring closed captions and fan resources for additional evidence (using topic descriptive text)

Finding an optimal representation

- Teams report improvement from processing more frames (**Wuhan)**
- Combining different feature types (local/global)
    - **BUPT:** Use CNN for both local and global features + 3 local features
- Direct comparsion CNN vs SIFT
    - **InsightDCU:** SIFT/BovW outperforms CNN only runs, features from convolutional layers better than fully connected
- Combination methods
    - **PKU-ICST:** fuse CNN, SIFT BOW and text (captions)

Finding an optimal representation (2)

- **LAHORE en SHEFFIELD:** 4 different combinations of 4 different local features and 4 matching methods
  - (i) combining hsvSIFT features with GMM matching rank list,
  - (ii) SIFT features with Bhatacharya distance for similarity measurement,
  - (iii) Combination of Colour SIFT descriptor with LUCENE,Terrier matching algorithm,
  - iv) HOG(Histogram of Oriented Gradients) features alone, matching: euclidean distance.
- **TRIMPS:** compared
  - 1. BOW: oppo-SIFT + Streamed-KMeans + FastANN
  - 2. RCNN global features (euclidean distance)
  - 3. Selective Search + CNN + LSH
  - 4. HOGgles + local features
- **TU_CHEMNITZ**: explored classification of audio track (as in 2014)

Finding an optimal representation (3)

- **UMQG:** (Queensland)
  - New approach based on object detection and indexing
  - 1. video decomposition, extracting objects
  - 2. describing objects (CNN)
  - 3. matching query image with nearest object
  - Codebook, quantization
  - Result: approach cannot rival yet standard SIFT/BOW approach

Dealing with query images

- How to exploit the mask (focus vs background)
  - **Wuhan:** manual selection of ROI on different query images: helped significantly.
- Combining sample images
  - Not mentioned in papers
- Exploiting the full query video clip (for query expansion)
  - Successfully applied by **PKU_ICST** and **NERCS**
  - Full clips are also mined for interactive runs (Chemnitz, Wuhan)

Matching

- Typically: Inverted files for fast lookup in sparse BovW space (Lucene),
- Experiments with similarity function:
  - **BUPT** Query adaptive late fusion (equals manual tuned system)
  - **Wuhan:** Asymmetrical query adaptive matching
- Pseudo relevance feedback, query expansion
  - Mentioned in several papers

Postprocessing the ranked list (1)

- **InsightDCU:** weak geometry consistency check for spatial filtering helped

- **NII-HITACHI:** postprocessing experiments
  - 1. query adaptive weighting, DPM and BOW (weight based on NN)
  - 2. DPM (deformable part models) and Fast RCNN
  - 2nd system is slightly better than last year's system

- **Wuhan university:**
  - Apply face filter and color filter (as in 2014)
  - new: adjacent shot matching,
  - new: query text expansion/matching on captions

Postprocessing the ranked list (2)

- **NTT: spatial verification**
  - 1. Ensemble of weak geometric relations (multiple pairwise geometric constraints)
  - 2. Angle Free : Hough voting in 3D camera motion space
  - Methods are complementary and combination yields best results

- **TU Chemnitz:**
  - Indoor/Outdoor detector based on audio analysis for removing false matches
  - Sequence clustering (similar shots)

Interactive experiments

- **TU_CHEMNITZ:** 1 run; fast review of 3500 instances, improved on automatic
- **BUPT:** 1 run (performed lower than automatic)
- **INSIGHTDCU:** 1 run (outerperformed automatic)
- **ITI_CERTH:** 3 runs: BoW, saliency detection, combi (small differences)
- **PKU_ICST:** 2 rounds of relevance feedback on initial run. Fusion with original run

End of INS overview

Some questions

- Is 464 hours of video challenging enough?
- Should we decrease interactive search time?
- Should we explore natural language queries (cf. visualqa)?
  “the guy in the background with the moustache”
- Exploiting captions
  - How do we deal with the success of using the closed captions?
  - Need special run category?
- Any ideas for experimental contrast conditions that we want to focus on as a community? Any ideas for new data?
  - E.g. images vs video example, types of modalities,

Recommendations for the final paper

- Re-run a TV13 or TV12 on TV 14 data to help monitoring progress over the years.

- Perform a per topic or per topic class error analysis to get a better understanding about the pros and cons of certain techniques for particular target characteristics. *Why did it work or fail?*

INS 2016 plans

Continue with same test data and new set of 30 topics

Consider new type of topic: location + person

- Provide 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