TRECVID 2014 INSTANCE RETRIEVAL

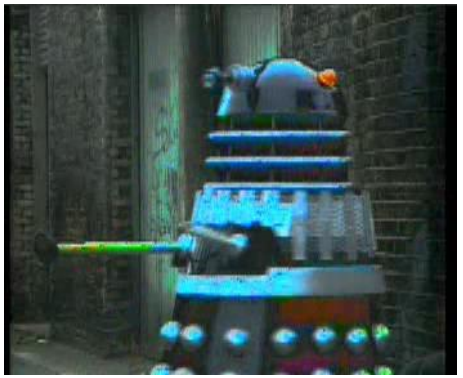
AN INTRODUCTION ….

Wessel Kraaij
TNO, Radboud University Nijmegen

Paul Over
NIST

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 example the images came
  - a target type (OBJECT/LOGO, PERSON)
  - <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

Topics: selection criteria

Tried to include targets with various degrees/sources of variability:

- **Inherent characteristics**: boundedness, size, rigidity, planar/non-planar, mobility,...
- **Locale**: multiplicity, variability, complexity,...
- **Camera view**: distance, angle, lighting,...

Kinds of targets (very similar to 2013's):

- rigid, non-planar objects, large and small
- logos, other objects manufactured to be identical
- people/animals

Topics:

Effect of examples – 5 conditions:

- A - example #1 only
- B - examples #1 and #2 only
- C - examples #1, #2, and #3 only
- D - all four examples only
- E - video examples (+ optionally image examples)

How were these interpreted? "A" -> any single image or just image #1? Etc.

Dropped topics:

9100: SLUPSK vodka - only 2 true positives

9113: vest – text was too restrictive

9117: pay phone - late change in text ("a" -> "this")

Topics – segmented example images

Source
"this woman"

Region of interest mask

Topics – 19 Objects

Topic: **True positives:**

99 494

A checkerboard band …

100 2

a SLUPSK … bottle

101 1568

a Primus … machine

102 398

this large vase …

103 1818

a … ketchup container

105 97

this dog, Wellard

Topics – 19 Objects (cont.)

Topic: **True positives:**

106 243

an …Underground logo

108 121

these 2 … heads

109 104

a Mercedes star logo

110 444

these etched glass doors

111 416

this dartboard

112 846

this Holmes … logo …

Topics – 19 Objects (cont.)

Topic: **True positives:**

113

a yellow-green ... vest

114 387

a ... public mailbox

117

a pay phone

118 4

a Ford Mustang ... logo

120 189

a wooden park bench ...

121 730

a Royal Mail ... vest

Topics – 16 Objects (cont.)

Topic:	True positives:
122	211

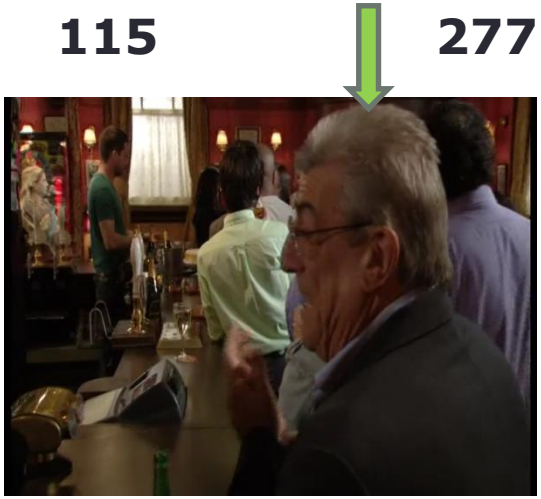
this round watch with black face and black leather band

Topics – 4 Persons

104 342

this woman

115 277



this man

116 238

this man

119 180

this man

Topics – 1 Location

107 **229**

this Walford East Station entrance

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

AXES	**Access to Media**
ATTlabs	AT&T Labs Research
BUPT_MCPRL	**Beijing University of Posts and Telecommunications**
ITI_CERTH	**Centre for Research and Technology Hellas**
VIREO	City University of Hong Kong
insightdcu	**Insight Centre for Data Analytics**
IRIM	IRIM Consortium
JRS	JOANNEUM RESEARCH
NU	Nagoya University
NII	National Institute of Informatics
NTT_CSL	NTT Communication Science Laboratories
ORAND	**ORAND S.A. Chile**
OrangeBJ	**Orange Labs International Center Beijing**
PKU-ICST	**Peking University ICST**
TUC_MI	**Technische Universität Chemnitz**
TelecomItalia	Telecom Italia
U_TK	University of Tokushima
TokyoTech-Waseda	Tokyo Institute of Technology, Waseda University
MIC_TJ	Tongji University
Tsinghua_IMMG	Tsinghua University
MediaMill	University of Amsterdam
Sheffield_UETLahore	University of Sheffield, Lahore U. of Engineering and Technology
NERCMS	Wuhan University

BLUE indicates team submitted interactive runs (up from 5)

Evaluation

For each topic (including dropped), the submissions were pooled and judged down to at least rank 120 (on average to rank 260, max 460), resulting in 262632 judged shots (~ 600 person-hrs).

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

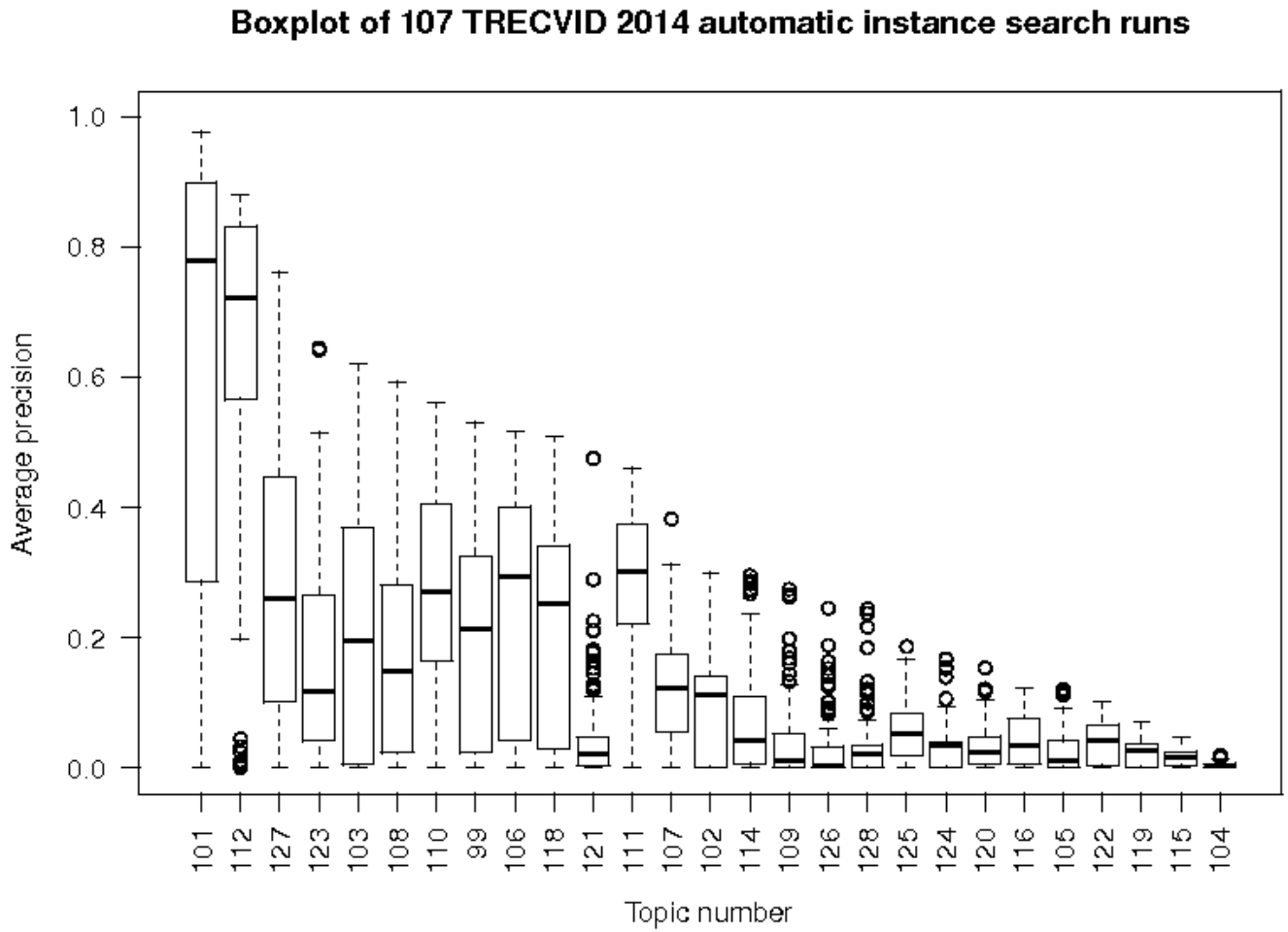
13248 clips (avg. 441.6 / topic) contained the topic target (5%)

True positives per topic: min 2 med 277.5 max 1818

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

Results by topic - automatic

Targets with single location in BLUE



#	Text
101	a Primus washing machine
112	this HOLMES lager logo ...
127	this ... bust of Queen Vic
123	a white plastic kettle ...
103	a ... ketchup container
108	these 2 ceramic heads
110	these etched glass doors
99	a checkerboard band ...
106	a London Underground logo
118	a Ford Mustang grill logo
121	a Royal Mail red vest
111	this dartboard
107	this Walford Station entrance
102	this large vase
114	a red public mailbox
109	a Mercedes star logo
126	a Peugeot logo
128	this F pendant
125	this wheelchair ...
124	this woman
120	a wooden park bench ...
116	this man
105	this dog, Wellard
122	this round watch ...
119	this man
115	this man
104	this woman

Randomization testing

Best run from each of the top 10 teams (automatic)

MAP											
0.325	F_D_NII_2	1	=	>>	>>	>>	>>	>>	>>	>>	>>
0.304	F_D_NU_1	2	=	>>	>>	>>	>>	>>	>>	>>	>>
0.234	F_D_NTT_CSL_1	3	=							>	>>
0.232	F_D_PKU-ICST_2	4	=						>	>	>>
0.227	F_D_MediaMill_1	5	=								>
0.227	F_D_BUPT_MCPRL_1	6	=								>>
0.213	F_D_IRIM_1	7	=								>>
0.197	F_D_VIREO_3	8	=								>
0.183	F_D_ORAND_4	9	=								
0.167	F_D_OrangeBJ_2	10	=								
			1	2	3	4	5	6	7	8	9	10

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

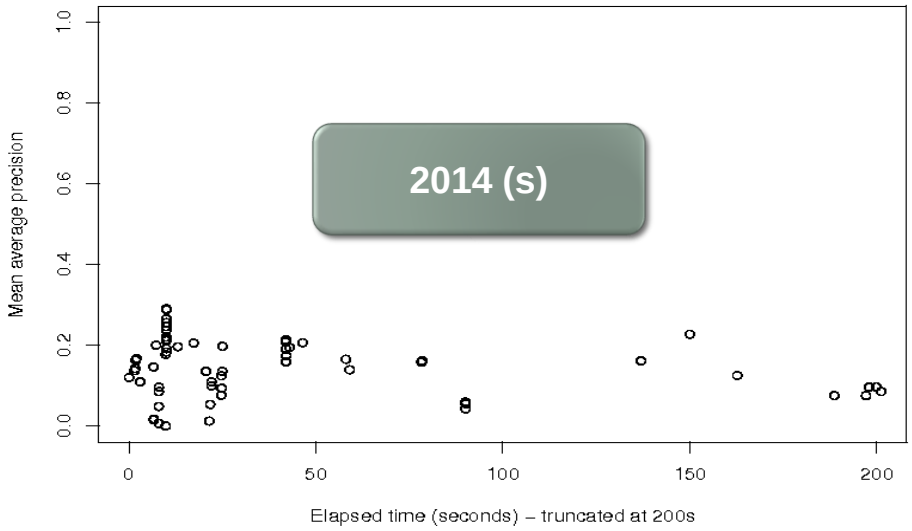
>> p < 0.01

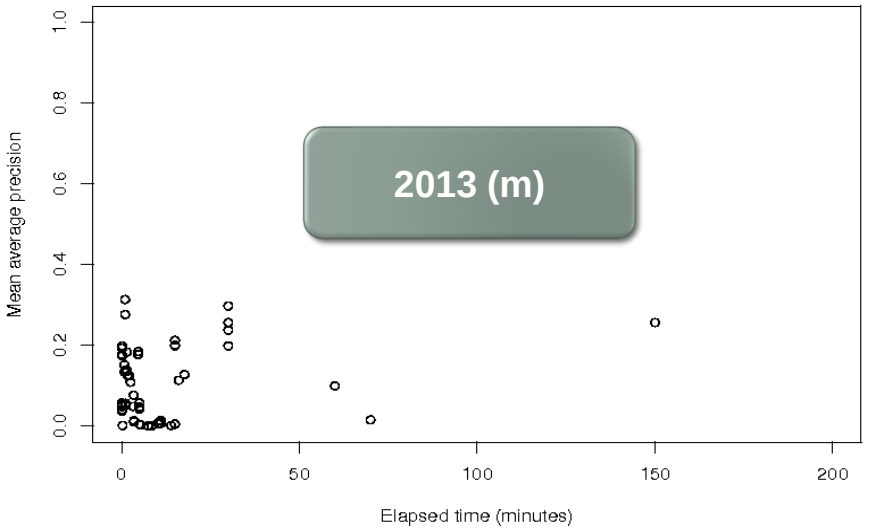
> p < 0.05

NIST National Institute of Standards and Technology

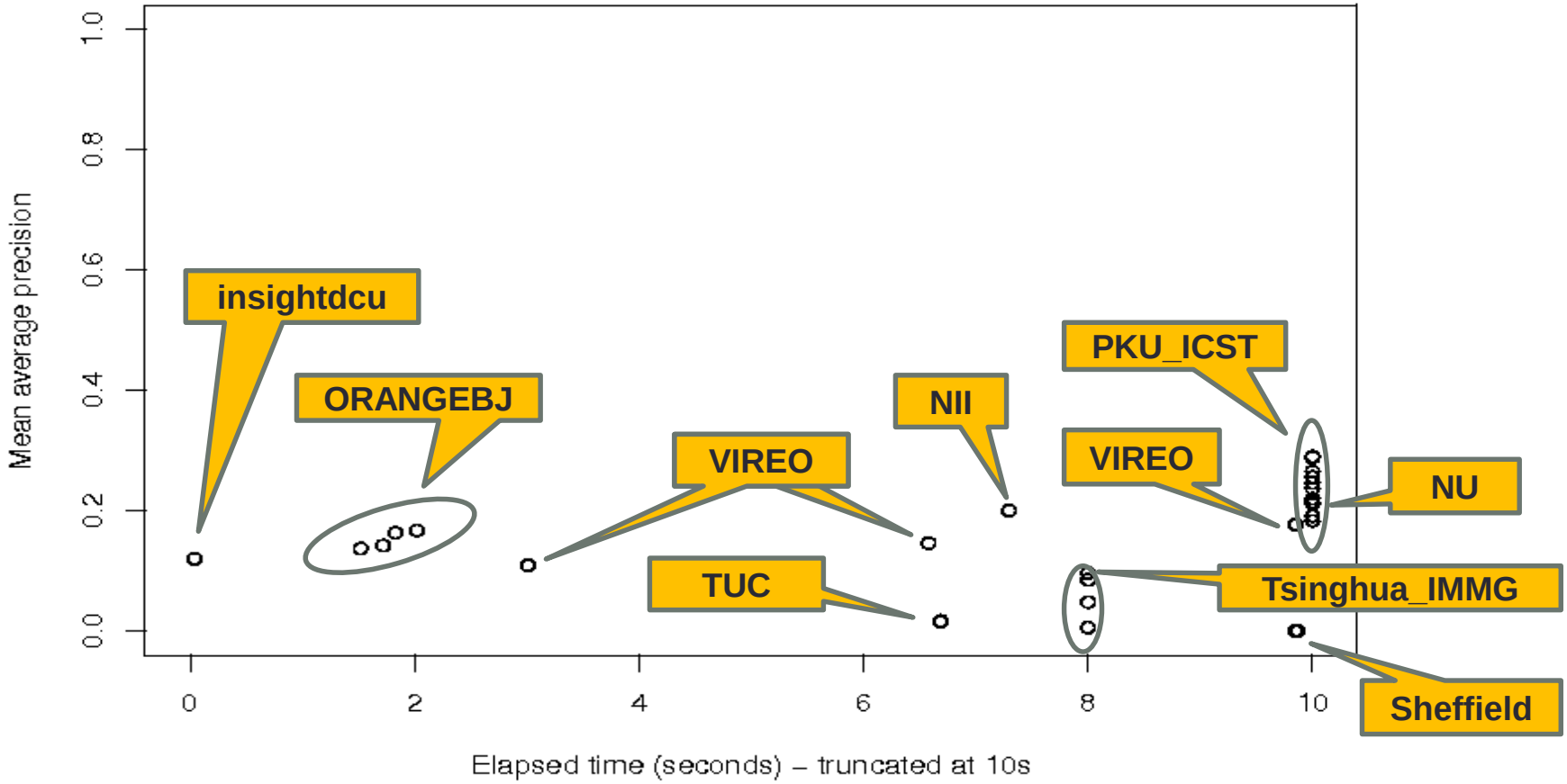
TNO innovation for life

MAP vs. query processing time (automatic)



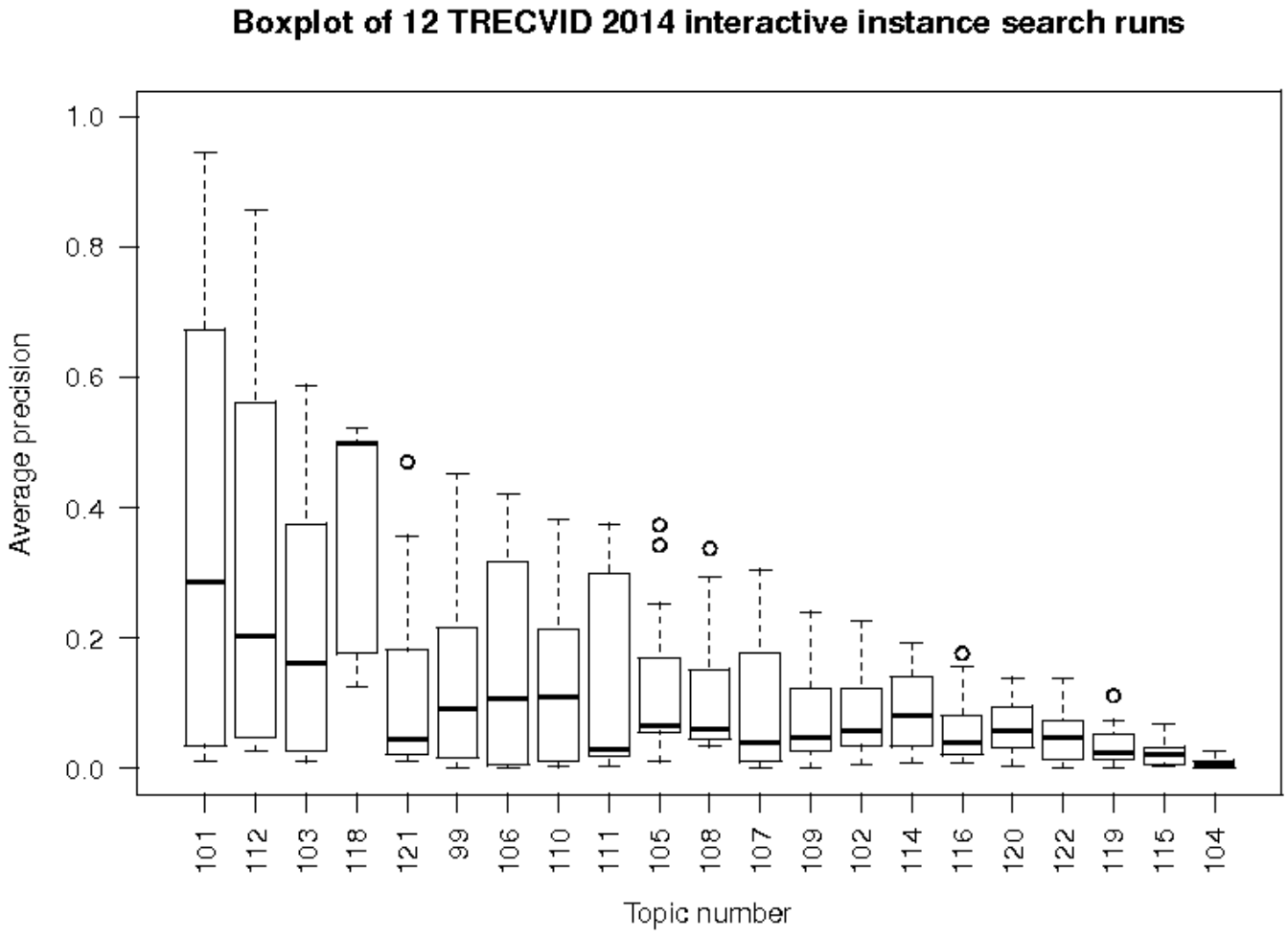


MAP vs. fastest query processing time
(<=10 s, automatic)



Results by topic - interactive

Targets with single location in BLUE



#	Text
101	a Primus washing machine
112	this HOLMES lager logo ...
103	a ... ketchup container
118	a Ford Mustang grill logo
121	a Royal Mail red vest
99	a checkerboard band ...
106	a London Underground logo
110	these etched glass doors
111	this dartboard
105	this dog, Wellard
108	these 2 ceramic heads
107	this Walford Station entrance
109	a Mercedes star logo
102	this large vase
114	a red public mailbox
116	this man
120	a wooden park bench ...
122	this round watch ...
119	this man
115	this man
104	this woman

Randomization testing

Best run from each of the top 10 teams (interactive)

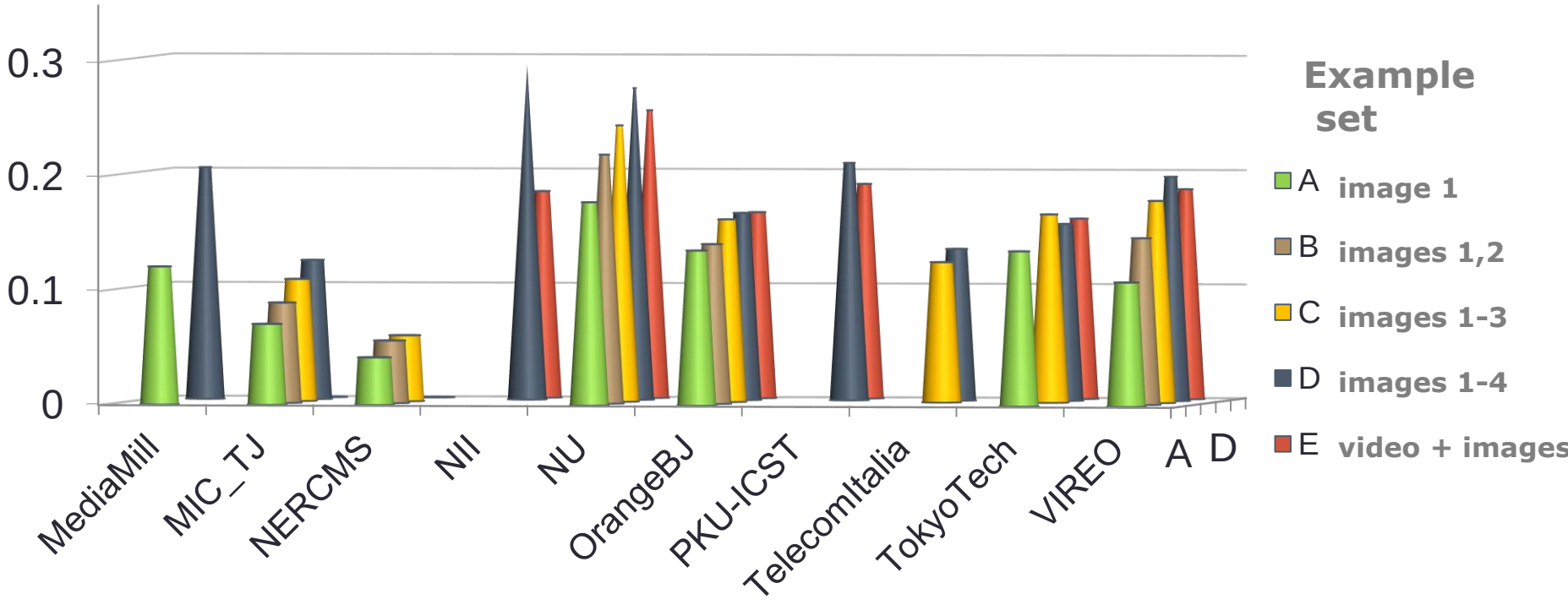
MAP									
0.317	I_D_PKU-ICST_3	1	=	>>	>>	>>	>>	>>	>>	>>
0.249	I_D_OrangeBJ_3	2		=		>	>	>>	>>	>>
0.237	I_D_BUPT_MCPRL_2	3			=		>	>>	>>	>>
0.174	I_D_ORAND_3	4				=		>>	>>	>>
0.135	I_D_insightdcu_2	5					=		>>	>>
0.108	I_D_AXES_1	6						=	>	>>
0.037	I_E_TUC_MI_1	7							=	
0.032	I_D_ITI_CERTH_1	8								=
			1	2	3	4	5	6	7	8

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

>> p < 0.01

> p < 0.05

Results by example set - automatic



Scores for multiple runs with same example set were averaged

Some observations about the task

- 2nd iteration on the Eastenders dataset: task seems healthy
  - Stable number of participants
  - Dataset is challenging enough, despite closed world setting
  - Systems produce meaningful results
  - Participants report progress
  - Persons are the moset difficult category
- Interactive search task helps focusing on efficiency

Overview of submissions (1)

- 19 out of 23 teams described INS runs for the TV notebook (Missing:ATTLABS, PKU_ICST, U_TK, Tsinghua_IMMG)
-
- 5 teams will present their INS system

2:10 - 2:35, National Institute of Informatics, Japan (NII)
2:35 - 3:00, Nagoya University (NU)
3:00 - 3:25, NTT Communication Science Laboratories (NTT_CSL)

3:50 - 4:15, Beijing University of Posts and Telecommunications (BUPT)
4:15 - 4:40, ORAND S.A. Chile (ORAND)

Overview of submissions (2)

- Nearly all systems use some form of SIFT local descriptors
  - Large variety of experiments adressing representation, fusion or efficiency challenges
  - Trend is moving to larger BoVW vocabularies, larger nr of keyframes (Nagoya: all)
- New in 2014: several experiments with CNN for intermediate features
- Increased focus on post-processing (spatial verification, feedback)
- Effectiveness of new methods not always consistent across teams (e.g. asymmetric similarity function)➔ further research is needed

Typical INS template system

- Processing clips
  - Keyframe choice (1 per shot – 5fps-all frames)
  - Keyframe downsizing?
- Representation
  - Global (HSV, LBP,CNN,..)
  - Local
    - Detection methods
    - Choice of descriptors
  - Cluster to BoVW
    - 1M words, hard/soft etc
- Matching
  - Similarity function(idf weighting,
  - Weighting ROI vs. background
- Postprocessing
  - spatial verification
  - Face/color filtering
- Feedback
- Fusion of scores
  - Average pooling

Each design choice has an impact on speed and effectiveness

Dealing with topic info

- How to exploit the mask (focus vs background)
  - **MediaMill**: compared **mask**, full and fused
  - **BUPT:** boundary region of mask contains relevant local points (also **InsightDCU:** padding**)**
  - **Vireo**: background context modelling (stare model), helps
- Combining sample images
  - Several teams use joint average querying (Arandjelovic/Zisserman) to combine samples into a single query
- Exploiting the full video clip (for query expansion)
  - **NII:** tracked interest points in ROI, helps a bit (but interlaced video raised issues)
  - **OrangeBJ:** no gains
  - **Tokyotech:** tracking and warping the mask: small gain
  - **VIREO:** tracking objects in query video helps if video quality is good (often not the case)

Finding an optimal representation

- Teams try to process more frames (**IRIM, Nagoya)**
- Combining different feature types (local/global)
  - **IRIM:** review of techniques and their results
  - **BUPT:** combines BoVW and CNN

- Combining multiple keypoint detectors and multiple descriptors
  - **Nagoya:** a single descriptor (Hessian Affine ROOTSift) is almost as good as a combination of 6, yet is more efficient!
  - **ORAND:** No quantization codebook, keep raw keypoints (faced scale issue)
  - **Sheffield:** compared SIFT, HOG, Global features

Finding an optimal representation (2)

- Experiments with MPEG VS **TU Chemnitz, TelecomItalia:** OK for mid size rigid objects

- Exploring the potential of CNN **(INSIGHTDCU)**: promising experiments with small scale dataset. Seems to be useful as a representation that could help improve BOVW. Not sufficiently discriminative for primary search keys.

Matching

- Typically: Inverted files for fast lookup in sparse BovW space (Lucene),
- Experiments with similarity function:
  - **NII:** asymmetric similarity function (2013), tested by **IRIM** (no effect), **Nagoya** (helps)
  - **VIREO:** new normalization term in cosine similarity helps to increase recall
- Use of Collection statistics:
  - BM25 enhancements for weighting (**NTT-NII**): did help, as in tv13
  - IDF adjusted for burstiness (**INSIGHTDCU)**
- Pseudo relevance feedback, query expansion
  - **NTT-CSL**: Use ROI features for reranking (promising)

Post filtering

- **NII:** Improved spatial verification method
- **Nagoya:** Spatial verification helps
- **OrangeBJ:** Face detector for filtering hits for topics involving faces: did not help
- **Wuhan university:** Apply face filter and color filter
- **TU Chemnitz:** Indoor/Outdoor detector based on audio analysis for removing false matches

System architecture & Efficiency

- Bag of visual words, indexed video database
  - **Most systems**
  - sparse BovW, Lucene inverted file based scoring

- **JRS:** experimented with compact VLAT signatures: particular signature was not sufficiently discriminative
- **TU Chemnitz:** PostgreSQL on grid platform
- **MIC_TJ (Tongjing Univ)**: Hybrid parallelization using CPU's, GPU's and map/reduce (like in 2013)
- **ORAND** Approximate KNN on un-quantized local descriptors
- **Nagoya:** Efficient re-ranking methods (involving spatial verification)
- **CERTH:** complete index in RAM

Interactive experiments

- **OrangeBJ** (BUPT & Orangelabs) (1 interactive run) Strong performance using "Relative rerank method"
- **BUPT_MCPRL** (1 interactive run): automatic system without CNN, small gain
- **ORAND** (1 run): labels propagated to similar shots in same scene (similarity shot graph)
- **INSIGHTDCU** (2runs): using positive images for new queries outperformed using them for training SVM

Interactive experiments (2)

- **AXES** (2 runs): Pseudo relevance feedback, interactive check
- **TUC_MI** (Chemnitz) 2 runs: MPEG-7 color descriptor, not sufficiently discriminative
- **ITI_CERTH**: (2 runs) shots vs scene presentation: shot based presentation better results

End of INS overview

Some questions

- Any comments about
  - Closed world dataset?
  - Types of objects included in the queries
  - Did anybody use Eastender resources?

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 2015 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