TRECVID 2012 INSTANCE RETRIEVAL PILOT

AN INTRODUCTION ….

Wessel Kraaij
TNO, Radboud University Nijmegen

Paul Over
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:
  - example segmented images of the target (2-6)
  - a target type (PERSON, PLACE, OBJECT)
  - <topic title>
- Return a list of up to 1000 shots ranked by likelihood that they contain the topic target
- Automatic or interactive runs are acccepted

Differences between INS and SIN

INS	SIN
Very few training images (probably from the same clip)	Many training images from several clips
Many use cases require real time response	Concept detection can be performed off-line
Targets include unique entities (persons/locations/objects) or industrially made products	Concepts include events, people, objects, locations, scenes. Usually there is some abstraction (car)
Use cases: forensic search in surveillance/ seized video, video linking	Automatic indexing to support search.

Data …

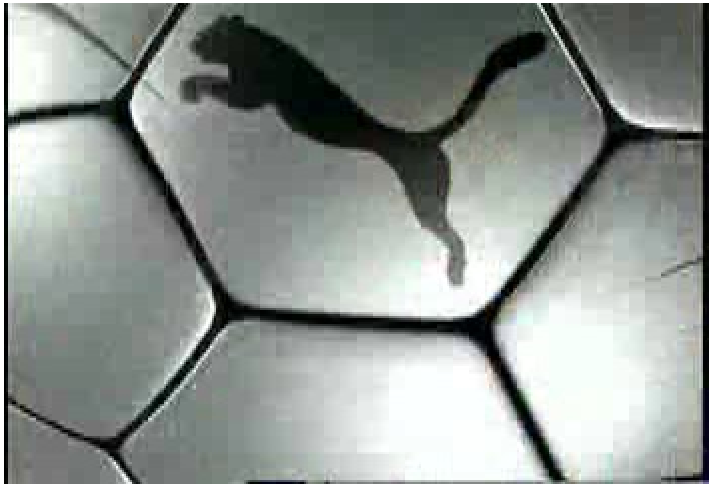
Robin Aly (Twente University),in consultation with NIST:

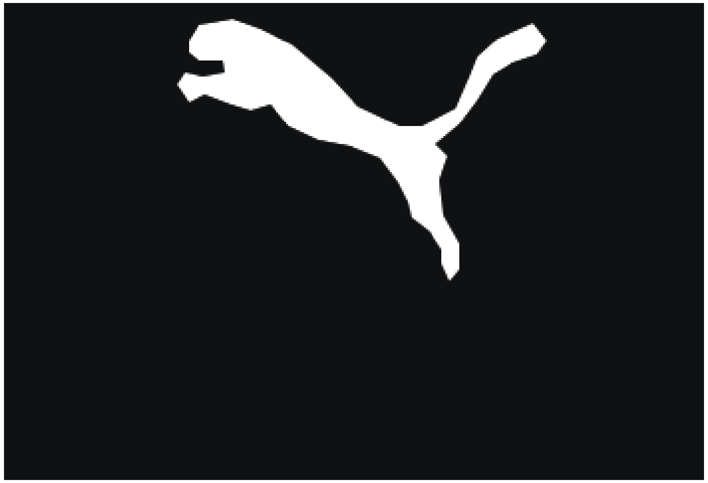
- designed text queries to retrieve videos containing many different instances of the same object, person, location.
- issued several queries against Flickr video available under Creative Commons licenses for research
- provided the query results (videos) divided into 74,958 10s segments to NIST (640x360 25 fps)

NIST:

- reviewed the most promising queries and the videos they returned
- created topics, each targeting a specific object, person, location
- chose example images from some videos and removed those from the test collection.

Topics – segmented example images





Source **Mask**

Topics – 15 Objects

Topic: **#Examples:**

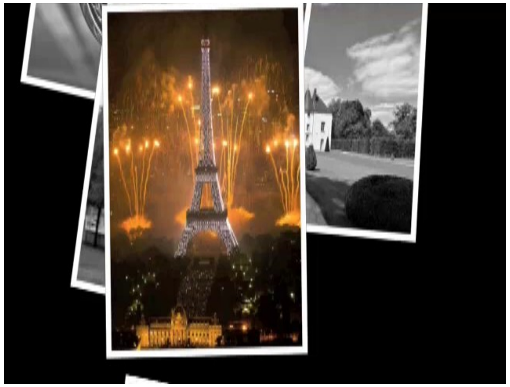
48 5

Mercedes star

49 6

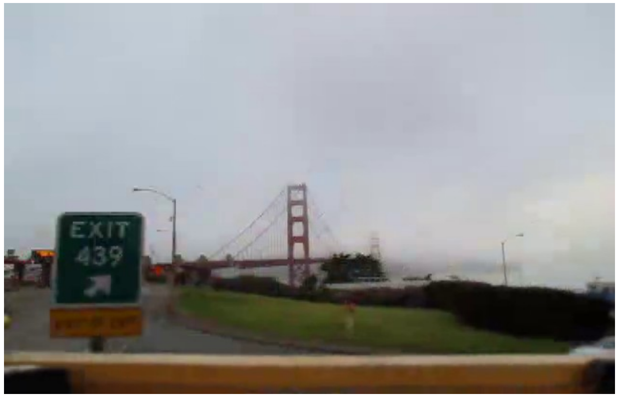
Brooklyn Bridge tower

50 4



Eiffel Tower

51 5

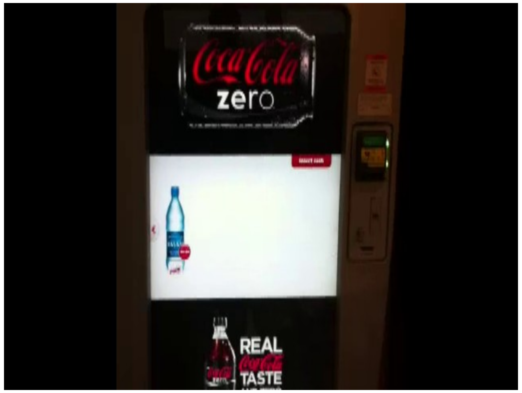


Golden Gate Bridge

52 4

London Underground logo

53 6



Coca-Cola logo

Topics – 15 Objects (cont.)

55 3

Sears/Willis Tower

57 3

Leshan Giant Buddha

58 2

US Capitol exterior

59 4

Baldachin in St.Peter’s

61 4

Pepsi logo (circle)

62 4

One World Trace Center

Topics – 15 Objects (cont.)

64 4

Empire State Building

67 4

MacDonald’s arches

68 6

PUMA logo animal

Topics – 5 Locations

54 5

Stonehenge

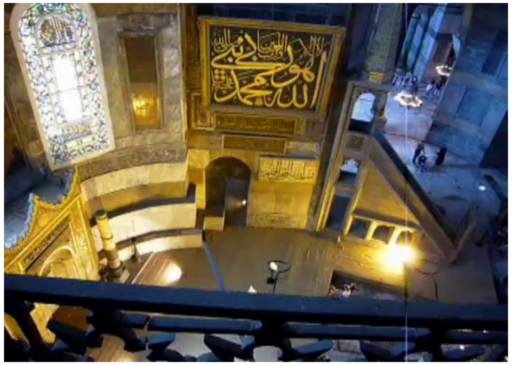
56 9

Pantheon interior

63 4

Prague Castle

65 8



Hagia Sophia interior

66 6

Hoover Dam exterior

Topics – 1 Person

60 **6**

Stephen Colbert

TV2012 24 Finishers (tv11:13)

PicSOM	Aalto U.
Bilkent	Bilkent U. RETINA Vision and Learning Group
CEALIST	CEA
VIREO	City U. of Hong Kong
PRISMA-Orand	Department of Computer Science, U. of Chile.
U_Tokushima	Dept. of Information Science & Intelligent Systems, Tokushima U.
DCU_IAD	Dublin City U., IAD
AXES	Access to Audiovisual Archives: www.axes-project.eu
FTRDBJ	France Telecom Orange Labs (Beijing)
MADM	German Research Center for Artificial Intelligence
ARTEMIS.Ubimedia	Institut TELECOM; TELECOM SudParis; France Alcatel-Lucent
PKU_ICST	Institute of Computer Science and Technology, Peking U.
JRS.VUT	JOANNEUM RESEARCH Forschungsgesellschaft mbH Vienna U. of Technology
IRIM	IRIM - Indexation et Recherche d'Information MultimÃ©dia GDR-ISIS
BUPT.MCPRL	Beijing U. of Posts and Telecommunications
NII	National Institute of Informatics
NTT_NII	NTT Communication Science Laboratories, National Institute of Informatics
IMP	Osaka Prefecture U.
RMIT	RMIT U. School of CS&IT
TNOM3	TNO
MediaMill	U. of Amsterdam
UCSB_UCR_VCG	U. of California, Santa Barbara
sheffield_harbin	U. of Sheffield
ATTLabs	Video and Multimedia Technologies Research Department, AT&T Labs Research

Team submitted interactive runs

Evaluation

For each topic, the submissions were pooled and judged down to at least rank 140 (on average to rank 225), resulting in 189,418 judged shots (525 hrs).

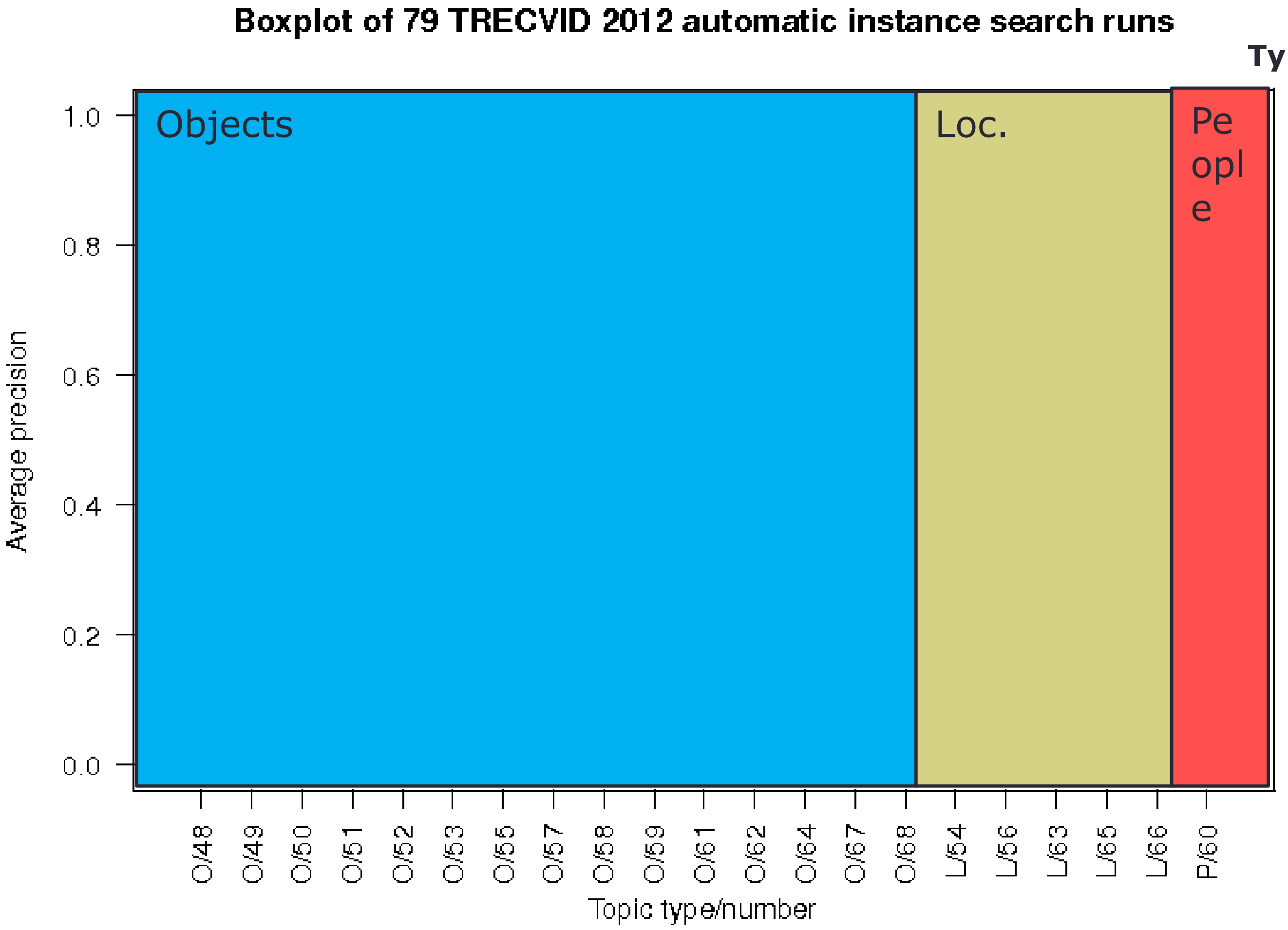
NIST assessors were given their topics in advance and asked to use internet resources to familiarize themselves with each topic target’s appearance.

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

1232 clips (avg. 58.7 / topic) contained the topic target(<1%)

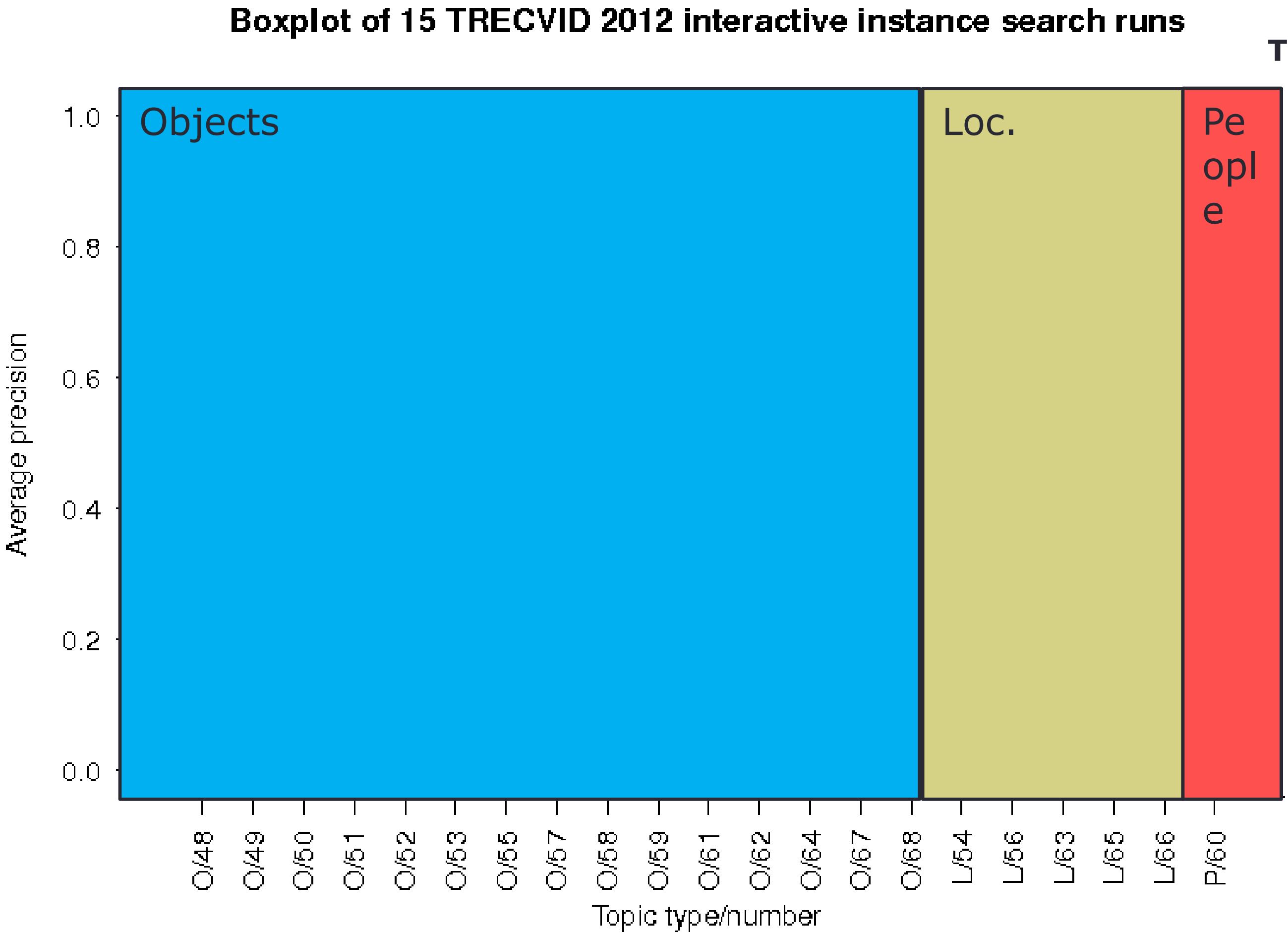
trec_eval_video was used to calculate average precision, recall, precision, etc.

Evaluation – results by topic/type - automatic



Type/#	Name [clips with target]
O/48	Mercedes star [50]
O/49	Brooklyn bridge tower [57]
O/50	Eiffel Tower [276]
O/51	Golden Gate Bridge [295]
O/52	London Underground logo [59]
O/53	Coca-cola logo (letters) [61]
O/55	Sears/Willis Tower [36]
O/57	Leshan Giant Buddha [21]
O/58	US Capitol exterior [79]
O/59	St. Peter’s baldachine [30]
O/61	Pepsi logo – circle [13]
O/62	One World Trade Center [9]
O/64	Empire State Building [51]
O/67	MacDonald’s arches [5]
O/68	Puma logo – animal [15]
L/54	Stonehenge [40]
L/56	Pantheon interior [43]
L/63	Prague Castle interior [25]
L/65	Hagia Sophia interior [10]
L/66	Hoover Dam exterior [12]
P/60	Stephen Colbert [45]

Evaluation – results by topic/type - interactive



Type/# Name [clips with target]

O/48 Mercedes star [50]
O/49 Brooklyn bridge tower [57]
O/50 Eiffel Tower [276]
O/51 Golden Gate Bridge [295]
O/52 London Underground logo [59]
O/53 Coca-cola logo (letters) [61]
O/55 Sears/Willis Tower [36]
O/57 Leshan Giant Buddha [21]
O/58 US Capitol exterior [79]
O/59 St. Peter’s baldachine [30]
O/61 Pepsi logo – circle [13]
O/62 One World Trade Center [9]
O/64 Empire State Building [51]
O/67 MacDonald’s arches [5]
O/68 Puma logo – animal [15]

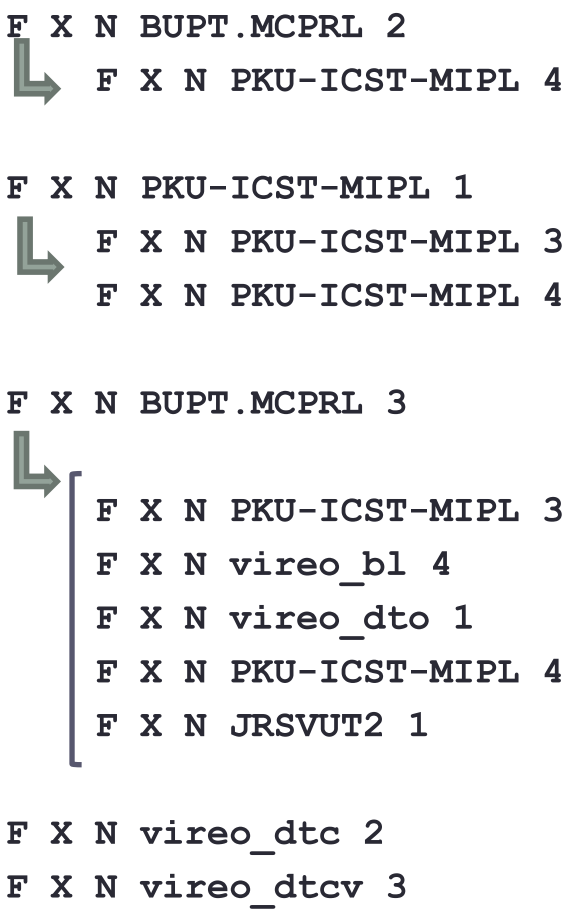
L/54 Stonehenge [40]
L/56 Pantheon interior [43]
L/63 Prague Castle interior [25]
L/65 Hagia Sophia interior [10]
L/66 Hoover Dam exterior [12]

P/60 Stephen Colbert [45]

Evaluation – top 20, based on MAP

	Automatic		MAP
F X N	BUPT.MCPRL	3	0.268
F X N	BUPT.MCPRL	2	0.245
F X N	PKU-ICST-MIPL	1	0.220
F X N	vireo_dtc	2	0.202
F X N	vireo_dtcv	3	0.200
F X N	PKU-ICST-MIPL	3	0.189
F X N	vireo_bl	4	0.188
F X N	vireo_dto	1	0.181
F X N	PKU-ICST-MIPL	4	0.173
F X N	JRSVUT2	1	0.172
F X N	IMP.h_f_e2	2	0.169
F X N	IMP.h_f_e1	4	0.169
F X N	NII	1	0.168
F X N	IMP.h_e2	1	0.165
F X N	JRSVUT3	3	0.161
F X N	JRSVUT4	4	0.160
F X N	IMP.h_e3	3	0.157
F X N	prisma-two180px	1	0.155
F X N	NTT-NII	1	0.150
F X N	NTT-NII	3	0.148

Randomization test



F X N BUPT.MCPRL 2
- F X N PKU-ICST-MIPL 4

F X N PKU-ICST-MIPL 1
- F X N PKU-ICST-MIPL 3
- F X N PKU-ICST-MIPL 4

F X N BUPT.MCPRL 3
- F X N PKU-ICST-MIPL 3
- F X N vireo_bl 4
- F X N vireo_dto 1
- F X N PKU-ICST-MIPL 4
- F X N JRSVUT2 1

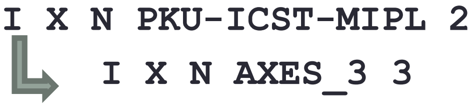
F X N vireo_dtc 2
F X N vireo_dtcv 3

The bold arrows denote statistically significant differences

Evaluation – top 20, based on MAP

Interactive			MAP
I X N	ICST-MIPL	2	0.270
I X N	FTRDBJ	4	0.251
I X N	AXES_2	2	0.229
I X N	AXES_4	4	0.202
I X N	AXES_1	1	0.190
I X N	AXES_3	3	0.173

Randomization test



I X N PKU-ICST-MIPL 2

I X N AXES_3 3

I X N AXES_4 4

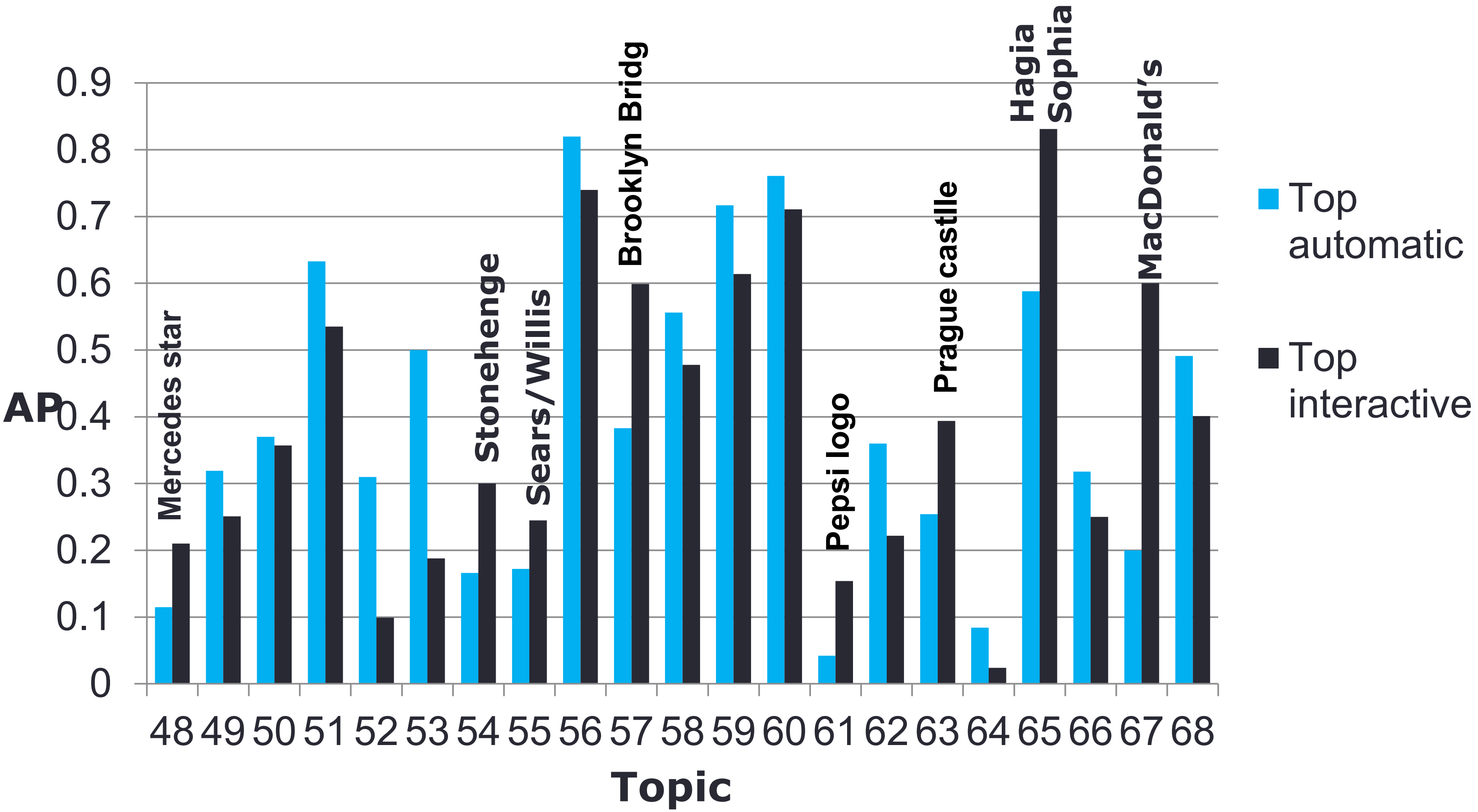
I X N AXES_2 2

I X N AXES_1 1

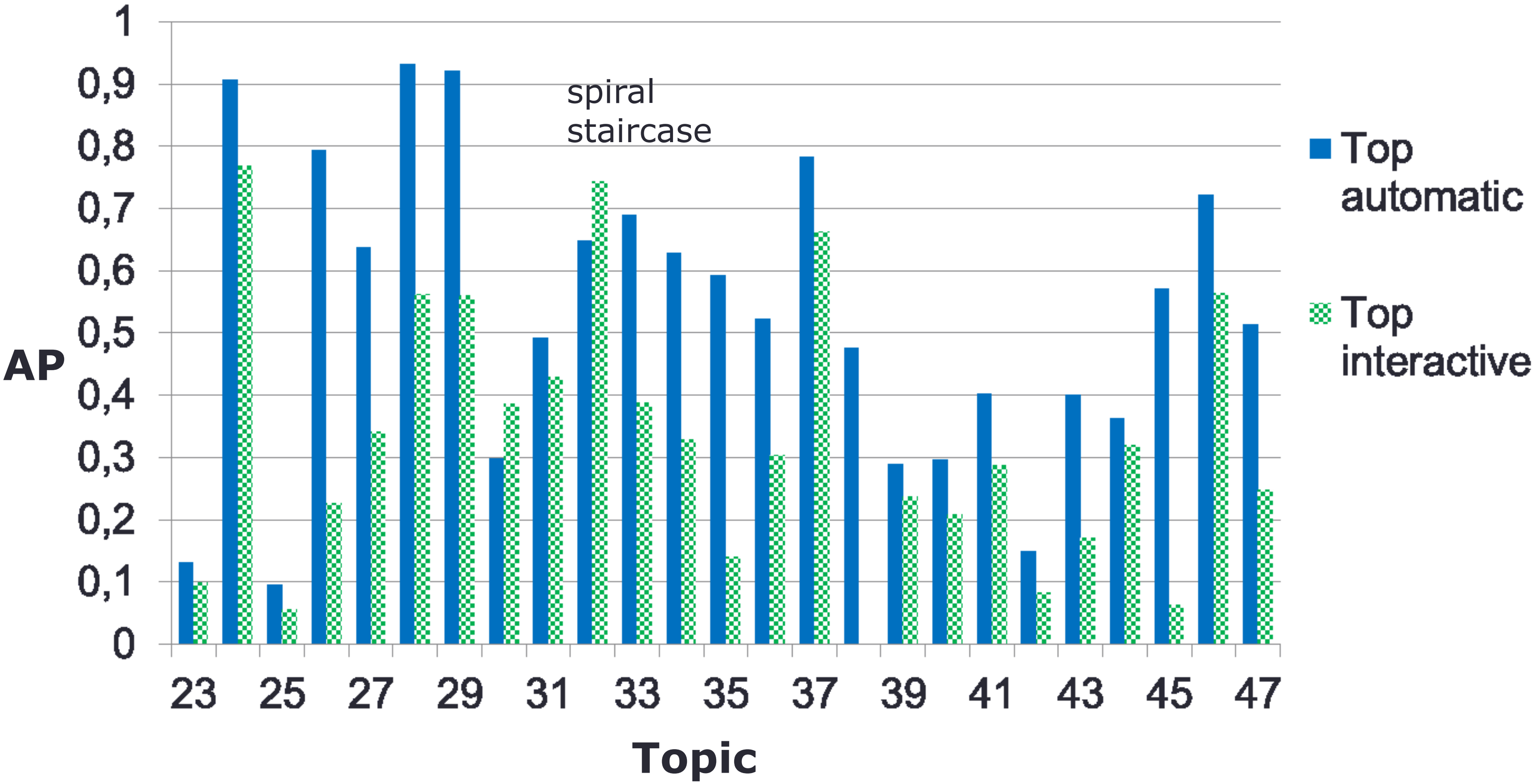
I X N FTRDBJ 4

The bold arrows denote statistically significant differences

Evaluation – top automatic vs interactive

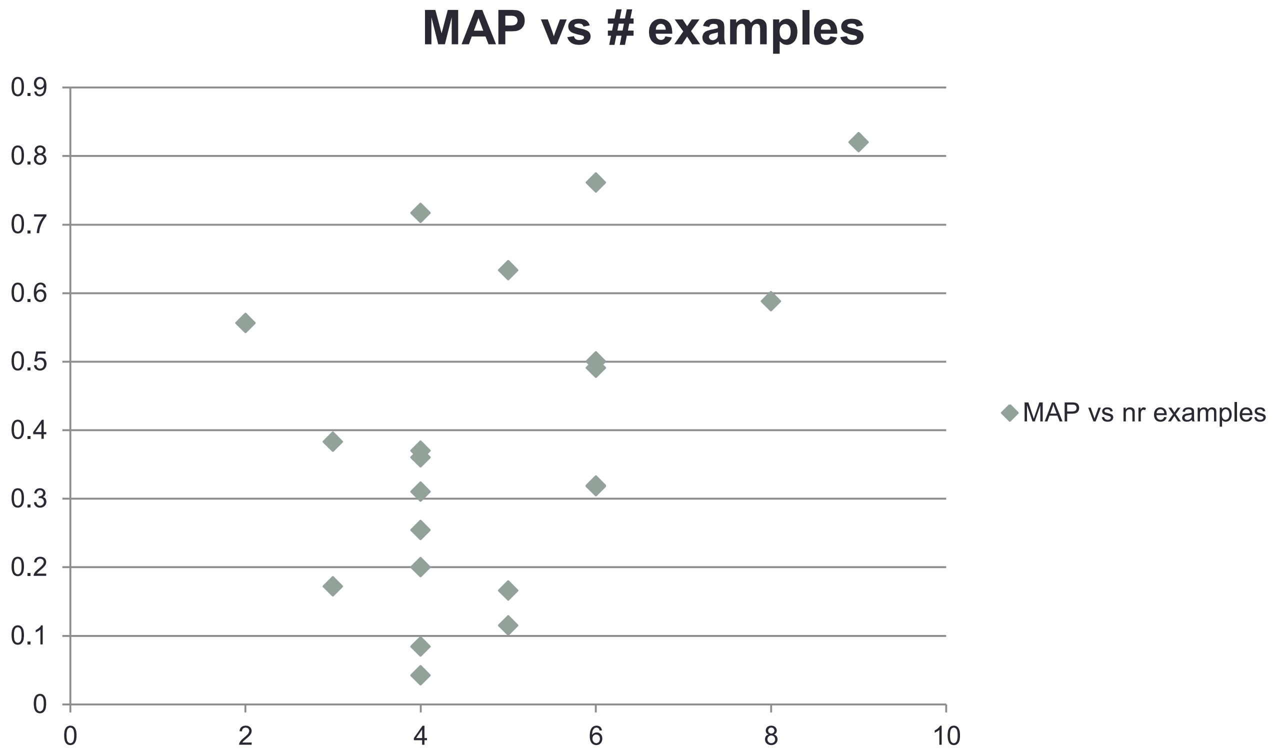


2011 Evaluation – top automatic vs interactive



Possible factors for query difficulty(1)

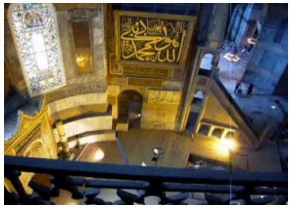
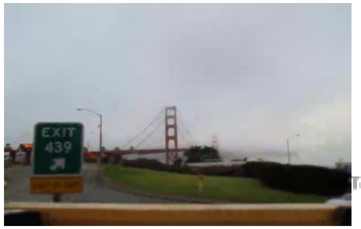
- Nr of sample images
- Pearson correlation 0.4



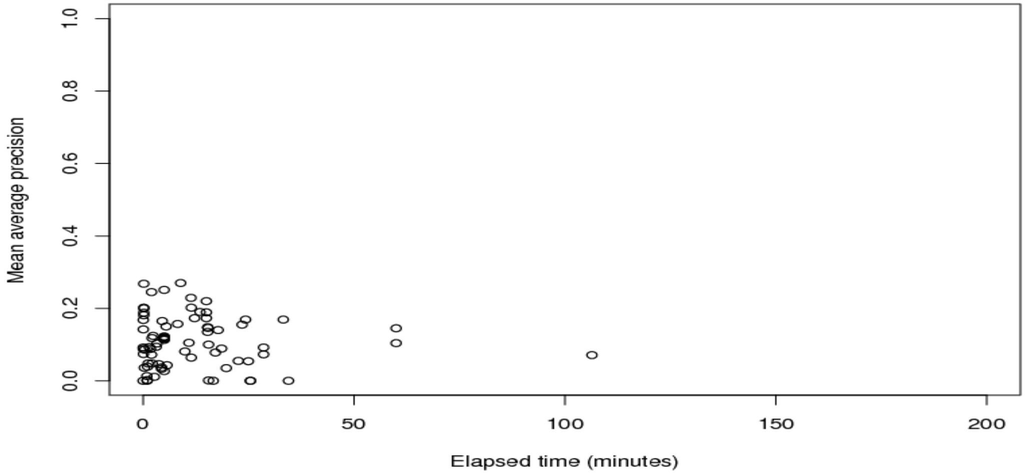
Possible factors for query difficulty(2)

- Easy topics
  - Whole frame
  - Simple background
  - Interior shots (constant illumination)

- Difficult topics
  - Small focus (ROI)
  - Complex background

Evaluation – time vs. effectiveness



- Ranges from 6 sec to 87 hours / topic
- Runs with subminute processing speed & map> 0.15:
- **BUPT 3** (0.27;0.1): Very rich combination of local (SIFT), regional (e.g. PHOG: capturing spatial layout) and global features, linear fusion, pseudo feedback
- **Vireo 1,2,3,4** (0.18-0.20;0.2): SIFT BOVW (100K), spatial consistency postfiltering. Inverted file contains all information necessary for postfiltering.

Overview of submissions

- All submissions use local descriptors, most BOVW
- A large variety of exploratory experiments with different objectives
- 18 out of 24 INS teams submitted a paper
- Main team experiments have been grouped by a number of themes
- Presentations by Univ Chile, NTT-NII and JRS

- Some teams did per topic error analysis (e.g. JRS)
- Some teams evaluated a TV11 system on TV12 data (e.g. NII)

Reusing techniques from text IR

- INS resembles an ad-hoc task in visual feature space
- Dimension reduction using visual words (1K -1 M)
- Inverted files for fast lookup (Lucene)
- Feature weighting: BM25, tf.idf, RSJ weights (NTT-NII)
- Pseudo relevance feedback
  - BUPT-MCPRL (not clear how effective)

System architecture & Efficiency

- Ad hoc search Pre-index all clips in a collection-defined feature space, analyse query in this space, rank the clips.
  - 1. All local features; 2. BOVW: ; SOM
- run-time query specific classifier Analyse query, enrich using external data, define query specific feature space. Rank clips according to this space
  - 3. local features for sample images
  - 4. rerank with internet sample image based classifier
- Teams: AXES, DCU IAD, JRS (3>>2), UvA, NII, NTT(3), IMP (1:hash based appr. NN), PRISMA (1: parallel approx NN), TNO (1: FLANN>> 2), UC SB& Riverside (2,4), Vireo (CityUHK) (2)

Dealing with query info

- How to exploit the mask (focus vs background)
  - UvA: fusion helps
  - Vireo: background context modelling (blurring context), helps
- Adding extra sample images from internet sources
  - AXES, PKU ICST
- Enlarging query samples
  - JRS, TNO: no increase
- Dealing with different samples
  - Early vs late fusion
  - Vireo: "video level fusion" helps
- Using type information
  - Nobody?

Finding an optimal representation

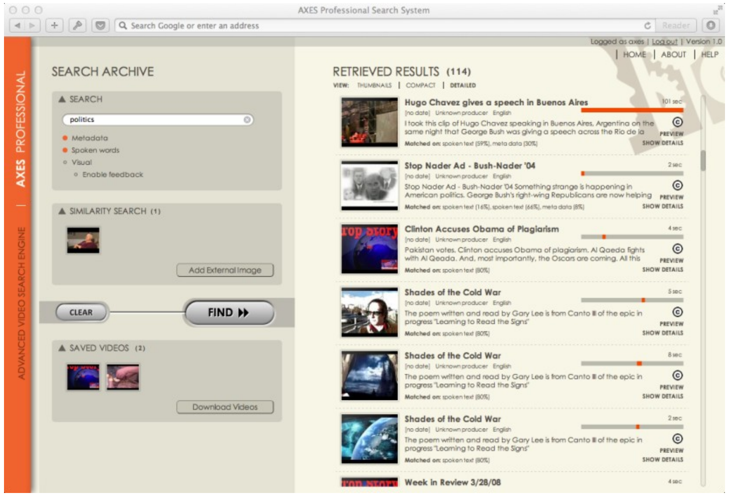
- Comparing different feature types
  - CEA: BOVW, HSV hist
  - Sheffield+Harbin: PHOW, SIFT
- Fusing many different descriptors
  - BUPT: HSV_hist, RGB_moment, SIFT, C_SIFT, Gabor, EDH, LBP, PHOG, HOG
- How to combine the features (fusion experiments)
  - CEA: **descriptor-first** vs query-first
  - IRIM: Fusing results of several labs, no significant difference between fusion strategies
  - JRS: fusion of SIFT and CSIFT runs (densely sampled vs Difference of Gaussian points) (fusion did not help overall)
  - Sheffield+Harbin: Battacharya vs Eucledian vst fidf

How to exploit spatial constraints

- BOVW approaches drop spatial information regarding local descriptors
- Postfiltering techniques:
  - Mediamill: spatial filtering helped for 7 topics, hurt others
  - DFKI: Hough refinement (checking scale and orientation of matched descriptors): “important increment”
  - Picsom(Aalto): pairwise matching of local descriptors (helped)
  - PKU ICST: 1. keypoint matching, 2. re-ranking by clustering top results and weeding out the outliers (good increment)
  - Vireo: 1. standard weak geometric consistency checking (WGC), 2. Delaunay Triangulation 3. region version of DT (all help)

Interactive experiments

- AXES (4 runs)
  - Fusion of subsystems: ASR, Google image based visual model, face recognition, object/location retrieval (all query-time)
  - **Tabbed** vs untabbed, **FB** or no FB
- PKU ICST (Peking Univ.)
  - 2000 visual words (SIFT), retrieve 1000 clips using multibag SVM, annotate 25 clips, retrain SVM, rerank (only 1 interactive run)
- France telecom (no description)



Hugo Chavez gives a speech in Buenos Aires 101 sec
[no date] Unknown producer English
I took this clip of Hugo Chavez speaking in Buenos Aires, Argentina on the same night that George Bush was giving a speech across the Rio de la
Matched on: spoken text [59%], meta data [30%]

Stop Nader Ad - Bush-Nader '04 2 sec
[no date] Unknown producer English
Stop Nader Ad - Bush-Nader '04 Something strange is happening in American politics. George Bush's right-wing Republicans are now helping
Matched on: spoken text [16%], spoken text [66%], meta data [8%]

Clinton Accuses Obama of Plagiarism 4 sec
[no date] Unknown producer English
Pakistan votes. Clinton accuses Obama of plagiarism. Al Qaeda fights with Al Qeada. And, most importantly, the Oscars are coming. All this
Matched on: spoken text [80%]

Shades of the Cold War 5 sec
[no date] Unknown producer English
The poem written and read by Gary Lee is from Canto III of the epic in progress "Learning to Read the Signs"
Matched on: spoken text [80%]

Shades of the Cold War 8 sec
[no date] Unknown producer English
The poem written and read by Gary Lee is from Canto III of the epic in progress "Learning to Read the Signs"
Matched on: spoken text [80%]

Shades of the Cold War 2 sec
[no date] Unknown producer English
The poem written and read by Gary Lee is from Canto III of the epic in progress "Learning to Read the Signs"
Matched on: spoken text [80%]

Week in Review 3/28/08 4 sec

Three pilot years for INS

- 2010: Sound and Vision data
  - Very low map figures
  - Resolution of many target objects was too low
  - Query type specific approaches
- 2011: Rushes data
  - More encouraging results
  - Part of the increased results maybe due to doubling the collection using CCD transformations
  - Decreased use of type specific approaches
- 2012: Flickr data
  - More realistic results
  - Some consolidation in successful approaches
- 2013: next slide

INS 2013 plans

- 464 hours (5 years) of the BBC EastEnders television series
  - MPEG-4
  - Closed-captioning text
  - Some metadata

- Made available by the BBC in collaboration with the EU AXES program for research in TRECVID

- Represents a “small world” with a slowly changing set of:
  - People (several dozen)
  - Locales: homes, workplaces, pubs, cafes, open-air market, pets,
  - Objects: clothes, cars, holdhold goods, personal possessions, etc

- Seen
  - from various viewpoints
  - in various combinations

INS 2013 plans

Possible topic types might include the following (where targets are identified only by the example images in the topic)

Find all shots with person X
Find all shots with locale Y
Find all shots with object Z

Find all shots with person X in local Y
Find all shots with person X1 and person X2
Find all shots with person X and object Z

Find all shots with Person X engaged in activity W

Find all shots with person X and person Y, talking/walking/arguing/dancing/making physical contact/... with eachother

. . .

INS 2013 plans

No training data provided

Participants may use publicly available EastEnder-specific and non-EastEnder-specific resources, as long as they

- notify NIST immediately so other participants can be made aware
- report use in workshop notebook paper/slides

Questions / Remarks for Discussion

- How can we measure progress?
- How can we structure the task & report template to maximize learning?
- How can we add temporal (video) aspects in the task design?
- INS might be a good track to re-introduce a subtask on localization, temporal and/or spatial