TRECVID 2011 INSTANCE RETRIEVAL PILOT

AN INTRODUCTION ….

Wessel Kraaij
TNO, Radboud University Nijmegen

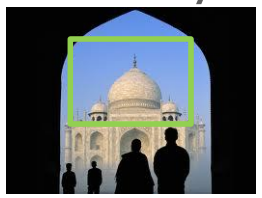
Paul Over
NIST

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)

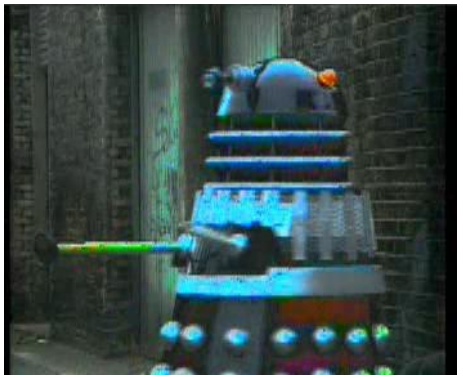



Differences between INS and SIN

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.

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, CHARACTER, PLACE, OBJECT)
- Return a list of up to 1000 shots ranked by likelihood that they contain the topic target

Data …

BBC rushes video – mostly travel show material, containing **recurring**

- objects
- people
- locations

All videos were chopped into 20 to 10s clips using ffmpeg, yielding

- 10 491 original short clips

Topics were created at NIST by

- watching most of the videos in fast forward mode
- noting repeated objects, persons, locations
- key difference with 2010: more true positives in collection, fewer "small" targets

Data – to increase the number of test clips

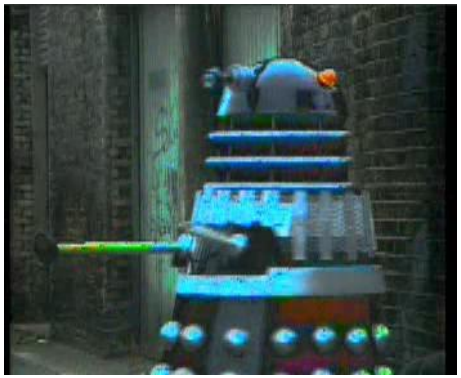
- 4 transformations were selected to mimic alternate image capture
  - **G** - Gamma: range = 0.3 : 1.8
  - **C** - Contrast: brightness-range = -20 : 20, contrast-range = -30 : 30
  - **A** - Aspect ratio: ratio-range = 0.5 : 2
  - **H** - Hue: hue-range = -20 : 20, saturation-range = 1 : 2

- 3 out of the 4 were chosen randomly for each original clip and all 3 applied to produce a transformed clip

- All original clips + transformed clips were renamed (1.mpg to 20982.mpg) to create the test collection

Data – example keyframes from test clips

Frames from originals


Trans-formations: **- C A H** **G C A -** **G – A H**

Frames From alterred originals



Data

Did systems generally recognize the original-transformed clip pairs as such or treat each clip independently?

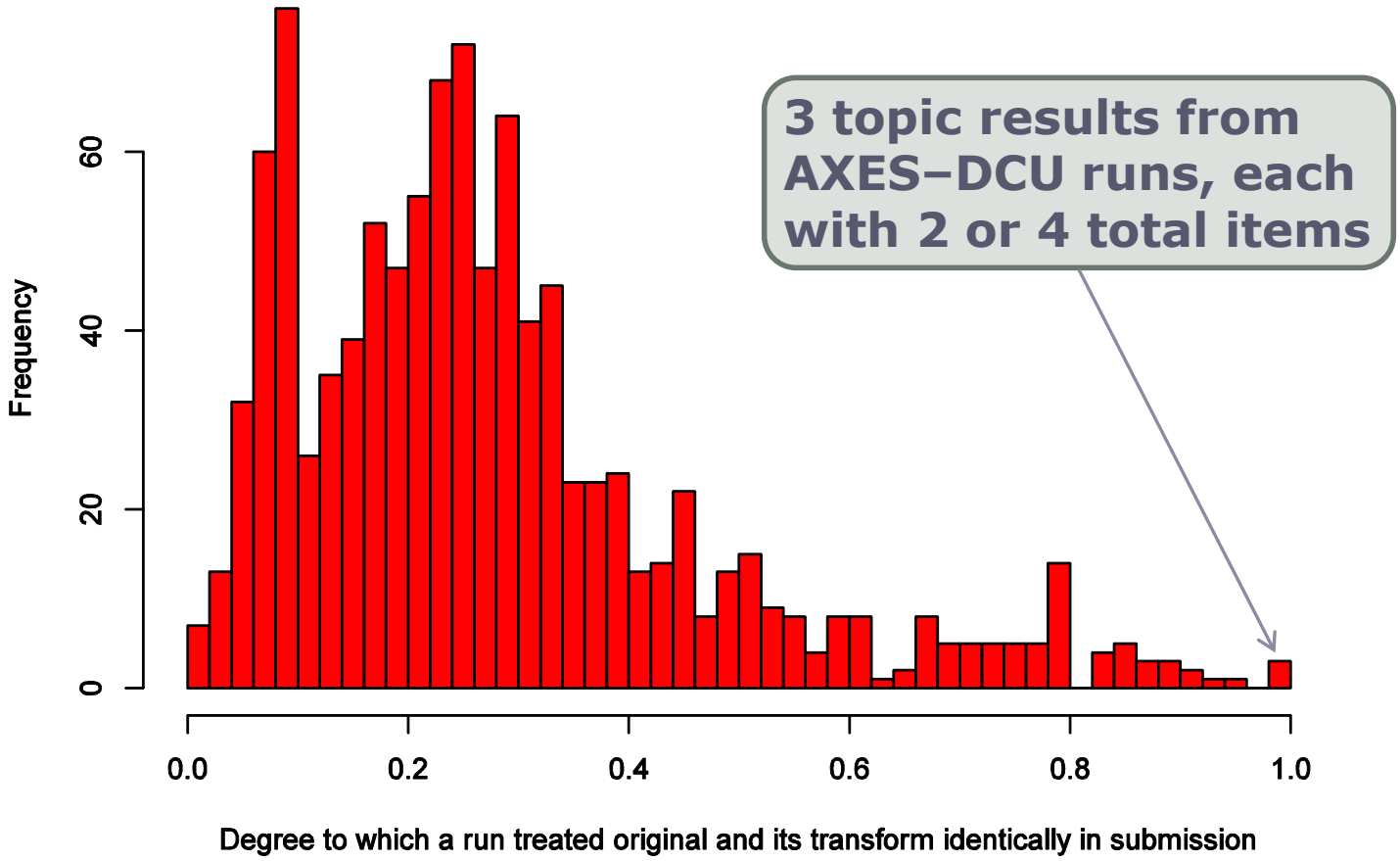
- For each topic we calculated the ratio of
  - clips where both the original and its transform were submitted

    to
  - total submitted clips

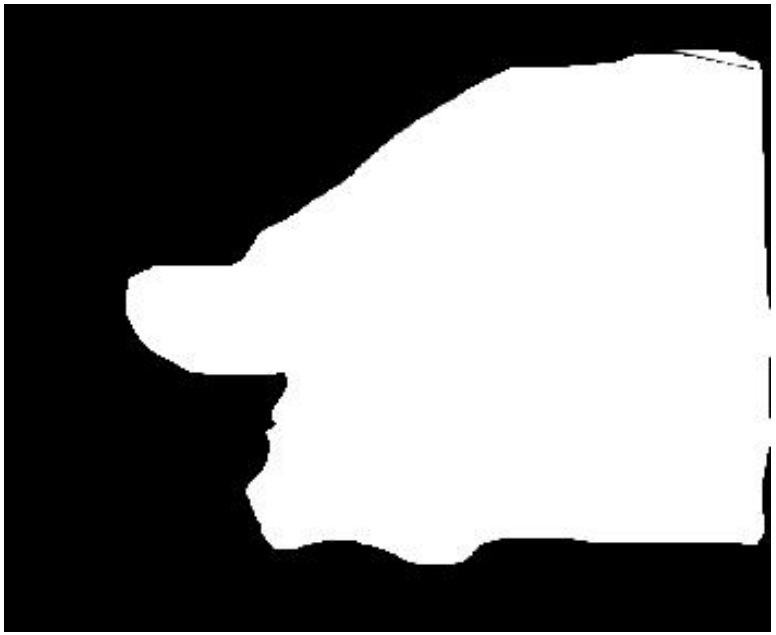
- If systems treated each each original and its transform identically then the ratio should == 1

Data

Original clip and transform generally treated differently



Topics – segmented example images



Topics – 6 People/characters

Topic# # of examples

Female presenter X

Carol Smiley

Tony Clark's wife

Linda Robson

Male presenter Y

Grey-haired lady

Topics – 17 Objects

airplane-shaped balloon

lantern

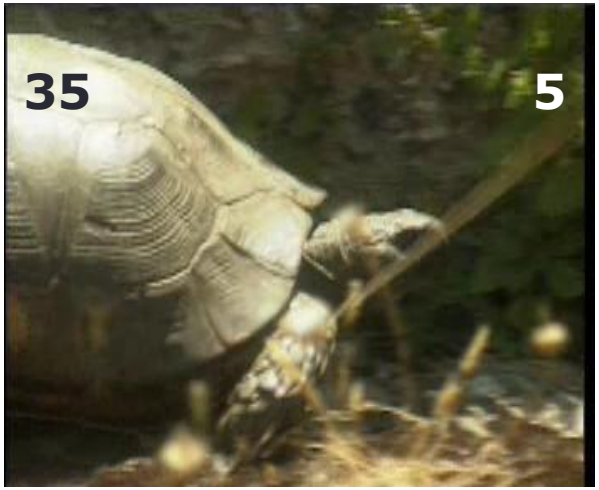
US flag

monkey

windmill from outside

all-yellow balloon

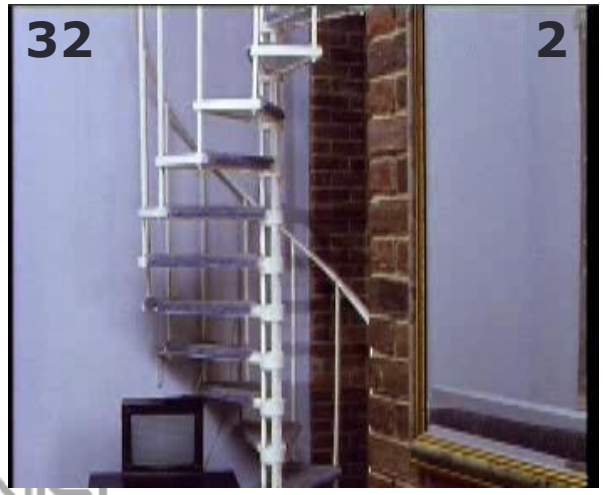
Topics – 17 Objects (cont.)



tortoise

cylindrical building

newsprint balloon



spiral staircase

the Parthenon

yellow dome with clock

National Institute of Standards and Technology

Topics – 17 Objects (cont.)

plane flying

SUV

trailer

fork

setting sun

Topics – 2 Locations

upstairs in the windmill

downstairs in the windmill

TV2011 Finishers

AXES-DCU	Access to Audiovisual Archives
ATTLabs	AT&T Labs Research
BUPT-MCPRL	Beijing University of Posts and Telecommunications-MCPRL
VIREO	City University of Hong Kong
FIU-UM	Florida International University
ARTEMIS-Ubimedia	Institut TELECOM SudParis, Alcatel-Lucent Bell Labs France
CAUVIS-IME-USP	Instituto de Matematica e Estatistica - USP
JRS-VUT	JOANNEUM RESEARCH and Vienna University of Technology
IRIM	Laboratoire d'Informatique de Grenoble
NII	National Institute of Informatics
TNO	Netherlands Organisation for Applied Scientific Research
NTT-NII	NTT Communication Science Laboratories-NII
tokushima_U	Tokushima University

Evaluation

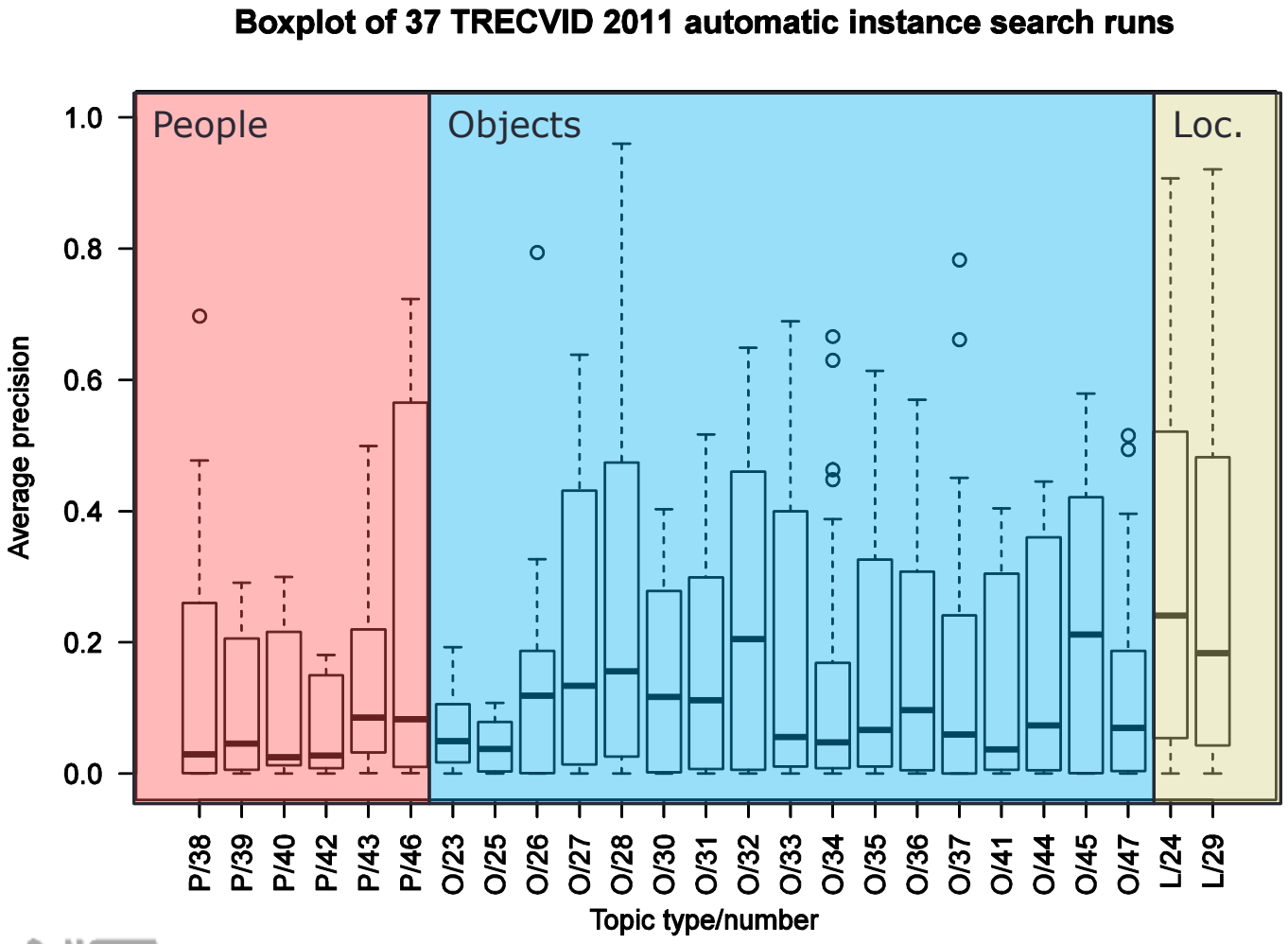
For each topic, the submissions were pooled and judged down to at least rank 100 (on average to rank 252), resulting in 114,796 judged shots.

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

1830 clips (avg. 73.2 / topic) contained the topic target.

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

Evaluation – results by topic/type - automatic



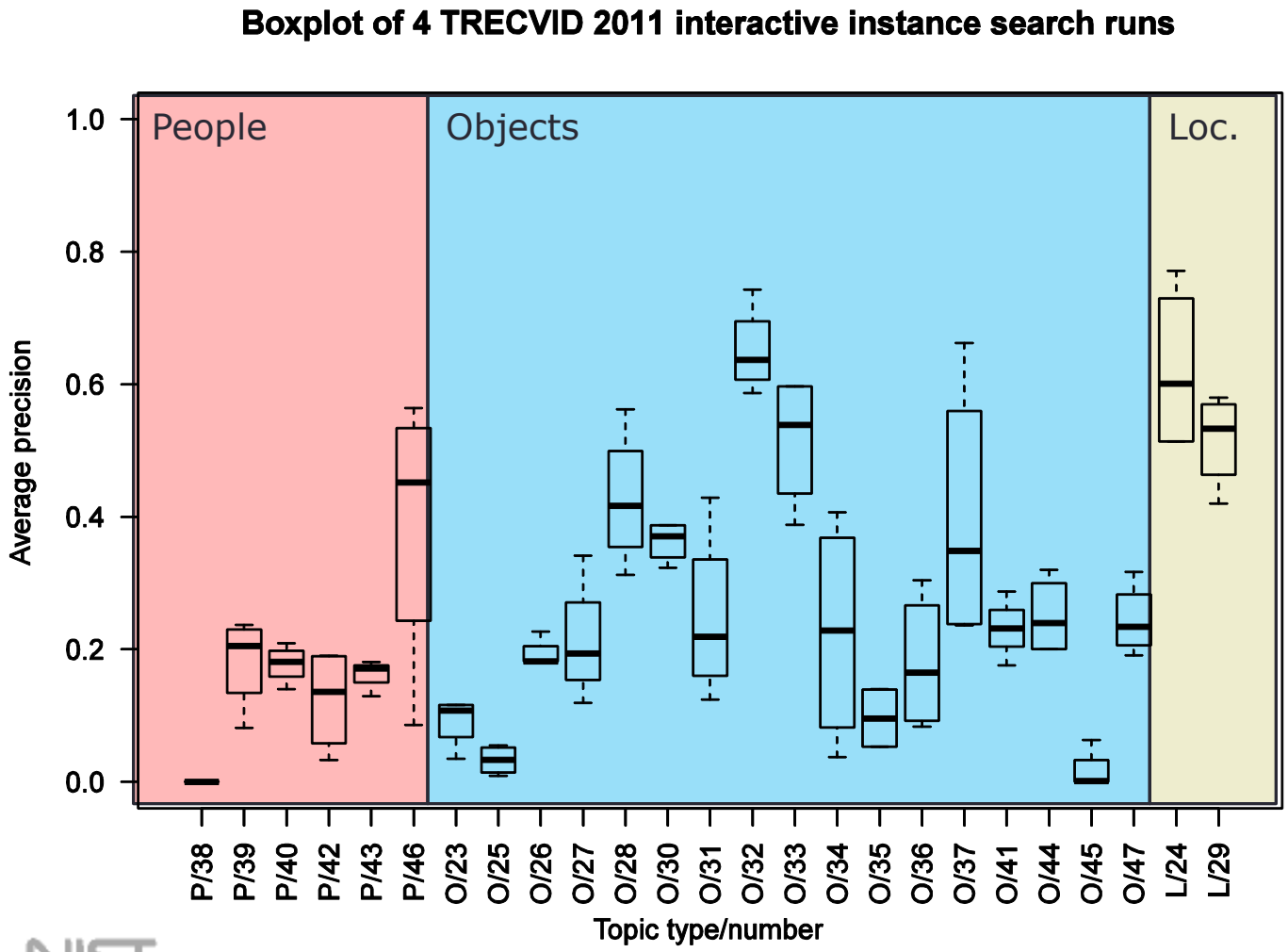
Type/# Name [clips with target]

P/38 Female presenter X [21]
P/39 Carol Smilie [34]
P/40 Linda Robson [43]
P/42 Male presenter Y [84]
P/43 Tony Clark's wife [287]
P/46 grey-haired lady [139]

O/23 setting sun [86]
O/25 fork [105]
O/26 trailer [22]
O/27 SUV [32]
O/28 plane flying [64]
O/30 yellow dome with clock [177]
O/31 the Parthenon [31]
O/32 spiral staircase [49]
O/33 newsprint balloon [45]
O/34 tall, cylindrical building [27]
O/35 tortoise [57]
O/36 all-yellow balloon [48]
O/37 windmill seen from outside [70]
O/41 monkey [108]
O/44 US flag [25]
O/45 lantern [28]
O/47 airplane-shaped balloon [70]

L/24 upstairs, in the windmill [109]
L/29 downstairs, in the windmill [69]

Evaluation – results by topic/type - interactive



Type/# Name [clips with target]

P/38 Female presenter X [21]
P/39 Carol Smilie [34]
P/40 Linda Robson [43]
P/42 Male presenter Y [84]
P/43 Tony Clark's wife [287]
P/46 grey-haired lady [139]

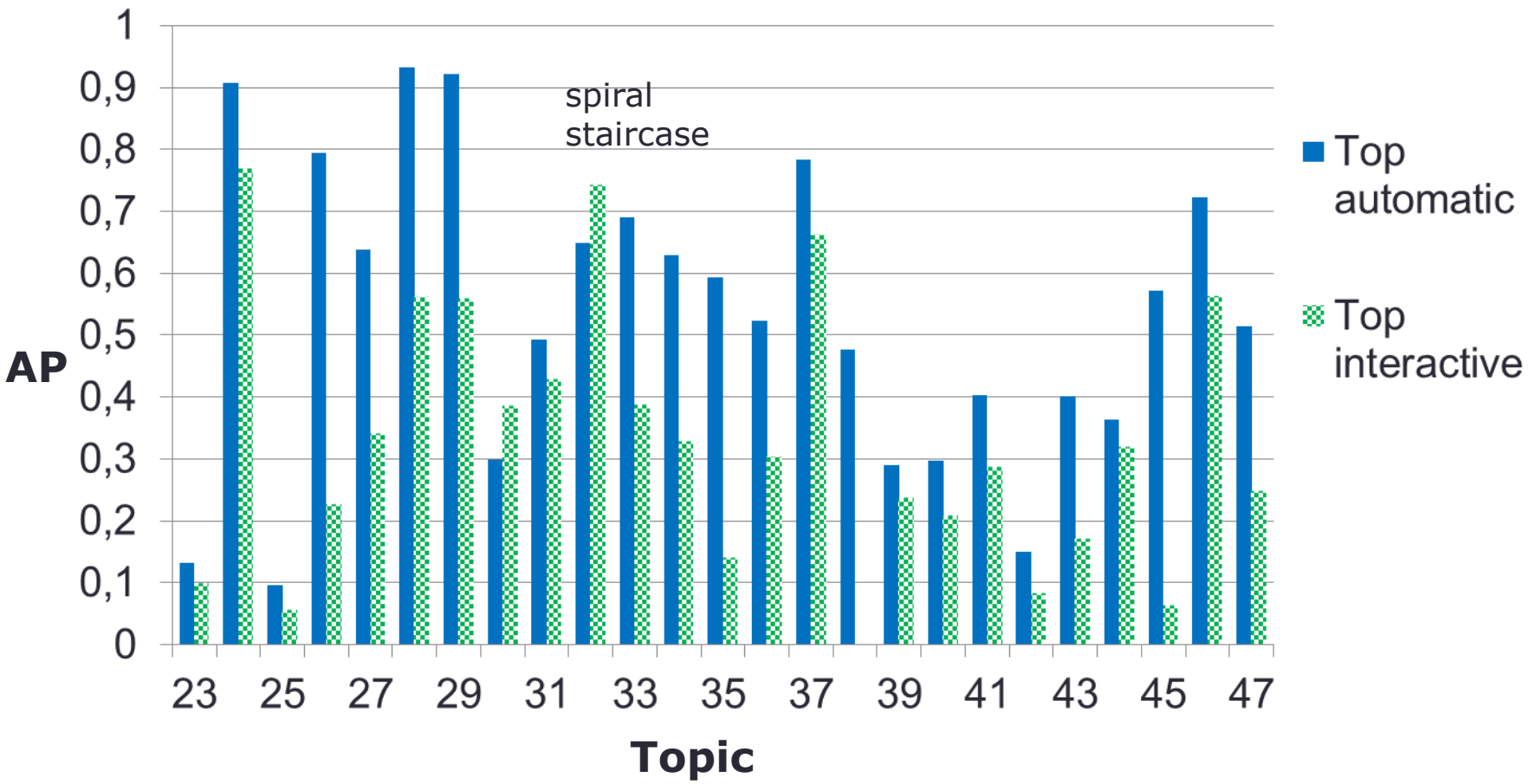
O/23 setting sun [86]
O/25 fork [105]
O/26 trailer [22]
O/27 SUV [32]
O/28 plane flying [64]
O/30 yellow dome with clock [177]
O/31 the Parthenon [31]
O/32 spiral staircase [49]
O/33 newsprint balloon [45]
O/34 tall, cylindrical building [27]
O/35 tortoise [57]
O/36 all-yellow balloon [48]
O/37 windmill seen from outside [70]
O/41 monkey [108]
O/44 US flag [25]
O/45 lantern [28]
O/47 airplane-shaped balloon [70]

L/24 upstairs, in the windmill [109]
L/29 downstairs, in the windmill [69]

Evaluation – top half, based on MAP

Automatic			MAP	Interactive			MAP
F X N	NII.Caizhi.HISimZ	4	0.531				
F X N	NII.Caizhi.HISim	3	0.491				
F X N	MCPRBUPT1	1	0.407				
F X N	MCPRBUPT2	2	0.353				
F X N	NII.SupCatGlobal	1	0.340				
F X N	MCPRBUPT3	3	0.328				
F X N	TNO-SURFAC2	1	0.325	I X N	AXES_DCU_1	1	0.327
F X N	vireo_f	1	0.312				
F X N	vireo_b	2	0.309				
F X N	vireo_s	3	0.299				
F X N	vireo_m	4	0.295				
F X N	TNO-SUREIG	3	0.274				
F X N	IRIM_1	1	0.274				
F X N	IRIM_3	3	0.259	I X N	AXES_DCU_2	2	0.265
F X N	IRIM_4	4	0.251				
F X N	JRS_VUT	4	0.170	I X N	AXES_DCU_3	3	0.250
F X N	IRIM_2	2	0.166	I X N	AXES_DCU_4	4	0.206
F X N	NII.Chanseba	2	0.115				
F X N	JRS_VUT	3	0.104				

Evaluation – top automatic vs interactive



Evaluation – randomization test results on top 10

Automatic

F X N NII.Caizhi.HISimZ 4

F X N NII.Caizhi.HISim 3

F X N MCPRBUPT1 1

F X N MCPRBUPT2 2
F X N NII.SupCatGlobal 1
F X N MCPRBUPT3 3
F X N TNO-SURFAC2 1
F X N vireo_f 1
F X N vireo_b 2
F X N vireo_s 3

Interactive

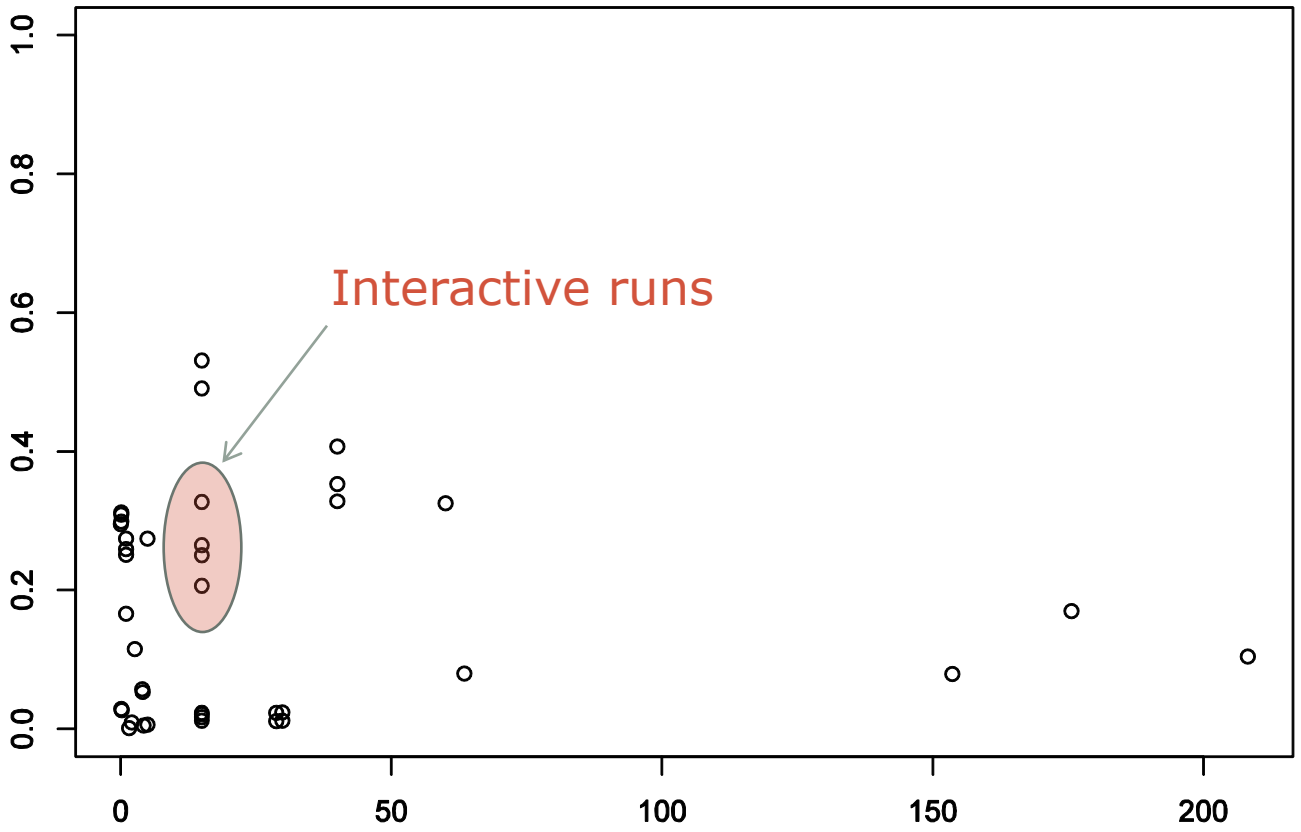
I X N AXES_DCU_1 1

I X N AXES_DCU_2 2
I X N AXES_DCU_3 3

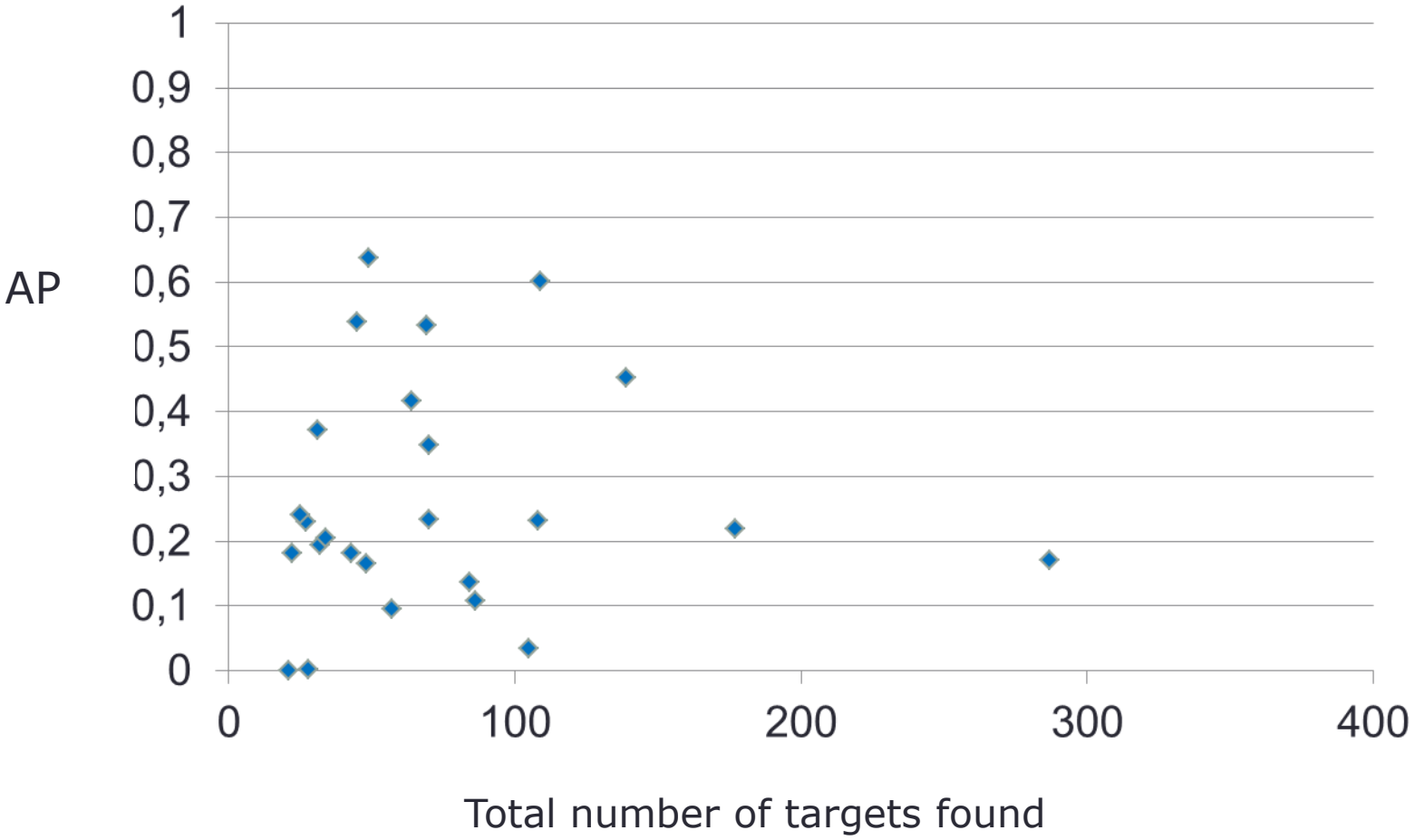
I X N AXES_DCU_4 4

The bold arrows denote significant differences

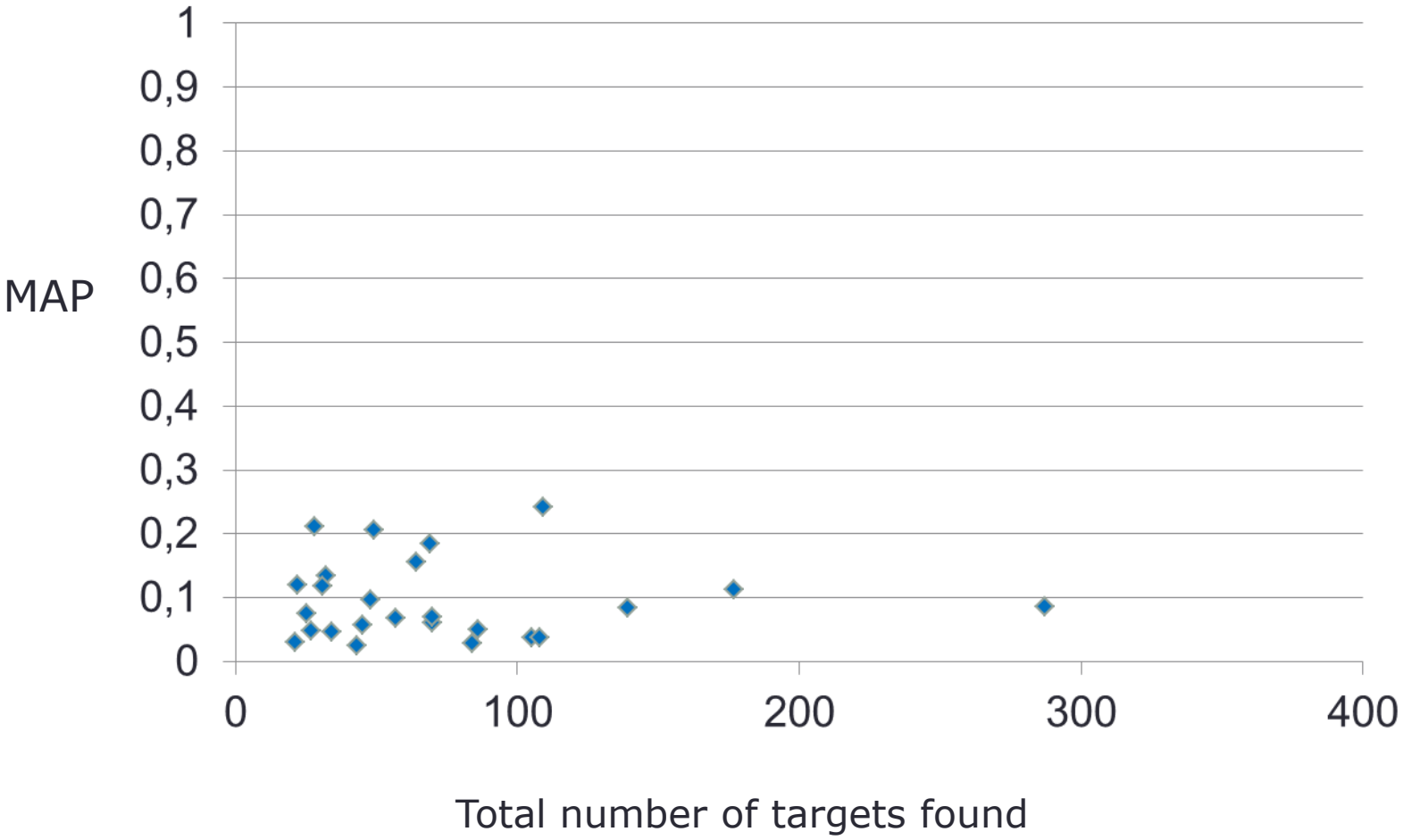
Evaluation – time vs effectiveness



Effectiveness vs. total found (interactive run median for each topic)



Effectiveness vs. total found (automatic run median for each topic)



Overview of submissions

- [TNO, NII and AXES-DCU presentations coming up]
- Three groups did not submit a notebook paper yet…

AT&T Labs Research

- INS system based on CCD system
  - Baseline run based on SURF local features
  - Normalization technique promoting matches from each query sample image near the top
- Outlier analysis
  - Weak performance for homogeneous visual characteristics (low contrast, few edges)
  - Objects small, using context hurts (balloons)
  - Objects small, using context helps (context is typical for dataset)
- Challenges identified:
  - When to include visual context features?

Beijing University of Posts and Telecommunications-MCPRL

- Features: HSV hist, RGB_moment, SIFT, SURF, CSIFT, Gabor Wavelet, Edge hist, LBP, HoG
- Higher weight for close-up shots (reranking)
- Specific normalization techniques for each modality
- Runs compare three (non specified) score merging strategies

VIREO: City University of Hong Kong

- Key differences with search task and CCD:
  - Region of interest specification
  - Wider definition of relevance than visual copies (eg person)
  - Multiple examples with varying conditions (unlike CCD)
- Approach:
  - SIFT (Lowe), BoW approach
  - One keyframe/s
- "We experimented four runs to contrast the following for instance search: **full matching** (vireo b) versus **partial matching** (vireo m); use of **weak geometric information** (vireo b) versus **stronger spatial configuration** (vireo s); use of **face matching** (vireo f)."
- No clear winning approach, performance depends on aspects such as size, context uniformity etc

Florida International University / University of Miami

- Texture features plus SIFT
- Based on Multiple Correspondence Analysis (MCA)
- Variants enhanced by
  - KNN reranking, MCA reranking, SIFT, 261 extra training images
- Question: Were the extra training images collected, selected and processed automatically?
- No real differences between variant runs.
- First participation in INS task

Instituto de Matematica e Estatistica University of Sao Paulo

- Representation: Pyramid histogram of visual words (PHOW) a variant of Dense SIFT (5 pixels distance)
- 600000 descriptors clustered into 300 visual words
- Frames represented as word frequency vector
- Similarity computation based on chi-square

- 1 run only

- Results: above median for location topics (where texture was important)

JOANNEUM RESEARCH and Vienna University of Technology

- Approach: fusion of four different techniques
  - Face detection (Viola Jones) followed by face matching (Gabor wavelets)
  - BoF (bag of visual words): codebook size 100
  - Mean Shift Segments: (color segmentation)
  - SIFT (Lowe):
- Fusion: take best result across all topic sample images for all four methods

- SIFT only run performed best, especially strong for location type

IRIM team (collaboration of LIG, Eurecom, INRIA etc etc)

- Two representations:
  - Bag of visual words (using SURF descriptors) codebook 16K
  - Bag of Regions (with HSV histogram as descriptor) codebook 2K
- Similarity
  - BoVW: complement of histogram intersection
  - BOR: L1-distance
- Limited use of the mask (only over 8 points for BOVW)
- Best results with merged BOVW/BOR and complete frame

Observations

- Many new participants (5 out of 13)
- Results provide a better baseline than last year
- Reuse of techniques from e.g. CCD
- Classical SIFT (b/w and colour) descriptors highly successful
- Difficult to automatically decide whether context should be ignored (mask), targets still too small
- Location type was easier
- Online search speed seems to be not an issue (NII)
- Indexing the collection with sufficient detail is a computational challenge

Questions / Remarks

- Possibly a 2-step interactive procedure could work?
  - First include context to find candidate shots
  - Second de-emphasize the masked image region to focus on the object in focus
- Why did only few groups do an interactive run?
  - Many systems were fast enough to allow interaction, e.g. weeding out incorrect hits.
- Which group did use the type information?
- Like last year, some participants added external training data
  - Perhaps we need to add run categories (just like (HLF/SIN)