TRECVID 2013 INSTANCE RETRIEVAL

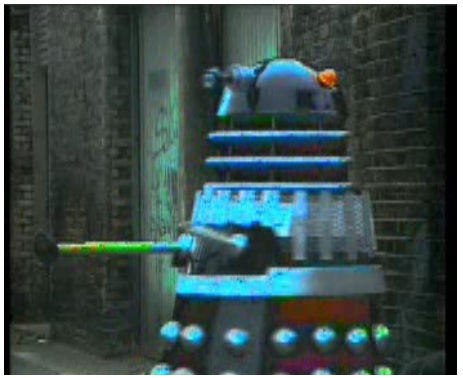
AN INTRODUCTION ….

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:
  - example segmented images of the target (4)
  - 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

Differences between INS and SIN

INS	SIN
Very few (4) training images (probably from the same clip)	Many (>> 100) 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.

INS CHALLENGE: Find objects, persons in video given a few visual examples in a few seconds

New data …

The BBC and the AXES project made **464 hours** of the BBC soap opera EastEnders available for research in **MPEG-4**

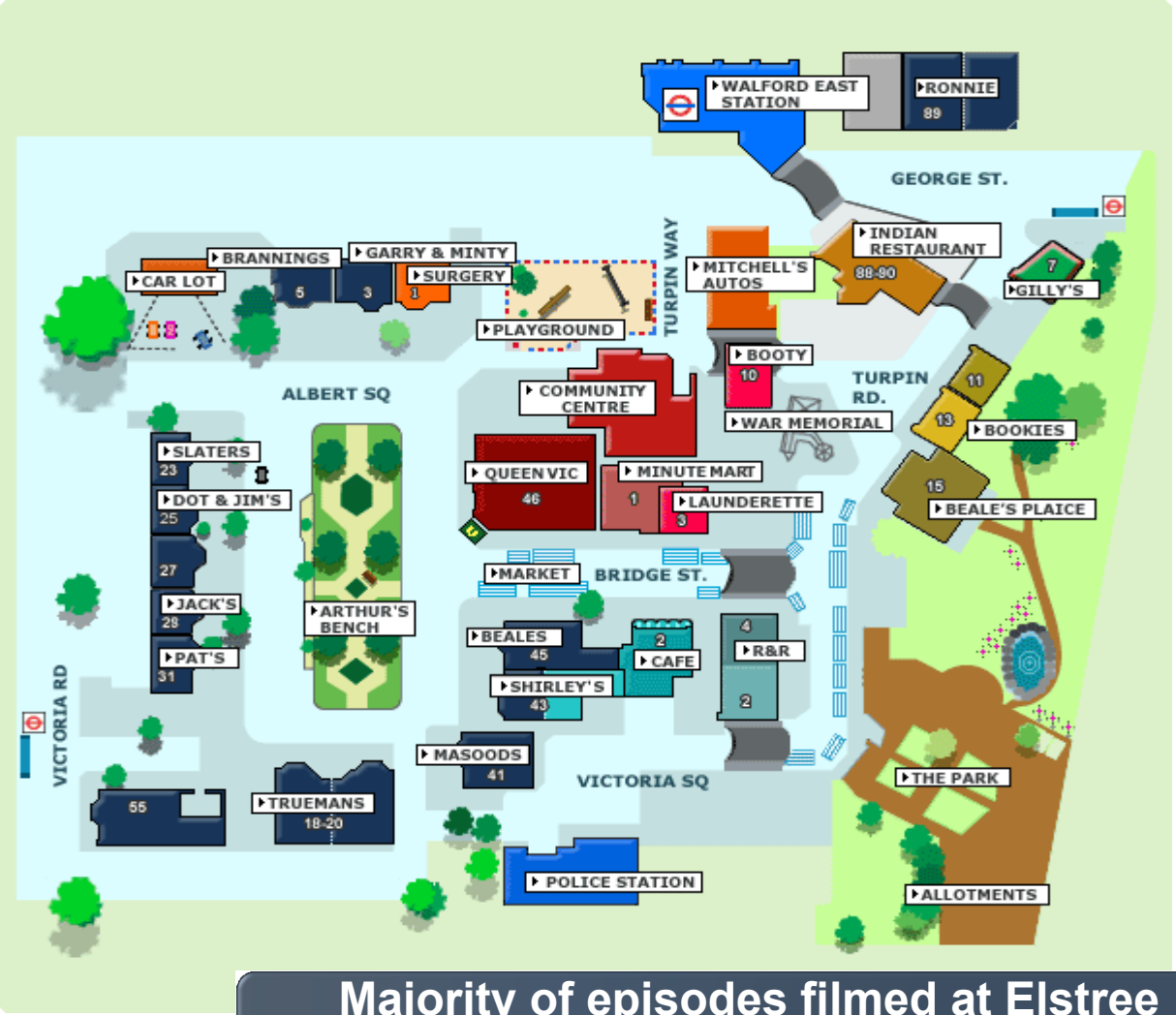
- 244 weekly "omnibus" files from 5 years of broadcasts
- 471527 shots
- Average shot length: 3.5 seconds
- Transcripts from BBC
- Per-file metadata

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 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 (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,...

Topics – segmented example images

Source

Mask

Example from TV12

Topics – 26 Objects

Topic: **True positives:**

69 **2300**

a 'no smoking' logo

70 **741**

a small red obelisk

71 **31**

an Audi logo

72 **261**

a metropolitan police logo

73 **674**

this ceramic cat face

74 **100**

a cigarette

Topics – 26 Objects (cont.)

75 82

a SKOE can

76 831

Queen Victoria bust

77 31

this dog

78 880

A JENKINS logo

79 390

this CD stand

80 251

this phone booth

Topics – 26 Objects (cont.)

81 213

a black taxi

82 61

a BMW logo

83 118

chrome/glass cafetiere

85 455

David fridge magnet

86 759

these scales

87 25

a VW logo

Topics – 26 Objects (cont.)

89 1266

this pendant

90 363

this wooden bench

91 782

a menu with stripes

93 75

these turnstiles

94 171

a tomato ketchup dispenser

95 440

a public trash can

Topics –26 Objects (cont.)

97 252

these checkerboard spheres

98 386

a P (parking automat) sign

Topics – 4 Persons

84 32

this man

88 1605

Tamwar

92 171

this man

96 161

Aunt Sal

INS 2013: 22 Finishers (tv12:24)

CEALIST	CEA LIST, Vision & Content Engineering Laboratory
IRIM	CEA-LIST,ETIS,EURECOM,INRIA-TEXMEX,LABRI,LIF,LIG,LIMSI-TLP,LIP6,LIRIS,LISTIC,CNAM
VIREO	City University of Hong Kong
AXES	**Access to Media**
iAD_DCU	Dublin City University University of Tromso
ITI_CERTH	**Information Technologies Institute, Centre for Research and Technology Hellas**
ARTEMIS	Institut Mines-Telecom; Telecom SudParis; ARTEMIS Department
JRS	JOANNEUM RESEARCH Forschungsgesellschaft mbH
BUPT_MCPRL	Multimedia Communication and Pattern Recognition Labs
MIC_TJ	Multimedia and Intelligent Computing Lab, Tongji University
NII	National Institute of Informatics
NTT_NII	NTT, NII
ORAND	**ORAND S.A. Chile**
FTRDBJ	**Orange Labs International Centers China**
IMP	Osaka Prefecture University
PKU-ICST	**Peking U.-ICST**
TNO_M3	TNO
TokyoTechCanon	Tokyo Institute of Technology Canon Inc.
thu.ridl	Tsinghua University School of Software, Department of Computer Science and Technology
sheffield	U. of Sheffield, UK Harbin Engineering Univ, PRC U. of Engineering & Technology (Lahore)
MediaMill	University of Amsterdam
NERCMS	Wuhan University

RED 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 253, max 460), resulting in 209,302 judged shots (~ 600 person-hrs).

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

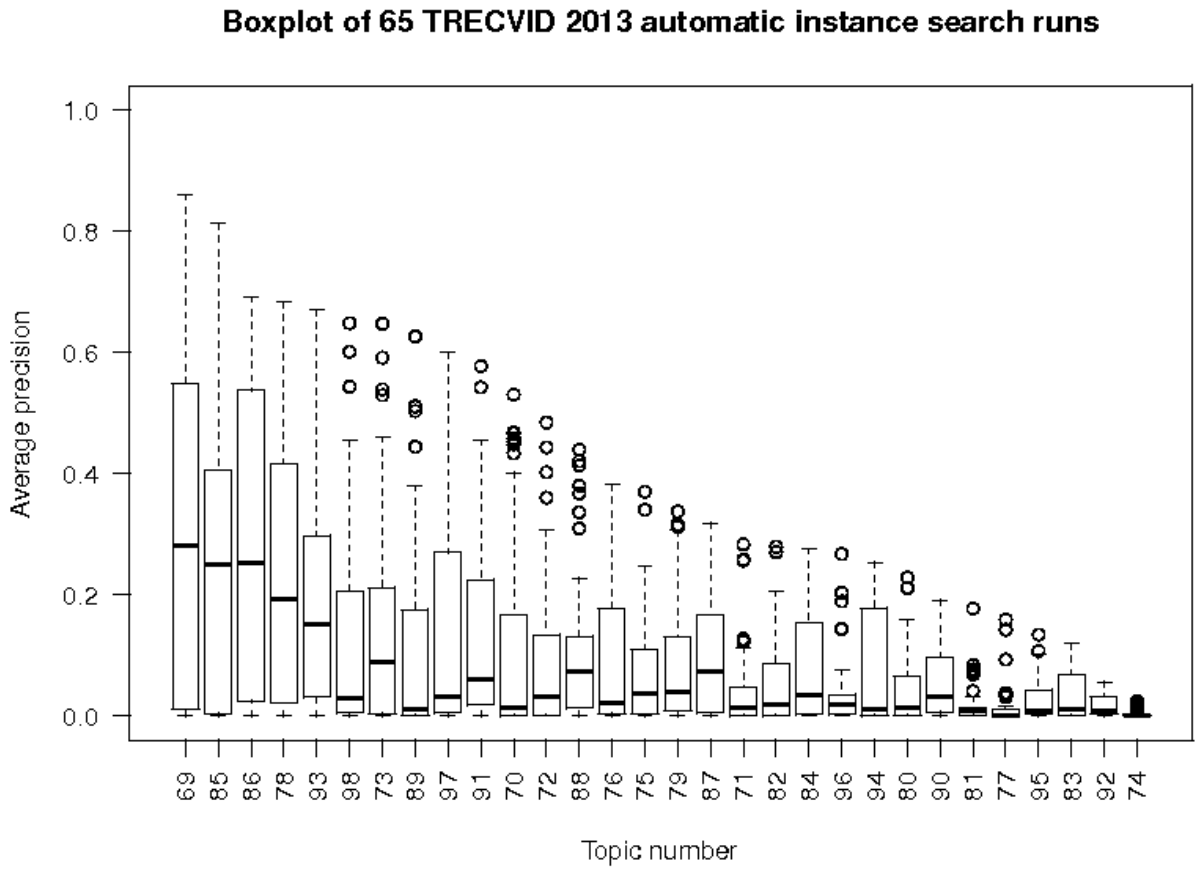
13907 clips (avg. 463.6 / topic) contained the topic target (6.6%)

True positives per topic: min 25 med 256.5 max 2300

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

➔ New INS run notebook pages are available in the active participants area.

Evaluation – results by topic - automatic



 # Name [clips with target]

Objects with single location in blue

69 a no smoking logo
85 this David magnet
86 these scales
78 a Jenkins logo
93 these turnstiles
98 a P (parking automat) sign
73 this ceramic cat face
89 this pendant
97 these checkerboard spheres
91 a Kathy's menu with stripes
70 a small red obelisk
72 a Metro Police logo
88 Tamwar
76 this monochrome bust of Victoria
75 a SKOE can
79 this CD stand in the market
87 a VW logo
71 an Audi logo
82 a BMW logo
84 this man
96 Aunt Sal
94 tomato-shaped ketchup bottle
80 this public phone booth
90 this wooden bench
81 a black taxi
77 this dog
95 a green public trash can
83 a chrome and glass cafetierre
92 this man
74 a cigarette

Evaluation – top 10, based on MAP

Automatic	MAP
NII-AsymDis_Cai-Zhi_2	0.313
NTT_NII_3	0.297
NII-AvgDist_Cai-Zhi_3	0.276
NII-GeoRerank_Cai-Zhi_1	0.256
NTT_NII_2	0.256
NTT_NII_1	0.237
PKU-ICST-MIPL_1	0.212
PKU-ICST-MIPL_3	0.200
PKU-ICST-MIPL_4	0.198
NTT_NII_4	0.198

Randomization test

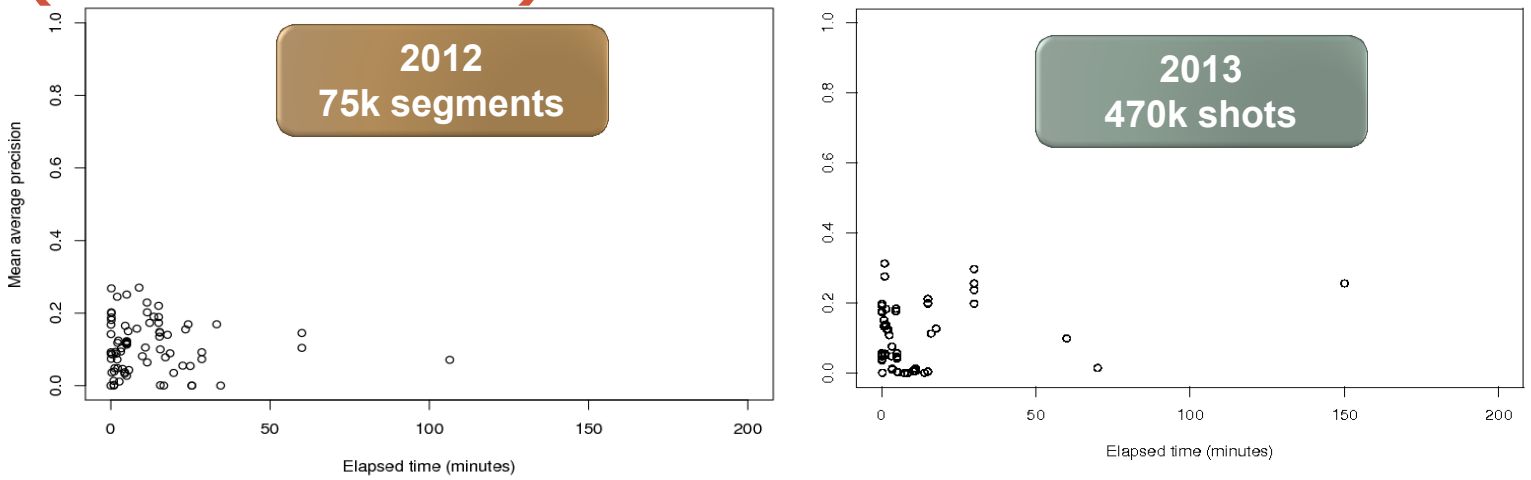
NII-AsymDis_Cai-Zhi_2 > NII-AvgDist_Cai-Zhi_3
> NTT_NII_4
> PKU-ICST-MIPL_1
> PKU-ICST-MIPL_4
> PKU-ICST-MIPL_3
> NII-GeoRerank_Cai-Zhi_1
> NTT_NII_4
> NTT_NII_1
> NTT_NII_4

NTT_NII_3 > NTT_NII_1
> NTT_NII_2
> NTT_NII_4
> PKU-ICST-MIPL_1
> PKU-ICST-MIPL_4
> PKU-ICST-MIPL_3

NTT_NII_2 > NTT_NII_4
> PKU-ICST-MIPL_3
> PKU-ICST-MIPL_4

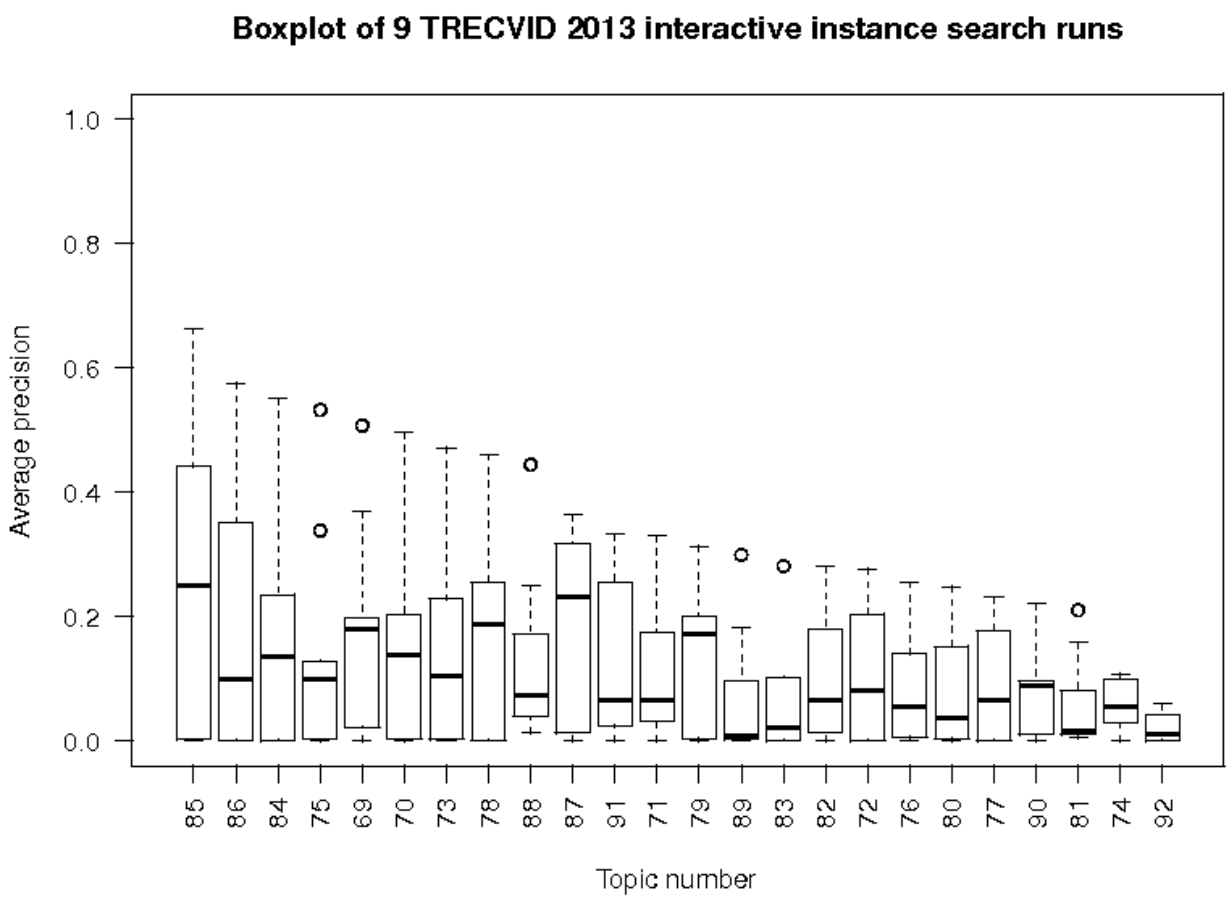
">" denotes statistically significant differences

MAP vs. query processing time (automatic)



- Ranges from 6 sec (0.1min) to 23 days/ topic
- Runs with <=1min processing speed & map=> 0.2:
- **NII**
  - *1M vwords, late fusion of 6 features, query adaptive similarity, aggregated feature vector for each clip, inverted file for speed up*
  - F_NO_NII-AsymDis_Cai-Zhi_2 (map=0.31;1min) *asymmetric similarity,*
  - F_NO_NII-AvgDist_Cai-Zhi_3 (map=0.28;1min)
- **Vireo**
  - F_NO_vireo_dtc_1 (map=0.2; 0.1min) *SIFT BOVW (250K), background context weighting strategy (stare), (quite similar to 2012 run)*

Evaluation – results by topic - interactive



 # Name [clips with target]

85 this David magnet
86 these scales
84 this man
75 a SKOE can
69 a no smoking logo
70 a small red obelisk
73 this ceramic cat face
78 a Jenkins logo
88 Tamwar
87 a VW logo
91 a Kathy's menu with stripes
71 an Audi logo
79 this CD stand in the market
89 this pendant
83 a chrome and glass cafetierre
82 a BMW logo
72 a Metro Police logo
76 this monochrome bust of Victoria
80 this public phone booth
77 this dog
90 this wooden bench
81 a black taxi
74 a cigarette
92 this man

Objects with single location in blue

Evaluation – all, based on MAP

Interactive	MAP
FTRDBJ_4	0.296
PKU-ICST-MIPL_2	0.245
orand-interactive_2	0.215
AXES_1_1	0.135
AXES_3_3	0.086
AXES_2_2	0.079
ITI_CERTH_2	0.009
ITI_CERTH_1	0.006
ITI_CERTH_3	0.005

Randomization test

FTRDBJ_4 > orand-interactive_2
> AXES_1_1
> AXES_2_2
> AXES_3_3
> ITI_CERTH_1
> ITI_CERTH_2
> ITI_CERTH_3

PKU-ICST-MIPL_2 > AXES_1_1
> AXES_2_2
> AXES_3_3
> ITI_CERTH_1
> ITI_CERTH_2
> ITI_CERTH_3

">" denotes statistically significant differences

Possible factors for query difficulty

- Easy topics
  - Simple visual context
  - Stationary target
  - Planar, rigid objects

- Difficult topics
  - Small target (ROI)
  - Moving target: differences in camera angle, location
  - Non planar, non rigid

Overview of submissions

- 17 out of 22 INS teams described INS runs for notebook

- All systems use some form of SIFT local descriptors
  - Large variety of experiments adressing representation, fusion or efficiency challenges

- Talks:
  - NII - National Insititute of Informatics ,Japan
  - Vireo – City University of Hong Kong
  - NTT-NII – Nippon Telegraph and Telephone Corp., NII

Typical INS template system

- Processing clips
  - Keyframe choice (1 per shot – 5fps)
  - Keyframe downsizing?
- Representation
  - Global (HSV, LBP,..)
  - Local
    - Detection methods (1-5)
    - Choice of descriptors (1-2)

Each design choice has an impact on speed and effectiveness

- Matching
  - #1: Object recognition based on nr of keypoint matches (Lowe), spatial verification
  - #2: BovW: clustering kp to codebook (size,hard/soft), choice of similarity function(idf weighting, ROI / background), spatial verification
- Fusion of scores

Finding an optimal representation

- Combining different feature types (local/global)
  - **CEA**: BOVW, HSV hist
  - **Sheffield/Harbin**: LBP, HOG, SIFT
  - **BUPT:** BoVW, BoVW+local
- VLAD quantization instead of BoVW: **AXES, ITI-CERTH** (VLAD>BovW)
- Combining multiple keypoint detectors and multiple descriptors
  - **NII:** Hesssian affine, Harris-Laplace, MSER // RootSIFt + C-Sift
  - **NTT:** SIFT+CSIFT

Special treatment of faces

- **AXES:** find additional faces with Google image search to extend training data
- **Orange labs Beijing:** BovW + face classifier based on "simile classifier based face descriptor": did help some topics, but slow

System architecture & Efficiency

- Object search, sequential video processing on the fly
  - **TNO**: Hadoop setup to speed up linear search
  - **JRS**: GPU based object search
  - **MIC_TJ**: Hybrid parallelization using GPU's and map/reduce

- Bag of visual words, indexed video database
  - **Most systems**: e.g. NII, NTT-NII, Vireo
  - sparse BovW, Lucene inverted file based scoring

Reusing techniques from text IR

- Inverted files for fast lookup in sparse BovW space (Lucene)
- NII:asymmetric similarity function
- Use of Collection statistics:
  - BM25 enhancements for weighting (NTT-NII): did help
  - Mining frequent cooccurring objects (VIREO)
- Pseudo relevance feedback, query expansion
  - PKU-ICST: to eliminate noisy hits
  - NTT-NII: no gains
  - IAD_DCU: helped to remove some false positives

Interactive experiments

- Orange labs Beijing (1 interactive run)
  - Interactive run significantly outperforms automatic runs (0.29 vs 0.19) "due to multiple feedback rounds"
- PKU ICST (Peking Univ.) (1 interactive run)
  - 2000 visual words (SIFT), retrieve 1000 clips using multibag SVM, annotate 50 clips, retrain SVM, rerank
  - Interactive run outperforms best automatic PKU run
- AXES (4 runs)
  - Fusion of subsystems: (metadata) closed captions, Google image based visual model, face recognition, object/location retrieval (all query-time)
  - Experiment focuses on different user types (post-docs, vs phd students)
- CERTH (3 runs)
  - VLAD quantization outperforms BovW
  - User interface benefits from scene segmentation module (linking related shots)

Some observations

- The task seems healthy after 3 pilot years
  - Stable number of participants
  - Interesting new dataset
  - Systems produce meaningful results
  - No ceiling reached yet
- Increased interest in interactive search
- INS might be a good track to re-introduce a subtask on localization, temporal and/or spatial
-

Some Questions

- How do participants judge the Eastenders dataset?
- Are the topics challenging enough?
- Factors affecting difficulty/success?
- Fan-site metadata:
  - Used?
  - How?
  - Successfully?

Recommendations for the final paper

- Re-run a TV12 or TV11 system on TV 13 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 2014 plans

Continue with same test data and new set of topics