TRECVID INSTANCE SEARCH (INS)

Shin'ichi Satoh and Zheng Wang
National Institute of Informatics, Japan

TRECVID (from TRECVID web site...)

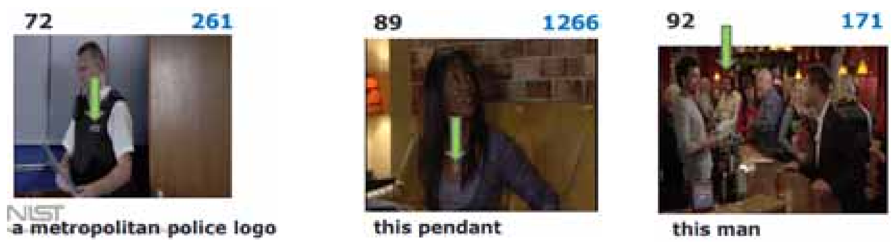
- Workshop series from 2001 to present
- Large-scale laboratory testing for content-based video analysis and retrieval
- Forum for the
  - exchange of research ideas
  - discussion of approaches: what works, what doesn't, and why
- Aims for realistic system tasks and test collections
  - ***unfiltered data***
  - focus on relatively high-level functionality
- Provides data, tasks, and uniform, appropriate scoring procedures

TRECVID Instance Search (INS)

- To find “instances” of some object, person, or location in video
  - specific object, person, or location
  - e.g., search for this particular dog, search for *Dot* (a certain person) appearing in *Kitchen1* (a certain location)
  - different manufactured objects which are indistinguishable
  - the certain person and the certain location are not recognizable at the same time
- Queries will be given as visual examples
- There exist couple of related benchmark datasets
  - Oxford Building, Paris (landmarks)
  - Flickr Logos (logos)
  - UKBench, Stanford Mobile Visual Search (specific objects)
  - etc.

Comparison with other benchmarks

- TRECVID INS determines data first: therefore very “wild”



- Other benchmarks define queries first, and then collect data: therefore objects clearly appear

Oxford 5K

- 5062 images collected from Flickr for particular Oxford landmarks.
- Manually annotated to generate a comprehensive ground truth for 11 different landmarks, each represented by 5 possible queries.

For each image and landmark, one of four possible labels was generated:

- Good - A nice, clear picture of the object/building.
- OK - More than 25% of the object is clearly visible.
- Bad - The object is not present.
- Junk - Less than 25% of the object is visible, or there are very high levels of occlusion or distortion.

- 7-220 good and ok images per query

Philbin, J. , et al., Object retrieval with large vocabularies and fast spatial matching, CVPR, 2007
http://www.robots.ox.ac.uk/~vgg/data/oxbuildings/

The Stanford Mobile Visual Search Dataset (SMVS)

- The data set consists of images for many different categories captured with a variety of camera-phones, and under widely varying lighting conditions. Database and query images alternate in each category.

Category	Database	Query
CD	100	400
DVD	100	400
Books	100	400
Video Clips	100	400
Landmarks	500	500
Business Cards	100	400
Text documents	100	400
Paintings	100	400

V. Chandrasekhar, et al., The Stanford Mobile Visual Search Data Set, MMSys, 2011
http://web.cs.wpi.edu/~claypool/mmsys-dataset/2011/stanford/

FlickrLogos-32

- The dataset FlickrLogos-32 contains photos showing brand logos and is meant for the evaluation of logo retrieval and multi-class logo detection/recognition systems on real-world images. They collected logos of 32 different logo brands by downloading them from Flickr. All logos have an approximately planar surface.

Partition	Description	Images	#Images
P_1 (training set)	Hand-picked images	10 per class	320 images
P_2 (validation set)	Images showing at least a single logo under various views	30 per class	3960 images
	Non-logo images	3000	
P_3 (test set = query set)	Images showing at least a single logo under various views	30 per class	3960 images
	Non-logo images	3000	
P_1, P_2 and P_3 are disjoint.			**8240 images**

http://www.multimedia-computing.de/flickrlogos/

University of Kentucky Benchmark (UKB)

- The University of Kentucky retrieval benchmark is a dataset which consist of 2550 classes, each class with 4 images with JPEG format.
- The pictures are from diverse categories such as animals, plants, household objects, etc.

D. Nister, et al., Scalable recognition with a vocabulary tree. CVPR, 2006

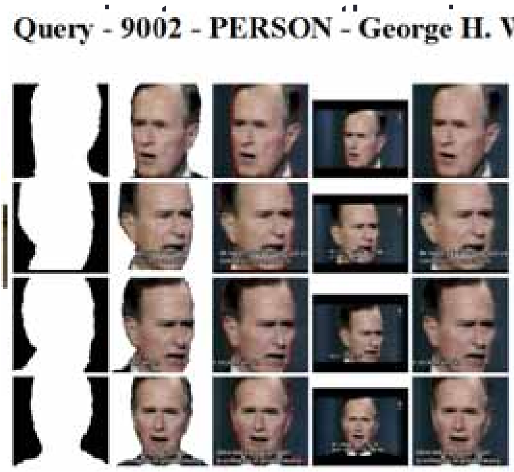
Comparison with other benchmarks

	TRECVID INS	The other benchmarks
Instance type	Different: object & person & location	Same: logo | landmark | …
Scale	Different scales in the images	Similar scale, main part of the image
Target Frequency	A wide range from 10 to 2000	Similar number, stable
Data Source	TV videos	Images collected from internet
Data type	Video, audio, text	image
Characteristic	determines data first: therefore very “wild”	define queries first, and then collect data: therefore objects clearly appear

TRECVID INS Data

- Collection of several hundreds hours of videos for each year
- Data should contain multiple occurrences of multiple specific objects.
- Search tasks should be reasonably difficult.

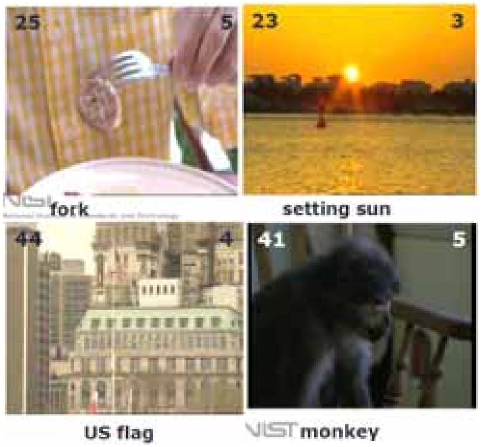
- Sound and Vision (2010): too difficult, too few repeated



TRECVID INS Data

- Collection of several hundreds hours of videos for each year
- Data should contain multiple occurrences of multiple specific objects.
- Search tasks should be reasonably difficult.

- BBC Rushes (2011): including retakes, artificial video transformations



TRECVID INS Data

- Collection of several hundreds hours of videos for each year
- Data should contain multiple occurrences of multiple specific objects.
- Search tasks should be reasonably difficult.

- Flickr Creative Commons (2012): retrieved by text queries, reasonable, but still hard to find repeated instances

Data

- Collection of several hundreds hours of videos for each year
- Data should contain multiple occurrences of multiple specific objects.
- Search tasks should be reasonably difficult.

- BBC EeastEnders (2013-present): drama series, “small world” many repeated instances (person, location, objects, ...)
- **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

Frequency of Ground-truth

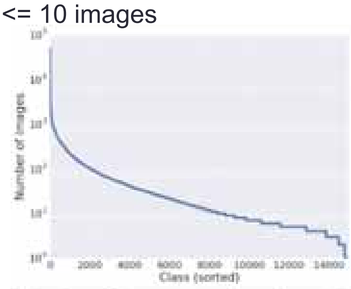
Ox5k 5K, 55 query, 11 classes, 7-220 imgs

SMVS 1200, 3300 query, 1200 classes

FL32 8240, 32 classes, average

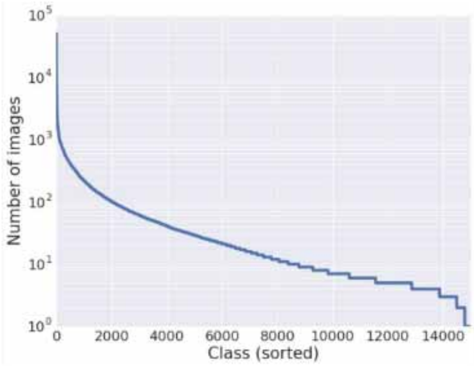
Google Landmark Retrieval

Few images per class
23% with <= 5 images
44% with <= 10 images



Number of images per class: long tail

Google Landmarks

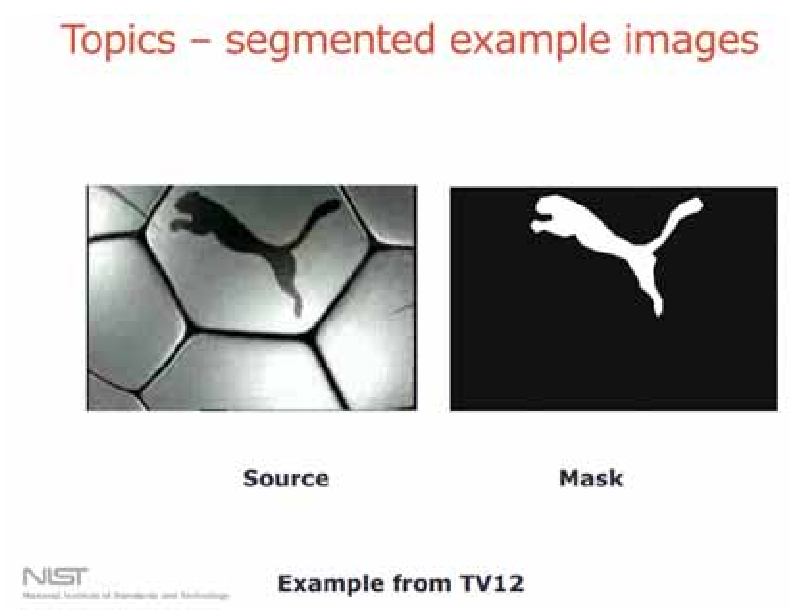


Number of images per class: long tail

- 13k unique landmarks from across the world
- 1M images
- 111k query images
- Automatically generated ground truth (GPS and visual features)

Queries in 2013-2015

- Couple of example images with masks
- Original videos are also given (since 2014)



Task in 2013-2015

- 2013-2015: specific object, person, or location

Topic: True positives:

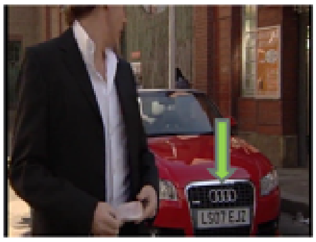
69 2300

a 'no smoking' logo

70 741

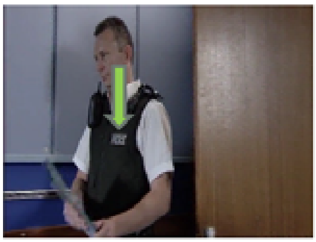
a small red obelisk

71 31



an Audi logo

72 261



a metropolitan police logo

73 674

this ceramic cat face

74 100

a cigarette

Task in 2013-2015

- 2013-2015: specific object, person, or location

89 1266

this pendant

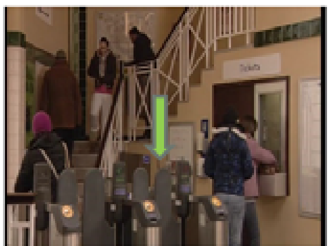
90 363

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

Task in 2013-2015

- 2013-2015: specific object, person, or location

104 342

this woman

115 277

this man

116 238

this man

139 33

this shaggy dog

140 95

a Walford Gazette banner

141 52

this guinea pig

Task in 2013-2015

- 2013-2015: specific object, person, or location

132 68

this brass piano lamp

133 112

this lava lamp

134 472

this cylindrical spice rack

149 286

this Walford Community Center entrance from street

151 94

this Walford Police Station entrance from street

Difficulties

Easy topics	Difficult topics
• Simple visual context • Stationary target • Planar, rigid objects	• Small target • Moving target: differences in camera angle, location • Non-planar, non-rigid

Typical INS template system in 2013-2015

representation (local/global)

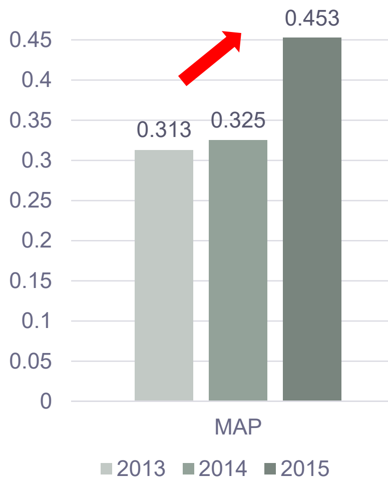
- HSV, LBP, HOG, CNN
- Hesssian affine, Harris-Laplace, MSER
- Detection methods, face classifier
- RootSIFt + C-Sift
- Cluster to BoVW (1M words, hard/soft etc)
- VLAD quantization

matching

- similarity function (asymmetric similarity)
- normalization term in cosine similarity
- BM25 enhancements for weighting
- IDF adjusted for burstiness
- Weighting ROI vs. background
- query adaptive late fusion

Post-processing

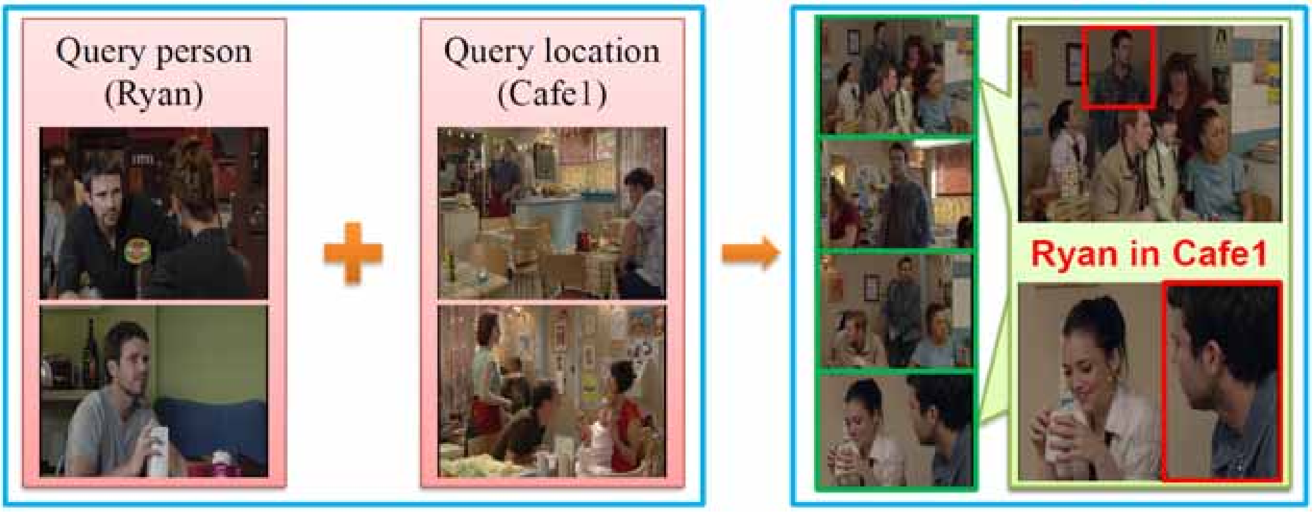
- spatial verification
- Face/color filtering
- Fusion of scores (Average pooling)
- Use ROI features for reranking



2013	**RootSIFt + C-Sift** **asymmetric similarity function**
2014	**larger BoVW vocabularies** **combines BoVW and CNN** **normalization term in cosine similarity** **ROI features for reranking**
2015	**Use CNN for both local and global features** **Query adaptive late fusion**

Task in 2016-present

- 2016-present: find a specific person in a specific location



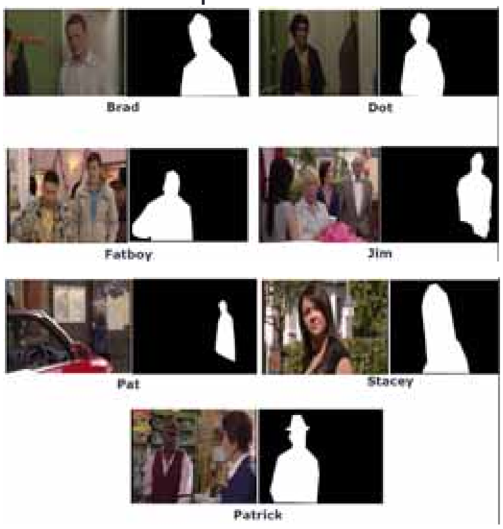
The figure refers to PKU_ICST at TRECVID 2017: Instance Search Task

Comparison with task in 2013-2015

	2013-2015	2016-present
Data Source	The same	
Topics	object / person / location	person + location
query	Image + mask	Person: image + mask Location: 6-12 images Related video shots
Characteristic	One condition	Two conditions together
Difficulty	Instance with different scales and types	Persons / locations have different views Person and location influence to each other, can not be searched out simultaneously

Topics in 2016

person



location

Topics in 2016

	Jim	Dot	Brad	Stacey	Pat	Patrick	Fatboy
Pub	x	x	x	x	x	x	x
Foyer	x	x	x	x	x		
LR1	x	x	x	x	x		x
Kitchen1	x	x	x	x	x	x	
Laundrette	x		x	x	x	x	x

30 x topics : find {jim, Dot, Brad, Stacey, Pat, Patrick, Fatboy} in {Pub,Foyer,LR1,Kitchen1,Laundrette}

Topics in 2017

person

location

Topics in 2017

	Peggy	Billy	Ian	Janine	Archie	Ryan	Shirley	Phil
Cafe1	x	x	x	x		x	x	x
Market			x	x	x		x	x
LR2	x	x			x		x	x
Kitchen2	x	x		x		x	x	x
Launderette	x	x	x	x	x	x	x	

30 x topics : find {Peggy, Billy, Ian, Janine, Archie, Ryan, Shirley, Phil} in {Cafe1,Market,LR2,Kitchen2,Launderette}

Task in 2016-present

- Example for person search 1

Person: [jane]

Difficulties：
1. different camera conditions
2. different styles

Query

Ranking results

Task in 2016-present

- Example for person search 2

Difficulties：
3. incomplete or occluded
4. face is too small
5. appears in adjacent frame

Person: [mo]

Query

Ranking results

Task in 2016-present

- Example for location search 1

Difficulties：
1. different views
2. blocked by persons

Location: [cafe2]

Query

Ranking results

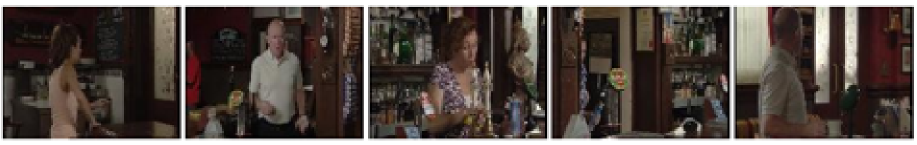
Task in 2016-present

- Example for location search 2

Difficulties：
1. different views
2. blocked by persons

Location: [pub]

Query



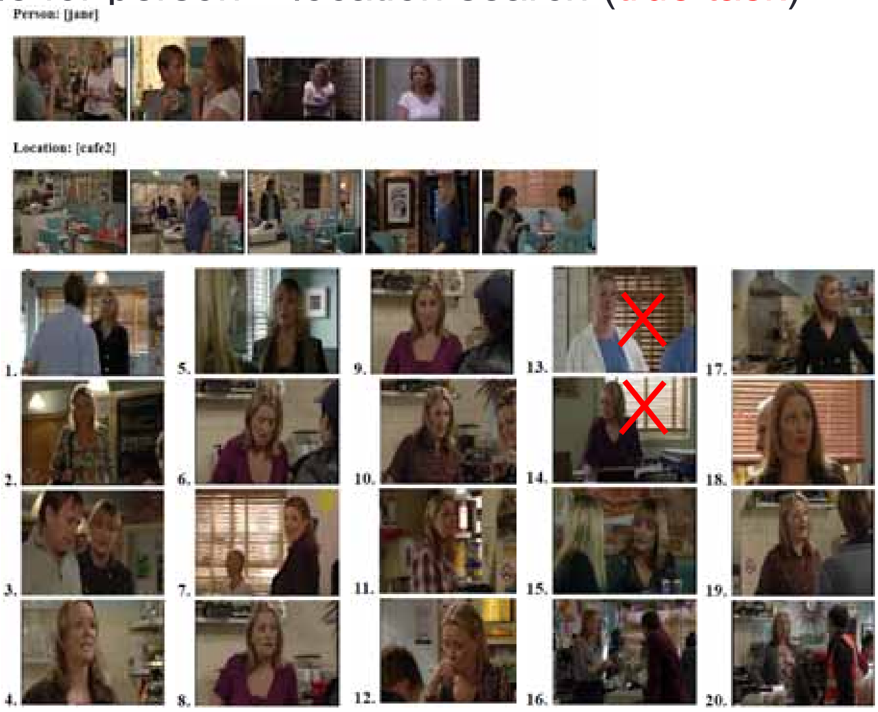
Ranking results

Task in 2016-present

- Example for person + location search (true task)

Query

Ranking results



Task in 2016-present

- additional difficulties for person + location : person search and location search are always in a dilemma.


person faces are non-front or occluded

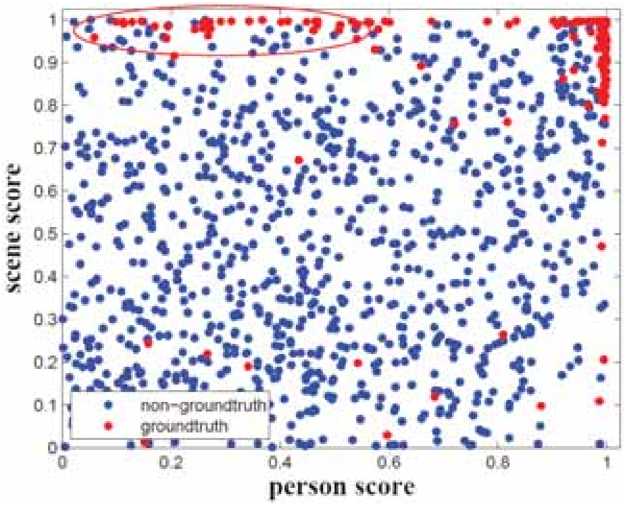

scenes are with low light or blur


although it is a wide-angle view scene, the person faces are very small

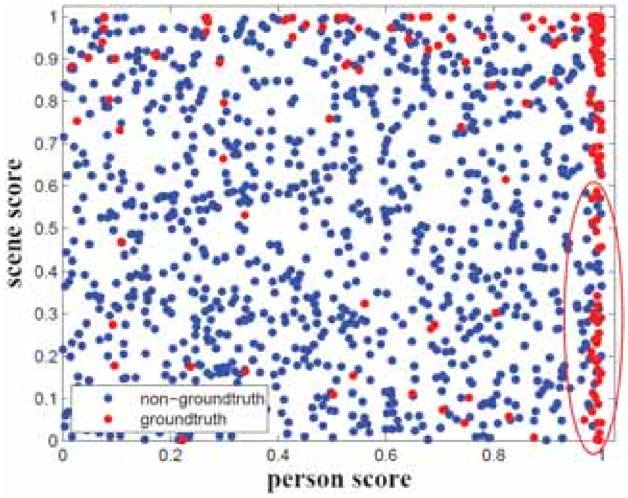

scenes are blocked by persons

Task in 2016-present

- additional difficulties for person + location : person search and location search are always in a dilemma.

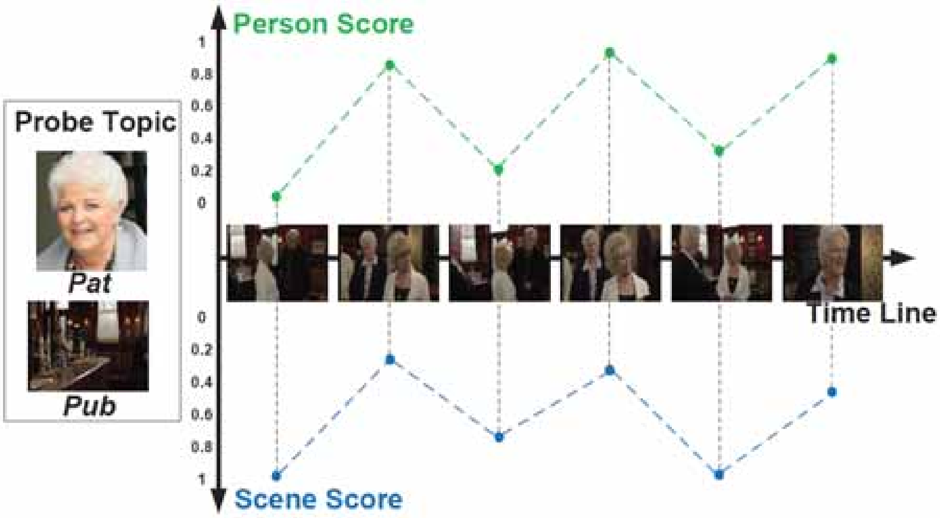


Topic 9170 in TRECVID INS 2016
high scene score V.S. low person score



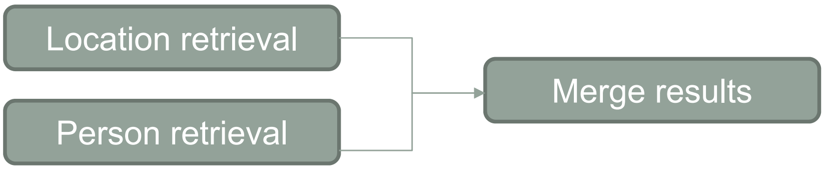
Topic 9210 in TRECVID INS 2017
low scene score V.S. high person score

Task in 2016-present



An example for consecutive shots in a time slice. Although the shots contain the target person in the target location, the person and location scores are not always high simultaneously.
Neighbor shots will be helpful.

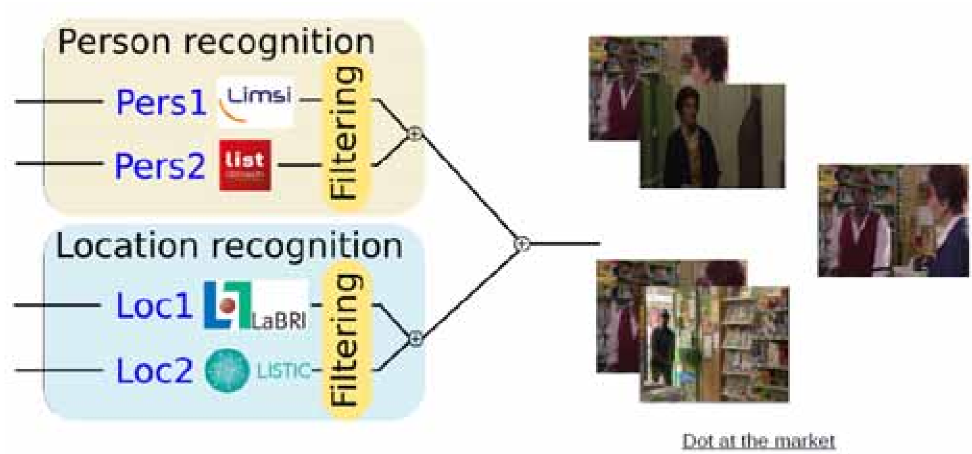
Systems Comparison



	Person retrieval	Location retrieval	Merge results
BUPT-MCPRL	face retrieval (dlib) person re-identification (Faster RCNN + fc layer feature) transcript-based	RootSIFT+AlexNet VGG-16 Places365	Peron guide location+ location guide person + random forest
NII-Hitachi-UIT	DPM+VGG-Face SVM with RBF kernel	BOW	scene tracking with person re-identification
IRIM	HOG detector + ResNet pre-trained on FaceScrub & VGG-Face Viola-Jones detector + FC7 of a VGG16 network	Bow + Filter out person Pretrained GoogLeNet Places365	Credits shots filtering Indoor/Outdoor shots filtering Shots threads filtering Late fusion
PKU_ICST	VGG-Face + Cosine + SVM+ Progressive training	AKM-based (6 kinds of BoW) DNN-based (VGGnet+GoogleNet+ResNet) + Progressive training	Peron guide location+ location guide person + highlight common clues Semi-supervised re-ranking

Typical Systems

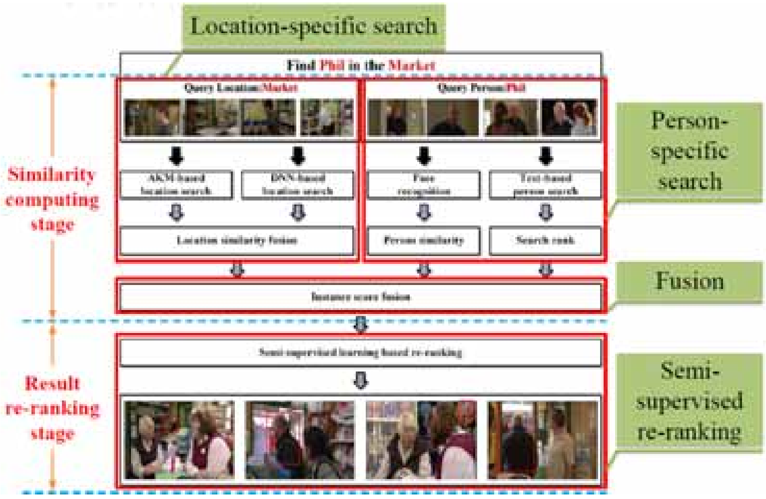
- IRIM at TRECVID 2017 (MAP = 0.4466)



Pers1： HOG detector + ResNet pre-trained on FaceScrub & VGG-Face
Pers2： Viola-Jones detector + FC7 of a VGG16 network
Loc1： Bow + Filter out person
Loc2： GoogLeNet Places365

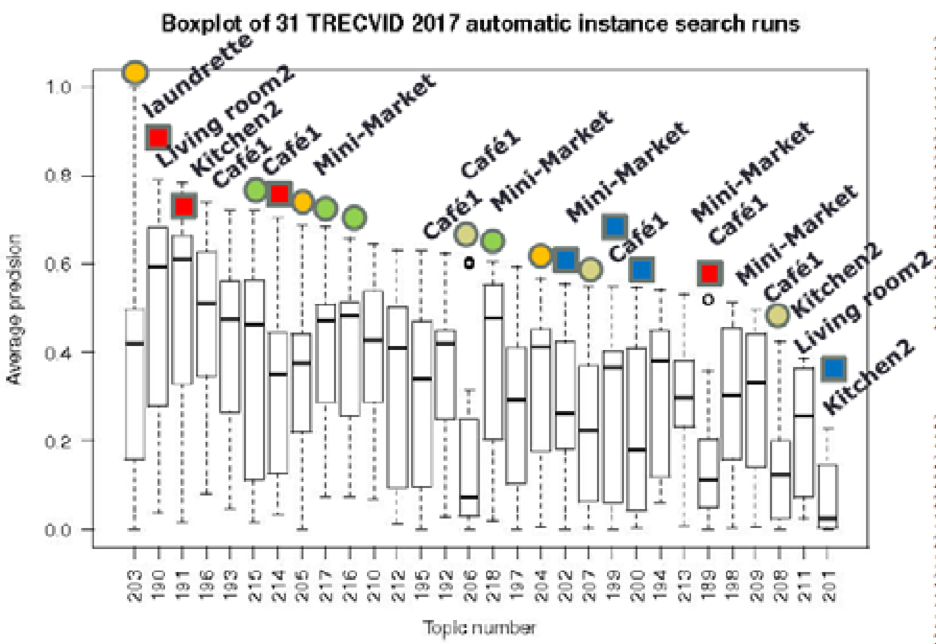
Typical Systems

- PKU_ICST at TRECVID 2017 (0.549)



Location-specific search: AKM-based (6 kinds of BoW) + DNN-based (VGGnet+GoogleNet+ResNet)
Person-specific search：VGG-Face + Cosine + SVM
Re-ranking：Semi-supervised re-ranking method (fusion)

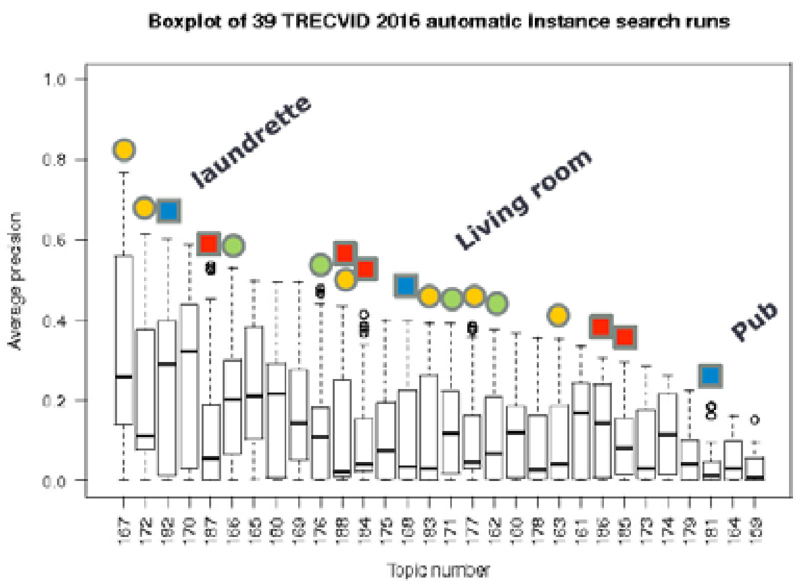
Analysis



What is the effect of person vs location on the performance ?
- **Mini-Market is hard**
- **Archie, Peggy, and phil are easy**
- **Janine and Ryan are hard**

#	Query
203	Find Archie in this Laundrette
190	Find Peggy in this LivingRoom 2
191	Find Peggy in this Kitchen 2
196	Find Ian at this Cafe 1
193	Find Billy in this Laundrette
215	Find Phil in this Cafe 1
214	Find Peggy in this Laundrette
205	Find Archie in this Mini-Market
217	Find Phil at this Kitchen 2
216	Find Phil in this Living Room 2
210	Find Shirley in this Laundrette
212	Find Shirley in this Kitchen 2
195	Find Billy in this Kitchen 2
192	Find Billy in this Cafe1
206	Find Ryan in this Cafe 1
218	Find Phil in this Mini-Market
197	Find Ian in this Laundrette
204	Find Archie in this Living Room 2
202	Find Janine in this Mini-Market
207	Find Ryan in this Laundrette
199	Find Janine in this Cafe 1
200	Find Janine in this Laundrette
194	Find Billy in this Living Room 2
213	Find Shirley in this Mini-Market
189	Find Peggy in this Cafe1
198	Find Ian in this Mini-Market
209	Find Shirley in this Cafe 1
208	Find Ryan in this Kitchen 2
211	Find Shirley in this Living Room 2
201	Find Janine in this Kitchen 2

Analysis



#	Query
167	Find Dot in this Living Room
172	Find Brad in this Living room
182	Find Fatboy in this Laundrette
170	Find Brad in this Laundrette
187	Find Pat at this Foyer
166	Find Dot at this Foyer
165	Find Dot in this Kitchen
180	Find Patrick in this Laundrette
169	Find Brad in this Kitchen
176	Find Stacey at this Foyer
188	Find Pat in this Living Room
184	Find Pat in this Pub
175	Find Stacey in this Laundrette
168	Find Brad in this Pub
183	Find Fatboy in this Living room
171	Find Brad at this Foyer
177	Find Stacey in this Living room
162	Find Jim at this Foyer
160	Find Jim in this Kitchen
178	Find Patrick in this Pub
163	Find Jim in this Living Room
161	Find Jim in this Laundrette
186	Find Pat in this Laundrette
185	Find Pat in this Kitchen
173	Find Stacey in this Pub
174	Find Stacey in this Kitchen
179	Find Patrick in this Kitchen
181	Find Fatboy in this Pub
164	Find Dot in this Pub
159	Find Jim in this Pub

What is the effect of person vs location on the performance ?

NIST
National Institute of Standards and Technology

"Ground truth" generation by pooling

- Genuine ground truth is *not* maintained
- Pool is composed of submitted runs by multiple teams
- Items in the pool are checked by human assessors
- Inferred AP (infAP) is computed as an unbiased estimate of AP
- Extended AP (infxAP) is then used based on stratified random sampling

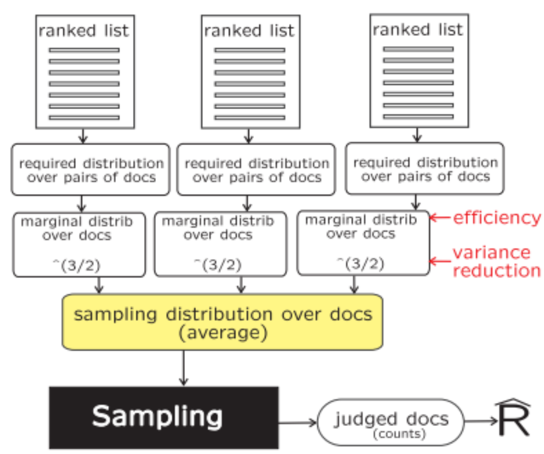


Figure 3: Sampling diagram

infAP: J. A. Aslam et al., A Statistical Method for System Evaluation Using Incomplete Judgments, SIGIR 2006
infxAP: E. Yilmaz et al. A Simple and Efficient Sampling Method for Estimating AP and NDCG, SIGIR 2008

Conclusion

- Brief explanation of TRECVID Instance Search
- Wild instance search benchmark because of “data first” approach
- Challenging task, while there still is a room to address, e.g.,
  - nature of video (consecutive shots, clips)
  - closed world information

- New data and new query structure may also be considered: discussed among participants of TRECVID