Person-Location Instance Search via Progressive Extension and Intersection Pushing

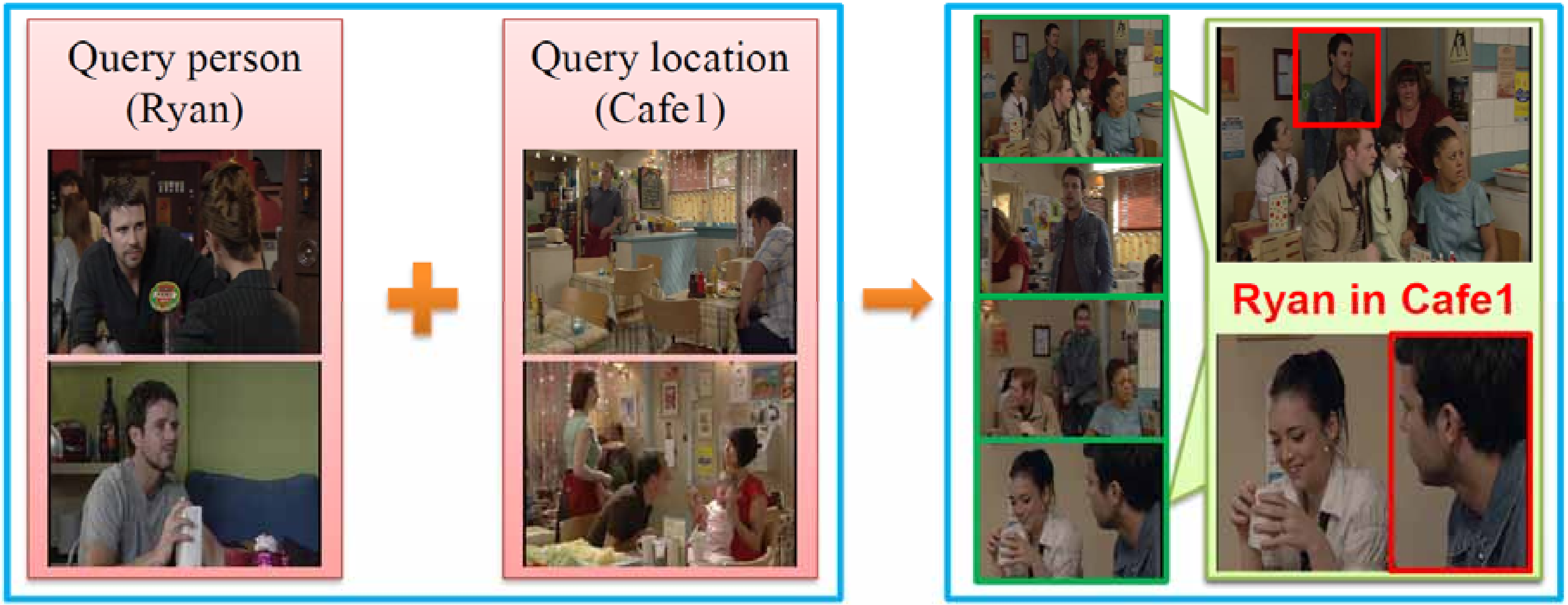
—— NII_Hitachi_UIT team at TRECVID2018 Instance Search Task

Zheng Wang and Shin'ichi Satoh

National Institute of Informatics, Japan

INS Task in 2016-present

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



The Figure refers to
[1] PKU_ICST at TRECVID 2017: Instance Search Task

INS Task Data

- 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

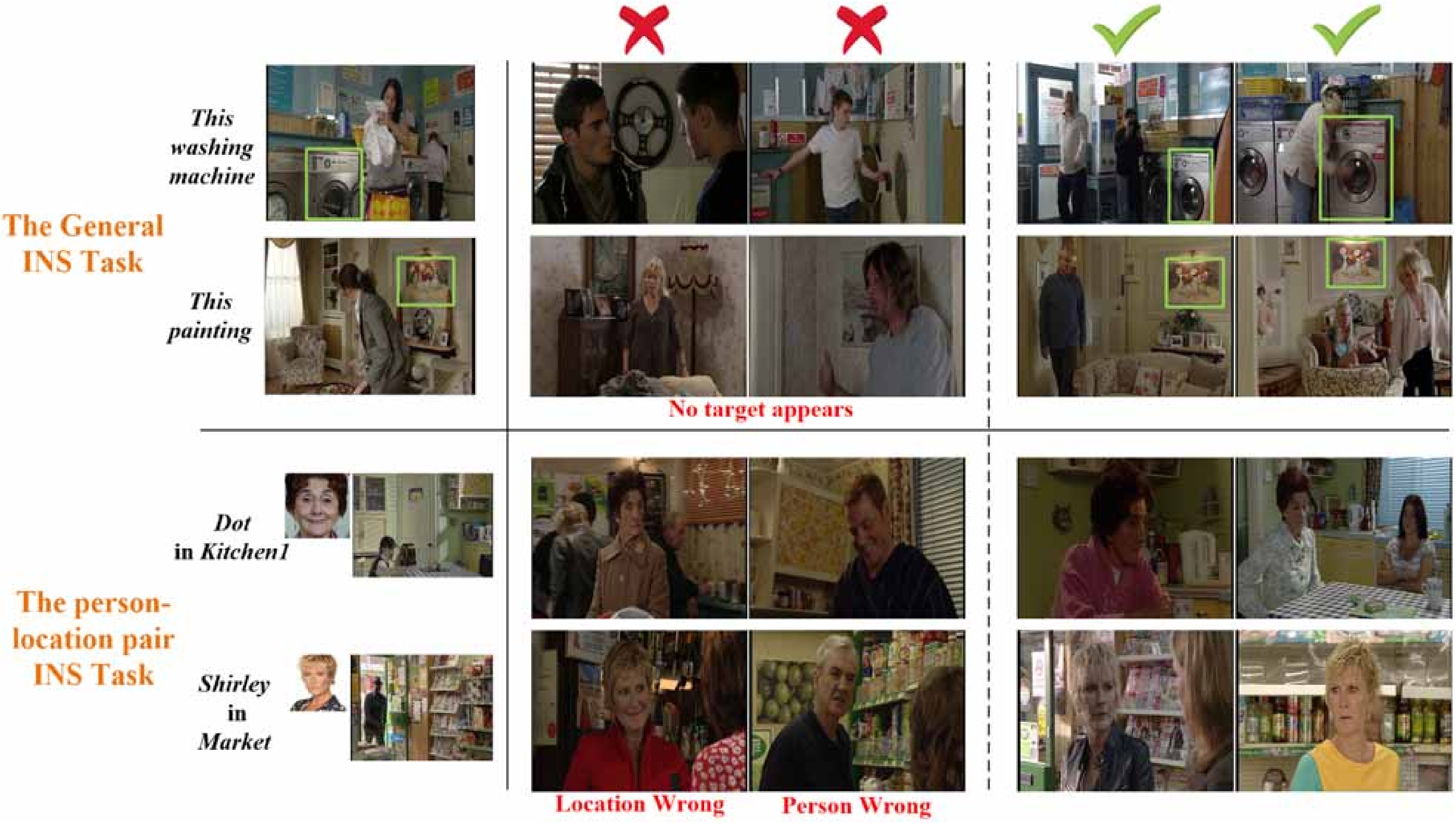
EastEnders’ world

Majority of episodes filmed at Elstree studios. Sometimes filmed on ‘location’.

Comparison with task in 2013-2015

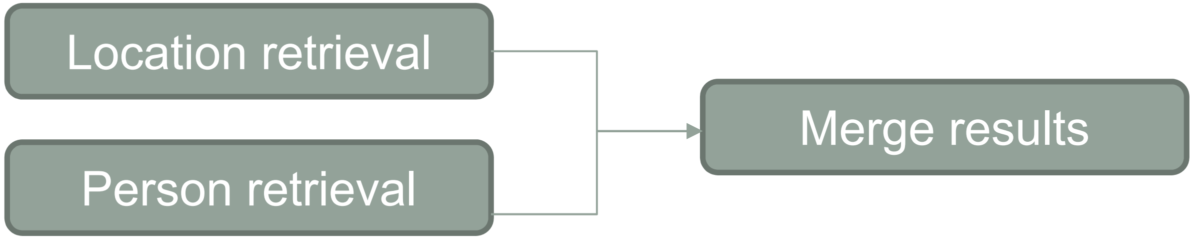
	2013-2015	2016-present
Data Source	The same	
Topics	object / person / location	person + location
query	Image + mask	Person: image + mask Location: 6-12 images Related video shots
Characteristic	One condition	Two conditions together
Difficulty	Instance with different scales and types	Persons / locations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

Example


✗
✗
✓
✓
The General INS Task
This washing machine
This painting
No target appears
The person-location pair INS Task
Dot in Kitchen1
Shirley in Market
Location Wrong
Person Wrong

Related Systems

The same routine:



Location retrieval

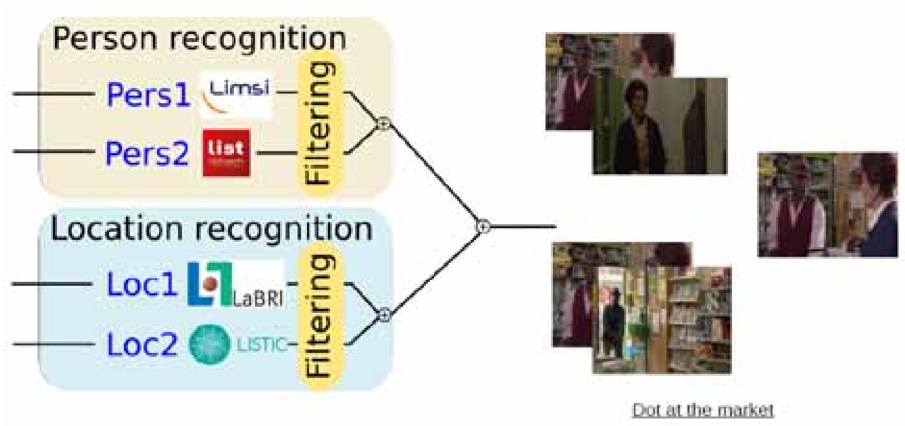
Person retrieval

Merge results

	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
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

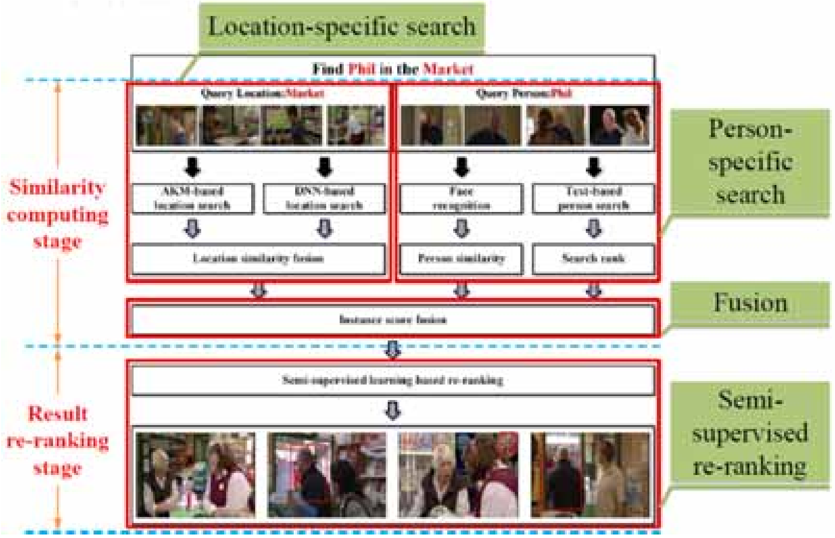
State-of-the-art Systems

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

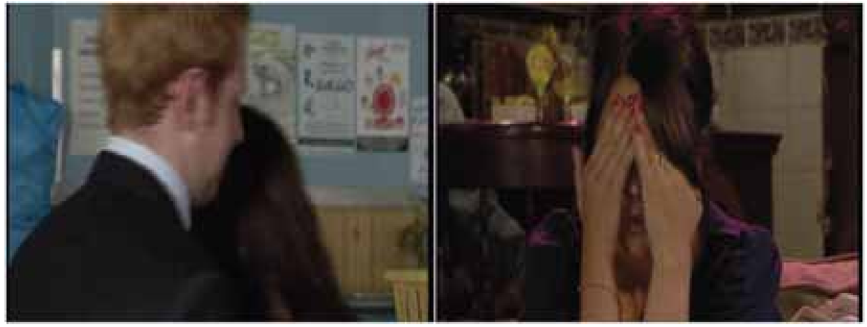
PKU_ICST at TRECVID 2017 (0.549)



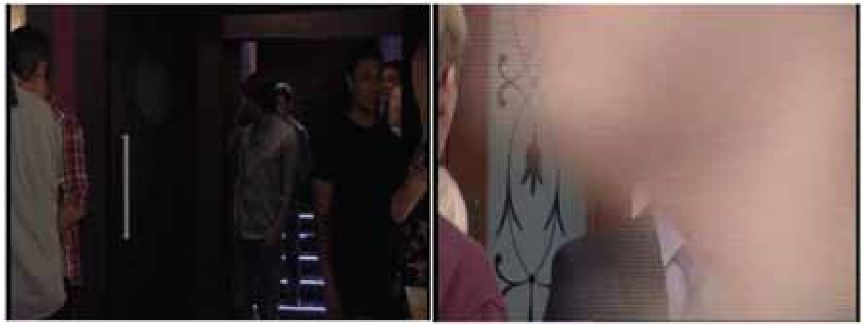
Location-specific search: AKM-based (6 kinds of BoW) + DNN-based (VGGnet+GoogleNet+ResNet)
Person-specific search： VGG-Face + Cosine + SVM
Re-ranking： Semi-supervised re-ranking method (fusion)

Difficulties

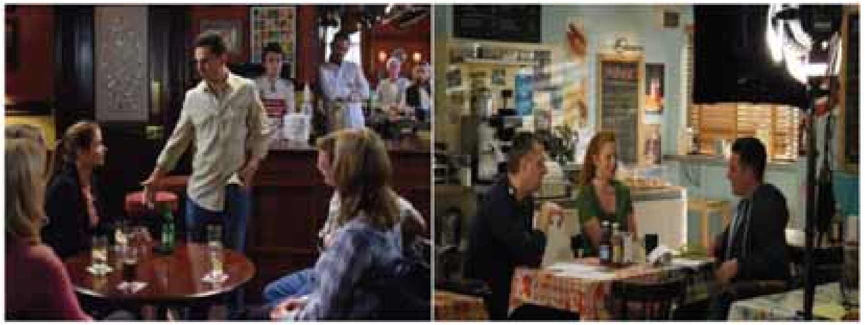
- additional difficulties for person + location : person search and location search are always in a dilemma.



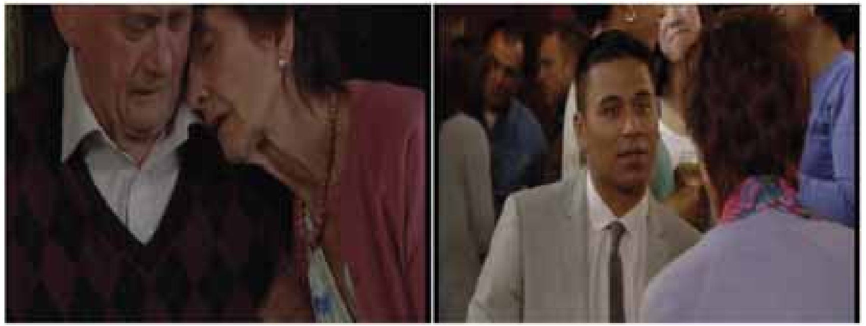
person faces are non-front or occluded



scenes are with low light or blur



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

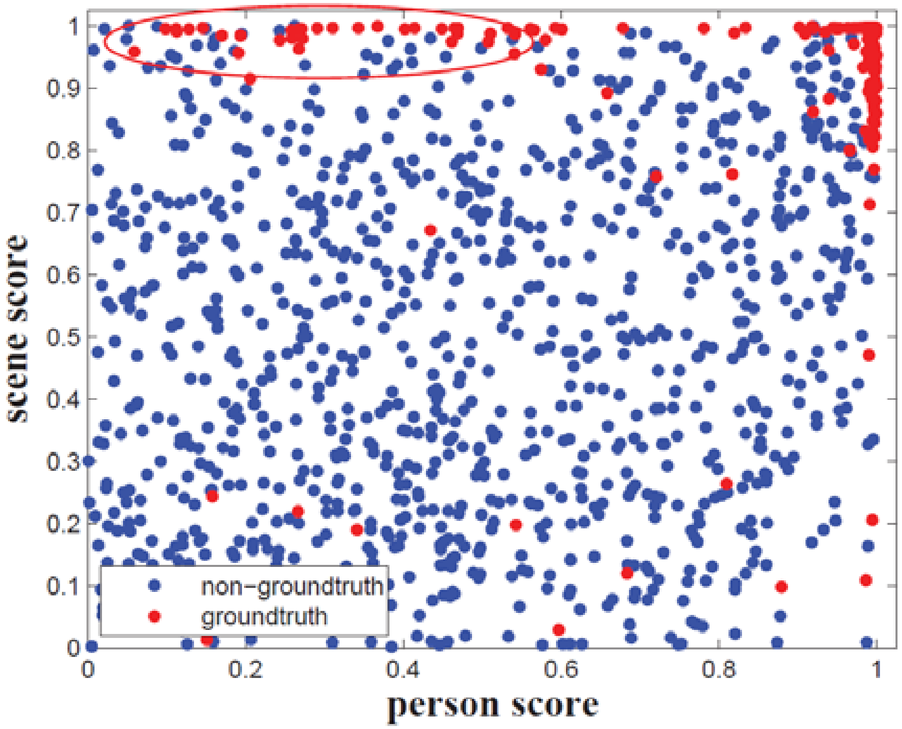


scenes are blocked by persons

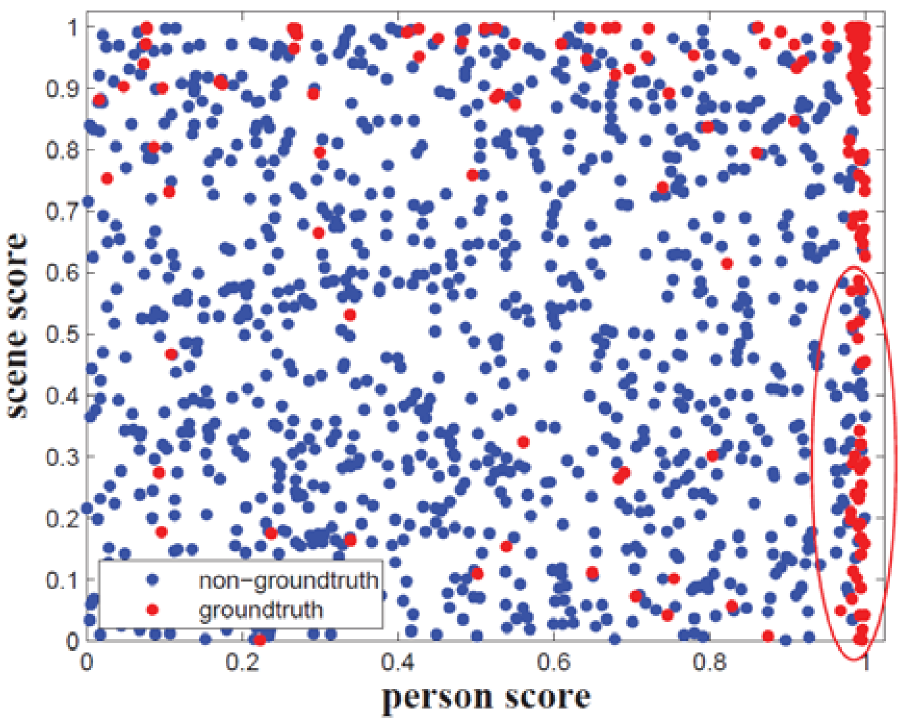
[2] J Lan, J Chen, Z Wang, C Liang, S Satoh, PS Instance Retrieval via Early Elimination and Late Expansion, ACM MM Workshop, 2017

Difficulties

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

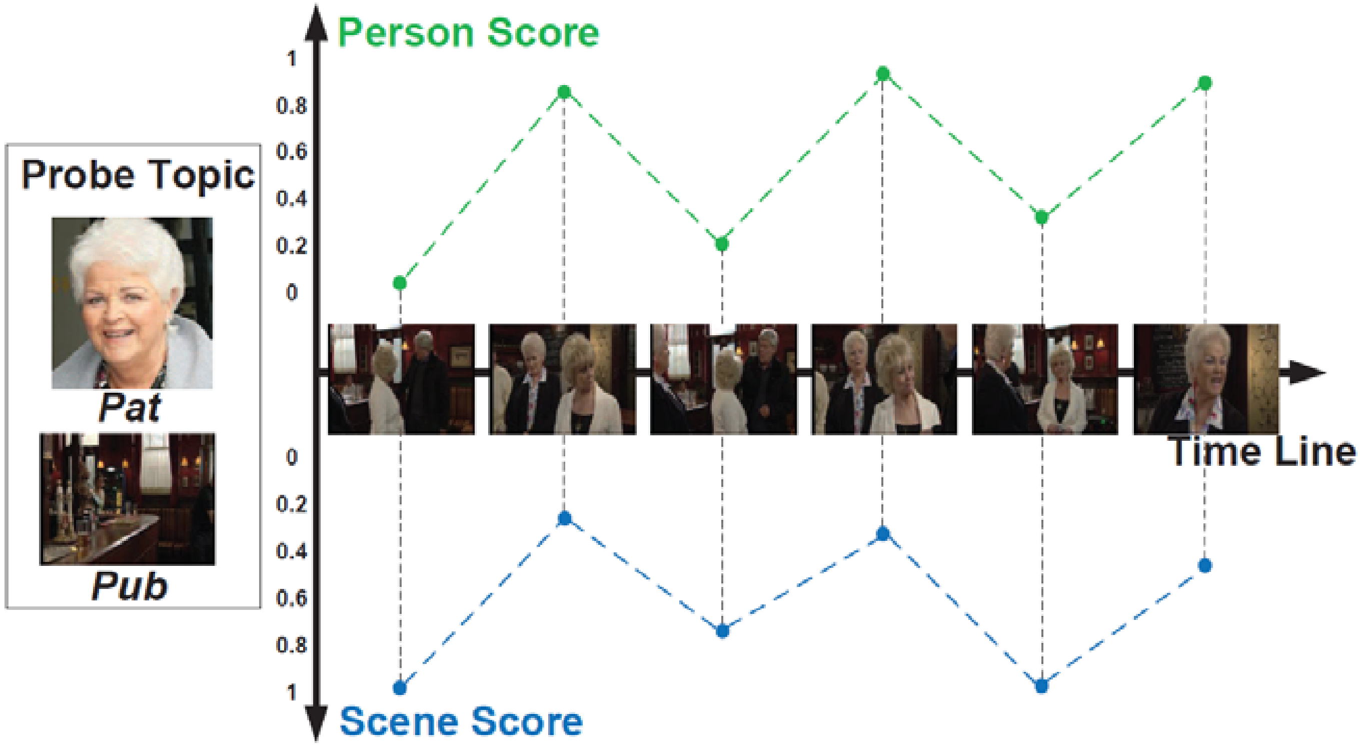


Topic 9170 in TRECVID INS 2016
high scene score V.S. low person score



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

Motivation

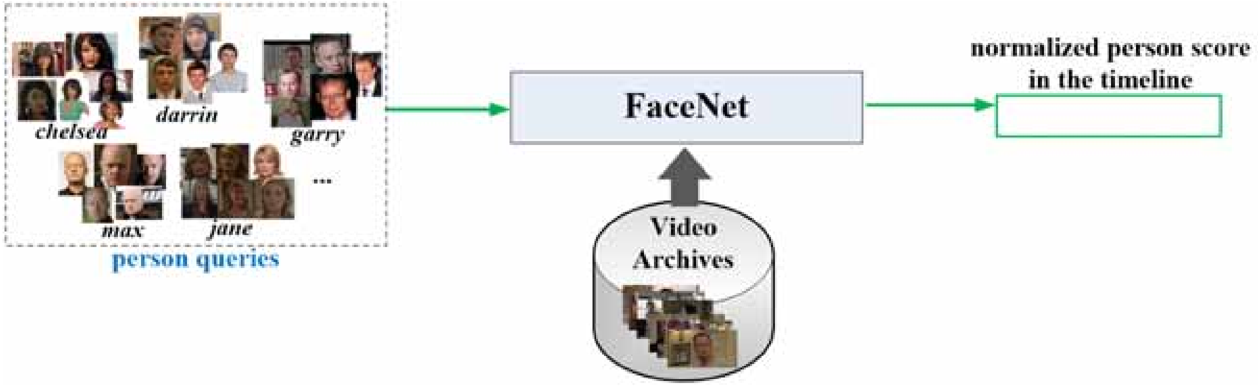


An example for consecutive shots in a time slice. Although the shots contain the target person in the target location, the person and location scores are not always high simultaneously.

Neighbor shots will be helpful.

Framework

- Person search

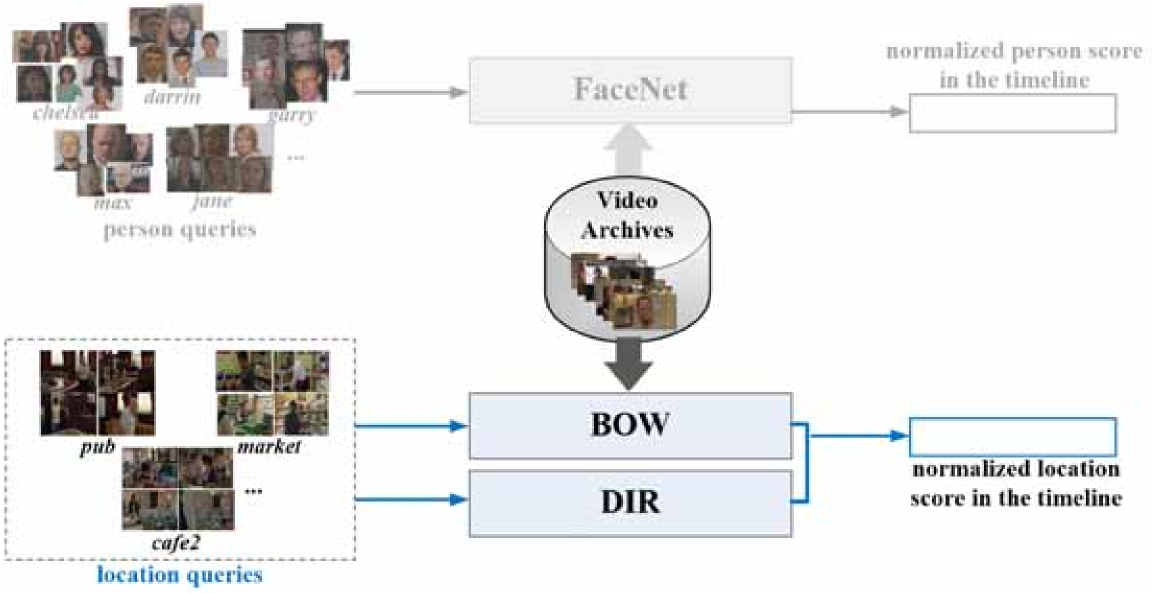


- We use face cues to do person search.
- FaceNet [3] framework is used for face recognition.
  - Multi-task CNN [4] is utilized for face detection and alignment.
  - The network is trained using softmax loss with the Inception-Resnet-v1 model.
  - The training data is VGGFace2.
  - No BBC EeastEnders data is used for training.
- For each query person, we collect 10 face images with different views.
- We use max-pooling strategy to achieve the similarity between one shot and one query topic.
- The similarity scores are normalized to [0, 1].

[3] FaceNet: A Unified Embedding for Face Recognition and Clustering, https://github.com/davidsandberg/facenet
[4] Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks, https://kpzhang93.github.io/MTCNN_face_detection_alignment/index.html

Framework

- Location search

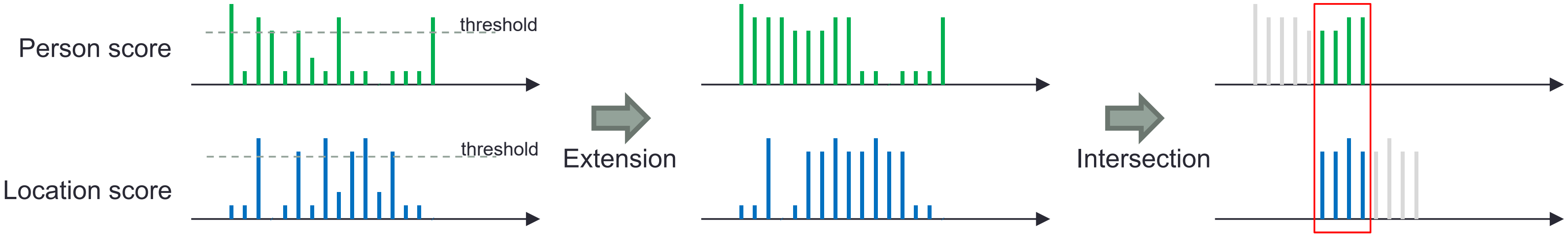


- We use two kinds of routines to obtain the location similarity scores.
- BOW stands for the hand crafted based routine. We exploit the method used in TRECVID INS 2017.
- DIR [5] stands for the deep learning based routine.
  - For each query location, we use query extension to combine global features of all corresponding query images.
- We use max-pooling strategy to achieve the similarity between one shot and one query topic.
- The similarity scores are normalized to [0, 1].

[5] Deep Image Retrieval: Learning global representations for image search, https://github.com/figitaki/deep-retrieval

Framework

- Fusion



Person score

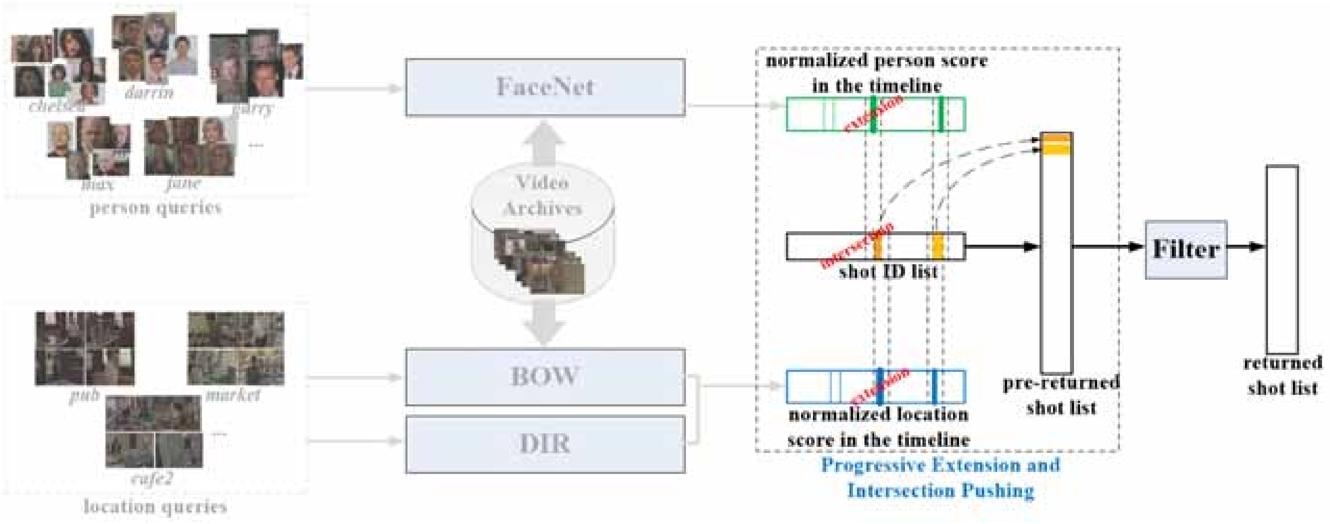
threshold

Location score

threshold

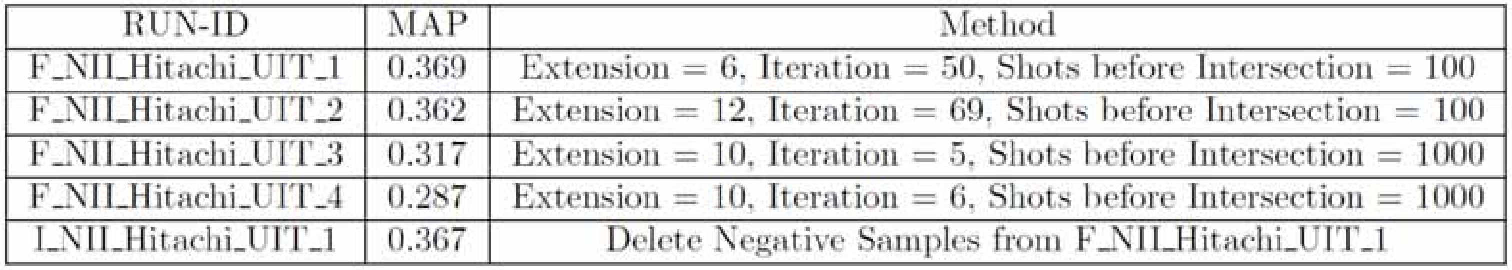
Extension

Intersection



- The method includes multiple iterations, so that we can obtain top results progressively.
- For each iteration,
  - We first extend the shot scores with high neighbor shot scores.
  - We do intersection to get the top results.

Results

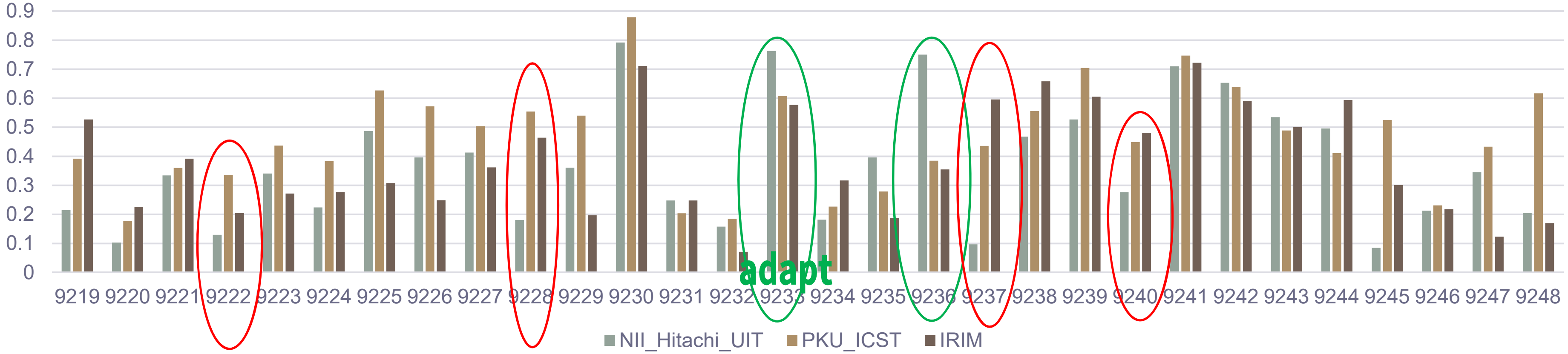


RUN-ID	MAP	Method
F_NII_Hitachi_UIT_1	0.369	Extension = 6, Iteration = 50, Shots before Intersection = 100
F_NII_Hitachi_UIT_2	0.362	Extension = 12, Iteration = 69, Shots before Intersection = 100
F_NII_Hitachi_UIT_3	0.317	Extension = 10, Iteration = 5, Shots before Intersection = 1000
F_NII_Hitachi_UIT_4	0.287	Extension = 10, Iteration = 6, Shots before Intersection = 1000
I_NII_Hitachi_UIT_1	0.367	Delete Negative Samples from F_NII_Hitachi_UIT_1

Extension: the number of neighbor shots extended.
Iteration: the times of intersection shots pushing.
Shots before Intersection: the number of shots selected before intersection.

Extension should be fine-grained
The iteration times should be large

Results-AP



Good results:

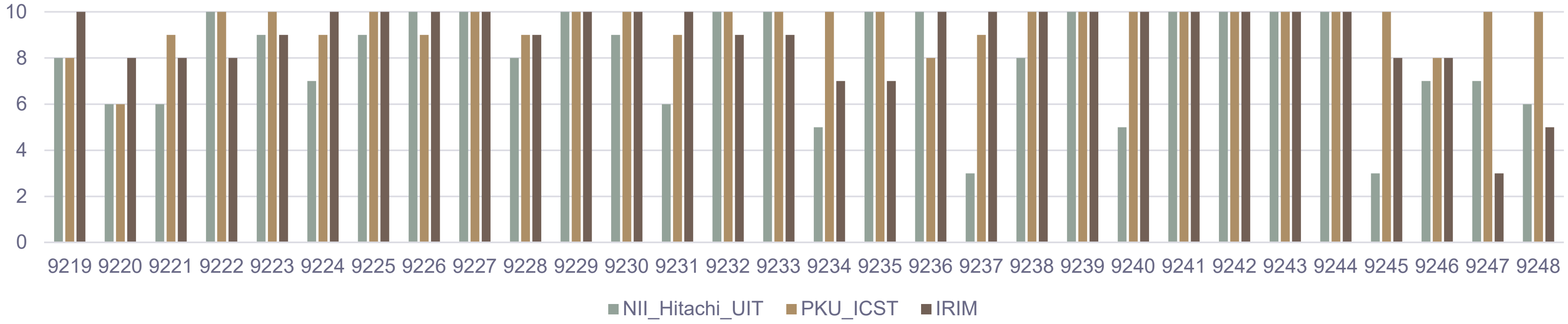
9233 Mo+Laundrette
9236 Darrin+Laundrette

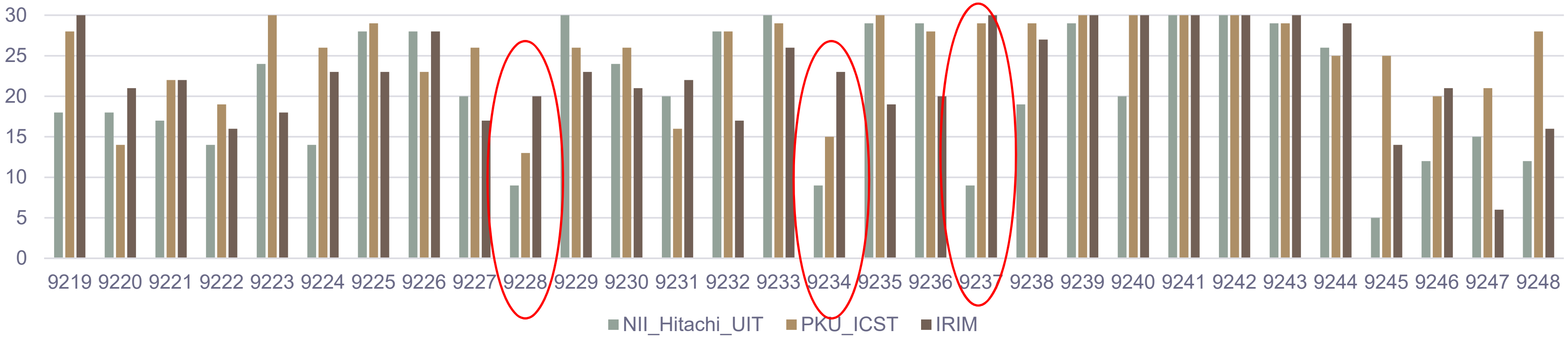
Bad results:

9222 Chelsea+Cafe2
9228 Garry+Cafe2
9237 Zainab+Cafe2
9240 Heather+Cafe2

Our method does not perform well in some scenes.
Our location search model does not adapt to the new INS domain.

Results-Hits at depth 10/30 in the result set





Bad results:

9228 Garry+Cafe2
9234 Darrin+Cafe2
9237 Zainab+Cafe2

For the top results, our method performs similar to the other methods.

Thanks