TRECVID 2020 INSTANCE RETRIEVAL INTRODUCTION AND TASK OVERVIEW

Wessel Kraaij
Leiden University; Netherlands
Organisation for Applied Scientific Research (TNO)

George Awad
Georgetown University; National Institute of Standards and Technology

Keith Curtis
National Institute of Standards and Technology

Disclaimer
The identification of any commercial product or trade name does not imply endorsement or recommendation by the National Institute of Standards and Technology.

Outline

- Task Definition
- Data
- Topics (Queries)
- Participating Teams
- Evaluation and Results
- General Observation

Task

From 2013 – 2015

- The objective of this task facilitates system development that automatically detects a specific object, person or location in any context using a small set of image and video examples.

From 2016 - 2018

- A different query type was used: *find a specific person in a specific location.*

In 2019 - 2021

- A new query type is being used: *find a specific person doing a specific action.*

System task:

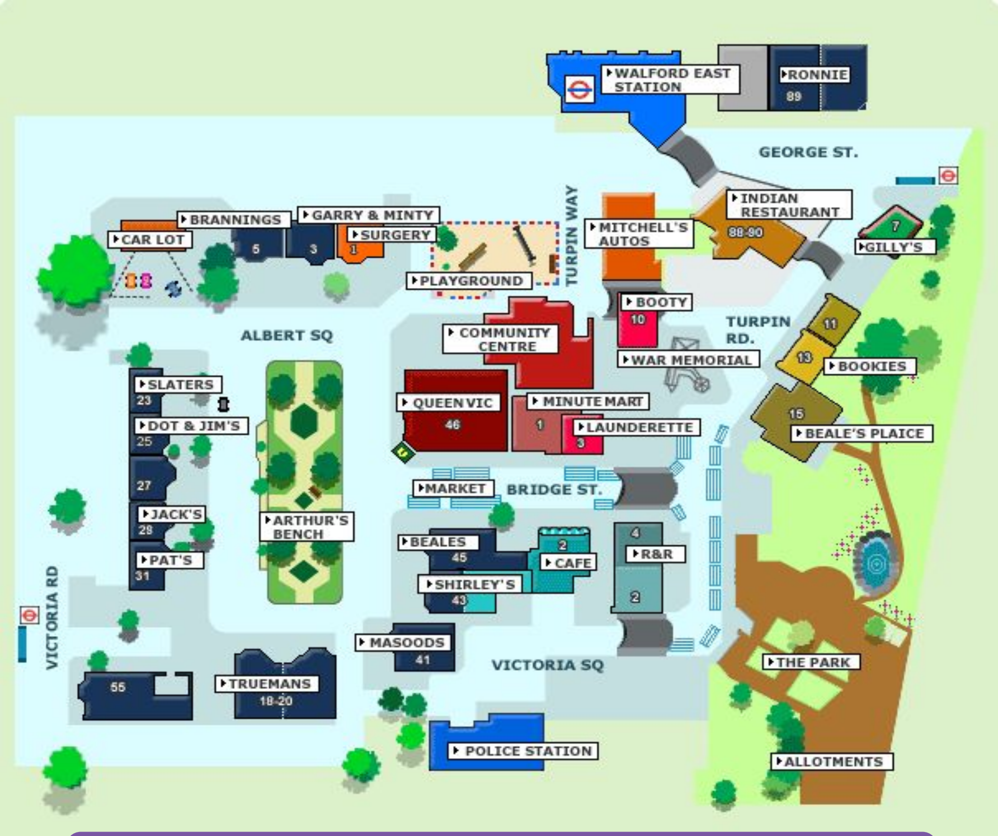
- Given a topic with:
  - 4 example images of the target person
  - 4 Region of Interest (ROI)-masked images of the target person
  - 4 to 6 video examples of a specific action
- Return a list of up to 1000 shots ranked by likelihood that they contain the target person doing the target action
- **Automatic** or **interactive** runs are accepted

Data

- The British Broadcasting Corporation (BBC) and the Access to Audiovisual Archives (AXES) project made **464 h** of the BBC soap opera EastEnders available for research
  - 244 weekly "omnibus" video files (MPEG-4) 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 at NIST

- Viewed several videos to develop a list of recurring people, actions and their overlapping.
- Listed in order the most frequent actions and most frequent person's performing them
- Created ≈90 topics targeting recurring specific persons doing specific actions.
- Chose 40 topics as a representative sample, including 20 unique topics for 2020 and 20 common topics for 2019 - 2021. Each topic includes images for target persons and example videos of the specific actions.
- Filtered out example shots from the submissions if it satisfies the topic.

Global Test Data Condition: Type of Training Data

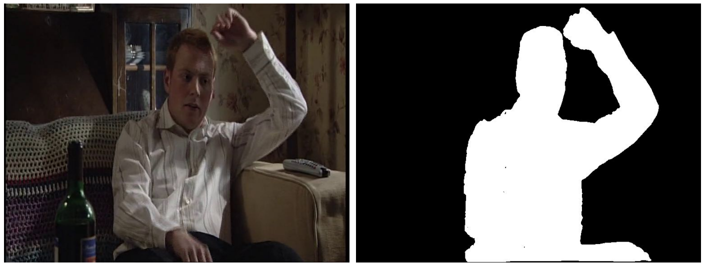
Effect of examples – 2 conditions:

- A – one or more provided images – no video
- E – video examples (+ optional image examples)

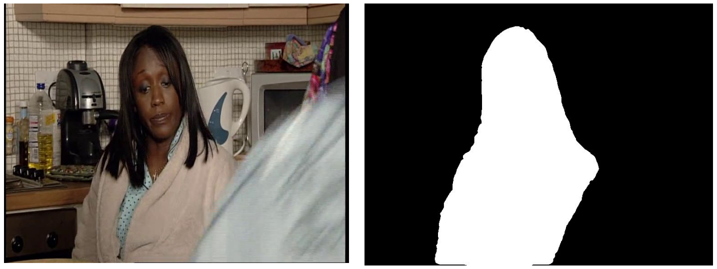
Sources of Training Data:

- A – Only sample video 0
- B – Other external data only
- C – Only provided images/videos in the official query
- D – Sample video 0 AND provided images/videos in the official query (A + C)
- E – External Data AND NIST provided images (sample video 0 OR official query images/videos)

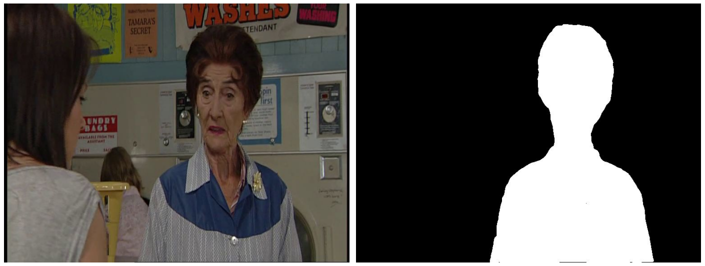
Topics – segmented ‘person’ example images



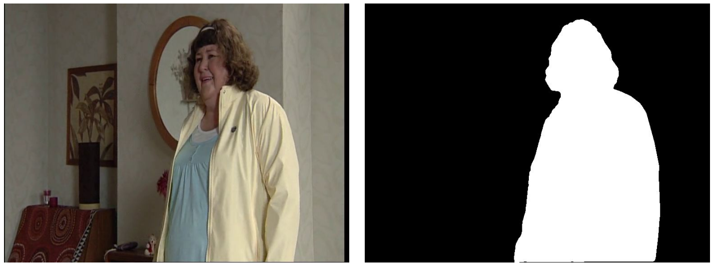
Bradley



Denise



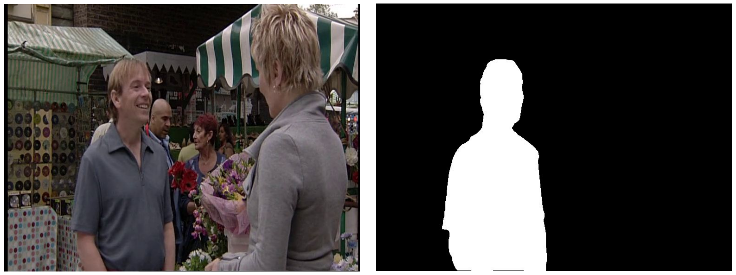
Dot



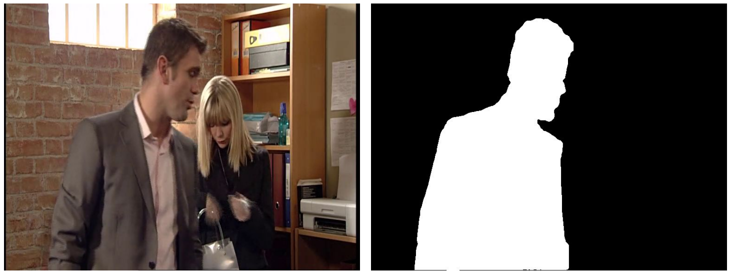
Heather

* All images are from the BBC Eastenders dataset licensed to TRECVID and participants by the BBC R&D

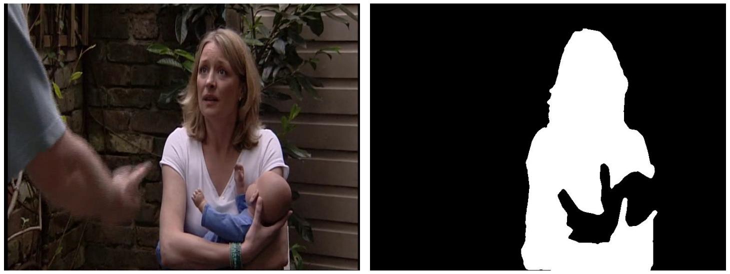
Topics – segmented ‘person’ example images



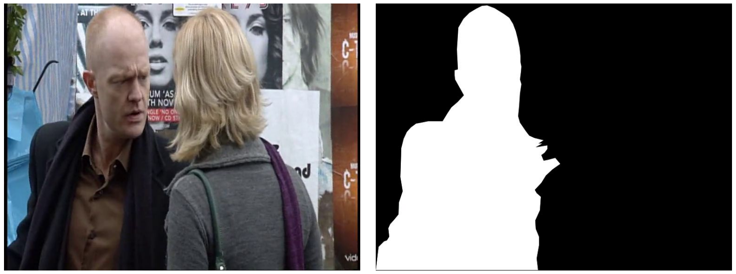
Ian



Jack



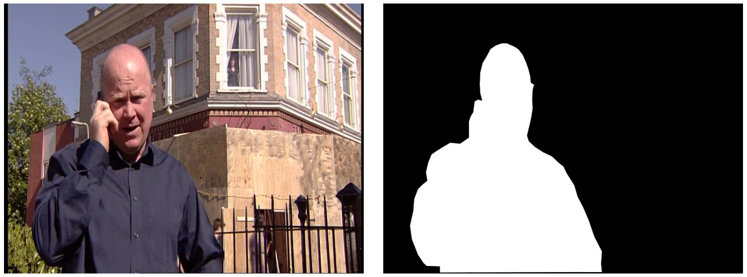
Jane



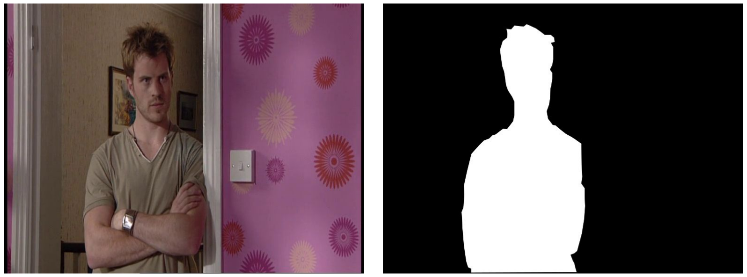
Max

* All images are from the BBC Eastenders dataset licensed to TRECVID and participants by the BBC R&D

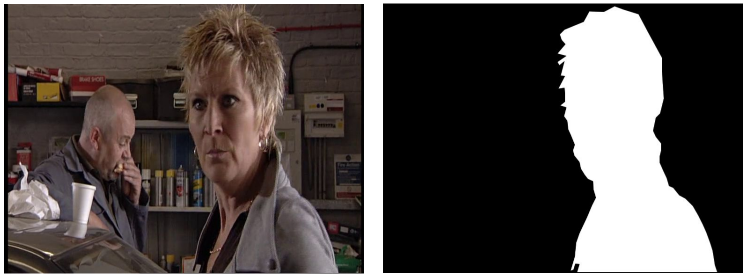
Topics – segmented ‘person’ example images



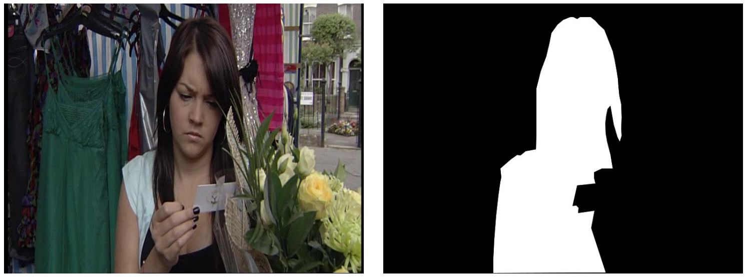
Phil



Sean



Shirley



Stacey

*** All images are from the BBC Eastenders dataset licensed to TRECVID and participants by the BBC R&D**

Sample Actions

Open door & enter

Sit on couch

* All images are from the BBC Eastenders dataset licensed to TRECVID and participants by the BBC R&D

Sample Actions

Drinking

Hugging

* All images are from the BBC Eastenders dataset licensed to TRECVID and participants by the BBC R&D

20 Unique Queries - 2020

	Max	Pat	Ian	Heather	Dot	Bradley	Billy	Stacey
Holding paper			x			x	x	
Sit on couch			x				x	
Holding cloth		x		x				
Drinking	x				x			
Smoking cigarette	x	x			x			
Holding phone	x							x
Crying			x	x				
Laughing		x						x
Go up / down stairs	x					x		

20 x unique queries : find {Max, Pat, Ian, Heather, Dot, Bradley, Billy, Stacey} doing {Holding paper, Sit on couch, Holding cloth, Drinking, Smoking cigarette, Holding phone, Crying, Laughing, Go up / down stairs}

20 Common Queries - 2020

	Sean	Max	Denise	Phil	Dot	Heather	Jack	Shirley	Stacey
Kissing			x				x		
Sit on couch				x		x			
Holding phone						x	x		
Drinking				x				x	
Open door & enter	x			x					
Open door & exit		x							x
Shouting	x							x	
Hugging			x						x
Close door without leaving					x		x		
Stand & talk at door		x			x				

20 x common queries : find {Sean, Max, Denise, Phil, Dot, Heather, Jack, Shirley, Stacey} doing {Kissing, Sit on couch, Holding phone, Drinking, Shouting, Hugging, Open door & leave, Open door & enter, Close door without leaving, Stand & talk at door}

INS 2020: 5 Finishers (out of 13)

Team	Organization	Run Types Submitted F: automatic, I: Interactive. Data Types: A: Images, E: Video
BUPT_MCPRL	Beijing University of Posts and Telecommunications, China	F_A (2)
UEC	The University of Electro-Communications, Japan	F_A (2)
PKU_WICT	Peking University, China	F_A (3), F_E (3), I_E (1)
WHU_NERCMS	Wuhan University, China	F_E (2)
NII_UIT	National Institute of Informatics, Japan (NII); University of Information Technology, VNU-HCM, Vietnam	F_A (4)

In total, 16 automatic and 1 interactive runs were submitted.

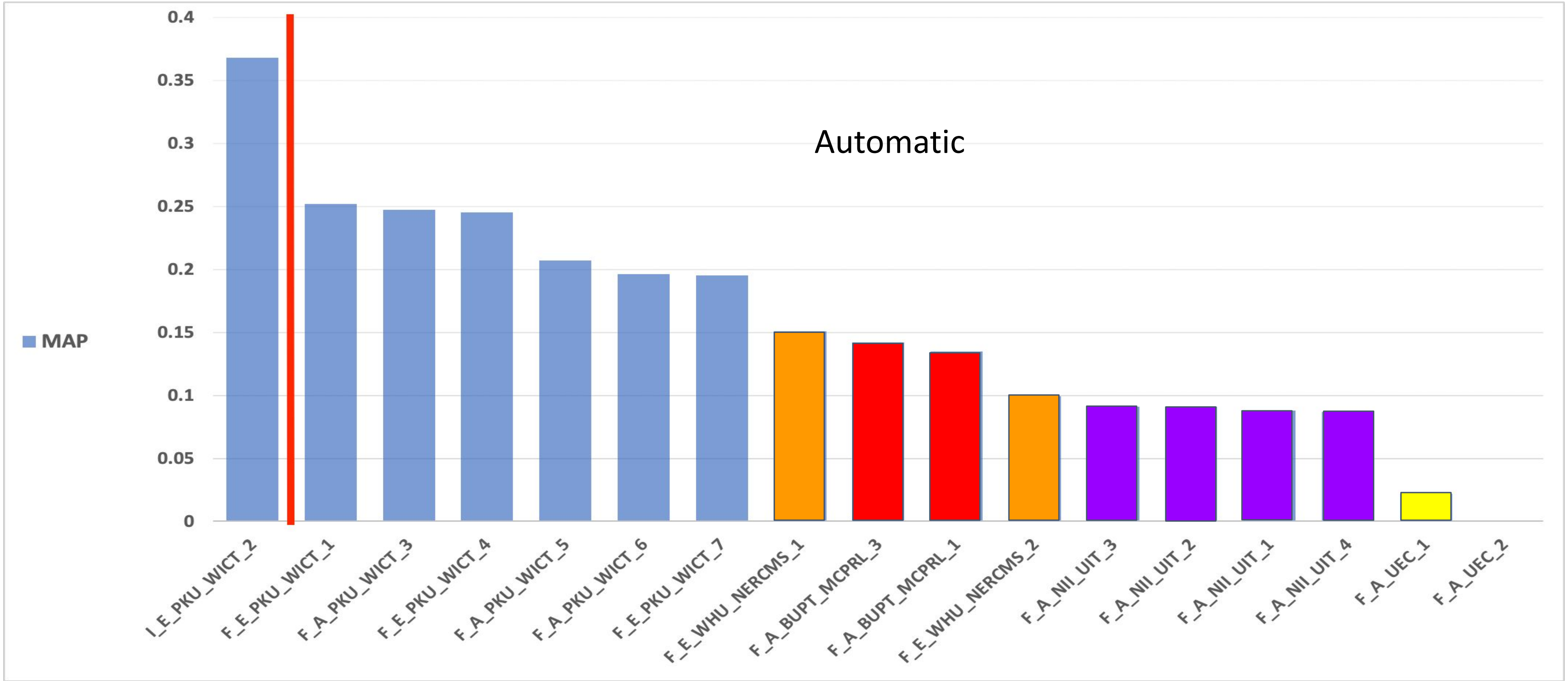
Evaluation

For each topic the results were pooled from all runs and judged up to max rank 520, resulting in 71 251 judged shots (≈ 473 person-h).

- 10 NIST assessors played the clips and determined if they contained the topic target or not.
- 4 920 clips (avg. 164 / topic) contained the topic target (6.905 %)
- True positives per topic: min 12 med 116 max 533
- The task is treated as a form of ranking and thus the trec_eval_video tool was used to calculate average precision, recall, precision, etc.
- To measure efficiency, speed was also measured.

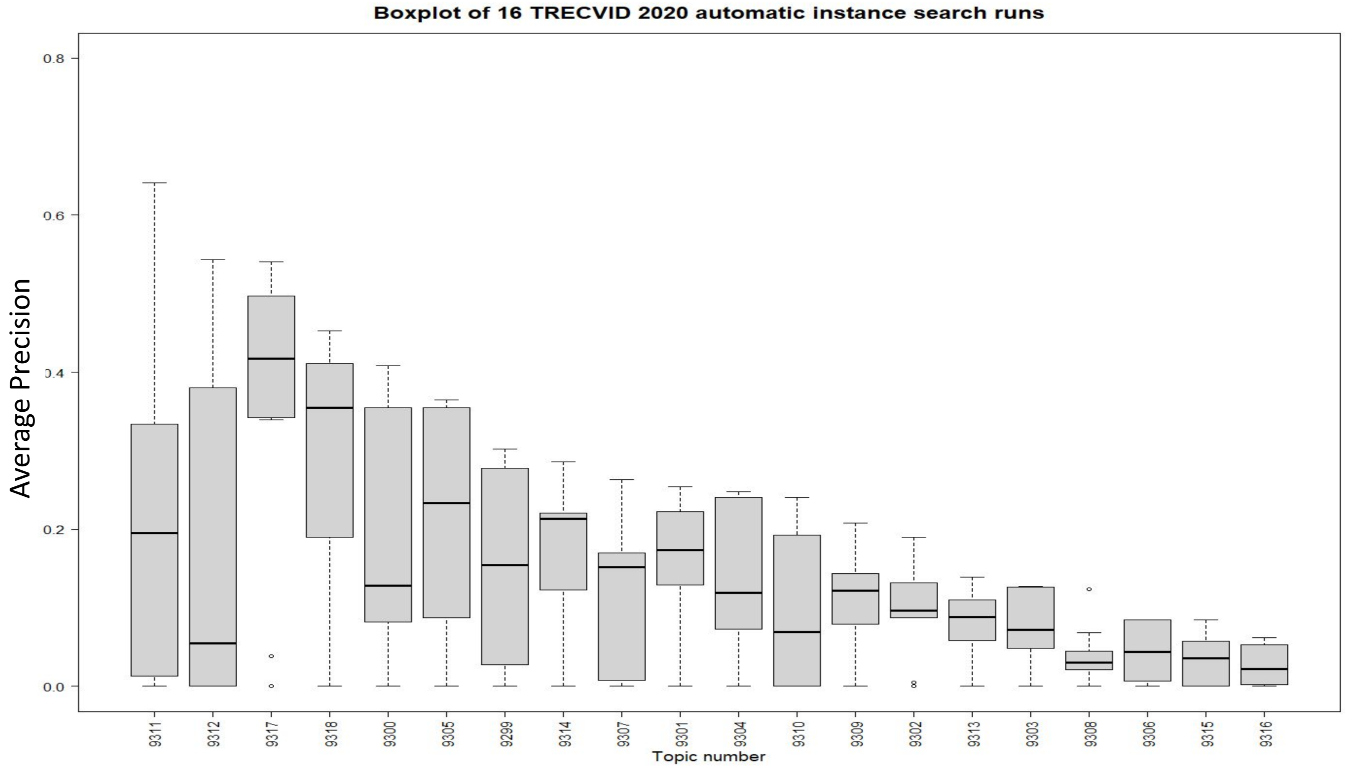
Results by Team

Interactive



Automatic

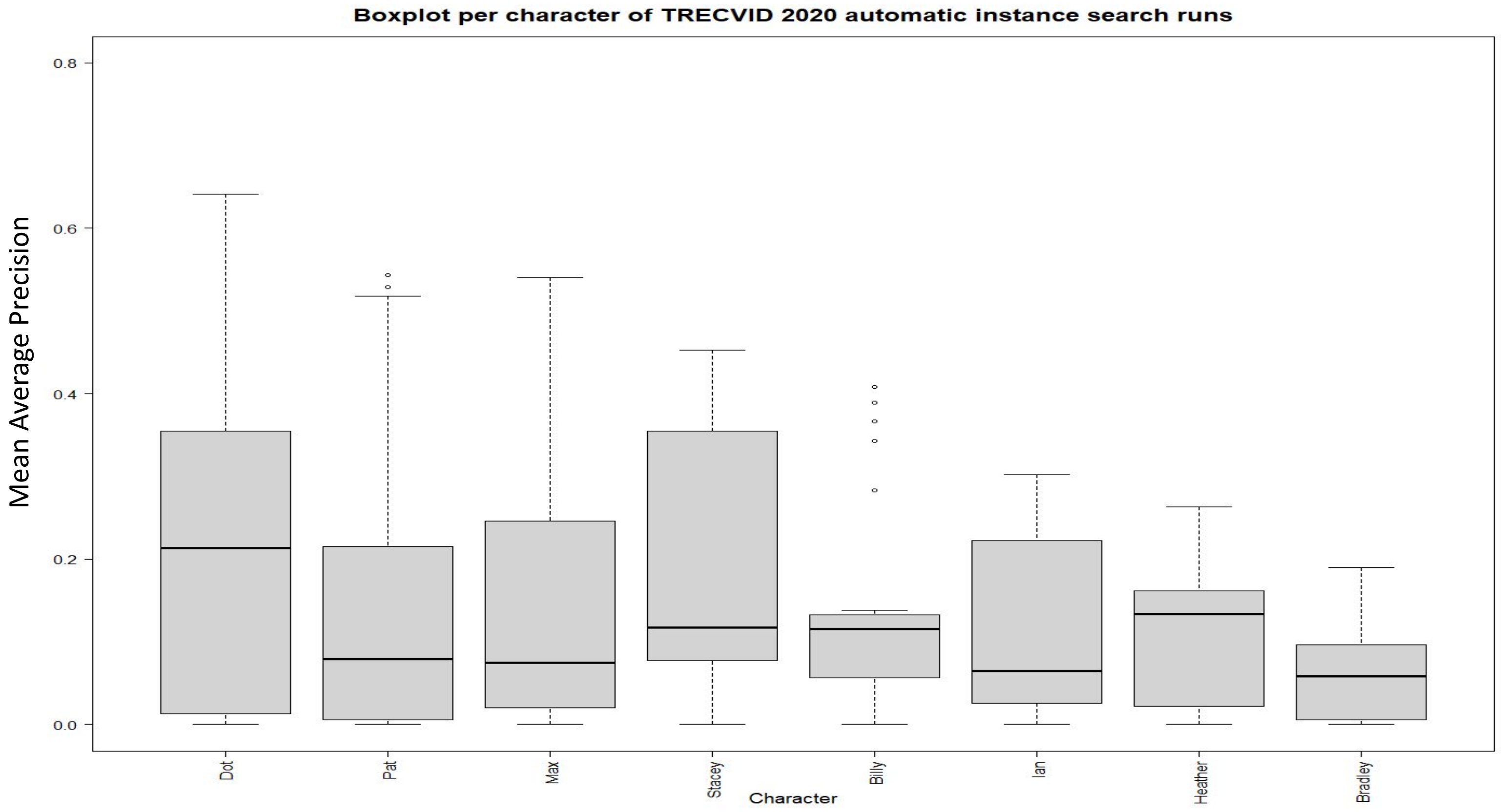
Results by Topics - Automatic



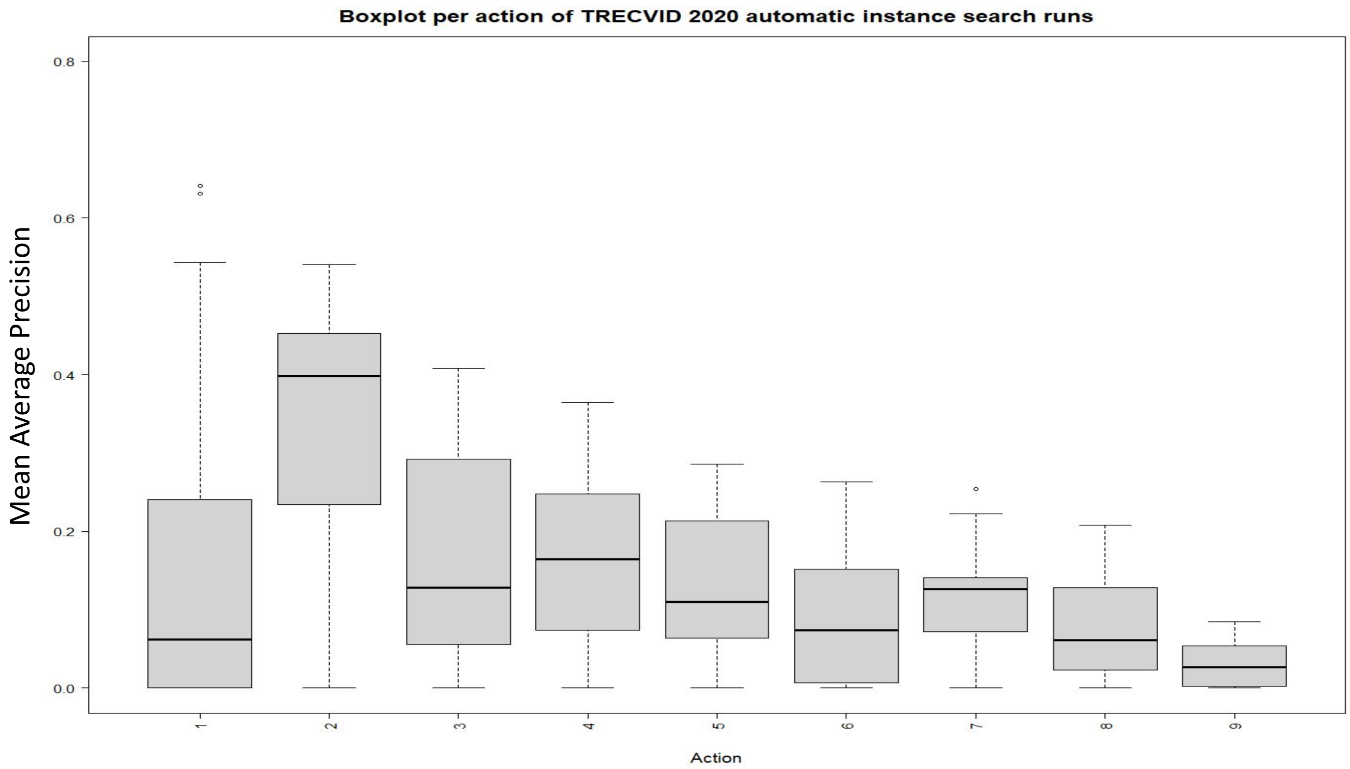
*Mean score of Average Precision per character/action

#	Query
9311	Find Dot Smoking Cigarette
9312	Find Pat Smoking Cigarette
9317	Find Max Holding phone
9318	Find Stacey Holding phone
9300	Find Billy Sit on couch
9305	Find Dot Drinking
9299	Find Ian Sit on couch
9314	Find Pat Laughing
9307	Find Heather Holding Cloth
9301	Find Ian Holding Paper
9304	Find Max Drinking
9310	Find Max Smoking Cigarette
9309	Find Heather Crying
9302	Find Bradley Holding Paper
9313	Find Stacey Laughing
9303	Find Billy Holding Paper
9308	Find Ian Crying
9306	Find Pat Holding Cloth
9315	Find Max Go Up / Down Stairs
9316	Find Bradley Go Up / Down Stairs

Results by Character - Automatic



Results by Action - Automatic


Boxplot per action of TRECVID 2020 automatic instance search runs

*Mean score of Average Precision by action

#   Action

1 – Smoking Cigarette
2 – Holding Phone
3 – Sit on Couch
4 – Drinking
5 – Laughing
6 – Holding Cloth
7 – Holding Paper
8 - Crying
9 – Go up / down Stairs

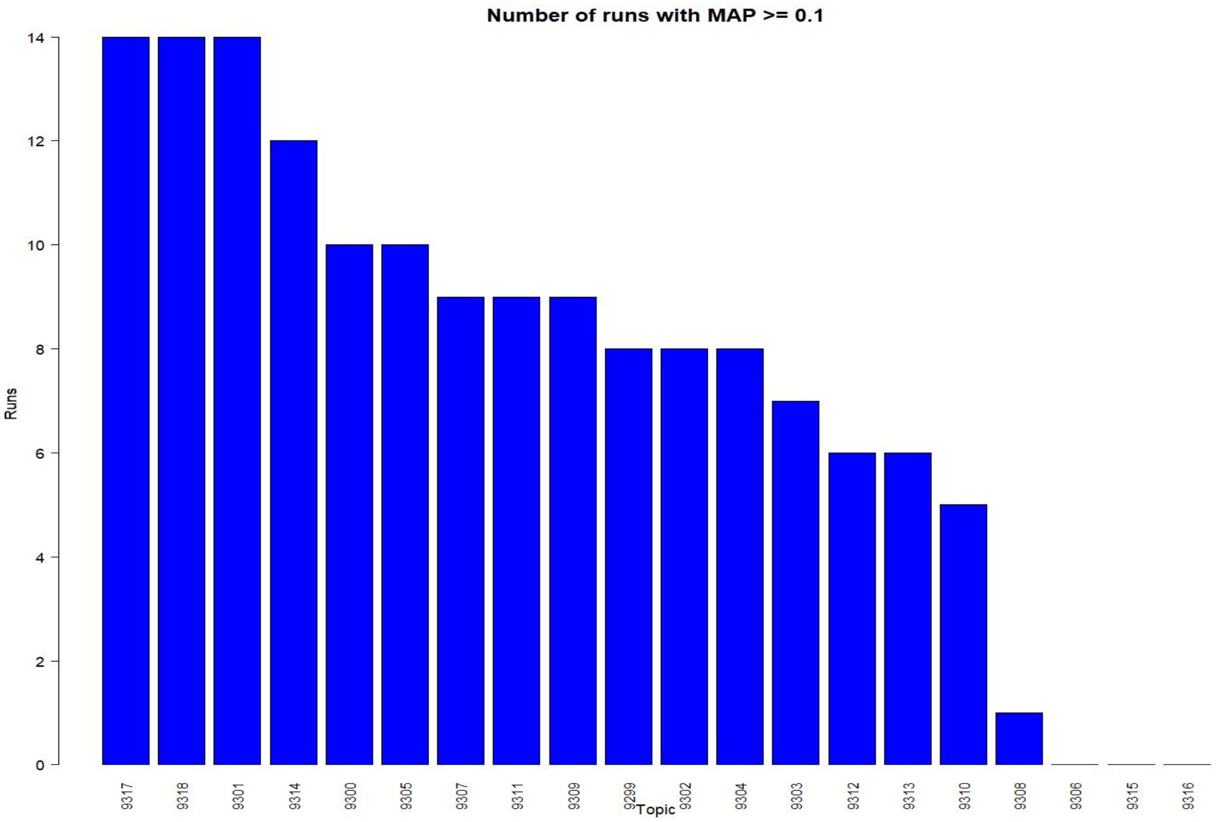
Some Observations…

- Poor results for topics involving Bradley could indicate that he is a difficult character to find. However - previous iterations of the INS task showed him to be among the easiest people to find. What gives?

- Actions involving Bradley consistently score poorly, whether it is Bradley or another character involved. Seems to be more a case of hard actions to identify.

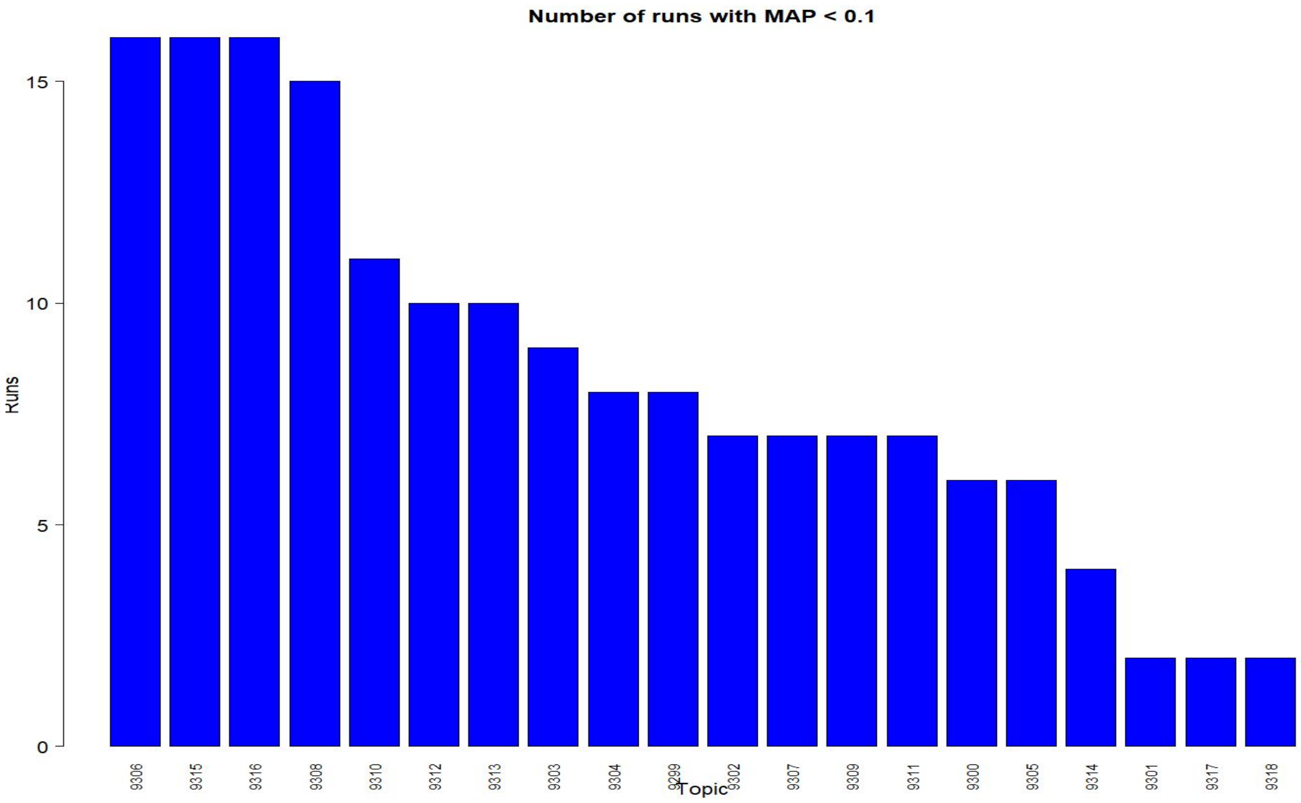
- Bradley Going up/down stairs - very poor results - but looking at frequent false positives on this topic reveal lots of instances of Bradley either standing on stairs or with stairs in the background. Appears to be a case of going up/down stairs being a very difficult action to identify.

Easier Topics



#	Query
9317	Find Max Holding phone
9318	Find Stacey Holding phone
9301	Find Ian Holding Paper
9314	Find Pat Laughing
9300	Find Billy Sit on couch
9305	Find Dot Drinking
9307	Find Heather Holding Cloth
9311	Find Dot Smoking Cigarette
9309	Find Heather Crying
9299	Find Ian Sit on couch
9302	Find Bradley Holding Paper
9304	Find Max Drinking
9303	Find Billy Holding Paper
9312	Find Pat Smoking Cigarette
9313	Find Stacey Laughing
9310	Find Max Smoking Cigarette
9308	Find Ian Crying
9306	Find Pat Holding Cloth
9315	Find Max Go Up / Down Stairs
9316	Find Bradley Go Up / Down Stairs

Harder Topics



Number of runs with MAP < 0.1

#	Query
9306	Find Pat Holding Cloth
9315	Find Max Go Up / Down Stairs
9316	Find Bradley Go Up / Down Stairs
9308	Find Ian Crying
9310	Find Max Smoking Cigarette
9312	Find Pat Smoking Cigarette
9313	Find Stacey Laughing
9303	Find Billy Holding Paper
9304	Find Max Drinking
9299	Find Ian Sit on couch
9302	Find Bradley Holding Paper
9307	Find Heather Holding Cloth
9309	Find Heather Crying
9311	Find Dot Smoking Cigarette
9300	Find Billy Sit on couch
9305	Find Dot Drinking
9314	Find Pat Laughing
9301	Find Ian Holding Paper
9317	Find Max Holding phone
9318	Find Stacey Holding phone

Some Observations…

- From the previous two bar charts we can safely say that holding phone is the easiest topic to find.
- Holding paper and drinking are also among the easiest topics to find.
- Go up / down stairs is the most difficult topic to find.

Some Frequent False Positives

Max Holding phone

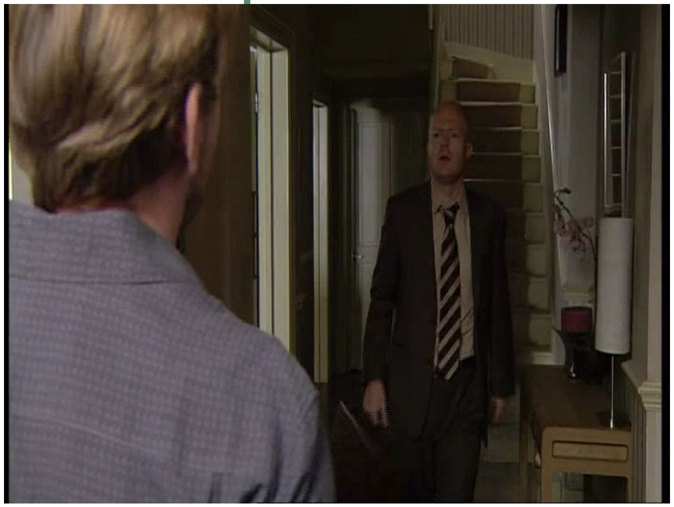
Phil talking on the phone. Phil has been misidentified as Max.

Ian Crying

Ian talking, but with darkened lighting which could be leading to systems misidentifying this as crying.

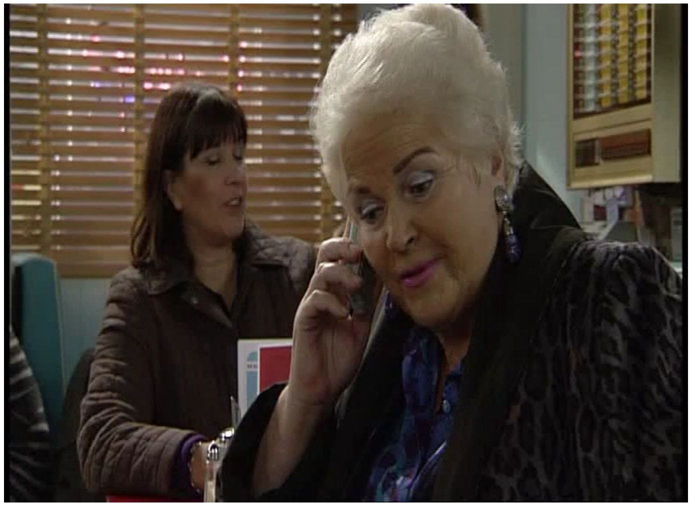
* All images are from the BBC Eastenders dataset licensed to TRECVID and participants by the BBC R&D

Some Frequent False Positives



Max Go up / down stairs

Max walking and talking with stairs in the background. Max's movement with stairs in background could easily be leading systems to identify this as going up or down the stairs.

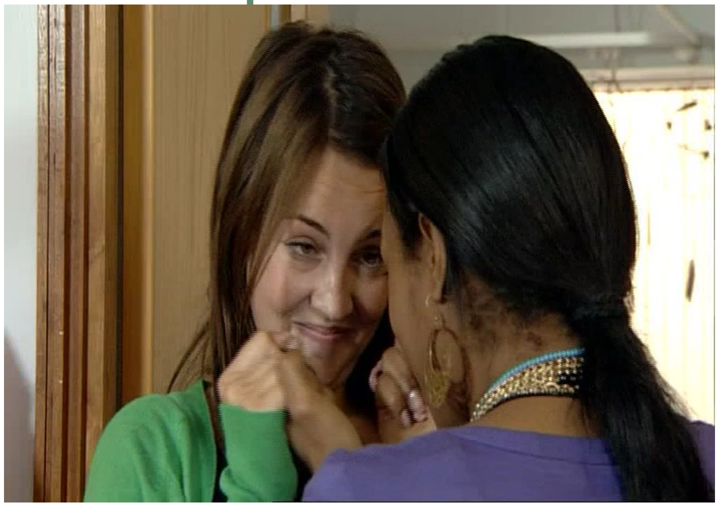


Pat Smoking Cigarette

Pat is talking on the phone with no cigarette visible in the scene.

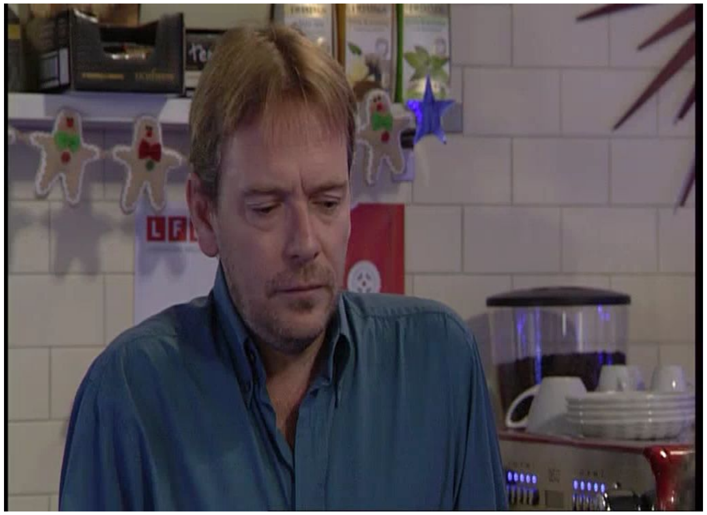
*** All images are from the BBC Eastenders dataset licensed to TRECVID and participants by the BBC R&D**

Some Frequent False Positives



Stacey Laughing

Stacey smiling and appearing to be emotional but not laughing. The smiling and apparent emotion could be leading systems to misidentify as laughter.

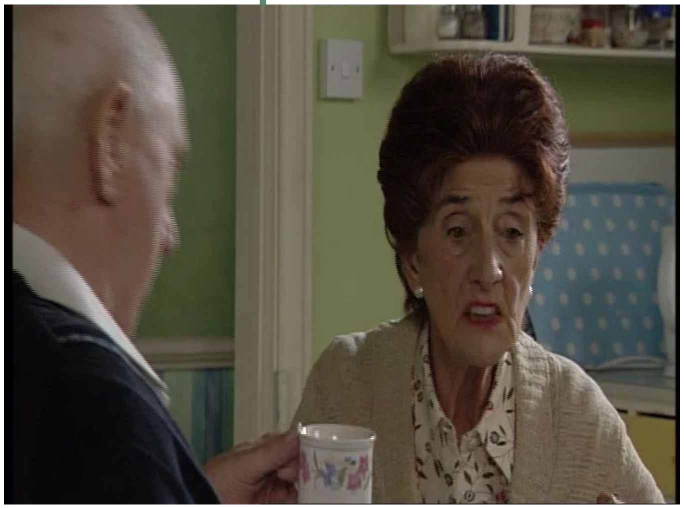


Ian Holding Paper

Ian in shot but is not holding paper. However there is paper on the wall in the background.

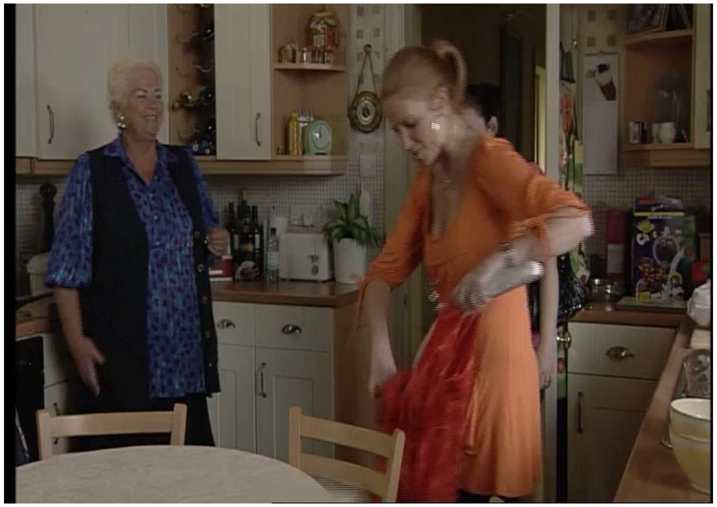
*** All images are from the BBC Eastenders dataset licensed to TRECVID and participants by the BBC R&D**

Some Frequent False Positives



Dot Drinking

Dot talking to Jim while Jim is drinking a cup of tea.



Pat Holding Cloth

Pat is talking with Bianca while Bianca is holding cloth – a dress.

*** All images are from the BBC Eastenders dataset licensed to TRECVID and participants by the BBC R&D**

Some Further Observations from Viewing Most Frequent False Positives of Worst Performing Topics

- Go up / down stairs - Systems tended to classify any shots with target person and a stairway in the background as a positive detection. More work needed on training systems to classify the action itself.

- Smoking – Systems classifying instances of the character holding their hand to their mouth or holding a cigarette in their mouth but not in the process of smoking. Seems to be a case of the smoking action itself being difficult to classify.

- Holding cloth - Fewer conclusions can be drawn. Many instances where target person Pat is misidentified as somebody else who is holding cloth. Also instances where target person Heather is in the scene but somebody else is holding cloth.

Automatic Run Results + Randomization Testing

Top 10 runs across all teams (automatic)

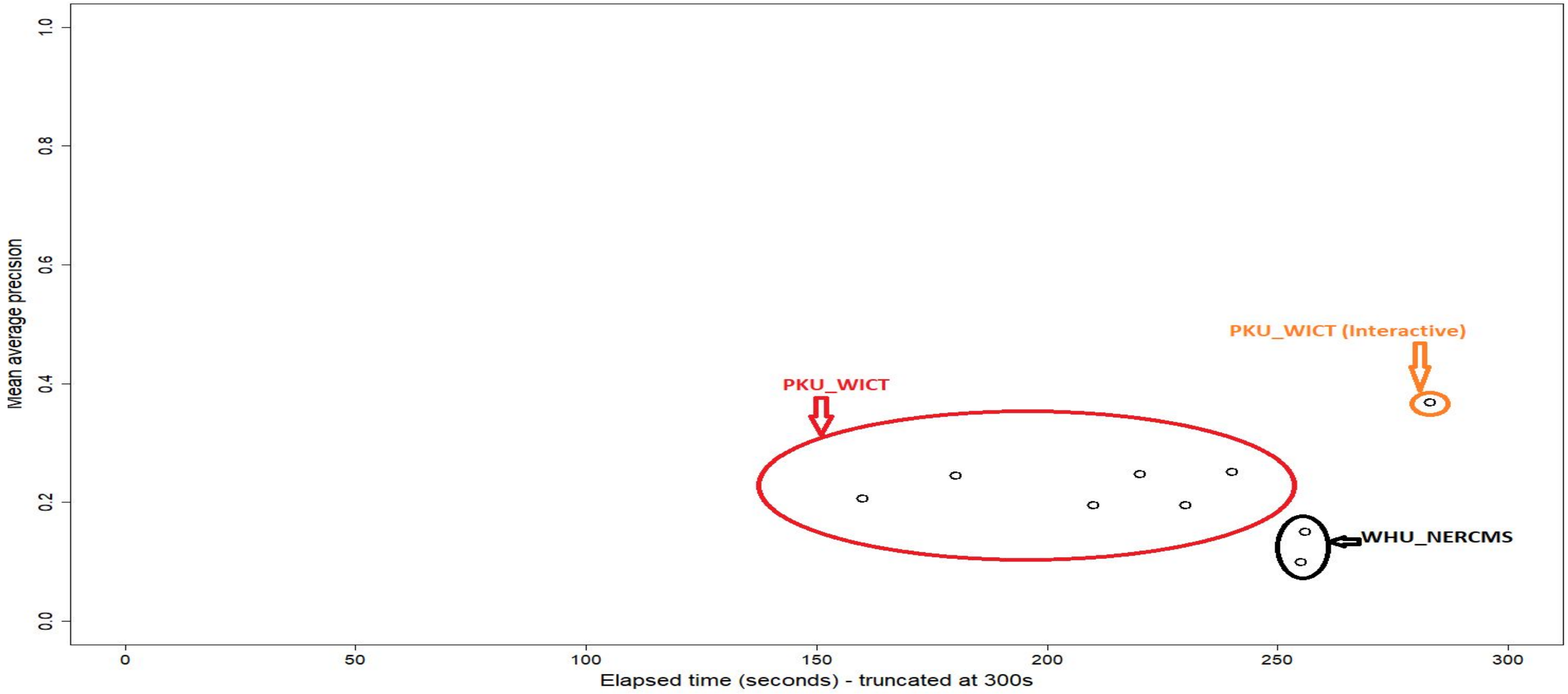
Rank	MAP											
1	0.252	F_M_E_E_PKU_WICT.20_1	=	>	>	>	>	>	>	>	>	>
2	0.247	F_M_A_E_PKU_WICT.20_3*		=		>	>	>	>	>	>	>
3	0.245	F_M_E_E_PKU_WICT.20_4*			=	>	>	>	>	>	>	>
4	0.207	F_M_A_E_PKU_WICT.20_5⤒				=			>	>	>	>
5	0.196	F_M_A_E_PKU_WICT.20_6⤒					=		>	>	>	>
6	0.195	F_M_E_E_PKU_WICT.20_7⤒						=	>	>	>	>
7	0.151	F_M_E_A_WHU_NERCMS.20_1^							=			>
8	0.142	F_M_A_E_BUPT_MCPRL.20_3^								=		>
9	0.135	F_M_A_E_BUPT_MCPRL.20_1†									=	
10	0.099	F_M_E_A_WHU_NERCMS.20_2†										=
			1	2	3	4	5	6	7	8	9	10

*^⤒† = difference not statistically significant

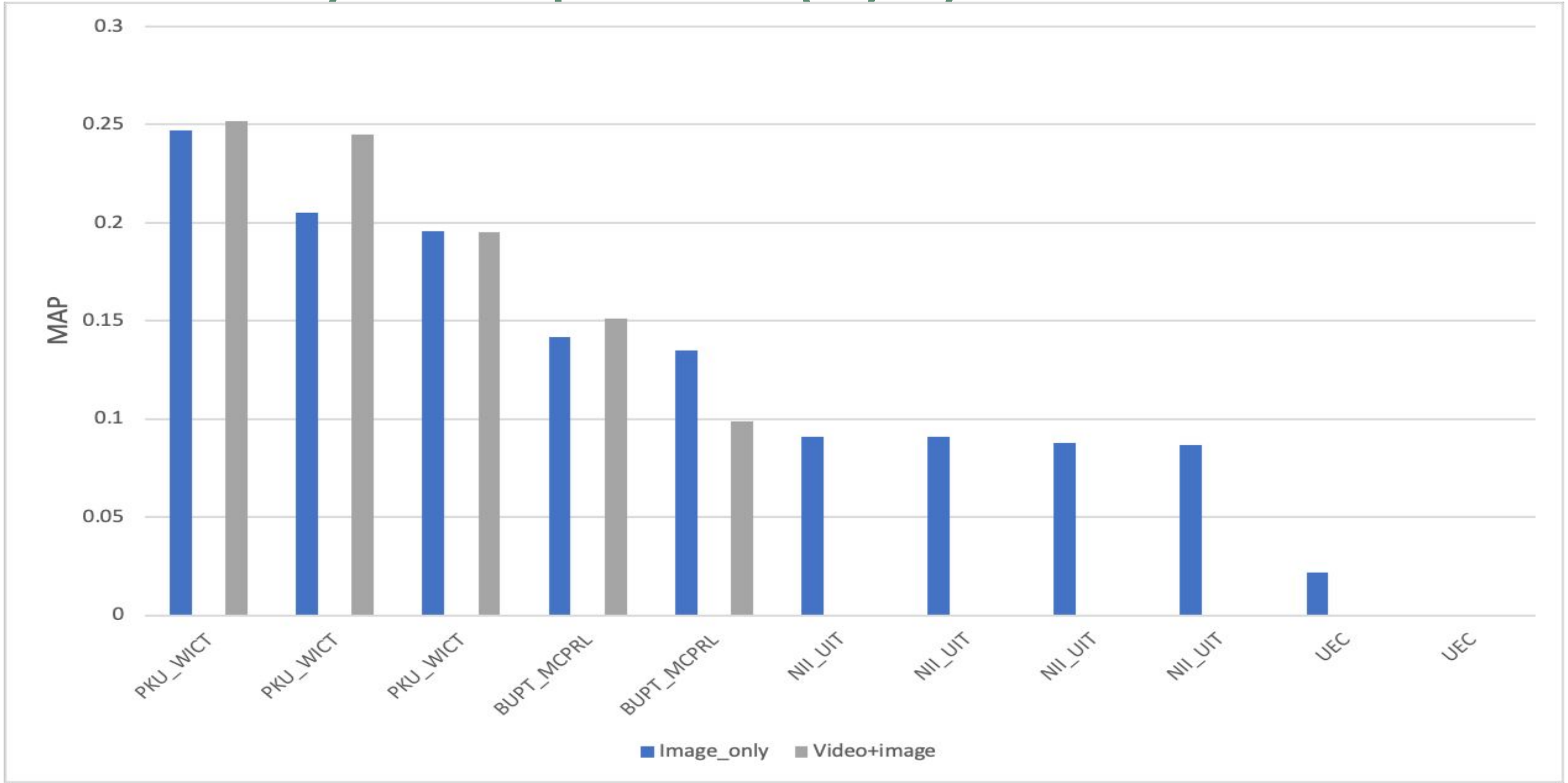
p = probability that row run scored better than the column run due to chance

> $p < 0.05$

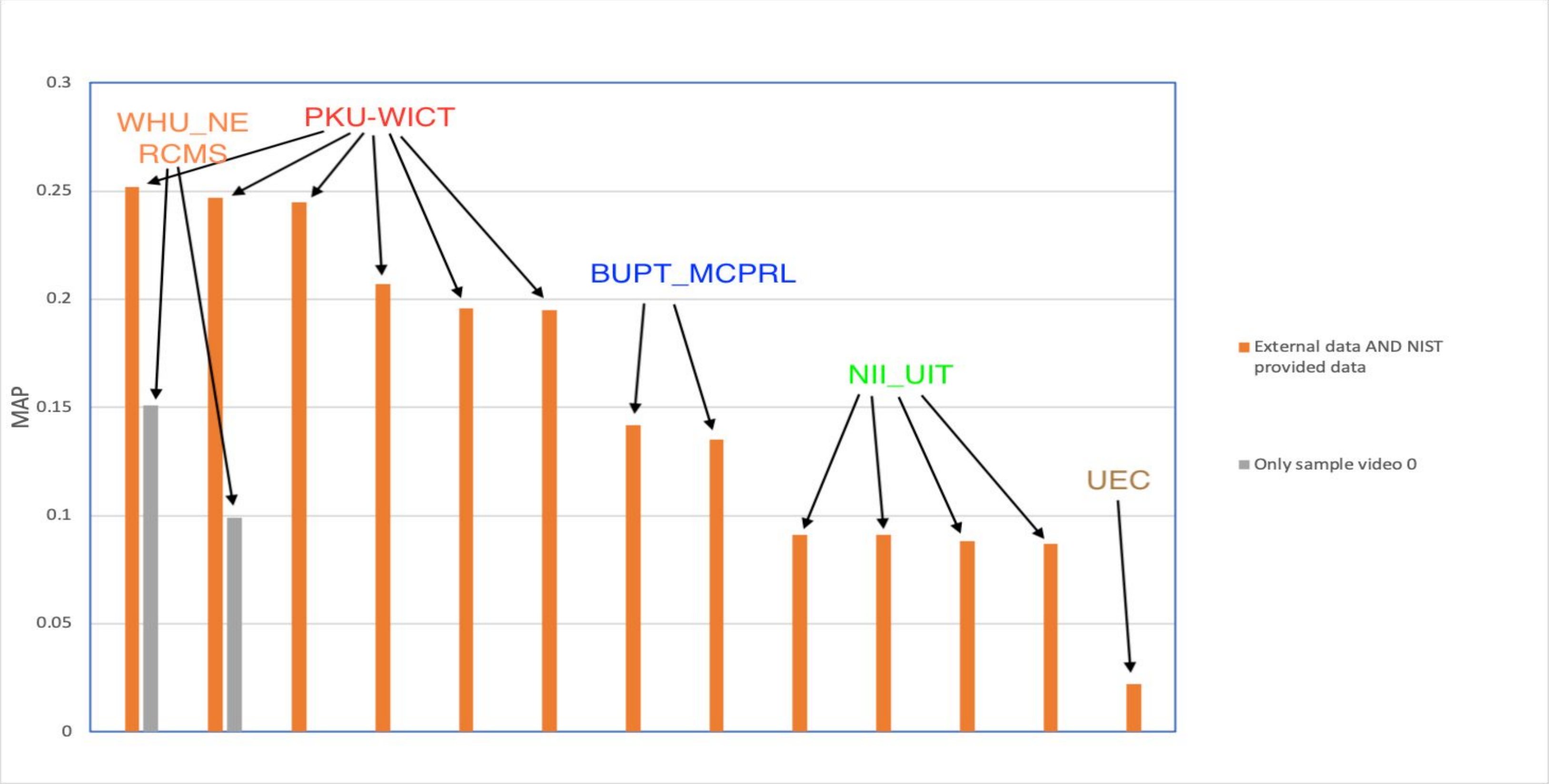
Mean Average Precision vs. Per Run Clock Processing Time



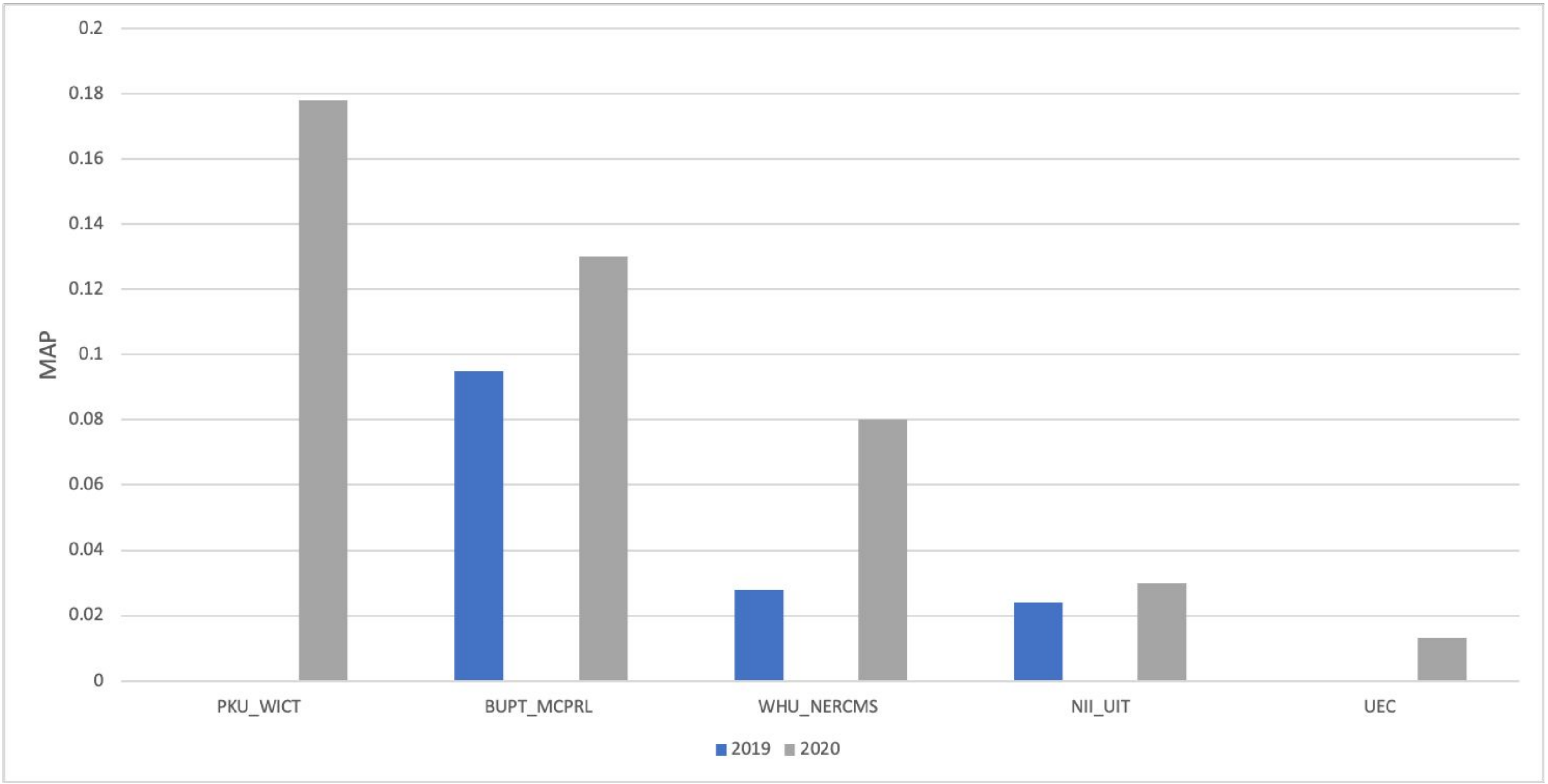
Results by Example Set (A/E) - Automatic



Results by Data Source



Progress Topics 2019-2020 - Results



Progress Topics 2019-2020 - Observations

- Very few teams submitted progress runs for both years of the task.
- Clear that all teams who did saw an improvement in 2020 – encouraging.
- Ideally all teams who submitted progress runs this year will do so again in 2021, allowing more solid conclusions to be drawn.

Some General Observations About the Task

- Slight decrease in number of participants and finishers, but higher % of participants finished the task.

- Many more teams now using E condition - training with video examples. Perhaps more necessary now with action recognition. But - Results from teams using both show little difference between image & video and image only!

- Interactive search task:
  - Limited participation – only one interactive run this year.

Some General Observations About the Task – Data Source

- Best results by far achieved using external development data (collected by teams) plus NIST provided data.
- Huge gap to results from systems trained using only external data or using only sample video 0.

Further Conclusions

- Person recognition has been a feature of the INS task since 2013 and is very mature by this stage. Very few frequent false positives misidentify the person.
- Action recognition is a new feature of INS task. The much increased difficulty of the new INS task is due to this. Requires much more work to reach an acceptable level of maturity.

Further Conclusions

- Visual Concepts very important.

- Easier tasks mostly those with obvious visual context (sit on couch, hold phone, hold paper, etc.)

- Harder tasks tend to be more independent from obvious visual context (crying, smoking, go up / down stairs hard to isolate from other scenes showing a set of stairs).

Approach summary of teams not presenting

- **University of Electro Communications – Tokyo**
  - #1 participation: baseline system
  - Combination of face detection, emotion, human object interaction and general action detection (no action specific training)
  - mAP 0.022 (A/F)
- **National Institute of Informatics /University of Information Technology (Vietnam)**
  - Reused 2019 face detector
  - Focus on removing false positives for action topics (wrong person)
  - Heuristic distance based method to assign face scores for each frame in a shot
  - Improved results on the Progress Set