TRECVID 2021 INSTANCE RETRIEVAL INTRODUCTION AND TASK OVERVIEW

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.

Table of Contents

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

Task

- ✓ 2013 – 2015

  The task asked systems to find a specific object, person or location in any context using a small set of image and video examples.

- ✓ 2016 - 2018

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

- ✓ 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

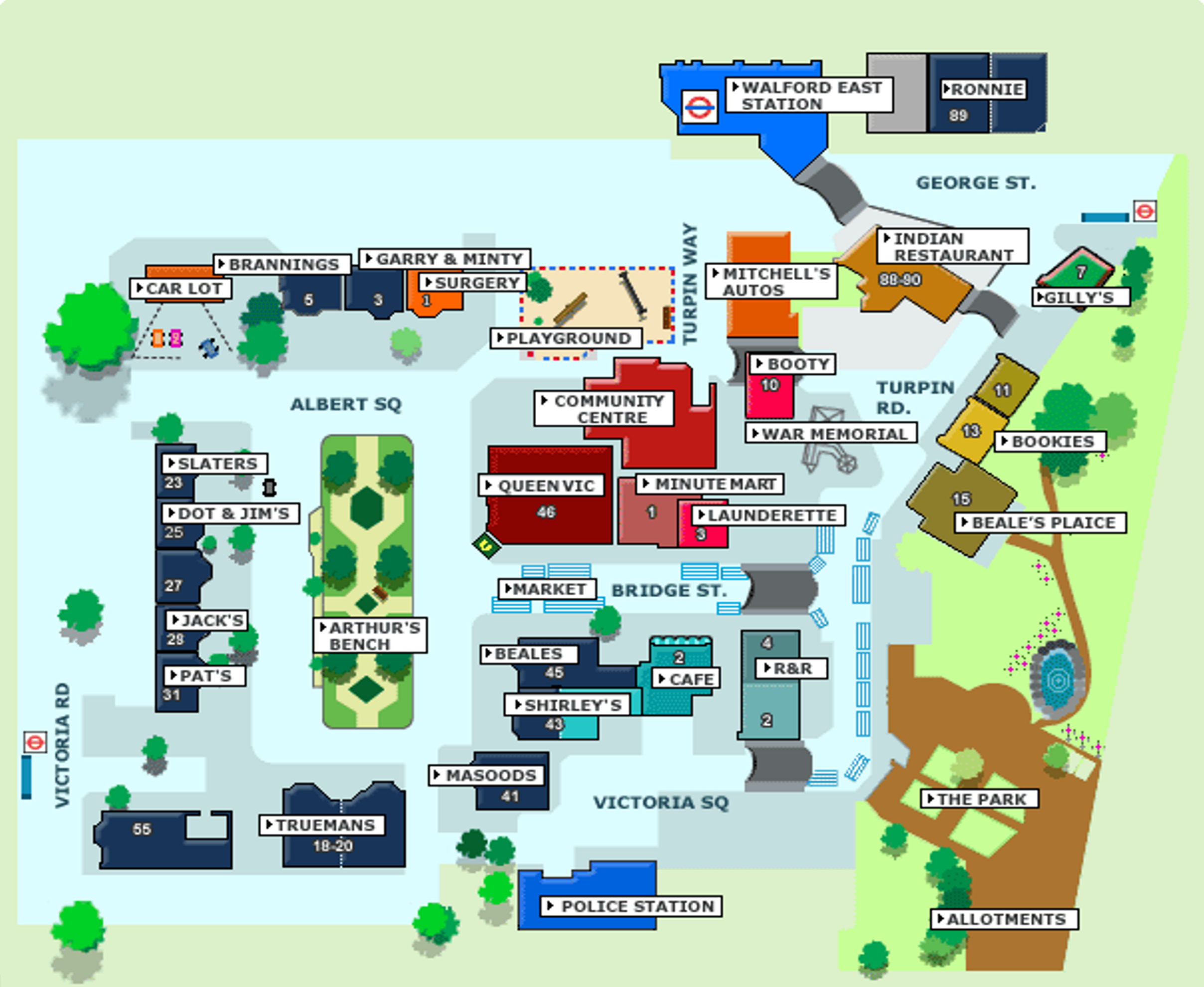
  Automatic or **interactive** runs are accepted

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” 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 2021 and 20 common topics for 2019 - 2021. Each topic includes images for target persons and example videos of the specific actions.
- Filtered example shots from the submissions if it satisfies the topic.

Global Test Condition: Type/source of Training Data

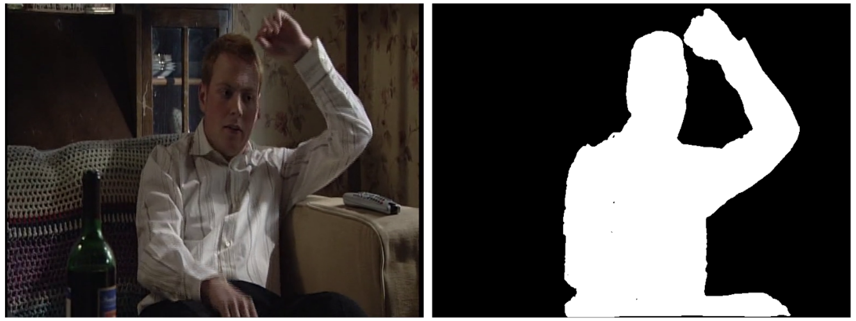
Effect of examples – 2 conditions:

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

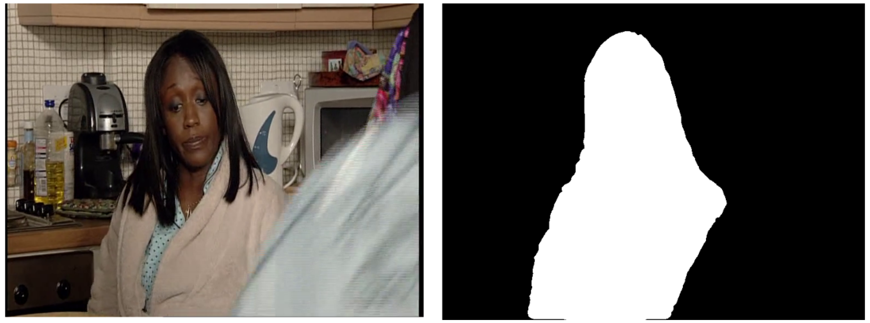
Sources of Training Data:

- 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)

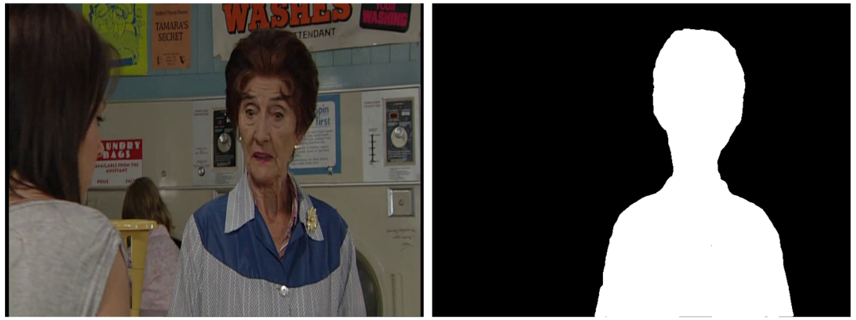
Topics – segmented ‘person’ example images



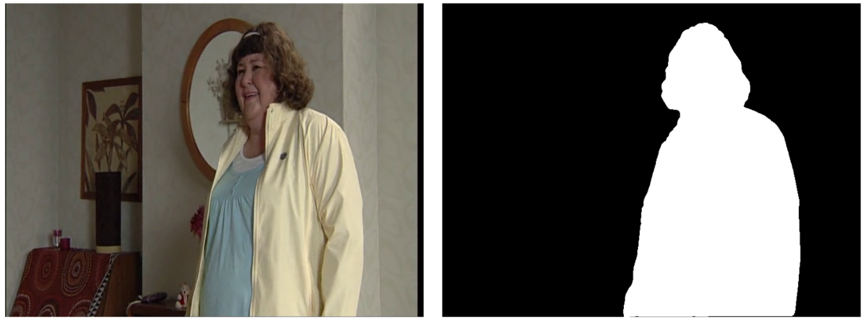
Bradley



Denise

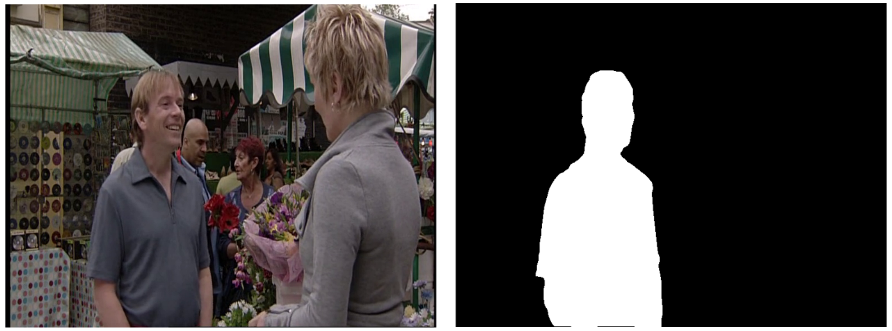


Dot

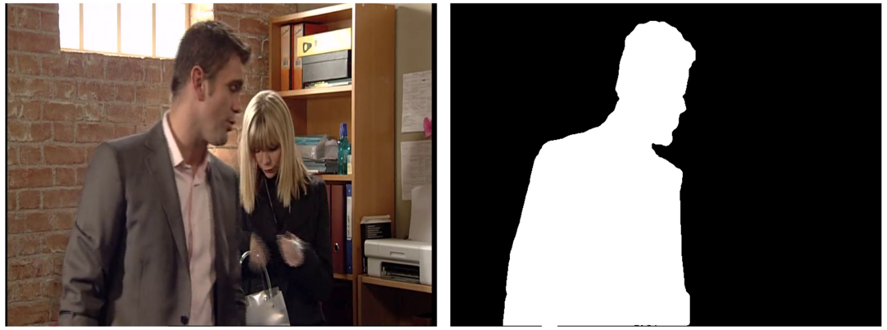


Heather

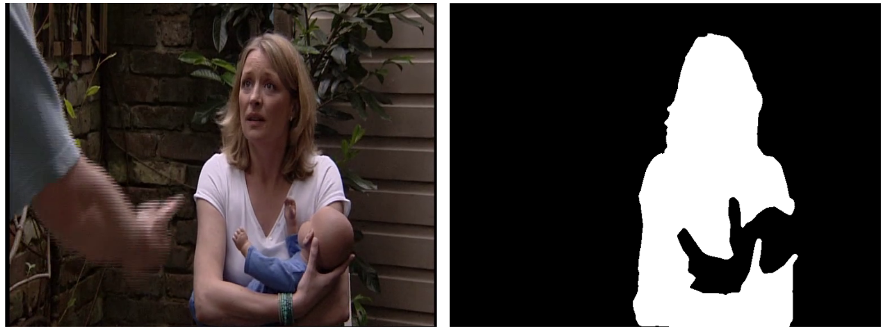
Topics – segmented ‘person’ example images



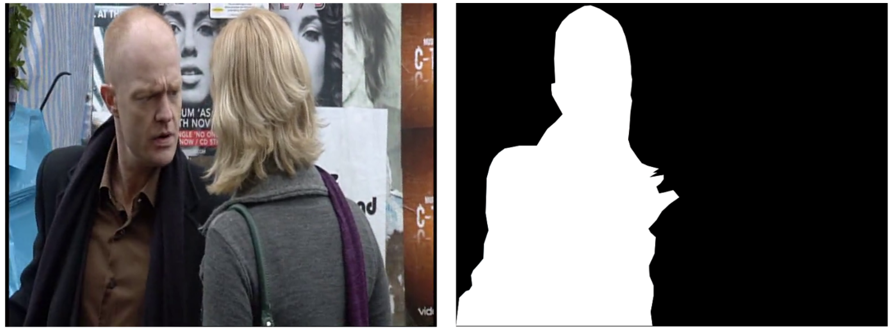
Ian



Jack

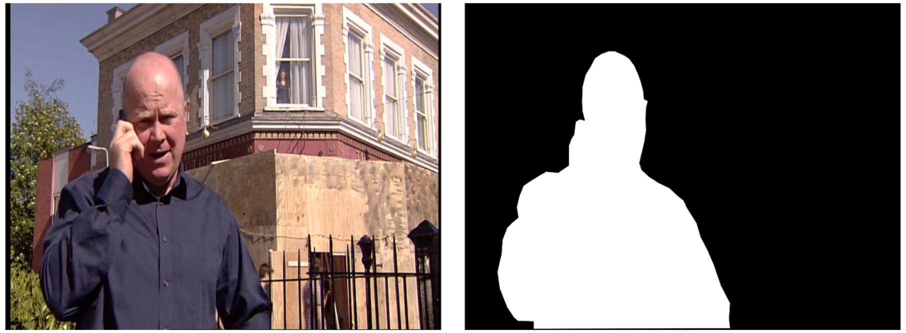


Jane

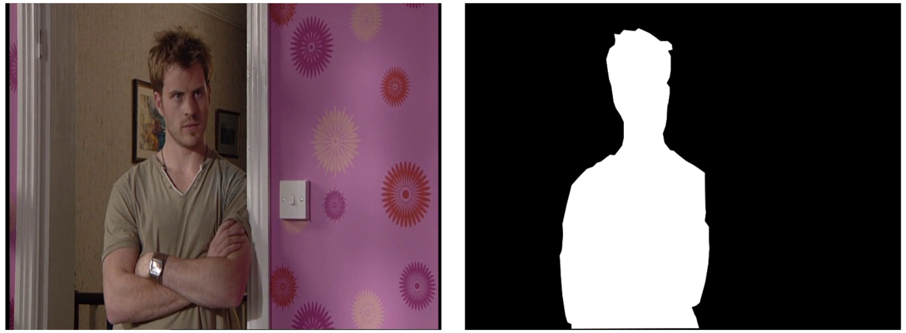


Max

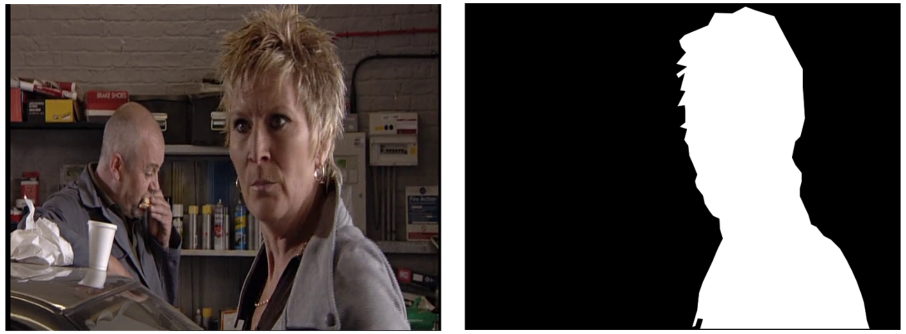
Topics – segmented 'person' example images



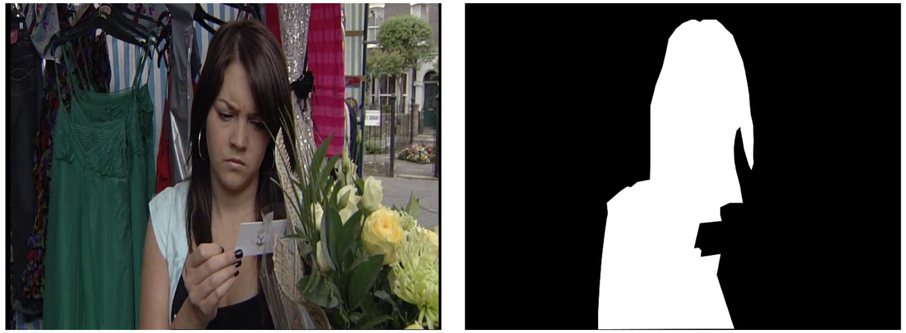
Phil



Sean



Shirley



Stacey

Sample Actions

Open door & enter

Sit on couch

Sample Actions

Drinking

Hugging

20 Unique Queries: 2021

	Max	Pat	Shirley	Bradley	Peggy	Stacey
Holding glass			X	X		X
Sit on couch	X				X	X
Holding cloth	X					X
Carrying bag	X				X	
Kissing	X					X
Holding phone			X	X	X	
Holding paper		X	X		X	
Open door and enter		X		X		

20 x unique queries : find {Max, Pat, Shirley, Bradley, Peggy, Stacey} doing {Holding glass, Sit on couch, Holding cloth, Carrying bag, Kissing, Holding phone, Holding paper, Open door and enter}

20 Common Queries: 2019 - 2021

	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 & leave		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 2021: 3 Finishers (out of 11)

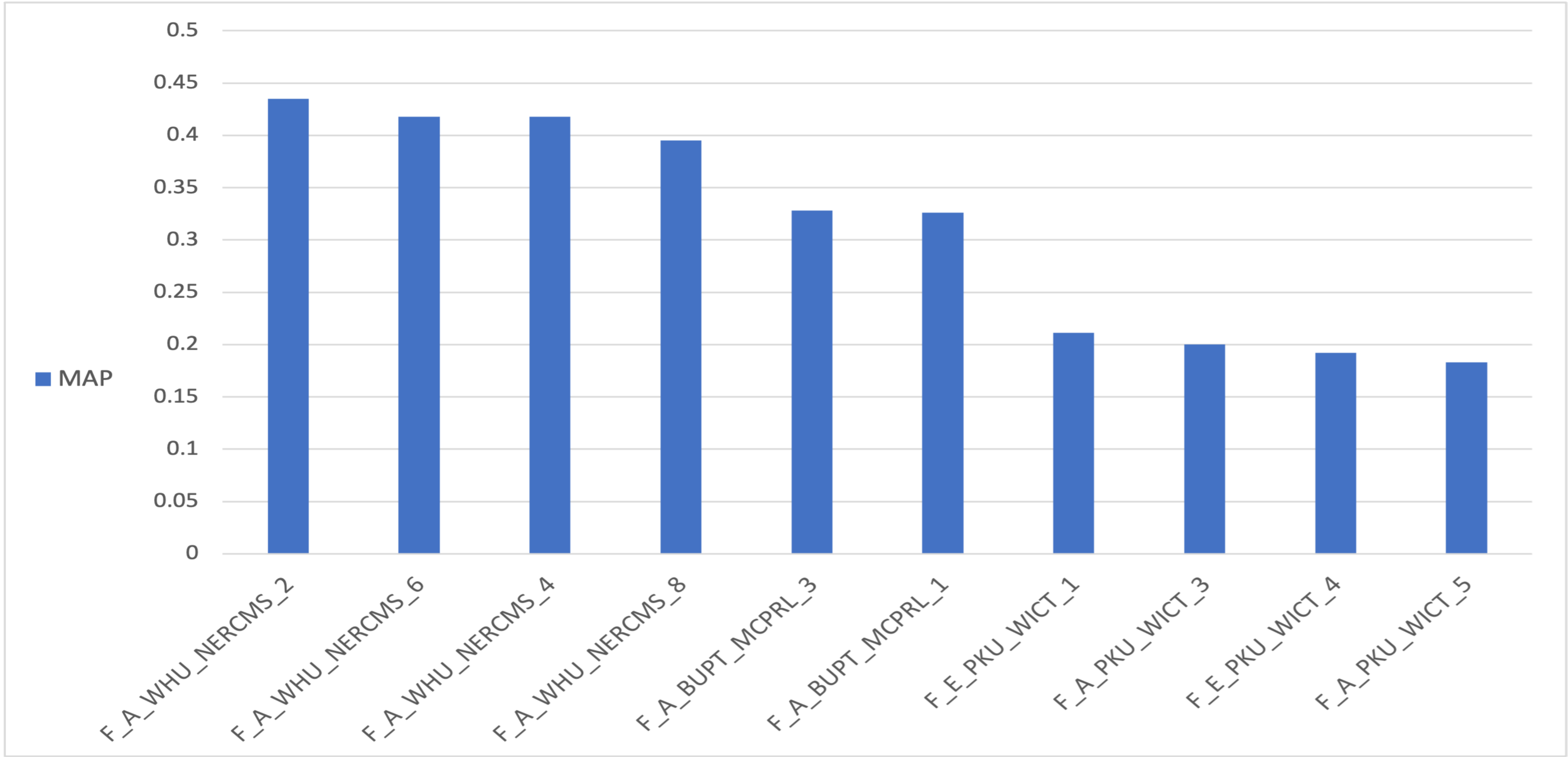
Team	Organization	Run Types F: automatic, I: Interactive (Main Task)	Run Types F: automatic, I: Interactive (Progress Task)
BUPT_MCPRL	Beijing University of Posts and Telecommunications	F_A (2)	F_A(2)
PKU_WICT	Peking University	F_A (2), F_E (2), I_E (1)	F_A (1), F_E (1), I_E(1)
WHU_NERCMS	Wuhan University	F_A (4), I_A(4)	F_A(2), I_A(4)

Evaluation

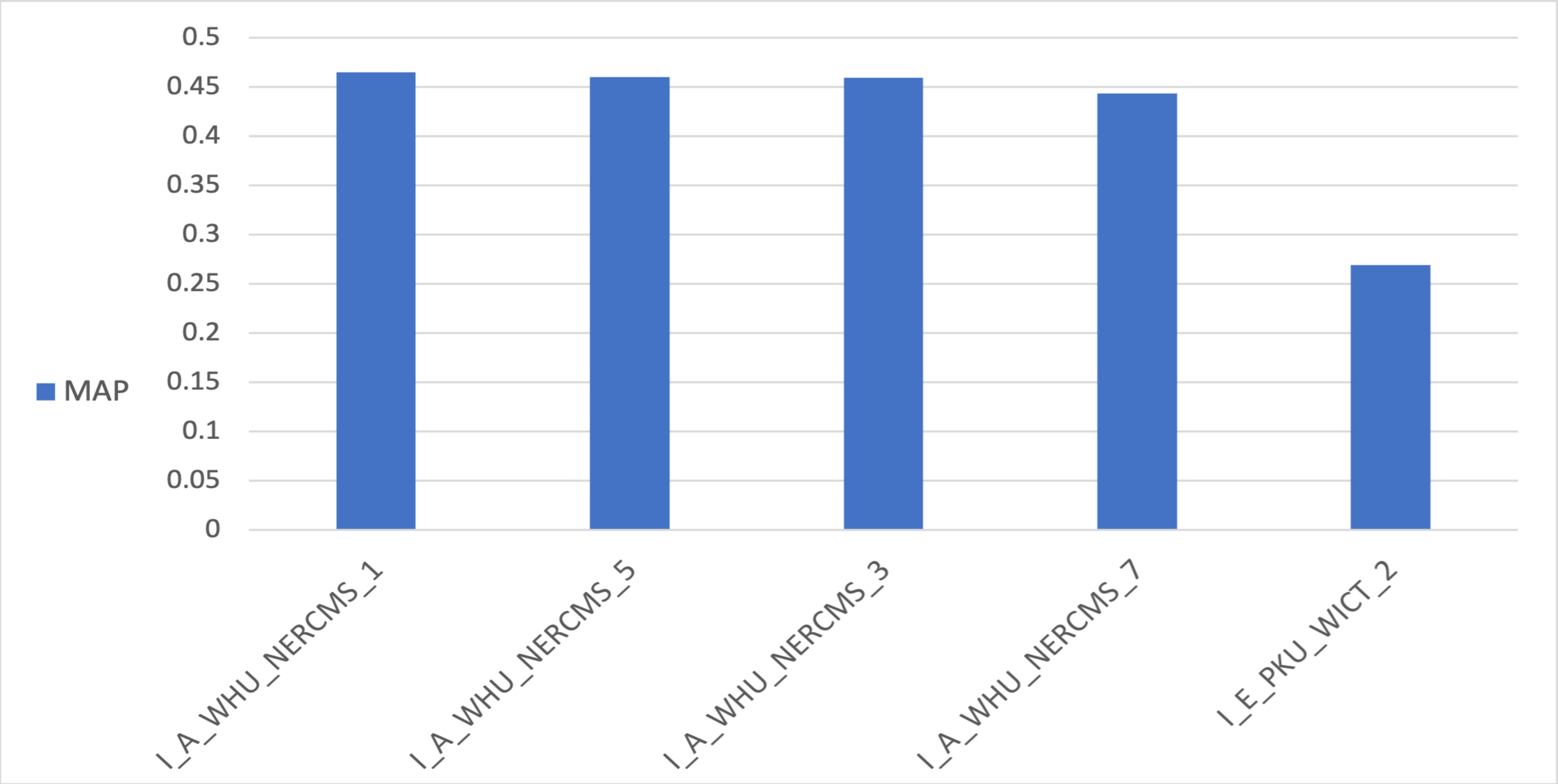
For each topic the submissions were pooled and judged down to max rank 800, resulting in 76305 judged shots (≈ 300 person-h).

- 10 NIST assessors played the clips and determined if they contained the topic target or not. Each assessor judged 3 separate topics.
- 8 508 clips (avg. 283.6 / topic) contained the topic target (11.15 %)
- True positives per topic: min 36 med 233 max 1024
- 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.
- In total, 10 automatic and 5 interactive runs were submitted.

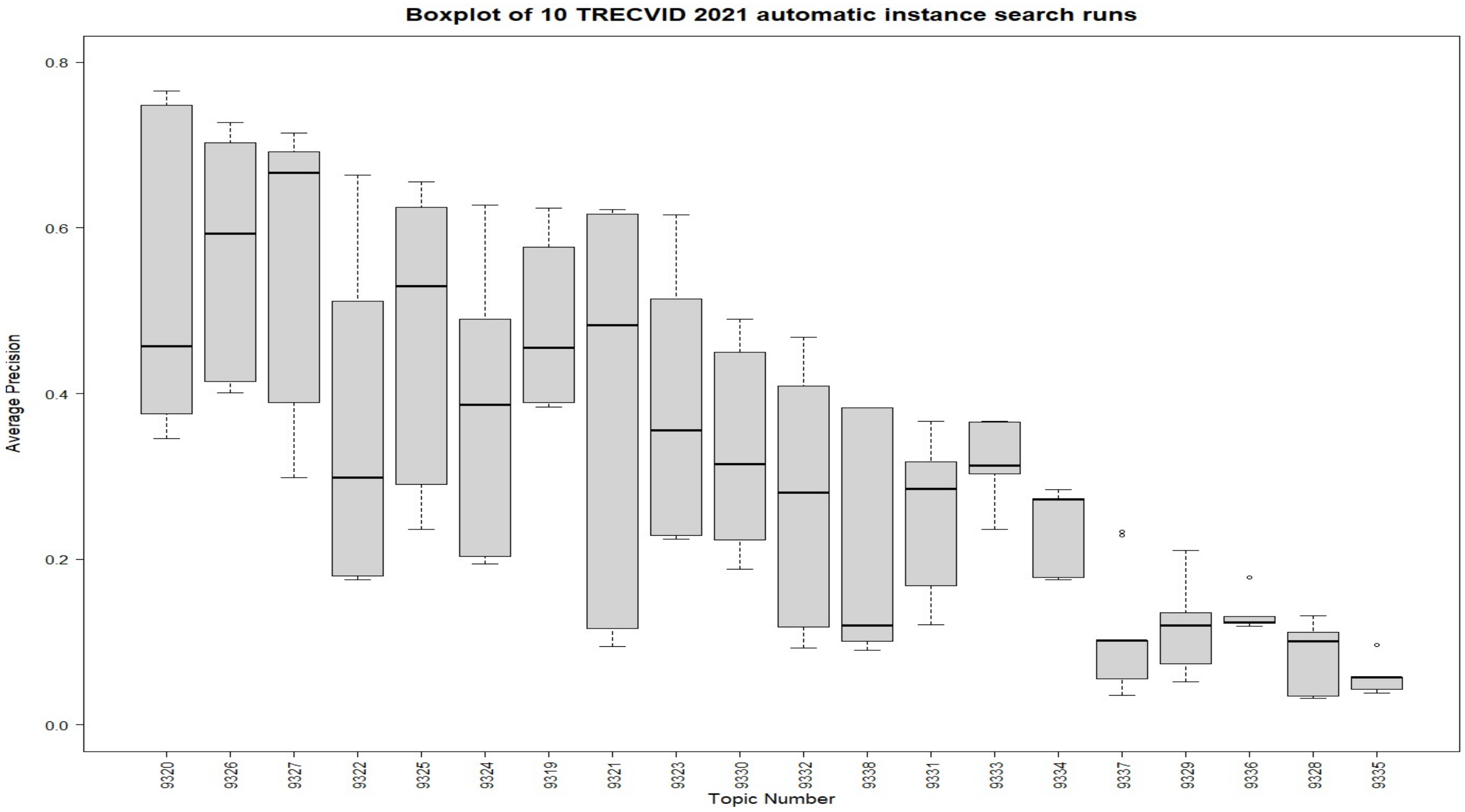
Results by Team (Automatic)



Results by Team (Interactive)



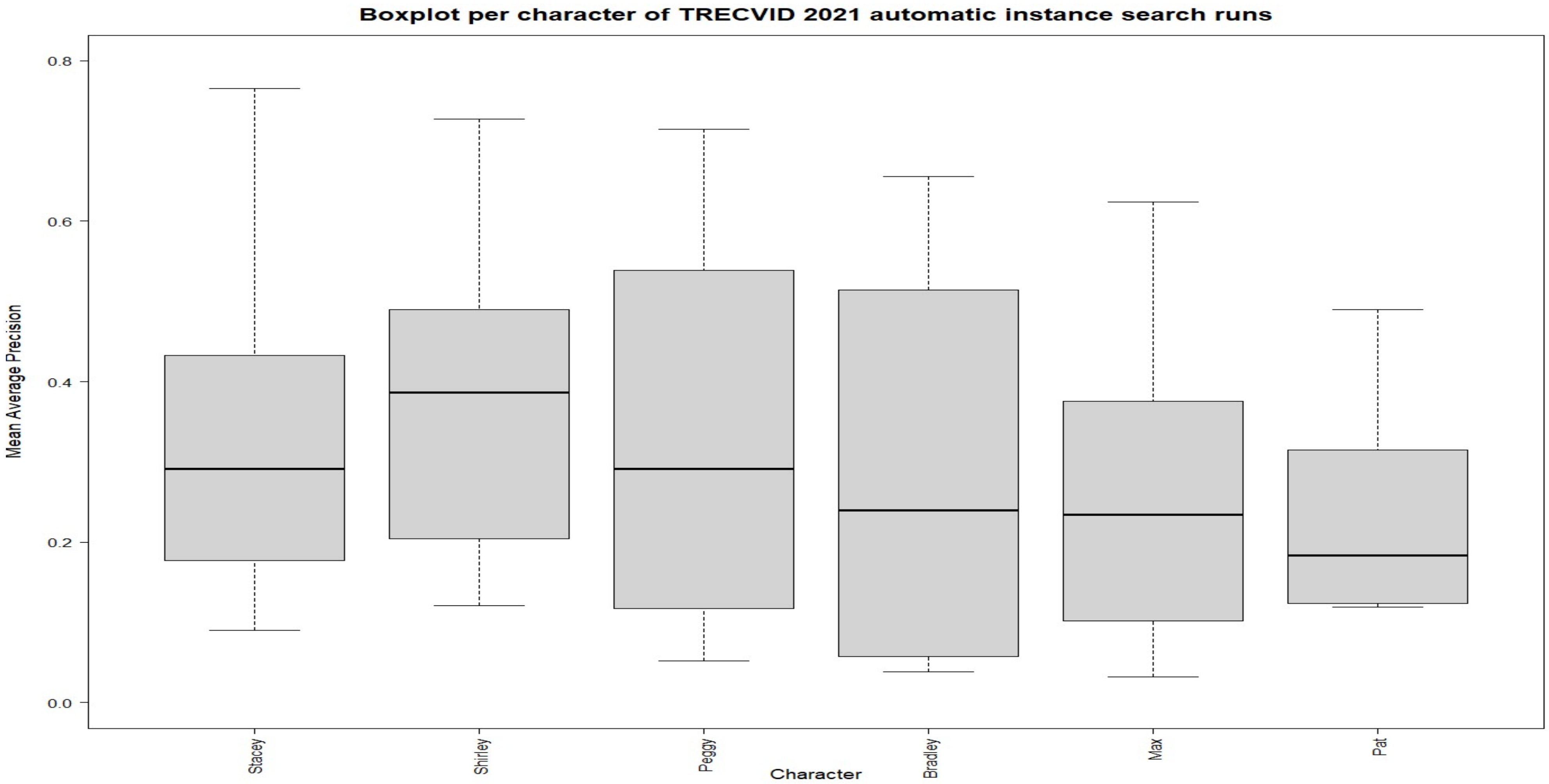
Results by Topics - Automatic



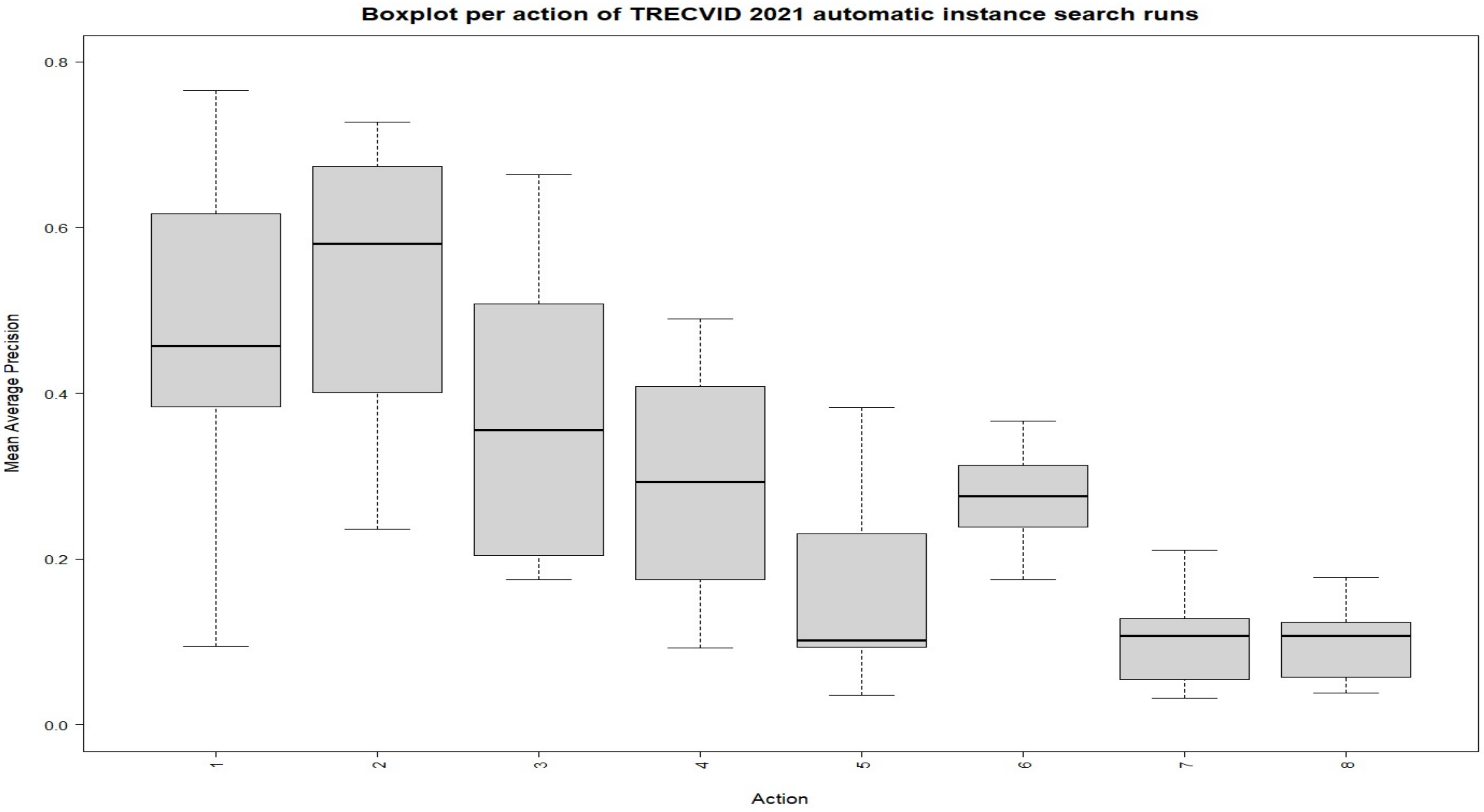
*Mean score of Average Precision per character/action

#	Query
9320	Find Stacey Sit on couch
9326	Find Shirley Holding Phone
9327	Find Peggy Holding Phone
9322	Find Stacey Holding glass
9325	Find Bradley Holding Phone
9324	Find Shirley Holding glass
9319	Find Max Sit on couch
9321	Find Peggy Sit on couch
9323	Find Bradley Holding glass
9330	Find Pat Holding Paper
9332	Find Peggy Holding Paper
9338	Find Stacey Holding Cloth
9331	Find Shirley Holding Paper
9333	Find Max Kissing
9334	Find Stacey Kissing
9337	Find Max Holding Cloth
9329	Find Peggy Carrying Bag
9336	Find Pat Open Door & Enter
9328	Find Max Carrying Bag
9335	Find Bradley Open Door & Enter

Results by Character - Automatic



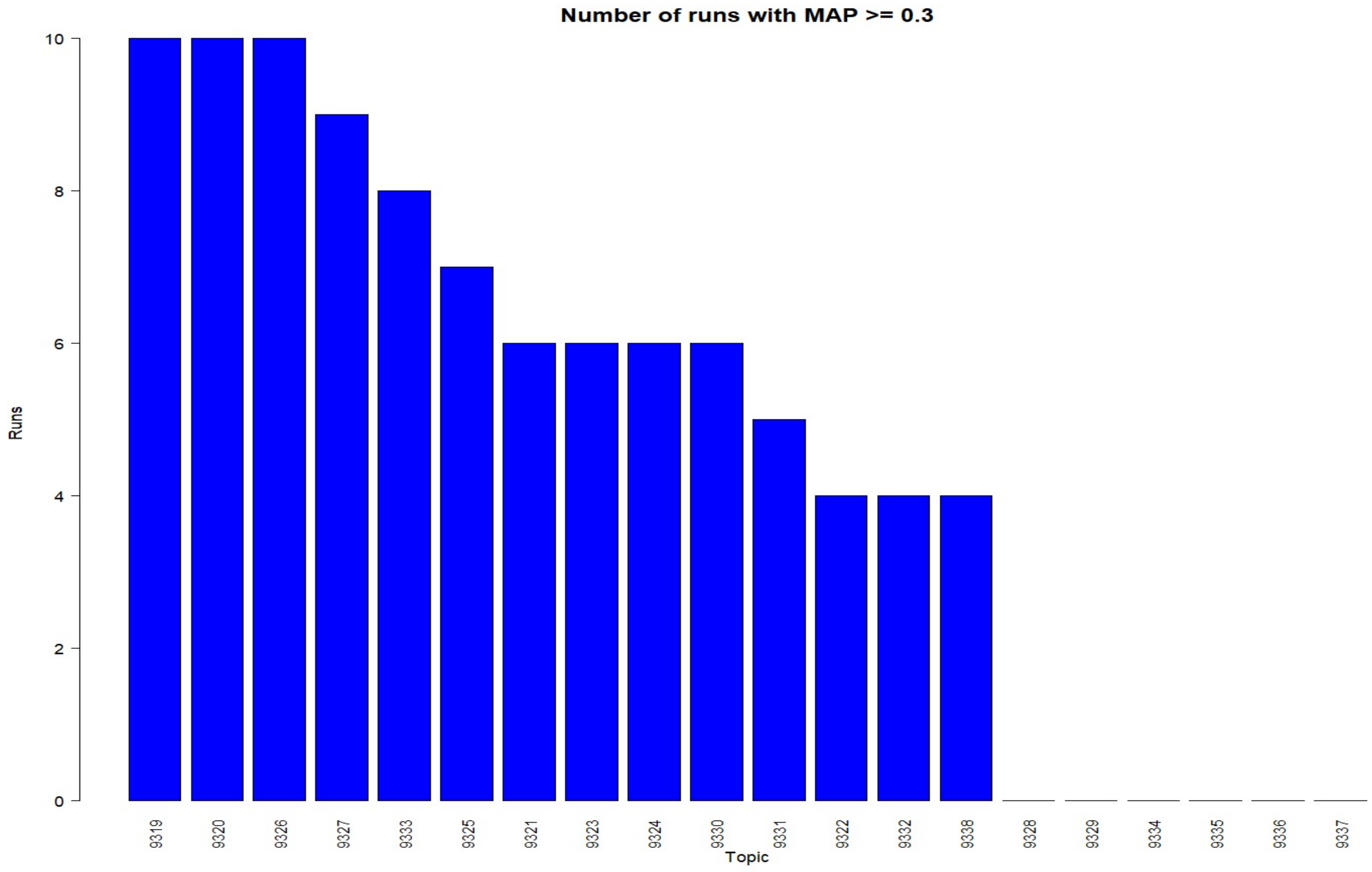
Results by Action - Automatic



*Mean score of Average Precision by action

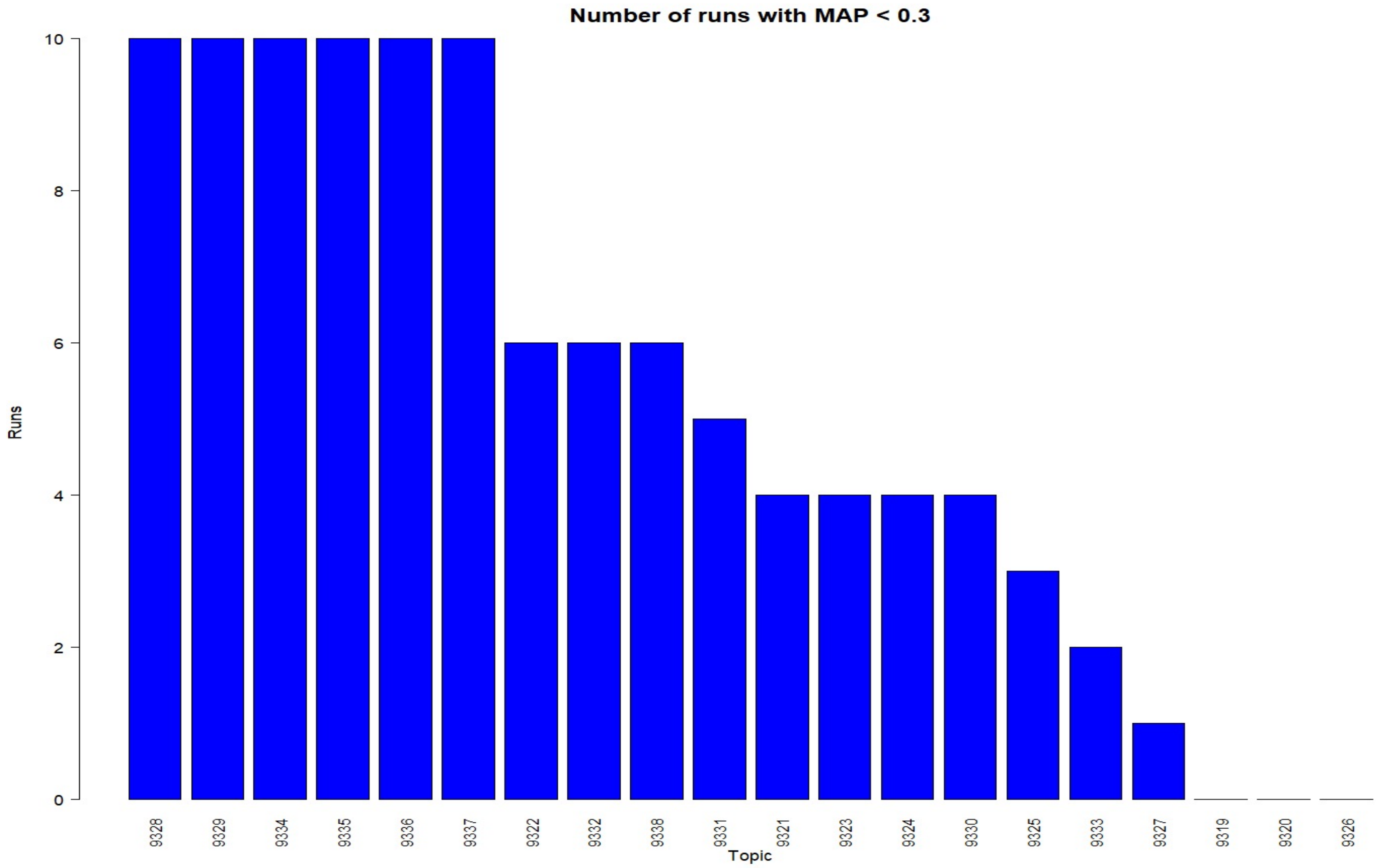
#	Action
1 –	Sit on Couch
2 –	Holding Phone
3 –	Holding Glass
4 –	Holding Paper
5 –	Holding Cloth
6 –	Kissing
7 –	Carrying Bag
8 –	Open Door & Enter

Easier Topics - Automatic



#	Query
9319	Find Max Sit on couch
9320	Find Stacey Sit on couch
9326	Find Shirley Holding Phone
9327	Find Peggy Holding Phone
9333	Find Max Kissing
9325	Find Bradley Holding Phone
9321	Find Peggy Sit on couch
9323	Find Bradley Holding glass
9324	Find Shirley Holding glass
9330	Find Pat Holding Paper
9331	Find Shirley Holding Paper
9322	Find Stacey Holding glass
9332	Find Peggy Holding Paper
9338	Find Stacey Holding Cloth
9328	Find Max Carrying Bag
9329	Find Peggy Carrying Bag
9334	Find Stacey Kissing
9335	Find Bradley Open Door & Enter
9336	Find Pat Open Door & Enter
9337	Find Max Holding Cloth

Harder Topics - Automatic

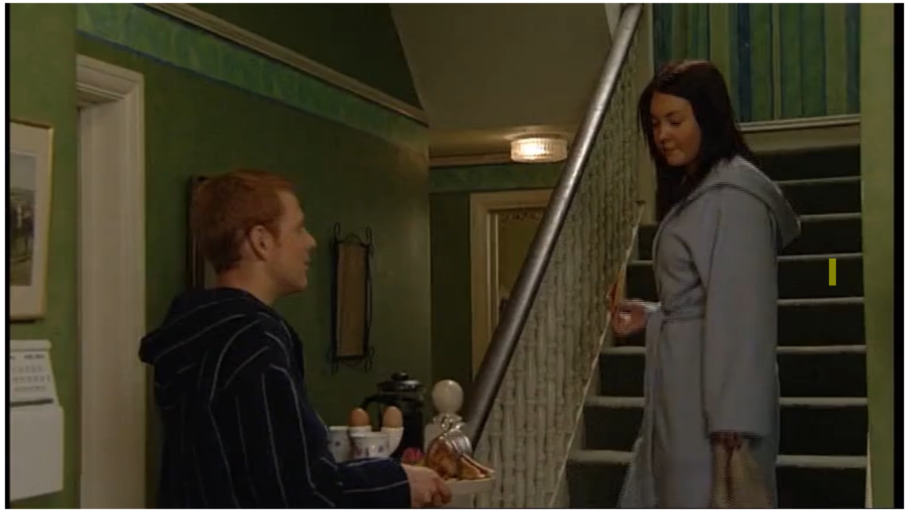


#	Query
9328	Find Max Carrying Bag
9329	Find Peggy Carrying Bag
9334	Find Stacey Kissing
9335	Find Bradley Open Door & Enter
9336	Find Pat Open Door & Enter
9337	Find Max Holding Cloth
9322	Find Stacey Holding glass
9332	Find Peggy Holding Paper
9338	Find Stacey Holding Cloth
9331	Find Shirley Holding Paper
9321	Find Peggy Sit on couch
9323	Find Bradley Holding glass
9324	Find Shirley Holding glass
9330	Find Pat Holding Paper
9325	Find Bradley Holding Phone
9333	Find Max Kissing
9327	Find Peggy Holding Phone
9319	Find Max Sit on couch
9320	Find Stacey Sit on couch
9326	Find Shirley Holding Phone

Some Observations…

- From the previous charts we can safely say that Holding phone, Sit on couch, and Holding glass are the easiest topics to find.
- Carrying bag, and Open door & enter are the most difficult topics to find.

Some Frequent False Positives



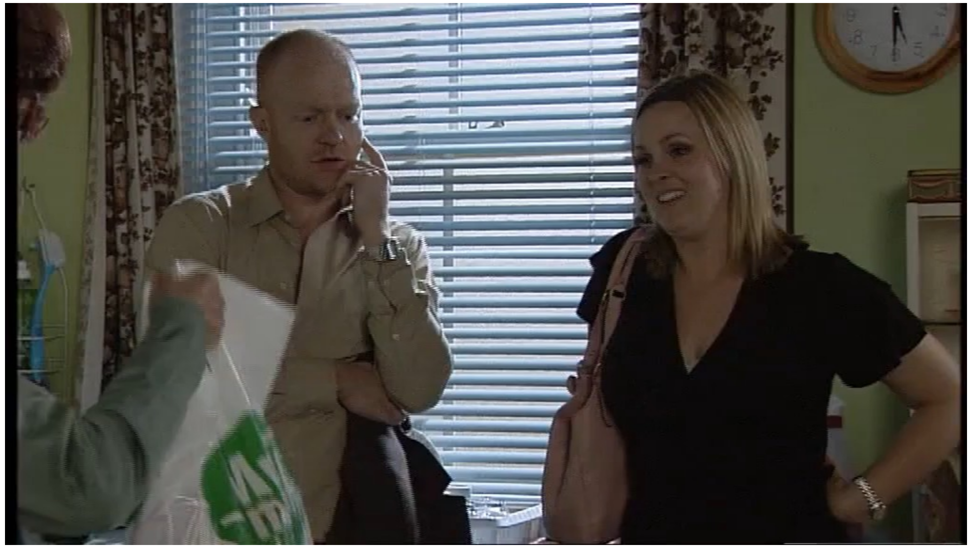
Bradley Holding Glass

Bradley is holding a tray instead with
cups on it

Bradley Holding Phone

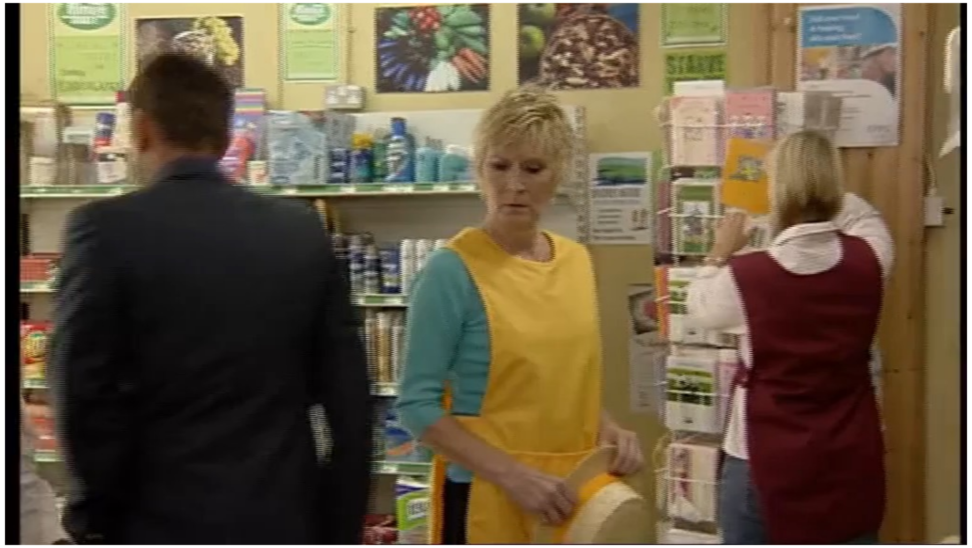
Bradley holds a joystick
(looks like a phone!)

Some Frequent False Positives



Max Carrying Bag

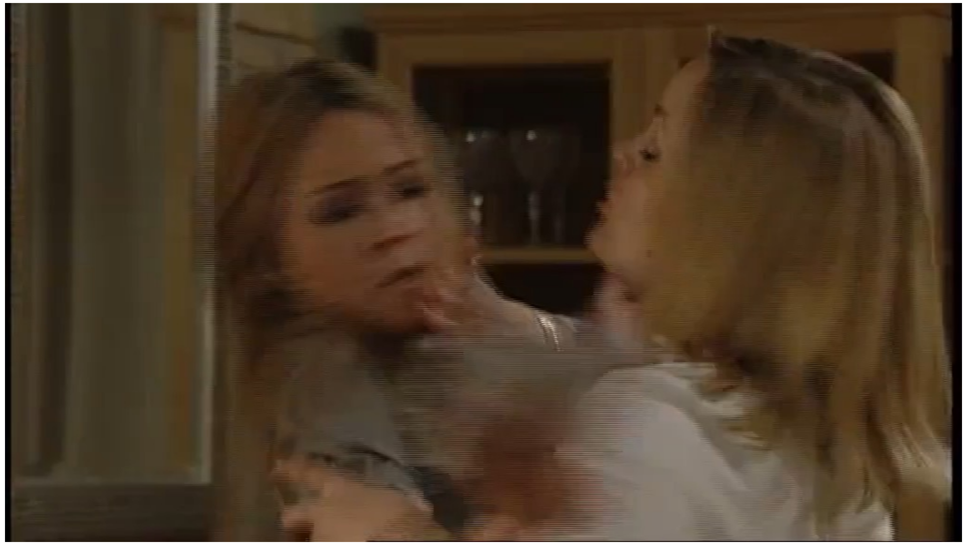
Max holding his jacket. But a person in front of him is showing a plastic bag



Shirley Holding Paper

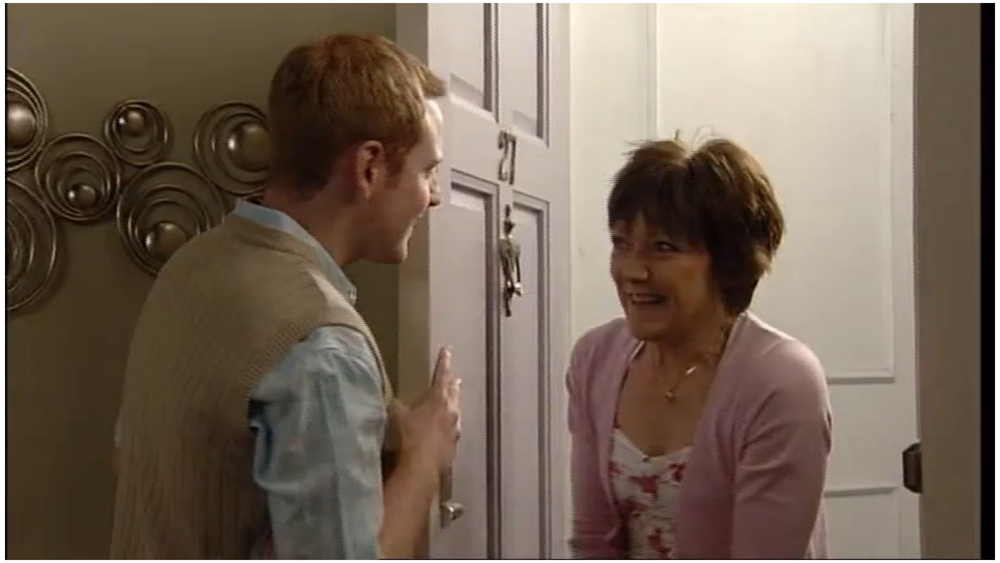
Shirley is holding a hat. However, there are paper material in the background

Some Frequent False Positives



Stacey Kissing

Stacey is shouting but is in very close proximity to Tanya



Bradley opens door and enter

Bradley opens the door but he is inside the room.

Some Frequent False Positives

Max Holding Cloth

Max is holding newspaper

Stacey Holding Cloth

Stacey holds purse, and duct tape

Automatic Run Results + Randomization Testing

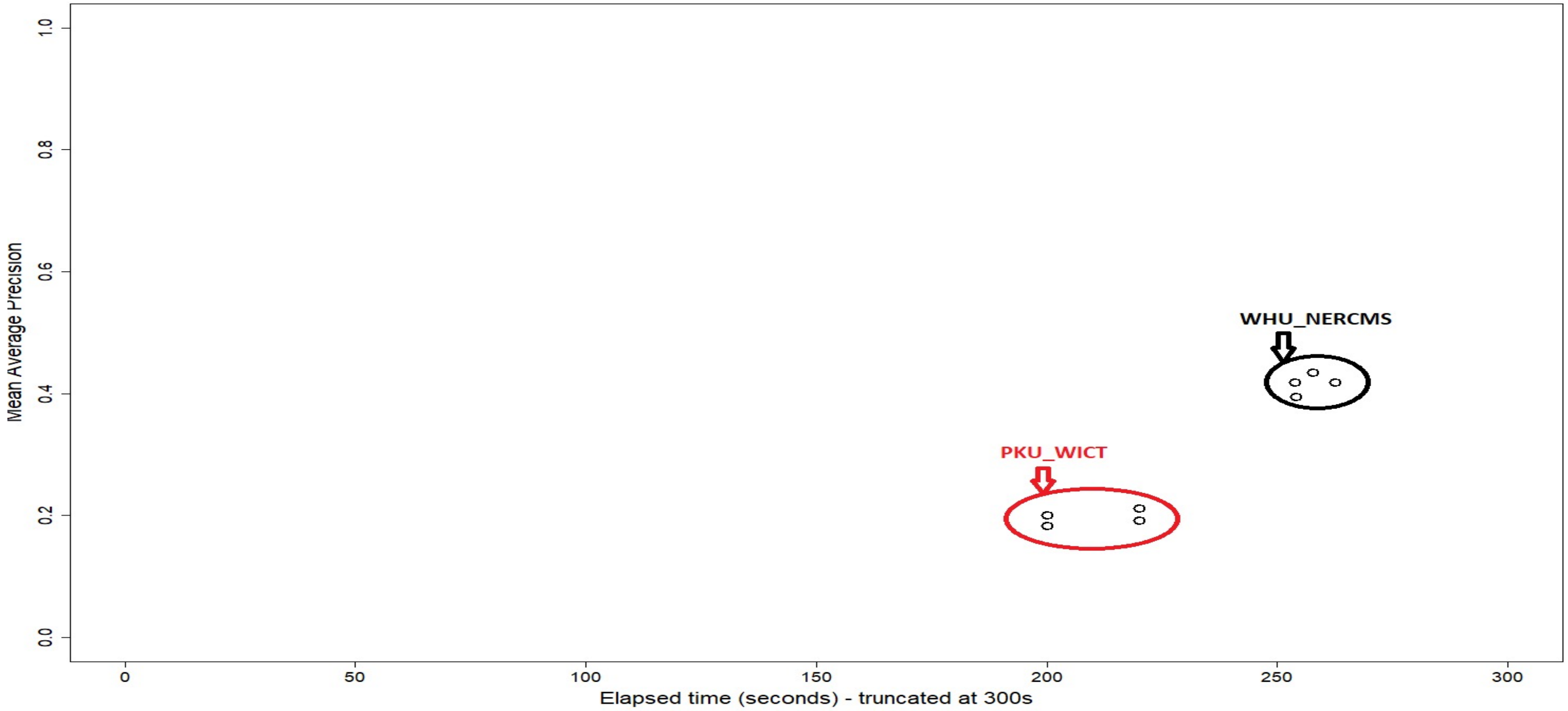
MAP	Top 10 runs across all teams (automatic)										
0.435	F_M_A_B_WHU_NERCMS.21_2*	=			>	>	>	>	>	>	>
0.418	F_M_A_B_WHU_NERCMS.21_6*		=		>	>	>	>	>	>	>
0.418	F_M_A_B_WHU_NERCMS.21_4*			=	>	>	>	>	>	>	>
0.395	F_M_A_B_WHU_NERCMS.21_8				=	>	>	>	>	>	>
0.328	F_M_A_B_BUPT_MCPRL.21_3⤒					=		>	>	>	>
0.326	F_M_A_B_BUPT_MCPRL.21_1⤒						=	>	>	>	>
0.211	F_M_E_E_PKU_WICT.21_1							=	>	>	>
0.200	F_M_A_E_PKU_WICT.21_3^								=		>
0.192	F_M_E_E_PKU_WICT.21_4^									=	>
0.183	F_M_A_E_PKU_WICT.21_5										=
		1	2	3	4	5	6	7	8	9	10

***^⤒ = difference not statistically significant**

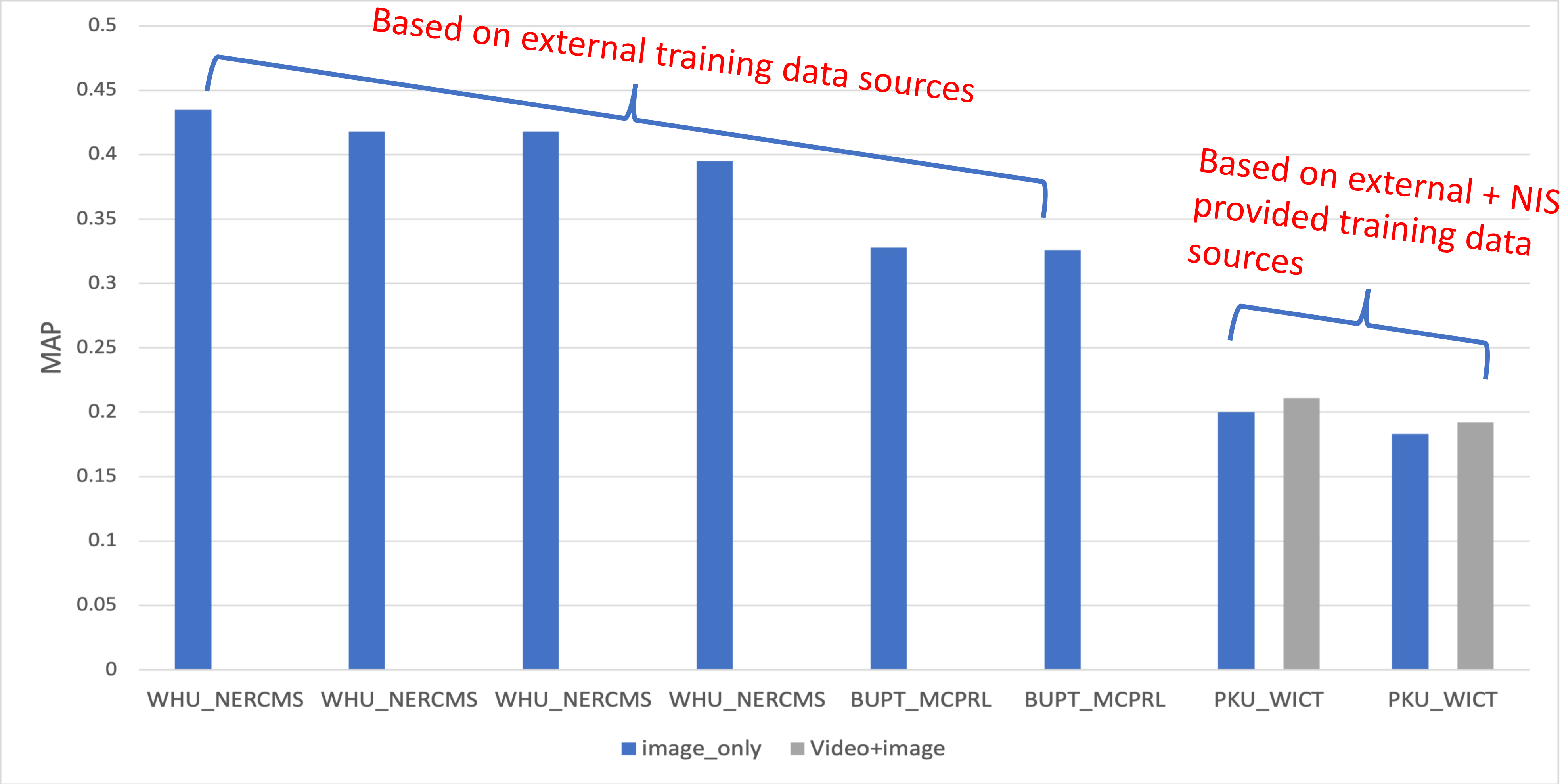
p = probability the row run scored better than the column run due to chance

> p < 0.05

Mean Average Precision vs. Per Run Clock Processing Time



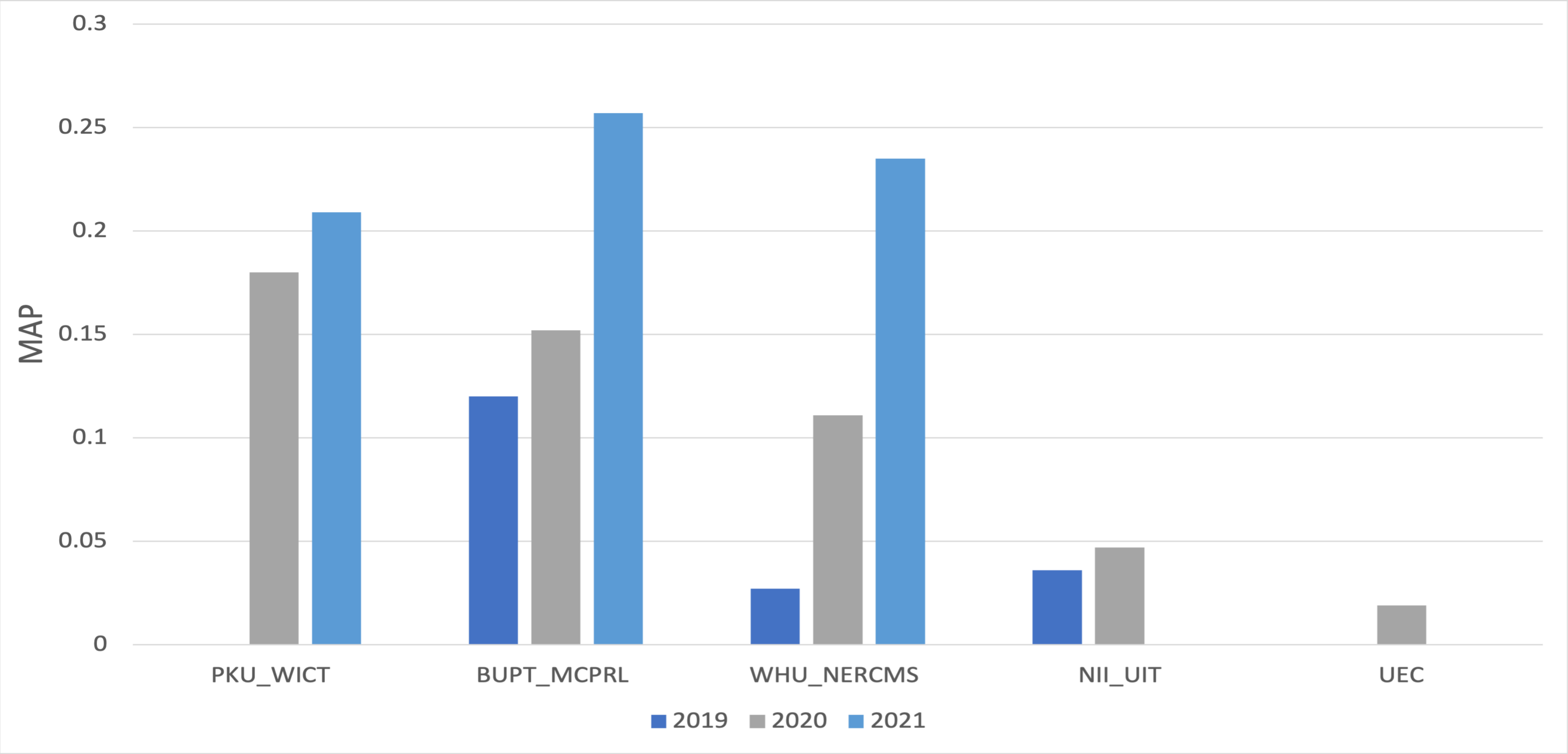
Results by Example Set (A/E) - Automatic



Based on external training data sources

Based on external + NIST provided training data sources

Progress Task 2019-2021: Max Performance



Progress Topics 2019-2021 - Observations

- Only two teams submitted progress runs for each year of the task.
- Those who did each saw encouraging improvement in results year-on-year.
- All teams who submitted progress runs for at least two years of the task also saw an improvement in results.

Some General Observations About the Task

- Decrease in number of participants and finishers, from 5 finishers out of 13 participants last year, to 3 finishers out of 11 participants this year.
- Most teams this year used A condition - training with image only, no video. But the only team which used both condition A + E achieved their best results using E condition – training with image and video.

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 and hard feature of INS task.
- Fine grained condition (action) detection is hard and needs more work (e.g. holding cloth vs phone vs paper, Or opening door and entering or exiting).
- The use of video data for training can be helpful in detecting actions that need temporal information (e.g. open door and enter)
- General improvements in performance have been seen year-on-year.

Further Conclusions

- Systems approaches tend to build two main modules for person retrieval and action retrieval, followed by fusion and ranking of results

- Action recognition approaches tends to follow:
  - Frame-level (actions that can be detected using single images)
  - Shot-level (actions that need more frames)
  - Human-Object interaction