University of BRISTOL · UNIVERSITY OF TORONTO · UNIVERSITÀ degli STUDI di CATANIA

Scaling Egocentric Vision: The EPIC-KITCHENS Dataset

University of BRISTOL

Dima Damen

EPIC KITCHENS

Scaling…

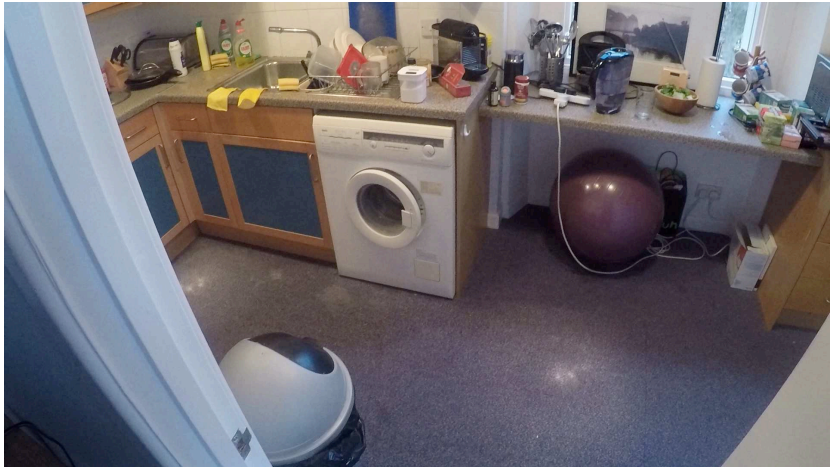
PASCAL2
Pattern Analysis, Statistical Modelling and Computational Learning

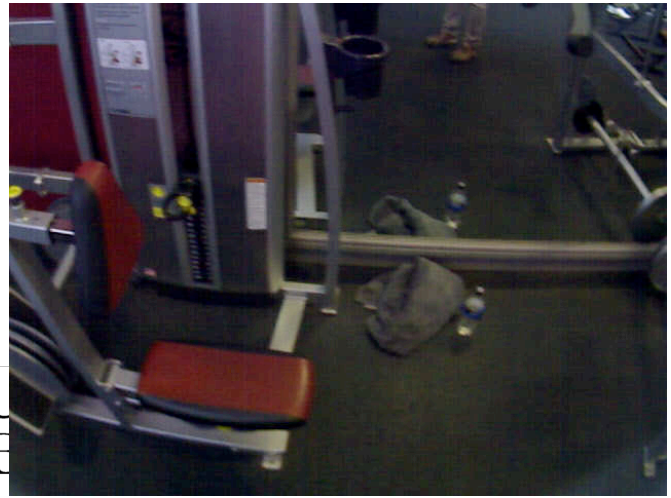
COCO
Common Objects in Context

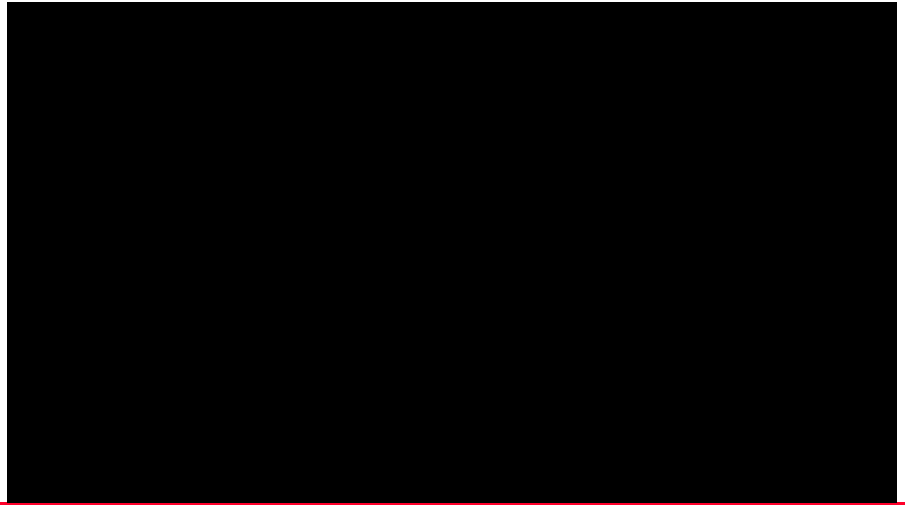
University of BRISTOL

EPIC KITCHENS

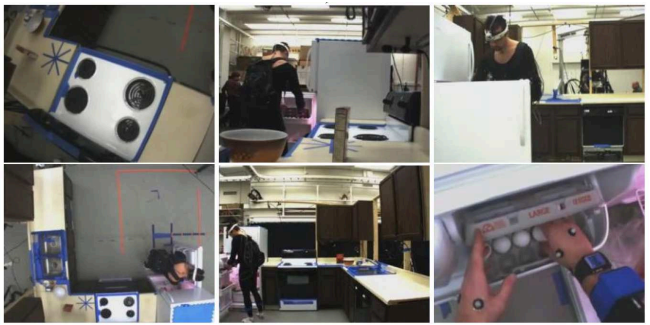
… Egocentric Vision



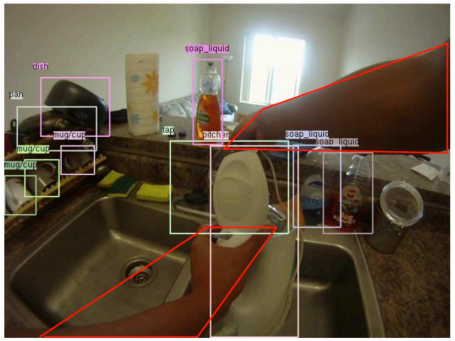




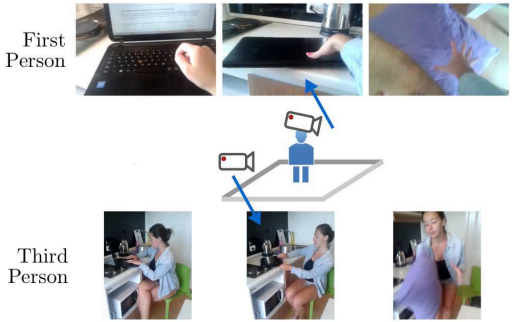
Scaling Egocentric Vision



CMU (2009)



ADL (2012)



Charades-Ego (2018)

BEOID (2014)

GTEA … (2011-)

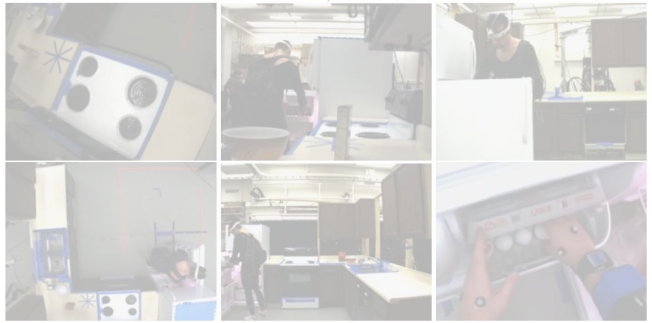
EGTEA+ (2018)

UNIVERSITY OF TORONTO

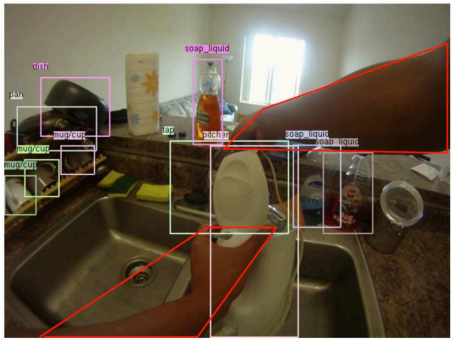
University of BRISTOL

UNIVERSITÀ degli STUDI di CATANIA

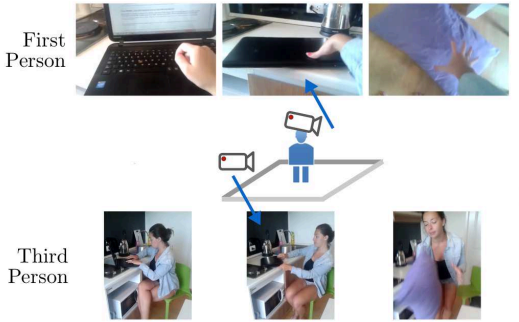
Native & Multiple Environments?



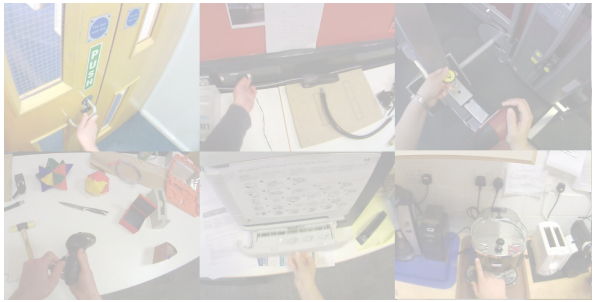
CMU (2009)



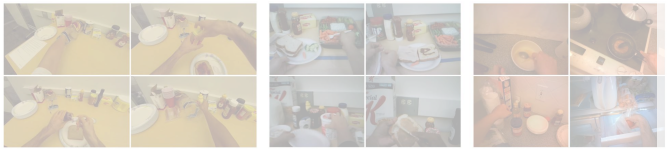
ADL (2012)



Charades-Ego (2018)



BEOID (2014)



GTEA … (2011-)

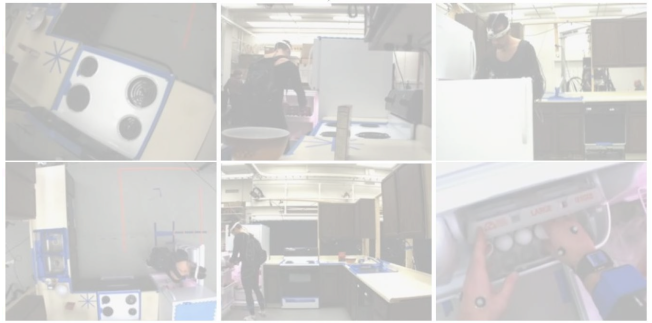
EGTEA+ (2018)

UNIVERSITY OF TORONTO

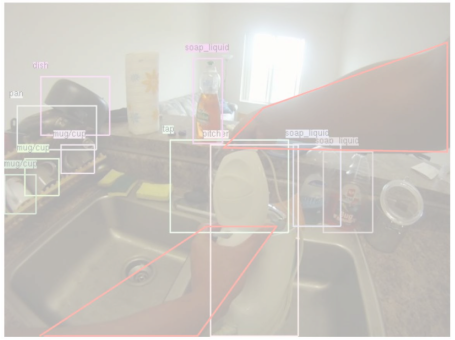
University of BRISTOL

UNIVERSITÀ degli STUDI di CATANIA

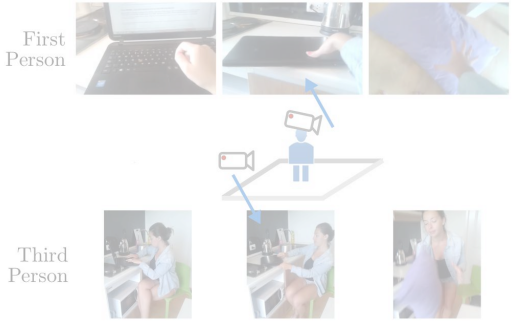
Non-scripted?



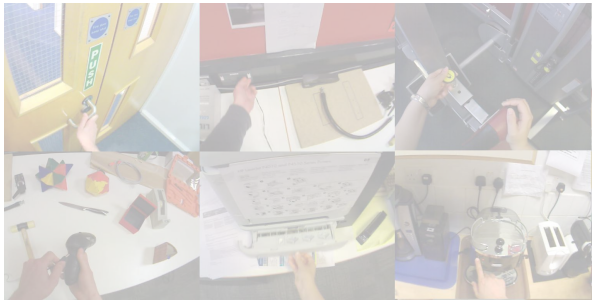
CMU (2009)



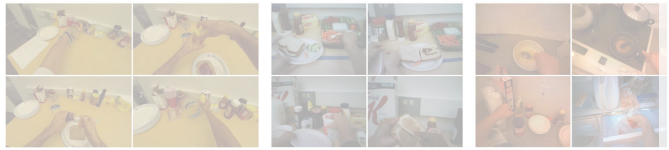
ADL (2012)



Charades-Ego (2018)



BEOID (2014)



GTEA … (2011-)

EGTEA+ (2018)

UNIVERSITY OF TORONTO

University of BRISTOL

UNIVERSITÀ degli STUDI di CATANIA

Scaling Egocentric Vision

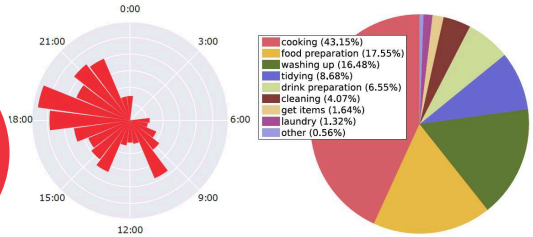
Dataset	Ego?
EPIC-KITCHENS	✓
EGTEA Gaze+ [16]	✓
Charades-ego [41]	70% ✓
BEOID [6]	✓
GTEA Gaze+ [13]	✓
ADL [36]	✓
CMU [8]	✓
YouCook2 [56]	×
VLOG [14]	×
Charades [42]	×
Breakfast [28]	×
50 Salads [44]	×
MPII Cooking 2 [39]	×

Scaling Egocentric Vision

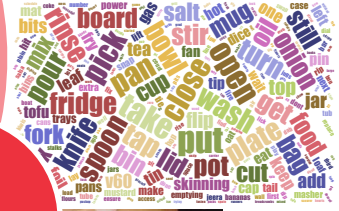
EPIC KITCHENS

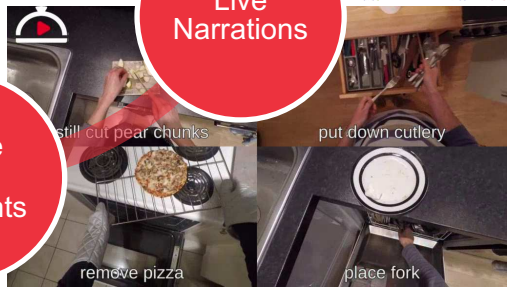
Data Collection

Native Environment, Natural Interactions



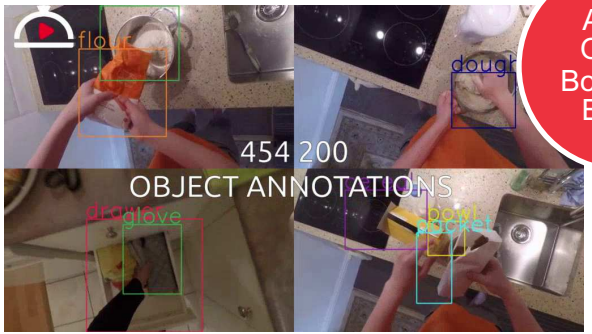
Live Narrations





Dense Action Segments

Active Object Bounding Boxes



Benchmark and Challenges

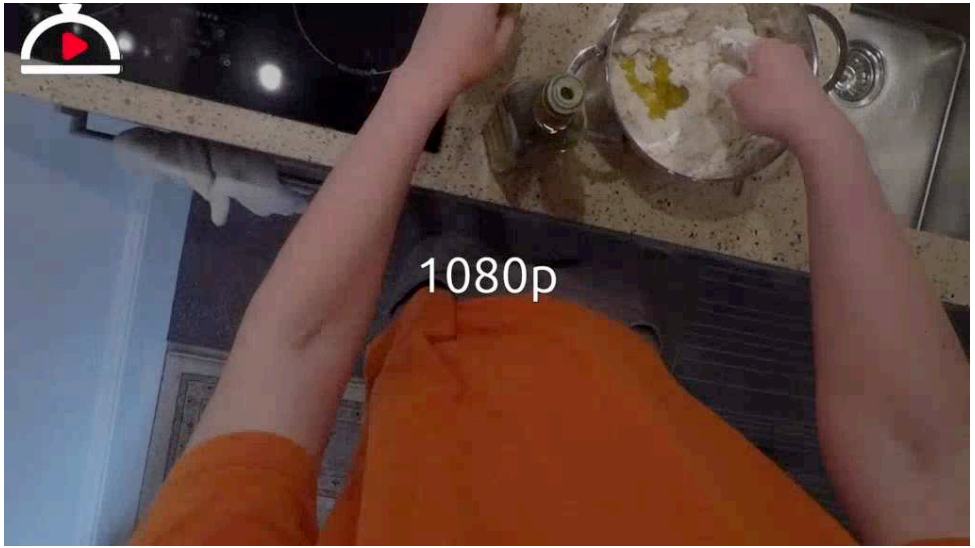
UNIVERSITY OF TORONTO

University of BRISTOL

UNIVERSITÀ degli STUDI di CATANIA

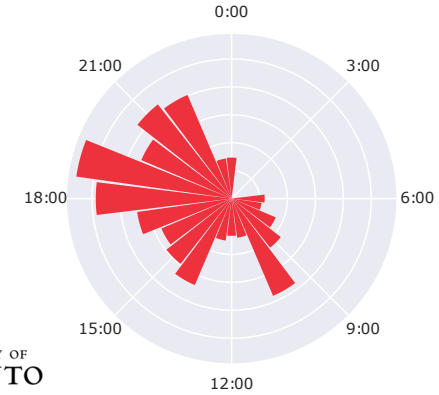
Data Collection

- Head-Mounted Go-Pro, adjustable mounting
- Recording starts immediately before entering the kitchen
- Only stopped before leaving the kitchen



Data Collection

- 32 kitchens
- Single-person environments
- 4 cities
- May – Nov 2017 – 55 hours
- 10 nationalities
- 3 days - all kitchen activities



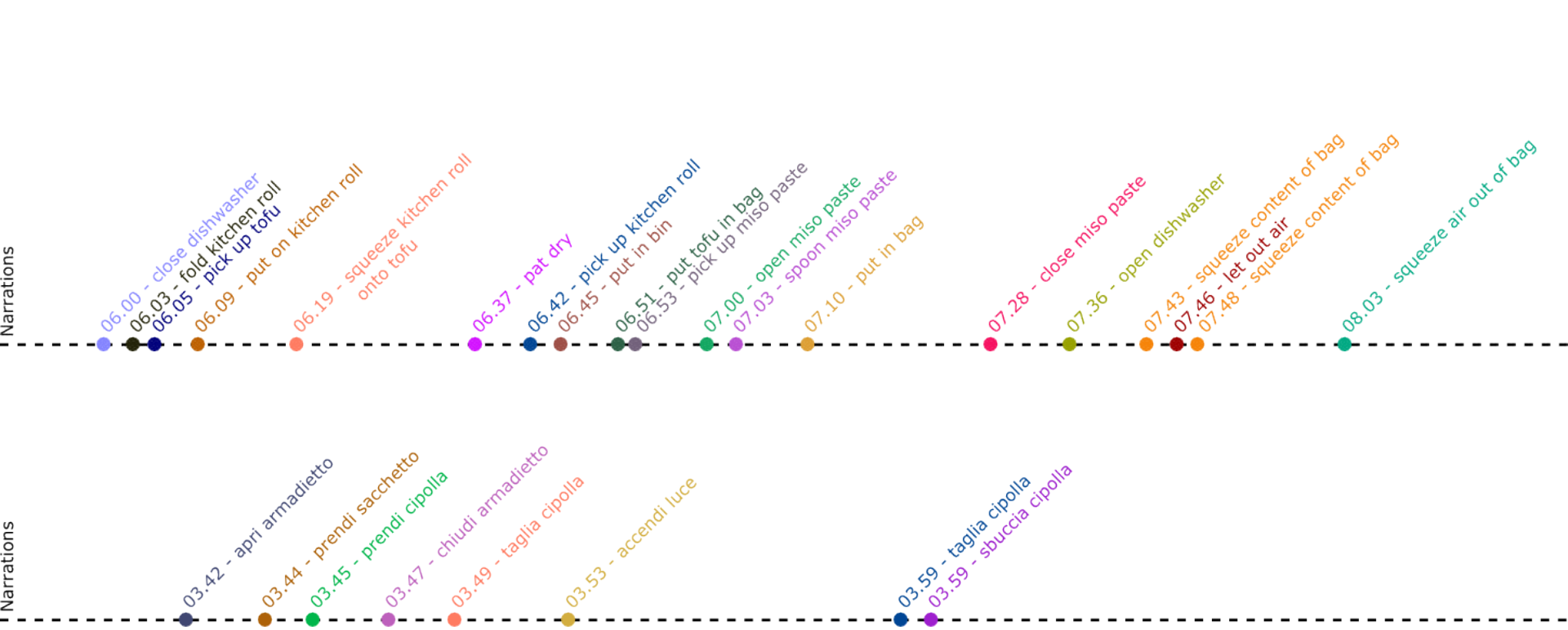
UNIVERSITY OF TORONTO

University of BRISTOL

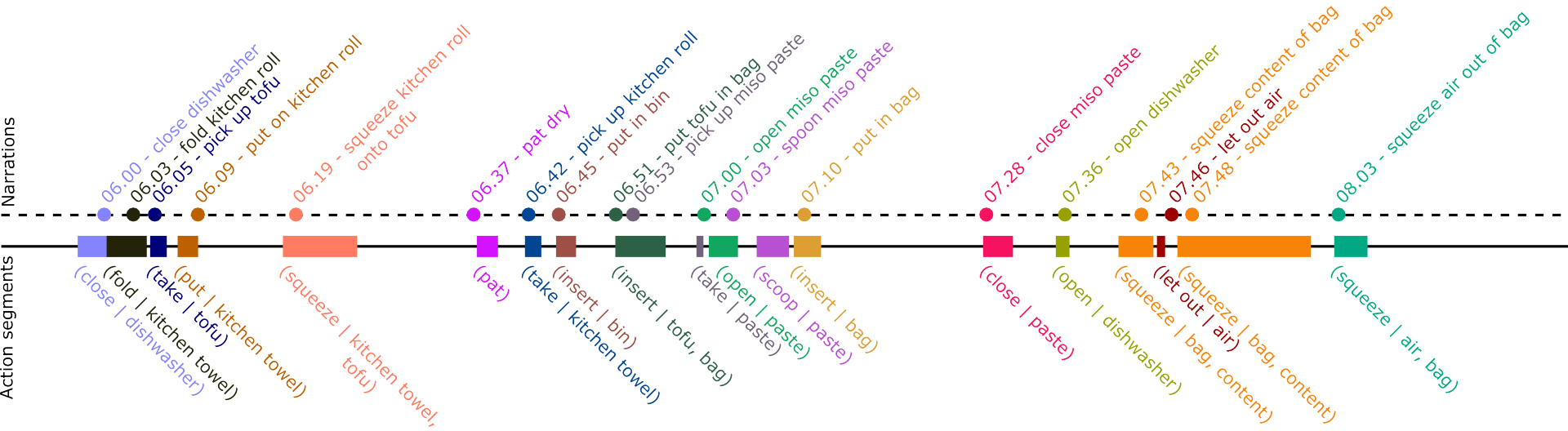
UNIVERSITÀ degli STUDI di CATANIA

Annotations (1) - Narrations

Annotations (1) - Narrations



Annotations (2) – Action Segments



UNIVERSITY OF TORONTO

University of BRISTOL

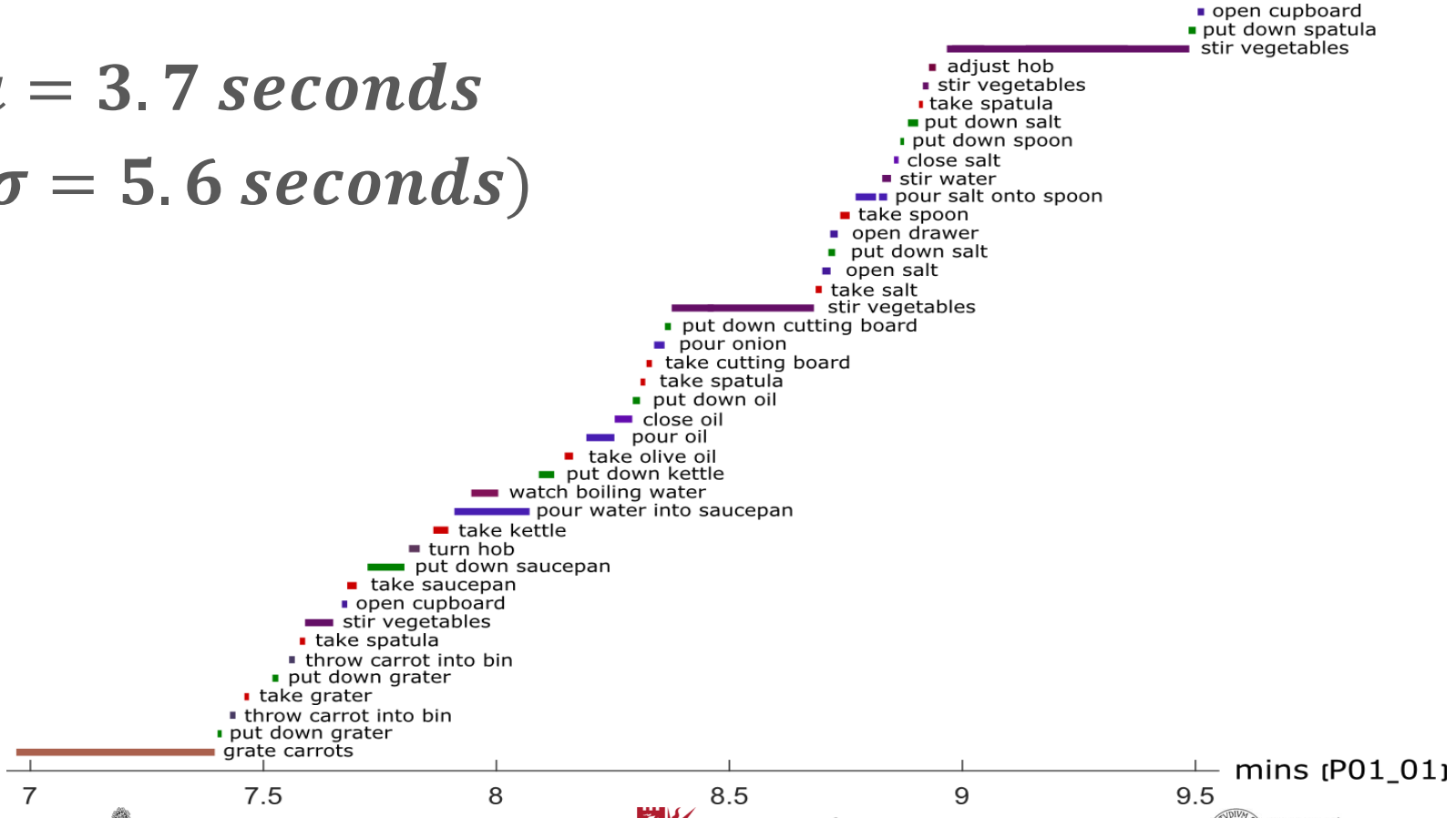
UNIVERSITÀ degli STUDI di CATANIA

39 000
ACTION SEGMENTS

Annotations (2) – Action Segments

$\mu = 3.7\ seconds$

$(\sigma = 5.6\ seconds)$

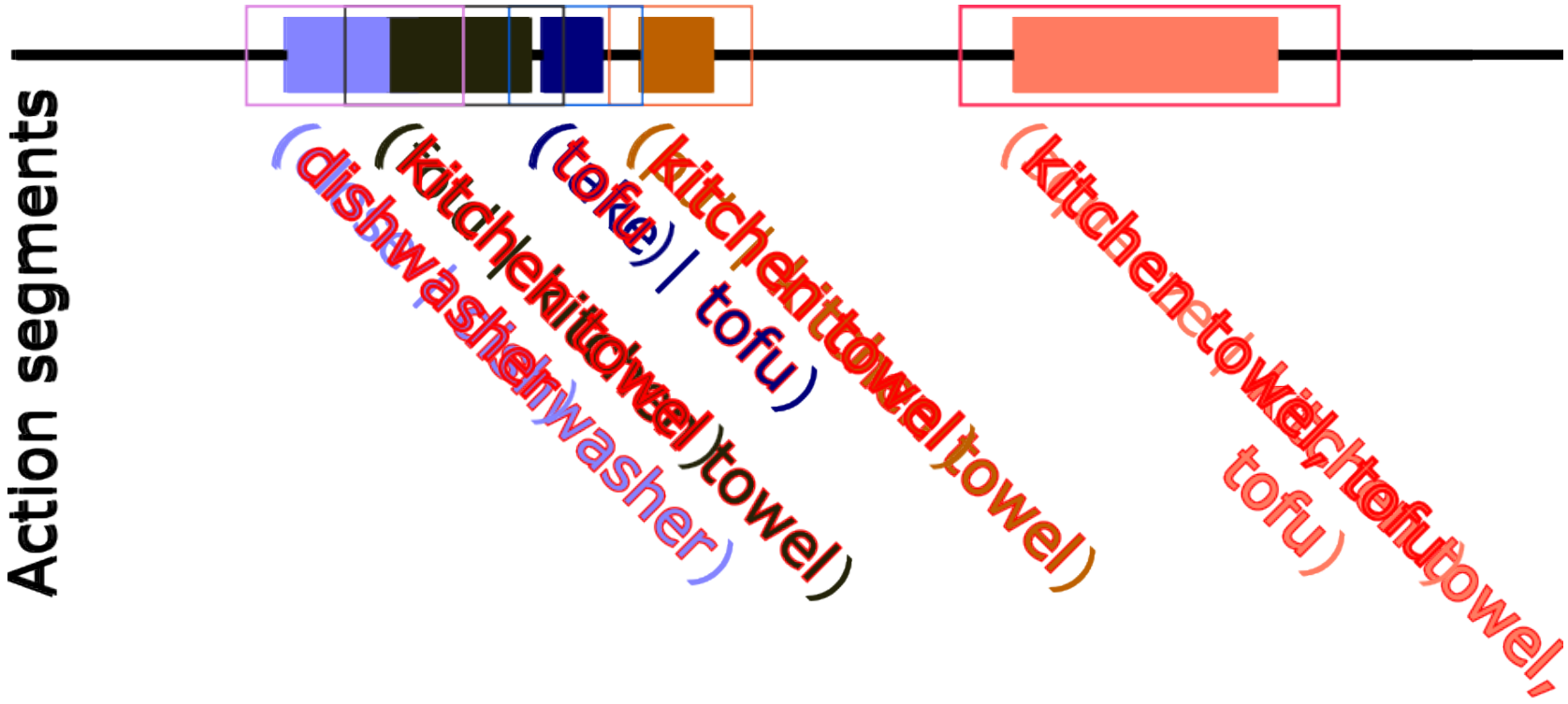


UNIVERSITY OF TORONTO

University of BRISTOL

UNIVERSITÀ degli STUDI di CATANIA

Annotations (3) – Object Bounding Boxes



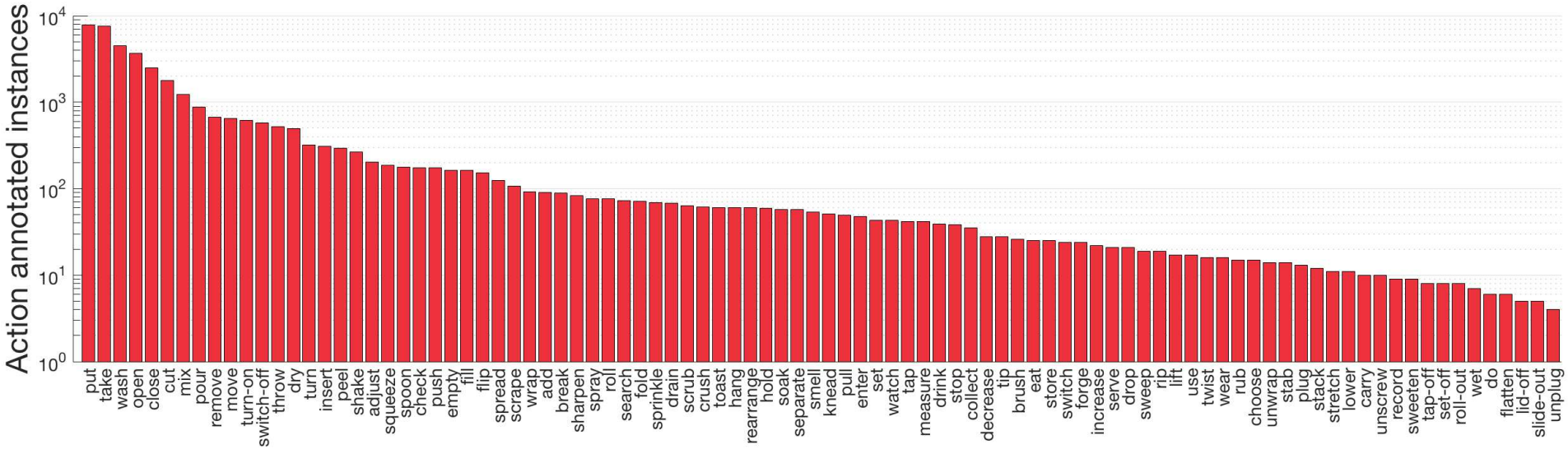
454 200
OBJECT ANNOTATIONS

Annotations (4) – Verb and Noun Classes

take, grab, pick, get, fetch, pick-up, ...

- 125 verb classes
- 331 noun classes

Annotations Statistics



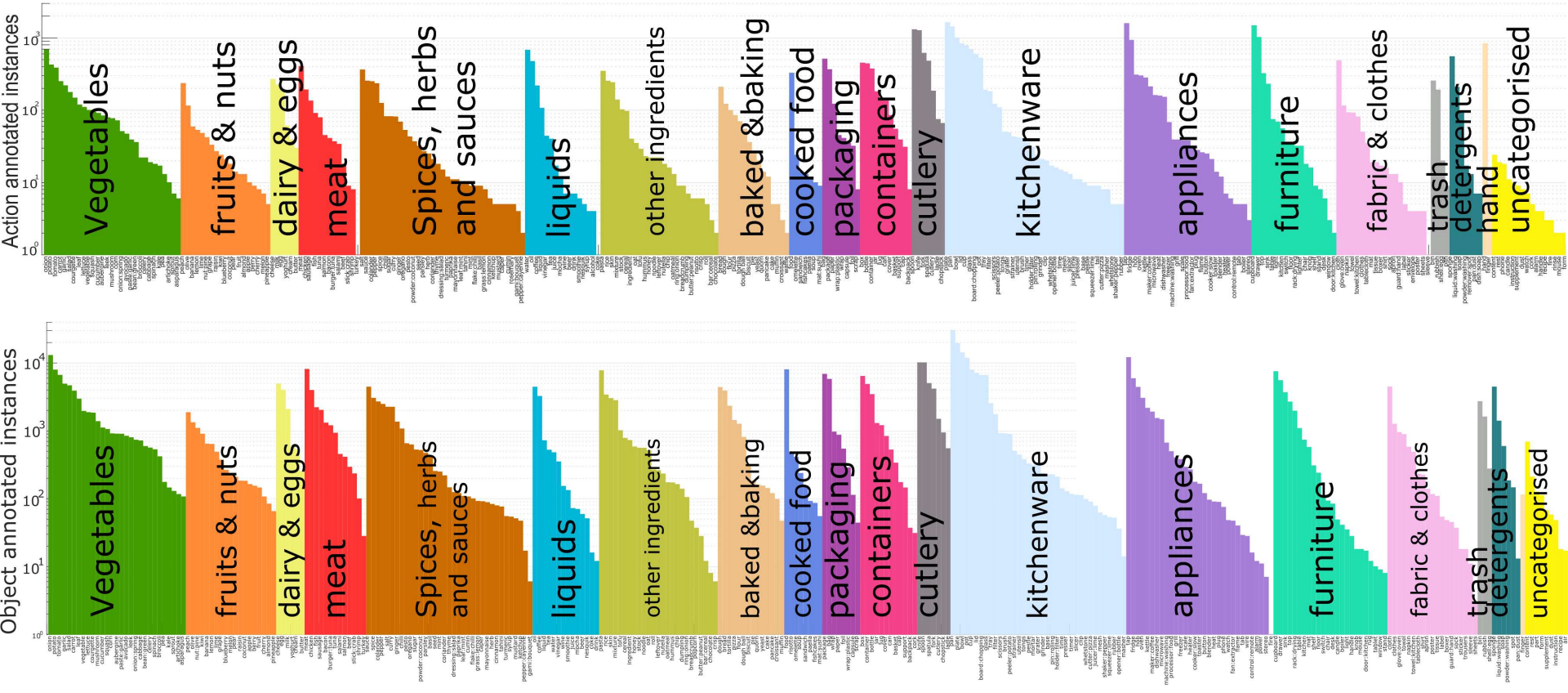
UNIVERSITY OF TORONTO

University of BRISTOL

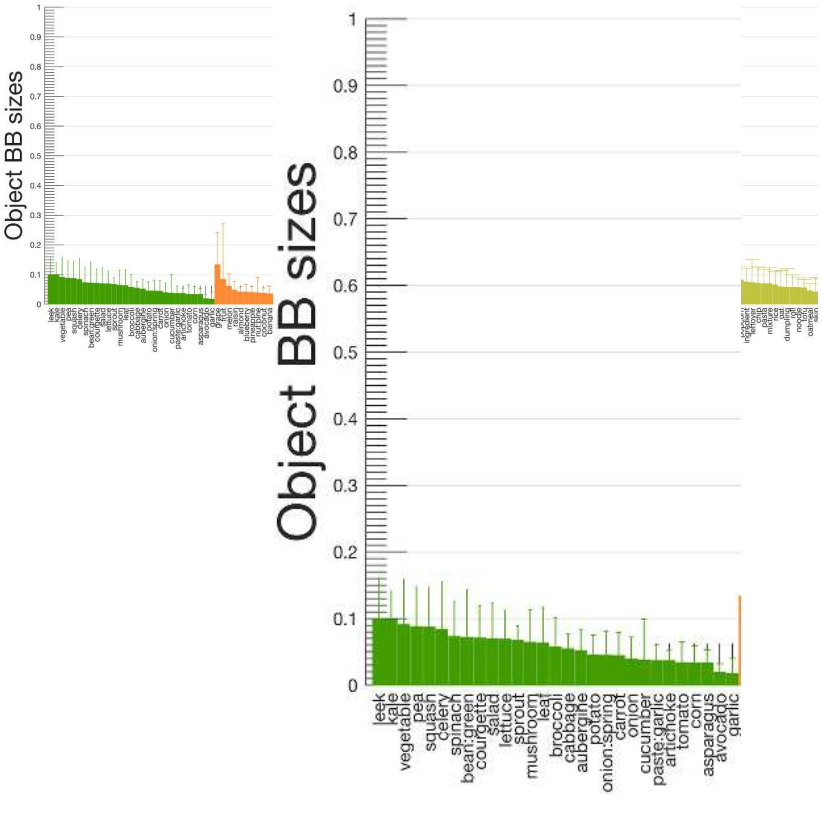
UNIVERSITÀ degli STUDI di CATANIA

EPIC KITCHENS

Annotations Statistics



Annotations (3) – Object Bounding Boxes



UNIVERSITY OF TORONTO

University of BRISTOL

UNIVERSITÀ degli STUDI di CATANIA

Train/Test Splits

- 20% - Seen Test Set
  - 28 Kitchens
- 7% - Unseen Test Set
  - 4 Kitchens

Table 4: Statistics of test splits: seen (S1) and unseen (S2) kitchens

	#Subjects	#Sequences	Duration (s)	%	Narrated Segments	Action Segments	Bounding Boxes
Train/Val	28	272	141731		28,587	28,561	326,388
S1 Test	28	106	39084	20%	8,069	8,064	97,872
S2 Test	4	54	13231	7%	2,939	2,939	29,995

Dataset Release

FHD video:

- 1920x1080 px
- 60FPS

RGB frames:

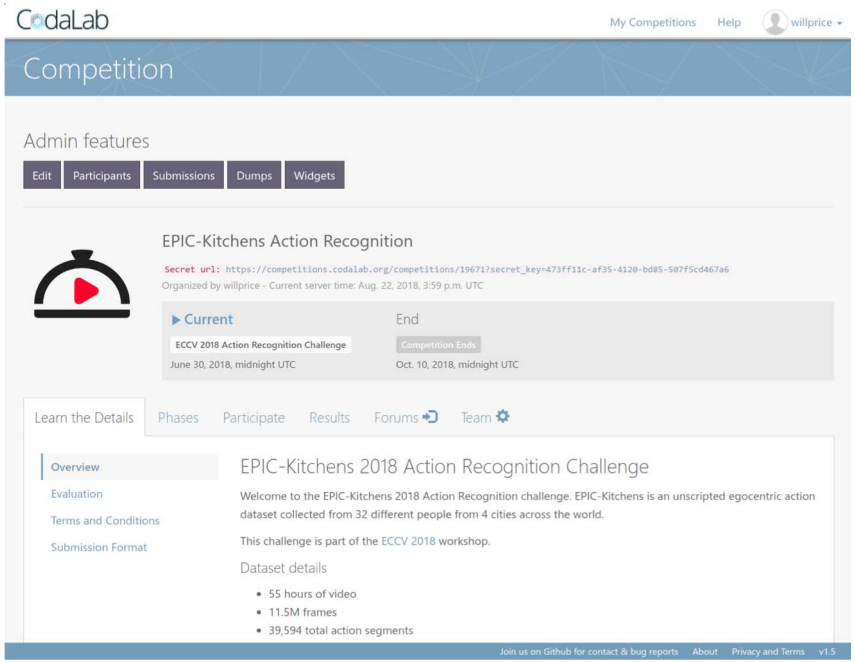
- 456x256 px
- 60FPS

TVL$_1$ optical flow (u, v) frames:

- 456x256 px
- 30FPS

Open Challenges

1. Object Detection Challenge
2. Action Recognition Challenge
3. Action Anticipation Challenge



Object Detection Challenge

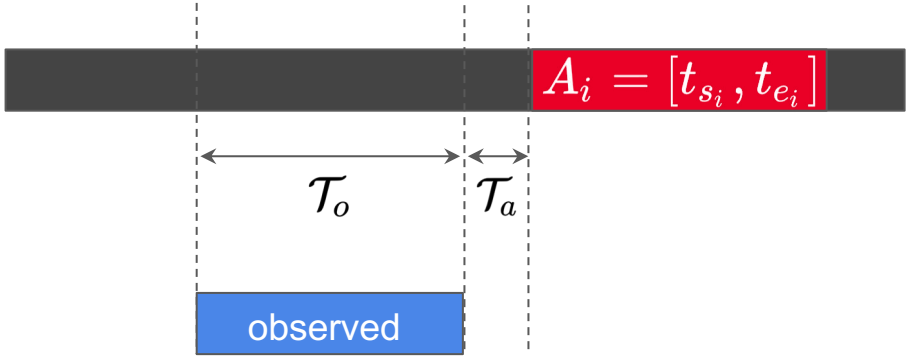
Action Recognition Challenge

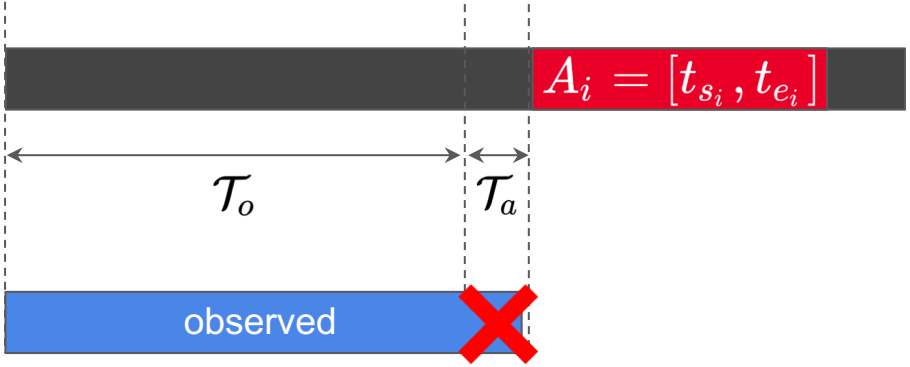
Given a trimmed action segment:
$(t_{\text{start}}, t_{\text{stop}})$
classify the action within.

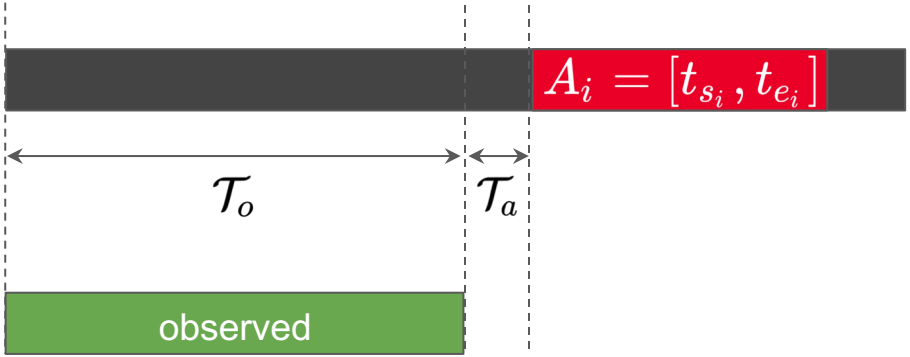
$$\hat{y}_{\text{verb}} = \text{open}$$
$$\hat{y}_{\text{noun}} = \text{oven}$$

$$\hat{y}_{\text{action}} = (\text{open}, \text{oven})$$

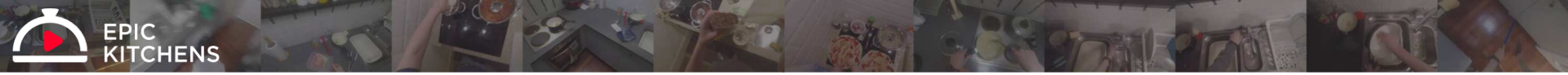
Action Anticipation Challenge







Anticipation time fixed to **1 second**



Our work using EPIC-Kitchens

University of BRISTOL

D Moltisanti et al (2019). Action Recognition from Single Timestamp Supervision in Untrimmed Videos. *Computer Vision and Pattern Recognition (CVPR)*

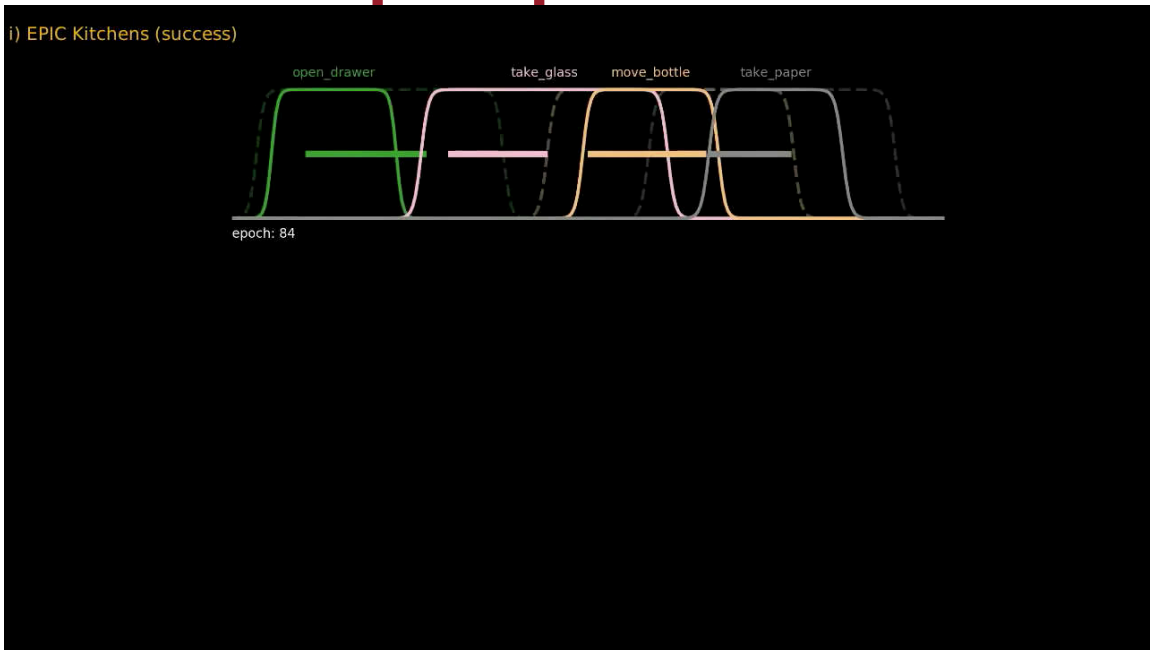
Action Recognition from Single Timestamp Supervision

with: Davide Moltisanti
Sanja Fidler

D Moltisanti et al (2019). Action Recognition from Single Timestamp Supervision in Untrimmed Videos. *Computer Vision and Pattern Recognition (CVPR)*

Action Recognition from Single Timestamp Supervision

with: Davide Moltisanti
Sanja Fidler



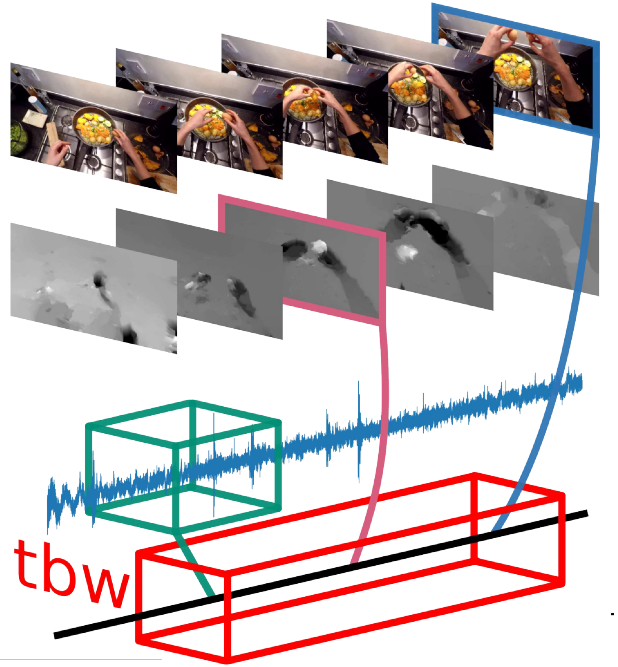
University of BRISTOL

D Moltisanti et al (2019). Action Recognition from Single Timestamp Supervision in Untrimmed Videos. *Computer Vision and Pattern Recognition (CVPR)*

EPIC KITCHENS

Audio-Visual Temporal Binding for Egocentric Action Recognition

with: Vangelis Kazakos
Arsha Nagrani
Andrew Zisserman



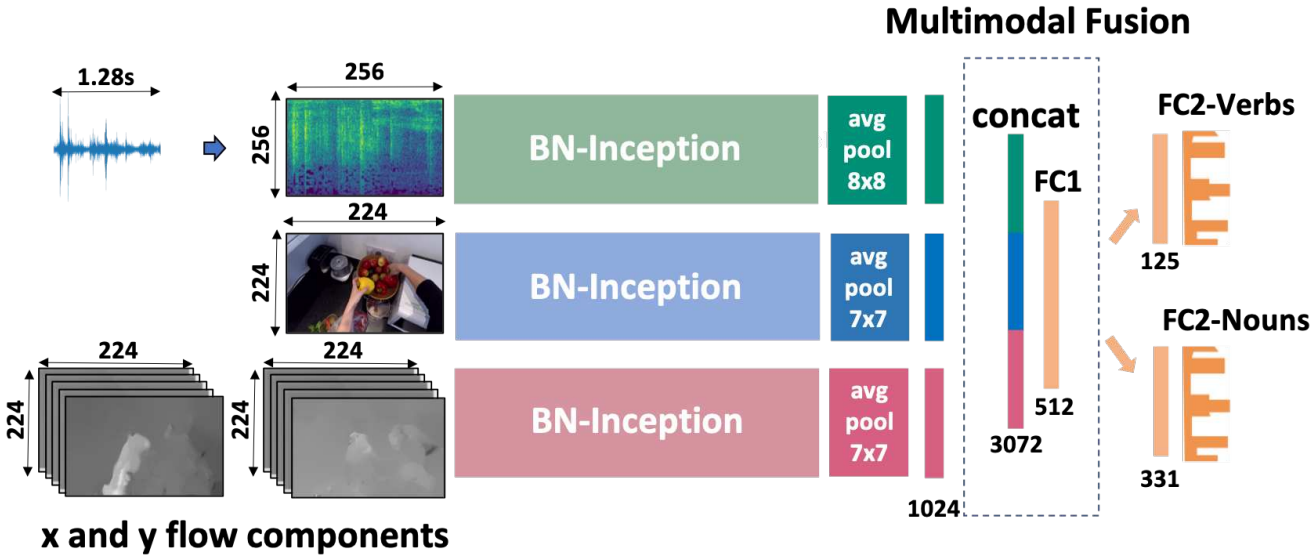
University of BRISTOL

E Kazakos, A Nagrani, A Zisserman, D Damen (2019). EPIC-Fusion: Audio-Visual Temporal Binding for Egocentric Action Recognition. ICCV

EPIC KITCHENS

Audio-Visual Temporal Binding for Egocentric Action Recognition

with: Vangelis Kazakos
Arsha Nagrani
Andrew Zisserman

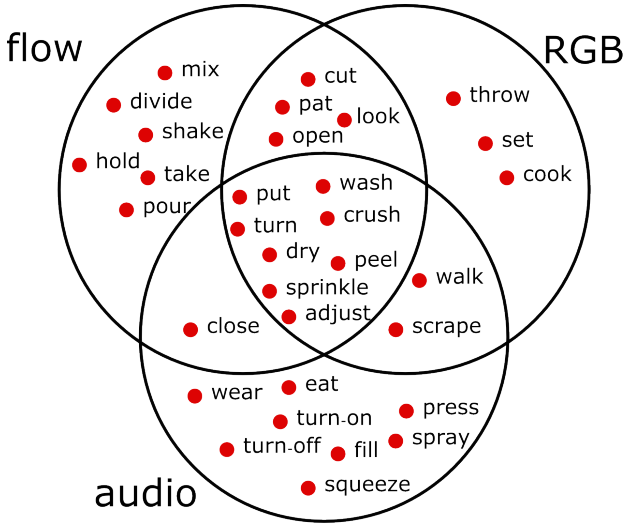


University of BRISTOL

E Kazakos, A Nagrani, A Zisserman, D Damen (2019). EPIC-Fusion: Audio-Visual Temporal Binding for Egocentric Action Recognition. ICCV

Audio-Visual Temporal Binding for Egocentric Action Recognition

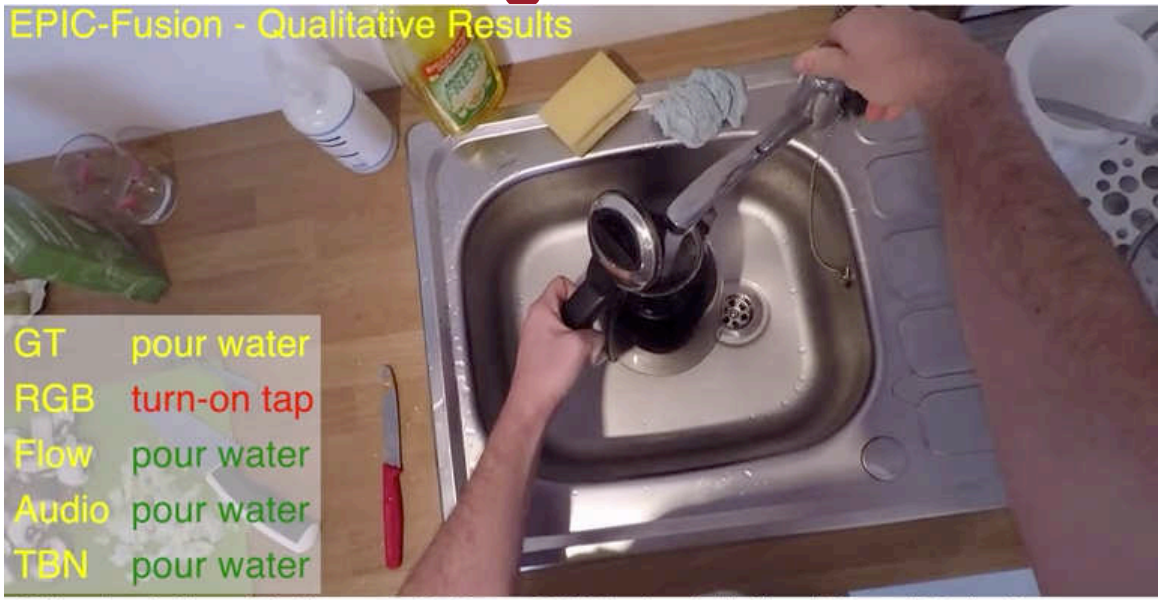
with: Vangelis Kazakos
Arsha Nagrani
Andrew Zisserman



E Kazakos, A Nagrani, A Zisserman, D Damen (2019). EPIC-Fusion: Audio-Visual Temporal Binding for Egocentric Action Recognition. ICCV

Audio-Visual Temporal Binding for Egocentric Action Recognition

with: Vangelis Kazakos
Arsha Nagrani
Andrew Zisserman



E. Kazakos, A. Nagrani, A. Zisserman, D. Damen, EPIC-Fusion: Audio-Visual Temporal Binding for Egocentric Action Recognition, ICCV 2019

E Kazakos, A Nagrani, A Zisserman, D Damen (2019). EPIC-Fusion: Audio-Visual Temporal Binding for Egocentric Action Recognition. ICCV

EPIC KITCHENS

Fine-Grained Action Retrieval through Multiple Part-of-Speech Embeddings

with: Michael Wray
Gabriela Csurka
Diane Larlus

In this work we focus on
Fine-Grained Action Retrieval

I put meat on a ball of dough ⟷

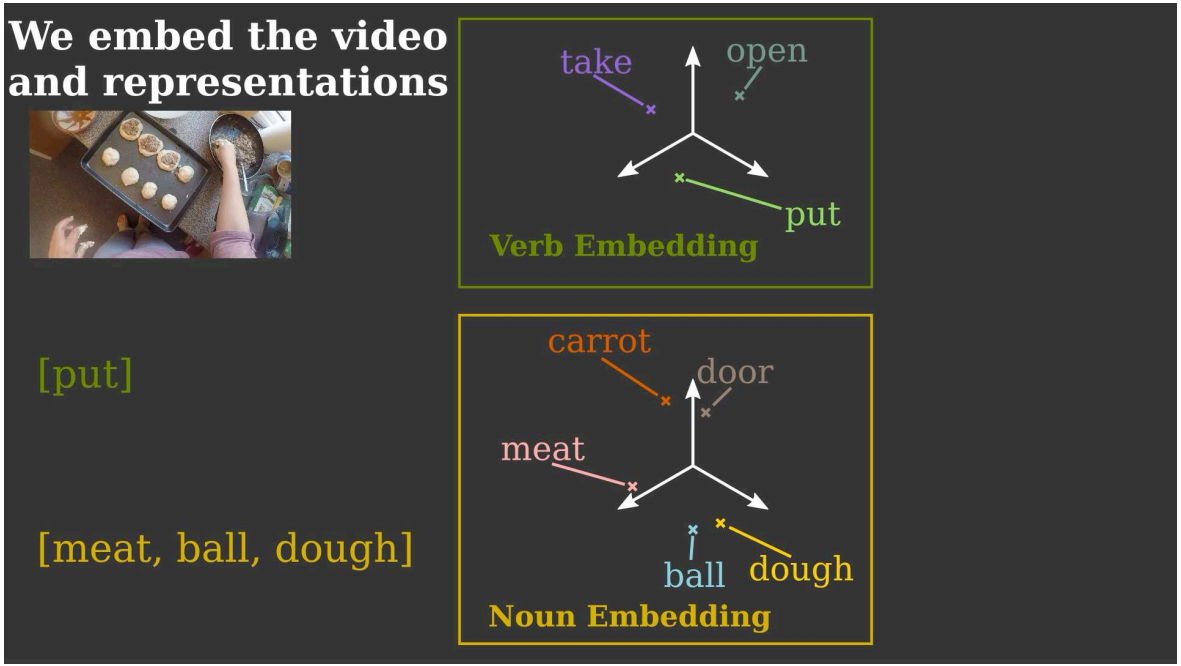
University of BRISTOL

M Wray, D Larlus, G Csurka, D Damen (2019). Fine-Grained Action Retrieval through Multiple Parts-of-Speech Embeddings. ICCV

EPIC KITCHENS

Fine-Grained Action Retrieval through Multiple Part-of-Speech Embeddings

with: Michael Wray
Gabriela Csurka
Diane Larlus

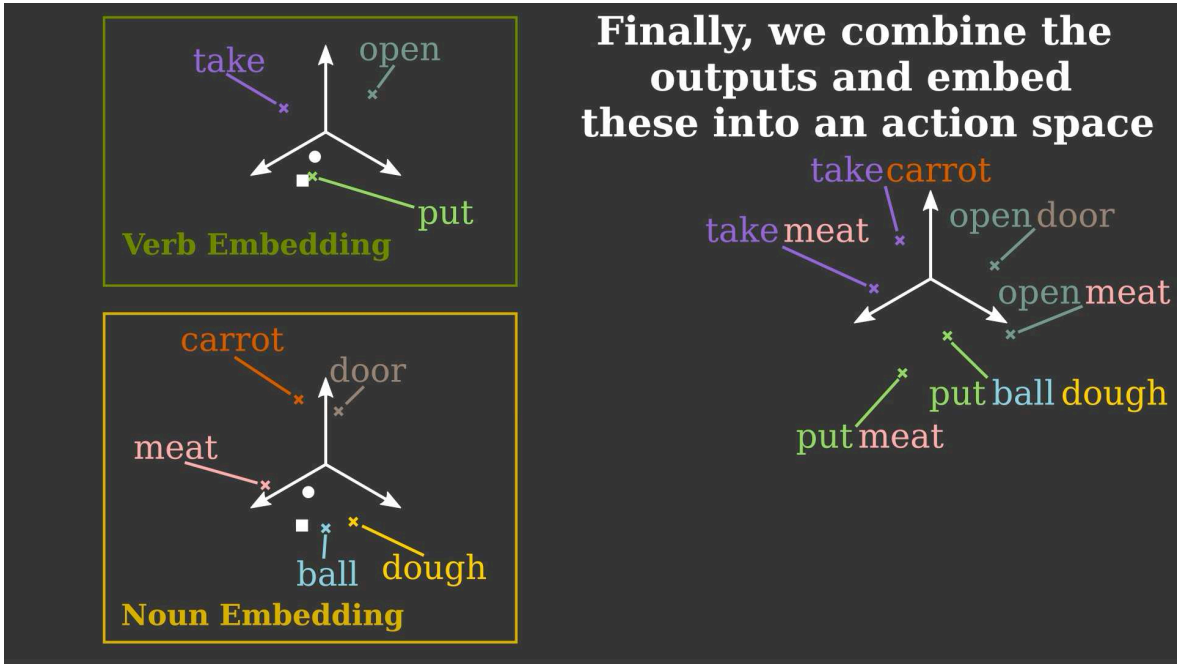


University of BRISTOL

M Wray, D Larlus, G Csurka, D Damen (2019). Fine-Grained Action Retrieval through Multiple Parts-of-Speech Embeddings. ICCV

Fine-Grained Action Retrieval through Multiple Part-of-Speech Embeddings

with: Michael Wray
Gabriela Csurka
Diane Larlus



University of BRISTOL

M Wray, D Larlus, G Csurka, D Damen (2019). Fine-Grained Action Retrieval through Multiple Parts-of-Speech Embeddings. ICCV

with: Will Price

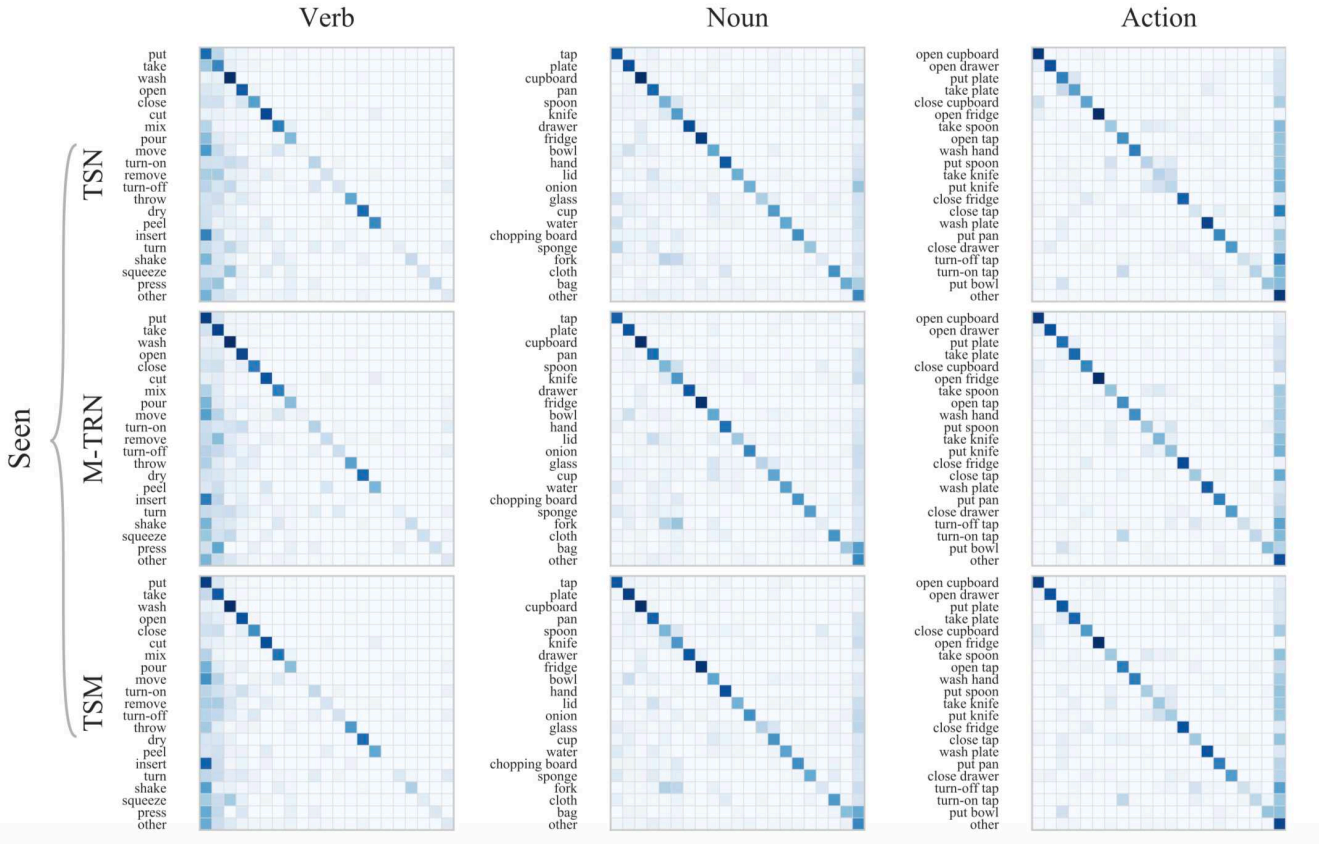
Evaluating Action Recognition Models

			Verb				Noun				Action			
			Top-1		Top-5		Top-1		Top-5		Top-1		Top-5	
BB	Model	Modality	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2
BN-Inception	TSN	RGB	47.97	36.46	87.03	74.36	38.85	22.64	65.54	46.94	22.39	11.30	44.75	26.32
		Flow	51.68	47.35	84.63	76.95	26.82	21.20	50.64	42.47	16.76	13.49	33.75	27.52
		Fusion	54.70	46.06	87.24	76.65	40.11	24.27	65.81	49.27	25.43	14.78	45.69	29.81
	TRN	RGB	58.26	47.29	87.14	76.54	36.32	22.91	63.30	44.73	25.46	15.06	45.66	28.99
		Flow	55.20	50.32	84.04	77.67	23.95	19.02	47.02	40.25	16.03	12.77	32.92	27.62
		Fusion	61.04	51.83	87.46	79.11	37.90	24.75	63.69	47.35	26.54	16.59	46.37	31.14
	M-TRN	RGB	57.66	45.41	86.91	76.34	37.94	23.90	63.78	46.33	26.62	15.57	46.39	29.57
		Flow	55.92	51.38	84.44	77.74	24.88	20.69	48.37	40.83	16.78	14.00	34.09	28.75
		Fusion	61.12	51.62	87.71	78.42	39.28	26.02	64.36	48.99	27.86	17.34	47.56	32.57
ResNet-50	TSN	RGB	49.71	36.70	87.19	73.64	39.85	23.11	65.93	44.73	23.97	12.77	46.14	26.08
		Flow	53.14	47.56	84.88	76.89	27.76	20.28	51.29	42.23	18.03	13.11	35.18	27.83
		Fusion	55.50	45.75	87.85	77.40	41.28	25.13	66.53	48.11	26.89	15.40	47.35	30.01
	TRN	RGB	58.82	47.32	86.60	76.92	37.27	23.69	62.96	46.02	26.62	15.71	46.09	30.01
		Flow	55.16	50.39	83.87	77.71	23.19	18.50	47.33	40.70	15.77	12.02	33.08	27.42
		Fusion	61.60	52.27	87.20	79.55	38.41	25.74	63.37	47.87	27.58	17.79	46.44	32.20
	M-TRN	RGB	60.16	46.94	87.18	75.21	38.36	24.41	64.67	46.71	28.23	16.32	47.89	29.74
		Flow	56.79	50.36	84.91	77.67	25.00	20.28	48.70	41.45	17.24	13.42	34.80	29.02
		Fusion	62.68	52.03	87.96	78.90	39.82	25.88	64.94	49.03	29.41	17.86	48.91	32.54
	TSM	RGB	57.88	43.50	87.14	73.85	40.84	23.32	66.10	46.02	28.22	14.99	49.12	28.06
		Flow	58.08	52.68	85.88	79.11	27.49	20.83	50.27	43.70	19.14	14.27	36.90	29.60
		Fusion	62.37	51.96	88.55	79.21	41.88	25.61	66.43	49.47	29.90	17.38	49.81	32.67

Table 1: Backbone (BB) comparison using 8 segments in both training and testing evaluating top-1/5 accuracy across tasks. S1 denotes the seen test set, and S2 the unseen test set. Cells are coloured on a per column basis: low high.

Eva

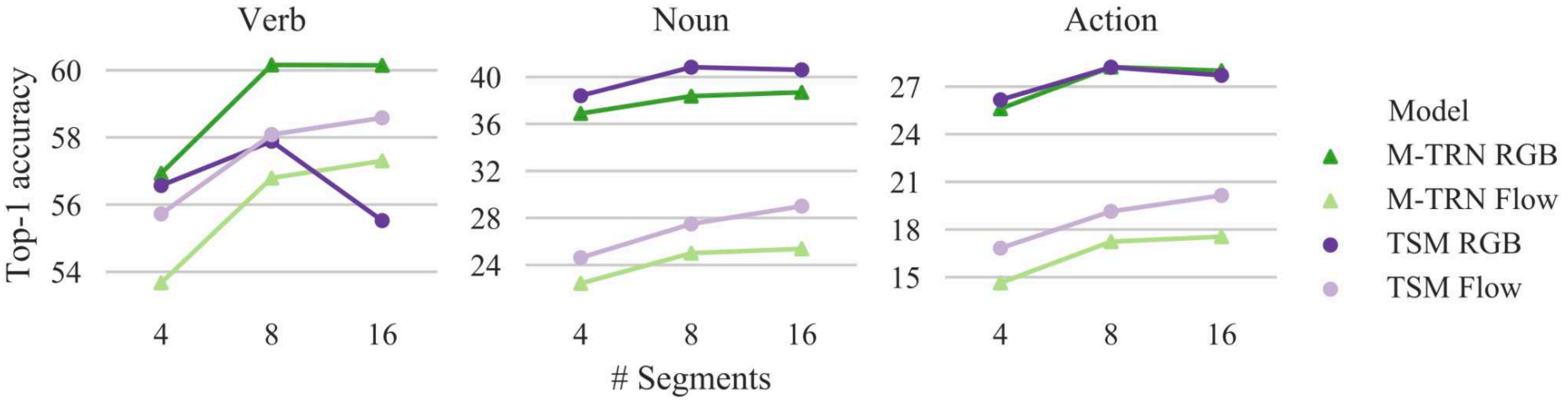
with: Will Price



W Price, D Damen (2019). An Evaluation of Action Recognition Models on EPIC-Kitchens. Arxiv

with: Will Price

Evaluating Action Recognition Models



University of BRISTOL

W Price, D Damen (2019). An Evaluation of Action Recognition Models on EPIC-Kitchens. Arxiv

with: Will Price

Evaluating Action Recognition Models

	GFLOP/s		Params (M)	
Model	RGB	Flow	RGB	Flow
TSN	33.12	35.33	24.48	24.51
TRN	33.12	35.32	25.33	25.35
M-TRN	33.12	35.33	27.18	27.21
TSM	33.12	35.33	24.48	24.51

Models Released

Table 3: Model parameter and FLOP/s count using a ResNet-50 backbone with 8 segments for a single video.

W Price, D Damen (2019). An Evaluation of Action Recognition Models on EPIC-Kitchens. Arxiv

For further info, datasets, code, publications…

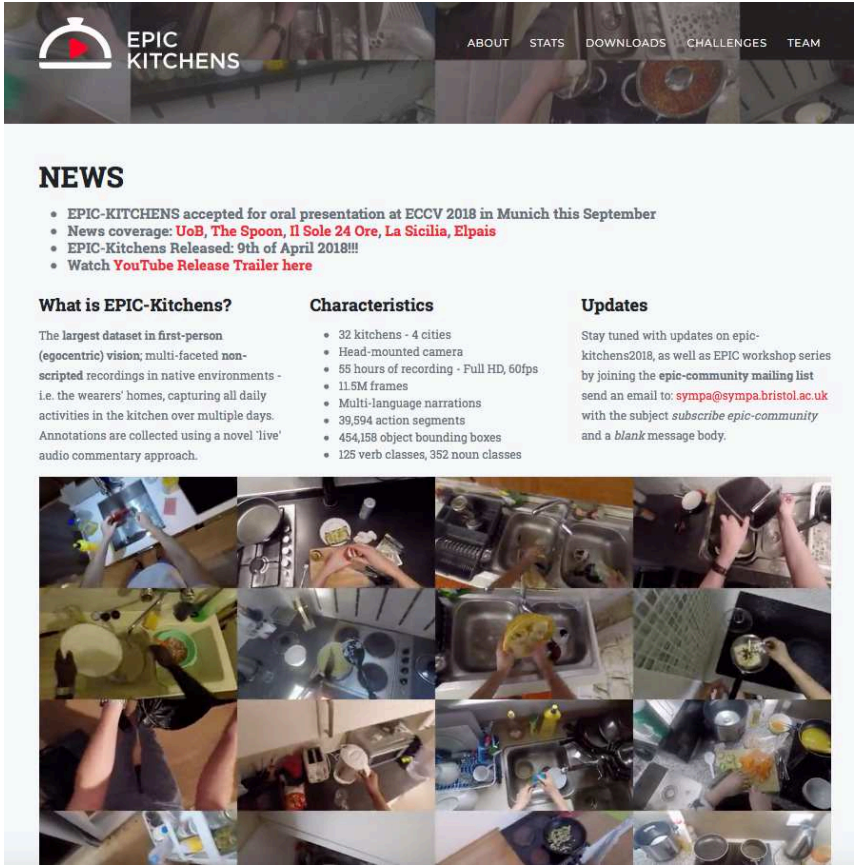
http://dimadamen.github.io

@dimadamen

http://www.linkedin.com/in/dimadamen

More?

http://epic-kitchens.github.io



UNIVERSITY OF TORONTO

University of BRISTOL

UNIVERSITÀ degli STUDI di CATANIA