

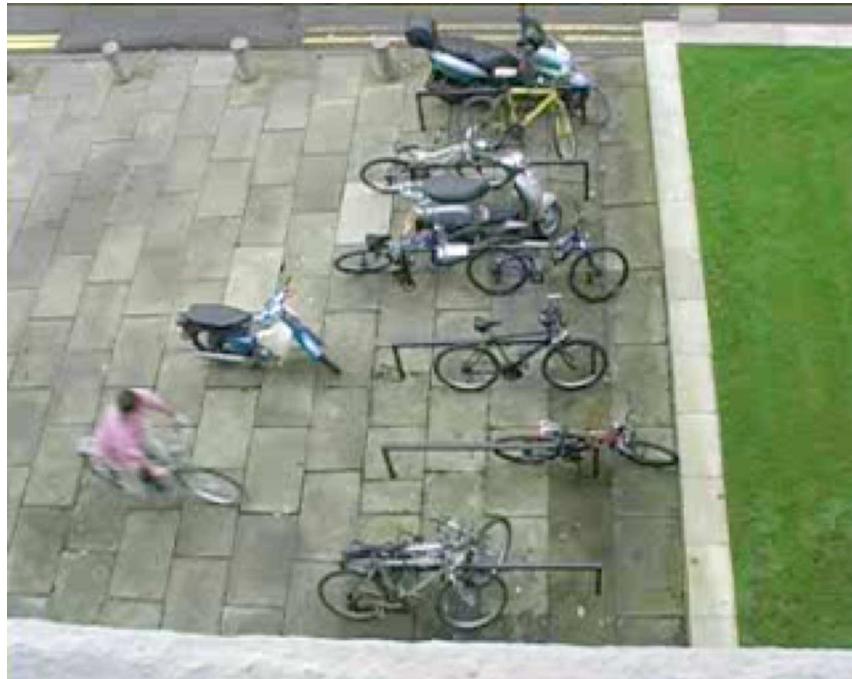


A fine-grained perspective onto object interactions

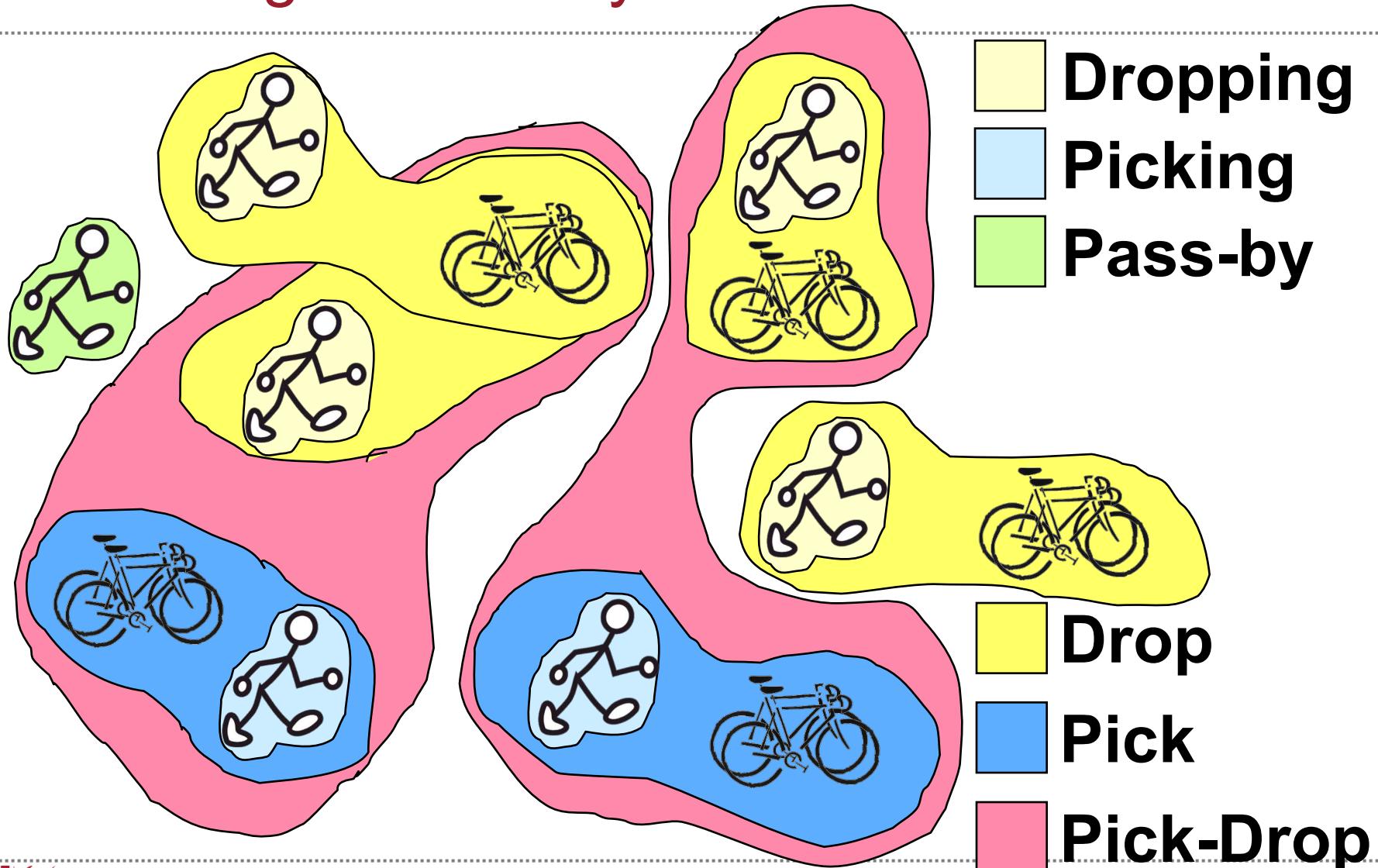
The Background Story...



The Background Story...

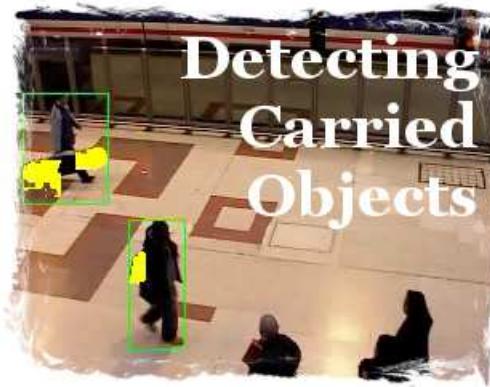
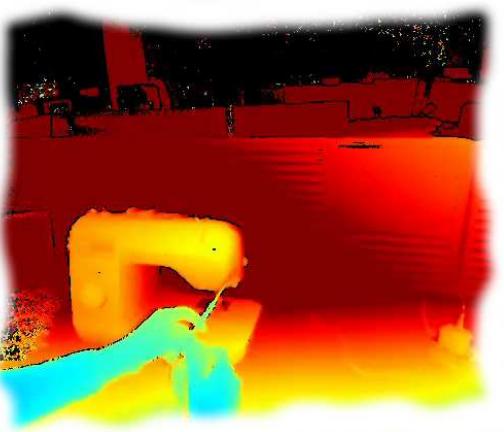


The Background Story...



The Background Story...

- A few more objects over the years...



Into First-Person Vision



Visual Sensing – the landscape



Visual Sensing – the landscape



Expensive

Visual Sensing – the landscape



Moveable

Visual Sensing – the landscape



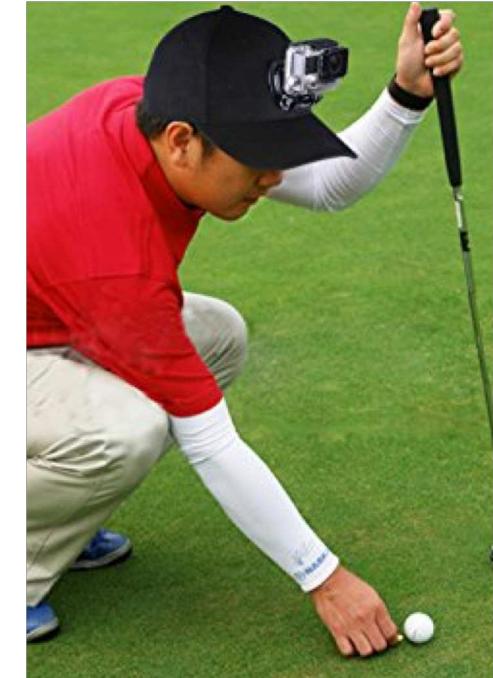
Wearable?



Wearable?



Wearable?

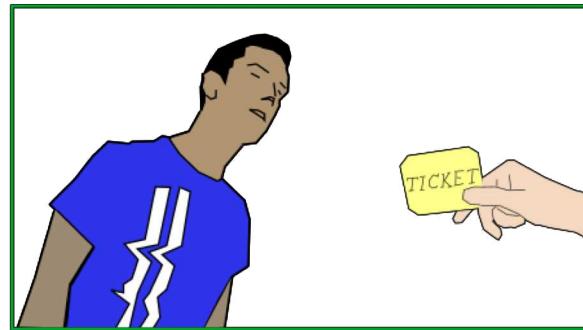


Wearable?



Wearable Cameras vs First Person View

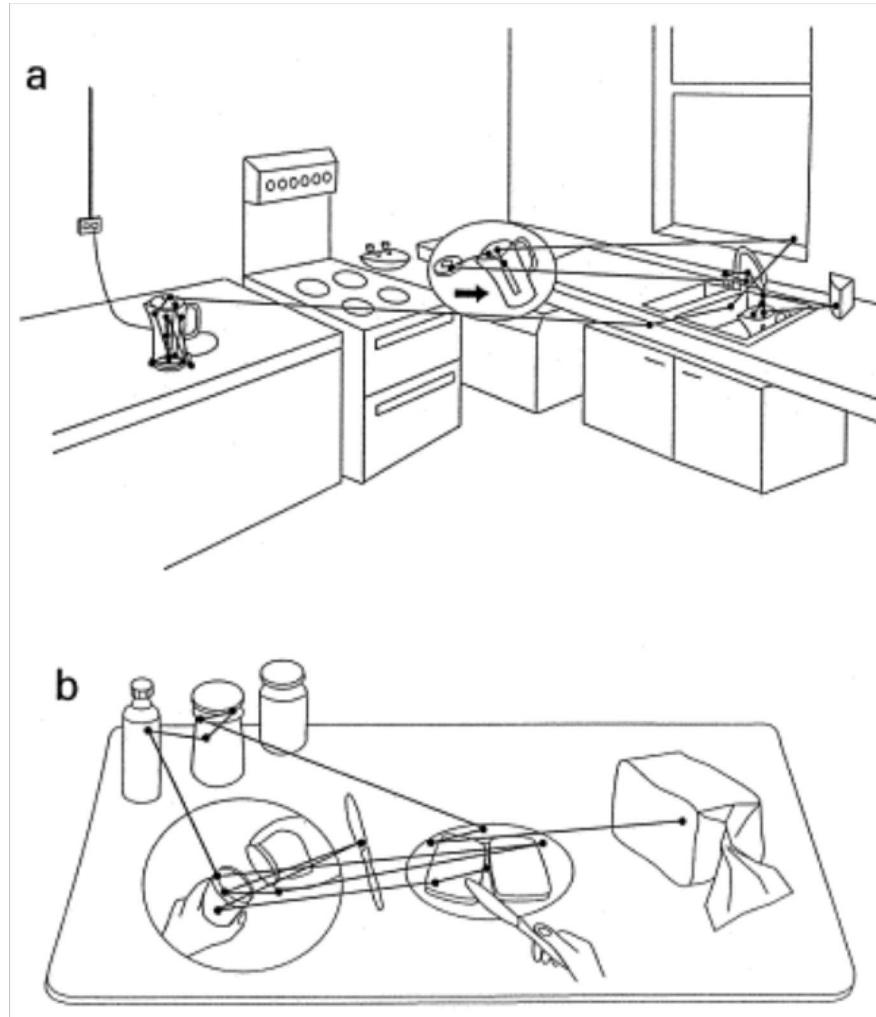
- OPV (Ordinal-Person Views)
 - FPV (First-Person View)
 - SPV (Second-Person View)
 - TPV (Third-Person View)



See for yourself!

- Videos...

Quick introduction to human gaze



Quick introduction to human gaze



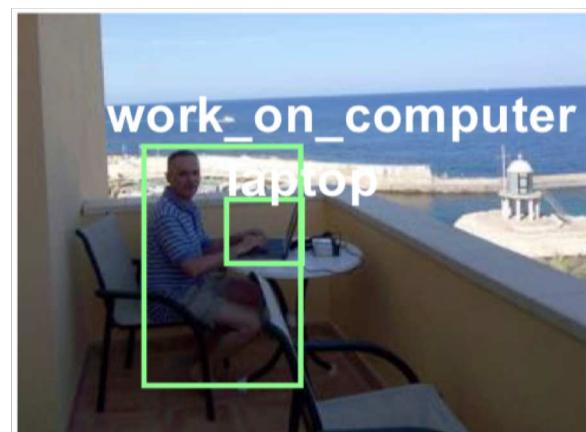
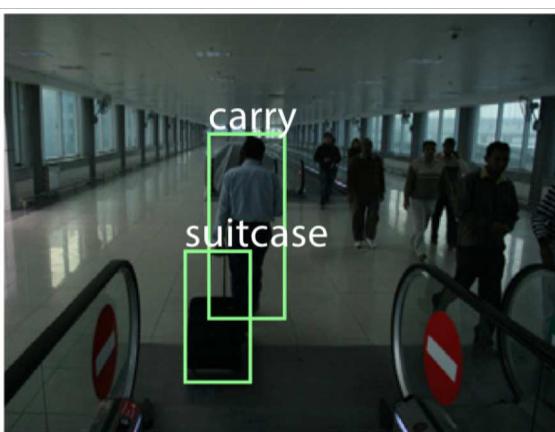
Hence...

a fine-grained perspective

to object interactions

from first-person video

Action Recognition/Understanding



ATTACKING	
AGENT	MAN
VICTIM	PLAYER
TOOL	CHAIR
PLACE	STADIUM

SIGNING	
AGENT	MAN
SIGNEDITEM	BOOK
TOOL	PEN
PLACE	SHOP

THROWING	
AGENT	MAN
ITEM	BASEBALL
DESTINATION	CATCHER
PLACE	BA

A man in a yellow jacket is looking at his phone with three others are in the background.

A man in a yellow jacket is looking at his phone with three others are in the background.

Action Recognition/Understanding



Action Recognition/Understanding

ACTIVITYNET

Expand all Collapse all

Drinking Beer(57)

- Food and drink preparation (337)
- Kitchen and food clean-up(85)
- Washing dishes(85)
- Sports, Exercise, and Recreation(3485)
- Socializing, Relaxing, and Leisure(1249)
- Arts and Entertainment(1039)
 - Dancing(379)
 - Tango(78)
 - Cheerleading(83)
 - Cumbia(76)
 - Breakdancing(79)
 - Belly dance(63)

Fishing (58949)

Drums (57517)

Choir (56753)

(55462) Pet (52614)

aircraft (50092)

e (49776) Dish (49708)

Highlight film (47867)

Holding a laptop Closing a laptop Put down laptop Taking a dish Taking a dish

DEEPMIND
KINETICS VIEWER

48 results for **auctioning**
Click on a thumbnail to play the video

tugransubasta.com

In this talk

- Why fine-grained?
 - You-Do, I-Learn
- Fine-Grained Problems
 - how ‘well’: Skill Determination in Video
 - when: Action Completion
 - when: Trespassing the Boundaries
 - which: Unequivocal Representation of Actions
- EPIC-KITCHENS 2018
 - Dataset collection and Challenges

In this talk

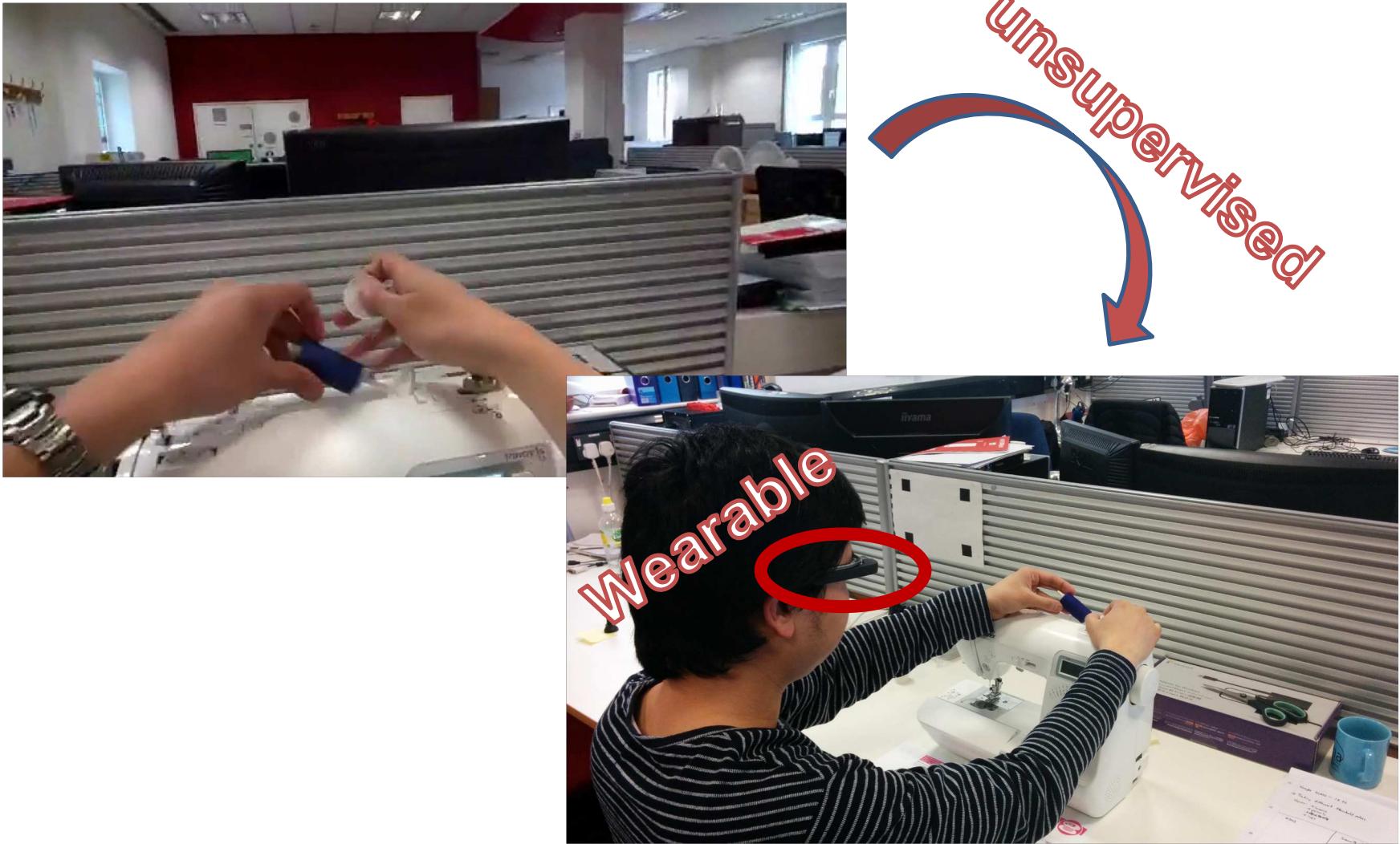
- Why fine-grained?
 - You-Do, I-Learn
- Fine-Grained Problems
 - how ‘well’: Skill Determination in Video
 - when: Action Completion
 - when: Trespassing the Boundaries
 - which: Unequivocal Representation of Actions
- EPIC-KITCHENS 2018
 - Dataset collection and Challenges

You-Do, I-Learn



D Damen, T Leelasawassuk, W Mayol-Cuevas (2016). You-Do, I-Learn: Egocentric Unsupervised Discovery of Objects and their Modes of Interaction Towards Video-Based Guidance. *Computer Vision and Image Understanding*

You-Do, I-Learn



You Do, I Learn - Demonstration



You Do, I Learn – Google Glass Prototype

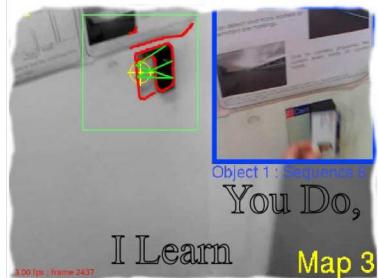
GlaciAR
Final Demo

Teesid Leelasawassuk, Dima
Damen and Walterio Mayol
University of Bristol

October 2014

More info...

Project You-Do, I-Learn



[Video1 \(2014\)](#), [Video2 \(2017\)](#)

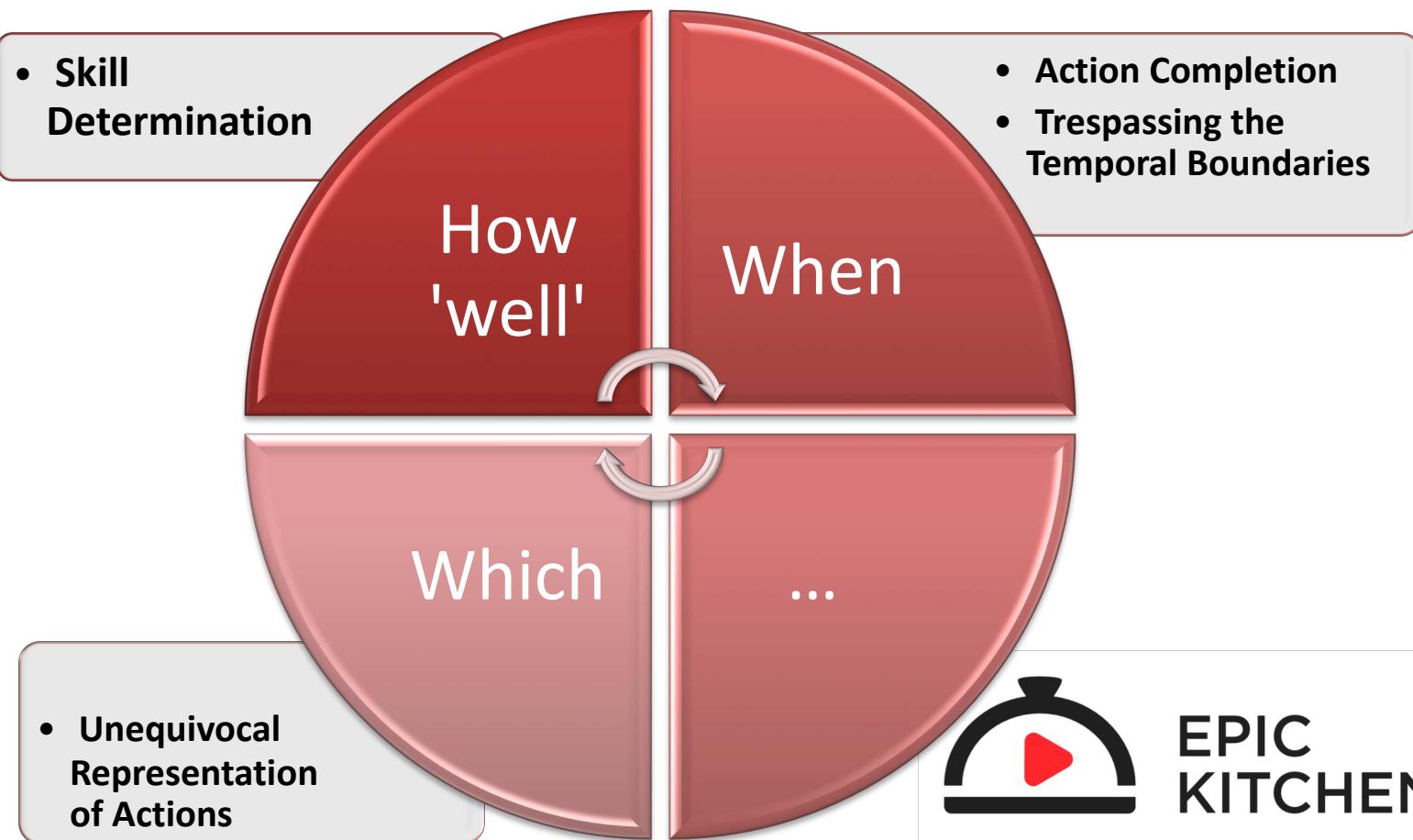
Automated capture and delivery of assistive task guidance with an eyewear computer: The GlaciAR system. T Leelasawassuk, D Damen, W Mayol-Cuevas. Augmented Human, Mar 2017 [pdf](#)

You-Do, I-Learn: Discovering Task Relevant Objects and their Modes of Interaction from Multi-User Egocentric Video. D Damen, T Leelasawassuk, O Haines, A Calway, W Mayol-Cuevas. British Machine Vision Conference (BMVC), Sep 2014. [PDF](#) | [Abstract](#) | [Dataset](#)

Multi-user egocentric Online System for Unsupervised Assistance on Object Usage. D Damen, O Haines, T Leelasawassuk, A Calway, W Mayol-Cuevas. ECCV Workshop on Assistive Computer Vision and Robotics (ACVR), Sep 2014. [PDF Preprint](#)

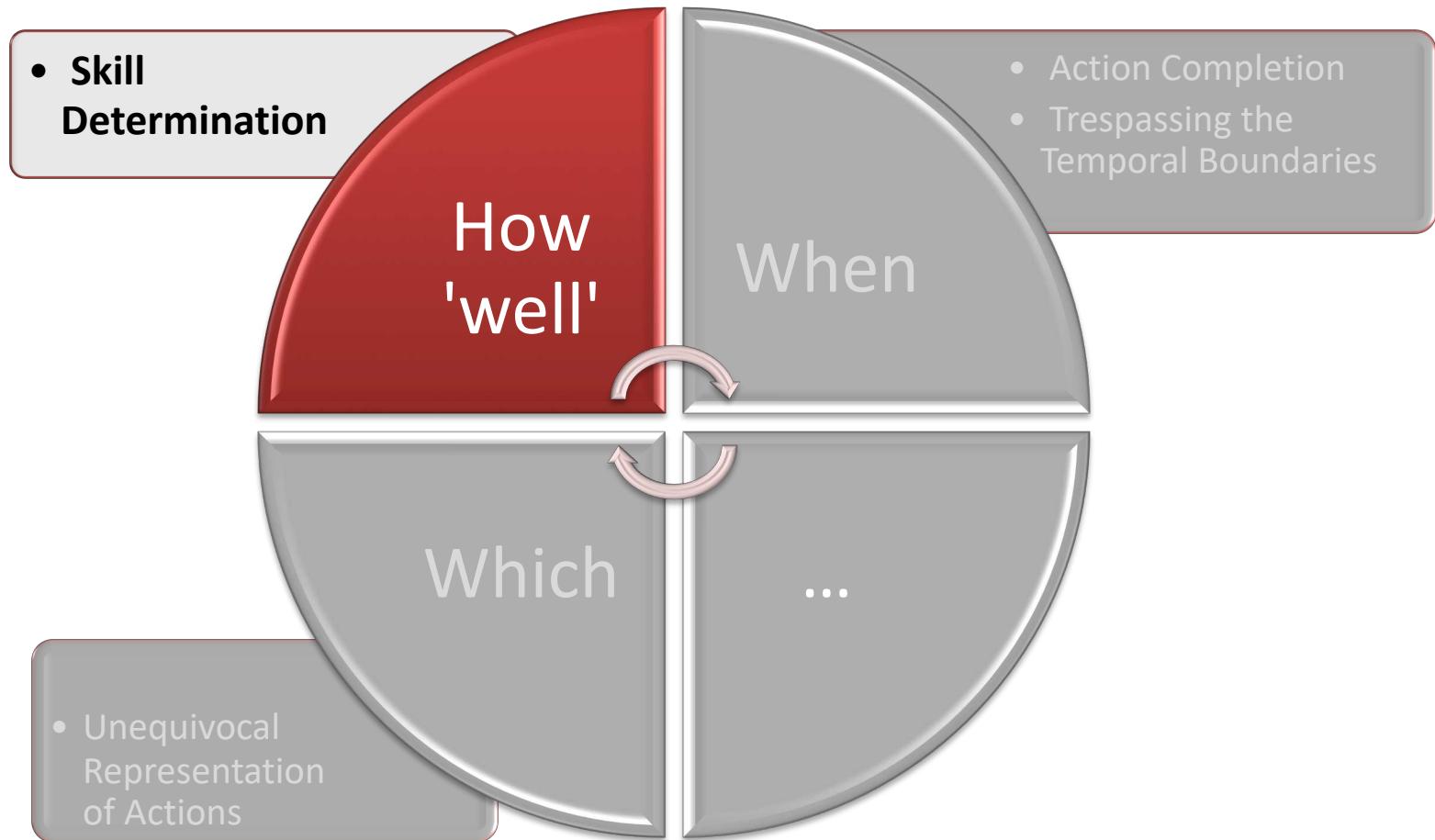
Estimating Visual Attention from a Head Mounted IMU. T Leelasawassuk, D Damen, W Mayol-Cuevas. International Symposium on Wearable Computers (ISWC), Sep 2015. [PDF](#)

Fine-Grained Object Interactions



**EPIC
KITCHENS**

Fine-Grained Object Interactions



Who's Better? Who's Best? Skill Determination in Video using Deep Ranking

with: Hazel Doughty
Walterio Mayol-Cuevas



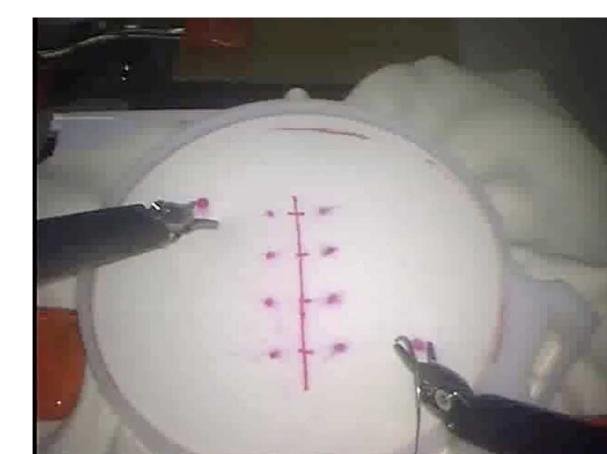
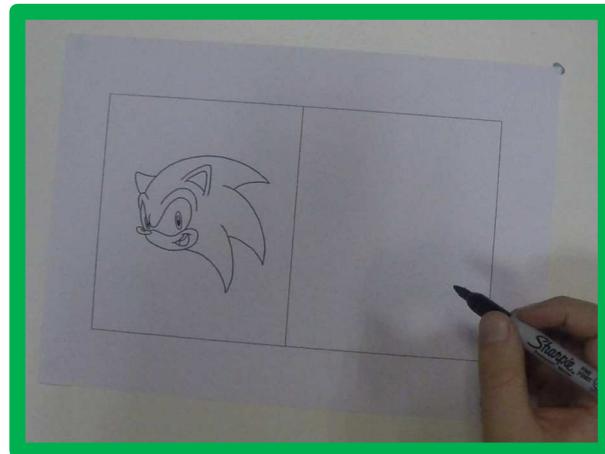
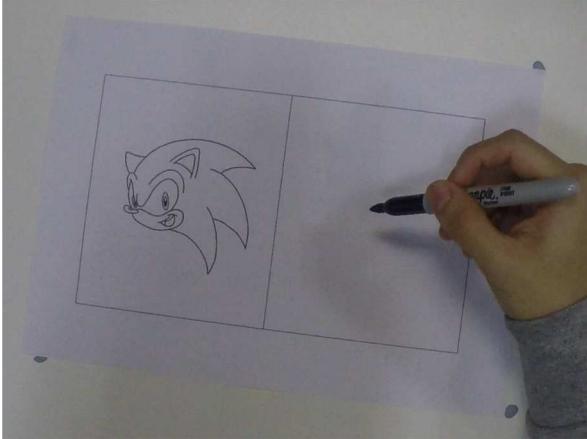
Assess relative skill for a collection of video sequences,
applicable to a variety of tasks.

Who's Better? Who's Best? Skill

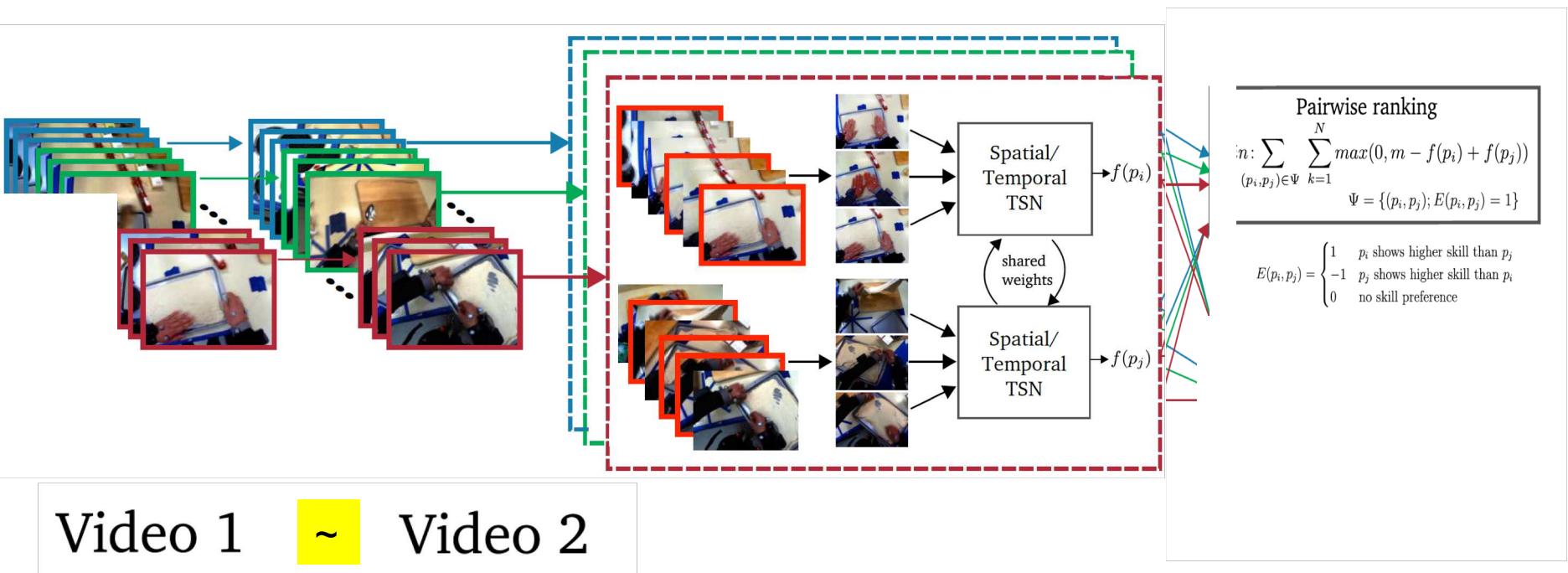
with: Hazel Doughty
Walterio Mayol-Cuevas

Determination in Video using Deep Ranking

Input: Pairwise annotations of videos, indicating higher skill or no skill preference



Who's Better? Who's Best? Skill Determination in Video using Deep Ranking



Who's Better? Who's Best? Skill Determination in Video using Deep Ranking

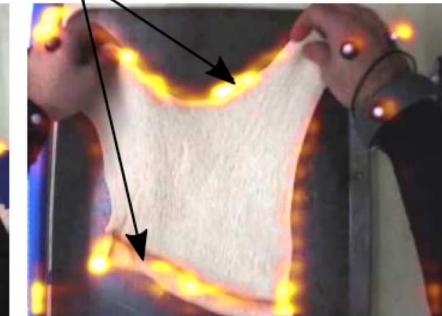
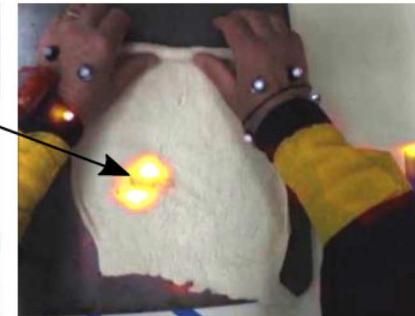
with: Hazel Doughty
Walterio Mayol-Cuevas

Dough Rolling

Holes in the dough

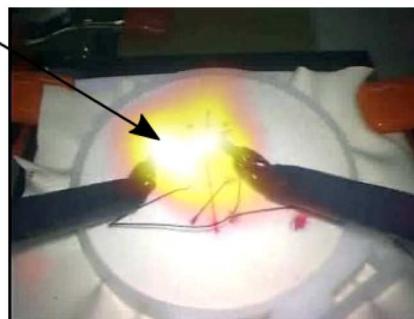
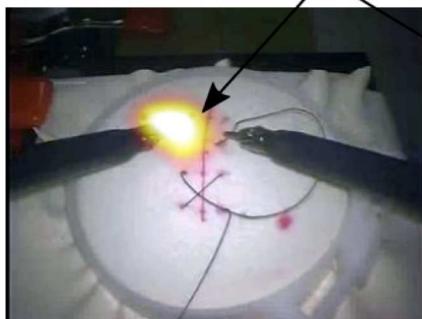


Curved or rolled edges

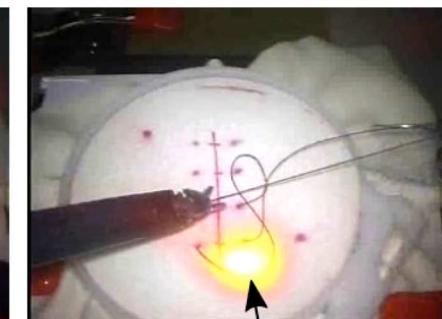
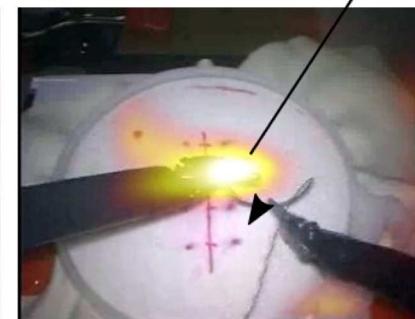


Surgery

Tissue damage



Abnormal needle pass



Loose Stitching

Best

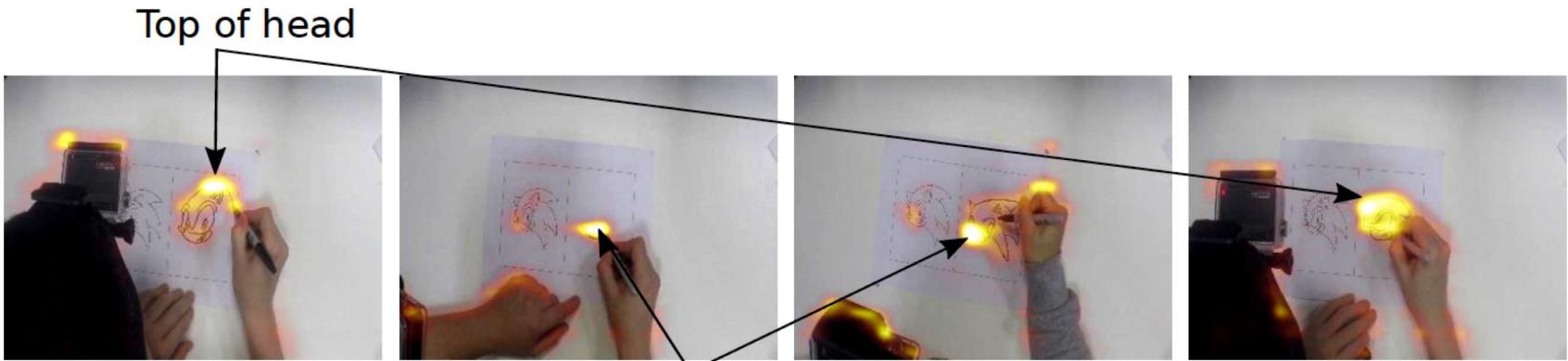


Worst

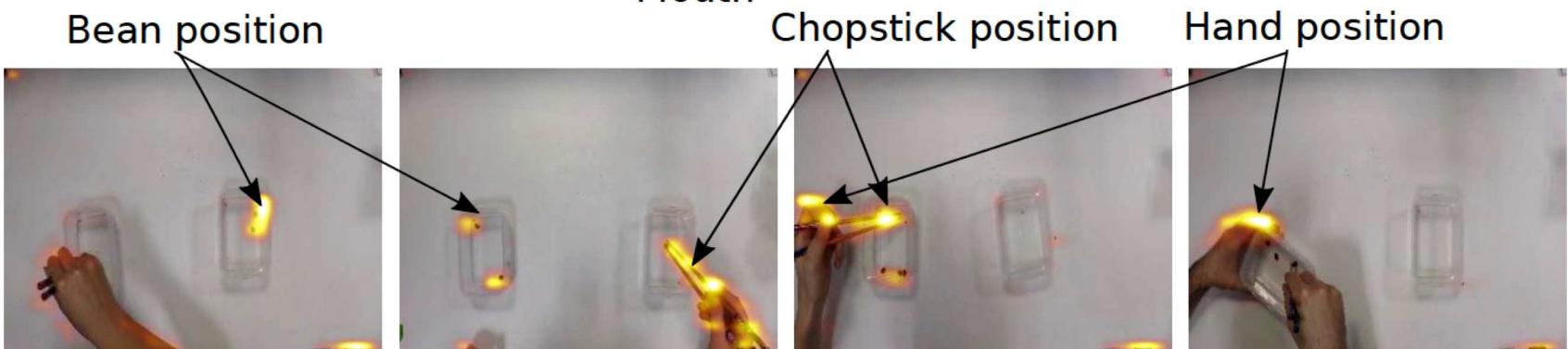
Who's Better? Who's Best? Skill Determination in Video using Deep Ranking

with: Hazel Doughty
Walterio Mayol-Cuevas

Drawing



Chopstick
Using

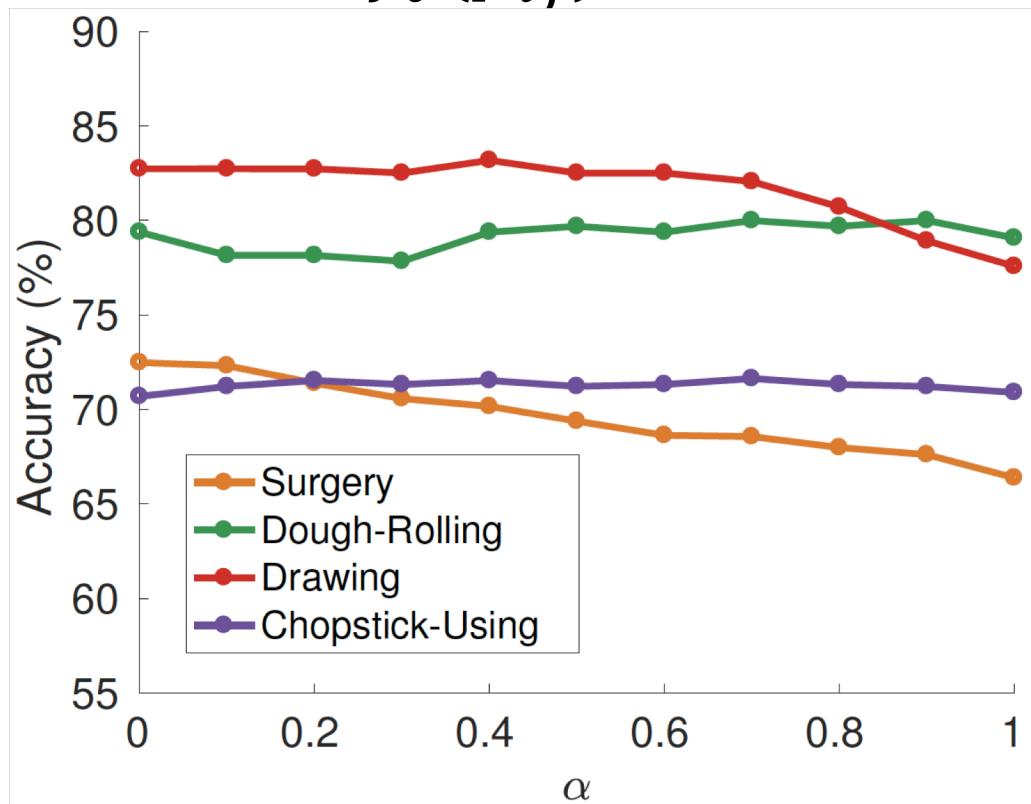


Best ← → Worst

Who's Better? Who's Best? Skill Determination in Video using Deep Ranking

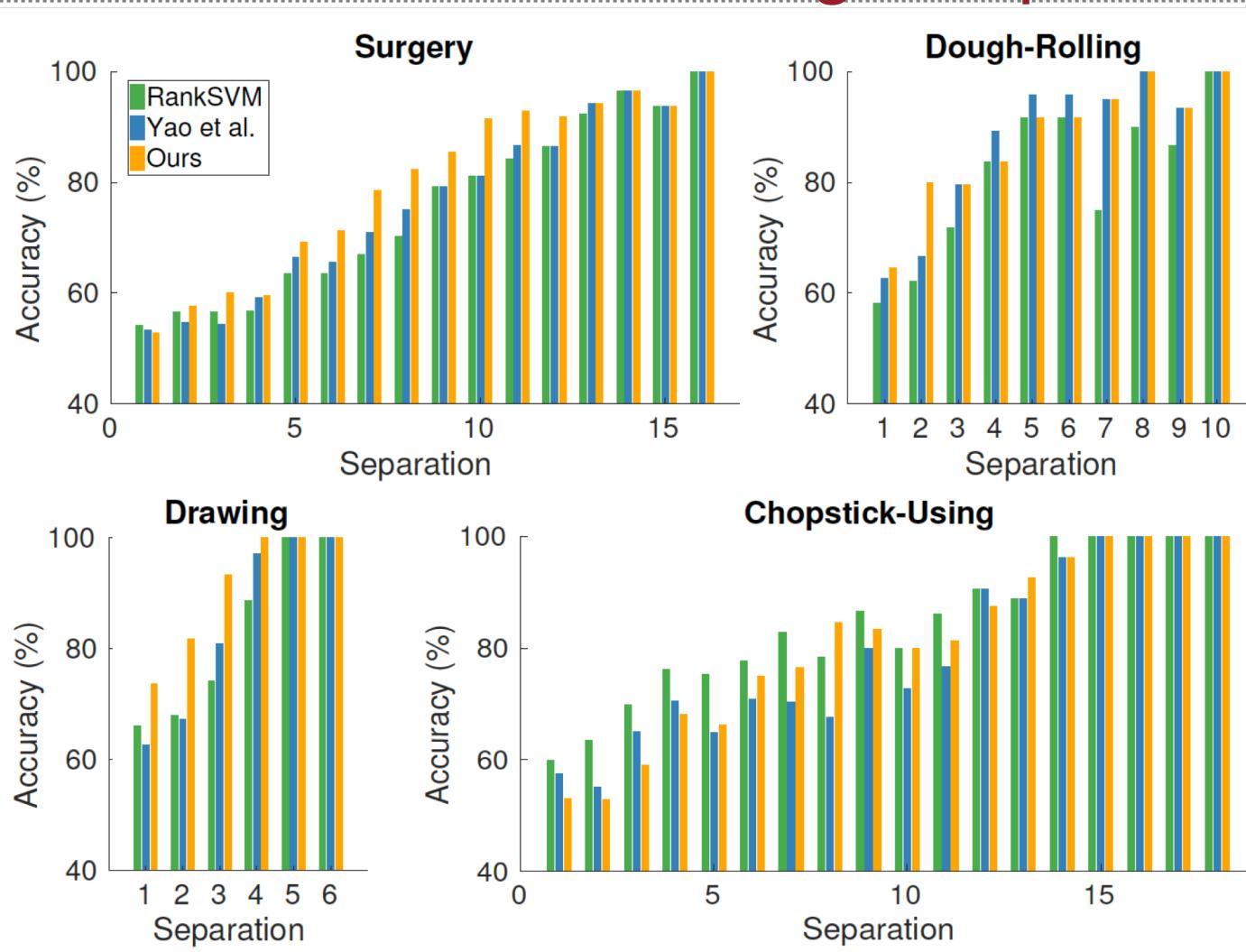
with: Hazel Doughty
Walterio Mayol-Cuevas

$$\frac{1}{\sigma} \sum_{j=1}^{\sigma} \alpha f_s(p_{ij}) + (1 - \alpha) f_t(p_{ij})$$



Who's Better? Who's Best? Skill Determination in Video using Deep Ranking

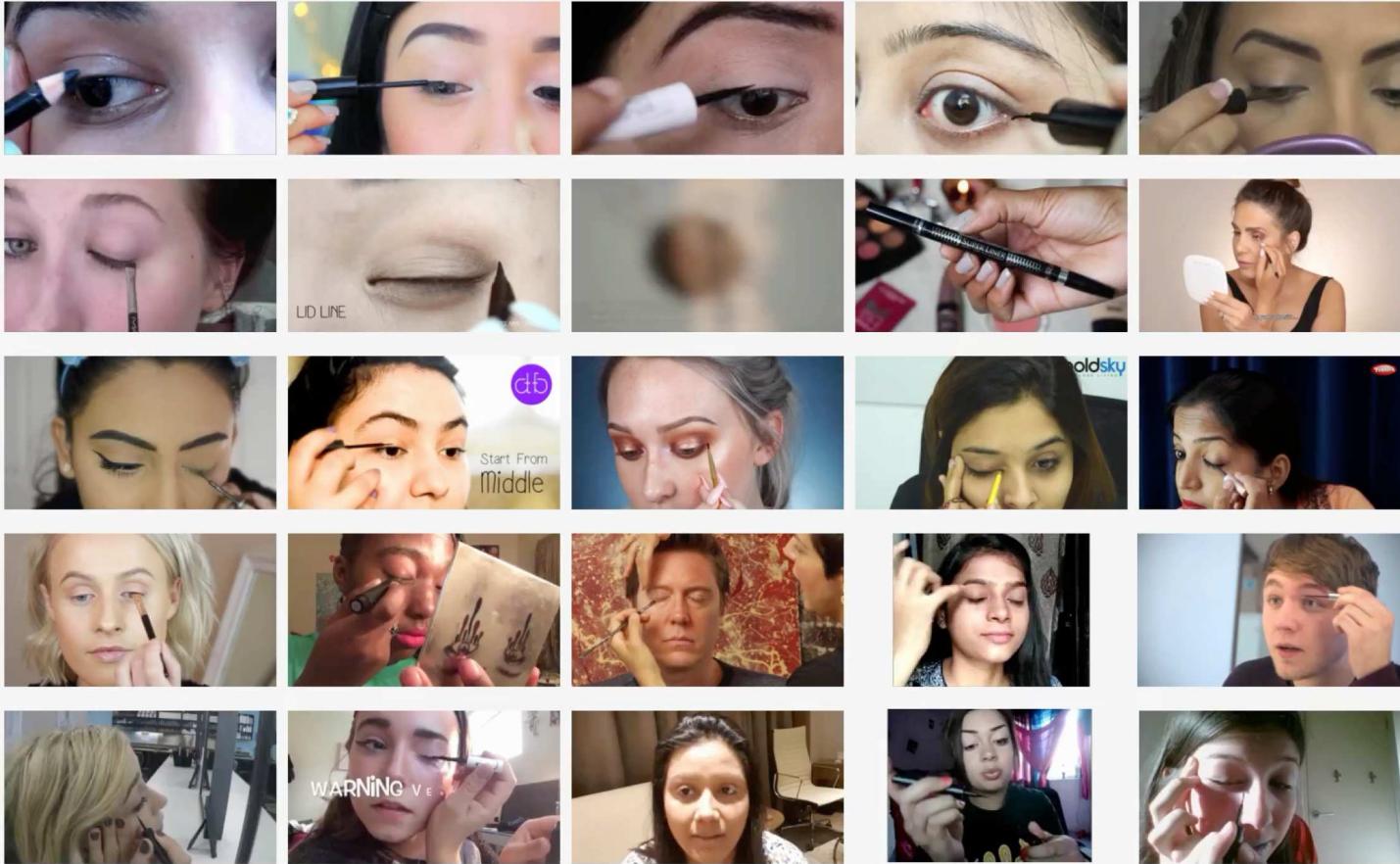
with: Hazel Doughty
Walterio Mayol-Cuevas



The Pros and Cons: Rank-Aware Temporal Attention

with: Hazel Doughty
Walterio Mayol-Cuevas

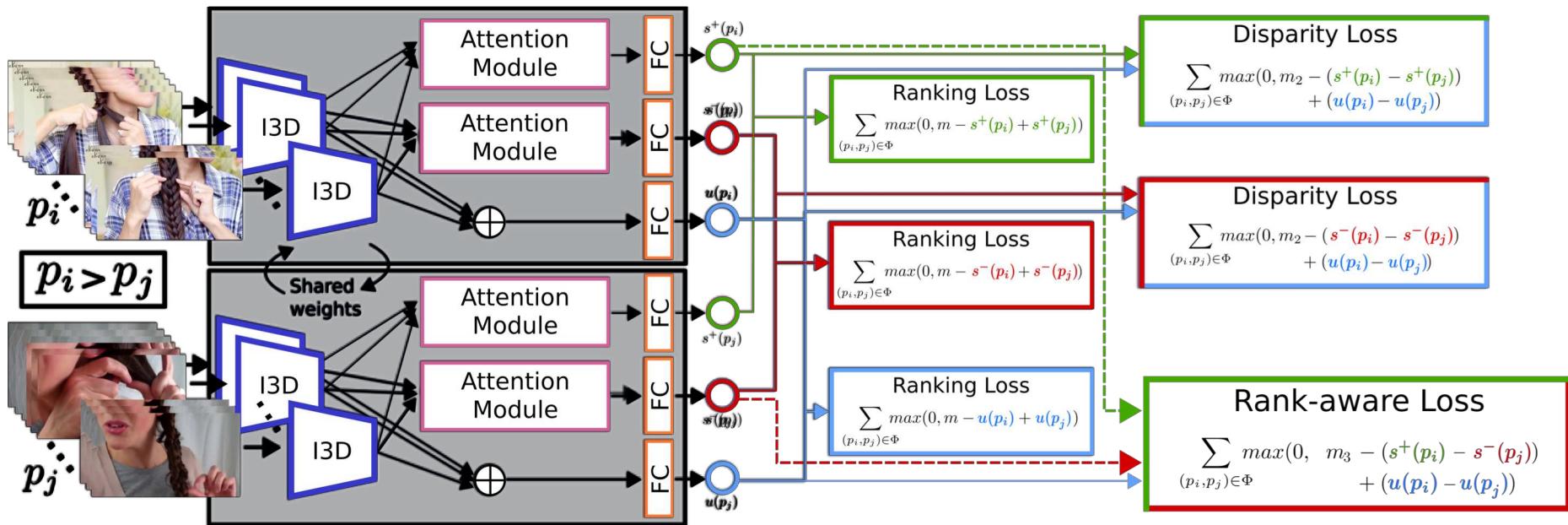
Best



Worst

The Pros and Cons: Rank-Aware Temporal Attention

with: Hazel Doughty
Walterio Mayol-Cuevas



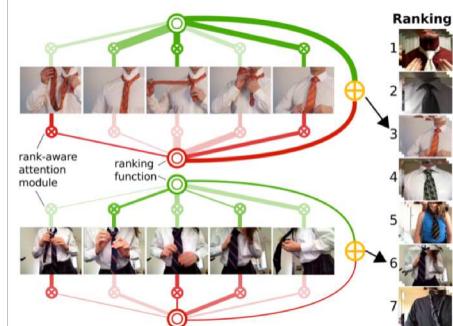
The Pros and Cons: Rank-Aware Temporal Attention

with: Hazel Doughty
Walterio Mayol-Cuevas



More info...

Project Skill Determination in Video

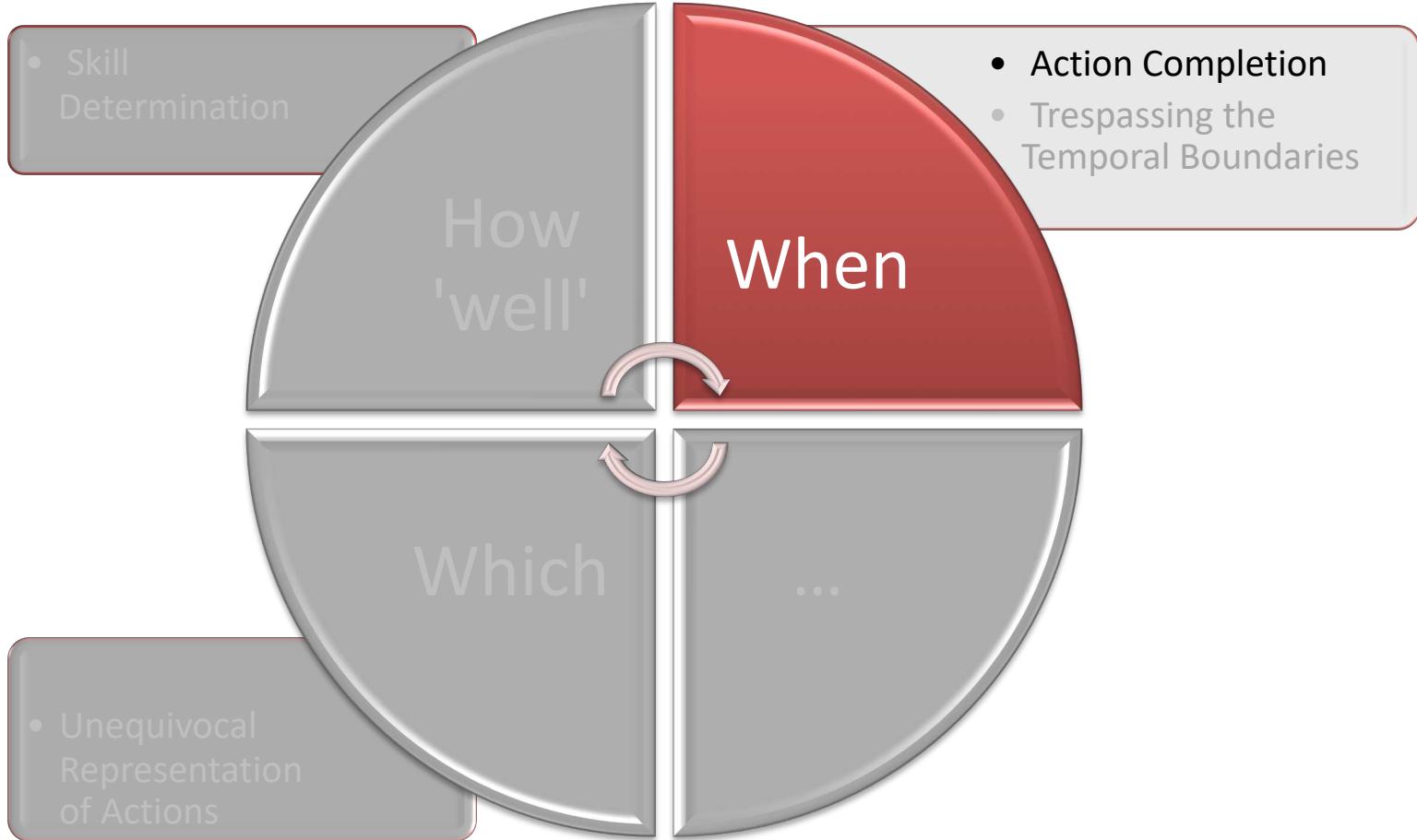


Video

The Pros and Cons: Rank-aware Temporal Attention for Skill Determination in Long Videos. H Doughty, W Mayol-Cuevas, D Damen. Arxiv (2018). [arxiv](#)

Who's Better? Who's Best? Pairwise Deep Ranking for Skill Determination. H Doughty, D Damen, W Mayol-Cuevas. CVPR (2018). [PDF](#) | [arxiv](#) | [Dataset](#)

Fine-Grained Object Interactions



Action Completion Detection



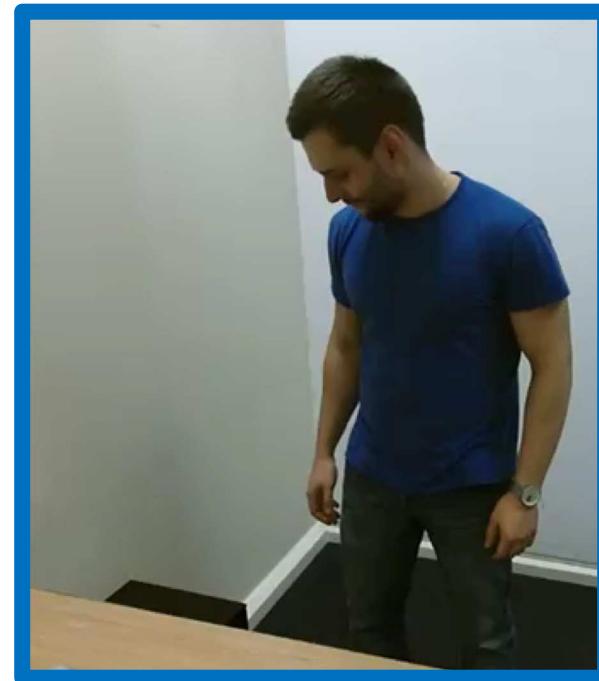
Understanding Completion

What if the observed action is not fully completed!?

Complete *pull*



Incomplete *pull*



Understanding Completion

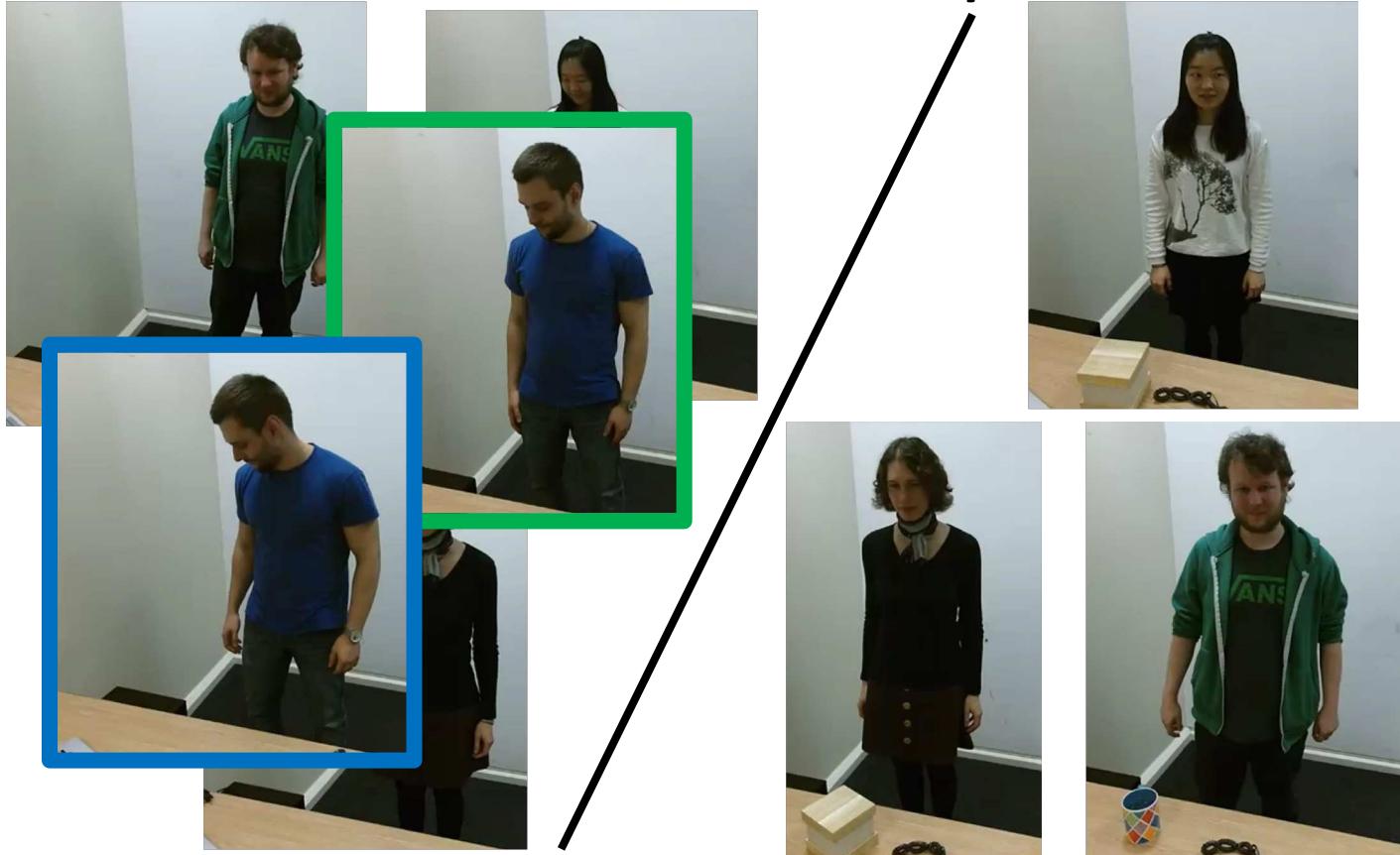
Pull-vs-pick



Complete *pull* and incomplete *pull* are introduced to *pull-vs-pick* classifier.

Understanding Completion

Pull-vs-pick



Both **complete pull** and **incomplete pull** are classified as **pull**.

Bristol Action Completion Dataset 2016

complete

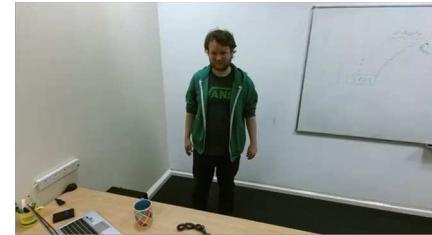


incomplete



switch

complete



incomplete



pull



plug



pick



open



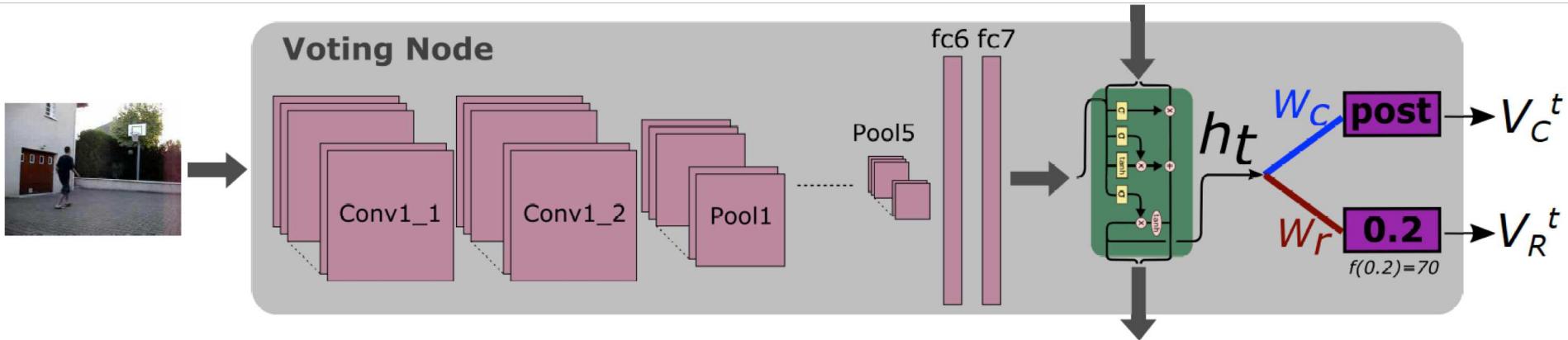
drink

Action Completion Detection



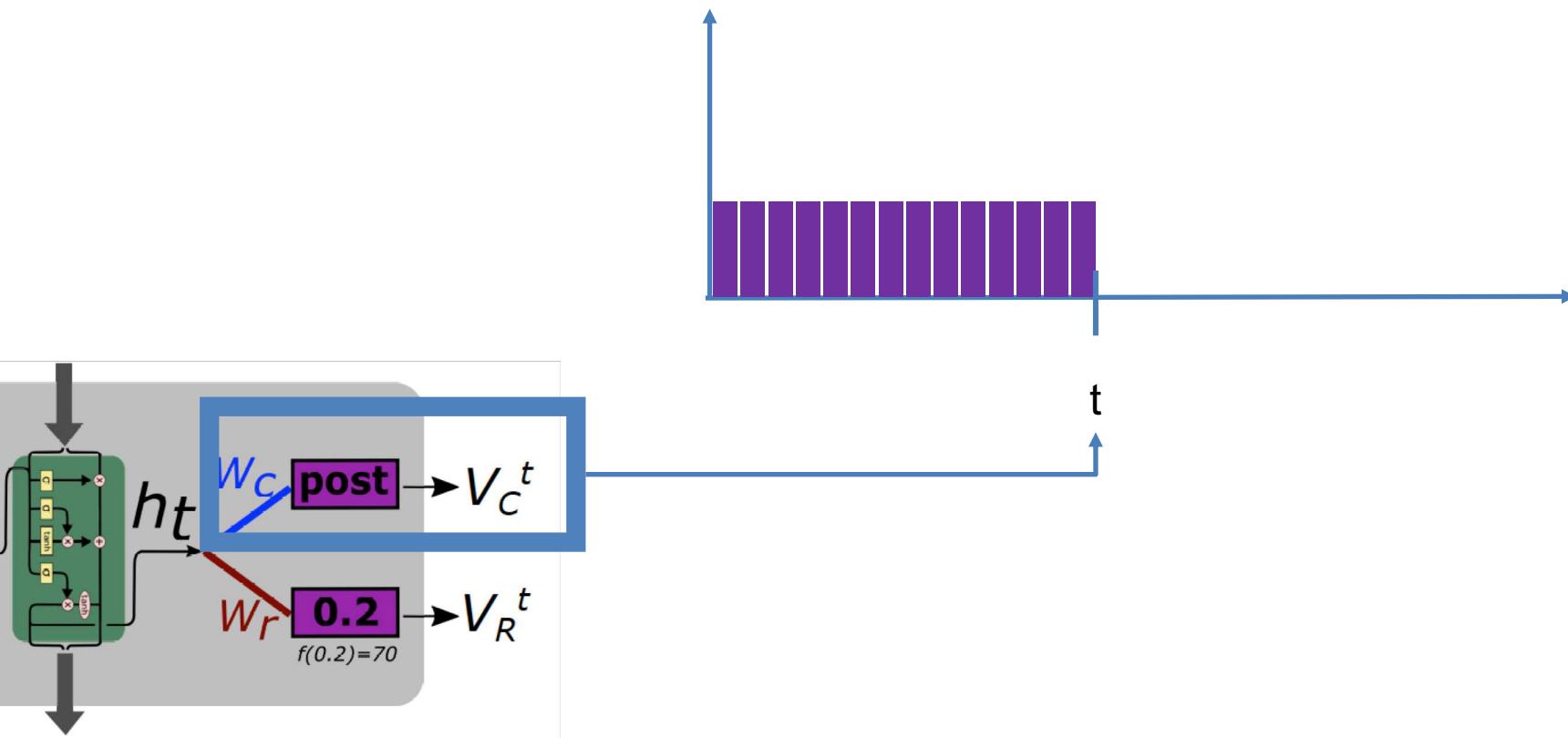
Action Completion Detection

- Each frame in the sequence, contributes to the completion moment detection via ‘voting’



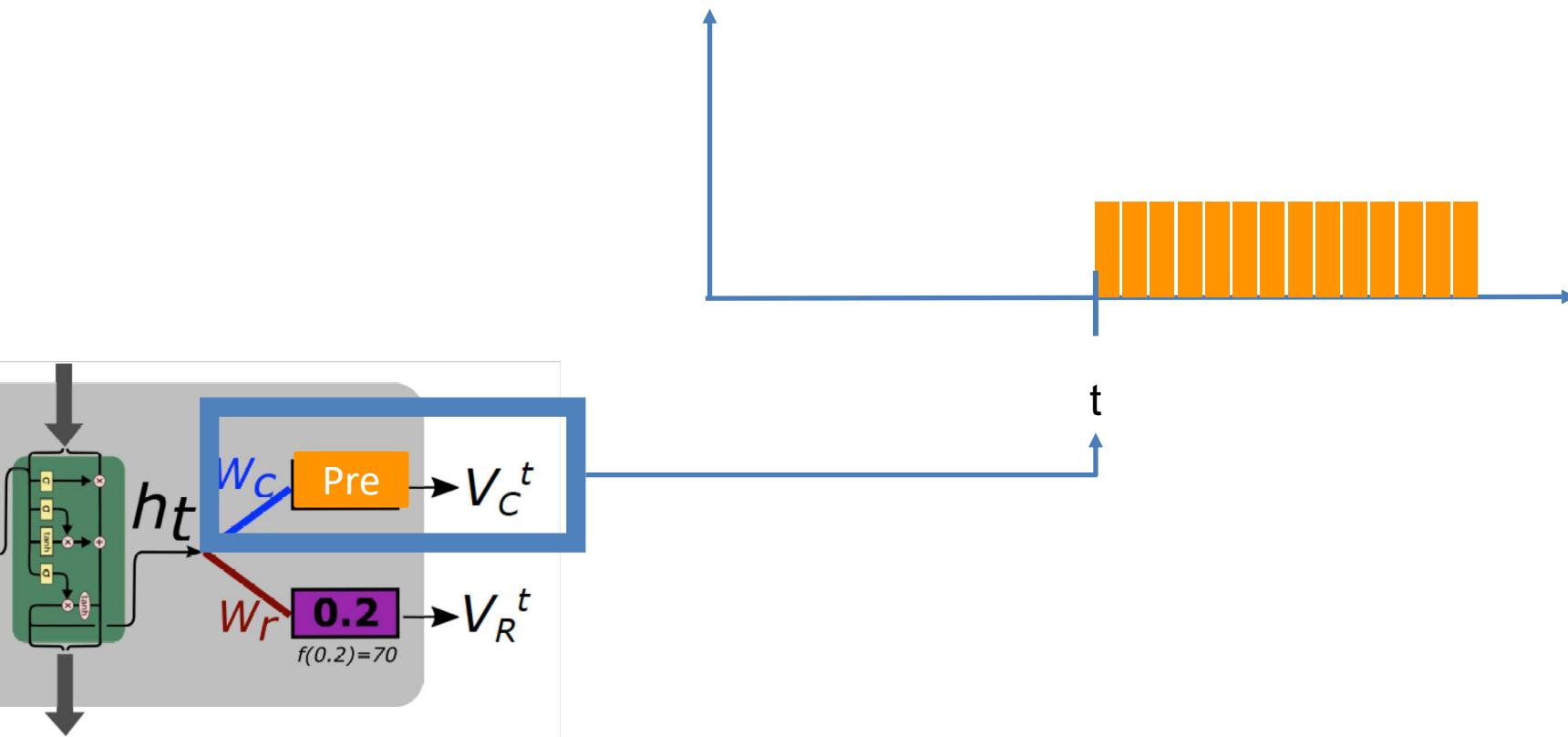
Action Completion Detection

1. Classification-Based Voting



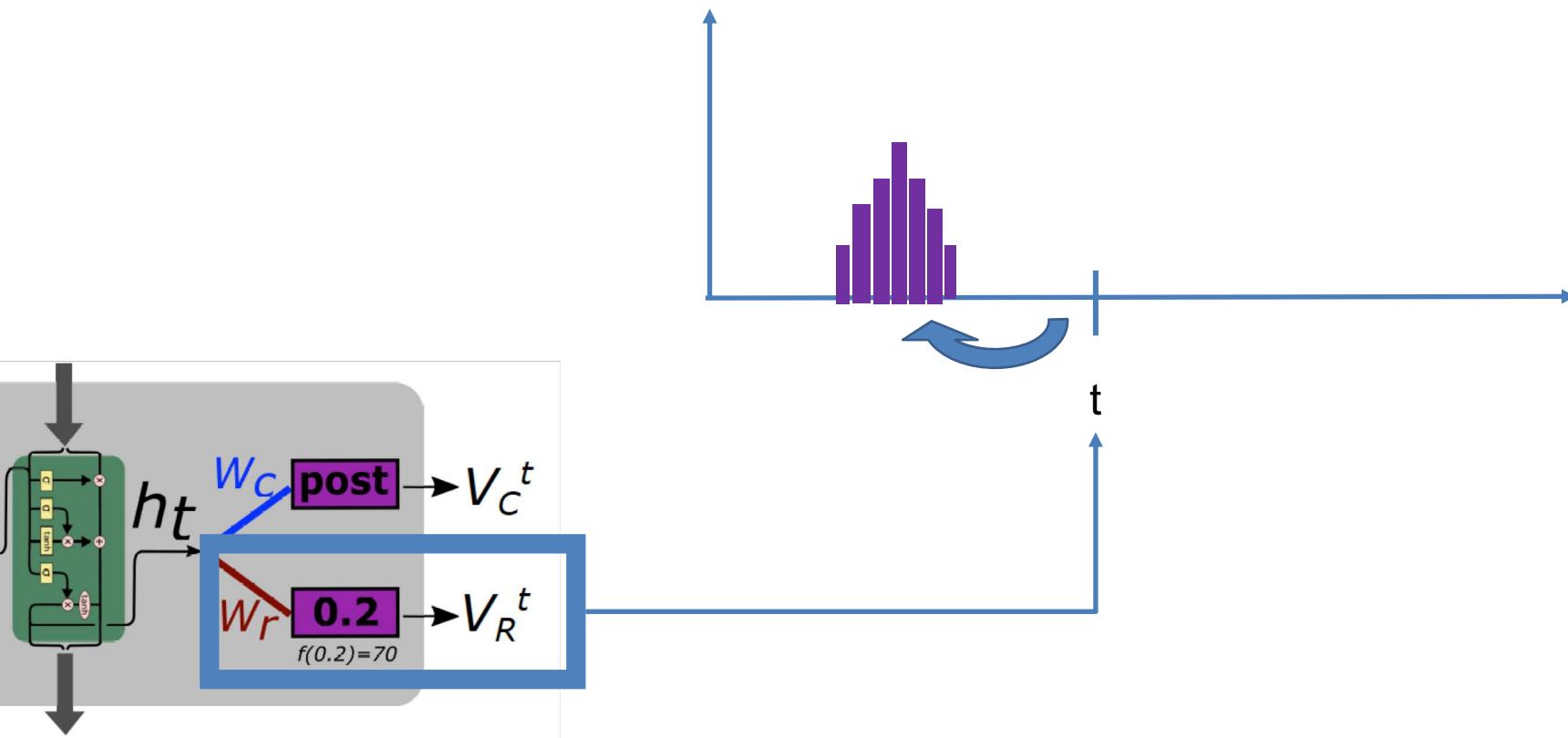
Action Completion Detection

1. Classification-Based Voting



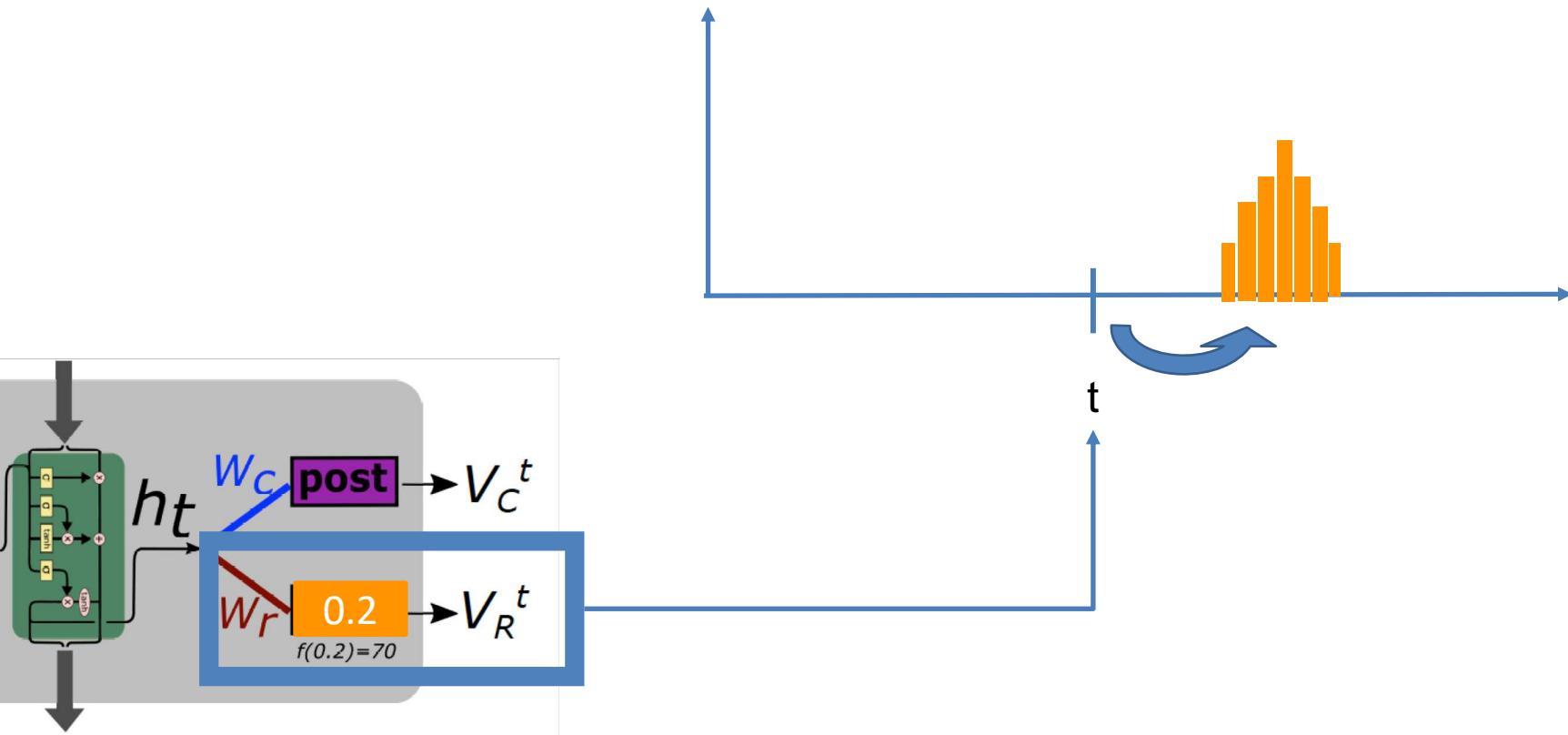
Action Completion Detection

2. Regression-Based Voting

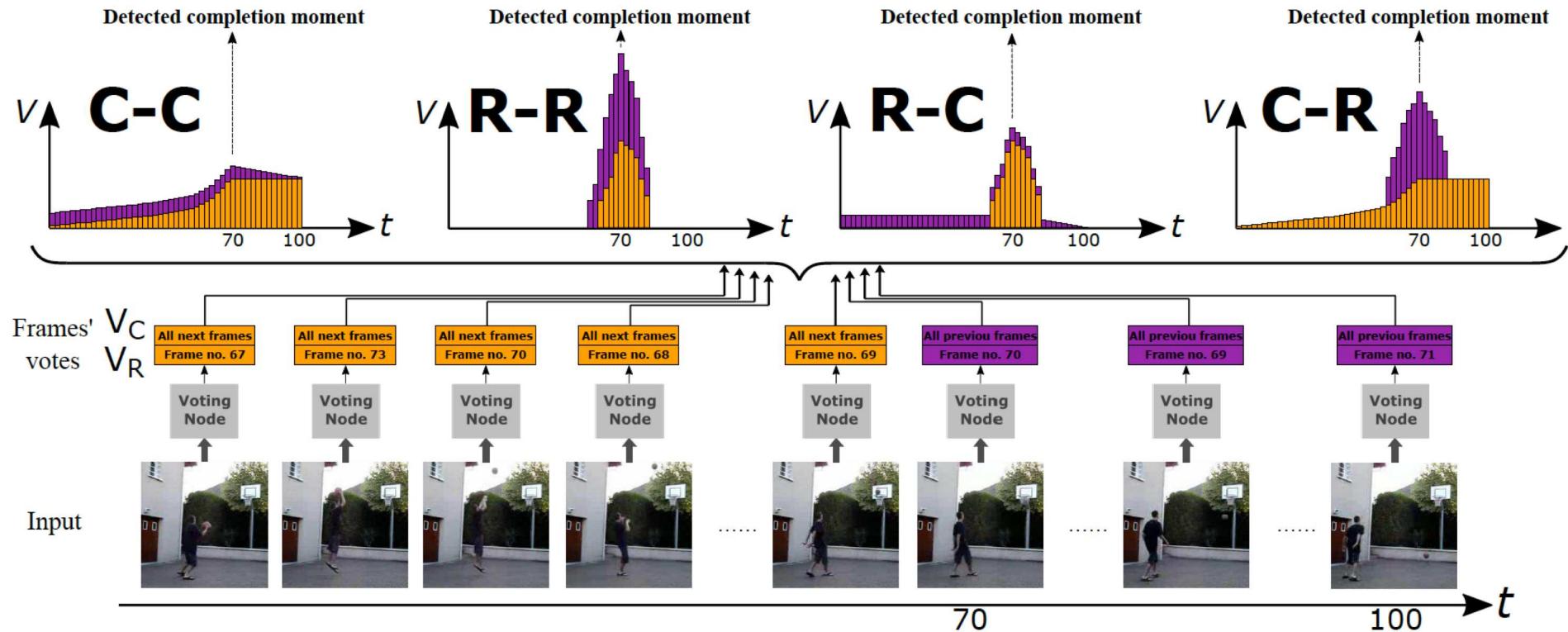


Action Completion Detection

2. Regression-Based Voting



Action Completion Detection



Action Completion Detection



Pre-V ←
 V_R^T ←
C-C ←
R-R ←
R-C ←
C-R ←
GT ←

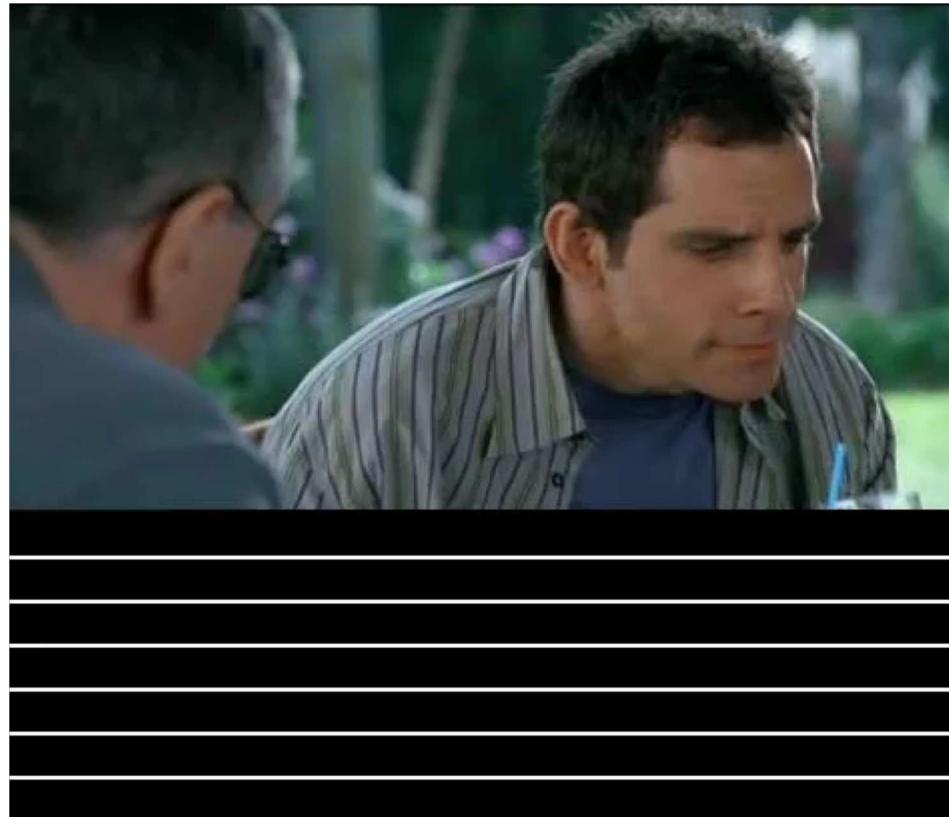
Action Completion Detection



Pre-V ←
 V_R^T ←
C-C ←
R-R ←
R-C ←
C-R ←
GT ←



Action Completion Detection



More info...

Project Action Completion

Action Completion Detection – Blowing Candles

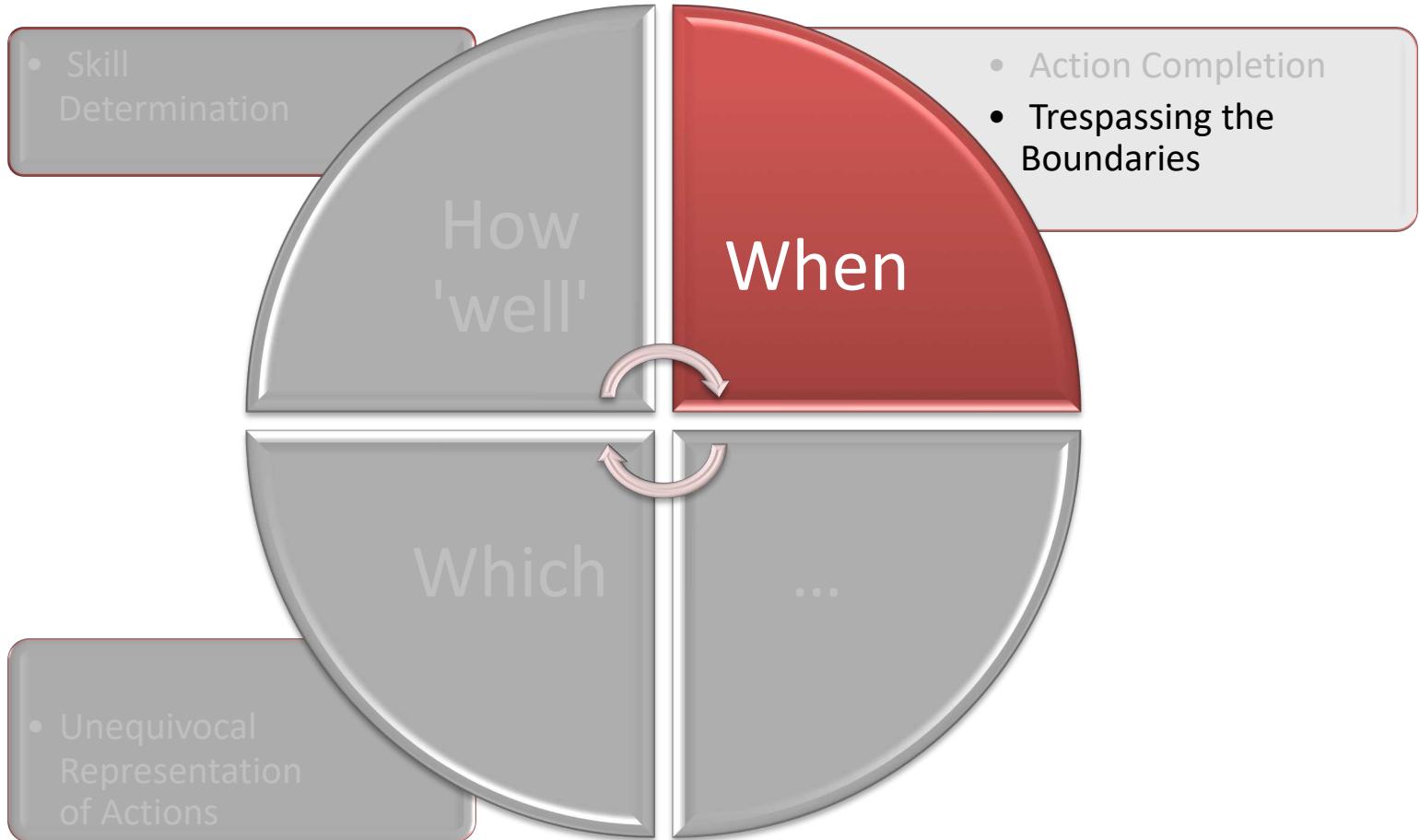


[Video](#)

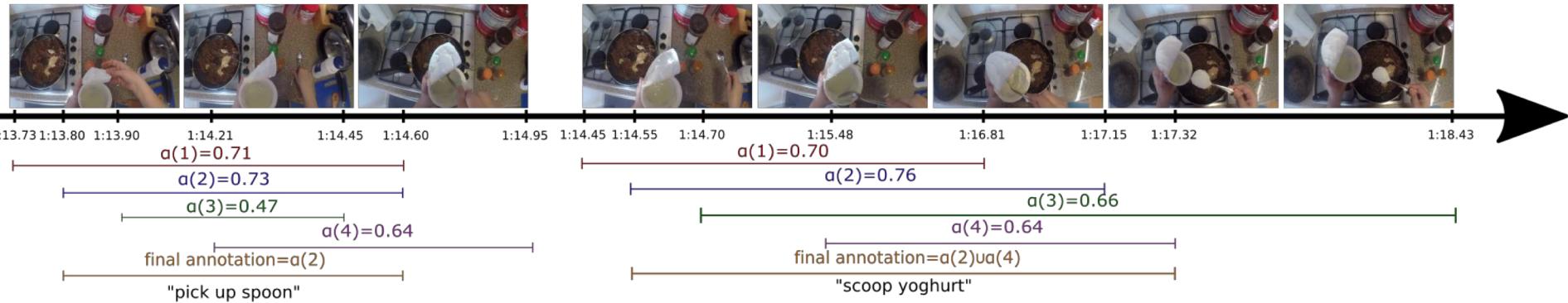
Action Completion: A Temporal Model for Moment Detection. F Heidarivincheh, M Mirmehdi, D Damen. Arxiv (2018) [Arxiv](#) | [Dataset](#)

Beyond Action Recognition: Action Completion in RGB-D Data. F Heidarivincheh, M Mirmehdi, D Damen. British Machine Vision Conference (BMVC), Sep 2016. [pdf](#) | [abstract](#) | [Dataset](#)

Fine-Grained Object Interactions



Temporal Boundaries for Object Interactions



- How robust are current state-of-the-art approaches to annotated boundaries in test segments?
- Modify test segment boundaries, maintaining significant overlap of segments $\text{IoU} > 0.5$
- **Correct in Green – Incorrect in Red**

Trespassing the Boundaries

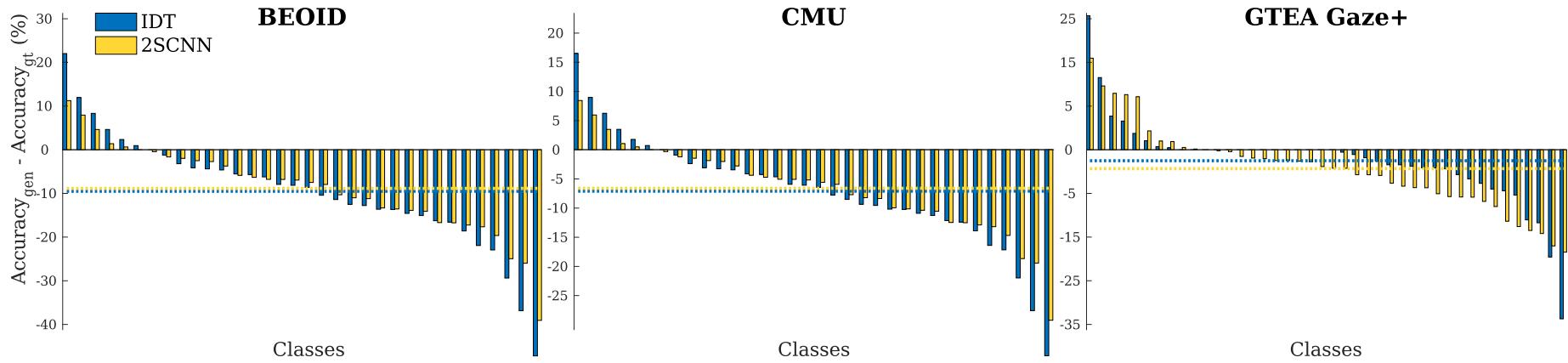
GTEA Gaze+

ground truth

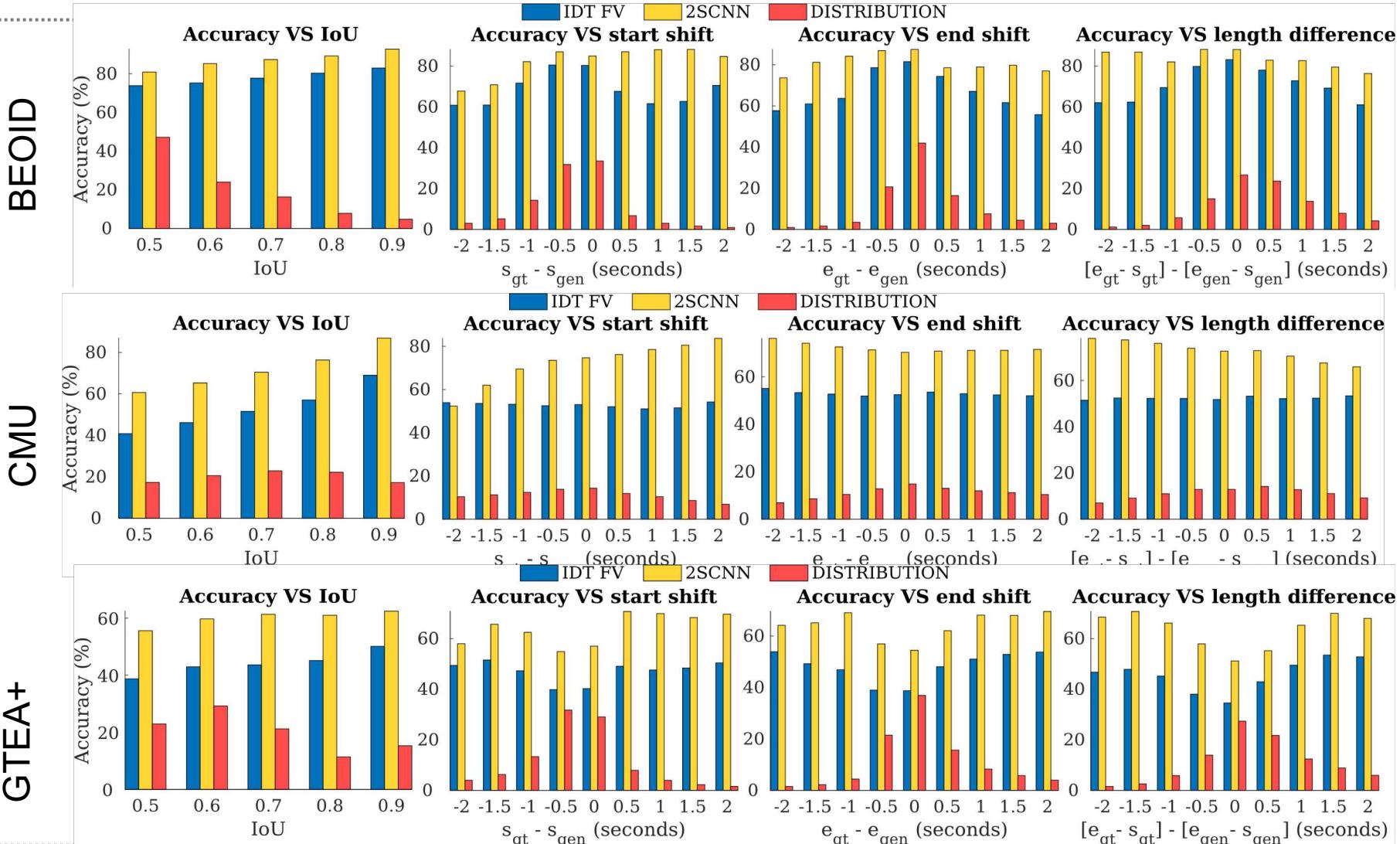


predicted class: take knife

Trespassing the Boundaries



Trespassing the Boundaries

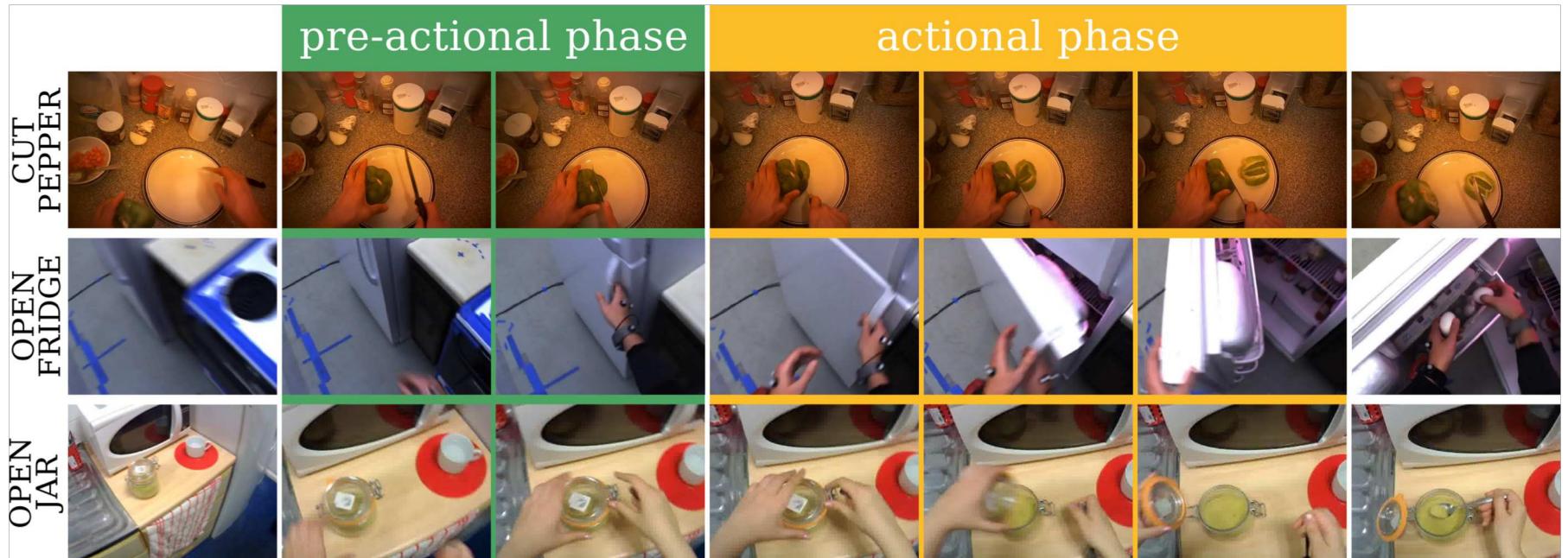


The Rubicon Boundaries

- Labelling approach proposal for temporally consistent annotations
- Decomposes an object interaction into two phases:
 - *pre-actional* phase
 - *actional* phase



The Rubicon Boundaries



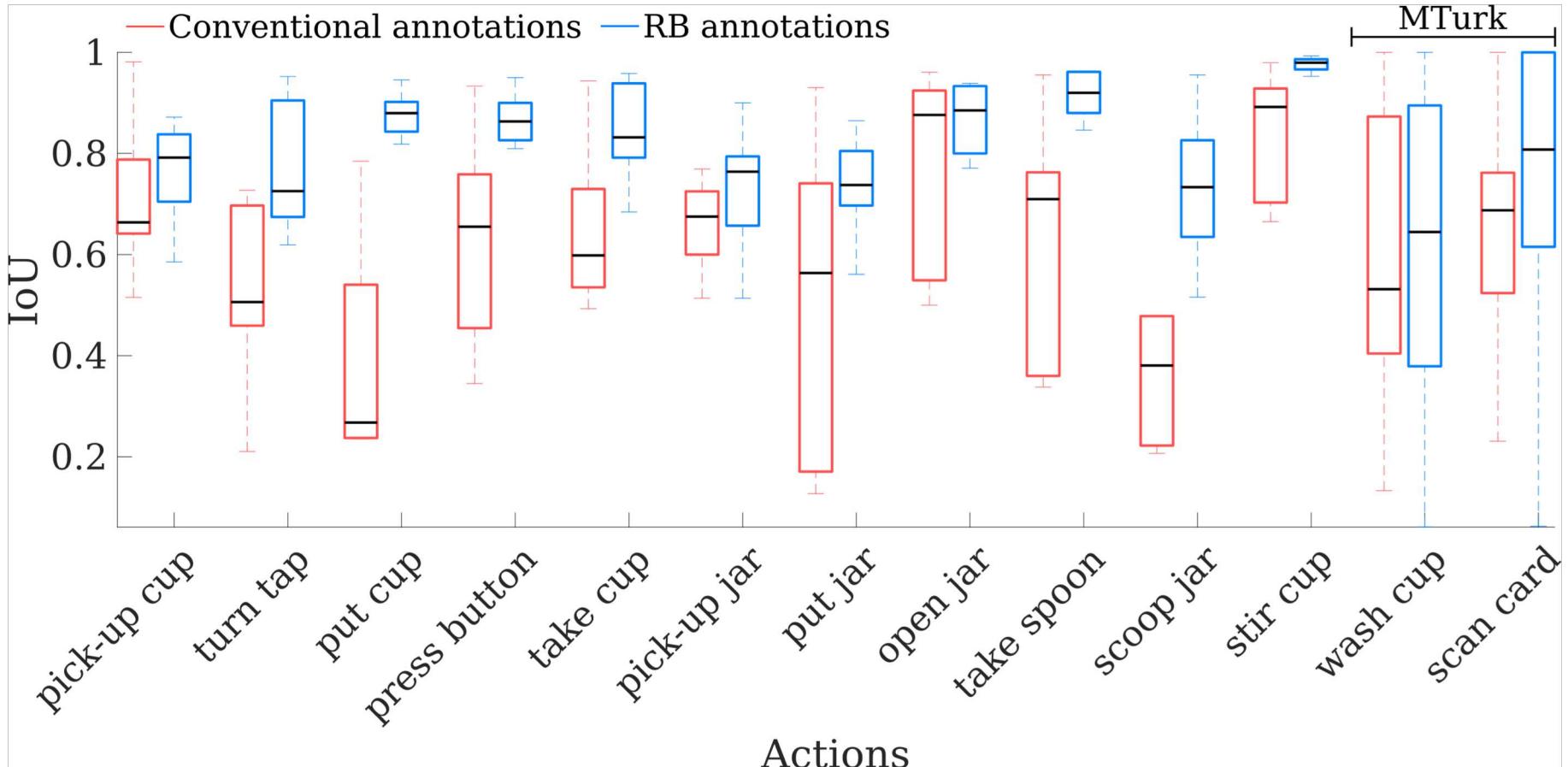
with: Davide Moltisanti
Michael Wray
Walterio Mayol-Cuevas

The Rubicon Boundaries

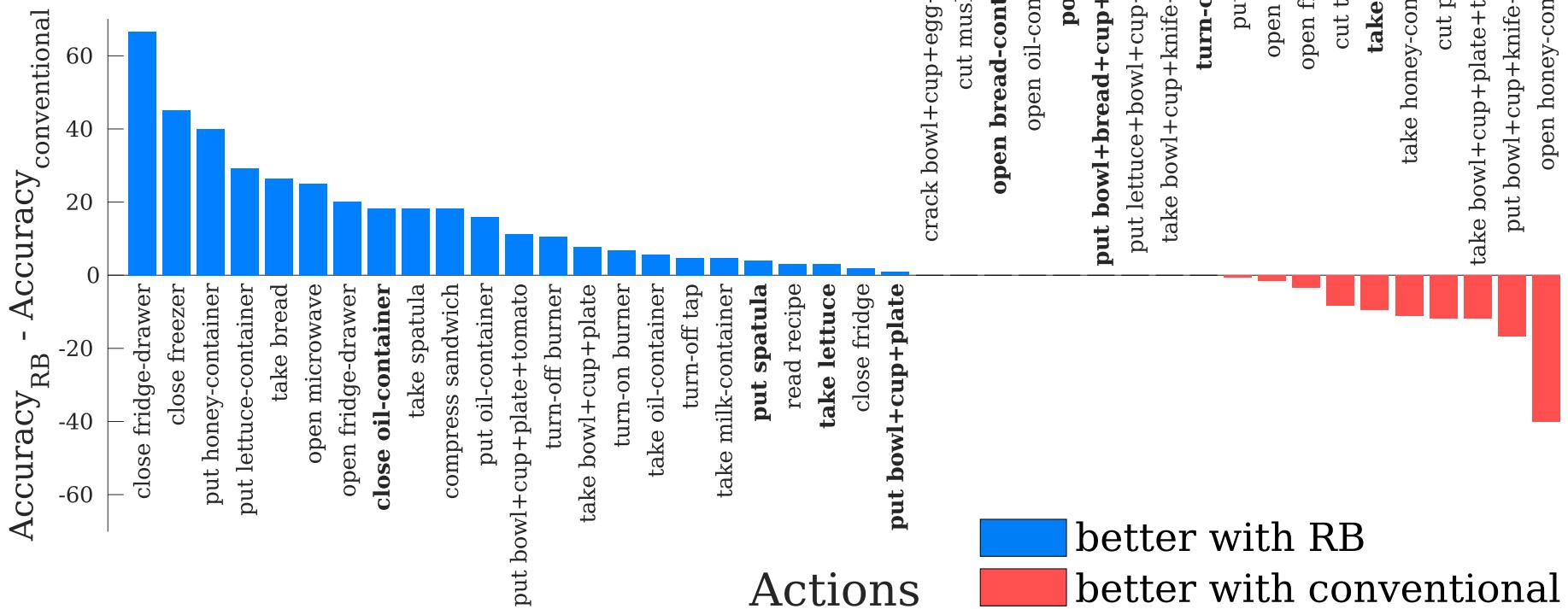
Cut pepper (GTEA Gaze+)



The Rubicon Boundaries

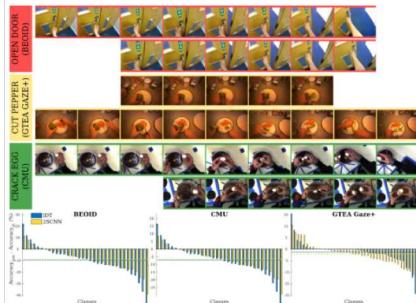


The Rubicon Boundaries



More info...

Project Trespassing the Boundaries of Object Interactions



Video

Trespassing the Boundaries: Labeling Temporal Bounds for Object Interactions in Egocentric Video. D Moltisanti, M Wray, W Mayol-Cuevas, D Damen. International Conference on Computer Vision (ICCV), 2017. [pdf](#) (camera ready) | [arxiv](#)

Upcoming (CVPR 2019)...

- Learning from Single timestamps



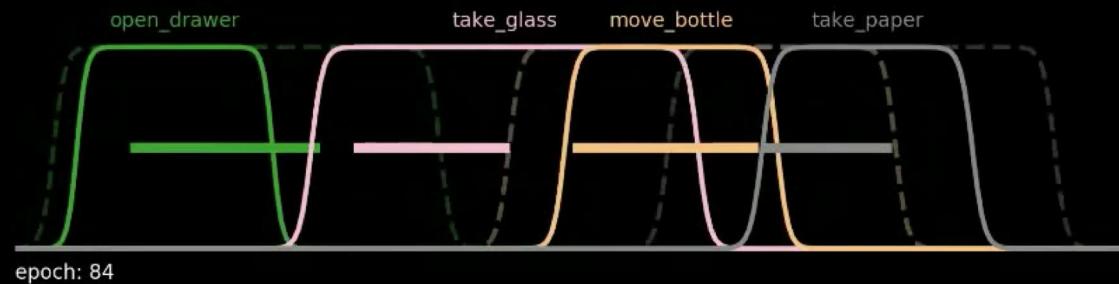
Upcoming (CVPR 2019)...

- Learning from Single timestamps

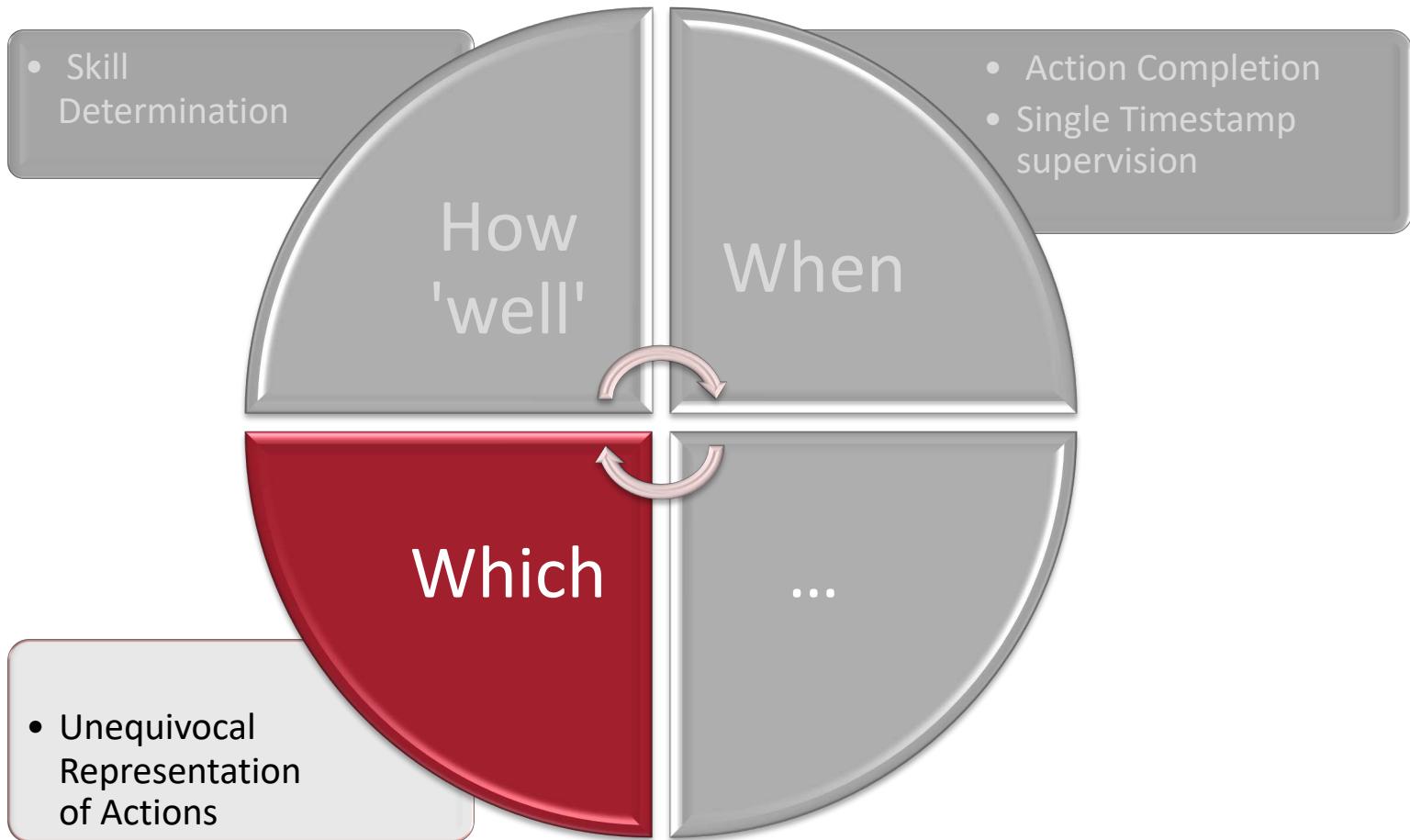


Upcoming (CVPR 2019)...

i) EPIC Kitchens (success)



Fine-Grained Object Interactions



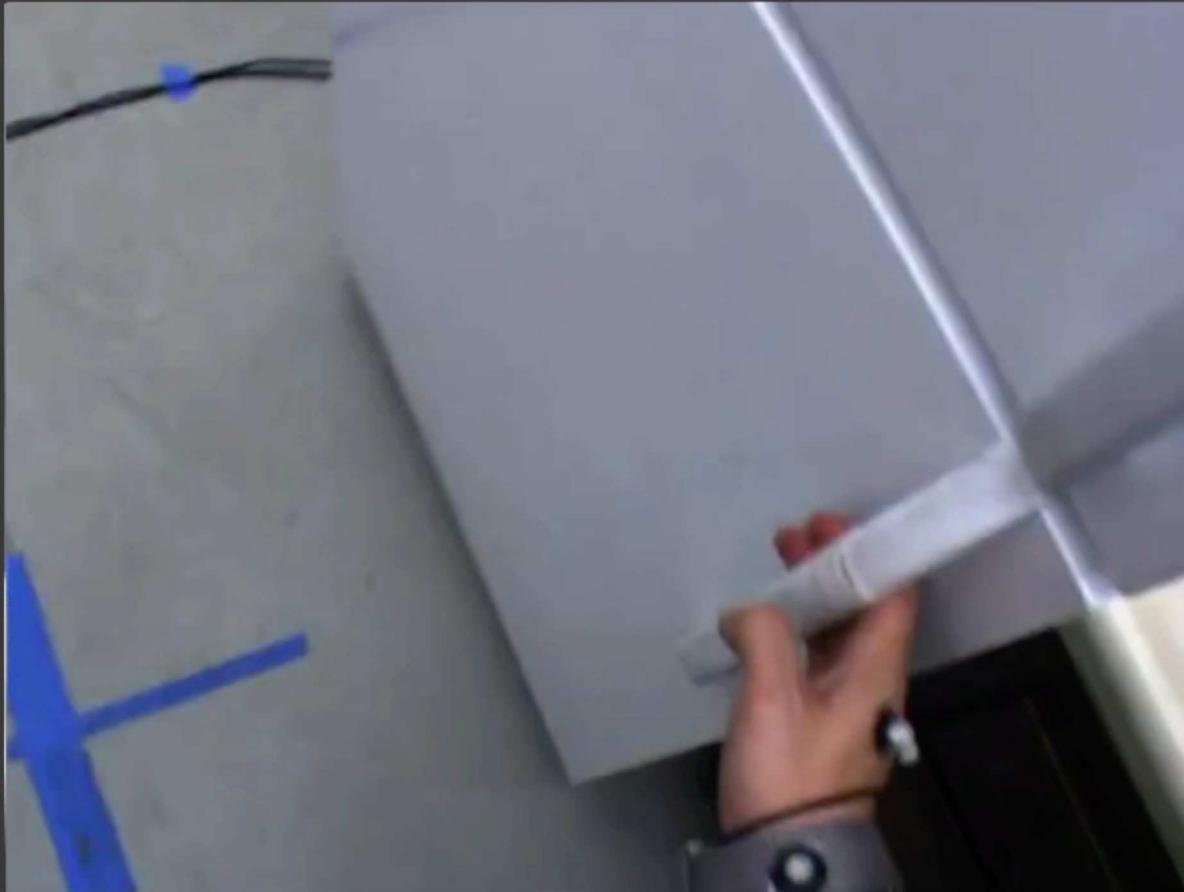
Towards an Unequivocal Representation of Actions

with: Michael Wray
Davide Moltisanti

- Think of an “open” action...

Towards an Unequivocal Representation of Actions

with: Michael Wray
Davide Moltisanti



Object Interactions – the Dilemma

Open

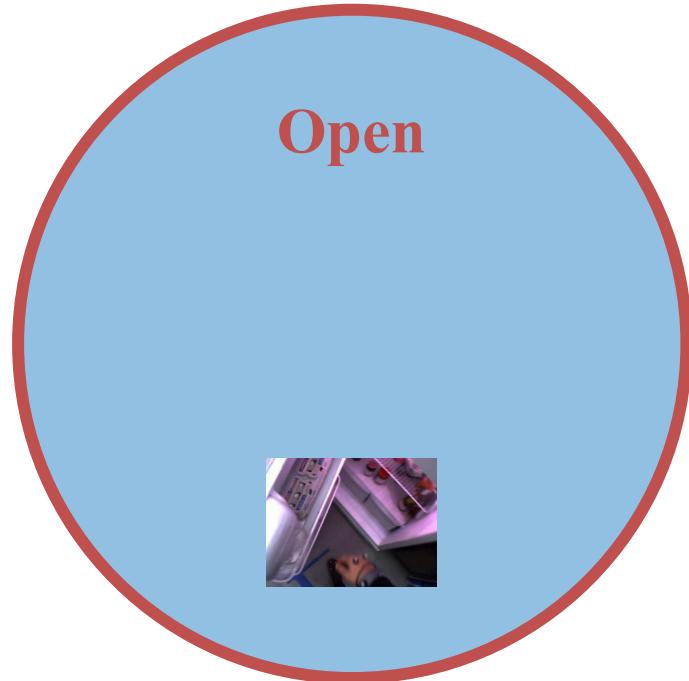


with: Michael Wray
Davide Moltisanti
Walterio Mayol-Cuevas

Object Interactions – the Dilemma



Object Interactions – the Dilemma

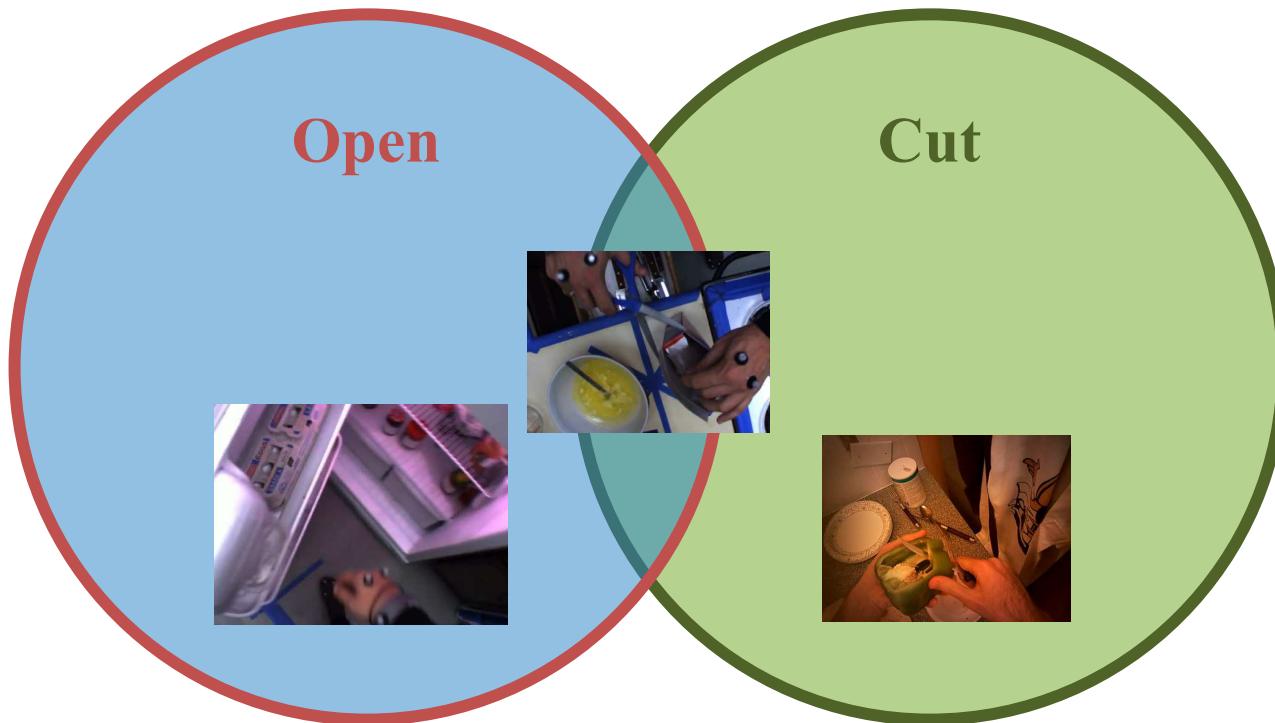


with: Michael Wray
Davide Moltisanti
Walterio Mayol-Cuevas

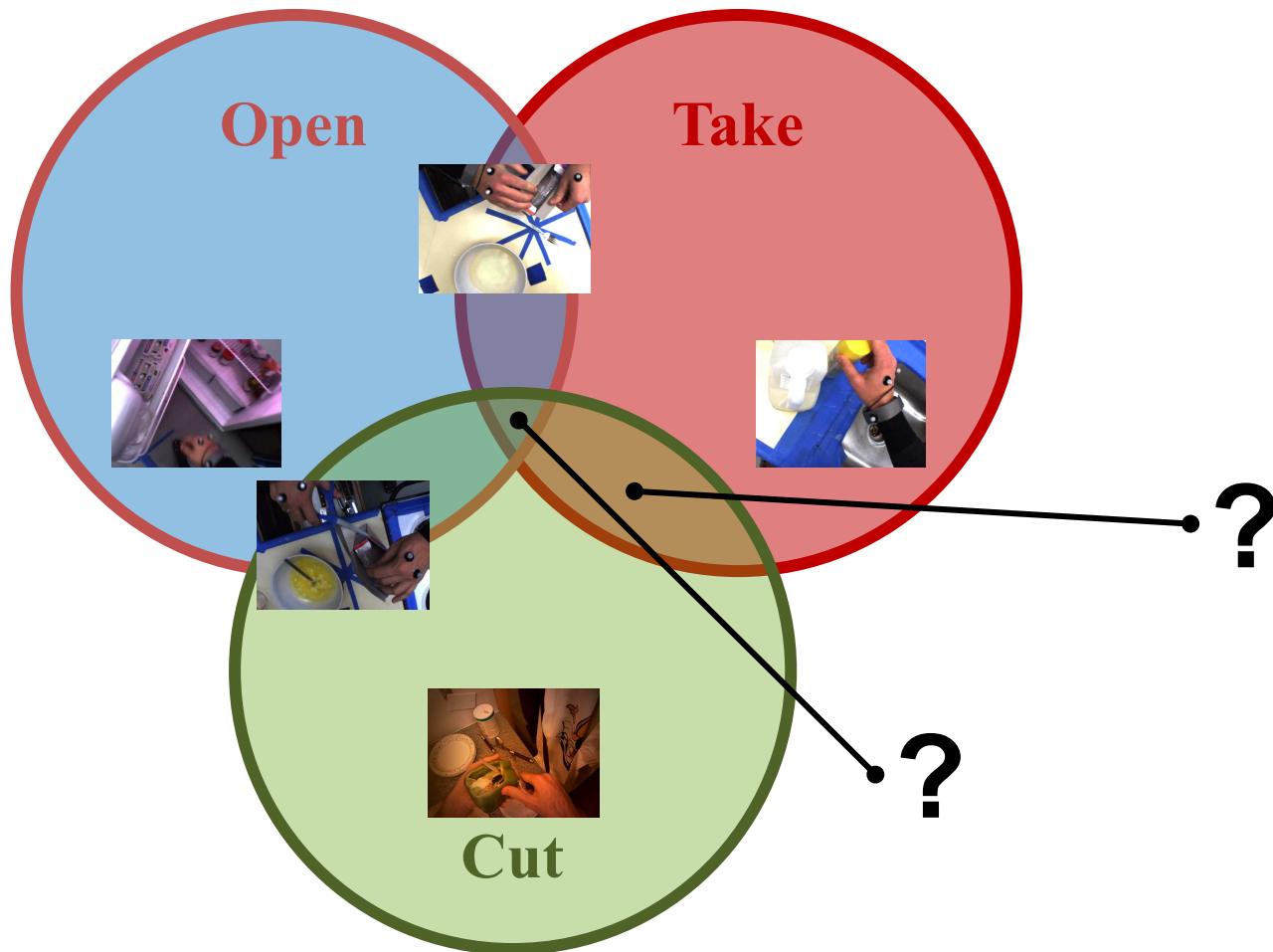
Object Interactions – the Dilemma



Object Interactions – the Dilemma



Object Interactions – the Dilemma



Towards an Unequivocal Representation of Actions

with: Michael Wray
Davide Moltisanti

- Action representations using a single verb is highly-ambiguous
 - Solution1: pre-selected non-overlapping verbs (SL)
 - run, walk, open, close
 - Solution2: Using nouns to disambiguate actions (V-N)
 - open-drawer, open-bottle, open-fridge
 - actions constrained to known nouns
 - Solution3: Multi-verb labels (ML, SAML)
 - open, hold, pull
 - How many verbs would be enough?

Towards an Unequivocal Representation of Actions

with: Michael Wray
Davide Moltisanti

- Soft-Assigned Multi-Label
 - Multi-label using verbs only
 - Each verb assigned a value between 0 and 1
 - Object agnostic
 - Trained with Sigmoid Binary Cross Entropy

Towards an Unequivocal Representation of Actions

with: Michael Wray
Davide Moltisanti

- Collected from AMT



- Annotators agree:
 - Relevant Verb -> Main action
 - Irrelevant Verb -> unrelated motion
- Annotators disagree:
 - Relevant motion but not the main action

Towards an Unequivocal Representation of Actions

with: Michael Wray
Davide Moltisanti

- Collected from AMT



SL

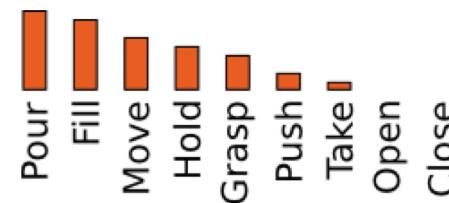
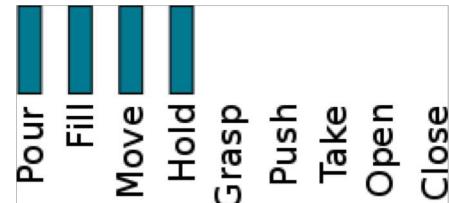
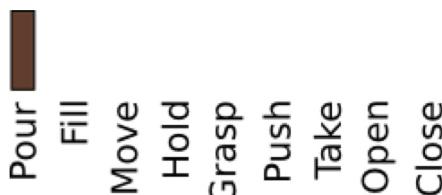
- Majority Vote.
- One-hot vector.

ML

- Threshold of 0.5.
- Binary Vector

SAML

- Full Annotation.
- Continuous Vector.



Towards an Unequivocal Representation of Actions

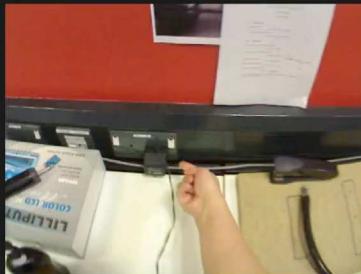
with: Michael Wray
Davide Moltisanti

Top 3 retrieved classes across all datasets.

Turn On/Off
Press
Rotate



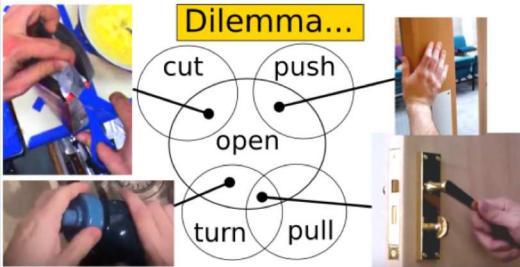
Turn On/Off
Press
Rotate



Labelling Method can differentiate turn On/Off tap by pressing and by rotating.

More info...

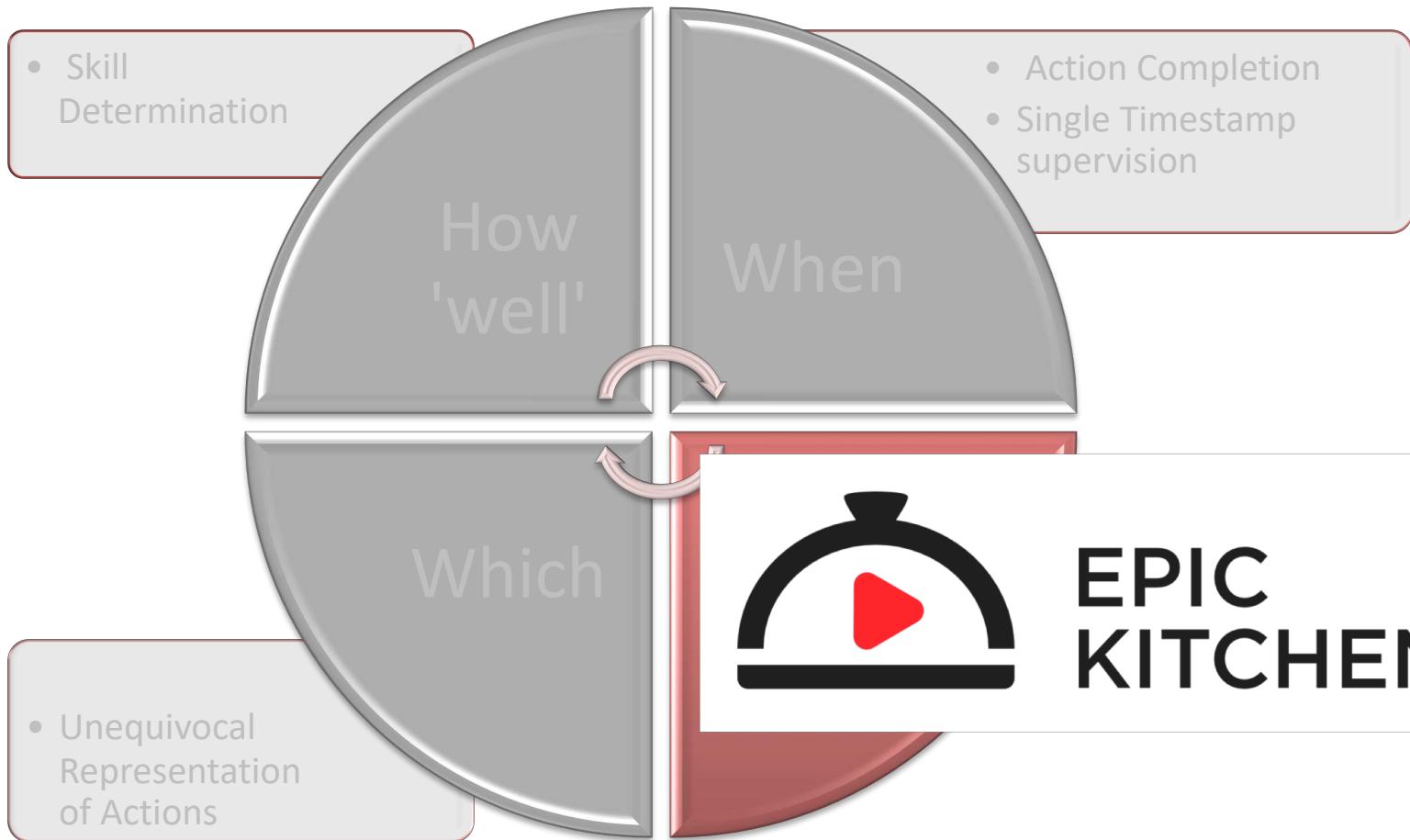
Project Towards an Unequivocal Representation of Actions



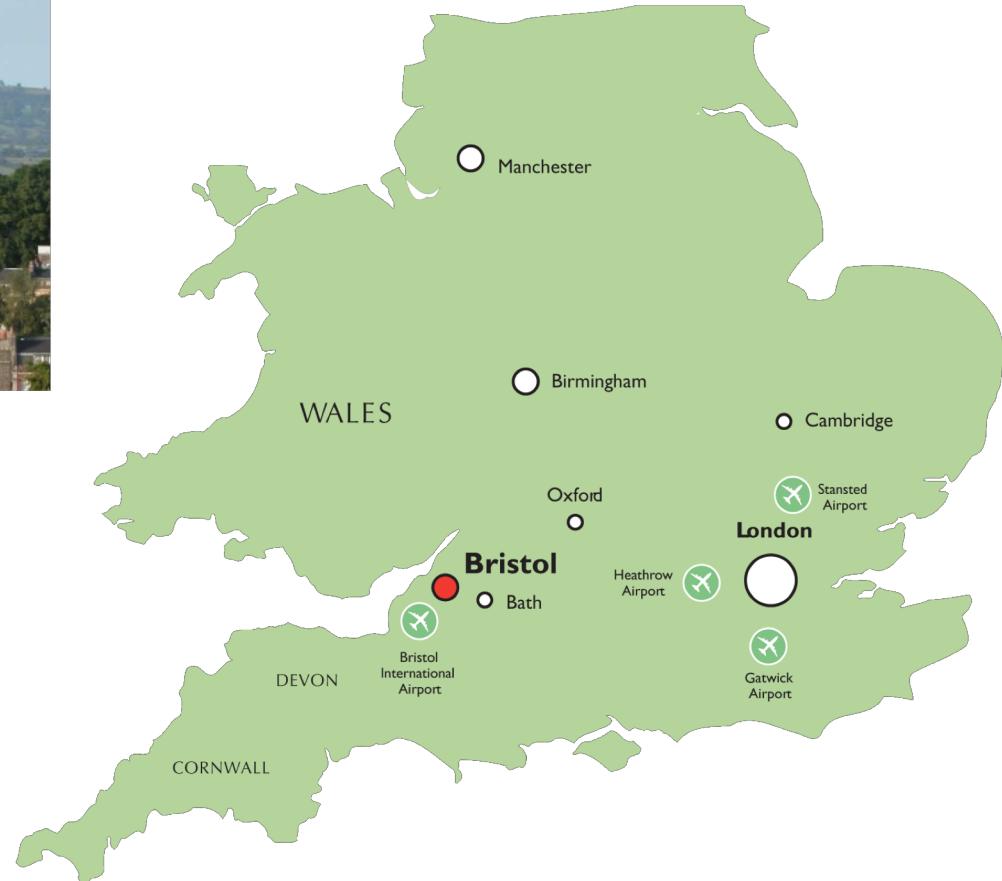
Towards an Unequivocal Representation of Actions. M Wray, D Moltisanti, D Damen. [ArXiv](#), 2018.

Improving Classification by Improving Labelling: Introducing Probabilistic Multi-Label Object Interaction Recognition. M Wray, D Moltisanti, W Mayol-Cuevas, D Damen. ArXiv, 2017. [arxiv](#)

Fine-Grained Object Interactions



Bristol and University of Bristol



Bristol and University of Bristol



Thank you...

For further info, datasets, code, publications...

<http://dimadamen.github.io>



@dimadamen



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

Join epic-community@bristol.ac.uk

- send an email to: sympa@sympa.bristol.ac.uk
- with the subject: subscribe epic-community
- and blank message content



Scaling Egocentric Vision: The **EPIC-KITCHENS** Dataset



Dima Damen



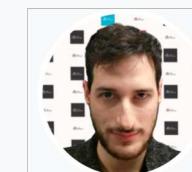
Hazel Doughty



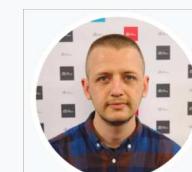
Giovanni M. Farinella



Sanja Fidler



Antonino Furnari



Evangelos Kazakos



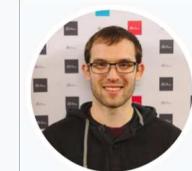
Davide Moltisanti



Jonathan Munro



Toby Perrett



Will Price



Michael Wray



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]	✗





EPIC
KITCHENS

Scaling Egocentric Vision

CodaLab
Competition

EPIC-Kitchens Object Detection
Secret url: <https://competitions.codalab.o>
Organized by hazel dougherty - Current server time: 5:45:00 UTC

▶ Current
ECCV 2018 Object Recognition Challenge
June 30, 2018, midnight UTC

Learn the Details Phases Participate Results



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



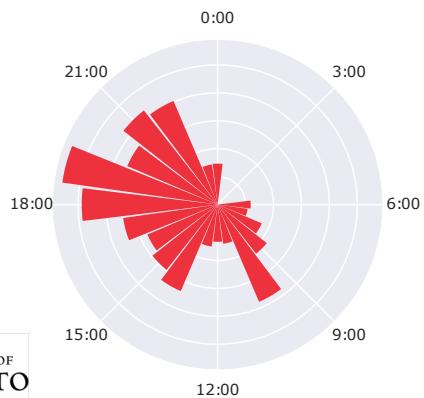


Data Collection

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



32
KITCHENS





Annotations (1) - Narrations

Narrations

06.00 - close dishwasher
06.03 - fold kitchen roll
06.05 - pick up tofu
06.09 - put on kitchen roll
06.19 - squeeze kitchen roll onto tofu

06.37 - pat dry
06.42 - pick up kitchen roll
06.45 - put in bin
06.51 - put tofu in bag
06.53 - pick up miso paste
07.00 - open miso paste
07.03 - spoon miso paste
07.10 - put in bag

07.28 - close miso paste
07.36 - open dishwasher
07.43 - squeeze content of bag
07.46 - let out air
07.48 - squeeze content of bag
08.03 - squeeze air out of bag

Narrations

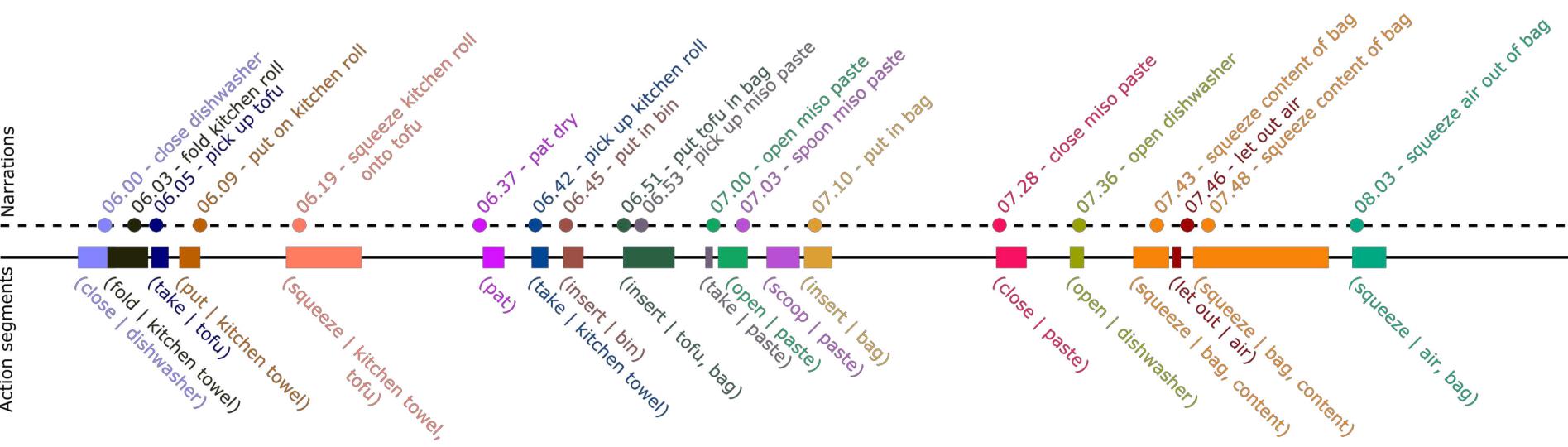
03.42 - apri armadietto
03.44 - prendi sacchetto
03.45 - prendi cipolla
03.47 - chiudi armadietto
03.49 - taglia cipolla
03.53 - accendi luce

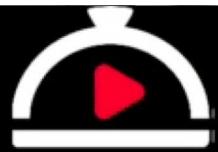
03.59 - taglia cipolla
03.59 - sbuccia cipolla





Annotations (2) – Action Segments



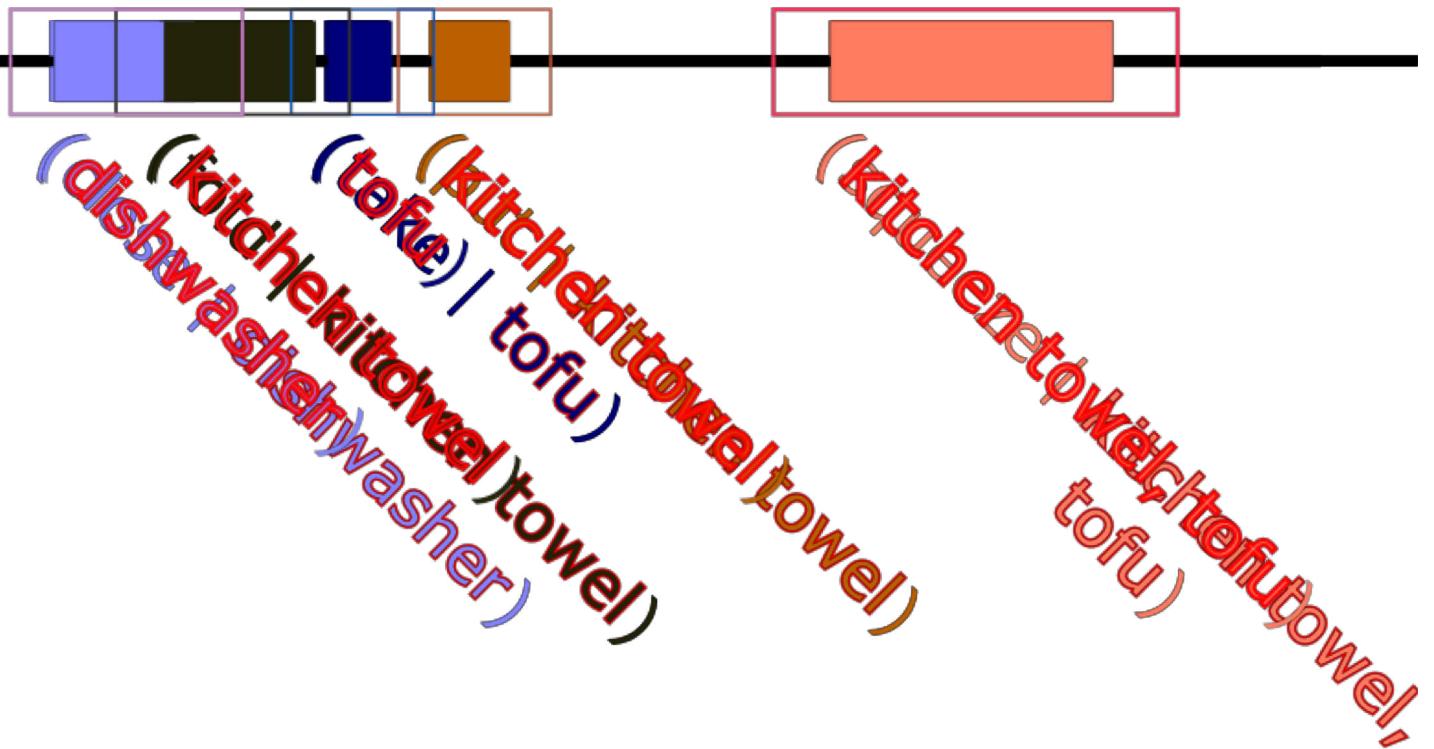


39 000
ACTION SEGMENTS



Annotations (3) – Object Bounding Boxes

Action segments





454 200
OBJECT ANNOTATIONS



Annotations (4) – Verb and Noun Classes

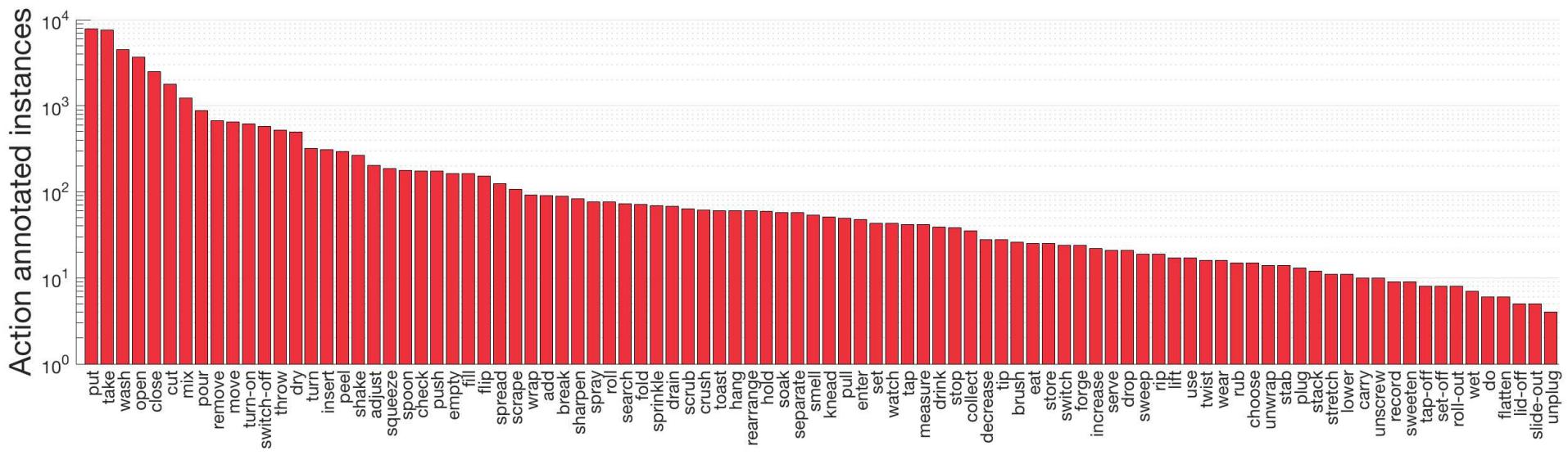
|take, grab, pick, get, fetch, pick-up, ...

- 120 verb classes
- 331 noun classes



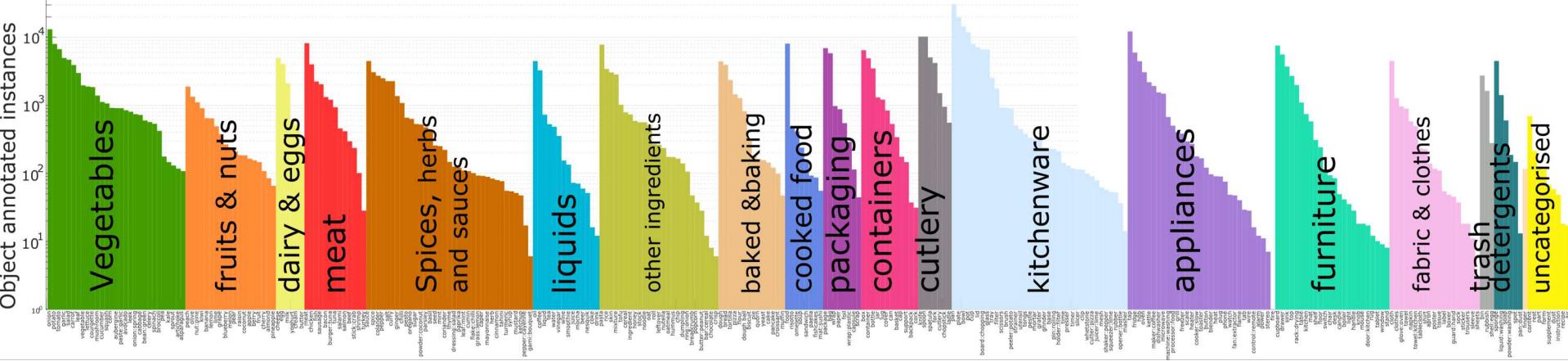
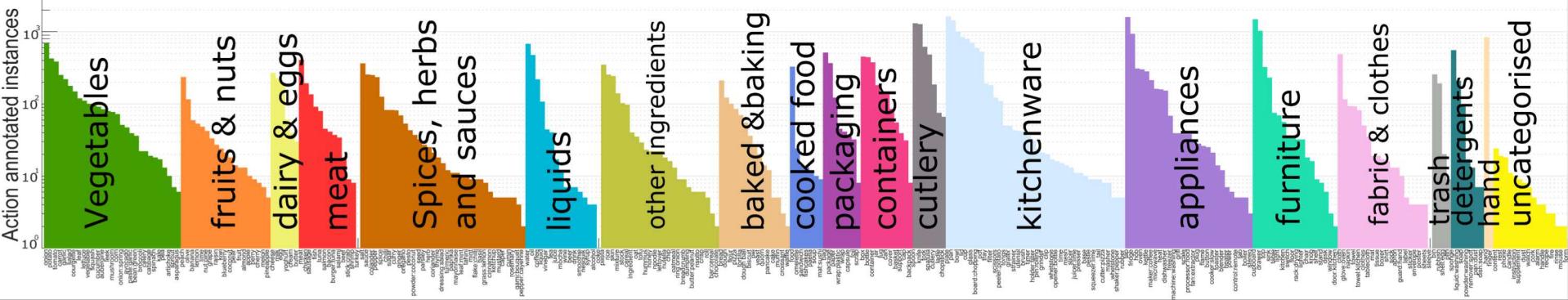


Annotations Statistics





Annotations Statistics





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



TVL₁ optical flow (u , v) frames:

- 456x256 px
- 30FPS





Open Challenges

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





Open Challenges

- Challenges open on CodaLab – 9 Sep
- First Challenge Results in CVPR 2019
- EPIC@CVPR2019
- ActivityNet@CVPR2019

The screenshot shows the CodaLab interface for the EPIC-Kitchens Action Recognition challenge. At the top, there's a navigation bar with 'My Competitions' and 'Help' links, and a user profile for 'willprice'. Below the header, the title 'Competition' is displayed above a section titled 'Admin features' with tabs for 'Edit', 'Participants', 'Submissions', 'Dumps', and 'Widgets'. A large image of a kitchen scene serves as the background for the competition area. In the center, there's a logo of a dome with a play button inside. To its right, the competition name 'EPIC-Kitchens Action Recognition' is shown, along with a 'Secret url' link and the current server time. Below this, a status bar indicates 'Current' (ECCV 2018 Action Recognition Challenge) and 'End' (Competition Ends), with specific dates: June 30, 2018, midnight UTC and Oct. 10, 2018, midnight UTC. At the bottom of the main content area, there are tabs for 'Learn the Details', 'Phases', 'Participate', 'Results', 'Forums', and 'Team'. The 'Learn the Details' tab is active, showing sections for 'Overview', 'Evaluation', 'Terms and Conditions', and 'Submission Format'. The 'Overview' section contains a brief welcome message about the dataset and its relation to the ECCV 2018 workshop. The 'Dataset details' section lists three bullet points: '55 hours of video', '11.5M frames', and '39,594 total action segments'. At the very bottom of the page, there are links for 'Join us on Github for contact & bug reports', 'About', 'Privacy and Terms', and 'v1.5'.





More?

<http://epic-kitchens.github.io>



NEWS

- EPIC-KITCHENS accepted for oral presentation at ECCV 2018 in Munich this September
- News coverage: [IoB](#), [The Spoon](#), [Il Sole 24 Ore](#), [La Sicilia](#), [Elpais](#)
- EPIC-Kitchens Released: 9th of April 2018!!!
- Watch [YouTube Release Trailer here](#)

What is EPIC-Kitchens?

The largest dataset in first-person (egocentric) vision; multi-faceted non-scripted recordings in native environments - i.e. the wearers' homes, capturing all daily activities in the kitchen over multiple days. Annotations are collected using a novel 'live' audio commentary approach.

Characteristics

- 32 kitchens - 4 cities
- Head-mounted camera
- 55 hours of recording - Full HD, 60fps
- 11.5M frames
- Multi-language narrations
- 39,594 action segments
- 454,158 object bounding boxes
- 125 verb classes, 352 noun classes

Updates

Stay tuned with updates on epic-kitchens2018, as well as EPIC workshop series by joining the [epic-community mailing list](#) send an email to: sympa@sympa.bristol.ac.uk with the subject *subscribe epic-community* and a *blank* message body.

