UNIVERSITY OF SURREY

Anastasia Anichenko, Frank Guerin, Andrew Gilbert

Interpretable Action Recognition on Hard to Classify Actions

The problem:

- ❑ We aimed to recognise activities in an interpretable way, by tracking positions of hands and principal objects
- ❑ We found particularly low performance for some categories due to the loss of other features (e.g. “Putting”, see right)

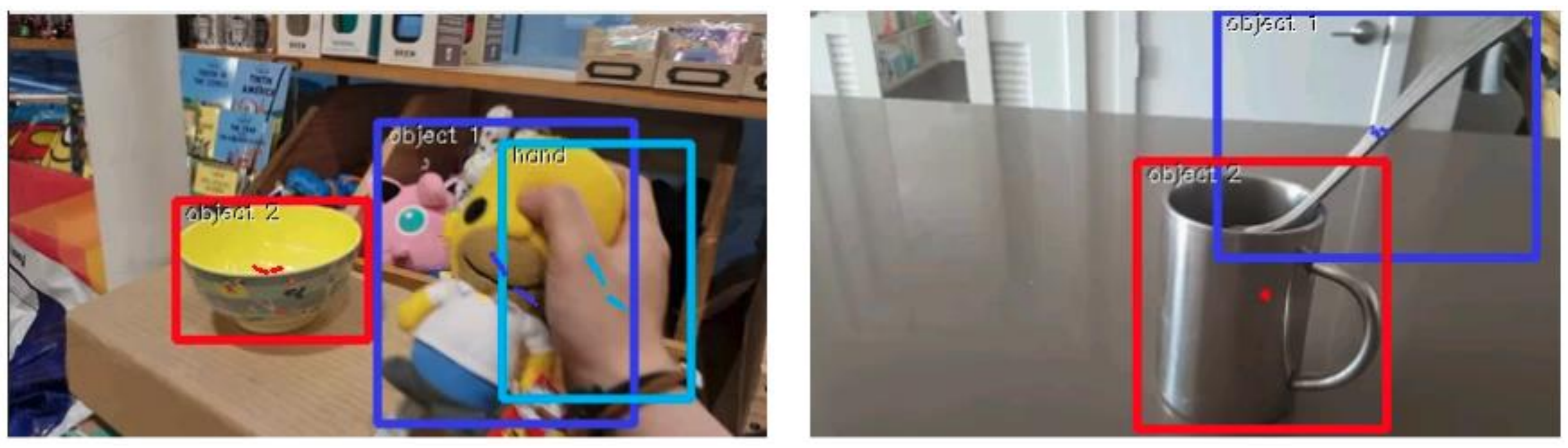
The proposed solution:

- ❑ Add object shape information: object detection model was fine-tuned to differentiate “Container” and “NotContainer”
- ❑ Add depth information: depth estimation model extracts depth for individual objects, add depth to our interpretable model

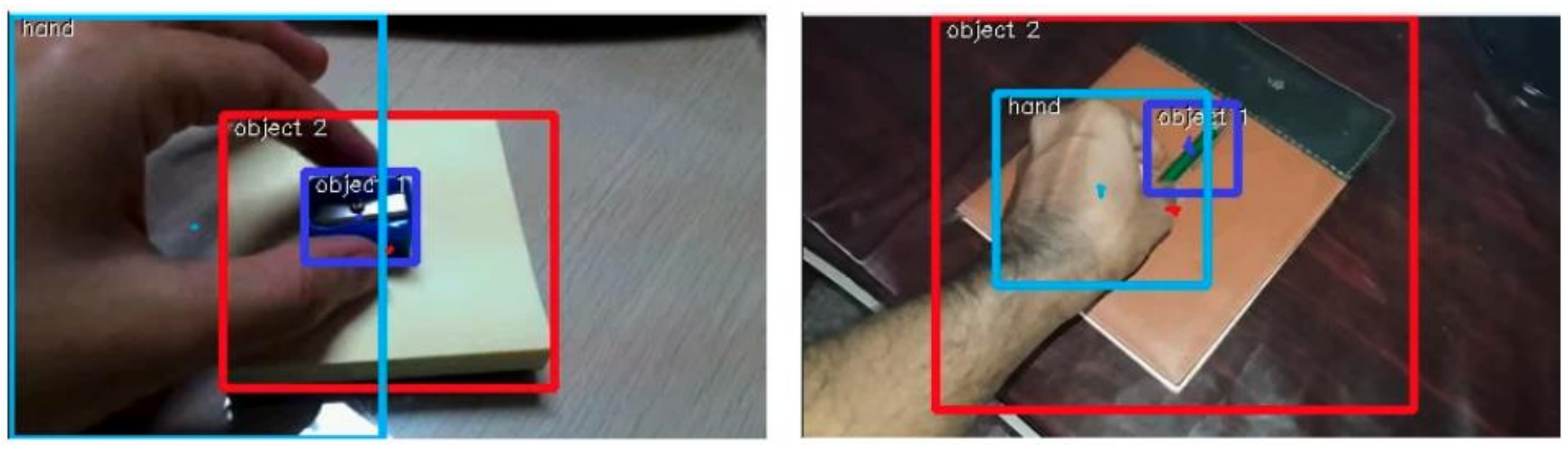
The result:

- ❑ Object shape information did not help much (it is very hard to make a generic container recognizer)
- ❑ Depth information made a significant improvement (reasonable quality depth information is easy to obtain)

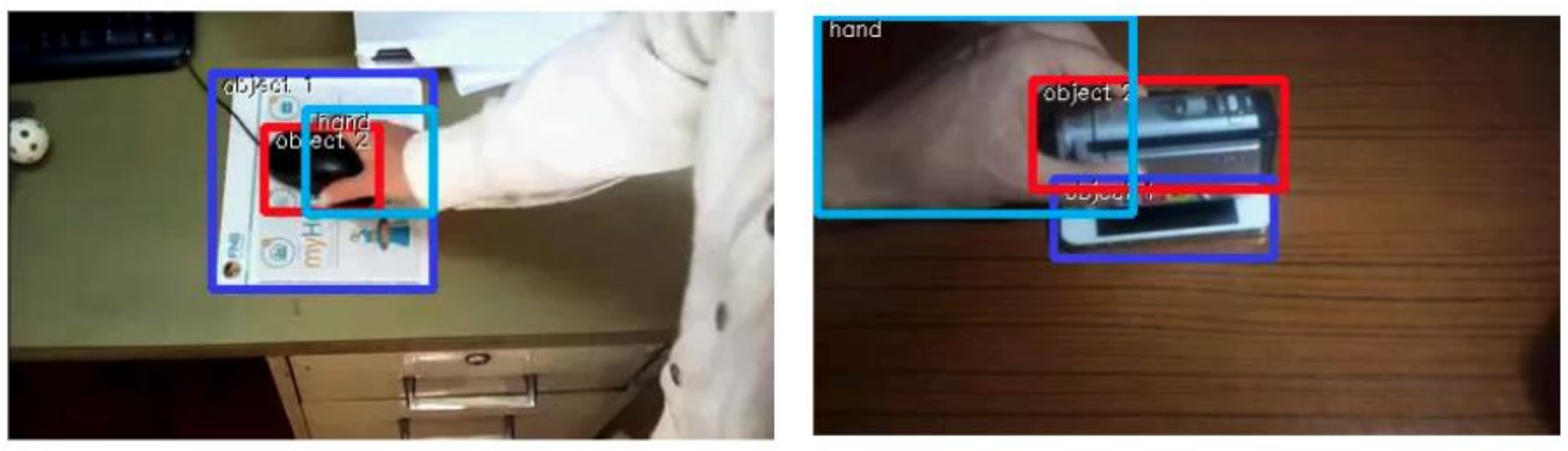
Hard to classify “Putting” activities from Something-Something V2



“Putting something into something"



"Putting something onto something"



"Putting something underneath something"

Our Interpretable Model

We first temporally segment the video into five ‘phases’ based on features that characterize that phase

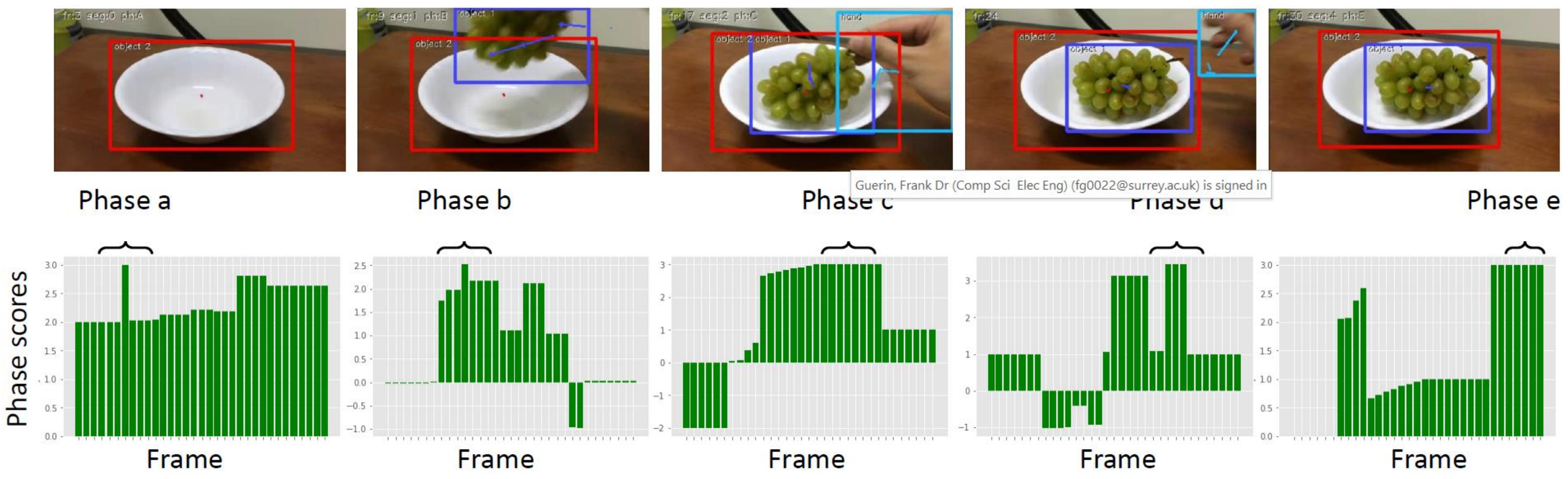
The object(s) is present in the scene, the manipulation has not happened;

The hand enters (possibly carrying an object);

The critical manipulation happens (e.g. object placed or picked);

The hand departs (possibly carrying an object);

The objects are present, with the result of the manipulation evident.



Once phases are assigned, we compute **feature vectors** characterising each phase

Feature vectors include relations among bounding boxes of the two principal objects and the hands, for example,

- ❑ Object size,
- ❑ Object movement since previous frame,
- ❑ Relative movement between two objects,
- ❑ Object moving with the hand,
- ❑ Object moving relative to the hand,
- ❑ etc.

We train a random forest classifier for each activity class

When doing multi-class classification, the highest probability random forest prediction is returned as the class.

Scan to see more details of our interpretable model

Results and Conclusions

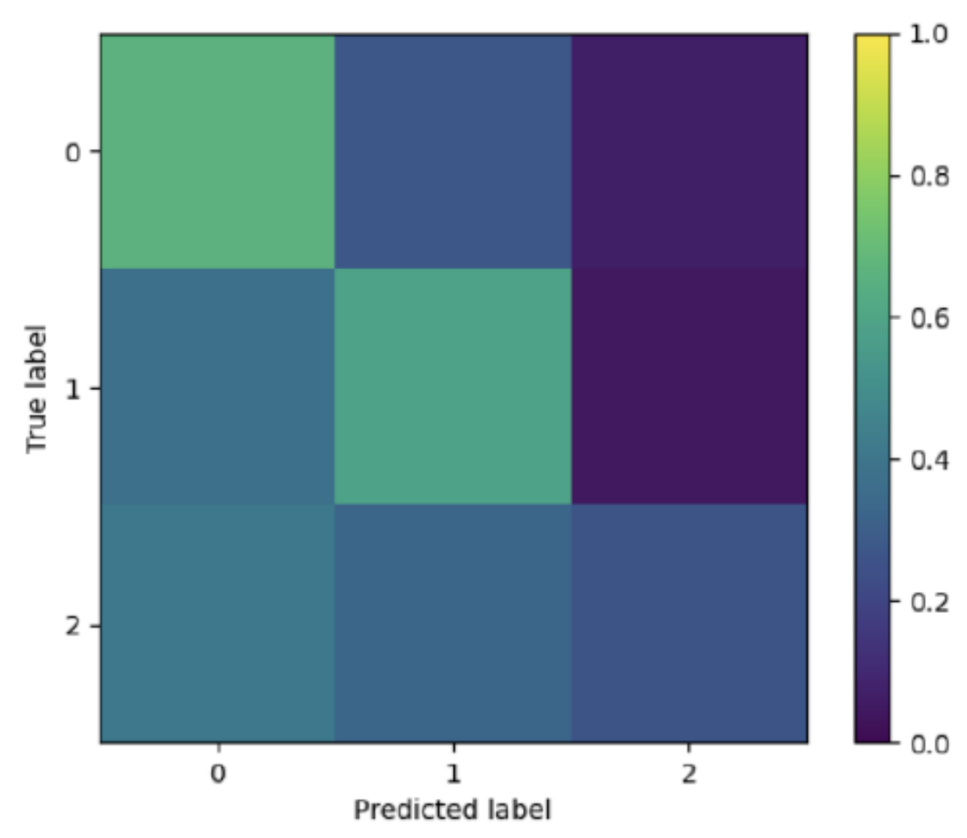
“Putting something into something"

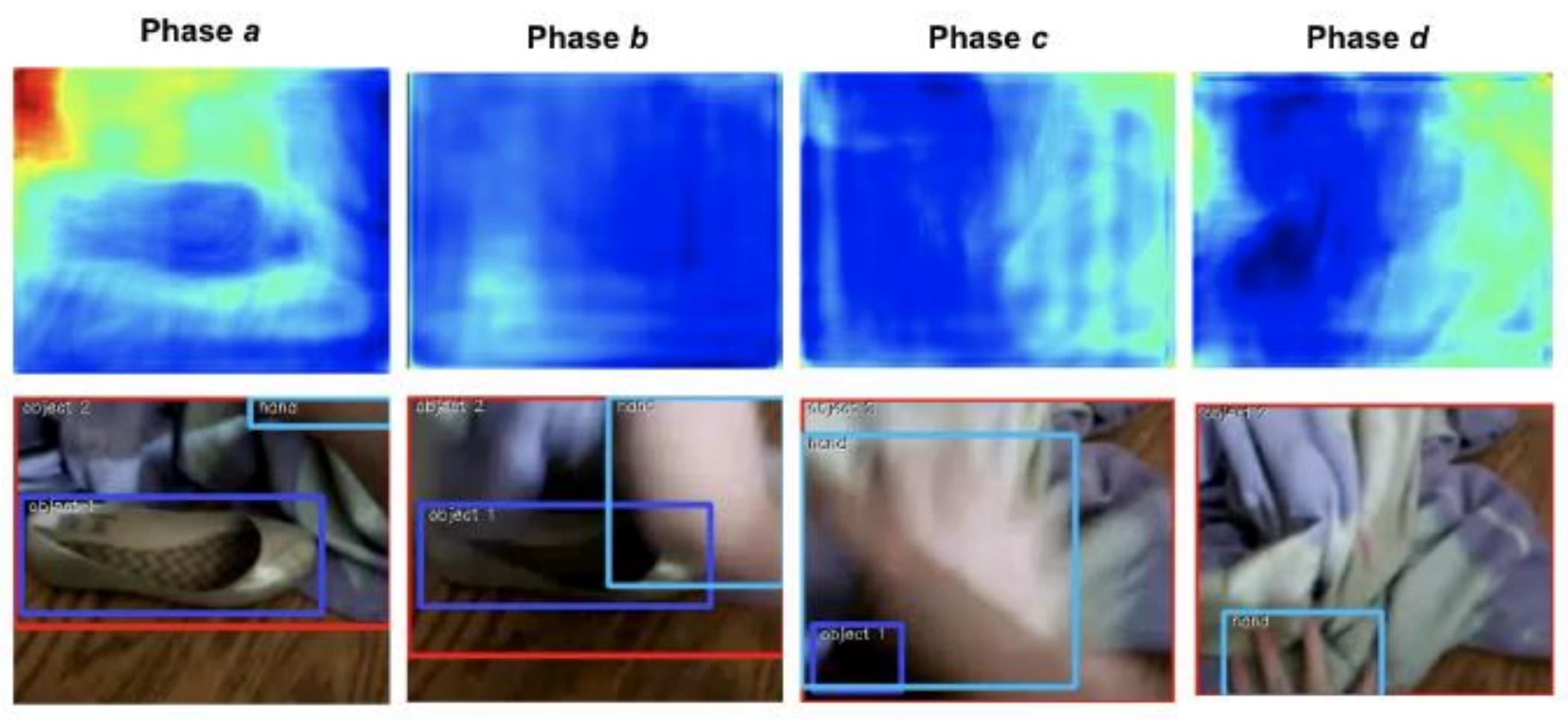
"Putting something onto something"

"Putting something underneath something"

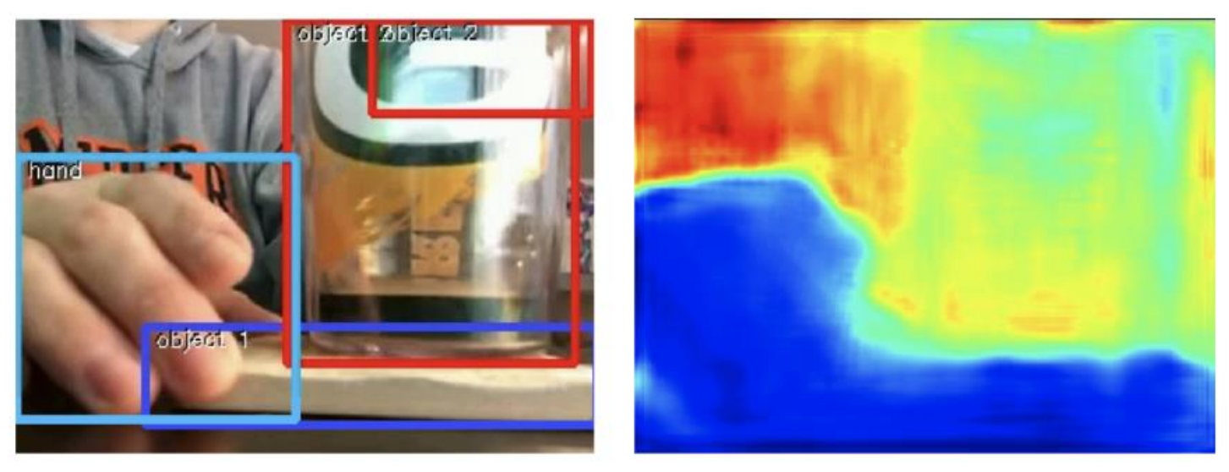
Metric	Precision				Recall			
SSV2 class	106	112	118	avg	106	112	118	avg
TDM (Ours) baseline [2]	0.69	0.47	0.29	0.48	0.63	0.59	0.24	0.49
3D CNN [3]	0.61	0.36	0.00	0.32	0.84	0.22	0.00	0.36
VideoMAE [5]	**0.89**	**0.71**	**0.71**	**0.77**	**0.86**	**0.80**	**0.60**	**0.76**
Ours + initial improvements	0.72	0.45	0.32	0.49	0.62	0.68	0.14	0.48
Ours + container detection	0.68	0.41	0.32	0.47	0.57	0.54	0.34	0.48
Ours + depth relations	0.72	0.45	0.37	0.51	0.66	0.59	0.26	0.50
Ours + container detection + depth relations	0.66	0.46	0.30	0.47	0.71	0.42	0.24	0.46

Table 1: Precision and recall rates for each model tested on the validation subset. Macro average is used instead of weighted average. The highest result is highlighted in black whilst the second highest is highlighted in blue.





Here we see that the depth maps are pretty useless. This is because these are unusual “in the wild” frames, which are probably not close to the samples the monocular depth estimator was trained on. This is a challenging example of “putting something underneath something” that got misclassified as “putting something into something”.



In this example “putting something underneath something” was done in side-view, so the original 2D features are actually more useful than the depth map. The depth map hinders recognition here.

Funded By LEVERHULME TRUST