Beyond Action Recognition: Action Completion in RGB-D Data

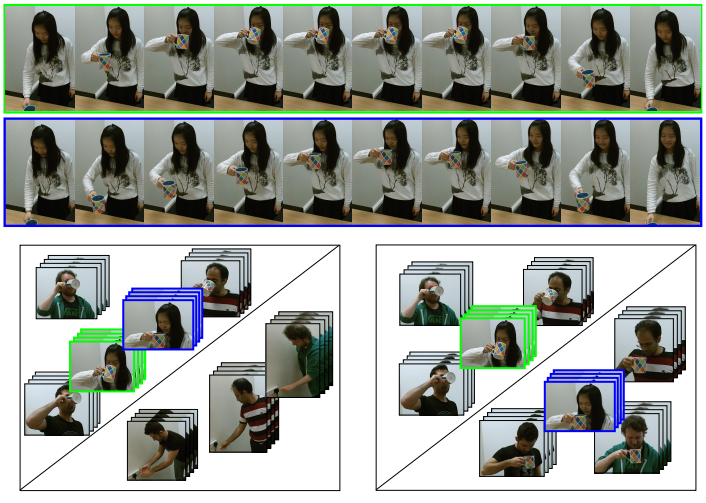
Farnoosh Heidarivincheh
farnoosh.heidarivincheh@bristol.ac.uk
Majid Mirmehdi
majid@cs.bris.ac.uk
Dima Damen
dima.damen@bristol.ac.uk

Department of Computer Science
University of Bristol
Bristol, UK

Robust motion representations for action recognition have achieved remarkable performance in both controlled and 'in-the-wild' scenarios. Such representations are primarily assessed for their ability to label a sequence according to some predefined action classes (e.g. *walk, wave, open*). Although increasingly accurate, these classifiers are likely to label a sequence, even if the action has not been fully completed, because the motion observed is similar enough to the training set. Consider the case where one attempts to drink but realises the beverage is too hot. A *drinking-vs-all* classifier is likely to recognise this action as *drinking* regardless. We introduce the term **action completion** as a step beyond the task of action recognition. It aims to recognise whether the action's goal has been successfully achieved. The notion of completion differs per action and could be infeasible to verify using a visual sensor, however, for many actions, an observer would be able to make the distinction by noticing subtle differences in motion.

We address incompletion in a supervised approach, using a new dataset that contains 414 complete as well as incomplete sequences, captured using a depth sensor, spanning 6 actions (*switch, plug, open, pull, pick* and *drink*). For each action, we varied the conditions so the action cannot be completed. For example, for *plug*, subjects were given a plug that does not match the socket, while for *pull*, a drawer was locked so could not be pulled, and similarly for the other actions. Given labelled *complete* and *incomplete* sequences of the same action, we build a model of completion of that action as a binary classifier for each of our actions.

Since the notion of completion differs per action, a general action completion method should investigate the performance of different types of features to accommodate the various action classes. For example, for the action *pick*, the difference between complete and incomplete actions originates from the subtle change in body pose when holding an object, or by observing an object in the hand. On the other hand, for the action *drink*, the speed at which the action is performed is better able to assess the completion. We propose a method that chooses the feature(s) suitable for recognising completion from a pool of depth features using 'leave-one-person-out' cross validation on the training set and automatically selecting the most discriminative feature(s).



(a) *drink* vs. *plug* (b) complete vs. incomplete (of *drink*)

Figure 1: For a complete *drink* (green) and an incomplete *drink* (blue) sequences from our dataset, both are classified as *drink* when using *drink* vs. *plug* classifier (a). The proposed supervised action completion model (b) identifies the incomplete sequence.

We present results on a pool of five features: Local Occupancy Pattern (LOP), Joint Positions (JP), Joint Relative Positions (JRP), Joint Relative Angles (JRA) and Joint Velocities (JV) encoded by the Fourier temporal pyramid [1]. On a sequence of experiments, we show that:

1. Complete Action Recognition - The various depth features produce high and comparable % accuracy for action recognition on our dataset.

2. Incomplete Action Recognition - These features, originally designed for action recognition, behave differently on *incomplete* action sequences with only some able to distinguish the subtle changes between *complete* and *incomplete* sequences of an action.

3. Complete vs. Incomplete Action Recognition - A binary classification was performed as complete vs. incomplete of the same action for each feature. Table 1 shows varying success rates of the different features for the tested actions.

4. Automatic Feature Selection - Using cross validation on training data, the features with the maximum accuracy were selected to build the completion model. By automatic feature selection, we achieve 95.7% accuracy for recognising action completion across the whole dataset - Table 2.

[1] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *CVPR*, 2012.

	LOP	JP	JRP	JRA	JV
switch	**100**	85.1	85.1	**100**	**100**
plug	83.6	87.7	78.1	79.5	**94.5**
open	**97.1**	95.6	**97.1**	95.6	**97.1**
pull	87.3	71.8	77.5	88.7	**94.4**
pick	92.8	94.2	**98.6**	**98.6**	95.7
drink	97	97	97	97	**100**

Table 1: Complete vs. incomplete action results.

	Subjects								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**total**
switch	100	100	100	100	100	100	100	100	100
plug	83.3	100	87.5	100	88.9	100	100	100	94.5
open	100	85.7	100	100	100	87.5	90	100	95.6
pull	88.9	100	100	100	100	87.5	80	100	94.4
pick	90	100	100	100	100	100	50	100	92.8
drink	77.8	100	100	100	100	100	100	100	97
								total	95.7

Table 2: General action completion results

Figure 2: Sample frames of correctly (a), (b) and incorrectly (c) classified test sequences. In (c), using JV solely, the hand seems to perform a *pull* in full even when the drawer remains unmoved.