Watch-n-Patch: Unsupervised Understanding of Actions and Relations

Chenxia Wu[1,2], Jiemi Zhang[1], Silvio Savarese[2], Ashutosh Saxena[1,2]
[1]Department of Computer Science, Cornell University. [2]Department of Computer Science, Stanford University.

We consider modeling human activities containing a sequence of actions (see an example in Fig. 1), as perceived by an RGB-D sensor in home and office environments. In the complex human activity such as *warming milk* in the example, there are not only short-range action relations, *e.g.*, *microwaving* is often followed by *fetch-bowl-from-oven*, but there are also long-range action relations, *e.g.*, *fetch-milk-from-fridge* is strongly related to *put-milk-back-to-fridge* even though other actions occur between them.

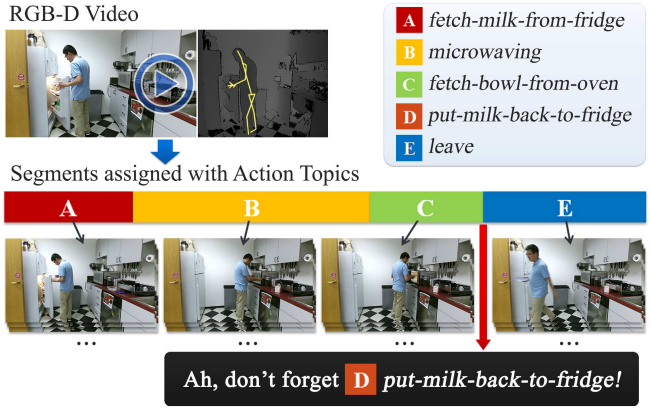


Figure 1: Our goal is to automatically segment RGB-D videos and assign action-topics to each segment. We propose a completely unsupervised approach to modeling the human skeleton and RGB-D features to actions, as well as the pairwise action co-occurrence and temporal relations. We then show that our model can be used to detect which action people forgot, a new application which we call *action patching*.

The challenge that we undertake in this paper is: Can an algorithm learn about the aforementioned relations in the activities when just given a completely *unlabeled* set of RGB-D videos?

Different from previous approaches, we consider a completely unsupervised setting. The novelty of our approach is the ability to model the long-range action relations in the temporal sequence, by considering pairwise action co-occurrence and temporal relations, *e.g.*, *put-milk-back-to-fridge* often co-occurs with and temporally after *fetch-milk-from-fridge*. We also use the more informative human skeleton and RGB-D features, which show higher performance over RGB only features for action recognition [2, 3].

In order to capture the rich structure in the activity, we draw strong parallels with the work done on document modeling from natural language (e.g., [1]). Fig. 2 illustrates our approach pipeline and Fig. 3 gives the graphic model of our learning approach. We consider an activity video as a document, which consists of a sequence of short-term action clips as *action-words* (the numbers in Fig. 2 and w_{nd} in Fig. 3). And an activity is about a set of *action-topics* indicating which actions are present in the video (the capital letters in Fig. 2 and z_{nd} in Fig. 3), such as *fetch-milk-from-fridge* in the *warming milk* activity. Action-words are drawn from these action-topics and has a distribution for each topic (green line in Fig. 2 and ϕ_k in Fig. 3). Then we model the following:

- *Action co-occurrence.* Some actions often co-occur in the same activity. We model the co-occurrence by adding correlated topic priors to the occurrence of action-topics, *e.g.*, *fetch-milk-from-fridge* and *put-milk-back-to-fridge* has strong correlations. (black line in Fig. 2 and μ, Σ in Fig. 3.)
- *Action temporal relations.* Some actions often causally follow each other, and actions change over time during the activity execution. We model the relative time distributions between every action-topic pair to capture the temporal relations. (red line in Fig. 2 and θ_{kl} in Fig. 3.)

We first show that our model is able to learn meaningful representations from the unlabeled activity videos. We use the model to temporally segment

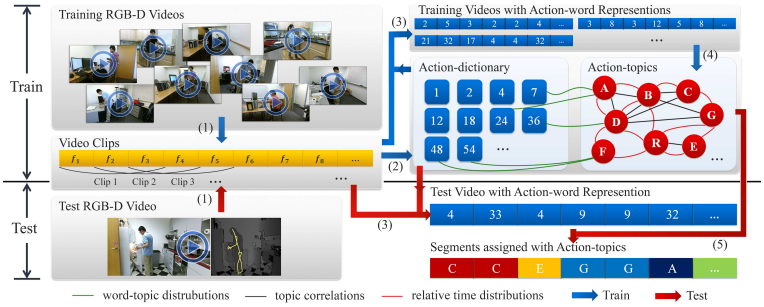
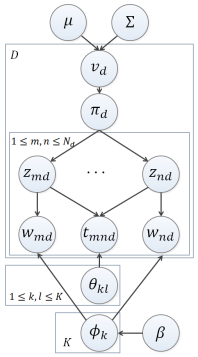


Figure 2: **The pipeline of our approach.** Training (blue arrows) follows steps (1), (2), (3), (4). Testing (red arrows) follows steps (1), (3), (5). The steps are: (1) Decompose the video into a sequence of overlapping fixed-length temporal clips. (2) Learn the action-dictionary by clustering the clips, where the cluster centers are action-words. (3) Map the clips to the action-words in the action-dictionary to get the action-word representation of the video. (4) Learn the model from the action-word representations of training videos. (5) Assign action-words in the video with action-topics using the learned model.



Symbols	Meaning
D	number of videos in the training database;
K	number of action-topics;
N_d	number of words in a video;
w_{nd}	n-th word in d-th document;
z_{nd}	topic-word assignment of w_{nd};
t_{nd}	the normalized timestamp of of w_{nd};
t_{mnd}	$= t_{md} - t_{nd}$ the relative time between w_{md} and w_{nd};
$\pi_{:d}$	the probabilities of topics in d-th document;
$v_{:d}$	the priors of $\pi_{:d}$ in d-th document;
ϕ_k	the multinomial distribution of the word from topic k;
μ, Σ	the mutivariate normal distribution of $v_{:d}$;
θ_{kl}	the relative time distribution of t_{mnd}, between topic k, l;

Figure 3: The graphic model of our approach (Left). Notations in our model (Right).

videos to segments assigned with action-topics. We show that these action-topics are semantically meaningful by mapping them to ground-truth action classes and evaluating the labeling performance.

We then also show that our model can be used to detect forgotten actions in the activity, a new application that we call *action patching*. We show that the learned co-occurrence and temporal relations are very helpful to infer the forgotten actions by evaluating the patching accuracy.

We also provide a new challenging RGB-D activity video dataset [1] recorded by the new Kinect v2, in which the human skeletons and the audio are also recorded. It contains 458 videos of human daily activities as compositions of multiple actions interacted with different objects, in which people forget actions in 222 videos. They are performed by different subjects in different environments with complex backgrounds.

In summary, the main contributions of this work are:

- Our model is completely unsupervised and non-parametric, thus being more useful and scalable.
- Our model considers both the short-range and the long-range action relations, showing the effectiveness in the action segmentation and recognition, as well as in a new application action patching.
- We provide a new challenging RGB-D activity dataset recorded by the new Kinect v2, which contains videos of multiple actions interacted with different objects.

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.
[2] Hema Swetha Koppula and Ashutosh Saxena. Anticipating human activities using object affordances for reactive robotic response. In *RSS*, 2013.
[3] Di Wu and Ling Shao. Leveraging hierarchical parametric networks for skeletal joints based action segmentation and recognition. In *CVPR*, 2014.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.

[1]The dataset and tools are released at http://watchnpatch.cs.cornell.edu