Anticipating Human Activities for Reactive Robotic Response

Hema S. Koppula* and Ashutosh Saxena*

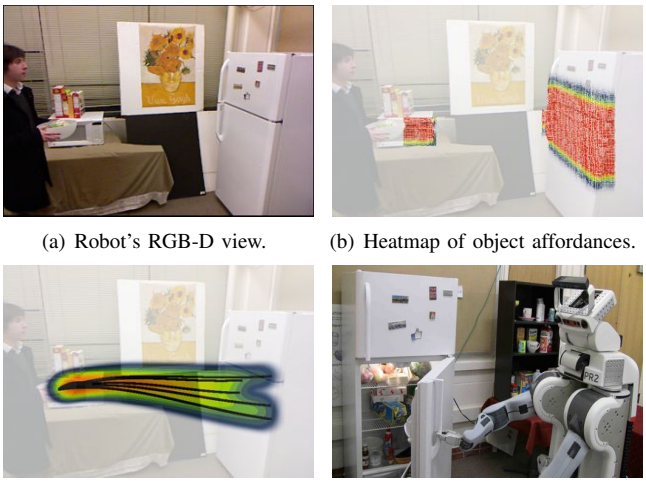
***Abstract*— An important aspect of human perception is anticipation, which we use extensively in our day-to-day activities when interacting with other humans as well as with our surroundings. Anticipating which activities will a human do next (and how to do them) in useful for many applications, for example, anticipation enables an assistive robot to plan ahead for reactive responses in the human environments. In this work, we represent each possible future using an anticipatory temporal conditional random field (ATCRF) that models the rich spatial-temporal relations through object affordances. We then consider each ATCRF as a particle and represent the distribution over the potential futures using a set of particles. In extensive evaluation on CAD-120 human activity RGB-D dataset, for new subjects (not seen in the training set), we obtain an activity anticipation accuracy (defined as whether one of top three predictions actually happened) of 75.4%, 69.2% and 58.1% for an anticipation time of 1, 3 and 10 seconds respectively. Finally, we also use our algorithm on a robot for performing a few reactive responses.** [1]

I. INTRODUCTION

For a personal robot to be able to assist humans, it is important for it to be able to detect what a human in currently doing as well as *anticipate* what she is going to do next and how. The former ability is useful for applications such as monitoring and surveillance, but we need the latter for applications that require reactive responses (e.g., see Figure 1). In this paper, our goal is to use anticipation for predicting future activities as well as improving detection (of past activities).

There has been a significant amount of work in detecting human activities from 2D RGB videos [2], [3], [4], from inertial/location sensors [5], and more recently from RGB-D videos [6], [7], [8]. The primary approach in these works is to first convert the input sensor stream into a spatio-temporal representation, and then to infer labels over the inputs. These works use different types of information, such as human pose, interaction with objects, object shape and appearance features. However, these methods can be used only to predict the labeling of an observed activity and cannot be used to anticipate what can happen next and how.

Our goal is to predict the future activities as well as the details of how a human is going to perform them in short-term (e.g., 1-10 seconds). We present an anticipatory temporal conditional random field (ATCRF), where we model the past with the CRF described above but augmented with the trajectories and with nodes/edges representing the object affordances, sub-activities, and trajectories in the future. Since there are many possible futures, each ATCRF represents only one of them. In order to find the most likely ones, we consider each ATCRF as a particle and propagate them over time, using the set of particles to represent the distribution over the future possible activities. One challenge is to use the discriminative power of the CRFs (where the observations are continuous and labels are discrete) for also producing the generative anticipation—labels over sub-activities, affordances, and spatial trajectories.



(a) Robot's RGB-D view. (b) Heatmap of object affordances. (c) Heatmap of trajectories. (d) Robot opening the door.

Fig. 1. **Reactive robot response through anticipation:** Robot observes a person holding an object and walking towards a fridge (a). It uses our ATCRF to anticipate the object affordances (b), and trajectories (c). It then performs an anticipatory action of opening the door (d).

We evaluate our anticipation approach extensively on CAD-120 human activity dataset [6], which contains 120 RGB-D videos of daily human activities, such as *having meal*, *microwaving food*, *taking medicine*, etc. Our algorithm obtains an activity anticipation accuracy (defined as whether one of top three predictions actually happened) of (75.4.1%,69.2%,58.1%) for predicting (1,3,10) seconds into the future. Our experiments also show good performance on anticipating the object affordances and trajectories. For robotic evaluation, we measure how many times the robot anticipates and performs the correct reactive response. Videos showing our robotic experiments and code are available at: `http://pr.cs.cornell.edu/anticipation/`.

II. APPROACH

Our goal is to anticipate what a human will do next given the current observation of his pose and the surrounding environment. Since activities happen over a long time horizon, with each activity being composed of sub-activities involving different number of objects, we first perform segmentation in time. Each temporal segment represents one sub-activity,

*Department of Computer Science, Cornell University. {hema,asaxena}@cs.cornell.edu

[1]This work was presented at RSS 2013 [1] and received the best student paper award. Submitting for consideration as an oral.

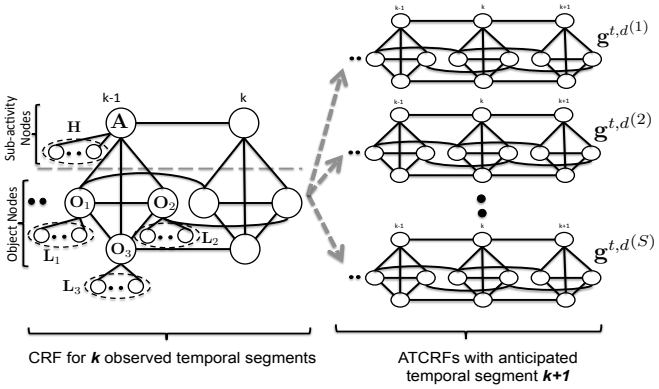


Fig. 2. Figure showing the CRF structure and the process of augmenting it to obtain multiple ATCRFs at time t for an activity with three objects. For the sake of clarity, frame level nodes are shown only for one temporal segment.

and we then model the activity using a spatio-temporal graph (a CRF) shown in Figure 2-left.

However, this graph can only model the present observations. In order to predict the future, we augment the graph with an 'anticipated' temporal segment, with anticipated nodes for sub-activities, objects (their affordances), and the corresponding spatio-temporal trajectories. We call this augmented graph an anticipatory temporal CRF (ATCRF).

Our goal is to obtain a distribution over the future possibilities, i.e., a distribution over possible ATCRFs. Motivated by particle filtering algorithm [9], we represent this distribution as a set of weighted particles, where each particle is a sampled ATCRF. Partial observations become available as the sub-activity is being performed and we use these partial observations to improve the estimation of the distribution. Since each of our ATCRF captures strong context over time (which sub-activity follows another) and space (spatial motion of humans and objects, and their interactions), each of our particles (i.e., possible future) is rich in its modeling capacity. Our experiments in Section III will show that this is essential for anticipating human actions.

Anticipated temporal segments are generated based on the available object affordances and the current configuration of the 3D scene. For example, if a person has picked up a coffee mug, one possible outcome could be drinking from it. Therefore, for each object, we sample possible locations at the end of the anticipated sub-activity and several trajectories based on the selected affordance.

III. EXPERIMENTS

We performed an extensive evaluation on CAD-120 human activity RGB-D dataset. For a new subject (not seen in the training set), we obtain an activity anticipation accuracy (defined as whether one of top three predictions actually happened) of 75.4%, 69.2% and 58.1% for an anticipation time of 1, 3 and 10 seconds respectively. Fig. 3 shows how the performance changes with the future anticipation time.

We also consider the following scenario for evaluating our algorithm on the robot for an assistance task: Robot is instructed to refill water glasses for people seated at a table, but when it anticipates an interaction with the cup, it waits for the interaction to complete before refilling. The robot

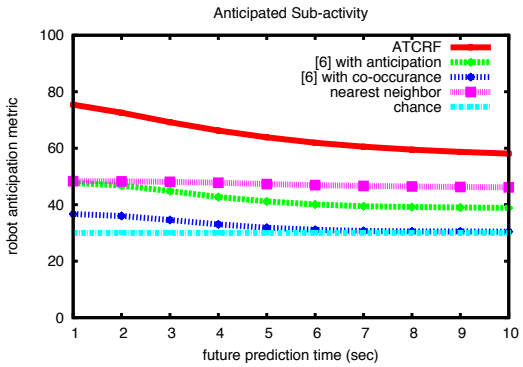


Fig. 3. Plot showing how *robot anticipation metric* changes with the future anticipation time.

considers the three top scored anticipations for taking the decision. We considered 40 pour instructions given during 10 interaction tasks, and obtained a success rate of 85%, which is the fraction of times the robot correctly identifies its response ('to pour' or 'not pour').

The accompanying video shows PR2 robot performing two assistive tasks based on the generated anticipations. In the first task, the robot assists in the activity by opening the fridge door when it sees the person approaching the fridge with an object. In the second task, the robot serves a drink without spilling by anticipating the person's interactions with the cup.

IV. CONCLUSIONS

In this work, we considered the problem of using anticipation of future activities. We modeled the human activities and object affordances in the past using a rich graphical model (CRF), and extended it to include future possible scenarios. Each possibility was represented as a potential graph structure and labeling over the graph (which includes discrete labels as well as human and object trajectories), which we called ATCRF. We used importance sampling techniques for estimating and evaluating the most likely future scenarios. We also extensively evaluated our algorithm, against baselines, on the tasks of anticipating activity and affordance labels.

REFERENCES

[1] H. S. Koppula and A. Saxena, "Anticipating human activities using object affordances for reactive robotic response," in *RSS*, 2013.
[2] K. Tang, L. Fei-Fei, and D. Koller, "Learning latent temporal structure for complex event detection," in *CVPR*, 2012.
[3] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele, "A database for fine grained activity detection of cooking activities," in *CVPR*, 2012.
[4] H. Pirsiavash and D. Ramanan, "Detecting activities of daily living in first-person camera views," in *CVPR*, 2012.
[5] J.-K. Min and S.-B. Cho, "Activity recognition based on wearable sensors using selection/fusion hybrid ensemble," in *SMC*, 2011.
[6] H. S. Koppula, R. Gupta, and A. Saxena, "Learning human activities and object affordances from rgb-d videos," *IJRR*, 2013.
[7] J. Sung, C. Ponce, B. Selman, and A. Saxena, "Unstructured human activity detection from rgbd images," in *ICRA*, 2012.
[8] B. Ni, G. Wang, and P. Moulin, "Rgbd-hudaact: A color-depth video database for human daily activity recognition," in *ICCV Workshop on CDC4CV*, 2011.
[9] M. Montemerlo, S. Thrun, and W. Whittaker, "Conditional particle filters for simultaneous mobile robot localization and people-tracking," in *ICRA*, 2002.