Anticipating Human Activities using Object Affordances for Reactive Robotic Response

Hema S. Koppula and Ashutosh Saxena.
Department of Computer Science, Cornell University.
{hema,asaxena}@cs.cornell.edu

***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) can enable an assistive robot to plan ahead for reactive responses in human environments. Furthermore, anticipation can even improve the detection accuracy of past activities. The challenge, however, is two-fold: We need to capture the rich context for modeling the activities and object affordances, and we need to anticipate the distribution over a large space of future human activities.**

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, we first show that anticipation improves the state-of-the-art detection results. 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.

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 [37, 31, 29], from inertial/location sensors [23], and more recently from RGB-D videos [21, 36, 27]. 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 enable robots 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). For example, if a robot has seen a person move his hand to a coffee mug, it is possible he would move the coffee mug to a few potential places such as his mouth, to a kitchen sink or just move it to a different location on the table. If a robot can anticipate this, then it would rather not start pouring milk into the coffee when the person is moving his hand towards the mug, thus avoiding a spill. Such scenarios happen in several other settings, for example, manufacturing scenarios in future co-robotic settings (e.g., [8, 28]).

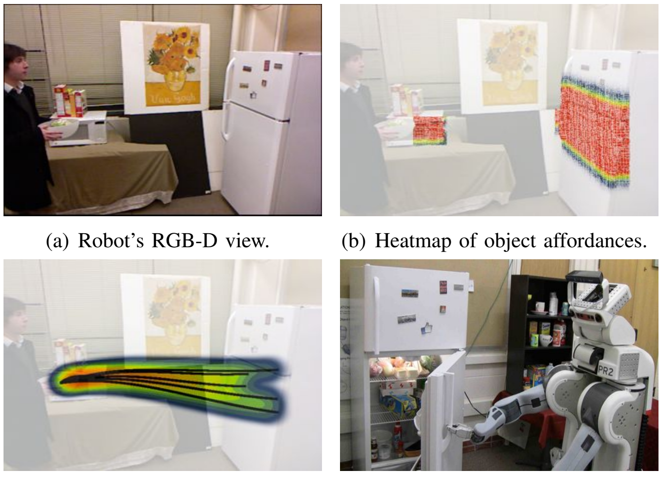
(a) Robot's RGB-D view.
(b) Heatmap of object affordances.
(c) Heatmap of trajectories.
(d) Robot opening the door.

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).

There are three aspects of activities that we need to model. First, we need to model the activities through a hierarchical structure in time where an activity is composed of a sequence of sub-activities [21]. Second, we need to model their inter-dependencies with objects and their affordances. We model the object affordances in terms of the relative position of the object with respect to the human and the environment.[1] Third, we need to anticipate the motion trajectory of the objects and humans, which tells us how the activity can be performed. Modeling trajectories not only helps in discriminating the activities,[2] but is also useful for the robot to reactively plan

[1] For example, a *drinkable* object is found near the mouth of the person performing the *drinking* activity and a *placeable* object is near a stable surface in the environment where it is being placed.

[2] For example, in stirring activity, the target position of the stirrer is immaterial but the circular trajectory motion is.

motions in the workspace.

For anticipation, we present an anticipatory temporal conditional random field (ATCRF), where we start with modeling the past with a standard CRF (based on [21]) 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.

We evaluate our anticipation approach extensively on CAD-120 human activity dataset [21], which contains 120 RGB-D videos of daily human activities, such as *microwaving food*, *taking medicine*, etc. We first show that anticipation improves the detection of *past* activities: 85.0% with vs 82.3% without. Our algorithm obtains an activity anticipation accuracy (defined as whether one of top three predictions actually happened) of (75.4%,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. Overview

In this section, we present an overview of our 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, and we then model the activity using a spatio-temporal graph (a CRF) shown in Figure 2-left, described in Section III-A.

what

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), formally defined in Section III-B.

Our goal is to obtain a distribution over the future possibilities, i.e., a distribution over possible ATCRFs. Motivated by particle filtering algorithm [25], 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. Section III-C describes this approach. 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. Later, our experiments in Section V 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. The location and trajectory generation are described in Section III-D and Section III-E respectively.

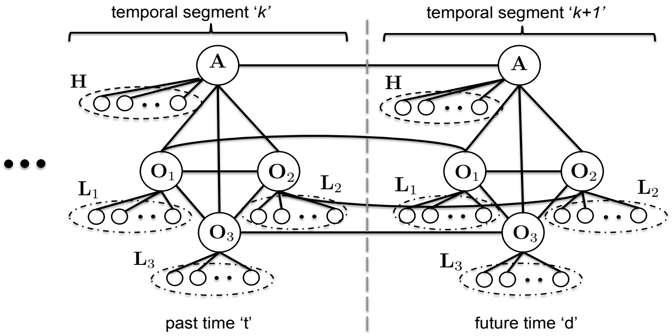

Fig. 2: An ATCRF that models the human poses $\mathcal{H}$, object affordance labels $\mathcal{O}$, object locations $\mathcal{L}$, and sub-activity labels $\mathcal{A}$, over past time 't', and future time 'd'. Two temporal segments are shown in this figure: k^{th} for the recent past, and $(k+1)^{th}$ for the future. Each temporal segment has three objects for illustration in the figure.

III. Our Approach

A robot observes a scene containing a human and objects for time t in the past, and its goal is to anticipate future possibilities for time d.

However, for the future d frames, we do not even know the structure of the graph—there may be different number of objects being interacted with depending on which sub-activity is performed in the future. Our goal is to compute a distribution over the possible future states (i.e., sub-activity, human poses and object locations). We will do so by sampling several possible graph structures by augmenting the graph in time, each of which we will call an anticipatory temporal conditional random field (ATCRF). We first describe an ATCRF below.

A. Modeling Past with an CRF

MRFs/CRFs are a workhorse of machine learning and have been applied to a variety of applications. Recently, with RGB-D data they have been applied to scene labeling [20, 1] and activity detection [21]. Conditioned on a variety of features as input, the CRFs model rich contextual relations. Learning and inference is tractable in these methods when the label space is discrete and small.

Following [21], we discretize time to the frames of the video[3] and group the frames into temporal segments, where each temporal segment spans a set of contiguous frames corresponding to a single sub-activity. Therefore, at time 't'

[3] In the following, we will use the number of videos frames as a unit of time, where 1 unit of time ≈ 71ms (=1/14, for a frame-rate of about 14Hz).

we have observed 't' frames of the activity that are grouped into 'k' temporal segments. For the past t frames, we know the structure of the CRF but we do not know the labels of the nodes in the CRF. We represent the graph until time t as: $\mathcal{G}^t = (\mathcal{V}^t, \mathcal{E}^t)$, where $\mathcal{E}^t$ represents the edges, and $\mathcal{V}^t$ represents the nodes $\{\mathcal{H}^t, \mathcal{O}^t, \mathcal{L}^t, \mathcal{A}^t\}$: human pose nodes $\mathcal{H}^t$, object affordance nodes $\mathcal{O}^t$, object location nodes $\mathcal{L}^t$, and sub-activity nodes $\mathcal{A}^t$. Figure 2-left part shows the structure of this CRF for an activity with three objects.

Our goal is to model the $P(\mathcal{H}^t, \mathcal{O}^t, \mathcal{L}^t, \mathcal{A}^t|\Phi^t_{\mathcal{H}}, \Phi^t_{\mathcal{L}})$, where $\Phi^t_{\mathcal{H}}$ and $\Phi^t_{\mathcal{L}}$ are the observations for the human poses and object locations until time t. Using the independencies expressed over the graph in Figure 2, for a graph $\mathcal{G}^t$, we have:

$$P_{\mathcal{G}^t}(\mathcal{H}^t, \mathcal{O}^t, \mathcal{L}^t, \mathcal{A}^t|\Phi^t_{\mathcal{H}}, \Phi^t_{\mathcal{L}}) = P(\mathcal{O}^t, \mathcal{A}^t|\mathcal{H}^t, \mathcal{L}^t)P(\mathcal{H}^t, \mathcal{L}^t|\Phi^t_{\mathcal{H}}, \Phi^t_{\mathcal{L}})$$

The second term $P(\mathcal{H}^t, \mathcal{L}^t|\Phi^t_{\mathcal{H}}, \Phi^t_{\mathcal{L}})$ models the distribution of true human pose and object locations (both are continuous trajectories) given the observations from the RGB-D Kinect sensor. We model it using a Gaussian distribution. The first term $P(\mathcal{O}^t, \mathcal{A}^t|\mathcal{H}^t, \mathcal{L}^t)$ predicts the object affordances and the sub-activities that are discrete labels—this term further factorizes following the graph structure as:

$$P(\mathcal{O}^t\!, \mathcal{A}^t|\mathcal{H}^t\!, \mathcal{L}^t) \propto \overbrace{\prod_{o_i\in\mathcal{O}} \Psi_{\mathcal{O}}(o_i|\ell_{o_i})}^{\text{object affordance}} \overbrace{\prod_{a_i\in\mathcal{A}} \Psi_{\mathcal{A}}(a_i|h_{a_i})}^{\text{sub-activity}} \prod_{v_i,v_j\in\mathcal{E}} \overbrace{\Psi_{\mathcal{E}}(v_i, v_j|\cdot)}^{\text{edge terms}} \tag{2}$$

Given the continuous state space of $\mathcal{H}$ and $\mathcal{L}$, we rely on [21] for powerful modeling using a discriminative framework for the above term.

B. ATCRF: Modeling one Possible Future with an augmented CRF.

We defined the anticipatory temporal conditional random field as an augmented graph $\mathcal{G}^{t,d} = (\mathcal{V}^{t,d}, \mathcal{E}^{t,d})$, where t is observed time and d is the future anticipation time. $\mathcal{V}^{t,d} = \{\mathcal{H}^{t,d}, \mathcal{O}^{t,d}, \mathcal{L}^{t,d}, \mathcal{A}^{t,d}\}$ represents the set of nodes in the past time t as well as in the future time d. $\mathcal{E}^{t,d}$ represents the set of all edges in the graph (see Figure 2.) The observations (not shown in the figure) are represented as set of features, $\Phi^t_{\mathcal{H}}$ and $\Phi^t_{\mathcal{O}}$, extracted from the t observed video frames. Note that we do not have observations for the future frames.

In the augmented graph $\mathcal{G}^{t,d}$, we have:

$$P_{\mathcal{G}^{t,d}}(\mathcal{H}^{t,d}, \mathcal{O}^{t,d}, \mathcal{L}^{t,d}, \mathcal{A}^{t,d}|\Phi^t_{\mathcal{H}}, \Phi^t_{\mathcal{L}}) = P(\mathcal{O}^{t,d}, \mathcal{A}^{t,d}|\mathcal{H}^{t,d}, \mathcal{L}^{t,d})P(\mathcal{H}^{t,d}, \mathcal{L}^{t,d}|\Phi^t_{\mathcal{H}}, \Phi^t_{\mathcal{L}}) \tag{3}$$

The first term is similar to Eq. (2), except over the augmented graph, and we can still rely on the discriminatively trained CRF presented in [21]. We model the second term with a Gaussian distribution.

C. Modeling the Distribution over Future Possibilities with ATCRFs.

There can be several potential augmented graph structures $\mathcal{G}^{t,d}$ because of different possibilities in human pose configurations and object locations that determines the neighborhood graph. Even the number of nodes to be considered in the future changes depending on the sub-activity and the configuration of the environment.

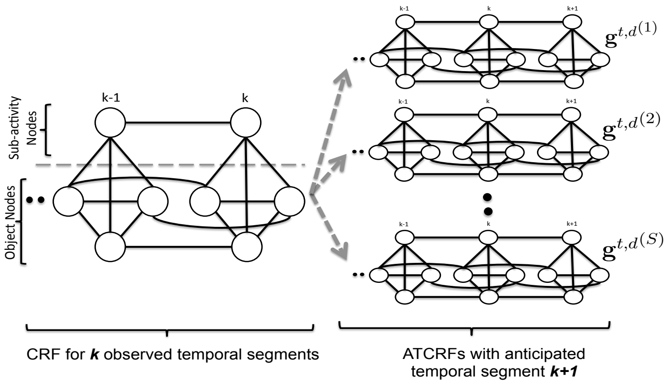

Fig. 3: Figure showing the process of augmenting the CRF structure to obtain multiple ATCRFs at time t for an activity with three objects. The frame level nodes are not shown for the sake of clarity. ?

Let $\mathbf{g}^{t,d}$ represent a sample augmented graph structure with particular values assigned to its node variables. I.e., one sample may represent that a person and object move in a certain way, performing a sub-activity with certain object affordances, and another sample may represent a person moving in a different way performing a different sub-activity.

Figure 3 shows the process of augmenting CRF structure corresponding to the seen frames with the sampled anticipations of the future to produce multiple ATCRF particles at time t. The frame level nodes are not shown in the figure. The left portion of the figure shows the nodes corresponding to the k observed temporal segments. This graph is then augmented with a set of anticipated nodes for the temporal segment $k+1$, to generate the ATCRF particles at time t. The frame level nodes of $k+1$ temporal segment are instantiated with anticipated human poses and object locations.

The goal of the robot is now to compute the distribution over these ATCRFs $\mathbf{g}^{t,d}$, i.e., given observations until time t, we would like to estimate the posterior distribution $p(\mathbf{g}^{t,d}|\Phi_t)$ from Eq. (3). However, this is extremely challenging because the space of ATCRFs is a very large one, so to even represent the distribution we need an exponential number of labels. We therefore represent the posterior using a set of weighted particles as shown in Eq. (4) and choose the weights using importance sampling as shown in Eq. (5).

$$p(\mathbf{g}^{t,d}|\Phi_t) \approx \sum_{s=1}^{S} \hat{w}^s_t \delta_{\mathbf{g}^{t,d(s)}}(\mathbf{g}^{t,d}) \tag{4}$$

$$\hat{w}^s_t \propto \frac{p(\mathbf{g}^{t,d(s)}|\Phi_t)}{q(\mathbf{g}^{t,d(s)}|\Phi_t)} \tag{5}$$

Here, $\delta_x(y)$ is the Kronecker delta function which takes the value 1 if x equals y and 0 otherwise, $\hat{w}^s_t$ is the weight of the sample s after observing t frames, and $q(\mathbf{g}^{t,d}|\Phi_t)$ is the proposal distribution. We need to perform importance sampling because: (a) sampling directly from $p(\mathbf{g}^{t,d}|\Phi_t)$ is not possible because of the form of the distribution in a discriminative framework, and (b) sampling uniformly would be quite naive because of the large space of ATCRFs and most of our samples would entirely miss the likely futures.

We now describe how we sample particles from the proposal

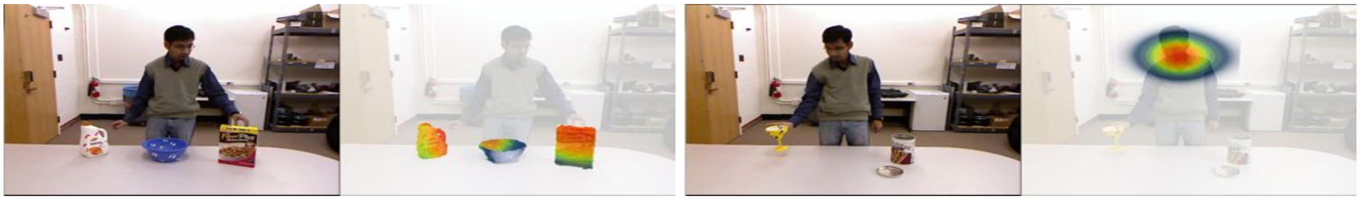

Fig. 4: Affordance heatmaps. The first two images show the *reachability* affordance heatmap (red signifies most likely *reachable* locations on the object) and the last two images show the *drinkability* affordance heatmap (red signifies the locations where the object is *drinkable*).

distribution $q(\mathbf{g}^{t,d}|\Phi_t)$ and how to evaluate the posterior for the generated samples.

Sampling. In order to generate a particle ATCRF, we need to generate possible human pose and object locations for the d future frames. We write the desired distribution to sample as:

$$q(\mathbf{g}^{t,d}|\Phi^t) = P_{\mathcal{G}^{t,d}}(\mathcal{H}^{t,d}, \mathcal{O}^{t,d}, \mathcal{L}^{t,d}, \mathcal{A}^{t,d}|\Phi^t_{\mathcal{H}}, \Phi^t_{\mathcal{L}})$$
$$= P_{\mathcal{G}^t}(\mathcal{H}^t, \mathcal{O}^t, \mathcal{L}^t, \mathcal{A}^t|\Phi^t_{\mathcal{H}}, \Phi^t_{\mathcal{L}})$$
$$P(\mathcal{H}^d, \mathcal{L}^d|\mathcal{O}^d, \mathcal{A}^d, \Phi^t_{\mathcal{H}}, \Phi^t_{\mathcal{L}},)P(\mathcal{O}^d, \mathcal{A}^d|\mathcal{O}^t, \mathcal{A}^t, \Phi^t_{\mathcal{H}}, \Phi^t_{\mathcal{L}}) \quad (6)$$

We first sample the affordances, one per object in the scene, and the corresponding sub-activity from the distribution $P(\mathcal{O}^d, \mathcal{A}^d|\Phi^t_{\mathcal{H}}, \Phi^t_{\mathcal{L}})$. This is discrete distribution generated from the training data based on the object type (e.g., cup, bowl, etc.) and object's current position with respect to the human in the scene (i.e., in contact with the hand or not). For example, if a human is holding an object of type 'cup' placed on a table, then the affordances *drinkable* and *movable* with their corresponding sub-activities (*drinking* and *moving* respectively) have equal probability, with all others being 0.

Once we have the sampled affordances and sub-activity, we need to sample the corresponding object locations and human poses for the d anticipated frames from the distribution $P(\mathcal{H}^d, \mathcal{L}^d|\mathcal{O}^d, \mathcal{A}^d, \Phi^t_{\mathcal{H}}, \Phi^t_{\mathcal{L}})$. In order to have meaningful object locations and human poses we take the following approach. We sample a set of target locations and motion trajectory curves based on the sampled affordance, sub-activity and available observations. We then generate the corresponding object locations and human poses from the sampled end point and trajectory curve. The details of sampling the target object location and motion trajectory curves are described in Section III-D and Section III-E respectively.

Scoring. Once we have the sampled ATCRF particles, we obtain the weight of each sample s by evaluating the posterior for the given sample, $q(\mathbf{g}^{t,d^{(s)}}|\Phi^t)$, as shown in Eq. (6) and normalize the weights across the samples.

D. Object Affordance Heatmaps

To represent object affordances we define a potential function based on how the object is being interacted with, when the corresponding affordance is active. The kind of interaction we consider depends on the affordance being considered. For example, when the active affordance of an object is *drinkable*, the object is found near the human's mouth, the interaction considered is the relative position of the object with respect to the human skeleton. In case of the affordance *placeable*, the interaction is the relative position of the object with respect to the environment, i.e., an object is *placeable* when it is above a surface that provides stability to the object once placed. The general form of the potential function for object affordance o given the observations at time t is:

$$\psi_o = \prod_i \psi_{dist_i} \prod_j \psi_{ori_j} \quad (7)$$

where ψ_{dist_i} is the i^{th} distance potential and ψ_{ori_j} is the j^{th} relative angular potential. We model each distance potential with a Gaussian distribution and each relative angular potential with a von Mises distribution. We find the parameters of the affordance potential functions from the training data using maximum likelihood estimation. Since the potential function is a product of the various components, the parameters of each distribution can be estimated separately. In detail, the mean and variance of the Gaussian distribution have closed form solutions, and we numerically estimate the mean and concentration parameter of the von Mises distribution.

We categorize these functions into three groups depending on the potentials used: (1) affordances *drinkable* and *reachable* have one distance potential per skeleton joint and one angular potential with respect to the head orientation, (2) affordances depending on the target object, such as *pourable* which depends on a *pour-to* object, have a distance potential and an angular potential with respect to the target object's location, (3) the rest of the affordances which depend on the environment, such *placeable* and *openable*, have a distance potential with respect to the closest surface and an angular potential with respect to the head orientation.

We generate heatmaps for each affordance by scoring the points in the 3D space using the potential function, and the value represents the strength of the particular affordance at that location. Figure 4 shows the heatmaps generated for the *reachable* and *drinkable* affordances. We obtain the future target locations of an object by weighted sampling of the scored 3D points.

E. Trajectory Generation

Once a location is sampled from the affordance heatmap, we generate a set of possible trajectories in which the object can be moved form its current location to the predicted target location. We use parametrized cubic equations, in particular Bézier curves, to generate human hand like motions [5]. We estimate the control points of the Bézier curves for the proposal distribution component from the trajectories in the training data. Figure 5 shows some of the anticipated trajectories for moving sub-activity.

Note that the aforementioned methods for the affordance and trajectory generation are only for the proposal distribution

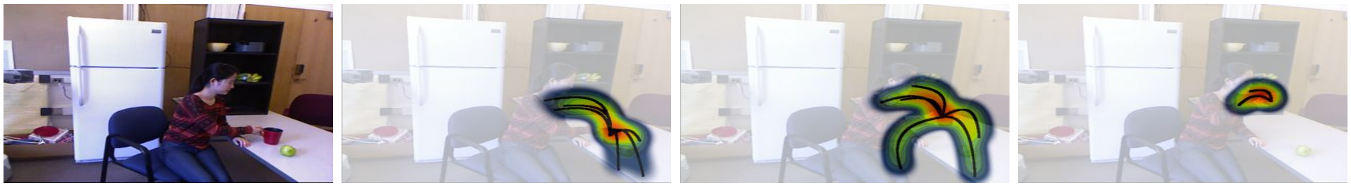

Fig. 5: Figure showing the heatmap of anticipated trajectories for moving sub-activity and how the trajectories evolve with time.

to sample. The estimated trajectories are finally scored using our ATCRF model.

IV. Related Work

Activity Detection. In recent years, much effort has been made to detect human activities from still images as well as videos. These works use human pose for action recognition by detecting local pose features [39] and modeling the spatial conguration between human body parts and objects [9, 40, 15, 16, 21]. There are also a few recent works which address the task of early recognition [32, 12]. Recently, with the availability of inexpensive RGB-D sensors, some works [41, 27, 36] consider detecting human activities from RGB-D videos. Koppula el al. [21] proposed a model to jointly predict sub-activities and object affordances by taking into account both spatial as well as temporal interactions between human poses and object interactions. However, all these work only predict the activities and affordance after the action is performed. None of these methods can anticipate what is going to happen next.

Anticipation of Human Actions. Anticipation or forecasting future human actions has been the focus of few recent works. Maximum entropy inverse reinforcement learning was used by [42, 18, 22] to obtain a distribution over possible human navigation trajectories from visual data, and also used to model the forthcoming interactions with pedestrians for mobile robots [42, 22]. However, these works focus only on human actions which are limited to navigation trajectories. Wang et al. [38] propose a latent variable model for inferring unknown human intentions, such as the target ball position in a robot table tennis scenario, to plan the robot's response. Dragan et al. [4] use inverse reinforcement learning to improve assistive teleoperation by combining user input with predictions of future goal for grasping an object and the motion trajectory to reach the goal. In comparison, we address the problem of anticipation of human actions at a fine-grained level of how a human interacts with objects in more involved activities such as *microwaving food* or *taking medicine* compared to the generic navigation activities or task-specific trajectories.

Learning Algorithms. Our work uses probabilistic graphical models to capture rich context. Such frameworks as HMMs [13, 26], DBNs [7], CRFs [30, 35], semi-CRFs [33] have been previously used to model the temporal structure of videos and text. While these previous works maintain their template graph structure over time, in our work new graph structures are possible. More importantly, our goal is anticipation and we use importance sampling for efficient estimation of the likelihood of the potential future activities.

Particle filters have been applied with great success to a variety of state estimation problems including object tracking [17, 11], mobile robot localization [6, 14], people tracking [34], etc. However, the worst-case complexity of these methods grows exponentially in the dimensions of the state space, it is not clear how particle filters can be applied to arbitrary, high-dimensional estimation problems. Some approaches use factorizations of the state space and apply different representations for the individual parts of the state space model. For example, Rao-Blackwellised particle filters sample only the discrete and non-linear parts of a state estimation problem. The remaining parts of the states are solved analytically conditioned on the particles by using Kalman filters [3, 10, 24, 34]. In our work, each of our particles is a CRF that models rich structure and lies in a high-dimensional space.

V. Experiments

Data. We use CAD-120 dataset [21] for our evaluations. The dataset has 120 RGB-D videos of four different subjects performing 10 high-level activities. The data is annotated with object affordance and sub-activity labels and includes ground-truth object categories, tracked object bounding boxes and human skeletons. The set of high-level activities are: {*making cereal*, *taking medicine*, *stacking objects*, *unstacking objects*, *microwaving food*, *picking objects*, *cleaning objects*, *taking food*, *arranging objects*, *having a meal*}, the set of sub-activity labels are: {*reaching*, *moving*, *pouring*, *eating*, *drinking*, *opening*, *placing*, *closing*, *scrubbing*, *null*} and the set of affordance labels are: {*reachable*, *movable*, *pourable*, *pourto*, *containable*, *drinkable*, *openable*, *placeable*, *closable*, *scrubbable*, *scrubber*, *stationary*}. We use all sub-activity classes for prediction of observed frames but do not anticipate *null* sub-activity.

Baseline Algorithms. We compare our method against the following baselines: *1) Chance.* The anticipated sub-activity and affordance labels are chosen at random.

2) Nearest Neighbor Exemplar. It first finds an example from the training data which is the most similar to the activity observed in the last temporal segment. The sub-activity and object affordance labels of the frames following the matched frames from the exemplar are predicted as the anticipations. To find the exemplar, we perform a nearest neighbor search in the feature space for the set of frames, using the node features described in [21].

3) Co-occurrence Method. The transition probabilities for sub-activities and affordances are computed from the training data. The observed frames are first labelled using the MRF model proposed by [21]. The anticipated sub-activity and affordances for the future frames are predicted based on the transition probabilities given the inferred labeling of the last frame.

4) ATCRF without $\{\mathcal{H}, \mathcal{L}\}$ anticipation (ATCRF-discrete). Our

TABLE I: Anticipation Results of Future Activities and Affordances, computed over 3 seconds in the future (similar trends hold for other anticipation times).

	Anticipated Sub-activity			Anticipated Object Affordance		
model	micro P/R	macro F1-score	robot anticipation metric	micro P/R	marco F1-score	robot anticipation metric
chance	10.0 ± 0.1	10.0 ± 0.1	30.0 ± 0.1	8.3 ± 0.1	8.3 ± 0.1	24.9 ± 0.1
Nearest-neighbor	22.0 ± 0.9	10.6 ± 0.6	48.1 ± 0.5	48.3 ± 1.5	17.2 ± 1.0	60.9 ± 1.1
Koppula et al. [21] + co-occurence	28.6 ± 1.8	11.1 ± 0.4	34.6 ± 2.8	55.9 ± 1.7	11.6 ± 0.4	62.0 ± 1.8
ATCRF-discrete	34.3 ± 0.8	12.2 ± 0.2	44.8 ± 1.1	59.5 ± 1.5	12.4 ± 0.3	67.6 ± 1.3
ATCRF	**47.7** ± 1.6	**37.9** ± 2.6	**69.2** ± 2.1	**66.1** ± 1.9	**36.7** ± 2.3	**71.3** ± 1.7

ATCRF model with only augmented nodes for discrete labels (sub-activities and object affordances).

Evaluation: We follow the same train-test split described in [21] and train our model on activities performed by three subjects and test on activities of a *new subject*. We report the results obtained by 4-fold cross validation by averaging across the folds. We consider the following metrics:
1) Labeling Metrics. For detecting and anticipating labels (for sub-activity and affordances), we compute the overall micro accuracy (P/R), macro precision, macro recall and macro F1 score. Micro accuracy is the percentage of correctly classified labels. Macro precision and recall are the averages of precision and recall respectively for all classes.
2) Robot Anticipation Metric. It is important for a robot to plan ahead for multiple future activity outcomes. Therefore, we measure the accuracy of the anticipation task for the top three predictions of the future. If the actual activity matches one of the top three predictions, then it counts towards positive.
3) Trajectory Metric. For evaluating the quality of anticipating trajectories, we compute the modified Hausdorff distance (MHD) as a physical measure of the distance between the anticipated object motion trajectories and the true object trajectory from the test data.[4]

Table I shows the frame-level metrics for anticipating sub-activity and object affordance labels for 3 seconds in the future on the CAD-120 dataset. We use the temporal segmentation algorithm from [21] for obtaining the graph structure of the observed past frames for all the methods. ATCRF outperforms all the baseline algorithms and achieves a significant increase across all metrics. Improving temporal segmentation further improves the anticipation performance [19]. We will now study our results on anticipation in the form of the following questions:

How does the performance change with the duration of the future anticipation? Figure 6 shows how the macro F1 score and the *robot anticipation metric* changes with the anticipation time. The average duration of a sub-activity in the CAD-120 dataset is around 3.6 seconds, therefore, an anticipation duration of 10 seconds is over two to three sub-activities. With the increase in anticipation duration, performance of the others approach that of a random chance baseline, the performance of our ATCRF declines. It still outperforms other baselines for all anticipation times.

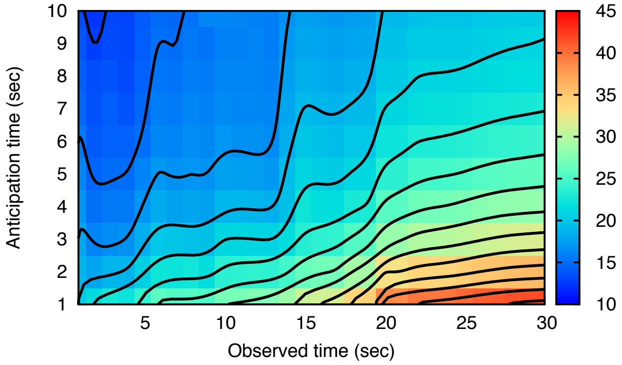

Fig. 7: Plot showing how macro F1 score depends on observed time and anticipation time.

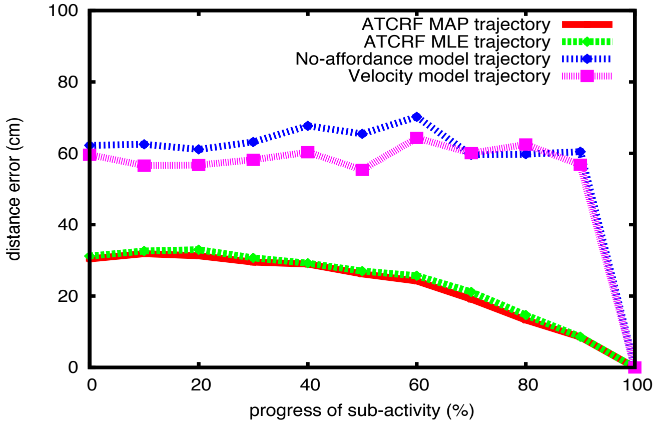

Fig. 8: Plot showing how the trajectory distance error (MHD) changes with the progress of the activity for our ATCRF (top particle and local mean) and other baselines (Kalman Filter velocity model using object affordance as target, and one without object affordance information).

How does the performance change with the duration of the past observations? Figure 7 shows how the macro F1 score changes with the past observation time and future anticipation time. The algorithm has lower performance when predicting longer into the future, but this improves as more observations from the activity become available. Therefore, context from the past helps in anticipating longer into the future.

How good are the anticipated trajectories? Since trajectories are continuous variables, we perform two types of estimation: MAP, where we take the highest scored particle generated by our model, and MLE where we take the weighted sum. Figure 8 shows how these distance errors, averaged over all the moving sub-activities in the dataset, change with the progress of the sub-activity. Figure 5 shows the sampled trajectories along with the heatmap corresponding to the distribution of trajectories. At the beginning of the sub-activity the anticipations correspond to moving the cup to other places on the table and near the mouth to drink. As the sub-activity

[4] The MHD allows for local time warping by finding the best local point correspondence over a small temporal window. When the temporal window is zero, the MHD is same as the Euclidean distance between the trajectories. We normalize the distance by the length of the trajectory in order to compare performance across trajectories of different lengths. The units of the MHD are centimeters.

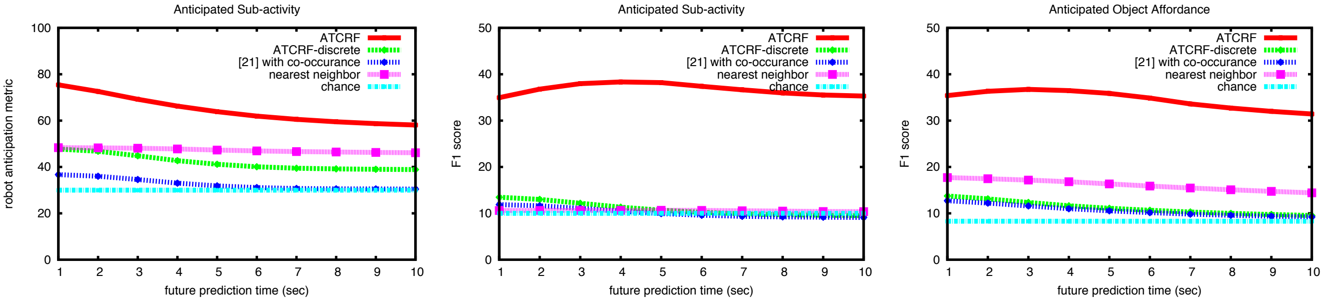

Fig. 6: Plots showing how *robot anticipation metric* and macro F1 score changes with the future anticipation time for all methods.

TABLE II: Online Detection Results of Past Activities and Affordances.

model	Past Sub-activity Detection			Past Object Affordance Detection		
	micro	macro		micro	macro	
	P/R	Prec.	Recall	P/R	Prec.	Recall
chance	10.0 (0.1)	10.0 (0.1)	10.0 (0.1)	8.3 (0.1)	8.3 (0.1)	8.3 (0.1)
Koppula et al. [21] - online	80.3 (1.5)	78.0 (1.3)	68.1 (2.6)	89.6 (0.8)	80.7 (2.8)	67.8 (1.4)
ATCRF-discrete	84.0 (1.3)	72.2 (2.3)	60.7 (2.3)	87.7 (1.0)	67.9 (2.4)	48.9 (2.6)
ATCRF	**84.7** (1.4)	**80.6** (1.0)	**75.6** (2.4)	**92.3** (0.7)	**84.8** (2.3)	**77.1** (1.1)

progresses, depending on the current position of the cup, a few new target locations become probable, such as moving the cup on to the lap (such instances are observed in the training data). These new possibilities tend to increase the distance measure as can be seen in the plot of Figure 8. However, on observing more frames, the intent of the human is inferred more accurately resulting in better anticipated trajectories, for example in Figure 5-last frame, anticipating only moving to drink trajectories.

Effect of anticipation on detection of past activities. Table II shows the detection results of the sub-activities and object affordances of the past temporal segments, computed in an online fashion. When we label each past segment, we observe that segment's features but not the future. The online metrics are computed by aggregating performance on the recent past of three segments. (Koppula et al. [21]'s method was to label a segment given past, present, as well as the future.) In this experiment, we assumed ground-truth segmentation and object tracks for consistent comparison across the methods. If we instead use an algorithm to segment [21], the overall performance drops, however similar trends hold. We see that both the anticipation methods (rows 3-4) improve the detection results over the one that does not anticipate (row 2). This shows that anticipating the future can improve present and past performance on detection.

A. Robotic Experiments

In this section we show how future activity predictions can help the robot perform appropriate actions in response to what the human is going to do next. By incorporating such reactive responses, the robot can better assist humans in tasks which they are unable to perform as well as work along side the humans much more efficiently.

We use a PR2 robot to demonstrate the following anticipatory response scenarios:

- Robot is instructed to refill water glasses for people seated at a table, but when it sees a person reaching a glass to drink, it waits for him to finish drinking before refilling, in order to avoid spilling.
- Robot opens the fridge door when a person approaches the fridge to place something inside the fridge.

PR2 is mounted with a Kinect as its main input sensor to obtain the RGB-D video stream. We used the OpenRAVE libraries [2] for programing the robot to perform the pre-programmed tasks described in the aforementioned scenarios by incorporating the anticipations generated with our ATCRFs. Figure 1 and Figure 9 show the snapshots of the robot observing the human, the anticipated actions and the response executed by the robot.

In our experiments, on the first scenario, we evaluate the success rate which is defined as the percentage of times the robot identifies the correct response. We have a new subject (not seen in the training data) performing the interaction task multiple times in addition to other activities which should not effect the robot's response, such as reaching for a book, etc. We considered a total of 10 interaction tasks which involve four objects including the cup, and 5 of these tasks were to reach for the cup and drink from it. The robot is given an instruction to pour water in the cup at four random time instants during each interaction tasks (40 total pour instructions). The robot makes a decision whether to execute the pouring task or not, based on the anticipated activity and object affordance. For example, if the robot anticipates a reaching action, where the cup is the reachable object, it will not perform the pouring action. The robot considers the three top scored anticipations for taking the decision following the *robot anticipation metric*.

We obtain a success rate of 85%, which is the fraction of times the robot correctly identifies its response ('pour' or 'not pour'). Out of the 6 failed instances, 3 instances are false-negatives, i.e., the robot anticipated an interaction with the cup when no interaction occurred in future. Videos showing the results of our robotic experiments and code are available at: `http://pr.cs.cornell.edu/anticipation/`.

VI. Conclusion

In this work, we considered the problem of using anticipation of future activities so that a robot could perform look-ahead planning of its reactive responses. 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 showed that anticipation can improve performance of

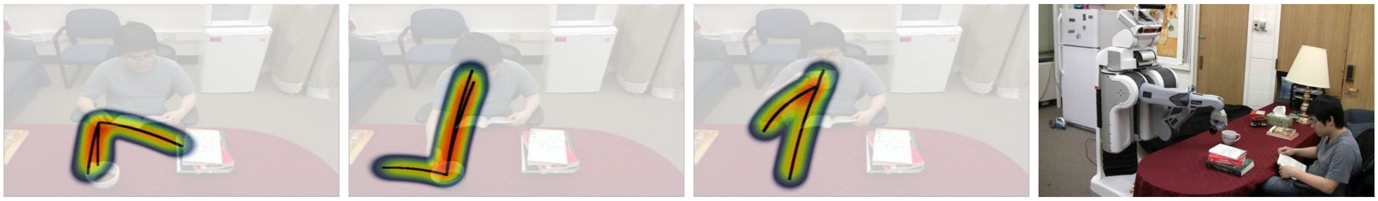

Fig. 9: Robot Anticipatory Response for refilling water task. See Figure 1 for opening fridge door task.

detection of even past activities and affordances. We also extensively evaluated our algorithm, against baselines, on the tasks of anticipating activity and affordance labels as well as the object trajectories in the future.

Acknowledgments

We thank Rudhir Gupta and Jerry Yeh for useful discussions. This work was partially supported by ARO award W911NF-12-1-0267, and by Microsoft Faculty Fellowship and Alfred P. Sloan Research Fellowship to one of us (Saxena).

References

[1] A. Anand, H. S. Koppula, T. Joachims, and A. Saxena. Contextually guided semantic labeling and search for 3d point clouds. *IJRR*, 2012.

[2] Rosen Diankov. *Automated Construction of Robotic Manipulation Programs*. PhD thesis, CMU, Robotics Institute, 2010.

[3] A. Doucet, N. de Freitas, K. P. Murphy, and S. J. Russell. Rao-blackwellised particle filtering for dynamic bayesian networks. In *UAI*, 2000.

[4] A. Dragan and S. Srinivasa. Formalizing assistive teleoperation. In *RSS*, 2012.

[5] J. J. Faraway, M. P. Reed, and J. Wang. Modelling three-dimensional trajectories by using bezier curves with application to hand motion. *JRSS Series C*, 56:571–585, 2007.

[6] D. Fox. Kld-sampling: Adaptive particle filters. In *NIPS*, 2001.

[7] S. Gong and T. Xiang. Recognition of group activities using dynamic probabilistic networks. In *ICCV*, 2003.

[8] E. Guizzo and E. Ackerman. The rise of the robot worker. *Spectrum, IEEE*, 49(10):34 –41, October 2012.

[9] A. Gupta, A. Kembhavi, and L. S. Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE PAMI*, 31(10):1775–1789, 2009.

[10] F. Gustafsson, F. Gunnarsson, N. Bergman, U. Forssell, J. Jansson, R. Karlsson, and P.-J. Nordlund. Particle filters for positioning, navigation, and tracking. *IEEE Trans. SP*, 2002.

[11] R. Hess and A. Fern. Discriminatively trained particle filters for complex multi-object tracking. In *CVPR*, 2009.

[12] M. Hoai and F. De la Torre. Max-margin early event detectors. In *CVPR*, 2012.

[13] S. Hongeng and R. Nevatia. Large-scale event detection using semi-hidden markov models. In *ICCV*, 2003.

[14] P. Jensfelt, D. Austin, O. Wijk, and M. Andersson. Feature based condensation for mobile robot localization. In *ICRA*, 2000.

[15] Y. Jiang, M. Lim, and A. Saxena. Learning object arrangements in 3d scenes using human context. In *ICML*, 2012.

[16] Y. Jiang, H. S. Koppula, and A. Saxena. Hallucinated humans as the hidden context for labeling 3d scenes. In *CVPR*, 2013.

[17] Z. Khan, T. Balch, and F. Dellaert. Mcmc-based particle filtering for tracking a variable number of interacting targets. *IEEE PAMI*, 2005.

[18] K. Kitani, B. D. Ziebart, J. A. Bagnell, and M. Hebert. Activity forecasting. In *ECCV*, 2012.

[19] H. S. Koppula and Ashutosh Saxena. Learning spatio-temporal structure from rgb-d videos for human activity detection and anticipation. In *ICML*, 2013.

[20] H. S. Koppula, A. Anand, T. Joachims, and A. Saxena. Semantic labeling of 3d point clouds for indoor scenes. In *NIPS*, 2011.

[21] H. S. Koppula, R. Gupta, and A. Saxena. Learning human activities and object affordances from rgb-d videos. *IJRR*, 2013.

[22] M. Kuderer, H. Kretzschmar, C. Sprunk, and W. Burgard. Feature-based prediction of trajectories for socially compliant navigation. In *RSS*, 2012.

[23] J.-K. Min and S.-B. Cho. Activity recognition based on wearable sensors using selection/fusion hybrid ensemble. In *SMC*, 2011.

[24] M. Montemerlo, S. Thrun, D. Koller, and B. Wegbreit. Fast-slam: A factored solution to the simultaneous localization and mapping problem. In *AAAI*, 2002.

[25] M. Montemerlo, S. Thrun, and W. Whittaker. Conditional particle filters for simultaneous mobile robot localization and people-tracking. In *ICRA*, 2002.

[26] P. Natarajan and R. Nevatia. Coupled hidden semi markov models for activity recognition. In *WMVC*, 2007.

[27] 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.

[28] S. Nikolaidis and J. Shah. Human-robot cross-training: computational formulation, modeling and evaluation of a human team training strategy. In *HRI*, 2013.

[29] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. In *CVPR*, 2012.

[30] A. Quattoni, S. Wang, L.-P. Morency, M. Collins, and T. Darrell. Hidden-state conditional random fields. *IEEE PAMI*, 2007.

[31] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele. A database for fine grained activity detection of cooking activities. In *CVPR*, 2012.

[32] M.S. Ryoo. Human activity prediction: Early recognition of ongoing activities from streaming videos. In *ICCV*, 2011.

[33] S. Sarawagi and W. W. Cohen. Semi-markov conditional random fields for information extraction. In *NIPS*, 2004.

[34] D. Schulz, D. Fox, and J. Hightower. People tracking with anonymous and id-sensors using rao-blackwellised particle filters. In *IJCAI*, 2003.

[35] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas. Conditional models for contextual human motion recognition. In *ICCV*, 2005.

[36] J. Sung, C. Ponce, B. Selman, and A. Saxena. Unstructured human activity detection from rgbd images. In *ICRA*, 2012.

[37] K. Tang, L. Fei-Fei, and D. Koller. Learning latent temporal structure for complex event detection. In *CVPR*, 2012.

[38] Z. Wang, M. Deisenroth, H. B. Amor, D. Vogt, B. Scholkopf, and J. Peters. Probabilistic modeling of human movements for intention inference. In *RSS*, 2012.

[39] W. Yang, Y. Wang, and G. Mori. Recognizing human actions from still images with latent poses. In *CVPR*, 2010.

[40] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *CVPR*, 2010.

[41] H. Zhang and L. E. Parker. 4-dimensional local spatio-temporal features for human activity recognition. In *IROS*, 2011.

[42] B. D. Ziebart, N. Ratliff, G. Gallagher, C. Mertz, K. Peterson, J. A. Bagnell, M. Hebert, A. Dey, and S. Srinivasa. Planning-based prediction for pedestrians. In *IROS*, 2009.