Anticipating the Future by Constructing Human Activities using Object Affordances

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 and 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 present a constructive approach for generating various possible future human activities by reasoning about the rich spatial-temporal relations through object affordances. We represent each possible future using an anticipatory temporal conditional random field (ATCRF) where we sample the nodes and edges corresponding to future object trajectories and human poses from a generative model. We then represent the distribution over the potential futures using a set of constructed ATCRF 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. [1]

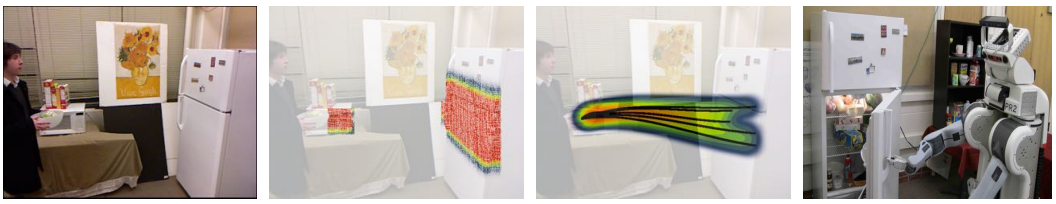
1 Introduction

Human activities are composed of many sub-activities that are performed in a sequence, in order to achieve a goal. For example, when the goal is to prepare cereal, the person first gets a bowl, reaches for cereal, moves it and pours cereal in to the bowl, etc. For many applications it is important 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, for example, an assistive robot (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 model three main aspects of the activities which allow us to construct various possible future activities that a person can perform and evaluate how likely these constructed future activities are. First, we model the activities through a hierarchical structure in time where an activity is composed of a sequence of sub-activities[6]. Second, we model their inter-dependencies with objects and their affordances. Third, we model the motion trajectory of the objects and humans, which tells us how the activity can be performed.

[1] Parts of this work have been published at RSS 2013 [1].



(a) Robot's RGB-D view. (b) Affordance heatmap. (c) Trajectory heatmap. (d) Robot's response.

Figure 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 present an anticipatory temporal conditional random field (ATCRF) where we model the past with a CRF as described in [6] and augment it with the trajectories and with nodes/edges representing the object affordances, sub-activities, and trajectories in the future. We construct many ATCRFs, each corresponding to a possible future, by sampling the object trajectories and human poses from a generative process modeling the activities and object affordnaces. Therefore, for each anticipation we have one sampled graph for modeling the spatio-temporal relations of the anticipated activity. In order to find the most likely future, we consider each constructed 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 [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.

2 Our Approach

After observing a scene containing a human and objects for time t in the past, our 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.

2.1 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 [9, 10] and activity detection [6]. Following [6], we discretize time to the frames of the video[2] 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' 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.

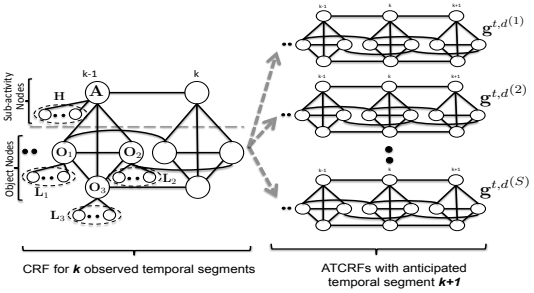


Figure 2: Figure showing the CRF structure and the process of augmenting it to obtain multiple ATCRFs at time t for an activity with three objects. For the sake of clarity, frame level nodes are shown only for one temporal segment.

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

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-left, we have:

$$P_{\mathcal{G}^t}(\mathcal{H}^t, \mathcal{O}^t, \mathcal{L}^t, \mathcal{A}^t | \Phi^t_{\mathcal{H}}, \Phi^t_{\mathcal{L}}) = P(\mathcal{O}^t, \mathcal{A}^t | \mathcal{H}^t, \mathcal{L}^t) P(\mathcal{H}^t, \mathcal{L}^t | \Phi^t_{\mathcal{H}}, \Phi^t_{\mathcal{L}}) \tag{1}$$

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:

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

Given the continuous state space of $\mathcal{H}$ and $\mathcal{L}$, we rely on [6] for powerful modeling using a discriminative framework for the above term.

2.2 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. The observations 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 [6]. We model the second term with a Gaussian distribution.

2.3 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. Let $\mathbf{g}^{t,d}$ represent a sample augmented graph structure with particular values assigned to its node variables. Figure 2 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 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 and choose the weights using importance sampling as shown in Eq. (4).

$$p(\mathbf{g}^{t,d} | \Phi_t) \approx \sum_{s=1}^{S} \hat{w}_t^s \delta_{\mathbf{g}^{t,d(s)}}(\mathbf{g}^{t,d}) \; ; \qquad \hat{w}_t^s \propto \frac{p(\mathbf{g}^{t,d(s)} | \Phi_t)}{q(\mathbf{g}^{t,d(s)} | \Phi_t)} \tag{4}$$

Here, $\delta_x(y)$ is the Kronecker delta function which takes the value 1 if x equals y and 0 otherwise, $\hat{w}_t^s$ 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.

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 as:

$$\begin{aligned} q(\mathbf{g}^{t,d} | \Phi^t) =& P_{\mathcal{G}^{t,d}}(\mathcal{H}^{t,d}, \mathcal{O}^{t,d}, \mathcal{L}^{t,d}, \mathcal{A}^{t,d} | \Phi^t_{\mathcal{H}}, \Phi^t_{\mathcal{L}}) = P_{\mathcal{G}^t}(\mathcal{H}^t, \mathcal{O}^t, \mathcal{L}^t, \mathcal{A}^t | \Phi^t_{\mathcal{H}}, \Phi^t_{\mathcal{L}}) \\ & P(\mathcal{H}^d, \mathcal{L}^d | \mathcal{O}^d, \mathcal{A}^d, \Phi^t_{\mathcal{H}}, \Phi^t_{\mathcal{L}},) P(\mathcal{O}^d, \mathcal{A}^d | \mathcal{O}^t, \mathcal{A}^t, \Phi^t_{\mathcal{H}}, \Phi^t_{\mathcal{L}}) \end{aligned} \tag{5}$$

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

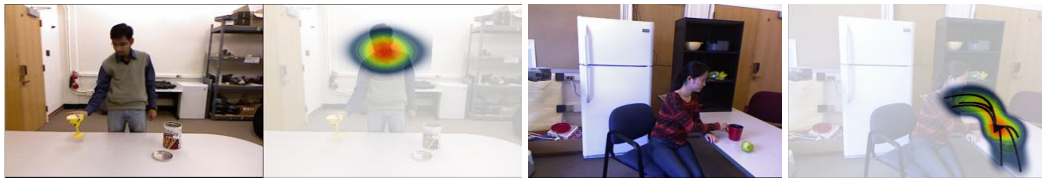


Figure 3: **Affordance and trajectory heatmaps**. The first two images show the *drinkability* affordance heatmap (red signifies the locations where the object is *drinkable*). The last two images show the heatmap of anticipated trajectories for *moving* sub-activity.

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).

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 below.

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. (5) and normalize the weights across the samples.

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 general form of the potential function for object affordance o given the observations at time t is:

$$\psi_o = \prod_i \psi_{dist_i} \prod_j \psi_{ori_j} \tag{6}$$

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.

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 3-left shows the heatmap generated for the *drinkable* affordances. We obtain the future target locations of an object by weighted sampling of the scored 3D points.

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 [11]. We estimate the control points of the Bézier curves from the trajectories in the training data. Figure 3-right shows some of the anticipated trajectories for moving sub-activity.

Note that the aforementioned methods for the affordance and trajectory generation are only for generating samples. The estimated trajectories are finally scored using our ATCRF model.

3 Experiments

Data. We use CAD-120 dataset [6] for our evaluations. The dataset has 120 RGB-D videos of four different subjects performing 10 high-level activities. The data is annotated with 12 object affordance labels and 10 sub-activity labels and includes ground-truth object categories, tracked object bounding boxes and human skeletons. 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 [6].

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

Evaluation: We follow the same train-test split described in [6] 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. For anticipating sub-activity and affordance labels, we compute the overall *accuracy*. Accuracy is the percentage of correctly classified labels. In many applications, it is often important to plan ahead for multiple future activity outcomes. Therefore, we define the *anticipation metric* as the accuracy of the anticipation task for the top three future predictions.

Table 1: Anticipation Results of Future Activities and Affordances, computed over 3 seconds in the future (similar trends hold for other anticipation times).

model	Anticipated Sub-activity		Anticipated Object Affordance	
	accuracy	anticipation metric	accuracy	anticipation metric
chance	10.0 ± 0.1	30.0 ± 0.1	8.3 ± 0.1	24.9 ± 0.1
Nearest-neighbor	22.0 ± 0.9	48.1 ± 0.5	48.3 ± 1.5	60.9 ± 1.1
[6] + co-occur.	28.6 ± 1.8	34.6 ± 2.8	55.9 ± 1.7	62.0 ± 1.8
ATCRF-discrete	34.3 ± 0.8	44.8 ± 1.1	59.5 ± 1.5	67.6 ± 1.3
ATCRF	**47.7** ± 1.6	**69.2** ± 2.1	**66.1** ± 1.9	**71.3** ± 1.7

Table 1 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 [6] 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. Please see [1] for more detailed evaluations. Improving temporal segmentation further improves the anticipation performance [12]. Videos showing the results of our robotic experiments and code are available at: `http://pr.cs.cornell.edu/anticipation/`.

4 Conclusion

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 labeled graph structure (which includes discrete labels as well as human and object trajectories), which we call *anticipatory temporal conditional random field* (ATCRF). We used importance sampling techniques for constructing possible future activities, and estimate the most likely future scenarios. We also extensively evaluated our algorithm 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] H. S. Koppula, A. Anand, T. Joachims, and A. Saxena. Semantic labeling of 3d point clouds for indoor scenes. In *NIPS*, 2011.

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

[11] 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.

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