rCRF: Recursive Belief Estimation over CRFs in RGB-D Activity Videos

Ozan Sener
School of Electrical & Computer Eng.
Cornell University

Ashutosh Saxena
Department of Computer Science
Cornell University

***Abstract*—For assistive robots, anticipating the future actions of humans is an essential task. This requires modelling both the evolution of the activities over time and the rich relationships between humans and the objects. Since the future activities of humans are quite ambiguous, robots need to assess all the future possibilities in order to choose an appropriate action. Therefore, a successful anticipation algorithm needs to compute all plausible future activities and their corresponding probabilities.**

In this paper, we address the problem of efficiently computing beliefs over future human activities from RGB-D videos. We present a new recursive algorithm that we call Recursive Conditional Random Field (rCRF) which can compute an accurate belief over a temporal CRF model. We use the rich modelling power of CRFs and describe a computationally tractable inference algorithm based on Bayesian filtering and structured diversity. In our experiments, we show that incorporating belief, computed via our approach, significantly outperforms the state-of-the-art methods, in terms of accuracy and computation time.

I. Introduction

Understanding human activities is an important skill for robots working with humans. Robots not only need to detect the activity that human is performing but also need to anticipate *what activity can a human possibly perform in the near future* in order to choose the right actions. Anticipation ability is especially important for assistive robots, and we have recently seen many successful collaborative robotics applications [**? ? ?**] using the most likely action(s) humans might take in near future. The set of the future possibilities is quite large, and robots need to be aware of all of them in addition to the most likely one. In this work, we focus on estimating the set of all possible future states with their likelihoods.

Anticipation is a challenging task, and it requires us to model the relationships between several objects and the human(s) in the scene, as well as their temporal evolution. Although the modelling assumptions and model parametrization varies, the common approach [**? ? ? ?**] is using Conditional Random Field (CRF) to represent the rich relations in the scene, and anticipating a single or a few most likely future states. Since the future is ambiguous, the most likely state might not be sufficient enough to assess the risk of each action. For example, consider a collaborative cooking scenario, the object that human is reaching is typically a distribution over many objects. Computing the trajectory, that is least likely to conflict with the human, is only possible via consideration of all future possibilities. The question, we address in this paper, is: *How can we estimate all plausible future activities and their probabilities in a scene modelled by a CRF?*

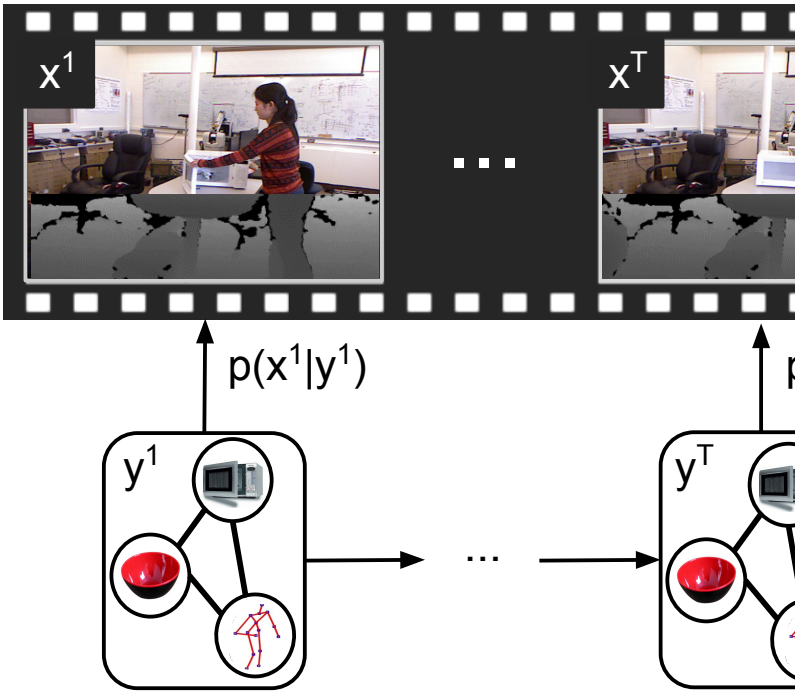


Fig. 1: Figure is showing the state and measurements at each time represented by a CRF. Our algorithm, rCRF, enables the application of recursive Bayesian estimation to CRF-based scene models. rCRF computes the full belief over human activity and object affordances $(y^1, \ldots, y^{T+M})$ by using RGB-D Video $(x^1, \ldots, x^T)$.

Bayesian filtering methods can accurately estimate a belief (set of probabilities) over variables of interest from sequential data. However, it is still very challenging to estimate a belief over a CRF for two reasons. Firstly, it is not tractable to enumerate the labels over a CRF model since the output space has a dimension exponential in the number of objects, labels, and the temporal length[1]. Secondly, there is a modelling difference between CRFs and Bayesian filtering framework. CRF is based on a discriminative setting whereas the Bayesian filtering mostly relies on the generative formulation.

In this paper, we present a recursive algorithm – Recursive

[1]Typically with 10 objects, 10 min. length (with 1 sec. long segments), 10 activity and 10 object labels, dimension is $(10^{10} \times 10)^{10\times 60} = 10^{6600}$.

CRF (rCRF) – which can efficiently estimate a full belief over a CRF-based temporal scene model. rCRF can be seen as an efficient belief estimation method which enables us to use CRF-based scene model in Bayesian filtering. It models the temporal evolution via Bayesian updates and models the measurements in the scene via CRF. In order to use CRFs in such a scenario, we solve two problems. First, we present an approximation to convert the discriminative likelihood of the CRF into a generative measurement equation. Second, we use structured diversity for tractable computation. To the best of our knowledge, rCRF is the only tractable method that can use a CRF-based scene model in a recursive Bayesian filtering.

We apply the rCRF to the problem of activity detection and anticipation from RGB-D data. As a CRF-based scene model, we use the model from [**?**] which represents the scene as a CRF over human activity and object affordances. We then use the RGB-D video to detect and anticipate activities via rCRF.

Our experiments show that we outperform the state-of-the-art methods for detection and anticipation, and the improvement in the anticipation accuracy is significant. In addition to the improvements in accuracy, we show that our anticipation also improves the computation time and runs near real-time.

In summary, the contributions of this work are:

- We present Recursive-CRF (rCRF) method that uses the rich modeling power of CRF in Bayesian filtering setting.
- We present a structured-diversity based approach to enable tractable computation of the belief.
- We apply our rCRF method to the problem of activity detection and anticipation in RGB-D videos.

II. Background and Related Work

Bayesian Recursive Filtering: Estimating a belief over variables of interest from partial observations is a widely studied problem [**?**]. Sequential Monte Carlo (SMC) —aka *particle filter*— is typically used to estimate beliefs in high-dimensional cases. SMC methods represent the belief as a set of samples and we refer the reader to [**?**] for rigorous analysis.

SMC methods are not directly applicable to spaces like CRF since the number of samples required is intractably high. One solution to this problem is the Rao-Blackwellised particle filter [**?**]. It uses a partition of the state variables $\mathbf{y}$ into two set of variables $\mathbf{y_1}$ and $\mathbf{y_2}$ such that the variables in one partition $\mathbf{y_2}$ can be estimated using the partition $\mathbf{y_1}$. Then Rao-Blackwellised particle filter [**?**] estimates the $\mathbf{y_1}$ via SMC and directly estimates $\mathbf{y_2}$ using $\mathbf{y_1}$. However, for our problem, we are not aware of any state decomposition which enables Rao-Blackwellised particle filter. Although there are discrimantive extensions of Bayesian models like recursive least squares[**?**], in this paper we only consider the states represented by CRFs. Moreover, we are not aware of any Bayesian smoothing formulation applied over CRFs.

One tractable application of the SMC framework to the CRF based scene analysis problems is the ATCRF [**?**] model. ATCRF [**?**] uses a set of heuristics to sample the particles. However, ATCRF faces the problem of computational limitations and requires computationally intractable number of samples for anticipation. We follow the Bayesian filtering theory and efficiently estimate the belief.

Structured Diversity and Variants of CRFs: CRFs are widely used to solve activity analysis problems [**?** **?**] in a discriminative setting. CRF models the conditional likelihood of the state given the observations, and the MAP solution can be found. Although this setting is powerful, it does not give any information about the belief other than the MAP state.

Other than the MAP solution, it is also tractable to compute the modes of the CRF [**?** **?** **?**]. These modes can be considered as an approximate state space, and the belief can be computed only for them. Indeed, this claim is empirically validated in many problems like parameter learning [**?**], empirical MBR [**?**] and discrimantive re-ranking [**?**].

Among the aforementioned approaches, Div-M-Best [**?**] is a method applicable to the sequential information. [**?**] starts by dividing the video into a set of frames and computes the diverse-most-likely solutions of each frame independently. Then, it combines the results via the temporal relations. On the contrary, we formulate the problem as recursive Bayesian smoothing and compute the samples based on temporal relations. Formally, given state variables $\mathbf{y}^1, \ldots, \mathbf{y}^T$ and observations $\mathbf{x}^1, \ldots, \mathbf{x}^T$, we directly sample $p(\mathbf{y}^t|\mathbf{x}^1, \ldots, \mathbf{x}^T)$, whereas, [**?**] samples $p(\mathbf{y}^t|\mathbf{x}^t)$. Since our sampling procedure uses the entire video, our samples are more accurate.

There are variants of CRFs that rely on sequential models as well such as, Dynamic CRF (dCRF) [**?**], Infinite Hidden CRF [**?**], Gaussian Process Latent CRF [**?**] and Hierarchical Semi-Markov CRF (HSCRF). Although they are applicable to videos, we are not aware of any tractable method to compute a belief over any of the aforementioned graphical model.

DCRF [**?**] learns the observation likelihood $-p(\mathbf{x^t}|\mathbf{y^t})-$ by using the low-dimensional nature of the features and follows Bayesian filtering. Since our features have very high-dimension (for N objects, we have $58N + 20N^2 + 103$ dimensional features), DCRF [**?**] is not directly applicable. However, it is possible to learn $p(\mathbf{y^t}|\mathbf{x^t})$ and *approximately* use the DCRF formulation by assuming observation and label likelihoods are equal. Moreover, This approach can be shown equivalent to finding local maximum of energy function defined by [**?**] following the formulation of Fox et al [**?**].

It is also common to compute a belief over latent nodes as in the case of infinte hidden CRF [**?**] and Gaussian Process Latent CRF[**?**]. However, they are not directly applicable to our problem since they can compute a belief only over the latent node. CRF-Filter [**?**] is a closely related approach which uses CRFs in a particle filtering scenario. However, it is based on sampling of a low dimensional state space and it is not applicable to our rich model either.

Human Activity Detection and Anticipation: Early works relied solely on human poses. These works range from jointly segmenting and recognizing sub-activities [**?** **?**] to choosing a relevant model out of activity models [**?**]. Main limitation of these methods is that they do not use the object information. Some methods successfully model and use the relations of the

human-poses and objects in the scene [**?** **?** **?** **?**]. However, a significant drawback of these works is missing the fact that object affordance is more important than object types for activities [**?**]. Indeed, object affordance based models had higher performance (e.g., [**?**]). A recent work modelled human activities with latent models [**?**] and also handled the disagreements among the activity annotations [**?**].

Another drawback of these methods is the requirement of the entire activity. Detecting the activity in its early stages is especially crucial for assistive robotics and surveillance systems. Although a few recent work adress the problem of activity detection with partial/early information [**?** **?**], these works do not perform anticipation. There are a few recent works addressing *what human will perform next* by using trajectory prediction. It is possible to predict the trajectory of the human using inverse reinforcement learning in 2D [**?** **?** **?**] or 3D [**?**]. However, these models rely on the low-dimensional structure of the 2D/3D coordinate space and therefore they do not apply to rich models like CRF.

Recent work on anticipatory temporal CRF [**?**] considers an anticipation with a CRF model. It anticipates the future via augmenting set of possible future observations to the CRF. It is also extended with an improved human motion model based on a Gaussian process [**?**]. However, their accuracy significantly drops for a long anticipation horizon since they fail to represent the uncertainty. Our method overcomes these problems by recursively estimating a full belief.

III. Overview

In this section, we summarize our method and explain how we estimate the full belief over the activities and object affordances. Moreover, we also give an illustrative example of the rCRF with a toy scene consisting of two objects (a microwave and a bowl) and a human in Figure **??**.

Reasoning about activities requires not only identifying the objects but also interpreting object-object relations and human-object relations. Indeed, we capture such rich information via CRF. As shown in the Figure **??**, each object and a human corresponds to a node in the graph on which we define the CRF. As a hidden variable, we are interested in object affordances such as *openable, graspable, movable, etc.*, and the activity human is performing such as *moving, opening, grasping, etc.*. We define the affordances as the actions that can be performed on/with the object [**?**]. We denote the affordance variables at time t as $\mathbf{O}_1^t, \ldots, \mathbf{O}_N^t$ for N objects and the activity variable as $\mathbf{A}^t$. Since they are not directly observed, we estimate them by using partial observations. We are using the 3D positions of the objects $\mathbf{L}_1^t, \ldots, \mathbf{L}_N^t$ and the human pose $\mathbf{H}^t$ as observations. The input video is temporally over-segmented prior to the application of the belief estimation, and the time instant t represent the t^{th} segment of the video. We explain the features and the potential functions we use while defining the CRF in Section **??**.

In addition to the spatial relations between objects and humans, we are also interested in their temporal evolution. In general, the problem of estimating a belief over set of hidden variables using the entire video corresponds to a Bayesian smoothing problem. Formally, we are interested in estimating states $\mathbf{y}^t = (\mathbf{O}_1^t, \ldots, \mathbf{O}_N^t, \mathbf{A}^t)$ given set of observations $\mathbf{x}^t = (\mathbf{L}_1^t, \ldots, \mathbf{L}_N^t, \mathbf{H}^t)$. We estimate the states through successive application of the recursive Bayesian updates. In order to tractably compute the Bayesian updates, we introduce two approximations in Section **??**. First, we compute the set of all plausible future states by using structured-diversity. Second, we use Jensen inequality in order to convert the discriminative likelihood into a generative one. After the introduction of these two machineries, we follow the recursive Bayesian estimation framework. As shown in Figure **??**, we first compute the Bayesian updates through the forward and backward messages, $\alpha^t(\mathbf{y}^t)$ and $\beta^t(\mathbf{y}^t)$. We then compute the posterior belief $p(\mathbf{y}^t|\mathbf{x}^1, \ldots, \mathbf{x}^T)$ by using the computed messages and the CRF-likelihood $p(\mathbf{y}^t|\mathbf{x}^t)$. As a final step of the iteration, we represent the belief via diverse samples of the posterior belief. Since the belief is recursively defined, we re-compute the messages and re-sample the belief until it converges.

IV. Belief Estimation with rCRF

In this section, we develop the Recursive Conditional Random Field (rCRF) to use CRF in a Bayesian filtering setting. rCRF jointly uses rich model of CRF and the recursive nature of the Bayesian filtering to compute an accurate belief. We first define our modelling assumptions in Section **??**, and then we introduce a link between the CRF likelihood and the measurement likelihood in Section **??** in order to compute the posterior belief. In Section **??**, we further show that the resulting posterior belief is equivalent to a CRF. Moreover, this equivalence enables efficient computation via the diversity based method [**?**] developed for CRFs.

A. Recursive Conditional Random Field

Consider a sequential estimation problem in which we are interested in variables $\mathbf{y}^t$ using observations $\mathbf{x}^t$ where t is the temporal variable. In our application, t is the temporal segment id. We note RGB-D camera reading as $\mathbf{x}^t$, and object and activity labels as $\mathbf{y}^t$. We now define the Recursive Conditional Random Field (rCRF) framework for such a problem following the assumptions of Hidden Markov Models.

Definition 1: Let $\mathcal{G}^t = (V^t, E^t)$ be set of graphs indexed by the temporal variable t and $\mathbf{y}^t$ is indexed by the vertices of $\mathcal{G}^t$ as $\mathbf{y}^t = (y_v^t)_{v \in V^t}$. Then, ($\mathbf{x}^{1\ldots T}$,$\mathbf{y}^{1\ldots T}$) is a ***Recursive Conditional Random Field*** with dynamics $p_v(\cdot|\cdot)$ when

1) For each t, $(\mathbf{y}^t, \mathbf{x}^t)$ is a CRF over $\mathcal{G}^t = (V^t, E^t)$
2) $p(\mathbf{y}^t|\mathbf{y}^1, \ldots, \mathbf{y}^{t-1}) = p(\mathbf{y}^t|\mathbf{y}^{t-1}) \quad \forall t$ (Markov)
3) $p(\mathbf{x}^t|\mathbf{y}^1, \ldots, \mathbf{y}^t, \mathbf{x}^1, \ldots, \mathbf{x}^{t-1}) = p(\mathbf{x}^t|\mathbf{y}^t) \quad \forall t$
4) $p(\mathbf{y}^t = \mathbf{y}|\mathbf{y}^{t-1} = \mathbf{y}') = p_v(\mathbf{y}|\mathbf{y}')$ (stationarity)

■

We visualize the graphical model representation of the rCRF in Figure **??**. In this work, we are interested in the belief over state variables at a given time instant t as:

$$bel^t(\mathbf{y}) = p(\mathbf{y}^t = \mathbf{y}|\mathbf{x}_1, \ldots, \mathbf{x}_T) \tag{1}$$

Here, T denotes the length of the video. Hence, in rCRF the belief of any frame is supported by the entire video. Moreover, the time instant t can be greater than the video length T as well. Hence, rCRF naturally supports anticipation setting.

We then decompose the belief by using the independence properties of the rCRF as:

$$bel^t(\mathbf{y}) \propto \underbrace{p(\mathbf{y}^t = \mathbf{y}|\mathbf{x}^1, \ldots, \mathbf{x}^t)}_{\alpha^t(\mathbf{y})} \underbrace{p(\mathbf{x}^{t+1}, \ldots, \mathbf{x}^T|\mathbf{y}^t = \mathbf{y})}_{\beta^t(\mathbf{y})} \tag{2}$$

Moreover, α^t and β^t can be computed recursively by using forward and backward messages. Following [**?**],

$$\begin{aligned} \alpha^t(\mathbf{y}^t) &= p(\mathbf{x}^t|\mathbf{y}^t) \sum_{\mathbf{y}^{t-1}} \alpha^{t-1}(\mathbf{y}^{t-1}) p(\mathbf{y}^t|\mathbf{y}^{t-1}) \\ \beta^t(\mathbf{y}^t) &= \sum_{\mathbf{y}^{t+1}} p(\mathbf{x}^{t+1}|\mathbf{y}^{t+1}) \beta^{t+1}(\mathbf{y}^{t+1}) p(\mathbf{y}^{t+1}|\mathbf{y}^t) \end{aligned} \tag{3}$$

with initializations $\alpha^1(\mathbf{y}^1) = p(\mathbf{x}^1|\mathbf{y}^1)$ and $\beta^T(\mathbf{y}^T) = 1$.

B. Computing the belief using an rCRF

Recursive definition in (**??**) has two significant drawbacks: firstly, CRF is modelling $p(\mathbf{y}^t|\mathbf{x}^t)$ instead of $p(\mathbf{x}^t|\mathbf{y}^t)$ and the transformation is not trivial. Secondly, computation of the messages require a summation over the entire output space, and it has an exponential dimension. In this section, we first compute the posterior of the observation given labels $p(\mathbf{x}^t|\mathbf{y}^t)$ by using the CRF posterior likelihood $p(\mathbf{y}^t|\mathbf{x}^t)$. Then, we show that the belief function at time t, $bel^t(\mathbf{y})$, can be approximately represented as a Gibbs measure over $\mathcal{G}^t$. Then, we conclude that the belief, $bel^t(\mathbf{y})$, is a CRF over the graph $\mathcal{G}^t$ with modified energy functions.

1) From $p(\mathbf{y}^t|\mathbf{x}^t)$ to $p(\mathbf{x}^t|\mathbf{y}^t)$: Since $(\mathbf{x}^t, \mathbf{y}^t)$ is a CRF, the posterior of the label given the observation follows [**?**];

$$p(\mathbf{y}^t|\mathbf{x}^t) \propto \exp\left(\sum_{i \in V^t} \theta_{x_i^t}(y_i^t) + \sum_{i,j \in E^t} \theta_{x_i^t, x_j^t}(y_i^t, y_j^t) \right) \tag{4}$$

where θ is the energy function defined over the node set $v \in V^t$ as θ_v and over the edge set $(u, v) \in E^t$ as $\theta_{u,v}$.

In order to transform $p(\mathbf{y}^t|\mathbf{x}^t)$ into $p(\mathbf{x}^t|\mathbf{y}^t)$, we use Bayes rule; $p(\mathbf{x}^t|\mathbf{y}^t) \propto \frac{p(\mathbf{y}^t|\mathbf{x}^t)}{\sum_{x^t} p(\mathbf{y}^t|\mathbf{x}^t) p(\mathbf{x}^t)}$ and compute $p(\mathbf{y}^t)$ as;

$$p(\mathbf{y}^t) = \sum_{\mathbf{x}^t} \exp\left(\sum_{i \in V^t} \theta_{x_i^t}(y_i^t) + \sum_{i,j \in E^t} \theta_{x_i^t, x_j^t}(y_i^t, y_j^t) \right) p(\mathbf{x}^t) \tag{5}$$

For tractability, we approximate the $p(\mathbf{y}^t)$ with its lower bound after applying the Jensen inequality as;

$$p(\mathbf{y}^t) \approx \exp\Big(\sum_{i \in V^t} \underbrace{\sum_{\mathbf{x}^t} \theta_{x_i^t}(y_i^t) p(\mathbf{x}^t)}_{\tilde{\theta}(y_i^t)} + \sum_{i,j \in E^t} \underbrace{\sum_{\mathbf{x}^t} \theta_{x_i^t, x_j^t}(y_i^t, y_j^t) p(\mathbf{x}^t)}_{\tilde{\theta}(y_i^t, y_j^t)} \Big) \tag{6}$$

We then estimate the inner summations $\tilde{\theta}(\cdot)$ from the training data using Monte Carlo method as $\tilde{\theta}(\cdot) = \frac{1}{N} \sum_{i=1}^{N} \theta_{\mathbf{x}^{(i)}}(\cdot)$ where N is the number of training samples and $\mathbf{x}^{(i)}$ is the i^{th} training sample. Therefore, we can compute the observation likelihood as: $p(\mathbf{x}^t|\mathbf{y}^t) \propto$

$$\exp\left(\sum_{i \in V^t} \theta_{x_i^t}(y_i^t) - \tilde{\theta}(y_i^t) + \sum_{i,j \in E^t} \theta_{x_i^t, x_j^t}(y_i^t, y_j^t) - \tilde{\theta}(y_i^t, y_j^t) \right) \tag{7}$$

2) Belief is a CRF: Here we compute the belief (**??**) in terms of forward and backward messages and CRF likelihood. We then show that the posterior belief is a CRF. This observation enables us to use efficient methods developed for CRFs.

In order to compute the belief (**??**), we decompose the system dynamics using the independence assumption in the graph in Fig. **??**. This gives us $p(\mathbf{y}^t|\mathbf{y}^{t-1}) = \prod_i p(y_i^t|y_i^{t-1})$. We then compute the belief function as $bel(\mathbf{y}^t) = \alpha^t(\mathbf{y}^t)\beta^t(\mathbf{y}^t)$ by using equations (**??**) and (**??**). After algebraic manipulations, the belief function can be approximated as follows (see supplementary material for a detailed derivation):

$$\begin{aligned} bel(\mathbf{y}^t) \propto \exp\Bigg[& \sum_{i,j \in E^t} \Big(\theta_{x_i^t, x_j^t}(y_i^t, y_j^t) - \tilde{\theta}(y_i^t, y_j^t) \Big) \\ & \sum_{i \in V^t} \Bigg(\theta_{x_i^t}(y_i^t) - \tilde{\theta}(y_i^t) + \sum_{\mathbf{y}^{t-1}} \alpha^{t-1}(\mathbf{y}^{t-1}) \log p(y_i^t|y_i^{t-1}) \\ & + \frac{1}{\gamma} \sum_{\mathbf{y}^{t+1}} \beta^{t+1}(\mathbf{y}^{t+1}) p(\mathbf{x}^{t+1}|\mathbf{y}^{t+1}) \log p(y_i^{t+1}|y_i^t) \Bigg) \Bigg] \end{aligned} \tag{8}$$

where $\gamma = \sum_{\mathbf{y}^{t+1}} \beta^{t+1}(\mathbf{y}^{t+1}) p(\mathbf{x}^{t+1}|\mathbf{y}^{t+1})$

One property to observe is the decomposition of the belief over the graph. Resulting belief function, (**??**), is a summation over energy terms defined over nodes $i \in V^t$ and edges $i, j \in E^t$. Hence, belief $bel^t(\cdot)$ is a Gibbs measure over $\mathcal{G}^t$. By using Hammersley-Clifford theorem [**?**], we conclude that the posterior belief in rCRF is also a CRF. In other words, belief is a CRF defined over the same graph with a modified energy.

3) Belief via Diverse-Most-Likely Samples: Since we computed the belief function and showed that it is equivalent to a CRF, we now need an efficient method for computing it.

We follow the observation that CRF-likelihood over a natural scene concentrates on a few diverse samples [**?**] because each scene only has a few plausible explanation. So, we compute the belief for only those samples. In other words, let's assume the set of all plausible solutions at time t is $\mathbf{Y}^t = \mathbf{y}^{t,1}, \ldots, \mathbf{y}^{t,M}$ where $\mathbf{y}^{t,i}$ is the i^{th} sample at time t. We then redefine the belief as;

$$\text{approx\_bel}^t(\mathbf{y}) = \begin{cases} \frac{bel^t(\mathbf{y})}{\sum_{\mathbf{y}' \in \mathbf{Y}^t} bel^t(\mathbf{y}')} & \text{if } \mathbf{y} \in \mathbf{Y}^t \\ 0 & \text{o.w.} \end{cases} \tag{9}$$

Since there are only a few plausible explanation of a visual observation and CRF-based belief concentrates only on those samples, proper selection of the samples $\mathbf{Y}^t$ is expected to work well in practice. These samples are typically selected as the diverse-most-likely solutions of the CRF. They are most-likely samples because we are only interested in the plausible

explanations. They are diverse because we are interested in the modes of the CRF other than set of samples around the MAP solution. Diversity is achieved via asserting samples to be at least δ unit apart from each other via the distance function Δ (we use hamming distance as a in our experiments). In other words, we solve the following optimization problem in order to get the samples which represent the belief;

$$\begin{aligned} \mathbf{y}^{t,i} &= \arg\max_{\mathbf{y}} bel^t(\mathbf{y}) \\ & s.t.\ \Delta(\mathbf{y}, \mathbf{y}^{t,j}) \geq \delta \quad \forall\, j < i \end{aligned} \tag{10}$$

This optimization is NP-hard in general; however, since we already showed $bel^t(\mathbf{y})$ is CRF, we use the existing diverse-m-best algorithms developed for CRFs. We use the Lagrange relaxation by Batra et al. [**?**]. We explain the details of solving this problem by using [**?**] in supplementary material.

In summary, we first compute the belief via (**??**) for all frames by using samples of the previous and the next frame as well as CRF likelihoods. Then, we compute the diverse samples of (**??**) by using [**?**]. After computing the samples, we compute the messages α^t and β^t by using the equations (**??**) and (**??**). We continue to re-sample the beliefs and re-compute the messages recursively until the convergence. Moreover, during the initialization, we only sample the observation function (**??**) since the messages are not available.

V. Human Activity Detection and Anticipation

In this section, we describe how we apply the rCRF framework to RGB-D videos for human activity detection and anticipation. We are interested in activities such as *reaching* and *moving*, and object affordances such as *reachable* and *movable* as explained in Section **??**. We follow the approach in [**?**], and start with temporally segmenting the video. This step can be considered as an oversegmentation in the temporal domain. It decreases the computation complexity and enables using motion information as an observation.

We then obtain the observations $\mathbf{x}^t = (L_1^t, L_2^t, H^t)$, by detecting the objects in the first frame and then tracking them. We obtain the human pose H^t through a skeleton tracker. We consider affordances and activities as state $\mathbf{y}^t = (O_1^t, \ldots, O_N^t, A)$ where N is the number of objects. We extracted set of features from the observations following the feature functions in [**?**] (*eg.* relative and absolute location of objects, human joints and their temporal displacements). After extracting the features, we define our CRF as a log-linear CRF and learn the energy function defined in (**??**) by using the Structural SVM [**?**] as in the case of [**?**]. We use the first order statistics for temporal dynamics as $p_v(y, y') = p(Y_v^t = y | Y_v^{t-1} = y') = \frac{\#(Y_v^t = y, Y_v^{t-1} = y')}{\#(Y_v^t = y')}$ where $\#(\cdot, \cdot)$ is number of the co-occurrence in training data.

After defining the observation, state and dynamics, we apply the rCRF framework. We also summarize the activity detection and anticipation application in Algorithm **??**.

Moreover, since the temporal relations are modeled as causal, we do not compute the backward messages during the anticipation. In anticipation, there is also no future observation.

Algorithm 1 Compute belief the over $(O_{1\ldots N}^t, A^t)$ for $t \in [1, T+\tau]$ in an RGB-D Video of length T

Initialization:
Compute $L_1^t, \ldots, L_N^t$, and H^t for $t \in [1, T]$ via [**?**].
Compute $p(L_{1\ldots N}^t, H^t | O_{1\ldots N}^t, A^t)$ for $t \in [1, T]$ via (**??**)
Compute the belief via (**??**) w/o messages ($\alpha = 1, \beta = 1$)
Detection:
repeat
  for $t \in [1, T]$ **do**
    Compute the forward/backward messages via (**??**)
    Compute the belief via (**??**) an sample via (**??**)
  end for
until convergence or number of iterations limit
Anticipation:
for $t \in [T+1, T+\tau]$ **do**
  Compute only the forward messages via (**??**)
  Sample the belief directly from the forward messages.
end for

Hence, the belief is defined solely by the forward messages. In order to compute the belief for future frames, we propagate the estimated belief. We propagate the belief to the next frame by sampling the next state of the each sample in the belief of the current frame via the temporal dynamics. Then, we choose diverse most likely samples out of the propagated samples via solving (**??**) with exhaustive search.

VI. Experimental Results

In order to experimentally evaluate the proposed rCRF model and the belief computation, we perform experiments on two applications. Firstly, we estimate a belief over the activity a human is performing and the affordances of the objects in the scene by using the RGB-D video. After computing the belief, we detect the most likely activity and affordance sequences and study the improvement in the detection accuracy. Secondly, we test the accuracy of the beliefs in the anticipation setting. Indeed, we show that it is possible to obtain high-quality detection and anticipation via rCRF.

Data: We use CAD-120 [**?**] dataset in order to evaluate our method. CAD-120 dataset includes 120 RGB-D videos of four different subject performing activities *reaching, moving, pouring, etc.* while interacting with objects having affordances *reachable, movable, pourable, etc.*. There are 10 activity classes and 12 object affordance classes.

Experimental Setup: For computing the features and learning the CRF parameters, we follow the approach and the code in [**?**]. Following the convention in [**?**], we use 4-fold cross-validation by training over the data from 3 subjects and testing on the remaining subject. We then average the results over 4-folds. We implemented the rCRF as we explain in Algorithm **??** with the following parameters obtainded via cross-validation; we sampled $M = 15$ diverse samples and ran the recursive message updates with the number of iterations limit as 5.

For the anticipation setting, In order to experiment the τ seconds into the future anticipation, we experiment over all

feasible anticipation scenarios. In other words, we anticipated the time instant $t+\tau$ by using the segments $1 \dots t$ for all $t < T - \tau$, where T is the length of the video. Then, we averaged the score over all feasible experiments.

Baseline Algorithms: In detection setting, we compare the detection results of the rCRF to MAP solution of the spatiotemporal CRF in [**?**]. We also included the state-of-the art activity detection results from Hu et al. [**?**]. Moreover, [**?**] is not based on object affordances and it only outputs activity detections. For the anticipation, we compare the rCRF with the state-of-the-art anticipation methods ATCRF [**?**] and GP-LCRF[**?**]. We also include DCRF[**?**]. In order to evaluate the contribution of the recursive modeling and the structured diversity separately, we also compare the rCRF with a recursive approach without diversity and a diversity-based approach without recursive modeling baselines.

The DivMBest algorithm in [**?**] uses the diverse sampling method to sample CRFs defined over each frame separately. DivMBest[**?**] then finds the most likely sequence via Viterbi algorithm. Since it is missing the recursive modeling, it serves as *structured diversity without recursive filtering* baseline. We replace the diversity-based sampling in our method with Gibbs sampler and consider it as *recursive filtering approach without structured diversity* baseline. For the Gibbs sampling, we sampled 50 samples per temporal segment. We denote the recursive approach with Gibbs sampling as *"rCRF w/o div"* while tabulating the results.

Evaluation Metrics: For activity detection, we compute the ratio of the correctly classified labels (*micro precision*) and the averages of the precision and recall values computed for each activity and object affordance classes (*macro precision* and *macro recall*). For anticipation, we record the ratio of the correctly classified labels *micro precision*, the average of the f-1 score that is computed for each activity and object affordance class (*macro f-1 score*), and the precision of the top 3 anticipated labels (*robot anticipation metric*). While computing the *robot anticipation metric*; if any of the top 3 anticipation is correct, it is counted as true positive.

Accuracy of the rCRF in detection setting. We evaluate the rCRF for activity detection and summarize the results in Table **??**. Table **??** suggests that the rCRF outperforms the MAP solution [**?**] and performs similarly with the state-of-the-art solution [**?**]. Since rCRF and [**?**] are using the same spatial relations, the performance difference is due to the modeling of the temporal relations in rCRF. We use first-order statistics as temporal dynamics, and they are quite accurate as shown in the heatmap in Figure **??**. They also capture semantic information like objects become stationary after being used.

Accuracy of the rCRF in anticipation setting. We evaluate the accuracy of the belief we compute via rCRF, both quantitatively and qualitatively. For qualitative evaluation, we show the segment that we are anticipating the belief over, as well as the belief we obtain in Figure **??**. Please note that, this visual information is not visible to the algorithm, and it is only included for the subjective evaluation.

As shown in the figure, anticipated belief is capturing the scene accurately. Belief is accurate even for the case of concurrent activities. For example, in the second column of the second row in Figure **??**, subject is reaching the microwave and moving the cleaner. Our method assigns similar likelihood values to both reaching and moving.

We also perform quantitative analysis over anticipation accuracy. We anticipate 3 seconds into the future and summarize the results in Table **??**. As shown in the Table **??**, rCRF outperforms the state-of-the-art heuristic method [**?**] and the GP-LCRF method [**?**] significantly as well as all other baselines. We believe this result is due to the accurate joint-modeling of the temporal relations and the CRF model. We further analysed this behaviour in the subsequent sections.

TABLE I: **Detection Performance over CAD-120.** We compare rCRF with MAP solution and baselines for detections accuracy.

	Sub-activity			Object Affordance		
	micro	macro		micro	macro	
	prec(%)	prec(%)	rec(%)	prec(%)	prec(%)	rec(%)
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
Hu et al.[**?**]	67.8±1.4	65.5±3.5	**63.5±6.6**	N/A	N/A	N/A
MAP Sol[**?**]	63.4±1.6	65.3±2.3	54.0±4.6	79.4±0.8	62.5±5.4	50.2±4.9
DivMBest[**?**]	64.0±1.3	61.7±2.1	56.4±2.7	80.1±1.0	76.2±2.5	53.2±3.2
DCRF[**?**]	61.2±2.1	62.8±2.8	54.3±1.5	71.9±2.9	80.6±2.4	62.5±3.6
rCRF w/o div	61.2±1.8	64.0±1.8	52.7±3.8	75.2±2.4	79.3±3.1	63.7±2.9
rCRF	**68.1±1.3**	**66.1±2.7**	57.2±3.9	**81.5±1.1**	**85.2±2.4**	**71.6±3.9**

TABLE II: **Anticipation performance for the anticipating 3 seconds in the future.** We compare rCRF with state-of-the-art anticipation algorithm and baselines for anticipation accuracy.

	Sub-activity			Object Affordance		
	micro	macro	robot ant.	micro	macro	robot ant.
Method	prec(%)	f1-scr(%)	metric(%)	prec(%)	f1-scr(%)	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
GP-LCRF [**?**]	52.1±1.2	43.2±1.5	76.1±1.5	68.1±1.0	44.2±1.2	74.9±1.1
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
DivMBest[**?**]	47.9±1.4	43.2±3.6	71.5±2.7	61.3±1.4	56.3±2.1	73.3±0.5
DCRF[**?**]	48.3±2.6	35.4±1.8	66.6±1.1	55.2±3.1	48.5±3.1	71.24±2.2
rCRF w/o div	49.6±2.1	39.7±2.6	65.1±1.1	56.2±1.9	47.4±3.1	70.8±2.5
rCRF	**54.3±3.9**	**45.8±2.7**	**76.5±2.6**	**78.7±3.4**	**74.9±3.8**	**82.1±2.9**

How important is the recursive modeling? DivMBest[**?**] is the application of the structured diversity without recursive modeling of the Bayesian filtering. In all experiments (Table **??** and **??**), rCRF outperforms the DivMBest [**?**]. We believe this is because rCRF samples $p(y^t|x^1, \dots, x^T)$ instead of $p(y^t|x^t)$ as in the case of [**?**]. In other words, DivMBest [**?**] samples without considering temporal relations; on the contrary, we sample the full belief directly.

Moreover, the improvement over the DCRF model shows the important of accurate recursive modeling. DCRF uses the recursive modeling without the proposed conversion of the discrimantive likelihood into generative one and it performs poorly. Hence, the proposed conversion is a necessary step.

We also studied the effect of anticipation horizon. We computed precision of all methods for horizons between 1 and 10 seconds and plotted in Figure **??** and **??**. We see significant improvements over longer anticipation time horizons.

In Figure **??** and **??**, accuracy of all algorithms decreases with the increasing horizon. One interesting observation is

decrease rate of DivMBest is steeper than others. Since DivMBest misses the recursive nature of the problem, accuracy of the belief it computes is limited; hence, the resulting belief does not stay informative with increasing horizon.

We further computed the entropy of the belief rCRF computes and plotted its average in Figure **??**. The decrease rate of the accuracy is much smaller than the increase rate of the entropy. In summary, recursive modeling is necessary for an accurate belief estimation and rCRF computes flatter yet still informative beliefs with increasing horizon.

How to efficiently cover the output space? In order to see the effect of structural diversity on covering the output space, we compare the rCRF with a version of it in which we replace diverse sampling with the Gibbs sampler. As expected, Gibbs sampler only sampled the small region around the posterior and failed to cover the output space. Within all experiments, rCRF outperforms Gibbs sampler baseline. Another interesting observation is, as shown in Figure **??**&**??**, although Gibbs sampler based method performed slightly better than other baselines for short horizon activity anticipation, it performed much worse for object affordance. We believe this is because of the dimensionality. Activity space has dimension 10^T whereas the object affordance space has dimension $12^{T \cdot M}$ where T is the length of the video and M is the number of objects. Hence, diversity plays bigger role with increasing dimension. Moreover, [**?**] uses the domain knowledge by selectively sampling points around the hand, etc. and it performs better than both baselines with increasing horizon. We believe this result is due to the efficient coverage of the output space with heuristics.

Computationally-efficient inference: We evaluated the computational efficiency by computing the average computation time for anticipating 3 second in the future via rCRF and the fastest available anticipation algorithm (the ATCRF[**?**]). Within our experiments, we did not include any pre-processing or feature extraction computation (they are same for all algorithms). Our experiments suggest that the rCRF is faster than [**?**] as shown in Table **??**. Hence, rCRF model outperforms the state-of-the-art anticipation algorithm in terms of speed in addition to the accuracy.

TABLE III: Computation time for anticipating 3 seconds in the future excluding pre-processing (*see supplementary material for details*).

ATCRF [**?**]	34.1s	rCRF	1.41s

Can rCRF generalize to RGB data?: Since there is no RGB activity dataset with object labels, it is hard to compare our algorithm in the RGB activity analysis setting. Removing the concept of the objet form the graph, makes it a chain-CRF and the inference and learning becomes straightforward. However, we still implement our rCRF over a linear-chain CRF for RGB activity analysis. We based our implementation on MPII cooking activity dataset [**?**] and use the publicly distributed features from the authors webpage. The shared features are HOG, HOF, dense trajetory features and MBH [**?**].

TABLE IV: Anticipation performance for the anticipating 3 seconds in the future in MPII Cooking Dataset[**?**].

Method	micro prec(%)	macro prec(%)	macro recall(%)
Chance	$1.5_{\pm 0.6}$	$1.5_{\pm 0.6}$	$1.5_{\pm 0.6}$
ATCRF [**?**]	$33.4_{\pm 3.3}$	$52.1_{\pm 4.6}$	$12.1_{\pm 1.4}$
DivMBest[**?**]	$34.4_{\pm 2.8}$	$55.3_{\pm 5.0}$	$14.3_{\pm 1.2}$
rCRF	$\mathbf{37.4_{\pm 2.9}}$	$\mathbf{63.2_{\pm 5.5}}$	$\mathbf{26.1_{\pm 2.6}}$

As shown in the Table **??**, our method outperforms all baselines and competing algorithms. We did not include Gibbs sampling here since the dimension of the activity space is rather low and the experiment over diversity is not informative. We believe this result is due to the accurate handling of temporal information in rCRF and it shows that it generalizes to other modalities.

VII. Conclusions

In this work, we consider the problem of using rich CRF-based scene models in Bayesian filtering setting. We presented the rCRF model which uses rich modelling power of CRFs in recursive Bayesian filtering. We further developed a computationally-tractable method based on Jensen inequality and structured diversity. We performed extensive experiments that show rCRF accurately anticipates the future beliefs over CRFs. We also experimentally demonstrated that the recursive framework significantly improves the accuracy of anticipation. Our rCRF not only resulted in more accurate anticipation but also improved the computation time.

Acknowledgement. We thank Hema Koppula for helpful discussions. This research was funded in part by Microsoft Faculty Fellowship (to Saxena), NSF Career award (to Saxena) and Army Research Office.

References

[] D. Batra, P. Yadollahpour, A. Guzman-Rivera, and G. Shakhnarovich. Diverse m-best solutions in markov random fields. In *ECCV* 2012.

[] K. Bousmalis, S. Zafeiriou, L Morency, and M. Pantic. Infinite hidden conditional random fields for human behavior analysis. *Neural Networks and Learning Systems, IEEE Trans*, 24(1): 170–177, 2013.

[] C. Chen, V. Kolmogorov, Y. Zhu, D. Metaxas, and C. Lampert. Computing the m most probable modes of a graphical model. In *AISTATS* 2013.

[] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR* 2005.

[] P. Del Moral. *Mean field simulation for Monte Carlo integration*. CRC Press, 2013.

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

[] A. D Dragan and S. S Srinivasa. Formalizing assistive teleoperation. *RSS*, 2008.

[] E.B. Fox. *Bayesian Nonparametric Learning of Complex Dynamical Phenomena*. Ph.D. thesis, MIT, Cambridge, MA, 2009.

[] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE PAMI*, (6): 721–741, 1984.

[] J.J. Gibson. *The ecological approach to visual perception*. Psychology Press, 1986.

[] 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.

[] J. M. Hammersley and P. Clifford. Markov fields on finite graphs and lattices. 1971.

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

[] M. Hoai, Z-Z. Lan, and F. De la Torre. Joint segmentation and classification of human actions in video. In *CVPR*, 2011.

[] N. Hu, G. Englebienne, Z. Lou, and B. Kröse. Learning latent structure for activity recognition. In *ICRA*, 2014.

[] N. Hu, Z. Lou, G. Englebienne, and B. Kröse. Learning to recognize human activities from soft labeled data. *RSS*, 2014.

[] Y. Jiang and A. Saxena. Modeling high-dimensional humans for activity anticipation using gaussian process latent crfs. In *RSS*, 2014.

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

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

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

[] H. Koppula, A. Jain, and A. Saxena. Anticipatory planning for humanrobot teams. ISER, 2014.

[] H. S Koppula and A. Saxena. Anticipating human activities using object affordances for reactive robotic response. *RSS*, 2013.

[] H. S Koppula and A. Saxena. Physically grounded spatio-temporal object affordances. In *ECCV*, pages 831–847. Springer International Publishing, 2014.

[] H. S. Koppula, R. Gupta, and A. Saxena. Learning human activities and object affordances from rgb-d videos. *IJRR*, 32 (8):951–970, 2013.

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

[] T. Lan, T.-C. Chen, and S. Savarese. A hierarchical representation for future action prediction. In *Proc. ECCV*, 2014.

[] E. L. Lawler. A procedure for computing the k best solutions to discrete optimization problems and its application to the shortest path problem. *Management Science*, 18(7):401–405, 1972.

[] B. Limketkai, D. Fox, and L. Liao. Crf-filters: Discriminative particle filters for sequential state estimation. In *ICRA*, 2007.

[] J. Mainprice and D. Berenson. Human-robot collaborative manipulation planning using early prediction of human motion. In *IROS*, 2013.

[] P. Matikainen, R. Sukthankar, and M. Hebert. Model recommendation for action recognition. In *CVPR*, 2012.

[] F. Meier, A. Globerson, and F. Sha. The more the merrier: Parameter learning for graphical models with multiple maps. In *ICML Workshop on Interactions between Inference and Learning*, 2013.

[] V. Premachandran, D. Tarlow, and D. Batra. Empirical minimum bayes risk prediction: How to extract an extra few% performance from vision models with just three more parameters. In *CVPR*, 2014.

[] A. Quattoni, S. Wang, L.-P. Morency, M. Collins, T. Darrell, and M. Csail. Hidden-state conditional random fields. *IEEE PAMI*, 29(10):1848–1852, 2007.

[] L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

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

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

[] S. Särkkä. *Bayesian filtering and smoothing*, volume 3. Cambridge University Press, 2013.

[] Q. Shi, L. Cheng, L. Wang, and A. Smola. Human action segmentation and recognition using discriminative semi-markov models. *IJCV*, 93(1):22–32, 2011.

[] C. Sminchisescu, A. Kanaujia, and D. Metaxas. Conditional models for contextual human motion recognition. *CVIU*, 104 (2):210–220, 2006.

[] C. Sutton, A. McCallum, and K. Rohanimanesh. Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data. *JMLR*, 8:693–723, 2007.

[] S. Thrun, W. Burgard, and D. Fox. *Probabilistic robotics*. MIT press, 2005.

[] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector machine learning for interdependent and structured output spaces. In *ICML*, 2004.

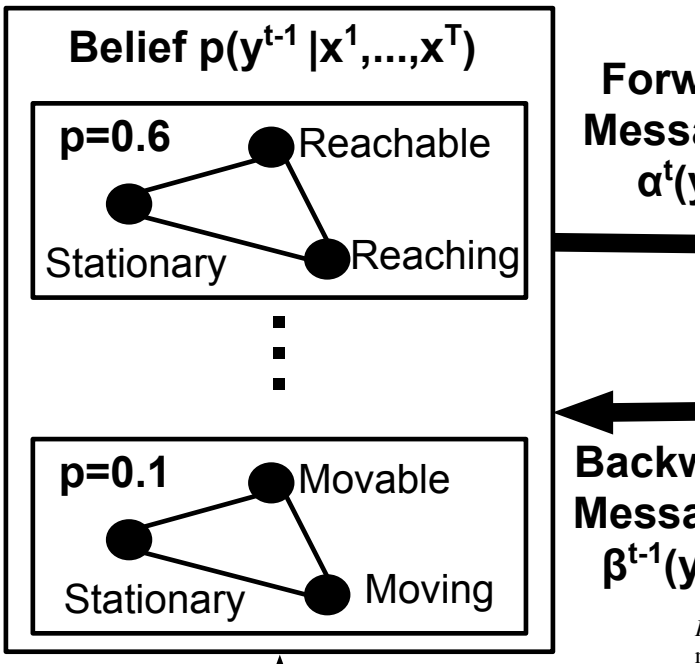
[] Y. Wang and Q. Ji. A dynamic conditional random field model for object segmentation in image sequences. In *CVPR*, 2005.

[] Z. Wang, K. Mülling, M. P. Deisenroth, H. B. Amor, D. Vogt, B. Schölkopf, and J. Peters. Probabilistic movement modeling for intention inference in human–robot interaction. *IJRR*, 32(7): 841–858, 2013.

[] P. Yadollahpour, D. Batra, and G. Shakhnarovich. Discriminative re-ranking of diverse segmentations. *CVPR*, 2013.

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

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



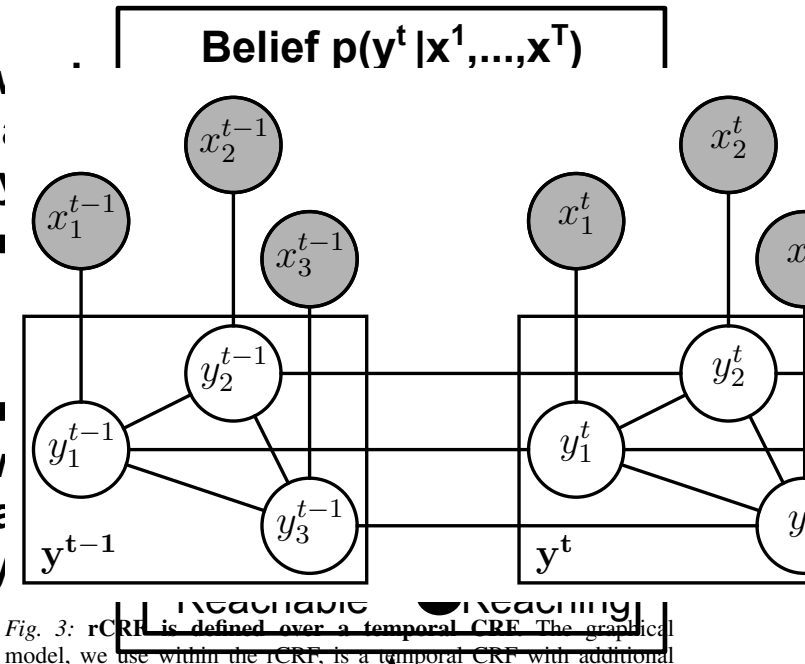
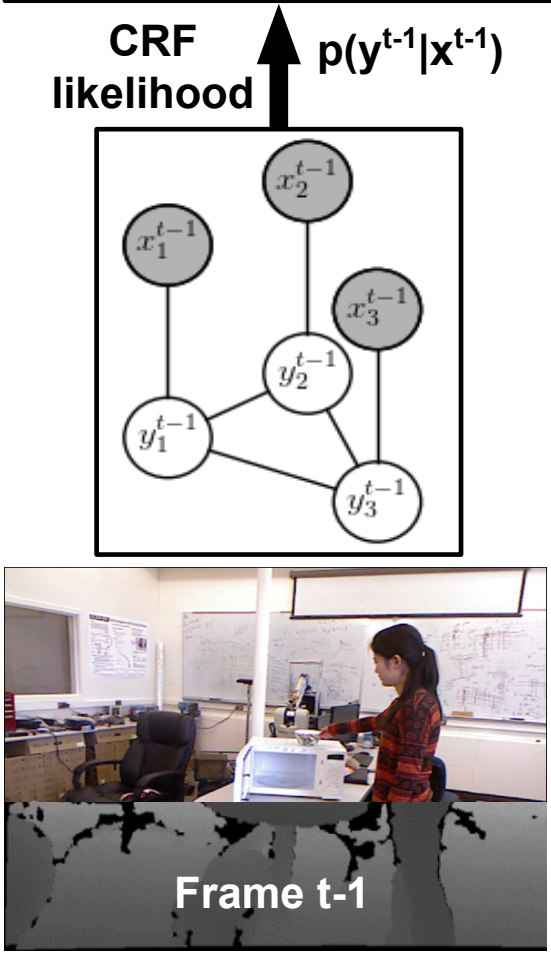
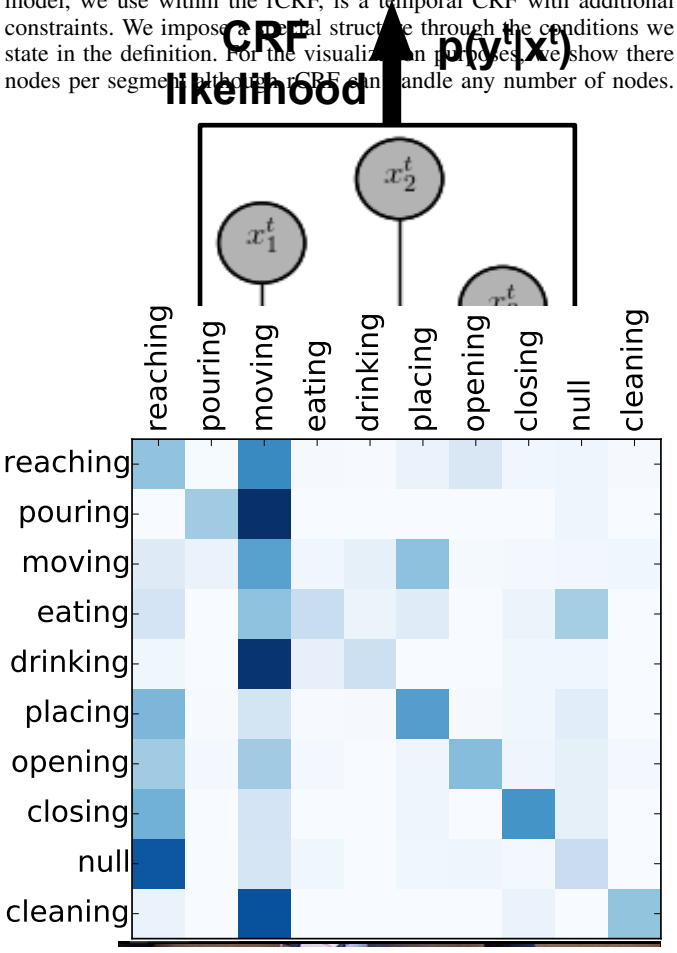


Fig. 3: **rCRF is defined over a temporal CRF.** The graphical model, we use within the rCRF, is a temporal CRF with additional constraints. We impose a spatial structure through the conditions we state in the definition. For the visualization purposes, we show there nodes per segment although the rCRF can handle any number of nodes.





(a) Human Activity

Fig. 2: **Computing the full belief by using rCRF.** Each iteration of the recursive estimation algorithm includes computing forward and backward messages, $\alpha^t(\mathbf{y}^t)$ and $\beta^{t-1}(\mathbf{y}^{t-1})$, by using the current samples and computing the belief $p(\mathbf{y}^t|\mathbf{x}^1,\ldots,\mathbf{x}^T)$ with the computed messages. Then, we re-compute the messages and re-sample the belief until the belief converges. *Here, we only have two objects as* $\mathbf{y}^t = (\mathbf{O}_1^t, \mathbf{O}_2^t, \mathbf{A}^t)$ *and* $\mathbf{x}^t = (\mathbf{L}_1^t, \mathbf{L}_2^t, \mathbf{H}^t)$

Fig. 4: Heatmap of the first-order statistics of activity and object affordance classes. They are used as temporal dynamics by rCRF.

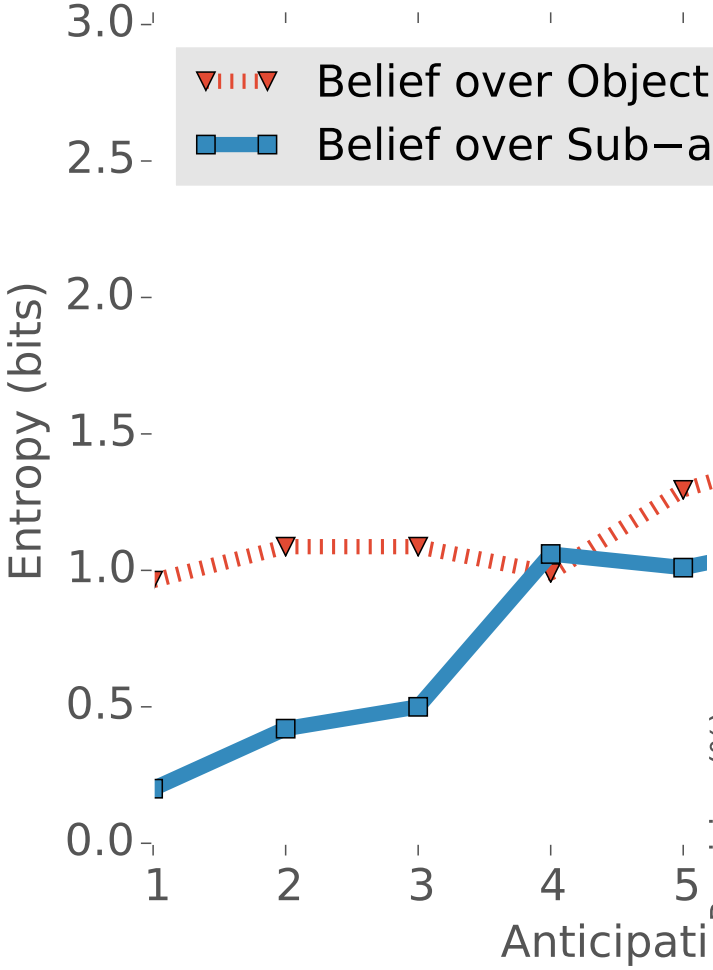


Fig. 5: Entropy of the belief vs. time *(uniform dist. has* ≈ 3.32 *bit entropy)*

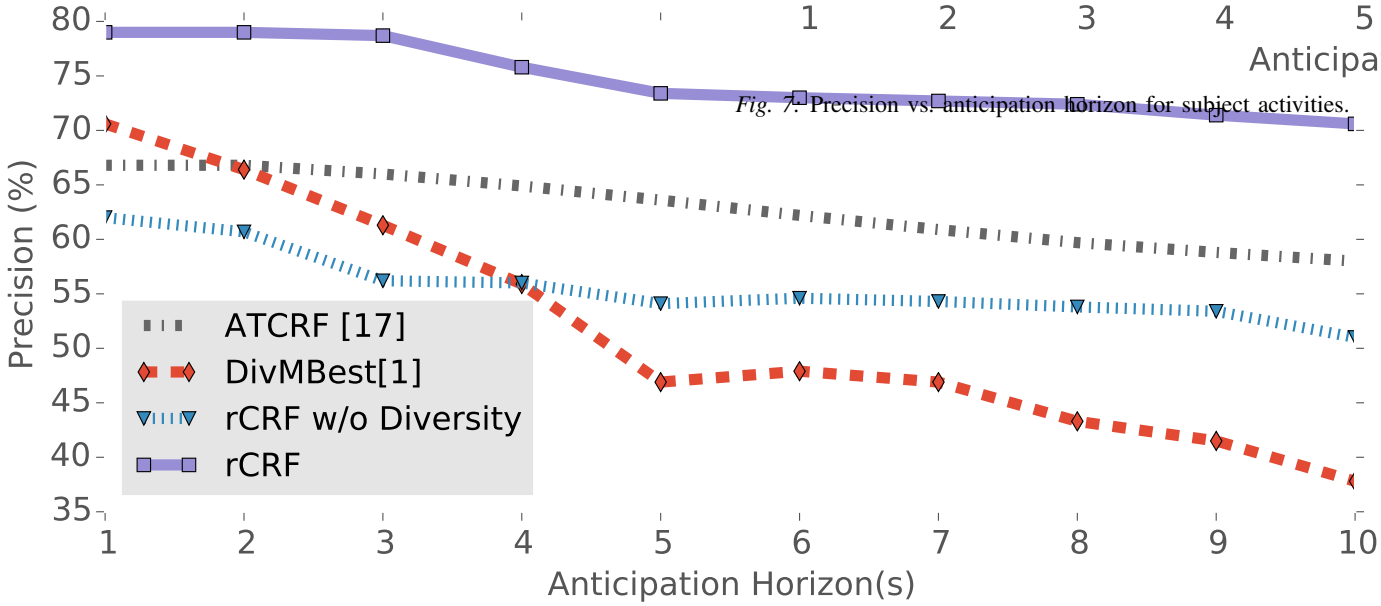


Fig. 6: Precision vs. anticipation horizon for object affordance.

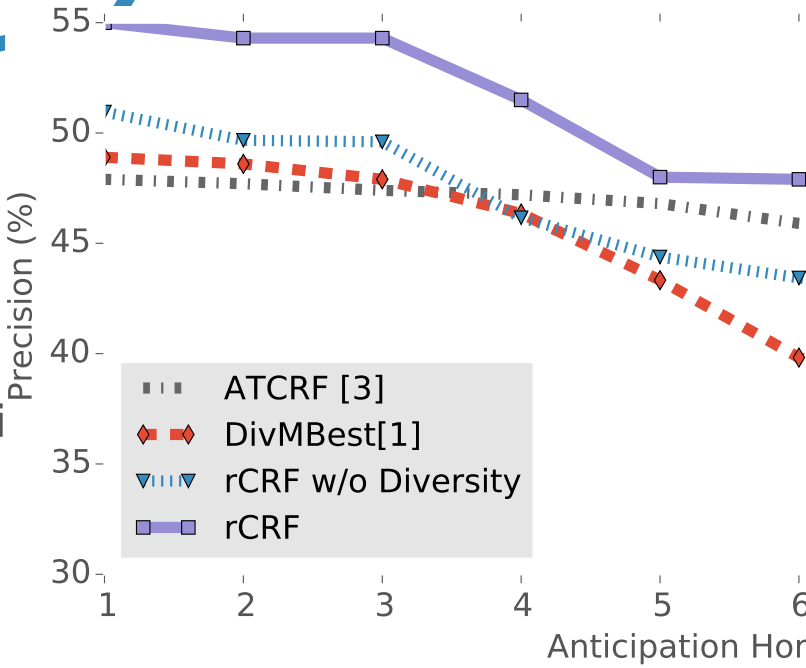


Fig. 7: Precision vs. anticipation horizon for subject activities.