Modeling High-Dimensional Humans for Activity Anticipation using Gaussian Process Latent CRFs

Yun Jiang and Ashutosh Saxena
Department of Computer Science, Cornell University, USA.
Email:{yunjiang,asaxena}@cs.cornell.edu

Abstract—**For robots, the ability to model human configurations and temporal dynamics is crucial for the task of anticipating future human activities, yet requires conflicting properties: On one hand, we need a detailed high-dimensional description of human configurations to reason about the physical plausibility of the prediction; on the other hand, we need a compact representation to be able to parsimoniously model the relations between the human and the environment.**

We therefore propose a new model, GP-LCRF, which admits both the high-dimensional and low-dimensional representation of humans. It assumes that the high-dimensional representation is generated from a latent variable corresponding to its low-dimensional representation using a Gaussian process. The generative process not only defines the mapping function between the high- and low-dimensional spaces, but also models a distribution of humans embedded as a potential function in GP-LCRF along with other potentials to jointly model the rich context among humans, objects and the activity. Through extensive experiments on activity anticipation, we show that our GP-LCRF consistently outperforms the state-of-the-art results and reduces the predicted human trajectory error by 11.6%.

I. Introduction

The ability to anticipate possible future moves of a human is a necessary social skill for humans as well as for robots that work in assembly-line environments (e.g., Baxter) or in homes and offices (e.g., PR2). With such a skill, the robots can work better with humans by performing appropriate tasks and by avoiding conflict. For instance, Koppula et. al. [20] used anticipation in assistive robotic settings, such as in the tasks of opening doors for people or serving drinks to people.

Human activity anticipation is a very challenging task, especially in *unstructured* environments with a large variety of objects and activities. Koppula et. al. [20] have shown that the rich context (such as object-object and human-object spatial relations) is important for predicting high-level human activities. However for anticipation and robotic planning, predicting *detailed* human motions is also crucial. In this work, our goal is to model the detailed human motions, along with the rich context, in anticipating the human activities. We specifically focus on *how to represent (and learn with) high-dimensional human configurations and their temporal dynamics.*

Recently, high-dimensional description of human motions is widely available through motion capture data or RGB-D cameras (e.g., Kinect), where a human configuration is specified by the joint locations and orientations and often has more than 30 degrees of freedom. While it captures human kinematics and dynamics accurately, modeling human motions in such space (much higher than 30 DOF when considering velocities and accelerations) often requires a detailed musculo-skeletal human model and a large number of spatial and timing constraints to produce smooth and realistic motions [4].

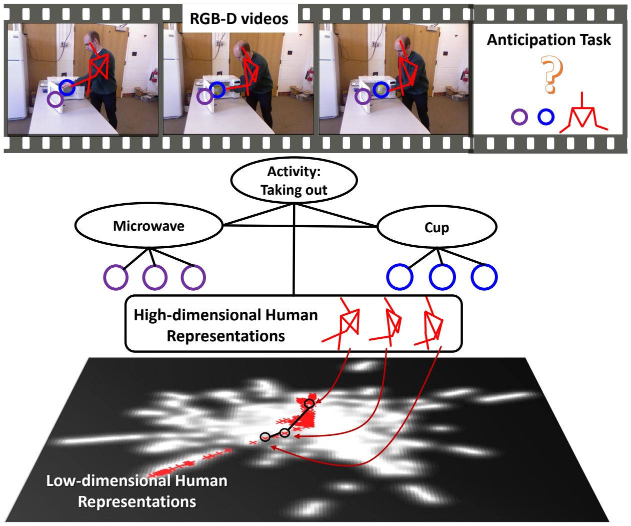

RGB-D videos
Anticipation Task
?
Activity: Taking out
Microwave
Cup
High-dimensional Human Representations
Low-dimensional Human Representations

Fig. 1: Given an RGB-D video of a human interacting with the environment, we are interested in predicting the future: what activity will he perform and how the environment and human pose will change. The key idea in this work is to compactly represent the high-dimensional human configuration in a low-dimensional space so that we can model relations between the human, activity and objects more effectively in a graphical model.

Such a high-DOF model does not lend itself to use in learning models where rich modeling of the human with the environment is needed. Therefore, some works assume a few static human poses are representative enough [6, 8, 13, 12] or simplify a human configuration to a 2D point for navigation task [3, 18, 44, 23] or to a 3D trajectory of one hand while keeping the rest body static neglecting kinematic constraints [20, 19]. In these works, human motions are under-represented and would fail when a more elaborate human motion prediction is required.

In this work, we design a model that can handle the two competing requirements: (a) having access to a high-dimensional model of a human skeleton so that physical feasibility and other kinematic criterion can be reasoned about, and (b) having a low-dimensional representation that allows its use in learning algorithms that model the human's relation to entities in the environment containing objects. The general idea is to learn a two-way mapping between the high-dimensional and low-dimensional representations of the human configurations. Meanwhile, we associate the mapping with a probability distribution over the low-dimensional space so that data-driven learning approaches can use the distribution to build a probabilistic model that captures the relationships between human and other entities in the environment.

We therefore propose a model GP-LCRF (stands for Gaussian Process Latent Conditional Random Field), which is a conditional random field (CRF) augmented with latent nodes in low-dimensional space corresponding to the compressed high-dimensional nodes in the CRF. The correspondence is modeled using a Gaussian process that explicitly defines a mapping function and a likelihood function over the two spaces. The likelihood function is incorporated as the potential function for the edges between the low-dimensional and high-dimensional nodes. In our application to the task of anticipating human activities, GP-LCRF models human configurations as the high-dimensional nodes and learns their compact representation as latent nodes. Edges between latent nodes are used to model the continuity of human motions. Inspired by [20], GP-LCRF also models objects and sub-activities as other nodes and their spatial-temporal relations as edges. In this way, GP-LCRF provides a joint graphical model for humans, objects and sub-activities together.

We show that, during training the model, the likelihood could be decomposed into two disjoint sets, one for learning the mapping of the latent space and the other for learning the parameters of the CRF. During inference, our goal is to anticipate future human actions, for which we use Gibbs sampling for inferring probable human configurations.

We test our model on CAD-120 human activity dataset [22], which contains 120 RGB-D videos of daily human activities such as eating food, cleaning, etc. Through extensive experiments, we show that our GP-LCRF achieves the state-of-the-art results while comparing against multiple baselines with or without modeling humans. Particularly, we reduce the human trajectory prediction error by 11.6% (with a p-value of 0.0107), which could make significant difference to robot planning that uses the predicted future human motions.

In summary, our contributions are:

- We propose a new graphical model, GP-LCRF, that bypasses the difficulty of modeling high-dimensional variables in traditional CRFs by augmenting the graph with latent nodes corresponding to the low-dimensional representations of those variables.
- We apply GP-LCRF to the human activity anticipation task. GP-LCRF is able to capture both the rich context in the environment and the temporal dynamics of humans.
- We test our model on a large human activity dataset (with various objects and activities) and achieve the state-of-the-art results.

II. Related Work

Modeling human configurations has attracted great attention in many computer vision and robotic applications. There is a significant body of work building a full human body model for tracking and reconstruction human motions. For example, Navaratnam et al. [30] learn a mapping from image features to human joint angles to estimate human pose in 2D images. Demircan et al. [4] utilize a detailed subject-customized bio-mechanical model and multiple markers from motion capture data to reconstruct realistic human motions in real-time. Kulić et al. [24] consider clustering human configurations into motion segments. In the field of character animation, human model is also used to synthesize and plan the high-dimensional human motions [37, 28]. There are some works focusing on learning a low-dimensional representation of humans that can be used to interpolate new human motions [34, 7]. While these works can model detailed human configuration well, the high-dimensional representation is not suitable for probabilistic modeling in the problem of human activity anticipation, and thus they are complimentary to ours.

Some works reduce the large space by using a few static poses. Grabner et al. [6] and Gupta et al. [8] utilize imaginary human actors to detect objects and human workspace in 2D images. Jiang et al. [13, 11, 12] apply hallucinated human configurations to a robotic task of arranging 3D indoor scenes. In another previous work of anticipating human activities [20, 19], human motions are implicitly modeled through object trajectories. Some other works [3, 18, 44, 23] predict possible human navigation trajectories in 2D from visual data. In all the aforementioned works, human motions are under-represented and would fail when a more elaborate prediction on human movements is required.

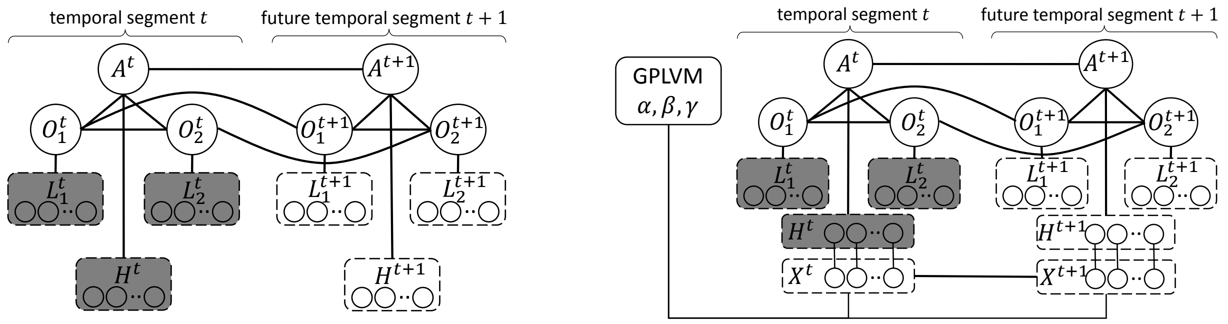
Some other works consider human robot collaborations without anticipating human activities [1], focus on high-level actions [31], or consider object affordances for manipulation [16]. These works are orthogonal to ours.

In terms of capturing the context in human activities, conditional Random Fields (CRFs) [25] have emerged as a popular way to model contextual relations. Many variants augment CRFs with latent variables in order to model hidden states, such as latent CRFs [33] that have been applied to object recognition [36, 14, 42], scene understanding [35], gesture recognition [40] and grounding natural language to robotic tasks [29]. However, in these models, the predefined latent space is discrete and small to keep the learning and inference tractable. Our GP-LCRF, on the contrary, learns the latent space in a non-parametric way and admits continuous latent values.

III. Overview

We define the anticipation task as follows: Given an RGB-D video of a human interacting with the surrounding environment, our goal is to predict what will happen to the environment in a time span in terms of the next sub-activity label, object affordance labels and object trajectories. Modeling future human configurations, in the context of the activity and the environment, is a key ingredient for a good anticipation.

Human configuration has two sides of nature: It is *high-dimensional* in terms of the degree of freedom a human body



(a) ATCRF [20]

(b) Our GP-LCRF

Fig. 2: Graphical representations of the original ATCRF [20] and our GP-LCRF. Our model adds latent low-dimensional nodes $\mathcal{X}$ to the model, which are related to the original high-dimensional human configuration nodes $\mathcal{H}$ through Gaussian Process latent variable model with parameters α, β, γ. Shaded nodes indicate observations. In both models, temporal segment t is given for anticipation with observed human poses $\mathcal{H}^t$ and object locations $\mathcal{L}^t$, and the goal is to infer the next segment $t+1$ where nothing is observed.

possesses. One would need the location and orientation of each joint of a human skeleton to fully describe a static human pose, and the velocities and accelerations to describe a sequence of human motions. The high-dimensional representation is a guarantee for generating realistic human poses/motions. We need it to perform (self-)collision detection, inverse kinematics and path planning.

However, most dynamic human behaviors are intrinsically *low-dimensional*, as our arms and legs operate in a coordinated way and they are far from independent with each other. Many daily behaviors such as walking and jumping have been represented in low-dimensional space [7, 34]. The low-dimensional representation is a requisite for probabilistic models of human motions. The distribution of human poses can be used to synthesize or predict new poses.

Our main idea is to keep both the high- and low-dimensional representation of human dynamics in anticipating human activities. Our learning model thus has two parts:

Learning low-dimensional human dynamic distributions. For each human pose, indexed by i, we use h_i to denote its high-dimensional and x_i for its corresponding low-dimensional representation. The correspondence is specified by a mapping function, i.e. $h_i = f(x_i)$. Additionally, we are also interested in associating the mapping with a probabilistic model, so that we can generate new human dynamics (x_i, h_i) from the learned distribution. Hence, the objective of this part is to learn the parameters of f as well as a likelihood function $L(x_i, h_i)$ from the training data $\{h_i\}$.

Modeling the spatial and temporal context of human activities. We use a graphical model, following [20] to capture the three important aspects in a human activity—sub-activities $\mathcal{A}$, objects $\mathcal{O}$ and humans $\mathcal{H}$. Given a video segment t,[1] each entity is represented by a node in the graph modeling its prior distribution and the edges in the graph model their relations, as shown in Fig. 2. The whole video is a repetition of such a graph. Edges between consecutive segments are used to model temporal dynamics. In particular, for each human pose in segment t, in addition to the original human node h_i^t, we add a low-dimensional latent node x_i^t. The edges between h_i^t and $\mathcal{O}^t$ or $\mathcal{A}^t$ are used to capture human-object and human-activity relations, while the edges between x_i^{t-1} and x_i^t are for modeling the human dynamics. This graphical model thus defines a joint distribution $P(\mathcal{A}, \mathcal{O}, \mathcal{H}, \mathcal{X})$ as a product of parameterized edge potentials. We learn those parameters from labeled data and then sample future segments from this distribution for anticipation.

By combining these two parts, our proposed GP-LCRF possesses many advantages: First, we can now use the context of high-dimensional data that is difficult to model for a traditional CRF. Second, as the low-dimensional representation is modeled as latent nodes and the mapping is learned in an unsupervised way, our model does not require any extra label/data to learn. Third, being able to learn the distribution of the low-dimensional latent node makes our GP-LCRF a generative model that suits the anticipation problem. Before presenting our GP-LCRF, we first briefly review the background of the two parts in the following.

IV. Preliminaries

A. Dimensionality Reduction with a Gaussian Processes

Consider a general setting for regression problems: Our goal is to learn a mapping $h = f(x)$ for a set of N training pairs (x_i, h_i). However, from a Bayesian point of view, instead of mapping to one point, a Gaussian process (GP) "maps" x to a distribution of h. Let μ be the mean of the training data $\mu = \sum h_i / N$, and let $H_k = [h_{1,k} - \mu_k, \ldots, h_{N,k} - \mu_k]^T$ be the feature vectors of the k^{th} dimension. In a GP model, H_k can be viewed as one sample from a multivariate Gaussian distribution:

$$P(H_k | \{x_i\}) = \frac{1}{(2\pi)^{N/2} |K|^{1/2}} \exp(-\frac{1}{2} H_k^T K^{-1} H_k) \quad (1)$$

K is the covariance matrix of all inputs x_i. We can use many non-linear kernel functions, such as the popular "RBF kernel", to admit non-linear mappings:

$$K_{i,j} = k(x_i, x_j) = \alpha \exp(-\frac{\gamma}{2} ||x_i - x_j||^2) + \delta_{x_i, x_j} \beta^{-1}$$

[1] Frames of a video are grouped into temporal segments, and each segment spans a set of contiguous frames, during which the sub-activity and object affordance labels do not change.

where δ_{x_i,x_j} is the Kronecker delta. Using this kernel means that the two points, x_i and x_j, that are correlated in the latent space, will also be highly correlated after the mapping. The parameter α imposes a prior on how much the two points are correlated, γ is the inverse width of the similarity function, and β reflects how noisy the prediction is in general.

In a more general setup, only $h_1, \dots h_N$ are given and the goal is to determine the mapping function f as well as the corresponding x_i. This can be solved using Gaussian process latent variable models (GPLVM) proposed in [26]. GPLVM maximizes the likelihood of the training data, based on Eq. (1), to learn the parameters of the kernel function (γ, α, β) and the latent variables $x_1, \dots, x_N$.

Since GPLVM provides a probabilistic model of nonlinear mappings and generalizes well for small datasets, it has been extended to model human motions in many works. For example, it is integrated with a dynamical model to capture dynamical patterns in human motions [39] so that it can provide a strong prior for tracking human activities [38, 43, 5]. In this work, we also adopt GPLVM as a dimensionality reduction approach, however, our goal is to incorporate this with Latent CRFs to model high-dimensional human motions and *rich context* in the environment at the same time.

B. Capturing Context with CRF

Conditional Random Fields (CRFs) have been widely applied to a variety of vision and robotic applications for its ability to model rich contextual relations. Given a graph, where $\mathcal{V}$ denotes the nodes and $\mathcal{E}$ for the edges, A CRF defines the joint distribution of $\mathcal{V}$ conditioned on observed features:

$$P(\mathcal{V}) \propto \prod\nolimits_i \psi_i(v_i) \prod\nolimits_{(i,j)\in\mathcal{E}} \psi_{i,j}(v_i, v_j)$$

where ψ_i denotes the unary potential and $\psi_{i,j}$ denotes the pairwise potentials. For instance, in scene labeling, each segment is modeled as a node whose value indicates the object class label, and the edges between nodes describe the object-object (spatial) context [2]. The potentials are usually parameterized as a inner product of the feature and parameter vectors. The parameters are learned by maximizing the joint likelihood of training data.

In the task of anticipating human activities, previous work [20, 19] propose a novel CRF, ATCRF, to model the relationships between sub-activity labels $\mathcal{A}$, object affordance labels $\mathcal{O}$, object locations $\mathcal{L}$ and human poses $\mathcal{H}$, as shown in Fig. 2-(a). Inside one temporal segment, a graph structure is defined as $\mathcal{G}^t = (\mathcal{V}^t, \mathcal{E}^t)$ with $\mathcal{V}^t = \{\mathcal{O}^t, \mathcal{H}^t, \mathcal{A}^t\}$. The edges capture object-activity, object-object and human-activity relations. They also model temporal relations with edges between the sub-activities nodes and the same object nodes from the two consecutive segments.

V. Representing Humans

In this work, we define the high-dimensional representation of a human pose h as:

- Joint locations: We include the head location and the local locations (with respect to the head) of eight upper-body joints (neck, torso, left and right shoulders, elbows and hands). This gives 27 features in total.
- Head 3D orientation in a global coordinate system.
- Velocity and acceleration: To distinguish motions with different directions (e.g., moving left vs. right), we additionally include the difference of consecutive poses. For all the 30 features above, the velocity and acceleration at time t are computed as $h^t - h^{t-1}$ and $h^t - 2h^{t-1} + h^{t-2}$.

In total, each h is a 90-dimension vector.

Curse of $\mathcal{H}$'s high dimensionality. In ATCRFs [20], anticipation is conducted by sampling future segments, including possible sub-activities, object affordance labels and object trajectories. Because of the high dimensionality of the human configuration $\mathcal{H}$, instead of sampling the full human body they *assume* that the sampled object trajectory is always reachable by hands and only samples the hand location around the object. As a result, the anticipated human configuration is not realistic and this inaccuracy propagates to the computation of features and the potential function scores. Furthermore, because they assume that h is fully determined by the object trajectory, its temporal potential $\psi(h^t, h^{t-1})$ overlaps with $\psi(o^t, o^{t-1})$ and thus does not capture the human dynamics.

VI. GP-LCRF for Human Activity Anticipation

We propose a model, GP-LCRF, that learns a probabilistic mapping between the high- and low-dimensional representation of human dynamics based on Gaussian processes. Then it embeds the compactly represented humans as latent nodes in a CRF to capture a variety of context between the human, objects and activities.

Our GP-LCRF introduces a layer of latent nodes in a CRF: each node h_i is now linked to a latent node x_i and their relation is defined by a GPLVM with parameters (α, β, γ). Because latent nodes have much lower dimensions, we can model the edges between latent nodes (e.g., (x_i^t, x_i^{t+1})) instead of attempting to capture it with high-dimensional nodes directly. (The high-dimensionality of the human nodes makes the edge distribution ill-conditioned.) Figure 2 shows the corresponding graphical model.

GP-LCRF differs from other latent CRFs in two aspects:

Prior. We adopt GPLVM to impose a Gaussian process prior on the mapping and a ℓ_2-norm prior on the latent nodes. This prior regulates the mapping so that the high-dimensional human configurations h_i that are close in the original space would remain close in the latent space x_i. This property of local distance preservation is very desirable in many applications, especially for time series analysis.

Non-parametric. In many latent CRFs, the values of latent nodes are discrete and finite [33]. Some other works consider a non-parametric Bayesian prior over the latent values but they do not handle dimensionality reduction. In our GP-LCRF, the latent space is completely determined by the training data, making it more adaptive to various applications.

A. Likelihood of GP-LCRF

As shown in Fig. 2-(b), a GP-LCRF is a repetition of small graphs (one per each temporal segment). A segment

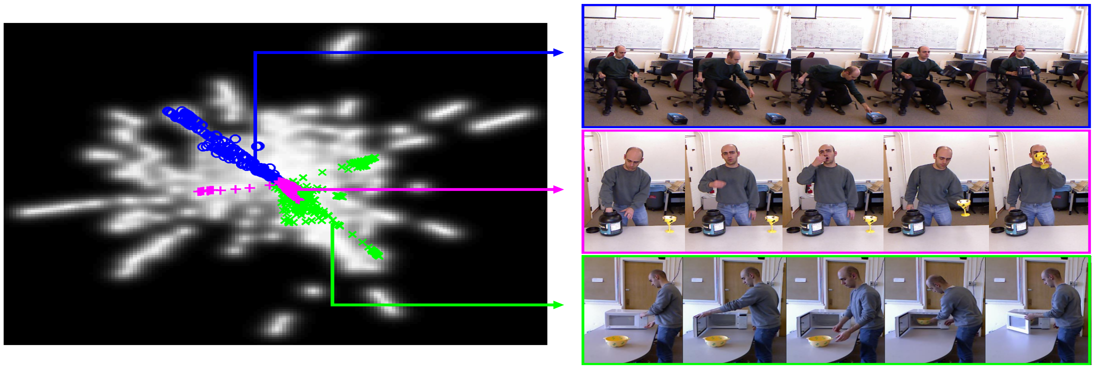


Fig. 3: An example of the learned mapping from the high-dimensional human configurations to a 2-dimensional space. The intensity of each pixel x visualizes the its probability $-\frac{D}{2}\ln\sigma^2(x) - \frac{1}{2}||x||^2$. We also plot the projected 2D points for different activities in different colors. We can see that human configurations from the same activity are mapped to a continuous area in the 2D space while the ones from different activities are separated.

t contains one sub-activity label node $\mathcal{A}^t$, object affordance label nodes $\mathcal{O}^t = \{O_i^t\}$, object location nodes $\mathcal{L}^t = \{L_i^t\}$, high-dimensional human configurations $\mathcal{H}^t = \{h_i^t\}$ and low-dimensional human representations $\mathcal{X}^t = \{x_i^t\}$.

Following the independence assumptions imposed by the edges in the graph, the likelihood of one temporal segment $P(\mathcal{A}^t, \mathcal{O}^t, \mathcal{X}^t | \mathcal{L}^t, \mathcal{H}^t)$ is,

$$\begin{aligned} P_t \;\propto\; & \psi(\mathcal{A}^t, \mathcal{H}^t)\prod\nolimits_i \psi(\mathcal{A}^t, o_i^t)\prod\nolimits_i \psi(o_i^t, L_i^t) \\ & \prod\nolimits_{(i,j)} \psi(o_i^t, o_j^t)\prod\nolimits_i \phi(x_i^t, h_i^t) \end{aligned} \tag{2}$$

where the first four terms capture human-activity relations, object-activity relations, object affordances and object-object relations respectively. These potentials are parameterized as log-linear functions of feature vectors [20]. We define the last term, potential of the mapping between x_i and h_i as the likelihood derived from GPLVM:

$$\phi(x_i, h_i) = \exp L(x_i, h_i) \tag{3}$$

$$L(x,h) = -\frac{||h - f(x)||^2}{2\sigma^2(x)} - \frac{D}{2}\ln\sigma^2(x) - \frac{1}{2}||x||^2 \tag{4}$$

where

$$\begin{aligned} f(x) &= \mu + H^T K^{-1}\mathbf{k}(x) \\ \sigma^2(x) &= k(x,x) - \mathbf{k}(x)^T K^{-1}\mathbf{k}(x) \\ \mathbf{k}(x) &= [k(x,x_1), \ldots, k(x,x_N)]^T \end{aligned}$$

The three terms in $L(x,h)$ measure the discrepancy between the given h and the prediction $f(x)$, the uncertainty of the prediction, and the prior of the latent value x.

We now consider the temporal relations between the two consecutive temporal segments $t-1$ and t:

$$P_{t-1,t} \propto \psi(\mathcal{A}^t, \mathcal{A}^{t-1})\prod\nolimits_i \psi(o_i^t, o_i^{t-1})\phi(x_i^t, x_i^{t-1}) \tag{5}$$

where the first two terms capture the temporal transitions of sub-activity labels and object affordance labels. They are also parameterized as log-linear functions of features [20]. We define the last term, the temporal transitions of latent nodes, as Gaussian distributions:

$$\phi(x_i^t, x_i^{t-1}) \propto \mathcal{N}(||x_i^t - x_i^{t-1}||^2; 0, 1) \tag{6}$$

Hence, the overall likelihood of a GP-LCRF is

$$L_{\text{GP-LCRF}} \propto \prod\nolimits_{t=1}^{T} P_t \times \prod\nolimits_{t=2}^{T} P_{t-1,t} \tag{7}$$

Using this function, we learn the parameters by maximize the training data's likelihood and to predict the future activities and human dynamics by sampling from this distribution.

B. Learning

During training, given all observations ($\mathcal{H}$ and $\mathcal{L}$) and labels ($\mathcal{A}$ and $\mathcal{O}$), our goal is to learn the parameters in every potentials *and latent nodes* $\mathcal{X}$ by maximizing the likelihood in Eq. (7), which can be written into two parts:

$$\begin{aligned} L_{\text{GP-LCRF}} = & \prod_t \Bigg(\psi(\mathcal{A}^t, \mathcal{H}^t)\prod_i \psi(\mathcal{A}^t, o_i^t)\psi(o_i^t, L_i^t) \\ & \prod_{(i,j)} \psi(o_i^t, o_j^t)\psi(\mathcal{A}^t, \mathcal{A}^{t-1})\prod_i \psi(o_i^t, o_i^{t-1}) \Bigg) \\ \times & \prod_t \left(\prod_i \phi(x_i^t, h_i^t)\phi(x_i^t, x_i^{t-1}) \right) \end{aligned}$$

The first pair of parentheses contains the CRF terms, with parameters denoted by Θ_{CRF}. (They are similar to the terms in ATCRF.) The second pair of parentheses contains all terms related to latent nodes in GP-LCRF with parameters including K, α, γ, β, denoted by Θ_{latent}. Note that Θ_{CRF} and Θ_{latent} are two *disjoint* sets.

Therefore, learning can be decomposed into two independent problems: 1) learning Θ_{CRF} by using the cutting-plane method in the structural learning for SVM [15], same as [20]; 2) learning Θ_{latent} by minimizing the negative log-likelihood, given by:

$$\begin{aligned} & -\ln P(\{x_i\}, \alpha, \gamma, \beta | \{h_i\}) \\ = & -\ln P(\{h_i\}|\{x_i\}, \alpha, \gamma, \beta)P(\{x_i\})P(\alpha, \gamma, \beta) \\ = & \frac{D}{2}\ln|K| + \frac{1}{2}\sum_{k=1}^{D} H_k^T K^{-1} H_k + \frac{1}{2}\sum_{i=1}^{N}||x_i||^2 + \ln\alpha\beta\gamma \end{aligned}$$

where the priors on the unknowns are: $P(x) = \mathcal{N}(0, I)$ and $P(\alpha, \beta, \gamma) \propto \alpha^{-1}\beta^{-1}\gamma^{-1}$. We use numerical optimization method L-BFGS [32] to minimize it.

TABLE I: Anticipation Results, computed over 3 seconds in the future averaged by 4-fold cross validation. The first six columns are in percentage and a higher value is better. The last column is in centimeters and a lower value is better.

Algorithms	Anticipated sub-activities			Anticipated object affordances			Anticipated traj.
	micro-P/R@1	macro-F1@1	Pre@3	micro-P/R@1	macro-F1@1	Pre@3	MHD@1 (cm)
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	48.1±0.9
ATCRF-KGS [20]	47.7±1.6	37.9±2.6	69.2±2.1	66.1±1.9	36.7±2.3	71.3±1.7	31.0±1.0
ATCRF [19]	49.6±1.4	40.6±1.6	74.4±1.6	67.2±1.1	41.4±1.5	73.2±1.0	30.2±1.0
HighDim-LCRF	47.0±1.8	37.2±2.8	68.5±2.1	65.8±1.8	37.3±2.4	70.6±1.6	29.3±0.9
PPCA-LCRF	50.0±1.5	40.7±1.4	74.2±1.2	67.8±1.7	41.7±1.3	73.4±1.0	28.7±0.9
Our 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**	**26.7±0.9**

C. Inference for Anticipation

Given the observed segment t, we predict the next future segment $t+1$ in the following way: We first sample possible object trajectories, represented in locations $\mathcal{L}^{t+1}$. Then we sample human configurations $\mathcal{H}^{t+1}$ and $\mathcal{X}^{t+1}$. We now use the sampled $\mathcal{L}^{t+1}$ and $\mathcal{H}^{t+1}$ as observations and infer the most likely sub-activity labels $\mathcal{A}^{t+1}$and object affordance labels $\mathcal{O}^{t+1}$ by maximizing the conditional likelihood in Eq. (7). All the samples together form a distribution over the future possibilities and we use the one with maximum a posterior (MAP) as our final anticipation.

We now present how to sample $\mathcal{H}^{t+1}$ and $\mathcal{X}^{t+1}$ in particular. (Sampling other terms is similar as in [20].) Given object locations, we generate a human motion of either moving or reaching an object. In both cases, the hand trajectory is given and the problem is formulated as: Given a target hand location ℓ^*, compute the most likely human configurations where both x and h are unknown. A good pose should reach to the target as well as being reasonable which can be measured by the likelihood from GPLVM, $L(x, h)$ in Eq. (4). Hence, we define the objective function as:

$$\arg\min_{x,h} -L(x,h) + \lambda||\ell^* - \ell(h)||^2 \tag{8}$$

where λ is the penalty of the new pose deviating from the target. In our implementation, we start with a simple IK solution h_0, and use the inverse mapping function $g(h) = x$ (given by GPLVM with back constraints [27]) to compute its corresponding x_0. In this way, the first term in Eq. (4) is always zero and can be neglected. So the new objective becomes a function of h only:

$$\arg\min_h \frac{D}{2}\ln\sigma^2(g(h)) + \frac{1}{2}||g(h)||^2 + \lambda||\ell^* - \ell(h)||^2 \tag{9}$$

We then use L-BFGS to optimize it.

VII. Experiments

Data. We test our model on the Cornell Activity Dataset-120 (CAD-120), same as used in [20, 19]. It contains 120 3D videos of four different subjects performing 10 high-level activities, where each high-level activity was performed three times with different objects. It contains a total of 61,585 total 3D video frames. The dataset is labeled with both sub-activity and object affordance labels. The sub-activity labels are: {*reaching, moving, pouring, eating, drinking, opening, placing, closing, scrubbing, null*} and the affordance labels are: {*reachable, movable, pourable, pour-to, containable, drinkable, openable, placeable, closable, scrubbable, scrubber, stationary*}.

Baselines. We compare against the following baselines:
1) *Chance*. Labels are chosen at random.
2) *ATCRF-KGS* [20]. ATCRF with fixed temporal structure.
3) *ATCRF* [19]. ATCRF with sampled temporal structures.
4) *HighDim-LCRF*. In this method, we do not compress the human configuration into a low-dimensional representation but directly model human dynamics in the high-dimensional space. We replace $\phi(x_i^t, h_i^t)$ with a Gaussian based on the distance between h_i^t to its nearest neighbor h^* in the training data. For an anticipated frame, we use inverse kinematics to generate a new pose that is closest to the target trajectory (without considering its GPLVM likelihood). We also change $\phi(x_i^{t-1}, x_i^t)$ to $\phi(h_i^{t-1}, h_i^t) \sim \mathcal{N}(||h_i^{t-1} - h_i^t||^2; 0, 1)$.
5) *PPCA-LCRF*. We use probabilistic principal component analysis (PPCA) instead of GPLVM for dimensionality reduction of human configurations. PPCA only learns a linear mapping and do not impose any prior on the latent space and the mapping. We verify through experiments that it does not model low-dimensional human dynamics well and thus is outperformed by our GP-LCRF model.

Evaluation. We 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 and evaluated by the following metrics (same are used in [20, 19]):
1) *Labeling Metrics (on top#1 prediction)*. For anticipated sub-activity and affordance labels, we compute the overall micro accuracy (P/R) and macro F1 score. Micro precision/recall is equal to the percentage of correctly classified labels. Macro precision and recall are averaged over all classes.
2) *Pre@3*. In practice a robot should plan 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 label matches one of the top three predictions, then it counts towards positive.
3) *Trajectory Metric (on top#1 prediction)*. For anticipated human trajectories, we compute the modified Hausdorff distance (MHD) to the true trajectories. MHD finds the best local point correspondence of the two trajectories over a small temporal window to compute distance between those points. The distance is normalized by the length of the trajectory.

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 can see that our proposed GP-LCRF outperforms all the baseline algorithms and achieves a

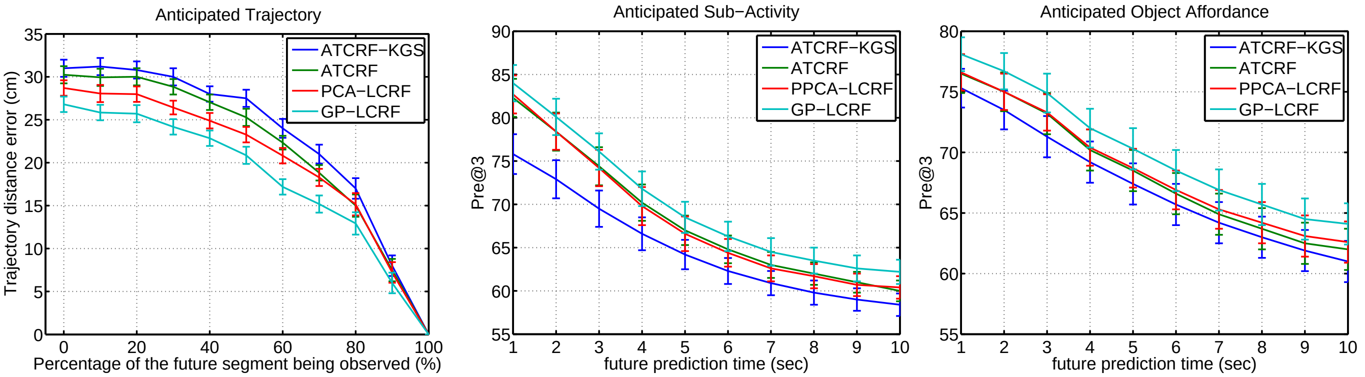


Fig. 4: Plots showing (from left to right): a) how the trajectory distance error changes with the observed percentage in the segment to anticipate increases from 0% to 100%; b) The Pre@3 of the anticipated sub-activity labels as a function of the length of future prediction time in seconds; c) The Pre@3 of the anticipated object affordance labels as a function of the length of future prediction time in seconds.

consistent increase across all metrics. Especially as our GP-LCRF aims to model human configurations better, we can see that the anticipated human trajectory error is reduced from 30.2 cm to 26.7 cm which is a 11.6% improvement and has a p-value of 0.0107 indicating the difference is statistically very significant. We now inspect the results in detail from the following aspects:

The importance of dimensionality reduction. Table I shows that when not using any dimensionality reduction, HighDim-LCRF performs even worse than ATCRF even though it tries to model the human temporal dynamics. This is because that in the high-dimensional space, $\phi(h^{t-1}, h^t)$ can be noisy and over-fitted, thus modeling it actually hurts the performance.

On the other hand, with dimensionality reduction, PPCA-LCRF outperforms HighDim-LCRF, however it only achieves comparable results as ATCRF. This shows that the quality of the dimensionality reduction is quite important. Figure 5 illustrates a learned mapping of human configurations. Although both mapped to a 2D space, compared to GPLCRF in Fig. 3, PPCA learns a flat mapping and does not distinguish different motions well enough. For instance, the motions in the activity of 'taking medicine' (in magenta) and 'microwaving food' (in green) are very different, however they are mapped to an overlapped area using PPCA in Fig. 5. As a result, the effect of the dimensionality reduction in PPCA-LCRF is not as significant as our GP-LCRF.

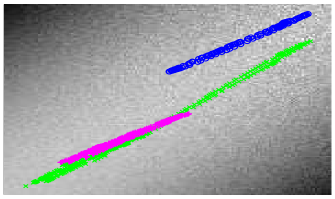


Fig. 5: The learned mapping using PPCA. The colored points corresponding to the activities in Fig. 3.

Sensitivity of the results to the degree of dimensionality reduction. We investigate the performance of GP-LCRF with different dimensions of the latent space, from 1-D to 5-D in Fig. 6, in terms of the trajectory distance error. We can see that under various learning conditions (where the anticipated segment is observed in different percentages), GP-LCRF with latent dimensions of two to five all give similar performance. Dimensions of one has an obvious performance drop but is still better than ATCRF. However, with the observation's percentage increase to 80%, the gap diminishes as the anticipation problem becomes easier. Evaluations with the labeling metrics share similar trends. Hence, this shows that our GP-LCRF is very robust to the choices of the latent dimensions.

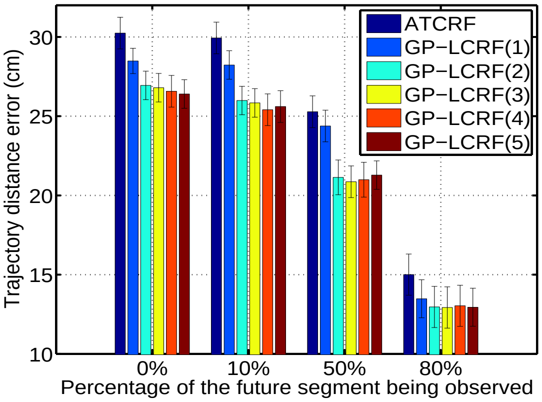


Fig. 6: The trajectory distance error of GP-LCRF with different dimensions of latent space (from 1D to 5D, shown in the parentheses). We evaluate the performance under different conditions where the percentage of the future segment observed is 0%, 10%, 50% and 80%, i.e., the task is to anticipate is the rest of 100%, 90%, 50% and 20% of that segment respectively.

The impact of the observation time. The first plot in Fig. 4 shows how the trajectory distance error, averaged over all the moving sub-activities in the dataset, changes with the increase of the observed part (in percentage) in the segment to be anticipated. While all approaches achieve better predictions with increased observation time, our GP-LCRF consistently performs better than the others, especially in the range of 20% to 60%. Because this part, unlike the beginning where the evidence of human motions is too weak to be useful and unlike the near end where the evidence human-object interactions weighs more than humans alone, is where the momentum of human motions can be captured from the observation by our model (through the velocity and acceleration features described in Sec. V) and be fully utilized for anticipation.

Results with change in the future anticipation time. The last two plots in Fig. 4 show the changes of Pre@3 with

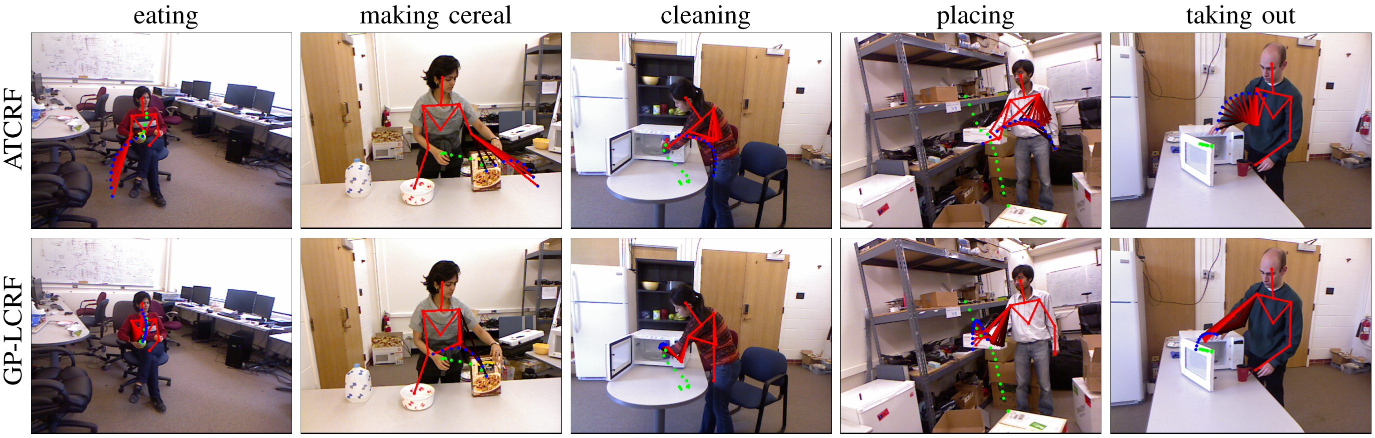


Fig. 7: Top-ranked trajectories predicted by ATCRF (top) and our GP-LCRF (bottom) for different activities. In each image, the ground-truth trajectory is shown in green dots, predicted trajectory in blue, and the anticipated human skeletons in red in the order of from dark to bright.

the anticipation time lengthened. The longer the anticipation time, the harder the task gets and thus the performances of all approaches decrease. However, the improvement of our GP-LCRF against ATCRF grows from 1.7% to 2.2% for sub-activity anticipation and from 1.6% to 2.1% for object affordance anticipation. This demonstrates the potential of modeling human kinematics well in long-term anticipations.

How does modeling human dynamics improve anticipated trajectories? In addition to the quantitative results in Fig. 4-(a), we also sample some qualitative results showing the top-ranked predicted trajectories in Fig. 7 using ATCRF (top) and using our GP-LCRF (bottom). In each image, we illustrate the predicted hand trajectories in blue dots, the ground-truth trajectories in green dots and human skeletons in red. We performed an ablative analysis and we now discuss some failures in the original ATCRF but are avoided by our GP-LCRF, arranged in three major categories:

1) Unrealistic skeletons leading to impossible trajectories: In the first two cases/columns, the trajectories sampled by ATCRF are both not reachable (without making any effort such as bending over or leaning forward). As ATCRF does not consider any human kinematics and it simply changes the hand location to match the trajectory, the forearms in these two cases are stretched out. The features computed from these false human skeletons are erroneous and thus wrong trajectories are picked out. Our GP-LCRF, however, generates kinematically-plausible skeletons (because of availability of high-dimensional configurations in the model) so that the out-of-reach trajectories will have high penalty in the likelihood $L(x, h)$ and out-ranked by those reachable ones.

2) Unnatural poses leading to unlikely trajectories: In other cases, such as the third column in Fig. 7 where the subject picked up a rag on the table along the green dots to clean the microwave, both trajectories are physically possible but the top one requires raising the right hand to cross the left hand making a very unnatural pose. Because GP-LCRF learns the distribution of human poses from the training data, it assigns a low probability to uncommon poses such as the top one and prefers the bottom poses and the trajectory instead.

3) Not modeling motions leading to discontinuous trajectories: How human body moved in the past gives such a strong cue that sometimes we can have decent anticipated trajectories purely based on the continuity and smoothness of human motions. For instance, the two subjects are lifting the box (4th column) and reaching towards the microwave door (last column). While our GP-LCRF chooses trajectories matching the moving directions best, ATCRF which does not model human temporal relations (i.e., no edges between $\mathcal{H}^{t-1}$ and $\mathcal{H}^{t}$) produces trajectories with sudden changes in the direction.

Runtime. On a 16-core 2.7GHz CPU, our code takes 11.2 seconds to anticipate 10 seconds in the future, thus achieving *near real-time (*1.12*X)* performance.

VIII. Conclusion and Discussion

In this paper, we proposed a new model, GP-LCRF, in order to compactly model human configurations and dynamics in the task of anticipating human activities. The key idea is to have access to high-dimensional representations of humans for generating detailed and realistic human poses in the future, and meanwhile to have a compact low-dimensional representation for the ease of modeling and learning the context among humans, objects and the activity. Bearing this in mind, our GP-LCRF models the low-dimensional representations as latent nodes through a Gaussian process, and models the relations between human and other entities in a CRF. We tested our model on 120 RGB-D videos of different activities and it out-performed the state-of-the-art results consistently, and reduced the predicted human trajectory error by 11.6%.

This improvement could make a significant difference to robots working in the presence of humans. With more accurate human trajectory prediction, robots can plan more relevant actions and paths [21, 10]. Furthermore, with real-time anticipation, our work can be used for human-robot interaction, such as to improve the efficiency of collaborative tasks [41, 9], or to avoid intrusion/collision during navigation [17].

Acknowledgments

We thank Hema Koppula and Ozan Sener for useful discussions. This work was supported by ARO award W911NF-12-1-0267, ONR award N00014-14-1-0156 and NSF Career Award to Saxena.

References

[1] D. J Agravante, A. Cherubini, A. Bussy, and A. Kheddar. Human-humanoid joint haptic table carrying task with height stabilization using vision. In *IROS*, 2013.

[2] A. Anand, H. Koppula, T. Joachims, and A. Saxena. Contextually guided semantic labeling and search for 3d point clouds. *IJRR*, 32(1):19–34, 2012.

[3] M. Bennewitz, W. Burgard, G. Cielniak, and S. Thrun. Learning motion patterns of people for compliant robot motion. *IJRR*, 24(1):31–48, 2005.

[4] E. Demircan, T. Besier, S. Menon, and O. Khatib. Human motion reconstruction and synthesis of human skills. In *Advances in Robot Kinematics: Motion in Man and Machine*, pages 283–292. Springer, 2010.

[5] A. Geiger, R. Urtasun, and T. Darrell. Rank priors for continuous non-linear dimensionality reduction. In *CVPR*, 2009.

[6] H. Grabner, J. Gall, and L. J. Van Gool. What makes a chair a chair? In *CVPR*, 2011.

[7] Keith Grochow, Steven L. Martin, Aaron Hertzmann, and Zoran Popovic. Style-based inverse kinematics. *ACM Trans. Graph.*, 23(3):522–531, 2004.

[8] A. Gupta, S. Satkin, A. A. Efros, and M. Hebert. From 3d scene geometry to human workspace. In *CVPR*, 2011.

[9] K. P Hawkins, N. Vo, S. Bansal, and A Bobick. Probabilistic human action prediction and wait-sensitive planning for responsive human-robot collaboration. In *HUMANOIDS*, 2013.

[10] A. Jain, B. Wojcik, T. Joachims, and A. Saxena. Learning trajectory preferences for manipulators via iterative improvement. In *NIPS*, 2013.

[11] Y. Jiang and A. Saxena. Hallucinating humans for learning robotic placement of objects. In *ISER*, 2012.

[12] Y. Jiang and A. Saxena. Infinite latent conditional random fields for modeling environments through humans. In *RSS*, 2013.

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

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

[15] T. Joachims, T. Finley, and CN. J. Yu. Cutting-plane training of structural svms. *Machine Learning*, 77(1):27–59, 2009.

[16] D. Katz, A. Venkatraman, M. Kazemi, D. Bagnell, and A. Stentz. Perceiving, learning, and exploiting object affordances for autonomous pile manipulation. In *RSS*, 2013.

[17] W.G. Kennedy, M. D Bugajska, M. Marge, W. Adams, B. R Fransen, D. Perzanowski, A. C Schultz, and J. G. Trafton. Spatial representation and reasoning for human-robot collaboration. In *AAAI*, 2007.

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

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

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

[21] H. Koppula and A. Saxena. Anticipatory planning for human-robot teams. In *ISER*, 2014.

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

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

[24] D. Kulić, C. Ott, D. Lee, J. Ishikawa, and Y. Nakamura. Incremental learning of full body motion primitives and their sequencing through human motion observation. *IJRR*, 31(3): 330–345, 2012.

[25] J. Lafferty, A. McCallum, and F.C.N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, 2001.

[26] N. D Lawrence. Gaussian process latent variable models for visualisation of high dimensional data. In *NIPS*, 2004.

[27] N. D Lawrence and J. Quiñonero-Candela. Local distance preservation in the gp-lvm through back constraints. In *ICML*, pages 513–520. ACM, 2006.

[28] W.Y. Lo and M. Zwicker. Real-time planning for parameterized human motion. In *SIGGRAPH/Eurographics Symposium on Computer Animation*, 2008.

[29] D. K. Misra, J. Sung, K. Lee, and A. Saxena. Tell me dave: Context-sensitive grounding of natural language to mobile manipulation instructions. In *RSS*, 2014.

[30] R. Navaratnam, A. W. Fitzgibbon, and R. Cipolla. The joint manifold model for semi-supervised multi-valued regression. In *ICCV*, 2007.

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

[32] J. Nocedal and S. Wright. Numerical optimization, series in operations research and financial engineering. *Springer, New York*, 2006.

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

[34] A. Safonova, J. K Hodgins, and N. S Pollard. Synthesizing physically realistic human motion in low-dimensional, behavior-specific spaces. *ACM Transactions on Graphics (TOG)*, 23(3): 514–521, 2004.

[35] A. Saxena, S.H. Chung, and A. Ng. Learning depth from single monocular images. In *NIPS 18*, 2005.

[36] P. Schnitzspan, S. Roth, and B. Schiele. Automatic discovery of meaningful object parts with latent crfs. In *CVPR*, 2010.

[37] H. Sidenbladh, M. J Black, and L. Sigal. Implicit probabilistic models of human motion for synthesis and tracking. In *ECCV*, 2002.

[38] R. Urtasun, D. J Fleet, and P. Fua. 3d people tracking with gaussian process dynamical models. In *CVPR*, 2006.

[39] J M Wang, D J Fleet, and A Hertzmann. Gaussian process dynamical models for human motion. *PAMI*, 30(2):283–298, 2008.

[40] S.B. Wang, A. Quattoni, L.P. Morency, D. Demirdjian, and T. Darrell. Hidden conditional random fields for gesture recognition. In *CVPR*, 2006.

[41] R. Wilcox, S. Nikolaidis, and J. A. Shah. Optimization of temporal dynamics for adaptive human-robot interaction in assembly manufacturing. In *RSS*, 2012.

[42] C. Wu, I. Lenz, and A. Saxena. Hierarchical semantic labeling for task-relevant rgb-d perception. In *RSS*, 2014.

[43] A. Yao, J. Gall, L. V Gool, and R. Urtasun. Learning probabilistic non-linear latent variable models for tracking complex activities. In *NIPS*, 2011.

[44] 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.