Temporal and Object Relations in Unsupervised Plan and Activity Recognition

Richard G. Freedman and **Hee-Tae Jung** and **Shlomo Zilberstein**
College of Information and Computer Sciences
University of Massachusetts Amherst
{freedman, hjung, shlomo}@cs.umass.edu

Abstract

We consider ways to improve the performance of unsupervised plan and activity recognition techniques by considering temporal and object relations in addition to postural data. Temporal relationships can help recognize activities with cyclic structure and are often implicit because plans have degrees of ordering actions. Relations with objects can help disambiguate observed activities that otherwise share a user's posture and position. We develop and investigate graphical models that extend the popular latent Dirichlet allocation approach with temporal and object relations, examine the relative performance and runtime trade-offs using a standard dataset, and consider the cost/benefit trade-offs these extensions offer in the context of human-robot and human-computer interaction.

1 Introduction

There has been growing interest lately in developing *plan recognition* (PR) and *activity recognition* (AR) techniques for a wide range of applications. In particular, these methods are essential for effective human-robot interaction (HRI) since robots need to predict what other agents in the environment are doing (Lösch et al. 2007; Sung et al. 2012). For example, when lending an object to a person, a robot cannot simply execute time-stamped commands. There is often considerable uncertainty about the pace and way in which people operate. The robot must therefore observe and understand what the person is doing. A poorly-timed or chosen response can hinder progress or have worse consequences in scenarios such as patient monitoring or search-and-rescue.

The prevailing techniques in AR often employ statistical methods and graphical models such as hidden Markov models (HMM) and latent variable mixture models (Sukthankar et al. 2014). Due to the designs of these graphical models, independence assumptions enable efficient statistical inference of the latent variables' values. For example, the Viterbi algorithm (Viterbi 1967) is often used for HMM's and variational inference (Krstovski and Smith 2013) and Gibbs sampling (Griffiths 2002) for mixture models. However, each of these methods has weaknesses stemming from these same independence assumptions. The HMM's latent Markov chain emphasizes temporal relations between lower-order observations; that is, the current state heavily relies on the order of the recent states in the sequence which only allows recognition of rigidly structured activities. On the other hand, latent variable mixture models such as the Latent Dirichlet Allocation (LDA) topic model (Blei, Ng, and Jordan 2003) omit structure completely and use bag-of-words models. Such models assume that every observation is independent of one another relying on the distribution of all observations in the sequence. Without any dependence on ordering, activities with cyclic structure or temporal constraints are hard to recognize.

Similar concerns have been raised in natural language processing (NLP), an area suggested to have much in common with plan recognition (Geib and Steedman 2007; Freedman, Jung, and Zilberstein 2014). For NLP, the local temporal dependencies enforced by HMM's place a strong emphasis on syntactic properties of phrases without any consideration of semantics. The global dependencies enforced by LDA topic models instead emphasize the semantic features of text without acknowledging its syntax. The composite model (Griffiths et al. 2004) has been developed to bring HMM's and LDA together for a single model that takes both syntax and semantics into account. We suggest that the composite model may also be used for sequences of observations to bring together both temporally local and global relationships for improved PR and AR.

Besides using temporal information, recognition systems can benefit from information regarding relations between the observed users and objects in the environment because the observing robot's interactions with the user will likely involve handling the same objects. For example, when moving furniture (Mörtl et al. 2012), both agents will need to handle the same object in order to coordinate carrying it. Furthermore, the object may provide information about the observed user's plan/action that is not available from the user's posture and position. For example, in a kitchen environment (Song et al. 2013; De la Torre et al. 2009) the most notable difference between cleaning a spill and mopping sauce from a plate is holding a napkin versus using a slice of bread. This has also been addressed in the field of robotics under *tool affordances*, a psychological theory stating that people view the functionality of objects based on their features (Gibson 2001).

We propose graphical models extending LDA and the composite model that enable the inclusion of object information during PR and AR. Although processing compositions of models will increase the computational complexity, the use of temporal and object-related information seems cru-

$$\begin{tikzcd}
\textbf{LDA} \arrow[r, "\textbf{temporal relations}"] \arrow[d, "\textbf{object relations}"] & \textbf{Composite Model} \arrow[d, "\textbf{object relations}"] \\
\textbf{Parameterized LDA} \arrow[r, "\textbf{temporal relations}"] & \textbf{Parameterized Composite Model}
\end{tikzcd}$$

Figure 1: Proposed models enhancing LDA

cial for applications involving cooperation between humans and machines. There are many factors that may be considered when performing PR and AR, and we must consider the cost/benefit ratio for each one. We hypothesize that both temporal and object relations provide unique information compared to postural data as well as to each other. We investigate these trade-offs when performing PR and AR with temporal- and object-related information in addition to postural data obtained from a RGB-D (red, green, blue-depth) sensor. The main contributions of this work are the development of suitable graphical models extending LDA for PR and AR with respect to these factors, an initial implementation of these models which has been used for testing on a standard dataset, and an analysis of the associated trade-offs.

The rest of the paper is organized as follows. Section 2 provides an overview of related work in the fields of PR, AR, and generative graphical models. Section 3 presents the graphical models extending LDA. These models are summarized in Fig. 1. We then describe initial experiments to illustrate the runtime costs and recognition performance in Section 4. We conclude with a discussion and future work.

2 Related Work

Real-time PR and AR systems have been developed for a wide range of domain specific applications including videogames (Cheng and Thawonmas 2004; Synnaeve and Bessière 2011), monitoring in smart environments (Cook, Krishnan, and Rashidi 2013), and human-robot interaction (Koppula and Saxena 2013). In gaming applications, PR is often performed at a higher-level where actions are predefined, while low-level user data is captured and higher-level actions must be inferred for other applications.

Approaches in AR for extracting these higher-level interpretations from low-level signal data are often statistical. The approach of Kelley et al. (2008) learns a collection of HMM's for each activity based on contextual information and then selects the one that best explains the observed sensor readings. However, it has often been the case that the single dimensionality of HMM's is not enough to capture deeper structure in plans. Hierarchical variations such as Cascading HMM (White, Blaylock, and Bölöni 2009), Abstract HMM (Bui, Venkatesh, and West 2002), and Hierarchical HMM (Fine, Singer, and Tishby 1998; Bui, Phung, and Venkatesh 2004) have thus been developed, often assuming a given hierarchical task network (HTN).

Not all applications can be well defined with such structure ahead-of-time, though. Accommodating for noisy sensor readings can greatly increase the size of the HTN. Additionally, for domain-inspecific environments like one's home where observed agents can perform a virtually endless number of tasks, an unsupervised approach is more appropriate to determine the commonly performed activities. Latent variable mixture models can cluster the data into bins signifying certain actions (Huỳnh, Fritz, and Schiele 2008).

LDA has become a frequently used model in AR (Jung et al. 2015; Chikhaoui, Wang, and Pigot 2012; Rieping, Englebienne, and Kröse 2014). Huỳnh, Fritz, and Schiele (2008) used them with on-body sensors to recognize daily routine activities. They used the learned model to develop a system that can automatically annotate future recorded data from the sensors. Wang and Mori (2009) trained a LDA model using annotated video sequences to recognize predefined actions—they named the approach *Semilatent Dirichlet Allocation* since it was not completely unsupervised like traditional methods. Zhang and Parker (2011) used LDA without any modifications to cluster readings from a RGB-D sensor attached to a robot. While Wang and Mori simply used the pixelated image data with flow fields, Zhang and Parker compressed the three-dimensional point-cloud data using local spatio-temporal features into vectors of four-dimensional cuboids. However, both approaches map their representations to codebooks with a finite set of symbols. This limits future observations which must also be mapped to this codebook; new unique inputs will consequently be misassigned to an input symbol. Freedman, Jung, and Zilberstein (2014) suggested a third representation to avoid this limitation by simply discretizing the joint-angles describing the observed agent's posture. They studied how discretization granularity affects the size of the library of available input symbols and the impact on the performance of LDA.

Although Freedman, Jung, and Zilberstein's representation is more robust with respect to observing the acting agent, it lacks other information regarding the surrounding environment. We propose revising LDA to handle spatio-temporal features as done by Zhang and Parker. We choose to revise the model rather than the representation because Zhang and Parker noted that *any changes to the environment changed the compression enough to alter the recognition performance without additional training*. We seek to make our model generalizable to other domains despite such changes, which implies using tuples of inputs in place of compressed inputs. Koppula and Saxena (2013) recently developed a supervised AR approach modifying conditional random fields for spatio-temporal features by pairing postures and object affordances as well. Other variations of LDA have been developed for various applications and purposes (Andrzejewski et al. 2011; Mimno et al. 2009; Wang and McCallum 2006; Blei and McAuliffe 2007; Wallach 2006; Griffiths et al. 2004).

3 LDA Topic Model and Its Extensions

LDA is a probabilistic topic model that generates a set of D documents from a set of T topics. Each topic is a distribution $\vec{\phi} = \{\phi_1, \ldots, \phi_T\}$ over the set of all word tokens in a V-token vocabulary and each document d with N_d tokens is composed of a distribution $\vec{\theta} = \{\theta_1, \ldots, \theta_D\}$

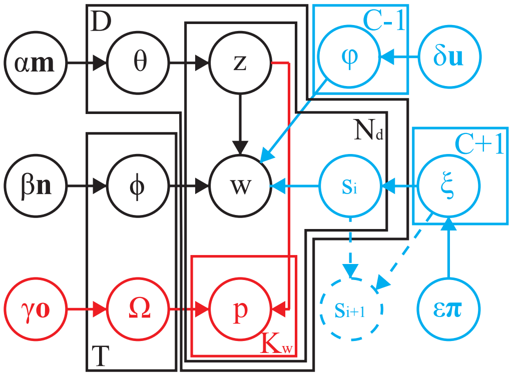

Figure 2: Graphical model representation of the parameterized composite model with LDA, the composite model, and parameterized LDA as subgraphs.

over the topics. Entries of $\vec{\phi}$ and $\vec{\theta}$ are drawn from a Dirichlet distribution with a hyperparameter in $H = \{\alpha\vec{m}, \beta\vec{n}\}$ (each a product of a scalar pseudocount and vector prior mean) (Steyvers and Griffiths 2007). Our ultimate extension of LDA is shown as a graphical model in Fig. 2, but only the subgraph shown in black is LDA itself. The generating process is lines $1, 2, 8, 9, 11, 12,$ and 15 of Algorithm 1, which also shows the ultimate extension covered in the paper.

To perform PR and AR, a word token is a pose where V is the number of discretized poses read by a RGB-D sensor similar to those described by Freedman, Jung, and Zilberstein (2014). We only use ten joint angles (the head, hands, and feet are rigidly connected to the body) and derive the angles from joints that share a link rather than all from the head. A document is composed of a sequence of observed postures representing a single plan execution. Each topic is considered to be an action or activity that composes a plan.

The sensor receives the postural data so that only the poses $\vec{w} = w_{1,1}, \ldots, w_{1,N_1}, w_{2,1}, \ldots, w_{D,N_D}$ are observed. We thus need to infer the latent activity variables $\vec{z}$ which will serve as an AR system. Then we approximate $\vec{\theta}$, which serves as a PR system. For training, we use collapsed Gibbs sampling to assign values to $\vec{z}$. Using the conjugate prior property of the Dirichlet distribution, we can integrate out the parameters to approximate the sampling likelihood

$$\mathrm{P}\left(z_i \left| \vec{z}_{\setminus i}, \vec{w}, H\right.\right) \propto \frac{N_{z_i}^{\setminus i}\left(w_i\right) + \beta n_{w_i}}{\sum_{v=1}^{V} N_{z_i}^{\setminus i}\left(v\right) + \beta} \cdot \frac{N_d^{\setminus i}\left(z_i\right) + \alpha m_{z_i}}{\sum_{t=1}^{T} N_d^{\setminus i}\left(t\right) + \alpha}$$

$= f_\phi \cdot f_\theta$ where $N_t^{\setminus i} : \{1, \ldots, V\} \rightarrow \mathbb{Z}^{\geq 0}$ is the number of times pose v is assigned activity t excluding the pose at the sampled index and $N_d^{\setminus i} : \{1, \ldots, T\} \rightarrow \mathbb{Z}^{\geq 0}$ is the number of times a pose in sequence d is assigned activity $t \in T$ excluding the pose at the sampled index. As $\left(\sum_{t=1}^{T} N_d^{\setminus i}\left(t\right) + \alpha\right)$ is a constant value $N_d - 1 + \alpha$ with respect to the sampled variable z_i, it only serves as a normalizing factor of the sampling likelihood and can thus be excluded. Then, to use LDA as a PR and AR system, we similarly derive the predictive probability of new observation sequences $\vec{w}' = \left\langle w_{D+1,1}, \ldots, w_{D+D',N_{D+D'}}\right\rangle$ given the training data and new observations up to the current one:

Algorithm 1:
Generative Process for Parameterized Composite Model

```
1  for each topic $t \in \{1, \ldots, T\}$ do
2      draw $\phi_t \sim$ Dirichlet $(\beta\vec{n})$
3      draw $\Omega_t \sim$ Dirichlet $(\gamma\vec{o})$
4  for each state $c \in \{2, \ldots, C\}$ do
5      draw $\varphi_c \sim$ Dirichlet $(\delta\vec{u})$
6  for each state $c \in \{0, \ldots, C\}$ do
7      draw $\xi_c \sim$ Dirichlet $(\epsilon\vec{\pi})$
8  for each document $d \in \{1, \ldots, D\}$ do
9      draw $\theta_d \sim$ Dirichlet $(\alpha\vec{m})$
10     assign $s_{d,0} \leftarrow 0$
11     for each index $i \in \{1, \ldots, N_d\}$ do
12         draw topic $z_{d,i} \sim \theta_d$
13         draw state $s_{d,i} \sim \xi_{s_{d,i-1}}$
14         if $s_{d,i} = 1$ then
15             draw word token $w_{d,i} \sim \phi_{z_{d,i}}$
16         else
17             draw word token $w_{d,i} \sim \varphi_{s_{d,i}}$
18         for each index $j \in \left\{1, \ldots, K_{w_{d,i}}\right\}$ do
19             draw parameter $p_{d,i,j} \sim \Omega_{z_{d,i}}$
```

$$\mathrm{P}\left(z'_i \left| \vec{z}, \vec{z}'_{<i}, \vec{w}, \vec{w}'_{<i}, H\right.\right) \propto f_\phi^+ \cdot f_\theta^+ = \frac{N'^{<i}_{z'_i}\left(w'_i\right) + N_{z'_i}\left(w'_i\right) + \beta n_{w'_i}}{\sum_{v=1}^{V}\left(N'^{<i}_{z'_i}\left(v\right) + N_{z'_i}\left(v\right)\right) + \beta} \cdot \frac{N'^{<i}_{D+d}\left(z'_i\right) + \alpha m_{z'_i}}{\sum_{t=1}^{T} N'^{<i}_{D+d}\left(t\right) + \alpha}$$

where we first perform Gibbs sampling on the previous new observations' activity assignments $\vec{z}'_{<i}$ as done during training. Although the previous activity assignments were already classified and likely piped to the response system, we still resample them to refine our likelihoods for better recognizing future actions as well as getting the most likely distribution to approximate each θ_{D+d} for PR.

3.1 Parameterized LDA

The parameterized extension of LDA considers documents whose word tokens may contain a list of parameters. That is, a single word in a document is now a K-ary proposition of the form $w_i\left(p_{i,1}, \ldots, p_{i,K_{w_i}}\right)$ where K_w is the arity of word token w, which may vary between identical word tokens, allowing an overloaded definition. We assume that these arguments are elements of a second vocabulary of Q items. Each topic has an additional distribution $\vec{\Omega} = \{\Omega_1, \ldots, \Omega_T\}$ over this second vocabulary which is drawn from a Dirichlet distribution with hyperparameter $\gamma\vec{o} \in H$. Thus its graphical model is the black and red subgraphs in Fig. 2 with generative process detailed by lines 1-3, 8, 9, 11, 12, 15, 18, and 19 from Algorithm 1.

This model essentially runs an additional K_{w_i} LDA topic models simultaneously that all share the same topic. That is, each parameter is sampled as in LDA from a distribution only conditioned on the topic of the current word token Ω_{z_i}. We note that parameterized LDA applies the bag-of-words model to the parameters so that each argument is independent of the others and the order does not matter.

For our application, each parameter is an object with which the observed agent is interacting. Identifying which objects should be considered as parameters is left for future research. Currently, we simply use objects that are within a fixed distance from the observed agent's joints identified by the RGB-D sensor. Song et al.'s (2013) extraction of clauses for recognizing activities with a Markov logic network identifies objects by such proximity to the observed agent's hands. By considering other joints, we can also extract localization information such as one's position in a room (near a refrigerator) and consider items near the head or feet which may become involved as the activity progresses (such as picking up an object from the ground).

Poses and objects can remove ambiguity that the other one would indicate about the activity alone. For example, Freedman, Jung, and Zilberstein (2014) acknowledged that some discretized poses for activities such as squatting and jumping appeared identical, i.e. generated similar word tokens. If the ground is an object within the vicinity of such a pose, then we are more likely to recognize the activity as squatting than jumping. This is because, like postures in PR and AR as well as most non-prepositional words in NLP, *objects provide semantic context to the activity and plan*. Inversely, Jain and Inamura (2013) show that a single object can be used in more than one activity depending on its orientation and utilized affordances. To accommodate the different orientation and/or affordance, it is likely that the pose of the agent will be different as in their example of using the back end of a screwdriver as a hammer—the arm would alternate between rising and falling rather than rotating in place.

We assume our sensor can perform object recognition in addition to reading postural data so that we observe pose-object pairs $\overrightarrow{(w,\vec{p})} = \langle (w_{1,1}, \vec{p}_{1,1}), \ldots, (w_{D,N_D}, \vec{p}_{D,N_D}) \rangle$ where $\vec{p}_{d,i} = \left\langle p_{d,i,1}, \ldots, p_{d,i,K_{w_{d,i}}} \right\rangle$. Hence we still need to infer just the latent activity variables $\vec{z}$. We continue to use Gibbs sampling to assign values to $\vec{z}$ using the same techniques described for approximating the sampling likelihood in the LDA topic model:

$$\mathrm{P}\left(z_i \middle| \vec{z}_{\setminus i}, \overrightarrow{(w,\vec{p})}, H\right) \propto f_\phi \cdot f_\theta \cdot \prod_{j=1}^{K_{w_i}} \frac{A_{z_i}^{\setminus i}(p_{i,j}) + \gamma o_{p_{i,j}}}{\sum_{q=1}^{Q} A_{z_i}^{\setminus i}(q) + \gamma}$$

$= f_\phi \cdot f_\theta \cdot f_\Omega$ where $A_t^{\setminus i} : \{1, \ldots, Q\} \rightarrow \mathbb{Z}^{\geq 0}$ is the number of times object q is assigned activity t excluding the parameters in the pose-object pair at the sampled index. As in LDA, we may omit $(\sum_{t=1}^{T} N_d^{\setminus i}(t) + \alpha)$ because it is constant with respect to the sampled variable z_i. For a new sequence of pose-object pair observations $\overrightarrow{(w,\vec{p})}' = \left\langle (w_{D+1,1}, \vec{p}_{D+1,1}), ..., (w_{D+D',N_{D+D'}}, \vec{p}_{D+D',N_{D+D'}}) \right\rangle$, the predictive probability of a single observation given the training data and the new observations up to now is:

$$\mathrm{P}\left(z'_i \middle| \vec{z}, \vec{z}'_{<i}, \overrightarrow{(w,\vec{p})}, \overrightarrow{(w,\vec{p})}'_{<i}, H\right) \propto f_\phi^+ \cdot f_\theta^+ \cdot f_\Omega^+ =$$

$$f_\phi^+ \cdot f_\theta^+ \cdot \prod_{j=1}^{K_{w'_i}} \frac{A'^{<i}_{z'_i}(p'_{i,j}) + A_{z'_i}(p'_{i,j}) + \gamma o_{p'_{i,j}}}{\sum_{q=1}^{Q} \left(A'^{<i}_{z'_i}(q) + A_{z'_i}(q)\right) + \gamma}$$

As with LDA for AR, we must perform Gibbs sampling on the previous new observations' activity assignments $\vec{z}'_{<i}$.

3.2 Composite Model

The composite model (Griffiths et al. 2004) integrates LDA with a HMM by setting one of the C states to call LDA for selecting a word; we will remain consistent with Griffiths et al.'s notation and let this be state 1. The remainder of the states contain their own distributions $\vec{\varphi} = \{\varphi_2, \ldots, \varphi_C\}$ over the vocabulary of word tokens so that they may select words during the document generation. It has been empirically supported that most the probability mass of $\vec{\varphi}$ is found about stopwords, tokens with very high frequencies that usually have to be removed before running LDA. Otherwise, they often appear in every topic by random chance. Stopwords typically serve a syntactic purpose in documents rather than a semantic purpose (which is what LDA captures). The HMM is *able to capture structure through its dependency on the previous state in the latent Markov chain* represented by the transition functions $\vec{\xi} = \{\xi_0, \xi_1, \ldots, \xi_C\}$ where the initial state of the chain is determined by distribution ξ_0. The graphical model is formed by the black and blue subgraphs in Fig. 2 and has generative process composed of lines 1, 2, 4-7, and 8-17 of Algorithm 1. Unlike the notation, we must use different variable names from Griffiths et al. since we already defined parameterized LDA.

In AR and PR, it is possible to encounter "stopwords" if some subset of poses is very common in the observation sequences (Freedman, Jung, and Zilberstein 2014). Although these are typically removed prior to training and testing, the framework laid out by these poses can be beneficial for recognition tasks. A transition between certain states may be used as a boundary if the observation sequence needs to be segmented into distinct activities. It also provides an ordering for the poses associated with activities so that some structures such as loops in plans may be easier to identify.

Similar to LDA, we only observe the poses $\vec{w}$ extracted from the postural data read by the RGB-D sensor. Unlike LDA, we now have two latent variables per observation which we need to infer: the latent activity variables $\vec{z}$ and the latent HMM state variables $\vec{s}$. Because Gibbs sampling only samples one random variable at a time, we must alternate between sampling z_i and s_i. As in the other topic models, we use the conjugate prior to approximate the sampling likelihood of the latent activity:

$$\mathrm{P}\left(z_i \middle| \vec{z}_{\setminus i}, \vec{w}, \vec{s}, H\right) \propto f_\phi^{\mathbf{1}(s_i=1)} \cdot f_\varphi^{\mathbf{1}(s_i\neq 1)} \cdot f_\theta$$

$$= f_\phi^{\mathbf{1}(s_i=1)} \cdot \left(\frac{N_{s_i}^{\setminus i}(w_i) + \delta u_{w_i}}{\sum_{v=1}^{V} N_{s_i}^{\setminus i}(v) + \delta}\right)^{\mathbf{1}(s_i\neq 1)} \cdot f_\theta$$

where $\mathbf{1}(x)$ is the indicator function that returns 1 if x is true and 0 otherwise, $N_c^{\setminus i} : \{1, \ldots, V\} \rightarrow \mathbb{Z}^{\geq 0}$ is the number of times pose v is assigned HMM state c excluding the pose at the sampled index, and $N_t^{\setminus i}$ only considers poses assigned to HMM state 1 (i.e., word tokens used in LDA). Because the current HMM state is fixed during this computation, $f_\varphi^{\mathbf{1}(s_i\neq 1)} \cdot \left(\sum_{t=1}^{T} N_d^1(t) + \alpha\right)$ is constant with respect to the sampled random variable and may thus be omitted

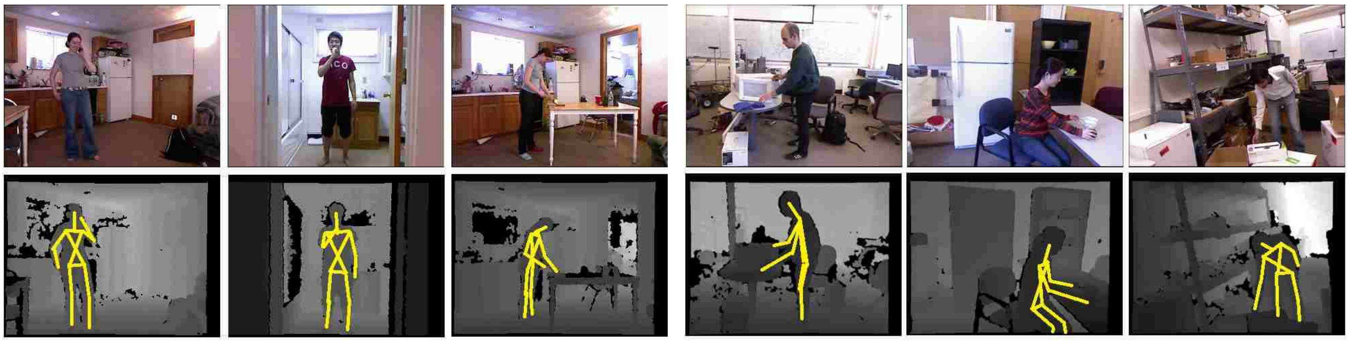

Figure 3: A selection of screenshots and extracted human postures from the CAD-120 Dataset. Image taken from the repository at *http://pr.cs.cornell.edu/humanactivities/images/all_activity_pic_combined.jpg*.

from the sampling likelihood. After reassigning the latent activity, we sample the HMM state using the sampling likelihood:

$$\mathrm{P}\left(s_i \,\middle|\, \vec{s}_{\backslash i}, \vec{w}, \vec{z}, H\right) \propto f_{\phi}^{\mathbf{1}(s_i=1)} \cdot f_{\varphi}^{\mathbf{1}(s_i\neq 1)}$$

$$\cdot \frac{F_{s_{i-1}}^{\backslash i \backslash i+1}(s_i) + \varepsilon \pi_{s_i}}{\sum_{c=1}^{C} F_{s_{i-1}}^{\backslash i \backslash i+1}(c) + \varepsilon} \cdot \left(\frac{F_{s_i}^{\backslash i+1}(s_{i+1}) + \varepsilon \pi_{s_{i+1}}}{\sum_{c=1}^{C} F_{s_i}^{\backslash i+1}(c) + \varepsilon}\right)^{\mathbf{1}(i < N_d)}$$

where $F_{c_1}^{\backslash i} : \{1, \ldots, C\} \to \mathbb{Z}^{\geq 0}$ is the number of times the transition from HMM state c_1 to HMM state c has occurred excluding the transitions both to and from the HMM state at the sampled index. For this sampling likelihood, only $\left(\sum_{c=1}^{C} F_{s_{i-1}}^{\backslash i \backslash i+1}(c) + \varepsilon\right)$ is constant with respect to s_i so that it may be ignored during the computation. After training, the predictive probability of an observation in $\vec{w}'$ given the training data and new observations up to the current one requires computing the joint probability of both the latent activity and the latent HMM state:

$$\mathrm{P}\left(z_i{}', s_i{}' \,\middle|\, \vec{z}, \vec{z}'_{<i}, \vec{s}, \vec{s}'_{<i}, \vec{w}, \vec{w}'_{<i}, H\right) \propto f_{\theta}^{+} \cdot f_{\xi}^{+} \cdot \left(f_{\phi}^{+}\right)^{\mathbf{1}(s_i=1)}$$

$$\cdot \left(f_{\varphi}^{+}\right)^{\mathbf{1}(s_i\neq 1)} = f_{\theta}^{+} \cdot \frac{F'^{<i}_{s'_{i-1}}(s'_i) + F_{s'_{i-1}}(s'_i) + \varepsilon \pi_{s'_i}}{\sum_{c=1}^{C}\left(F'^{<i}_{s'_{i-1}}(c) + F_{s'_{i-1}}(c)\right) + \varepsilon}$$

$$\cdot \left(f_{\phi}^{+}\right)^{\mathbf{1}(s_i=1)} \cdot \left(\frac{N'^{<i}_{s'_i}(w'_i) + N_{s'_i}(w'_i) + \delta u_{w'_i}}{\sum_{v=1}^{V}\left(N'^{<i}_{s'_i}(v) + N_{s'_i}(v)\right) + \delta}\right)^{\mathbf{1}\left(s'_i \neq 1\right)}$$

When performing Gibbs sampling on the previous new observations, we again alternate between activity assignments $\vec{z}_{<i}$ and HMM state assignments $\vec{s}_{<i}$.

3.3 Parameterized Composite Model

The parameterized composite model combines the parameterized LDA topic model with the composite model by simply having the HMM state call parameterized LDA instead of LDA for generating the next word. However, we assume that object the parameters only have semantic information and cannot be used for syntax. Thus the parameters are generated from the latent topic even if the HMM state is not 1. From a PR perspective, we interpret this as the passing of arguments between consecutive actions in a plan. When local ordering between actions is necessary, the order is usually important because a subset of the next action's prerequisites are only satisfied by the effects of the current action. This is the key idea behind the classic UCPOP planner (Penberthy and Weld 1992). Fig. 2 displays this hybrid graphical model and Algorithm 1 explains the generative process.

The parameters are conditionally independent of the pose and HMM state given the activity so that we simply multiply the sampling likelihood of the activity by the approximate likelihood of the parameters given the topic:

$$\mathrm{P}\left(z_i \,\middle|\, \vec{z}_{\backslash i}, \overrightarrow{(w, \vec{p})}, \vec{s}, H\right) \propto f_{\phi}^{\mathbf{1}(s_i=1)} \cdot f_{\varphi}^{\mathbf{1}(s_i\neq 1)} \cdot f_{\theta} \cdot f_{\Omega}$$

Due to the independence assumptions depicted by the directed edges in Fig. 2, the sampling likelihood for the HMM state does not change from the composite model. The new terms provided by the parameters $\vec{p}_i$ are constant with respect to all HMM state random variables as long as the activity z_i is observed. On the other hand, the joint predictive probability does receive an update based on the observed parameters and pose:

$$\mathrm{P}\left(z'_i, s'_i \,\middle|\, \vec{z}, \vec{z}'_{<i}, \vec{s}, \vec{s}'_{<i}, \overrightarrow{(w, \vec{p})}, \overrightarrow{(w, \vec{p})}'_{<i}, H\right) \propto$$

$$f_{\theta}^{+} \cdot \left(f_{\phi}^{+}\right)^{\mathbf{1}(s_i=1)} \cdot f_{\Omega}^{+} \cdot \left(f_{\varphi}^{+}\right)^{\mathbf{1}(s_i\neq 1)} \cdot f_{\xi}^{+}$$

4 Experiments

We implemented the four models described above such that each component is a module with the same underlying framework as portrayed by Fig. 2 and their formulas throughout Section 3. Although not as efficient as some state-of-the-art implementations of LDA, these implementations offer a basis for a fair comparison of the methods with respect to their relative runtimes and performance. These factors are important for PR and AR systems when used in actual applications because there are often real-time constraints on the machine's response time. A computer or robot must be able to successfully identify the observed user's actions and/or plan before providing a valid response. Hence we focus on the relative trade-offs between performance and runtime for these models to investigate to what extent including temporal and object information is practical. We assume that the proportions will hold if more efficient implementations are made using state-of-the-art versions of LDA such as Mallet (McCallum 2002).

For these comparison tests, we use the standard Cornell Activity Dataset 120 (CAD-120) (Lab 2013). It contains 124 recordings of short plan executions (which they call activities) with a total duration of approximately eleven minutes. Fig. 3 illustrates a sample of the activities performed in the dataset along with the extracted stick figure representation of the observed agent. Each RGB-D recording is fully annotated with orientation and position information for the acting agent's posture, objects used, object affordance labels based on how they are used in each frame, activity labels, and segmented subactivity labels for supervised learning.

We ignore the activity and subactivity labels when converting the dataset to a corpus of documents since our models are intended for unsupervised learning approaches. We use each frame's orientation data to generate the observed agent's posture as a word token using the modified joint-angle representation described above with granularity parameter 21. This gives us a vocabulary containing $V = 42588$ unique word tokens out of 65133 total in the corpus. Objects in each frame are depicted using a two-dimensional bounding box from the RGB image based on SIFT features without using depth. Hence we are only able to identify parameters as objects whose bounding boxes are within 150 millimeters of a joint of the observed user in the x- and y-directions (this accounts for the bounding box not always capturing the entire object). Due to the design of CAD-120, the lack of the z-direction in these proximity calculations does not greatly affect the list of parameters for each word.

To ensure optimal performance of each topic model with CAD-120, we trained the topic models using a sweep of parameter settings for T and C and then selected the ones yielding the greatest log-evidence of generating the training dataset. While hyperparameters α, β, γ, δ, ε, and $\vec{m}$ were optimized throughout the Gibbs sampling process, others could not be optimized due to biases they introduced during training. The number of activities T and HMM states C were held constant once initialized. Initial hyperparameter concentration values were always set to $\alpha = T$, $\beta = \delta = 0.02V$, $\gamma = 2Q$ where $Q = 10$, and $\varepsilon = C$; prior means $\vec{m}$, $\vec{n}$, $\vec{o}$, $\vec{u}$, $\vec{\pi}$ were always set to a uniform distribution. A burn-in period of twenty-five iterations was applied to make sure that the state sequences were truly random before optimizing ε in the transition functions. We trained the models on 99 sequences (80%) and then tested them on the remaining 25. To avoid an anomaly, five such partitions were randomly generated a priori for use in each parameter setting, and the same training and testing partitions were used across all models for a fair comparison. The number of word tokens in each partition $P1$ through $P5$'s test set is 11388, 12438, 14855, 12406, and 11647 respectively.

4.1 Runtime Performance

The empirical time to test and train each model is displayed in seconds. We also provide the number of Gibbs sampling iterations used to converge to the maximized log-evidence during training since this also varied per model and likely had an impact on the runtime. Since the train-test set choice did not impact runtime significantly, we only present the results of partition $P1$ in Table 1 due to space limitations.

Table 1: Elapsed Runtimes for Optimally-Trained Models ($P1$)

	Gibbs (iter)	Train (sec)	Test (sec)
LDA	50	362.398	1318.926
PLDA	100	1122.416	2411.291
Comp	540	4080.849	2915.388
PComp	540	4945.438	3436.523

Table 2: Log-Evidence of Test Set with Optimally-Trained Models

	LDA	PLDA	Comp	PComp
$P1$	-112096.2	-111055.4	-111097.6	-109512.0
$P2$	-123786.9	-123084.8	-122722.0	-122156.1
$P3$	-148437.2	-147803.2	-146915.9	-146763.1
$P4$	-121191.7	-119860.2	-118929.7	-119477.9
$P5$	-115952.3	-114871.8	-113931.9	-114142.1

Although the training set is larger than the testing set, the testing times take longer for the less computationally intensive models due to the resampling of every newly observed pose before classifying the next observation. These times show that the inclusion of temporal relations can greatly increase the training time needed to perform PR and AR. On the other hand, the inclusion of object relations appears to increase the amount of time to a lesser degree. Any of the models extending LDA take about two times longer or more to perform recognition, though.

4.2 Recognition Performance

We measure the recognition performance by the log-evidence of the testing set after training. For generative models, this value tells us the log of the probability that following the step-by-step procedures described within Algorithm 1 would have actually generated the observation sequences in the testing sequence. A higher log-evidence implies that the model is a better fit for the observed data. Table 2 provides these log-evidence values. In all cases, the modifications to LDA develop a model which better fits the data. This implies that LDA itself is a simplification of the generation process and omits information which was used for determining how to solve the task. Although the results provide evidence for the case that temporal relations are more informative than object relations, we believe this may not be concluded due to CAD-120's method for annotating objects (SIFT features in two-dimensional space); a dataset with more precise recording of objects in the environment would be necessary for confirmation. It is particularly worth noting that the parameterized composite model outperforms both parameterized LDA and the composite model in three of the train-test partitions and outperforms one of them in the other two partitions. Thus the *use of both temporal and object relations appears to often be more informative than either relation alone*. This synergistic effect supports our hypothesis that these relations contain mutually exclusive information regarding the observed agent's activity.

Table 3: Number of activities T and HMM states C in the optimally trained models for each partition of CAD-120.

	$P1$	$P2$	$P3$	$P4$	$P5$
LDA	$T = 61$	$T = 63$	$T = 45$	$T = 59$	$T = 55$
PLDA	$T = 89$	$T = 83$	$T = 89$	$T = 89$	$T = 89$
Comp	$T = 62$ $C = 4$	$T = 62$ $C = 6$	$T = 62$ $C = 2$	$T = 62$ $C = 5$	$T = 62$ $C = 6$
PComp	$T = 92$ $C = 6$	$T = 92$ $C = 4$	$T = 89$ $C = 3$	$T = 92$ $C = 2$	$T = 86$ $C = 3$

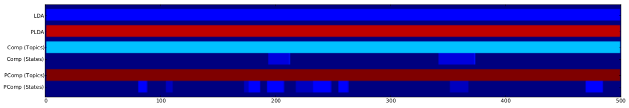

Figure 4: Visualization of each model's inferred topic and HMM state assignments for an execution of 'having a meal' which breaks down into 'moving' (1-83; 101-134; 156-194; 241-276; 346-370; 385-444), 'eating' (84-100; 371-384), 'reaching' (135-155; 333-345), 'drinking' (195-240), and 'placing' (277-332).

4.3 Topic and State Investigation

Because unsupervised learning algorithms identify their own patterns in data, it is important to study the learned clusters and ensure that the results are coherent. The learned topic and state distributions may not resemble what a human would classify as a distinct category, but some distinctions should be evident within and between the distributions. This will also provide us with an opportunity to compare what kinds of information are captured by the different models. To begin, we refer to Table 3 for the actual number of topics and HMM states used in the optimally trained models for each train-test partition. We observe that the models extending LDA were assigned very similar numbers of topics T regardless of the partition. This number is the edge case of our parameter sweep which implies that increasing the range would have further improved performance. However, the models' log-likelihoods of generating the training data were reaching the asymptotic limit and the log-evidence would most likely only experience marginal increase.

We thus want to identify in which ways the different topic models classify the plan executions in CAD-120. Because they have different log-evidence values, it would seem to be the case that each one is recognizing something unique in comparison to the other models. Fig. 4 displays an inferred breakdown by topic and HMM state for a plan execution in the training set of $P1$. The first thing to notice is that LDA infers a single topic for all postures in a single execution. This implies that it is only able to classify the overall activity, but we must consider that the length of an average recording in CAD-120 is less than twenty seconds (600 frames). Although Freedman, Jung, and Zilberstein (2014) explained that the entire execution represents the plan so that θ is a distribution of actions/activities throughout the entire recording, it is possible that these documents could be too short for such analysis which would generate unimodal θ (making PR and AR identical processes). Despite this potential setback, we use it since there are no other datasets to our knowledge that record object relations.

We also observe that even the parameterized variations infer a single topic for the entire document. However, these topics are different from those in LDA due to the inclusion of object relations. Table 3 shows that there are more topics in these models, allowing more precise cases during recognition, and this diversity may be confirmed by viewing the learned object distributions $\vec{\Omega}$ for each topic. Fig. 5 illustrates $\vec{\Omega}$ for the optimally trained parameterized LDA on $P1$ as a heat map (the parameterized composite model's heat map is very similar). We identify three distinct types of distributions. The unimodal distribution contains one red dot per column. This likely indicates an activity whose poses commonly interact with a single object. The uniform distribution is a solid blue column. This presents a lack of preference for objects which seems to indicate that either no objects are involved or the poses are distinct enough from those in other activities that the objects involved do not matter. The bimodal distribution has two dots each ranging in hue from light blue to orange. We only find this type pairing the objects 'box' and 'bowl' which are commonly found together in the 'making cereal' activities in CAD-120.

Hence the uniform activity inferences are still more specific in their description due to the pose-object pairs. Although we hypothesized that objects with similar affordances would be clustered together in each activity's Ω distribution, it may be the case that ten objects are too few to identify these higher-level functional purposes. CAD-120 only provides annotated affordance labels based on the activity so that a single object will appear to change during the plan execution if we were to use their affordance labels in place of the objects themselves. *A more robust dataset with longer plan execution recordings and a greater variety of objects will be necessary for a full analysis of the impact of object relations in our models for unsupervised PR and AR.*

On the other hand, the inferred states by the (parameterized) composite model have a more distinguishable trend. The majority of the inferences are dark blue which represent state 1 where (parameterized) LDA is used for sampling the posture/word token. However, there are a few streaks where the color changes and a different state is inferred. Fig. 4's caption reveals that most the streaks appear as a transition between two annotated subactivities. This alludes to the syntactic properties observed by Griffiths et al. (2004) when they introduced the composite model for analyzing text documents. This appears to imply that despite the shorter length, *the underlying framework may be learned even if the actual subactivities cannot be distinguished*. The learned transition functions $\vec{\xi}$ were almost always unimodal favoring state 1. If it did not, then it favored transitioning back to itself (which would explain the streak of a single color rather than a variety of colors). As both $\vec{\xi}$ and $\vec{\Omega}$ were similar between the parameterized composite model and its respective submodels, there appears to be little overlap between the information gained from temporal and object relations.

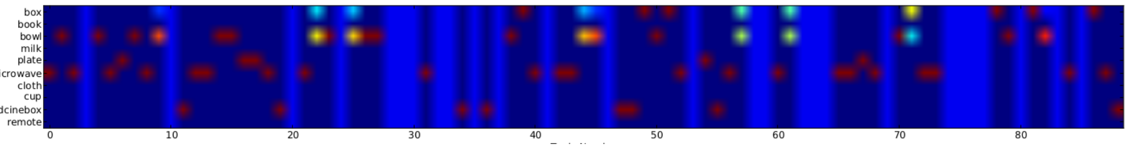

Figure 5: A heat map for the distributions over objects Ω_t for each activity t in the optimally trained parameterized LDA with respect to $P1$. Each column is single topic where red is greatest probability mass and dark blue is least probabilty mass.

5 Discussion

We have presented variations of LDA that incorporate temporal- and/or object-related information in an attempt to improve its performance in PR and AR tasks. Many applications involve interaction with humans so that recognition needs to be performed as accurately as possible. As these applications are very broad and diverse, unsupervised methods such as the ones proposed are necessary to avoid annotating large collections of training data for each situation and environment. Initial results suggest that temporal relations can improve performance at a high computational cost while object relations may have a less profound improvement with little additional computational cost, but both relations together usually improve the performance the most.

Future Work There are many directions in which this work may be continued. Besides additional experimentation, it will be useful to consider other models that can capture these relations such as Topics over Time (Wang and McCallum 2006) and the Bigram Topic Model (Wallach 2006). There are also algorithms developed specifically for speeding up Gibbs sampling during LDA (Porteous et al. 2008; Steele, Tassarotti, and Tristan 2015), and it would be very beneficial if these algorithms can be generalized to our variations to handle the observed runtime increase. One recently studied optimization method which we will strongly consider is the use of acceleration via GPUs which can impact the runtime by approximately two orders of magnitude (Steele, Tassarotti, and Tristan 2015) — such runtime boosts would make handling real-time constraints with our proposed models feasible. There would then be sufficient time to also study alternative PR methods. One such method could take advantage of the composite model's segmentation for matching linear combinations of segments instead of just $\vec{\theta}$. As real-time constraints are also important to test in practice, we plan to develop more efficient implementations and test our models using real software and robots in interactive simulations with humans. Furthermore, we mentioned earlier that object relations can be used to extract additional information for use in planning systems that determine response behaviors. Related questions include how to extract this information and what representations are best to use.

6 Acknowledgments

This work was supported in part by the National Science Foundation grant number IIS-1405550.

References

Andrzejewski, D.; Zhu, X.; Craven, M.; and Recht, B. 2011. A framework for incorporating general domain knowledge into latent Dirichlet allocation using first-order logic. In *Proc. of the 22nd International Joint Conference on Artificial Intelligence*, 1171–1177.

Blei, D. M., and McAuliffe, J. D. 2007. Supervised topic models. In *Neural Information Processing Systems*, volume 7, 121–128.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022.

Bui, H. H.; Phung, D. Q.; and Venkatesh, S. 2004. Hierarchical hidden Markov models with general state hierarchy. In *Proc. of the 19th National Conference on Artificial Intelligence*, 324–329.

Bui, H. H.; Venkatesh, S.; and West, G. 2002. Policy recognition in the abstract hidden Markov model. *Journal of Artificial Intelligence Research* 17(1):451–499.

Cheng, D. C., and Thawonmas, R. 2004. Case-based plan recognition for real-time strategy games. In *Proc. of the 5th Game-On International Conference*, 36–40.

Chikhaoui, B.; Wang, S.; and Pigot, H. 2012. ADR-SPLDA: Activity discovery and recognition by combining sequential patterns and latent dirichlet allocation. *Pervasive and Mobile Computing* 8(6):845–862.

Cook, D.; Krishnan, N.; and Rashidi, P. 2013. Activity discovery and activity recognition: A new partnership. *IEEE Transactions on Cybernetics* 43(3):820–828.

De la Torre, F.; Hodgins, J.; Montano, J.; Valcarcel, S.; Forcada, R.; and Macey, J. 2009. Guide to the Carnegie Mellon University multimodal activity (CMU-MMAC) database. Technical report, Robotics Institute, CMU.

Fine, S.; Singer, Y.; and Tishby, N. 1998. The hierarchical hidden Markov model: Analysis and applications. *Machine Learning* 32(1):41–62.

Freedman, R. G.; Jung, H.-T.; and Zilberstein, S. 2014. Plan and activity recognition from a topic modeling perspective. In *Proc. of the 24th Int'l Conference on Automated Planning and Scheduling*, 360–364.

Geib, C. W., and Steedman, M. 2007. On natural language processing and plan recognition. In *Proc. of the 20th International Joint Conference on Artificial Intelligence*, 1612–1617.

Gibson, E. 2001. *Perceiving the Affordances: A Portrait of Two Psychologists*. Taylor & Francis.

Griffiths, T. L.; Steyvers, M.; Blei, D. M.; and Tenenbaum, J. B. 2004. Integrating topics and syntax. In *NIPS*, 537–544.

Griffiths, T. 2002. Gibbs sampling in the generative model of latent Dirichlet allocation. Technical report, Stanford University.

Huỳnh, T.; Fritz, M.; and Schiele, B. 2008. Discovery of activity patterns using topic models. In *Proc. of the 10th International Conference on Ubiquitous Computing*, 10–19.

Jain, R., and Inamura, T. 2013. Bayesian learning of tool affordances based on generalization of functional feature to estimate effects of unseen tools. *Artificial Life and Robotics* 18(1-2):95–103.

Jung, H.-T.; Freedman, R. G.; Foster, T.; Choe, Y.-K.; Zilberstein, S.; and Grupen, R. A. 2015. Learning therapy strategies from demonstration using latent dirichlet allocation. In *Proc. of the 20th International Conference on Intelligent User Interfaces*, 432–436.

Kelley, R.; Nicolescu, M.; Tavakkoli, A.; King, C.; and Bebis, G. 2008. Understanding human intentions via hidden markov models in autonomous mobile robots. In *Proc. of the 3rd ACM/IEEE International Conference on Human-Robot Interaction*, 367–374.

Koppula, H. S., and Saxena, A. 2013. Learning spatio-temporal structure from RGB-D videos for human activity detection and anticipation. In *Proc. of the Int'l Conference on Machine Learning*.

Krstovski, K., and Smith, D. A. 2013. Online polylingual topic models for fast document translation detection. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, WMT'13, 252–261.

Lab, R. L. 2013. Cornell activity datasets: CAD-60 & CAD-120. http://pr.cs.cornell.edu/humanactivities/data.php. [Online].

Lösch, M.; Schmidt-Rohr, S.; Knoop, S.; Vacek, S.; and Dillmann, R. 2007. Feature set selection and optimal classifier for human activity recognition. In *Proc. of the 16th International Symposium on Robot and Human interactive Communication*, 1022–1027.

McCallum, A. K. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Mimno, D.; Wallach, H. M.; Naradowsky, J.; Smith, D. A.; and McCallum, A. 2009. Polylingual topic models. In *Proc. of Empirical Methods in Natural Language Processing*, 880–889.

Mörtl, A.; Lawitzky, M.; Kucukyilmaz, A.; Sezgin, M.; Basdogan, C.; and Hirche, S. 2012. The role of roles: Physical cooperation between humans and robots. *Int. J. Rob. Res.* 31(13):1656–1674.

Penberthy, J. S., and Weld, D. S. 1992. UCPOP: A sound, complete, partial order planner for ADL. In *Proc. of the 3rd Int'l Conference on Knowledge Representation and Reasoning*, 103–114.

Porteous, I.; Newman, D.; Ihler, A.; Asuncion, A.; Smyth, P.; and Welling, M. 2008. Fast collapsed gibbs sampling for latent Dirichlet allocation. In *Proceedings of the Fourteenth ACM International Conference on Knowledge Discovery and Data Mining*, 569–577.

Rieping, K.; Englebienne, G.; and Kröse, B. 2014. Behavior analysis of elderly using topic models. *Pervasive and Mobile Computing* 15(0):181–199.

Song, Y. C.; Kautz, H.; Allen, J.; Swift, M.; Li, Y.; Luo, J.; and Zhang, C. 2013. A Markov logic framework for recognizing complex events from multimodal data. In *Proc. of the 15th ACM International Conference on Multimodal Interaction*, 141–148.

Steele, G. S.; Tassarotti, J.; and Tristan, J. 2015. Efficient training of LDA on a GPU by mean-for-mode Gibbs sampling. Private Communication.

Steyvers, M., and Griffiths, T. 2007. Probabilistic topic models. In Landauer, T.; McNamara, S. D.; and Kintsch, W., eds., *Latent Semantic Analysis: A Road to Meaning*. Laurence Erlbaum.

Sukthankar, G.; Geib, C.; Bui, H. H.; Pynadath, D.; and Goldman, R. P. 2014. *Plan, Activity, and Intent Recognition: Theory and Practice*. Elsevier Science.

Sung, J.; Ponce, C.; Selman, B.; and Saxena, A. 2012. Unstructured human activity detection from RGBD images. In *Proc. of the IEEE International Conference on Robotics and Automation*, 842–849.

Synnaeve, G., and Bessière, P. 2011. A Bayesian model for plan recognition in RTS games applied to starcraft. In *Proc. of the 7th International Conference on Artificial Intelligence and Interactive Digital Entertainment*, 79–84.

Viterbi, A. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory* 13(2):260–269.

Wallach, H. M. 2006. Topic modeling: Beyond bag-of-words. In *Proc. of the 23rd Int'l Conference on Machine Learning*, 977–984.

Wang, X., and McCallum, A. 2006. Topics over time: A non-Markov continuous-time model of topical trends. In *Proc. of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 424–433.

Wang, Y., and Mori, G. 2009. Human action recognition by semi-latent topic models. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 31(10):1762–1774.

White, B. A.; Blaylock, N.; and Bölöni, L. 2009. Analyzing team actions with cascading HMM. In *Proc. of the 22nd Int'l Florida Artificial Intelligence Research Society Conference*, 129–134.

Zhang, H., and Parker, L. E. 2011. Four-dimensional local spatio-temporal features for human activity recognition. In *Proc. of the International Conference on Intelligent Robots and Systems*, 2044–2049.