Temporal and Object Relations in Plan and Activity Recognition for Robots Using Topic Models

Richard G. Freedman and **Hee-Tae Jung** and **Shlomo Zilberstein**
School of Computer Science
University of Massachusetts Amherst
Amherst, MA 01003, USA
{freedman, hjung, shlomo}@cs.umass.edu

Introduction

Plan recognition (PR) and activity recognition (AR) systems are essential for effective human-robot interaction (HRI) since the robot needs to predict what other agents in the environment are doing (Lösch et al. 2007). Even when robots are designed to perform simple tasks such as lending an object to a person (Levine and Williams 2014), they cannot follow simple time-stamped commands. There is often considerable uncertainty about the way in which people operate and the duration of time they need to complete each action. Consequently, a robot that needs to perform complementary actions must be able to observe and understand what the person is doing, and the AR process must be executed in real time in order to respond to the current situation in a timely manner. Topic models such as Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003) have been used for this purpose (Sung et al. 2012; Zhang and Parker 2011). The goal of this work is twofold. First, we explore a new way of representing RGB-D sensor readings for use with LDA as an integration of both a PR and AR system. We follow this with a proposition of extensions of LDA that take temporal and/or object relational information into account during the recognition process.

Topic Modeling for Recognition Tasks

The prevailing techniques to achieve real-time PR and AR performance often employ 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. However, the HMM's latent Markov chain emphasizes temporally local relations between observations; that is, the current state heavily relies on the order of the recently observed states in the sequence. Humans seldom act in such a structural manner. As shown by partially ordered plans, some actions have preconditions and effects that allow them to be performed independently. Hence human agents may perform subtasks in an order that the robot did not learn, or the human may perform some extraneous actions that would serve as noise in the execution sequence. In real-time systems, considering all the execution sequences with extraneous actions and independent subsequence orderings can be a bottleneck.

Latent variable mixture models such as topic models omit structure completely and use bag-of-words models. This assumes that every sensor observation is independent of one another for a more global relationship that relies on the distribution of all observations in the sequence. By treating plans, sequences of actions that solve a task, as a bag-of-words, we analogously use LDA to learn activities in the plan as distributions over the observations; the learned activity topics can be used for AR (Huỳnh, Fritz, and Schiele 2008). The distribution of topics approximates the "gist" of the plan which may be used in PR as a guide or heuristic when identifying the executed plan in a library of plans.

Representing RGB-D Data for Topic Modeling

The raw data recorded by the RGB-D sensor is in the form of homogeneous transform matrices that specify how the coordinates change position between frames. From these matrices, we derive sets of triples representing the human body at key points of motion called joint-angles. We consider fifteen joints so that every human pose is in $[-\pi, \pi]^{45}$. This is an uncountably infinite vocabulary with a very small likelihood of duplicates, but finding each activity's distribution over the poses requires a countable vocabulary with some duplicates in the collection of plan executions. Wang and Mori (2009) and Zhang and Parker (2011) created codebooks to accomplish this by clustering their formatted inputs and selecting the center of each cluster as a token in their vocabulary — all inputs in the same cluster are assigned this token value. However, the codebook is not updated after training so that all unique inputs in the test set are assigned a token from the training set. This makes the recognition systems strongly domain dependent since any changes to the environment and/or observed agents will not be considered accordingly.

We make the vocabulary finite and increase the likelihood of duplicate poses by *discretizing the space of all poses* with respect to a *granularity* parameter rather than deriving it from the training data. For granularity $g \in \mathbb{N}$, we map each angle a to integer $0 \leq i < g$ such that $(i\,/\,g) \cdot 2\pi \leq a + \pi < ((i+1)\,/\,g) \cdot 2\pi$. This reduces the vocabulary to $\{0, 1, \ldots, g-1\}^{45}$ which is still large for small g, but we must consider that many poses do not represent feasible

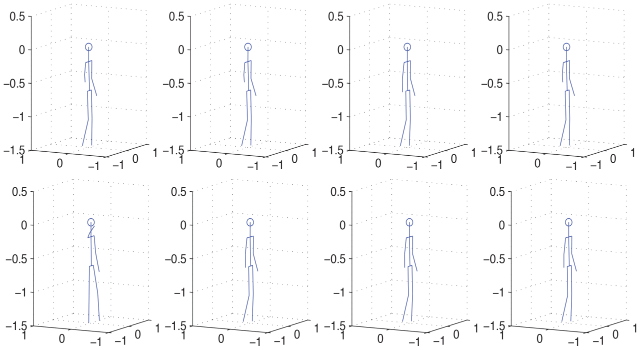

Figure 1: Most likely poses for topic ten from the fifteen-topic model at granularity thirty-five. The poses appear to be walking.

body postures. An advantage of discretizing the space over the use of a codebook is that it is possible to encounter new poses for which the system was not trained. This is important for domain independent HRI applications where PR and AR will be used to observe a constantly changing set of users.

While larger granularities clearly reduce the number of duplicate poses, another interesting factor to consider is its parity (Freedman, Jung, and Zilberstein 2014). This phenomenon may be explained through kinematics. When an even granularity is used to discretize the space, small joint movements near the vertical axis (where $a = 0$) will be assigned to one of two different groups: $(g\,/\,2)$ if $a \geq 0$ and $(g\,/\,2) - 1$ if $a < 0$. On the other hand, an odd granularity will always assign these movements to $((g - 1)\,/\,2)$. For naturally small body movements and oscillations about the vertical axis such as an arm swaying slightly at the user's side, the mapping between two groups rather than one creates significantly more combinations for even granularities compared to odd ones. Figure 1 shows a sample interpretable topic produced using granularity thirty-five.

Incorporating Temporal and Object Relations

While the proposed use of LDA has shown empirical success in learning activities whose poses represent distinct actions (Freedman, Jung, and Zilberstein 2014), there are still issues which must be addressed to improve the performance of the integrated PR and AR system. The more evident concern arises from the bag-of-words model's independence assumption. Without any dependence on ordering, activities with cyclic structure are hard to recognize unless each iteration has a proportional contribution of observations to the global distribution, and there is clearly some degree of correlation between consecutive poses due to continuous motion of the body. Similar concerns have been raised in natural language processing, an area suggested to have much in common with plan recognition (Geib and Steedman 2007). 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 instead emphasize the semantic features of text without acknowledging its syntax. The *composite model* (Griffiths et al. 2004) was developed to bring HMM's and LDA together for a single model that takes both syntax and semantics into account. We believe 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. The extensions of LDA for the composite model are blue in the graphical model of Figure 2.

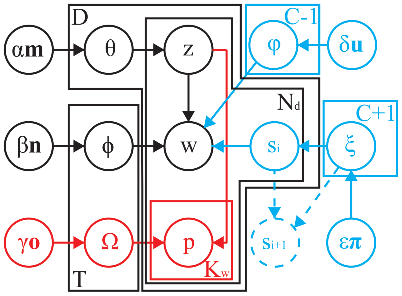

Figure 2: Graphical model representation of the parameterized composite model. The dotted lines represent conditional dependencies between random variables that are not visible due to the plate notation used.

Besides the lack of temporal information, we found that actions with similar poses such as squatting and jumping were clustered together since their inputs are indistinguishable (Freedman, Jung, and Zilberstein 2014). This can be attributed to a lack of environmental information such as nearby objects — squatting poses are closer to the ground than jumping poses. Another reason for including relations between observed users and objects in the environment is that the robot's interactions with the user will likely involve handling the same objects. For example, in joint tasks such as moving furniture (Mörtl et al. 2012), both agents will need to handle the same furniture object in order to coordinate carrying it. The generalized use of objects has also been addressed in the field of robotics under tool affordance (Gibson 2001; Jain and Inamura 2013; Jung, Ou, and Grupen 2010). We presently consider grounded object instances as parameters for a token. That is, a single word in a sequence 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. We assume that these arguments are elements of a *second vocabulary representing only the objects*. Our extension of LDA for these parameterizations are colored red in the graphical model of Figure 2.

Future Work

In addition to the new model proposed in this paper, there are other variations of LDA which have been developed to account for temporal information such as Topics over Time (Wang and McCallum 2006) and n-gram tokens (Wallach 2006). We will investigate which is most appropriate for postural input. Furthermore, real-time constraints are crucial for HRI and increasing model complexity will slow down the performance. Porteous et al. (2008) have developed a sampling method to improve the runtime of LDA, and we intend to extend their work for our models. Most importantly, we plan to incorporate our integrated PR and AR framework into an actual robot for testing in real environments with human subjects; it is necessary that our methods perform well empirically in the motivating domain.

References

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

Freedman, R. G.; Jung, H.-T.; and Zilberstein, S. 2014. Plan and activity recognition from a topic modeling perspective. In *Proceedings of the 24th International Conference on Automated Planning and Scheduling*.

Geib, C. W., and Steedman, M. 2007. On natural language processing and plan recognition. In *Proceedings 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 *Advances in Neural Information Processing 17*, 537–544.

Huỳnh, T.; Fritz, M.; and Schiele, B. 2008. Discovery of activity patterns using topic models. In *Proceedings 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.; Ou, S.; and Grupen, R. A. 2010. Learning to perceive human intention and assist in a situated context. In *Workshop on Learning for Human-Robot Interaction Modeling, Robotics: Science and Systems*.

Levine, S. J., and Williams, B. C. 2014. Concurrent plan recognition and execution for human-robot teams. In *Proceedings of the 24th International Conference on Automated Planning and Scheduling*.

Lösch, M.; Schmidt-Rohr, S.; Knoop, S.; Vacek, S.; and Dillmann, R. 2007. Feature set selection and optimal classifier for human activity recognition. In *Proceedings of the16th International Symposium on Robot and Human interactive Communication*, 1022–1027.

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

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 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, 569–577. New York, NY, USA: ACM.

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 *Proceedings of the IEEE International Conference on Robotics and Automation*, 842–849.

Wallach, H. M. 2006. Topic modeling: Beyond bag-of-words. In *Proceedings of the 23rd International 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 *Proceedings 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 Transactions on Pattern Analysis and Machine Intelligence Special Issue on Probabilistic Graphical Models in Computer Vision* 31(10):1762–1774.

Zhang, H., and Parker, L. 2011. 4-dimensional local spatio-temporal features for human activity recognition. In *Proceedings of the International Conference on Intelligent Robots and Systems*, 2044–2049.