How Robots Can Recognize Activities and Plans Using Topic Models

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

Abstract

The ability to identify what humans are doing in the environment is a crucial element of responsive behavior in human-robot interaction. We examine new ways to perform plan recognition (PR) using natural language processing (NLP) techniques. PR often focuses on the structural relationships between consecutive observations and ordered activities that comprise plans. However, NLP commonly treats text as a bag-of-words, omitting such structural relationships and using topic models to break down the distribution of concepts discussed in documents. In this paper, we examine an analogous treatment of plans as distributions of activities. We explore the application of Latent Dirichlet Allocation topic models to plan execution traces obtained from human postural data read by a RGB-D sensor. This investigation focuses on representing the data as text and interpreting learned activities as a form of activity recognition (AR). Additionally, we explain how the system may perform PR. The initial empirical results suggest that such NLP methods can be useful in complex PR and AR tasks.

1 Introduction

This paper presents an example in which techniques originally developed for *natural language processing* (NLP) can be used to allow robots to quickly recognize the activities performed by others. It has been suggested that *plan recognition* (PR) and natural language processing have much in common and are amenable to similar analyses. Geib and Steedman (2007) formally presented the following correspondence between PR and NLP:

- input is a set of observed actions (PR) or words (NLP),
- observations are organized into hierarchical data structures such as hierarchical task networks (HTNs, PR) or parse trees (NLP), and
- rules stating valid observation patterns for deriving the hierarchical data structure are represented through a library of plans (PR) or a grammar (NLP).

As implied by the HTN representation, PR techniques often focus on the structural relationships between consecutive observations and ordered activities that comprise plans. However, NLP commonly treats text as a *bag-of-words* and omits such structural relationships. A bag-of-words representation can loosely be related to a partially-ordered plan with no global ordering constraints. Local sequential ordering constraints can be clustered into a single word. This is similar to the way computer vision uses bag-of-words models with patches of pixels as a single word unit (Wang *et al.* 2006). Due to the combinatorial nature of representing all ordered sequences of partially-ordered plans, identifying them using rigidly structured recognition models such as HTNs, plan grammars (Geib and Steedman 2007), and hierarchical hidden Markov models (Fine *et al.* 1998; Bui *et al.* 2004) can be difficult. This means that many PR techniques are not easily able to recognize a large subset of plans, particularly those without a strong action ordering.

A now common method for studying the *distributions of topics* in bag-of-words models for text is *Latent Dirichlet Allocation* (LDA) (Blei *et al.* 2003). The topics used in LDA are themselves distributions over the vocabulary (set of words) pertaining to relevancy of concepts. Thus we examine ways to treat plans analogously – like bags-of-words – and analyze their distributions of topics using LDA. We hypothesize that, when the correct number of topics is selected, each topic will contain higher likelihoods of observed poses for a specific activity. The learned topics may additionally be used for *activity recognition* (AR), and we will focus on this aspect for the majority of the paper.

Wang and Mori (2009) performed a similar study using a variant of LDA with topic-annotated video data to perform AR. Their *Semilatent Dirichlet Allocation* model is a supervised method that predefines the topics and labels the image frames prior to learning. However, LDA itself is unsupervised and pixel-based representations can be vulnerable to confusion between postures as well as have difficulty accounting for scaling. Zhang and Parker (2011) also performed a similar study using LDA without modifications and a RGB-D sensor mounted on an actual robot. Their representation of the sensor input consists of identifying local spatio-temporal features and compacting them to vectors of four-dimensional cuboids. While this avoids the raster-image issues, they assign all the features to one of 600 discrete groups which is rather small (see Section 3). We instead consider pose information through human postural data obtained from a RGB-D sensor in order to avoid these representational drawbacks.

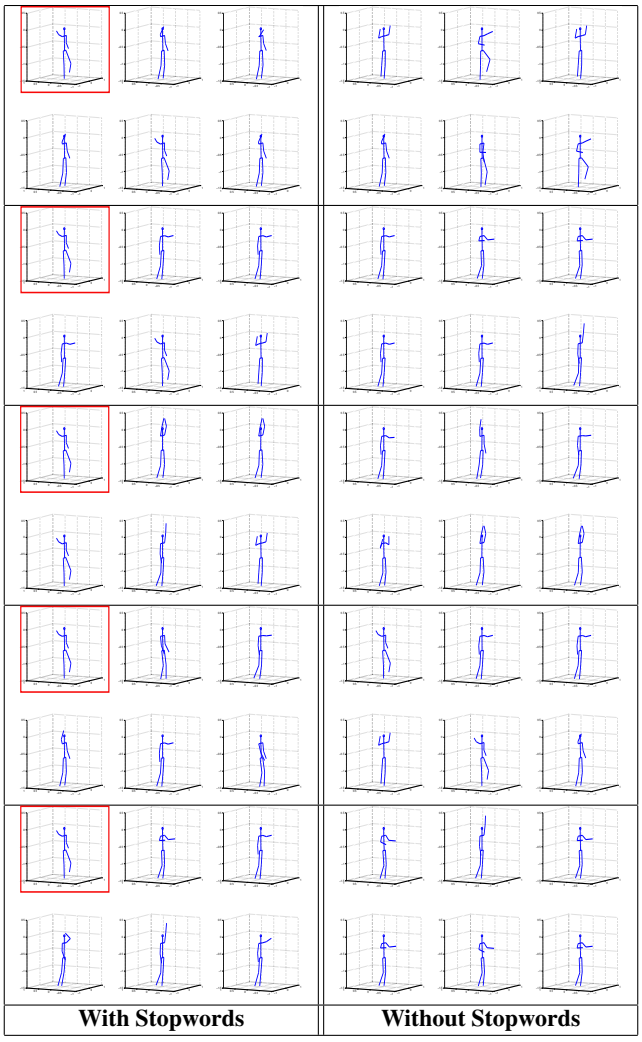

Figure 4: Most likely poses for the five-topic model using granularity three before and after removing stopwords. A red box is drawn around the stopword appearing in every topic.

number of topics will cluster many poses into a single activity yielding either an overarching theme (when reasonably small) or a collection of unrelated poses (when too small). A larger number of topics will sparsely store poses in each activity which will result in very specific actions or ambiguity where several actions are nearly identical. Hence we considered the following options: ten topics since we composed our documents using subsets of ten actions, fifteen topics in case the differences between left and right hands were distinguishable, and five topics since the lack of position data may make some poses look identical (such as standing and jumping).

With 13033 unique word tokens out of 16646, the distribution over poses and number of duplicate poses yielded good results for our corpus at granularity thirty-five. Figure 3 renders the most likely poses for four selected actions from the fifteen-topic model. The most likely poses captured in each topic are easily relatable to one-another and de-

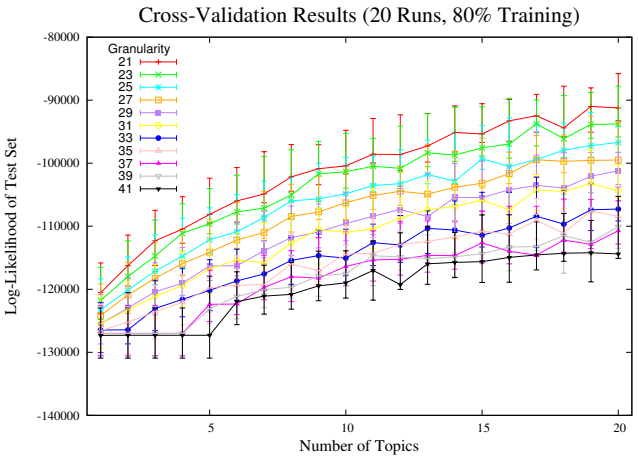

Figure 5: The results of running cross-validation on our dataset with twenty runs for each pairing of a subset of various granularities and numbers of topics.

pict particular actions. This typically holds for larger granularities. However, smaller granularities, especially with odd parity, appear to suffer from having too many duplicate poses that cluster into every action. This is due to their high frequency throughout all the recorded executions. This is especially the case for the generic standing pose at lower granularities; it accounts for almost half the word tokens in the corpus at granularity three. In NLP and information retrieval, such word tokens are referred to as *stopwords* and they are removed from the documents prior to training. By removing these stopwords, the actions are more easily distinguishable. Figure 4 shows the change in most common poses in the five-topic model with granularity three when all poses with frequency greater than 100 are regarded as stopwords – in particular, the most common pose in the top-left corner is no longer the same after removing the stopwords.

To explore the effectiveness of using different granularities and topics, we perform a twenty-run cross-validation over the forty recorded plan executions. In each run, we select thirty-two recordings (eighty percent) to train our LDA topic model and then use it to find the log-likelihood of the remaining eight recordings. The log-likelihood is derived from simulating the generative process described in Section 3 above; we use it to find the probability that the trained LDA topic model derives each of the eight recordings in the testing set. A plot of these results using up to twenty topics and odd granularities from twenty-one to forty-one is shown in Figure 5. We note two particular trends: (1) the overall log-likelihood decreases as the discretized space becomes more fine-grained and (2) the overall log-likelihood is increasing as more topics are used in the model.

The first trend is most likely a consequence of the increase in unique word tokens which also increases the chance of having poses exclusively appear in the test set. The model was not trained with such poses so that the probability of generating recordings with them is very low. The second trend typically implies that we should learn models with

more topics because we have not yet maximized the log-likelihood. However, we employ our knowledge of the domain to identify that twenty-topics is too many topics to consider. The recordings are only composed of approximately ten actions which should be analogous to a ten-topic model, but additional details can be extrapolated from the ten actions such as whether the left/right hand is used, whether the leg is bent or straight, and whether one's head is bent forwards or looking ahead. These extraneous features may be regarded as new topics when there are too many topics available in the model and could serve as a sign of overfitting the training set. Looking at the most likely poses for these activities provides evidence for this claim as there appear to be overlap between actions. We will look into formalizing signs of overfitting with respect to the number of topics in future research.

5 Discussion

Most plan recognition research has focused on the use of structural methods that enforce strict action ordering. However, many plans have partially ordered components and human agents can execute plans with extraneous actions that introduce noise. We investigated the treatment of plans as bags-of-words using sensor-level data from a RGB-D sensor by discretizing the information into a textual format that may then be analyzed using LDA topic models. This method shows potential for application in real-time PR and AR systems for HRI that can identify plans as distributions of actions just as natural language documents are composed of topics.

Future Research

This exploratory study has revealed several new directions for PR and AR research. One such direction involves taking advantage of the other data provided by the RGB-D sensor, primarily position. We only studied poses for our topic models in this work which resulted in ambiguities between some actions such as jumping and squatting or standing (when small like a hop). However, these nearly identical poses may be distinguished by their difference in vertical position. Likewise, we could identify orientation and destination which would enable us to integrate some of the past relational PR methods with our purely statistical method. A second direction is to investigate whether information from other types of sensors can yield word tokens to be applied to LDA for PR and AR. The RGB-D sensor's pose data represents a human form which is more intuitively mappable to activities, but other sensors may be able to provide equally useful information.

A third direction will be to perform a larger-scale study with more realistic parameters since the dataset used in this investigation only contains forty recorded plan executions in a controlled test environment. This would include more diverse plans, possible actions, and recorded subjects. There are also benchmark datasets that contain motion capture pose data such as the Carnegie Mellon Motion Capture Database (Hodgins). Although the encoded posture is different from the one retrieved by RGB-D sensors (the recorded joints are different), we plan to train a PR and AR system on them and possibly find a mapping between the encodings and/or generated word tokens. Not only would this give us access to a larger-scale study, but we would also have access to a large collection of data to train a more robust recognition system that may be used for plan recognition with a RGB-D sensor. It should also provide more insight into how many topics to use to best represent the data without overfitting.

Lastly, we are interested in modifying the LDA topic model to incorporate additional features besides just the pose data. For example, the objects with which users interact can have implications regarding the actions taken. This may disambiguate between poses as well; the aforementioned confusion between squatting and jumping would be easier to differentiate if it was known that the observed individual was using a jumprope. In addition to objects, subject features such as height and strength may also affect which actions people take to perform a planning task. We would be interested to see if this has any impact on the topic distributions. If the variation in topics is large enough between these features, then the different sets of available actions to each subject class may be regarded as different languages. Extensions of LDA such as Polylingual Topic Models (Mimno *et al.* 2009) exist that can be used for modeling topics across languages. It is important to know whether different groups of subjects should be considered differently when performing PR and AR so that general-purpose robots and other interaction systems will be better suited to cooperate with a greater variety of users.

This shows that besides the new directions for studying PR and AR, we must additionally consider how to integrate these systems with actual robots and use them in realistic situations. This raises questions regarding representation and real-time performance constraints. For example, how should the distribution of recognized activities from our topic model-based system be used for developing responsive behavior? After incorporating more contextual information such as objects, the actions could be converted into a STRIPS-like format for use in a planning system. Then for such a representation, to what extent can planning be performed alongside PR and AR to produce appropriate interactivitive experiences for those collaborating with the robot? A comparision of the inferred distribution (the plan) and the currently observed distribution may indicate which activities have yet to be performed. Whether those actions are not yet performed because the human is acting to satisfy some preconditions or because they form a subset of actions that may be performed in parallel (for partially ordered plans) will affect what goal conditions the robot should consider during planning. We will investigate these questions during our future endeavors. It is likely that the answers will influence research on PR and AR as much as this research will impact future work for fields in robotics such as HRI.

Acknowledgements

The authors thank the reviewers for their insightful comments that helped improve the manuscript. Support for this work was provided in part by NSF grant IIS-1116917 and ONR grant MURI-N000140710749.

References

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

Hung H. Bui, Dinh Q. Phung, and Svetha Venkatesh. Hierarchical hidden markov models with general state hierarchy. In *Proceedings of the 19th National Conference on Artificial Intelligence*, pages 324–329, 2004.

Shai Fine, Yoram Singer, and Naftali Tishby. The hierarchical hidden Markov model: Analysis and applications. *Machine Learning*, 32(1):41–62, 1998.

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

Robert P. Goldman, Christopher W. Geib, Henry Kautz, and Tamim Asfour. Plan Recognition – Dagstuhl Seminar 11141. *Dagstuhl Reports*, 1(4):1–22, 2011.

Raffay Hamid, S. Maddi, A. Bobick, and I. Essa. Structure from statistics - unsupervised activity analysis using suffix trees. In *IEEE 11th International Conference on Computer Vision*, pages 1–8, Oct 2007.

Jessica K. Hodgins. Carnegie mellon motion capture database. http://mocap.cs.cmu.edu.

Tâm Huỳnh, Mario Fritz, and Bernt Schiele. Discovery of activity patterns using topic models. In *Proceedings of the 10th International Conference on Ubiquitous Computing*, pages 10–19, 2008.

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

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

David Mimno, Hanna M. Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. Polylingual topic models. In *Proceedings of Empirical Methods in Natural Language Processing*, pages 880–889, 2009.

David V. Pynadath and Michael P. Wellman. Accounting for context in plan recognition, with application to traffic monitoring. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, pages 472–481, 1995.

Young Chol Song, Henry Kautz, James Allen, Mary Swift, Yuncheng Li, Jiebo Luo, and Ce Zhang. A markov logic framework for recognizing complex events from multimodal data. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, ICMI '13, pages 141–148, New York, NY, USA, 2013. ACM.

Mark Steyvers and Tom Griffiths. Probabilistic topic models. In T. Landauer, S. Dennis McNamara, and W. Kintsch, editors, *Latent Semantic Analysis: A Road to Meaning*. Laurence Erlbaum, 2007.

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

Marc Vilain. Getting serious about parsing plans: A grammatical analysis of plan recognition. In *Proceedings of the 8th National Conference on Artificial Intelligence*, pages 190–197, 1990.

Yang Wang and Greg Mori. 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, 2009.

Gang Wang, Ye Zhang, and Fei-Fei Li. Using dependent regions for object categorization in a generative framework. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1597–1604, 2006.

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