How Robots Can Recognize Activities and Plans Using Topic Models

Richard G. Freedman and **Hee-Tae Jung** and **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.

RGB-D sensors are a commonly used tool for observing human posture. As robots equipped with RGB-D sensors are used in a variety of domains, predefining the activities to be recognized may also be too limiting as we further discuss below. Section 2 follows with a brief background on PR, AR, and LDA. Section 3 then investigates ways to represent the data as text and shows how LDA may be used for performing PR and AR. Section 4 applies this to a small dataset we collected and interprets the learned topics within the plans. Recognizing plans and activities using the learned topic model is tested by cross-validation in this paper, and more rigorous evaluation of its performance will be a focus of future work. We conclude with a discussion of the approach and its possible extensions in Section 5.

Motivation

Human-robot interaction (HRI) studies how to improve the immersion of robots in social situations amongst humans. As explained by Lösch *et al.* (2007), Sung *et al.* (2012), and Zhang and Parker (2011), an integral component of successful interaction is the ability to predict what other agents are doing in the environment. Thus real-time PR using only onboard sensors is an important aspect of effective collaborative behavior in machines. With a reasonable interpretation of the actions of those around them, robots can appropriately plan responses for common engagements that will inevitably take place in the real world. Regardless of the domain, be it personal robotics for household chores, industrial robotics participating in the workforce, or aiding in dangerous activities such as search and rescue during natural disasters, we find situations involving direct correspondence between agents.

Most work with PR has not only been structural, but also represented at a higher level. That is, the representation of plans and actions assume that activities such as "move" and "lift" are already determined. Raw sensor data does not return such information; it needs to be extracted using AR-like approaches. One benefit of being able to identify topics from sensor data is that we can produce a wrapper that can lift the raw data to a higher level for use in well-studied structural PR techniques. In addition to integration with such techniques, we can take advantage of the simultaneously derived distribution of these topics to approximate the "gist" of the plan to recognize. This may be used as a guide or heuristic when identifying the executed plan in the library of plans.

Humans do not always act in 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 not specified by the robot's plan library, or the human may perform some extraneous actions that would serve as noise in the execution sequence. Being able to analyze the distribution of activity topics in an execution sequence introduces a computationally feasible method for handling noise and omitting ordering. Considering all the combinations of execution sequences in order to omit the noisy actions as well as reorder independent subsequences would require enormous effort. In real-time systems, this can be a considerable bottleneck.

2 Background

Plan and Activity Recognition

PR is the inverse of the planning problem. Rather than trying to derive a sequence of actions that can accomplish a task, we observe some sequence of actions or changes in the world state and try to identify the task. Past approaches to solving PR problems have ranged from purely logical to partially statistical methods. Logical methods often use lists of rules and relationships between actions to represent plans as structured objects such as grammars (Vilain 1990) and directed graphs. Statistical methods have extended the logic framework by inferring the likelihoods of different plans identified by the structured representations given various features of the problem (Pynadath and Wellman 1995). Song *et al.* (2013) combine the two for a statistical-relational approach that uses lifted information to develop a Markov Logic framework that enforces temporal constraints on observations from audio and visual input for PR. However, their method assumes that the library of plans and each plan's composition is known. These assumptions are commonly made in PR research.

While PR focuses on identifying the entire plan/task, AR is more specific and tries to recognize the single activities and/or actions that compose the plan (Goldman *et al.* 2011). Hamid *et al.* (2007) learned suffix trees over subsequences of events in order to use combinations of n-grams for multiple n to classify activities. They used these suffix trees on sequences of events representing interactions with key-objects in a kitchen to classify kitchen activities. Besides classifying sequences of events, one of the primary applications of AR is to produce higher-level interpretations of sensor data as described in the motivation above. The inference performed for AR is usually more machine learning centric due to the uncertainty involved in mapping raw sensor data to actual activities. Bayesian inference techniques similar to the ones described above for PR have also been used in AR. For example, Huỳnh *et al.* (2008) previously used topic models with wearable sensors to decompose a user's daily routine into its single-activity components without human annotation.

Latent Dirichlet Allocation

LDA is a probabilistic topic model that considers a set of documents D to be *generated* from a set of topics T. These topics provide a semantic interpretation of the documents without regarding syntax. The distributions of topic allocations over documents $\theta_{d \in D}$ and the distributions of words over topics $\phi_{t \in T}$ are each drawn from Dirichlet distributions specified by hyperparameters α and β respectively. Each word w_i in a document d is assigned a single topic $z_i \in T$ that is drawn from θ_d such that w_i would be drawn from ϕ_{z_i}. Only the words in each document $\vec{w}$ are observed; $\vec{z}$, $\vec{\theta}$, and $\vec{\phi}$ are all latent variables and the hyperparameters are selected as priors. Steyvers and Griffiths (2007) provide an in-depth explanation of this approach.

Through statistical sampling methods such as Gibbs sampling, it is possible to find assignments for the latent variables that (nearly) maximize the expected likelihood of gen-

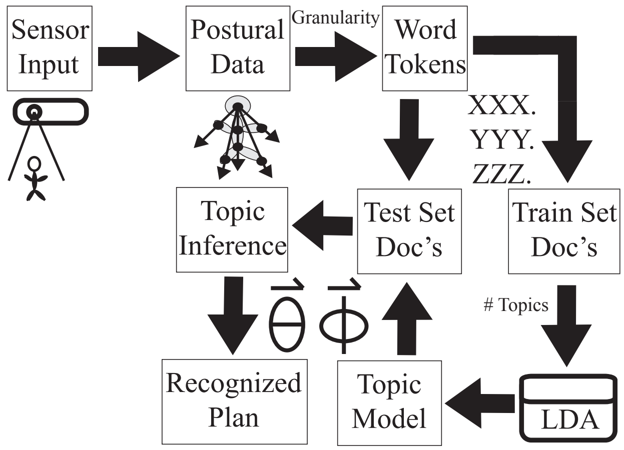

Figure 1: Applying LDA to PR and AR from recorded sensor data to learn actions (topics) and predict plans (documents).

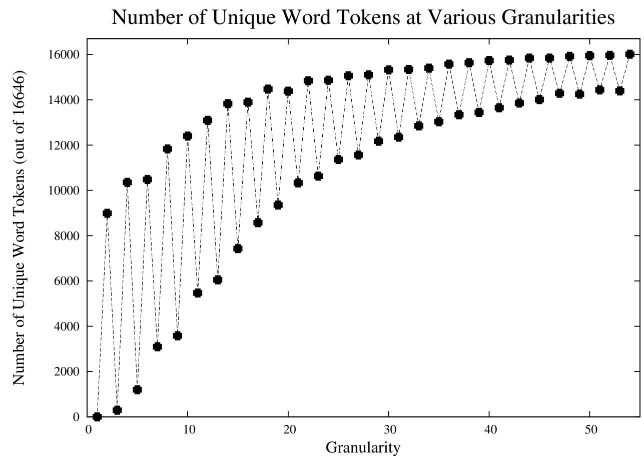

Figure 2: A plot of unique word tokens in the collection of forty recorded plan executions at various granularities.

erating the observed documents $\mathrm{P}\left(\vec{z},\vec{\theta},\vec{\phi}\,\middle|\,\vec{w},\alpha,\beta\right) =$

$$\frac{\mathrm{P}\left(\vec{w}\,\middle|\,\vec{z},\vec{\phi}\right)\cdot\mathrm{P}\left(\vec{z}\,\middle|\,\vec{\theta}\right)\cdot\mathrm{P}\left(\vec{\theta}\,\middle|\,\alpha\right)\cdot\mathrm{P}\left(\vec{\phi}\,\middle|\,\beta\right)}{\mathrm{P}\left(\vec{w}\,|\alpha,\beta\right)}$$

where the denominator is just a normalizing constant. This is referred to as *training the topic model*. When these assignments are found, $\vec{\theta}$ and $\vec{\phi}$ may be studied to learn more about the extracted topics (which requires a human's interpretation since LDA is unsupervised) and their presence in each document. We use the MALLET software (McCallum 2002) for training and performing our experiments.

3 Applying LDA to Plans

Using a RGB-D sensor that can approximate human postural data, we recorded a dataset of forty plan executions that were composed of subsets of ten actions (choice of hands and feet varied): standing, walking, reaching upwards, waving hand while stationary, waving while walking, throwing, jumping with hands down at one's side, jumping while reaching upwards, squatting, and kicking. The sequences of actions were generated to represent everyday human tasks in which a robot could assist or interact. For example, "reaching for an item on top of a tall bookshelf" is represented by the sequence of stand, reach upwards, jump while reaching upwards, stand. During the recording, an actor followed a narrator's instructions to perform the plan as specified. Each execution was recorded at thirty frames per second and lasted varying lengths less than one minute. We henceforth consider each "document" to be a recorded plan execution, each "topic" to be an action or activity, and each "word token" to be a pose extracted from a single frame. Figure 1 shows the pipeline used to process the data, which we describe below.

Textual Representation

The raw data recorded by the RGB-D sensor is in the form of homogenous 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*. Each triple contains the pitch, roll, and yaw that denote the vector whose initial point is the head and endpoint is another joint in the sensed agent's body (depicted under the "Postural Data" box in Figure 1). We consider fifteen joints and each word token is in $[-\pi,\pi]^{45}$ which is an uncountably infinite vocabulary with a very small likelihood of duplicate tokens. However, finding each activity's distribution over the word tokens $\vec{\phi}$ requires a countable vocabulary with some duplicate poses in the collection of plan executions. Wang and Mori (2009) and Zhang and Parker (2011) created codebooks to accomplish this by clustering the images of the training set and selecting the center of each cluster as a word token in their vocabulary; all images in the same cluster (including those in the test set) are assigned this token value.

We make the vocabulary finite and increase the likelihood of having duplicate word tokens by *discretizing the space* with respect to a *granularity* parameter. For granularity $g \in \mathbb{N}$, we map each angle φ to integer $0 \leq i < g$ such that $(i\,/\,g)\cdot 2\pi \leq \varphi + \pi < ((i+1)\,/\,g)\cdot 2\pi$. This reduces the vocabulary to $\{0,1,\ldots,g-1\}^{45}$ which is still large in size for small g, but we must consider that many of these poses do not represent feasible body structures; for example, the limitations of each joint's range of motion will prevent such word tokens that include hyperextended limbs. This is analogous to the fact that many combinations of orthographic letters do not form actual words used in a language. An advantage of using granularity to discretize the space over the use of a codebook is that word tokens appearing exclusively in the testing data may appear as new tokens rather than be assumed to be a previously encountered pose from the training data. That is, it is possible to encounter new poses for which the system was not trained. These can be handled by smoothing the multinomial parameters (Blei *et al.* 2003).

Figure 2 plots the number of unique word tokens in our collection of documents at various granularities. As ex-

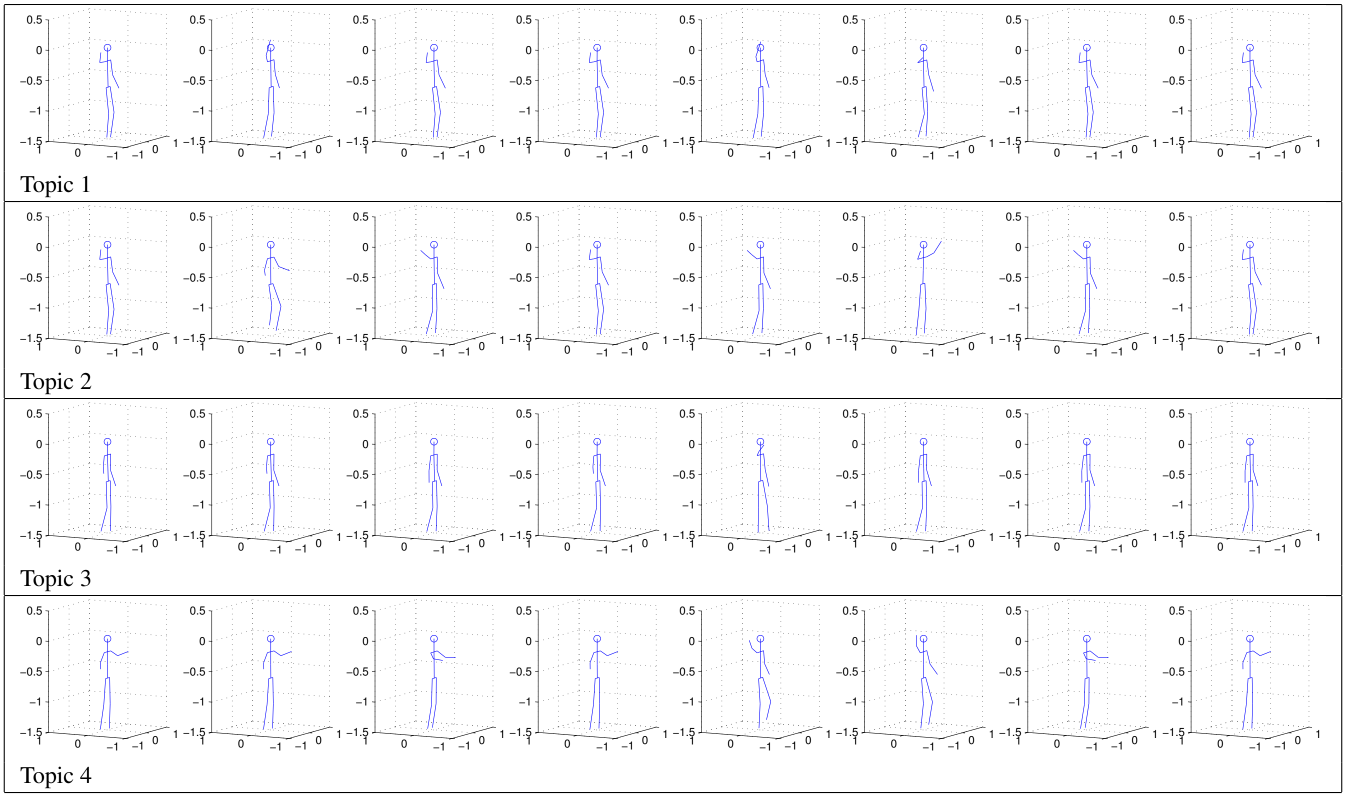

Figure 3: Most likely poses for four selected topics (actions) from the fifteen-topic model learned from the plan execution traces with granularity thirty-five. We interpret them as (top to bottom) throwing (right hand), hand raised, walking, and arms in carrying position.

pected, increasing the granularity reduces the number of duplicate poses since each interval is smaller. One interesting feature of the plot is the drastic difference between the number of unique tokens based on the parity of the granularity. 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 $\varphi = 0$) will be assigned to one of two different groups: $(g\,/\,2)$ if $\varphi \geq 0$ and $(g\,/\,2) - 1$ if $\varphi < 0$. 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 larger numbers of integer combinations for even granularities compared to odd ones.

Recognizing Activities and Plans

Performing AR and PR with our learned topic model requires finding the likelihood that the model would generate other action sequences that belong to this corpus. Because topic models are generative, it can find this likelihood for an unobserved execution sequence $\vec{w}'$ by simulating the generation process described in Section 2's LDA background. The new plan's distribution over actions θ' is drawn from the Dirichlet distribution with hyperparameter α used to draw each entry of $\vec{\theta}$, and $\vec{\phi}$ remains unchanged. Then each new pose w'_i is associated with action z'_i which is drawn from θ' such that w'_i would be drawn from the distribution $\phi_{z'_i}$.

As it simulates each generation step, it multiplies the current likelihood of generation with the likelihood of the simulated step. The values of the unobserved variables θ' and $\vec{z}\,'$ that maximize this generation likelihood are the *inferred values*. The process of inferring the values of $\vec{z}\,'$ is an AR system since it identifies the most likely actions for the observed poses. The distribution of activities θ' from which each z'_i is drawn represents a plan when viewed as a bag-of-words because the sequence of actions is no longer ordered; thus we consider the inference of this distribution to be a PR system. Hence LDA integrates both the AR and PR processes into a *single system for simultaneous inference* rather than channeling information from an AR system to a PR system. To perform this inference efficiently, Gibbs sampling may be used to obtain a good estimate. The use of log-likelihoods is also necessary to avoid underflow from multiplying so many probabilities together.

4 Experimental Results

For varying granularities between one and fifty-one, we ran LDA on our corpus of forty recorded plan executions with 2000 iterations of Gibbs sampling, initial hyperparameter values $\alpha = 50$ and $\beta = 0.01$, and hyperparameter optimization every ten iterations. The best choice for number of topics varies with respect to the context of the corpus. A smaller

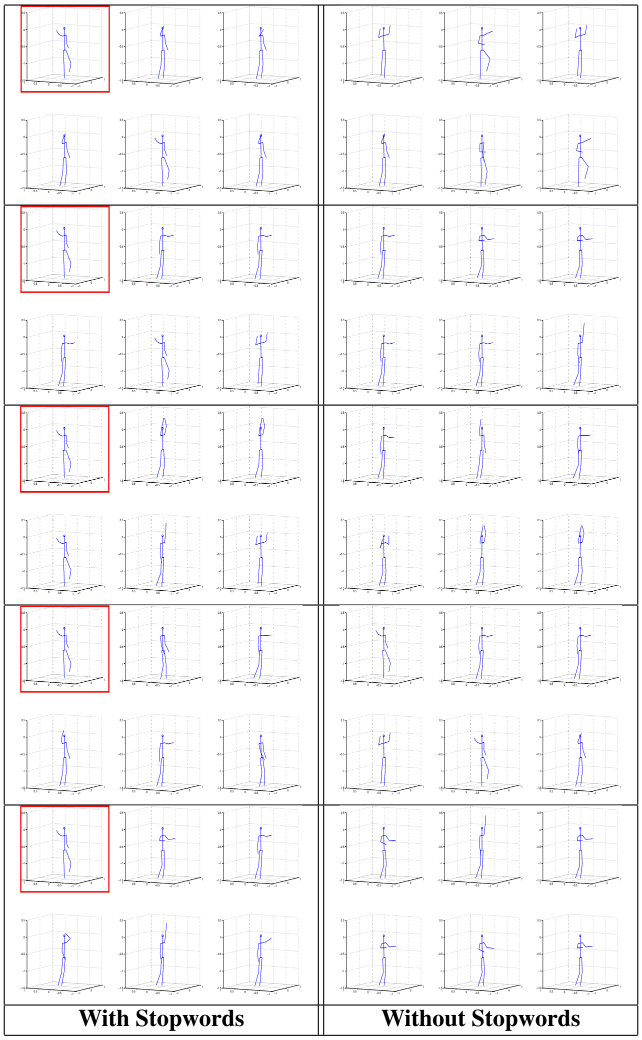

Figure 4: Most likely poses for the five-topic model using granularity three before and after removing stopwords. A red box is drawn around the stopword appearing in every topic.

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 depict 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.

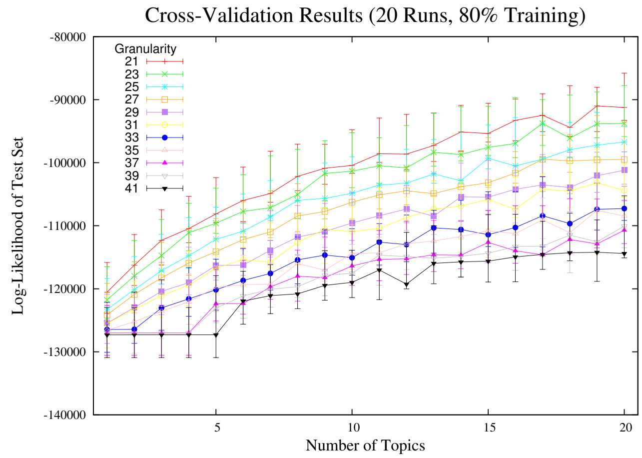

Figure 5: The results of running cross-validation on our dataset with twenty runs for each pairing of a subset of various granularities and numbers of topics.

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.