Plan and Activity Recognition from a Topic Modeling Perspective

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

School of Computer Science
University of Massachusetts
Amherst, MA 01003, USA
{freedman, hjung, shlomo}@cs.umass.edu

Abstract

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 human skeletal data of plan execution traces obtained from 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

It has been suggested that *plan recognition* (PR) and *natural language processing* (NLP) 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 will analogously treat plans 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 activities 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. We instead consider pose information through human skeletal coordinate data obtained from a RGB-D sensor similar to the Kinect in order to avoid these representational drawbacks.

Furthermore, RGB-D sensors are currently a mainstream tool for robots to observe human posture. As robots 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. 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. An integral component of successful interaction is knowing what other agents are doing in the environment. Thus robots need to be able to perform real-time PR using just their on-board sensors. 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 the activity such as "move" or "lift" is already known. 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.

Furthermore, 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).

Similar Bayesian inference techniques have also been used in AR which is closely related to PR. 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). 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. 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. 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 generating 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 skeletal coordinate 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, throwing, jumping, squatting, kicking, jumping while reaching upwards simultaneously, and waving while walking simultaneously. 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 single frame's skeletal pose. 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

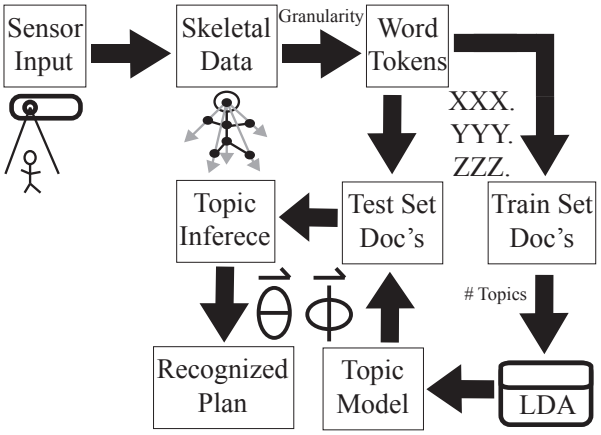

Figure 1: Applying LDA to PR and AR from recording sensor data to learning actions (topics) to predicting plans (documents).

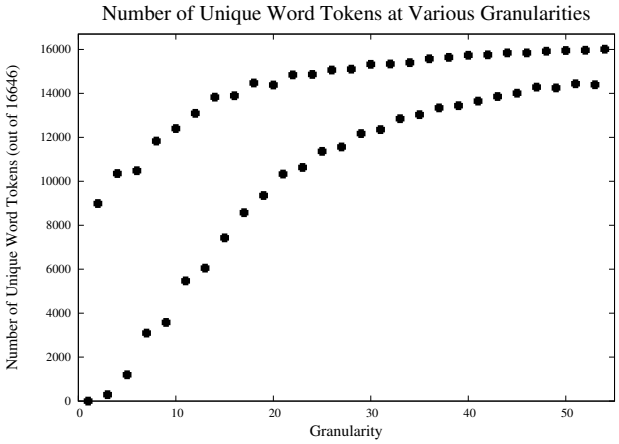

Figure 2: A plot of unique word tokens in the collection of forty recorded plan executions at various granularities.

point is the head and endpoint is another joint in the sensed agent's body. Since there are fifteen joints, each word token is in $[-\pi, \pi]^{45}$ which is an uncountably infinite vocabulary with a very small likelihood of duplicate tokens. However, finding activity distributions $\vec{\phi}$ requires a countable vocabulary with some duplicate poses in the collection of plan executions. Wang and Mori (2009) created a codebook 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.

Figure 2 plots the number of unique word tokens in our collection of documents at various granularities. As expected, 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' 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.

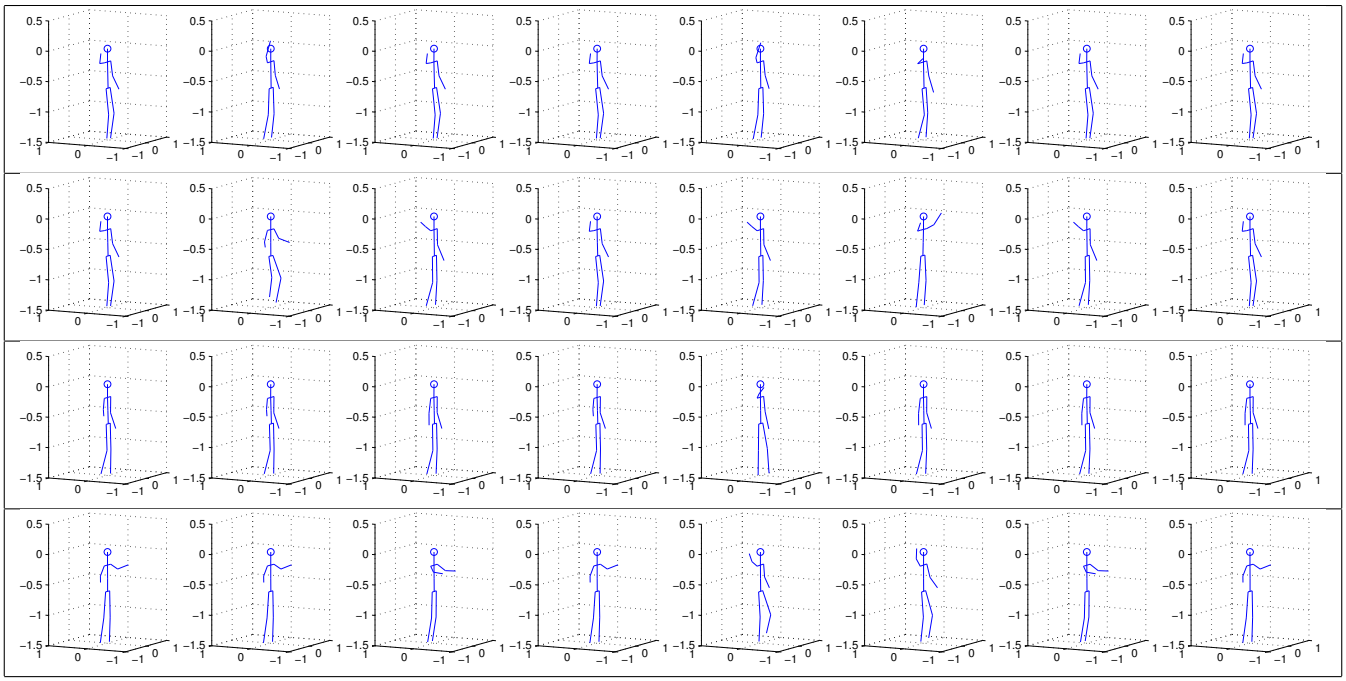

Figure 3: Most likely poses for selected actions from the fifteen-topic model learned from the plan execution traces with granularity thirty-five. Actions may be interpreted as (top to bottom) throwing with right hand, squatting/hand raised, walking, and arms in carrying position.

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 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.

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 which resulted in ambiguities between some actions

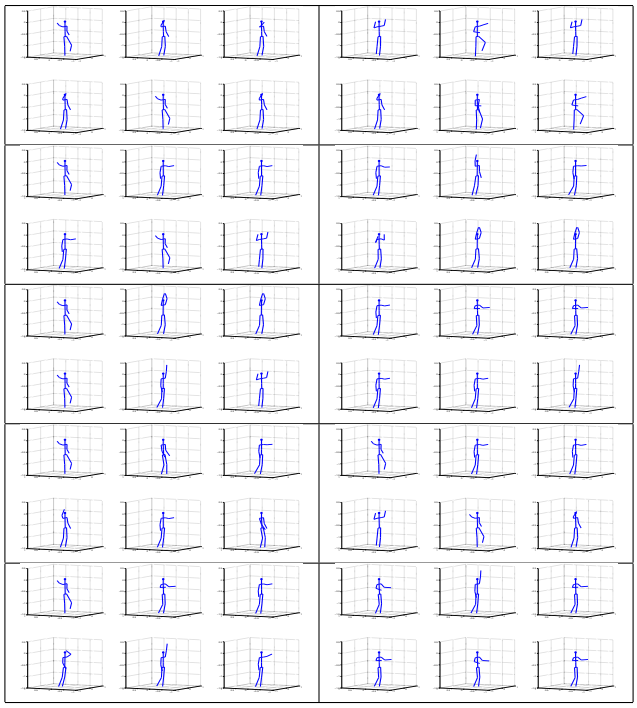

Figure 4: Most likely poses for the five-topic model using granularity three before (left) and after (right) removing stopwords.

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. As subject features such as height and strength may also affect which actions they 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.

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 Artifical 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 Artifical 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.

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.

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.

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.

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.