Human modeling for human-robot collaboration

Journal Title
XX(X):1–24

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Laura M. Hiatt[1] and Cody Narber[1] and Esube Bekele[1] and Sangeet S. Khemlani[1] and J. Gregory Trafton[1]

Abstract

Teamwork is best achieved when members of the team understand one another. Human-robot collaboration poses a particular challenge to this goal due to the differences between individual team members, both mentally/computationally and physically. One way in which this challenge can be addressed is by developing explicit models of human teammates. Here, we discuss, compare and contrast the many techniques available for modeling human cognition and behavior, and evaluate their benefits and drawbacks in the context of human-robot collaboration.

Keywords

Human-centered and Life-like Robotics, Cognitive Human-Robot Interaction, Cognitive Robotics

Introduction

Teamwork is best achieved when members of the team understand one another (Grosz and Kraus 1999). Human-robot collaboration poses a unique challenge to this goal: namely, providing robot teammates with the ability to recognize and understand what their human teammates are doing. While this is an ability humans naturally learn over time, robots need to be explicitly taught how to do this. This recognition problem is exacerbated by the differences between robot and human team members, both mentally/computationally and physically (Fong et al. 2001; Scassellati 2002). These differences mean that, when faced with the uncertainty of the real-world, robots cannot always count on human teammates to stay on-script, and cannot always easily anticipate how human teammates will react when something does go off-script.

One way in which this challenge can be addressed is by equipping robots with explicit models of their human teammates. There are many different techniques that are used to model human cognition and behavior, spanning different timescales and levels. In this paper, we begin by introducing a framework that we find useful for discriminating between and categorizing these approaches. We then discuss, compare and contrast the many techniques available for modeling human cognition and behavior, and evaluate their benefits and drawbacks in the context of human-robot collaboration.

Marr's Levels of Analysis

Marr (1982) recognized that information processing systems – whether human or not – can be explained at distinct levels of analysis: the computational level,

[1]Naval Research Laboratory, USA

Corresponding author:
Laura Hiatt
Naval Research Laboratory
4555 Overlook Ave., SW
Washington, DC 20375 USA.
Email: laura.hiatt@nrl.navy.mil

the algorithmic level, and the implementational level. The different levels correspond to different perspectives and applications of the processing system. Marr explains the levels intuitively through the example of a cash register, which can be conceived of as a simple information processing device with inputs and outputs. The computational level describes the primary function of the system, i.e., *what* it is doing. At the computational level, the cash register performs simple arithmetic, such as addition and subtraction. Hence, its mathematical functions are clear and transparent.

The algorithmic level, in contrast, specifies *how* a particular function is carried out, as well as the representations it depends on. The choice of which representation to use provides constraints on the kinds of processes the system carries out. For example, the cash register's internal representation of numerical information relies on the use of a particular representational scheme, e.g., binary, decimal, hexadecimal, tally marks, and so forth. The representation constrains the calculations that the system is able to perform, as well as how fast it performs them. It's difficult to perform multiplication using tally marks, for example, and it may take more or less time to perform addition and subtraction with binary vs. decimal numerals.

A description of the system at the implementational level specifies how the representations and procedures are physically realized. A cash register that uses algorithms operating over binary-encoded numerals implements those algorithms using, possibly, transistors, capacitors, and registers for storing and manipulating electrical signals that represent numerical information. This hardware can impose processing and memory limitations on the algorithms that can run on it, and so can be important to understand.

Marr's levels of analysis provide a useful framework within which to organize recent technological developments in modeling human performance for human-robot collaboration because they can provide a guide to clarify and understand what aspect of human behavior is being modeled. Techniques at the computational level, for example, are well suited to situations that would benefit from knowledge of ideal or typical human performance. Such approaches either assume perfect rationality, or smooth over human idiosyncrasies and noisy observations by providing general accounts and/or mathematical functions of human performance.

But, humans are prone to systemic errors and processing constraints, and many approaches to human-robot collaboration benefit from explicitly accounting for this. For such approaches, modeling techniques from the computational level are limited in their use, since they cannot capture human error or processing time. After all, knowing that a cash register can add and subtract does not tell you how long it takes to add twenty numbers, and so how long you should reasonably expect to wait in a long check-out line. Models of performance at the algorithmic level provide this analysis, and so they can help maintain interactions that are understanding of noisy human processes. On the other hand, this lower-level of granularity means that models at the algorithmic level are not as well-suited to understanding behavior that is wider in scope, or over longer timescales.

Approaches targeting the implementational level, in contrast, can capture how people's cognition can vary based on internal (i.e., stress) and external influences (i.e., caffeine), since both affect the physical "hardware" of cognition. Implementational level techniques such as biologically-plausible algorithms have a strong and promising place in understanding how the human mind works, but research at this level is not as well established as research at the other two levels. Because of this, their contributions are difficult to meaningfully use for human-robot collaboration.

Using these levels as a framework for organizing models of human behavior has several benefits. First, it allows researchers to compare and relate approaches with one another in a meaningful way. Viewing approaches in this framework illuminates strengths and weaknesses of different approaches to modeling humans that are not always brought out when considering them in their typical settings or domains, providing insight

into how each of the approaches can be expanded and improved. Second, this framework equips researchers to effectively select which type of model may be best for their particular situation. The level of analysis of a particular methodology provides useful guidelines and prescriptions for how the methodology can aid and facilitate interactions, and so can be considered as a starting guide for those looking for an appropriate methodology for modeling human-robot collaboration in their work.

Finally, to achieve truly robust and productive human-robot collaboration, robots should ideally understand what humans do in general (the computational level), how they do it (the algorithmic level) and how cognition is physically realized (the implementational level). By evaluating different models according to common criteria, researchers will able to more effectively combining approaches from different levels together to achieve better human-robot collaboration.

Other Considerations

When modeling approaches are applied to human-robot collaboration, they are subject to considerations that are not always relevant in other situations (such as the general plan or activity recognition problem). We have mentioned above two factors relevant to human-robot collaboration that we believe are related to the level at which human behavior is being modeled: how approaches handle noisy human processing; and what timescale the approach operates over.

Another axis we categorize approaches on is whether the model is hand-coded or learned automatically. Additionally, we consider how much data is needed to train the model. In general, of course, learning models is preferable to hand-coding them; however, due to the often high complexity of the models, and the often limited availability of data on human behavior in human-robot collaboration tasks, learning is not always feasible either theoretically or practically.

Finally, the purpose of this paper is to discuss meaningful ways of modeling and representing human behavior and activity, and, as such, we have included a somewhat subjective selection of work on human modeling in order to accomplish that goal. This is not meant to be a comprehensive view of human activity recognition, plan recognition, or modeling techniques in general.

Computational Models

We begin by discussing four types of models of human behavior that are computational under Marr's framework. These models can be used to capture general characterizations of human behavior, without being concerned about the specific processes or representations that cause the human to behave in that way.

The first methodology we discuss is simple probabilistic models, such as those describing cost functions that capture human behavior, or those that mathematically describe trajectories of human movement. Such approaches are useful when trying to make sense of a human's low-level signals to better anticipate or recognize their behavior.

The second computational methodology groups together several conventional approaches to machine learning. These approaches also tend to use mathematical functions and formulations to describe human behavior, but within more general learning frameworks. These more complicated formulations allow them to capture a broader array of behavior than simple probabilistic models, but their representation still does not generally provide explanations for how that behavior comes about.

We next describe knowledge-based models, which take a more symbolic approach to modeling human behavior. Such approaches typically include libraries or databases of what people, in general, do when faced with a certain goal. This allows them to capture human behavior at higher levels than the more mathematical-based models, allowing a richer, hierarchical-based understanding of what people are doing.

Finally, we describe Petri nets, a deterministic graphical model that treats the behavior of humans (or groups of humans) as a discrete event system,

and models how system resources (such as people or objects) move around over time as activities are executed. As such, Petri nets are successful at capturing high-level patterns of behavior, but, as with the other approaches in this section, provide little understanding for how or why that behavior is occurring.

Simple Probabilistic Models

Simple probabilistic models have the most success when modeling very low-level activities. The use of simple mathematical and statistical equations to describe higher-level behavior, instead of more symbolic or state-based approaches, is of limited use, because it is harder to understand how to use simple equations to describe and understand complicated human behavior directly. In low-level situations, however, simple probabilistic models can be a very effective and intuitive way to describe behavior.

Andrist et al. (2015) modeled human gaze behavior as normally distributed, with separate distributions for extroverts and introverts. The data used as the basis for the distribution was collected across several different human-human interactions. Morales Saiki et al. (2012) use a similarly simple, but effective, model, for describing how people walk down a hallway.

Such simple models can be made more powerful by combining them with filters such as Kalman and particle filters. These filters take mathematical models (or other models) and, essentially, unpack them over several timesteps, correcting their predictions given possibly noisy observations. This consideration of sequences of observations or measurements, applied to a single mathematical model, gives the models more predictive and explanatory power (Ristic et al. 2004) . Hecht et al. (2009), for example, used a flexible mass-spring model to capture human kinematics, and combine it with a particle filter to create a model of predicted human movement. Jenkins et al. (2007) also effectively performed higher-level action tracking from primitive human motion and kinematic pose estimates using a particle filter.

Dragan and Srinivasa (2012) provide a framework for inferring user intent that assumes that users execute actions by noisily optimizing a goal-dependent cost function. Then, when inferring a user's intent, the framework returns the goal corresponding to the trajectory with the lowest cost given the user's input. Cost functions are either learned via observation in offline training, or, for more complex domains, generated using assumptions of expertise and rationality (such as minimizing distance; others make this assumption as well, e.g., Fagg et al. 2004). The generality of the framework makes it suitable for a variety of tasks; in addition, this approach is notable for human-robot collaboration because it has been shown to increase the legibility and predictability of motion (Dragan et al. 2013). The approach, however, can be computationally expensive in high-dimensional spaces, and it depends highly on the availability and tractability of learning and representing an expressive cost function.

Conventional Approaches to Machine Learning

Traditional or conventional machine learning models are common and effective ways of learning human behavior. We consider here as conventional models any machine learning approach that: makes the basic assumption that the training and testing sets represent the same underlying distribution (Dai et al. 2007); lacks the ability to readily extend to natural generalizations; lacks the representational power to express hidden properties; and lacks the ability to infer existing causal relationships from a handful of relevant observations (Griffiths et al. 2010). These conventional approaches have had great success in recognizing human behavior across a variety of domains, including social robotics (Fong et al. 2003; Breazeal 2004; Tapus et al. 2007), learning from demonstration (Billard and Siegwart 2004) and capturing human affective state (Liu et al. 2008).

Some of the most common approaches to machine learning in modeling human activity use discriminative techniques like Decision Trees (DTs) C4.5 (Ravi

et al. 2005), Multiple-Layer Perceptions (MLPs) (Ermes et al. 2008), and Support Vector Machines (SVMs) (Castellano et al. 2012; Wang et al. 2005). Discriminative techniques use, as input, data features extracted from some form of modality (video, audio, wearable sensors, etc.) and apply a label. The mapping from features to label is learned using labeled training data; the model does not need to be constructed or specified ahead of time. The advantage to using these discriminative techniques is they can capture any human activity provided that they can learn from labeled training data that has meaningful features available to leverage in the model.

K-nearest neighbors (KNNs) is a non-parametric discriminative approach that classifies data without learning an explicit classification function. Given data features extracted from input data of various modalities, KNNs store all training examples and use distance metrics, such as Euclidean, to classify a new case to its closest points. As with the other discriminative techniques, KNNs can be applied to most situations where labeled data is available; for example, Mower et al. (2007) applied KNNs to using human physiological signals to capture user engagement in an HRI task. Notably, KNNs have been shown to be effective not only at classifying human behavior, but also at generating it by sampling in the discovered regions, which makes it very useful for human-robot collaboration (Admoni and Scassellati 2014).

Gaussian Mixture Models (GMMs) have similar inputs and outputs to KNNs; however, while KNNs result in a single label for each point, GMMs represent a point by a vector of class label likelihoods. GMMs thus can better handle some degree of uncertainty in a human's activity (Calinon et al. 2007). GMMs can also provide fast and early classification. Pérez-D'Arpino and Shah (2015) learn GMMs of human motions, based on the DOFs of the human arm, in order to anticipate their intended target more quickly. Here, the human's activity is constrained to reaching motions, and, in part because of this assumption, the authors are able to quickly and accurately anticipate the target of the reaching action.

A technique that has become very popular in recent years is deep learning. Deep learning is a technique that takes raw data and learns the best features to extract for classification, as opposed to relying on the modeler to specify how features should be extracted (Schmidhuber 2015). A common classification technique that uses deep learning is the convolutional neural network (CNN). CNNs typically use either raw images or video as input and apply image convolutions and other reduction techniques to extract a learned feature vector (Krizhevsky et al. 2012; Tang 2013). Deep learning has shown excellent results in pose recognition (Toshev and Szegedy 2014) and large-scale video classification (Karpathy et al. 2014). The drawback to using CNNs and other deep learning techniques is that because both the features and the model are discovered through the data, a significant amount of data must be gathered in order to avoid overfitting.

Two major benefits of using conventional machine learning approaches for human-robot collaboration are that the model does not need be specified ahead of time, and that there are many existing, off-the-shelf techniques to chose from. On the other hand, many of these approaches can require extensive data to train effectively, and the goodness of the model is very dependent on the data that is chosen, both in terms of the specific training data used and in how features are extracted from that data. Additionally, using large, expressive feature vectors can cause the training time to grow exponentially, making it difficult to train models in a timely manner. Finally, the lack of structure of many of these models means that temporal sequencing and dependencies of activities is not easily taken into account. For these reasons, conventional machine learning approaches, like simple probabilistic models, are best suited for modeling lower-level, short-term human activities, as opposed to longer-term tasks and goals.

Knowledge-Based Models

Knowledge-based models compare human behavior against ontological or knowledge-based representations of how tasks are typically executed in the world. An intuitive example of this is matching a human's behavior to a pre-specified planning library. By specifying general domain knowledge, these approaches can provide a clear view into a computational-level understanding of human behavior. Some of the earliest approaches to plan recognition used this approach (Lesh et al. 1999; Wilensky 1983; Perrault and Allen 1980).

Breazeal et al. (2009) approaches this mapping at a slightly deeper level, mapping human motions and activities onto the robot's own model of actions and activities. While this approach is useful when the robot and the human have very similar capabilities, both in terms of motions (i.e., both have arms to move) and activities (i.e., both represent plans the same way), it does not easily translate to situations where robots and humans are in less accordance. Many other approaches that assume a shared representation of domain knowledge have similar limitations (e.g., Levine and Williams 2014; Rota and Thonnat 2000), although some approaches have alleviated this by including in their representation multiple ways to complete tasks (a canonical example is Kautz and Allen 1986's specialization recipes).

Knowledge-based approaches to human modeling also can include temporal reasoning about human actions (e.g., Nevatia et al. 2003; Vu et al. 2003; Avrahami-Zilberbrand et al. 2005). This can help to further understand human behavior by taking expected task duration into account. It is worth noting, however, that task duration here is qualitatively (but subtly) different from the processing durations commonly associated with the algorithmic level: it is capturing the duration of the task and the human's behavior as viewed by an external observer, not the duration of the human's internal processing and behavior when reasoning about or performing the task.

In general, these approaches, while effective and generally fast (Maynord et al. 2015), place some strong assumptions on the human's behavior. They assume that their domain knowledge is complete; further, the extensive domain knowledge must be manually changed and updated for every new situation, which is ultimately unrealistic for fully functioning robotic agents (Carberry 2001; Turaga et al. 2008). Such assumptions also imply that these approaches cannot handle human actions or intentions outside of their knowledge, such as human error (although Avrahami-Zilberbrand et al. 2005 is one exception that can handle lossy observations).

On the other hand, there are clear and intuitive ways in which these approaches can be combined with other approaches in order to alleviate some of these shortcomings. Many more recent knowledge-based models are combined with statistical techniques to improve their functionality. Nevatia et al. (2003) use a hierarchical representation language to describe human activities, and use Bayesian networks to recognize primitive events and Hidden Markov Models (described in more detail below) to recognize higher-level events; other approaches also take this route (e.g., Blaylock and Allen 2014). In addition, recent work in goal reasoning (Gillespie et al. 2015) allows autonomous agents to dynamically adapt their goals to changing situations; if a cost or preference function for such changes were derived that was based on human behavior, it could allow these approaches to relax the assumption that the human's goal is known.

Petri Nets

Place/Transition Petri nets (PNs) are a type of deterministic graphical model that models humans and their interactions with the world as a discrete event system and tracks how the system's resources change over time. PNs consist of states (here called "places"), which contain sets of tokens. In human-robot collaboration, these tokens can be used to represent discrete resources, which are then consumed, produced, or transferred to a different place (representing, for

example, a change of location) by actions (Murata 1989). The places, transitions and tokens of PNs are modeled deterministically. A human's action can be recognized by comparing the actual state of the world with the overall state of the Petri Net (such as the locations of resources). If the states match, that action can be thought to have taken place; otherwise, a different action explaining the state of the world should be found.

Because PNs model the world in terms of resources, they can effectively handle concurrency (such as representing two agents as different resources). This also makes Petri nets well-suited to capturing high level activity such as people's (or groups of people's) movements over time, as opposed to lower-level activity (which would benefit from modeling resources continuously). The straight-forward representation of Petri nets also makes it intuitive to combine them hierarchically (Castel et al. 1996). As such, Petri nets are often used in surveillance applications, such as tracking cars in a parking lot to detect anomalies (Ghanem et al. 2003).

The structure of the Petri net must be hand-coded ahead of time, and the deterministic aspect of the model places a high burden on the structure very accurately representing the activity. The determinism of Petri nets also presents a challenge for its ability to accurately capture many tasks involving humans, although recent work has attempted to add in more tolerance of human variability. Albanese et al. (2008), for instance, worked towards plan variation in a surveillance setting by enabling the model to account for skipped actions. Lavee et al. (2010) accounted for temporal variation by allowing actions to occur at stochastic time intervals, enabling, for example, the PN to account for variance in humans' daily schedule or routines. More complicated activity can also be modeled by using a separate PN for each possible action (Ou-Yang and Winarjo 2011), with little additional computational overhead. This movement towards more explicitly accounting for uncertainty in Petri nets seems to correspond to an increasing use in human-robot collaboration. Ultimately, if this work continues, it also may place this approach at the boundary of the computational and algorithmic levels of Marr's framework. For now, however, the underlying theory is not yet established enough and so it remains at the computational level.

Computational Models Discussion

In general, these approaches all target modeling human behavior at the aggregate level, whether by constructing general plan libraries of what people typically do when completing a task, or by learning a single, idealized model of human behavior based on many data points. This type of knowledge can be effectively leveraged by many types of tasks in human-robot collaboration. In particular, many of these approaches are most useful when one is mostly concerned with identifying human behavior without evaluating it or reasoning about why the human is doing what they are doing. The knowledge-based approaches are an exception to this, as they can potentially provide some intuition about the validity of the human's actions. As we will see, however, algorithmic approaches are better positioned to provide this knowledge in a way that allows a robot teammate to help clarify and, potentially, correct the human's action.

Computational and Algorithmic Models

In this section we describe methodologies that span the bridge between Marr's computational and algorithmic levels. These models, like the above ones, are typically tailored towards capturing human behavior at a general level; however, the approaches in this section either have representations more suited to modeling human behaviors or intentions, or the algorithms are better at capturing less idealized and more idiosyncratic behavior. Therefore, even if work applying these methodologies to the algorithmic level is not well established, we believe the supporting theory is established enough to include them in this bridged category.

We begin by discussing Hidden Markov Models, and other Markov-based approaches. While HMMs are sometimes considered a conventional machine learning approach, we do not group them with those approaches because of (1) their ability to express latent variables; and (2) their inspectable model structure. These properties also put them on the border of computational and algorithmic levels. These approaches model behavior as transitions between states; such states are typically unobservable, and can be used to represent hidden human intentions and thoughts. These approaches have indirectly been shown to be robust to some classes of human idiosyncracies (such as forgotten actions), and we believe the potential is there for more. The same can be said for the second methodology we discuss, Dynamic Bayesian Networks, which are a generalization of Hidden Markov Models and have been shown to also be able to capture aspects of human beliefs and learning.

Topic models, on the other hand, do not model the structure of activities, and instead consider sets of observations holistically to decide how to label an activity. This allows them to be robust to variations in the patterns and timing of human activities, which is why we include them in this category. As with HMMs, topic models are sometimes considered as a conventional approach to machine learning, but we include them here because they also can express latent variables.

We finally discuss grammar-based models. Because of the unique way in which grammar-based models can probabilistically generate behavior, they can also capture idiosyncratic and non-idealized human behavior, and we include them here to highlight this flexibility.

Hidden Markov Models

Hidden Markov models (HMMs) are one common and effective way of modeling human behavior. An HMM consists of a set of possible states (which are hidden, or not directly observable) and probabilistic transitions between them. States can also probabilistically result in observations. Given a sequence of observations, statistical calculations (i.e., inference) can be used to find the most likely current state (Blunsom 2004). These calculations are simplified by the Markov assumption, which assumes that each state is independent of past states. Various forms of HMMs have been used in a variety of contexts, including robotics (Bennewitz et al. 2005), computational biology (Eddy 2004), and natural language processing (Blunsom 2004), among others.

The structure of the HMM is typically determined a priori; then, given data, the probabilisitic values of the transitions and observations can be learned using the Baum-Welch algorithm (Baum 1972), a form of Expectation-Maximization. Although the run-time of this algorithm grows exponentially with the number of states, it is also typically run offline and ahead of time. When using the model for inference online, the Viterbi algorithm allows the most likely sequence of hidden states resulting in the observations to be found in or near real-time (Rabiner 1989).

HMMs very naturally capture tasks that execute actions in a fairly rigid sequential order. Additionally, because states must be enumerated ahead of time, HMMs are well suited to recognizing tasks comprised of sequences that are easily written by hand. Thus, HMMs are a common choice for modeling human behavior in tasks involving manipulation and teleoperation (e.g., Yu et al. 2005; Aarno et al. 2005; Iba et al. 1999).

As the behavior or task being modeled becomes more complicated, the models themselves do, as well. Li and Okamura (2003) capture different intended actions of an operator haptically controlling a robot by defining a model where each action corresponds to a different trained HMM. The model with the highest cumulative likelihoods of observations is selected as the identified action; Kelley et al. (2008) have a similar structure. This structure allows these approaches to capture behavior as it is occurring, since it is not necessary to wait for the goal to be completed to observe it (Kelley et al. 2008).

The hierarchical and abstracting trend of HMMs also allows for recognition of behavior at higher levels (White et al. 2009; Bui et al. 2004, 2002). Nguyen et al. (2005) describe their application of hierarchical hidden Markov models to identifying human behavior within a controlled environment, and they also discuss the need for a Rao-Blackwellised particle filter (RBPF) to approximately solve the HMM for the most likely state and allow the approach to run in real time. Other structured, hierarchical-like formulations have also been used, such as coupled HMMs for modeling interactions (Natarajan and Nevatia 2007), layered HMMs for analysis at varying temporal granularity (Oliver et al. 2002), and factorial HMMs to provide better generalization capabilities (Kulić et al. 2007).

Even with the RBPF approximation and other recent work on tractability (Doucet et al. 2000; Bui 2003), however, these more sophisticated HMMs can still struggle with tractability and running in real-time. Another downside of the extra complexity is that it also requires more data to train in order to avoid overfitting the model to the observed data (Ghahramani 2001) .

Vasquez et al. (2008) work towards avoiding this issue by using an approximate learning algorithm based on the Growing Neural Gas algorithm to learn HMM parameters online and incrementally; they also use it to learn the structure simultaneously, removing the requirement of knowing the HMM structure a priori. Other heuristic approaches also work towards learning or refining a model's structure (Stolcke and Omohundro 1993; Freitag and McCallum 2000; Won et al. 2004). However, the problem of structure learning for HMMs is difficult to solve generally due to the very large search space, and is still an active area of research.

An expansion of HMMs that can also be used for human-robot collaboration is the Markov-based planning framework of Partially-Observable Markov Decision Processes (POMDP). The structure of a POMDP is very similar to an HMM, except that it also includes actions that can influence the probabilities of being in various states. The user's preferred sequences of actions (called policies) for different goals can be learned using Inverse Reinforcement Learning (Choi and Kim 2011; Makino and Takeuchi 2012) . The resulting model can then be used to infer or classify user intents (called policy recognition) in human-robot interaction; however, a very high computational expense makes it difficult for this to occur in real time (Ramırez and Geffner 2011). These approaches can also be used to generate human-like behavior by using the learned policy as the basis for developing a plan for the robot.

Markov planning frameworks can also be used to factor the human's behavior into the robot's actions. A variant of POMDPs, MOMDPs, includes the same hidden states to represent the human's intent or actions as POMDPs, but also includes fully-observable states to represent the robot's intents and actions. Nikolaidis et al. (2014), for example, models human users by classifying them by 'type' (such as "safe" or "efficient"); they then use that knowledge as part of a MOMDP, where the user type is considered the hidden variable, allowing the robot to better plan around its human counterpart.

Overall, Markov-based models like HMMs present both benefits and challenges to modeling humans for human-robot collaboration because of the Markov assumption and their sequential nature. Many HRI tasks, especially low-level ones like assistive teleoperation and shared control, are sequential in nature and so are easily and naturally represented as Markov chains; on the other hand, non-sequential tasks, such as plans that are partially ordered, are difficult to capture using these approaches.

The explicit sequences and connections between states also make it natural to predict what a human will do next based on what it is believed the human is currently doing; on the other hand, it prevents the models from capturing some of the more "human" aspects of human teammates, such as divergent beliefs, interruptions, distractions, or errors.

Dynamic Bayesian Networks

Dynamic Bayesian Networks (DBNs) are generalizations of Hidden Markov Models (Dagum et al. 1992; Murphy 2002). In DBNs, however, the unary state variables are replaced by multi-variate, mixed-observability Bayesian networks. Each such network/observation pair is referred to as a *time slice*. DBNs also adhere to the Markov assumption.

DBNs are more powerful than HMMs due to their ability to represent intradependencies between variables at the same time slice, as well as interdependencies between variables at different time slices. This allows models to capture correlations/causations between variables occurring at the same time, as well as how those variables may influence future state variables. These correspondences provide DBNs with the ability to combine input from multiple modalities, which can be very useful for human-robot collaboration where humans can simultaneously provide audio, visual and gestural data. In Oliver and Horvitz (2005), for example, the authors compared two systems that used data from video, audio and computer input. One of the systems trained a unique HMM for each data modality and the other system used a single DBN that had state variables and conditionals between variables across modalities. They found that the DBN outperformed the ensemble of HMMs when classifying office activities.

These correspondences also allow DBNs to be more robust to missing information, such as incomplete or partial observations of a human's actions, which is why we consider them to span the bridge to Marr's algorithmic level. Also supporting this position is work in cognitive science that leverages the flexibility of DBNs to capture human beliefs, desires and learning (Baker et al. 2011). This understanding can then be used to better interpret and predict the actions of others (Baker et al. 2005). Similar approaches have also been used to capture specific cognitive phenomena such as concept learning of shapes (Tenenbaum 1999), the production of co-verbal iconic gestures (Bergmann and Kopp 2010), and story understanding (Charniak and Goldman 1993).

Like with HMMs, the structure of DBNs must be defined a priori and by hand, typically by a domain expert. This is especially difficult in real-life situations where there are typically very large number of state variables with complex interdependencies. Also as with HMMs, structured learning research is being done to develop models automatically, but so far results are limited to specific problem domains (Gong and Xiang 2003; Pavlović et al. 1999).

An advantage of using a DBN is that generic learning and inference procedures can be used for any model. This gives researchers the flexibility to design models that are very representative of the behavior they are trying to capture without having to worry about how to solve them. On the other hand, the different optimizations available for learning and solving existing, well-defined, HMM structures are not available for general DBNs. Further, learning the parameters of large DBNs requires very large amounts of training data, which can make it difficult to take advantage of their representative power to capture sophisticated or long-term behavior. This limits their applicability for large-scale human-robot collaborations since it can be difficult to collect the large amount of data that would be required to train these networks. For that reason, DBNs as well as their HMM cousins tend to be used on specific, short-term or well-defined domains.

Topic Models

Topic models were originally designed for understanding text documents, and so are typically described using lexical terminology. We adopt this here to be consistent with the topic model literature.

Topic models are hierarchical, "bag-of-words" models that, canonically, ignore temporal-spatial dependencies between observations (Blei et al. 2003). There are two stages used to classify a sequence of observations. The first is identifying or learning a transformation function that bins each observation into

one of a set of possible words. Although theoretically any type of transformation function could be used for this, often K-means clustering is used.

The second level of classification involves classifying the distribution of words seen across a series of observations as a topic, or a mixture of topics. This represents the overall classification of, for example, an observed human activity. The mapping of distributions of words to topics can be done either in a supervised (as in semi-Latent Dirichlet Allocation; Wang and Mori 2009) or unsupervised manner (as in Latent Dirichlet Allocation (LDA), one of the most common topic models used; Blei et al. 2003). LDA has successfully been used to learn without supervision, for example, human activity from RGB-D video sequences (Zhang and Parker 2011).

Since most topic models ignore the spatio-temporal information of observations, they provide an understanding of the overall activity being performed that is robust to variations in lower-level activity order and position. This helps them be adept at recognizing human activity over larger timescales. In Huynh et al. (2008), topic models were used to automatically discover activity patterns in a user's daily routine; Rieping et al. (2014) take a similar approach. A daily routine cannot be easily specified by hand, as variable patterns exist for each activity within the routine, the routines range over a long period of time, and routines very often vary significantly for each instance. The robustness to temporal variation allows LDA to capture these routines, and is why we consider topic models to span the bridge between the Marr's computational and algorithmic levels.

Of course, the bag-of-words approach does not perform well in all domains, such as those where activities are typically performed according to dependencies between them. Due to their bag-of-words representations, much of the temporal information is lost unless it either encoded into the words, or the LDA model is modified to incorporate the temporal information in some way.

The first approach is taken by an extension to LDA, the Hierarchical Probabilistic Latent (HPL) model (Yin and Meng 2010), that encodes temporal information into the words of the model. Freedman et al. (2015) take the second approach, combining LDA and HMMs to perform human activity recognition using RGB-D data. The approach uses HMMs to incorporate in temporal information at the word level. Griffiths et al. (2004) take a similar approach.

The unsupervised aspect of LDAs helps them avoid many of the downfalls of the graphical models we have discussed, since the structure does not have to be hand-specified. On the other hand, it requires significant training data to discover activities on its own.

Grammar-Based Models

Context-free grammars use productions rules to specify the structure of a model. The rules specify how higher-level symbols (such as "drink coffee") can be composed by lower-level symbols (such as "pickup mug", "sip" some number of times, then "put down mug"). These production rules can then be used to model human behavior. For example, sequences of human actions can be recognized (such as by using HMMs and BNs; e.g., Ryoo and Aggarwal 2006) and then parsed using a grammar to better understand the structure of the behavior (Vilain 1990); these algorithms also tend to be efficient (Earley 1970; Turaga et al. 2008), making them useful for real-time human-robot collaboration. Similar to some of the knowledge-based approaches above, however, these approaches in general make strong assumptions about ordering, and assume human perfection and predictability.

More recent approaches to using grammars to model human behavior address various aspects of these criticisms. In particular, stochastic variants (SCFGs) relax some of the assumptions of infallibility and correctness of sensor data. Moore and Essa (2002) use SCFGs to recognize multi-task activities. By allowing for uncertainty, they can account for errors due to substitution, insertion or deletion of actions. While in their approach they use this ability to robustly handle errors in sensing, these errors could also be caused by human fallibility, adding an algorithmic component to

this computational model. Ivanov and Bobick (2000) and Ryoo and Aggarwal (2009) demonstrate a similar resistance to errors.

An extension of SCFGs, probabilistic state-dependent grammars (PSDGs) (Pynadath and Wellman 2000) include the consideration of the current state into the probability of a production. This allows preconditions and other state-dependent factors to affect its analysis. Given a fully observable state, the complexity of the recognizer is linear in the number of productions; however, given unobservable variables, the recognizer quickly fails to run in real time. It is also difficult to learn the model's parameters. Geib (2012) also includes the state into the grammar, but in a manner that allows for the better handling of loops in activities.

One challenge of using grammar-based approaches that remains unsolved is the difficulty of specifying and verifying the production rules and set of possible activity sequences that can be generated from it. Kitani et al. (2007) took a step towards addressing this problem by developing an algorithm that automatically learns grammar rules for an SCFG from noisy observations. In this work, however, the authors made several strong assumptions about the characteristics of the noise, and this remains a difficult problem to do generally (De La Higuera 2000). This is especially a problem for human-robot interaction where it is important that our model matches the human, and we view this as a major downside of using these approaches for modeling humans for human-robot collaboration. The difficulty in specifying and learning grammars also suggests that grammar-based approaches are best applied to structured domains; illustrating this, state of the art approaches are often demonstrated on regimented tasks such as the Towers of Hanoi (Minnen et al. 2003) and cooking (Kuehne et al. 2014).

Computational and Algorithmic Models Discussion

While understanding behavior at the computational level is sufficient and useful in many cases, most such approaches could benefit from some basic understanding of human reasoning, fallacy and error. In our opinion, this is a critical direction to head towards as we continue to expect more from our robot collaborators. Even on simple, routine tasks, humans make errors stemming from well-documented cognitive phenomena (Reason 1984) . The approaches described in this section, although mostly situated at the computational level, recognize, or have the potential to recognize, some of this human stochasticity. And while they do so with little intuition on how these errors come about, or without explicitly representing human reasoning, the tolerance they display increases their utility to human-robot collaboration, both by making their inferences more robust to noisy teammates, and by allowing them to help recognize when human teammates may unknowingly go off-script.

Algorithmic Models

The next set of methodologies we describe are classified as purely algorithmic in Marr's framework. These models strongly focus on capturing the representations and processes of human cognition, including the noisy aspect of human processing, and in large part assume that the question of *what* the human is doing is already known (such as by using one of the above computational accounts).

ACT-R/E (Trafton et al. 2013) is a cognitive architecture that models human cognition and memory at the process level. Each model written in ACT-R/E is typically specific to a task and must be verified against human performance for that task; once this initial work has been completed, however, models display a high fidelity to human behavior both in terms of reaction time and errors.

The model mReasoner captures the representations and processes people use to perform deductive and probabilistic reasoning. Given a set of facts or

statements about the world, mReasoner can predict the difficulty of drawing conclusions from those facts.

MAC/FAC (Forbus et al. 1995) models similarity-based retrieval. In terms of human-robot collaboration, this allows it to capture how people think of, for example, problems or decisions from the past that are similar to a problem they currently face. This potentially allows a robot can better reason about why a human went off-task: they could have remembered an alternate solution to the problem. The usefulness of this approach depends, however, on how much background knowledge it is given at the outset of an interaction.

ACT-R/E

ACT-R/E (Adaptive Character of Thought-Rational/Embodied) (Trafton et al. 2013) is a hybrid symbolic/sub-symbolic production-based system based on ACT-R (Anderson 2007). ACT-R/E is a cognitive architecture that is meant to model human cognition at the process level and to address how humans' limited-capacity brains handle the information processing requirements of their environment. ACT-R/E's goals are to maintain cognitive plausibility as much as possible while providing a functional architecture to explore embodied cognition, cognitive robotics, and human-robot interaction.

Given declarative knowledge (fact-based memories) and procedural knowledge (rule-based memories), as well as input from the world (visual, aural, etc.), ACT-R/E decides what productions in the model to fire next; these productions can change either the model's internal state (e.g., by creating new knowledge) or its physical one (e.g., by deciding to move its hand). It makes the decision of what production to fire next based on a) *symbolic* knowledge, such as who was where at what time; and b) *subsymbolic* knowledge, such as how relevant a fact is to the current situation, or how useful a production is expected to be when fired.

In terms of Marr's levels of analysis, models written in ACT-R/E are almost always at the algorithmic level: their representations and processes are well grounded in theory and empirical data. For example, ACT-R/E's declarative memory system (memory for facts) is based on the concept of *activation*. Activation values are a function of how frequently and recently that chunk has been accessed, as well as the extent to which the current context primes it (i.e., *spreading activation*). Cognitively, a chunk's activation represents how long an ACT-R/E model will take to remember a chunk, if it can even be remembered at all. The theory that underlies activation values has been empirically evaluated in numerous situations and across various tasks, and has been shown to be a remarkably good predictor of human declarative memory (Anderson et al. 1998; Anderson 1983; Schneider and Anderson 2011).

ACT-R/E models excel at capturing human behavior in relatively low level tasks or sub-tasks, such as predicting how long it takes someone to execute a task (Kennedy and Trafton 2007), predicting whether someone has forgotten a step in a task or a story (Trafton et al. 2011, 2012), or determining when someone needs additional information because their understanding of the situation is incorrect (Hiatt and Trafton 2010; Hiatt et al. 2011). These capabilities have clear value for human-robot collaboration.

ACT-R/E also, however, has several weaknesses for HRI. One weakness is that both models and parameters must be hand written and then verified against empirical data to confirm that the system as a whole captures human's behavior for a given task; there is currently no approach that learns the models in an unsupervised fashion. Mitigating these weakness, however, is that a relatively small amount of data is needed to verify any individual model. This, in large part, stems from the large amount of research and experimentation that has been done to develop and validate the architecture as a whole.

A second weakness is that ACT-R/E models are generally written at a task level that is quite specific and can be limited in scope. For example, understanding when and why someone forgets their medicine at a specific time is simple to model, but creating a model that takes into account a person's

entire daily medicine routine would be extremely complex and difficult to capture.

mReasoner

mReasoner is a cognitive model of human deductive and probabilistic reasoning (Khemlani and Johnson-Laird 2013; Khemlani et al. 2015). Its goal is to explain why humans find some inferences easy and other inferences difficult (Khemlani 2016). It is based on the theory that to reason, people construct and revise representations known as *mental models* from their knowledge, discourse, and observations. A mental model is an iconic simulation of a situation, i.e., its parts and relations corresponds to the structure of what it represents (Craik 1943; Johnson-Laird 2006).

mReasoner's fundamental operations include mechanisms for building, scanning, and revising models. It embodies two fundamental processes of inference (Kahneman 2011) required to build and manipulate mental models: *intuitions* concern how people rapidly construct models and draw inferences; *deliberations* concern how they revise their initial models and conclusions to correct errors. The system predicts that the more models that are required to carry out an inference, the more difficult and error-prone that inference will be. It also characterizes what those errors are. For example, the following reasoning problem:

- The fire happened *after* the alarm.
- The evacuation happened *before* the alarm.
- Did the fire happen after the evacuation? (Yes, necessarily.)

should be easy, because the system needs to generate just one model to solve it: one in which EVACUATION occurs, then ALARM, then FIRE. In contrast, for this reasoning problem:

- The fire happened *before* the alarm.
- The evacuation happened *before* the alarm.
- Did the fire happen after the evacuation? (Not necessarily, but it's possible.)

people often overlook the possibility that the FIRE can happen before the EVACUATION. mReasoner solves it by generating multiple models; it can explain people's difficulty as well as their erroneous responses.

mReasoner's ability to simulate common reasoning errors based on mental models makes it a useful tool in human-robot collaboration. Previous studies show that physical interactions are more robust when humans and robots rely on shared mental models of their environments and their tasks (Kennedy and Trafton 2007), and inferential interactions should follow the same constraints (Stout et al. 1999). When robots have access to a simulation of people's knowledge and their intuitive and deliberative inferential processes, they can anticipate human errors and provide remedial support.

One advantage of mReasoner for HRI is that its mechanisms are based on offline analyses of existing patterns in human reasoning; hence, as with ACT-R/E, it does not require extensive online training on large datasets compared to machine learning techniques. Nevertheless, its parameters allow the system to predict reasoning patterns from individual people (Khemlani and Johnson-Laird 2016). One limitation, however, is that it reasons over linguistic assertions instead of sensor data, and so robots that interface with it must be able to convert observations to a set of premises akin to the examples above. Carrying out those conversions in an efficient way remains a challenge for robots.

MAC/FAC

MAC/FAC is a model of similarity-based retrieval (Forbus et al. 1995). It captures how humans remember things that are similar to a given cue, such as remembering a story like one you just heard, or trying to solve a puzzle by remembering a similar one you solved in the past. Complicated concepts like this can have similarity across different levels, such as two, shallowly related stories about a dog named Rover, or two more deeply related stories that both have a protagonist travel after a life-changing event. The approach models how concepts like these are

represented in memory, as well as the operations that take place to retrieve them when needed.

The approach posits that retrieving these similar items happens in two stages: the fast MAC ("many are called") stage, where many items with surface similarity are selected as candidates; and the slower FAC ("few are chosen") stage, where deeper similarity and analogical reasoning are considered (Gentner 1983). This makes important predictions for human behavior during, for example, problem solving. If a human partner begins to solve a puzzle too quickly after a prompt, it is possible that they are solving it using a strategy from a different problem that is only shallowly related. On the other hand, if the human deliberates longer, they may have deeply mapped the current problem to one with a similar structure that therefore is likely to share a similar solution.

Similar to mReasoner, then, the approach can capture human errors and response times on certain types of tasks and so is a useful tool in human-robot collaboration. Additionally, as with mReasoner and ACT-R/E, because of the large amount of research and experimentation that has been done to develop this model, it can be validated in new situations with a relatively small amount of data. MAC/FAC also suffers from similar shortcomings, in that it models only one facet of human reasoning and so is most useful in conjunction with other approaches. The ubiquitous nature of the facet it captures, however, including analogical reasoning, is useful in many domains of human behavior (Liang and Forbus 2014; Lovett et al. 2009)

Algorithmic Models Discussion

The approaches discussed in this section model behavior at the algorithmic level of Marr's framework, and focus on capturing the representations and processes of human behavior to capture the *how* of human behavior and cognition. A fundamental advantage to basing human-robotic interactions on algorithmic accounts is that their explicit representations can allow the approaches to benefit related models and tasks. As a specific example, humans spontaneously use gesture to communicate about conceptual and spatiotemporal information, because gestures can improve thinking and reasoning (Bucciarelli et al. 2016; Wagner et al. 2004). Gestures, however, can be also outward signs of internal mental representations (Hostetter and Alibali 2008, 2010). Thus, using these algorithmic approaches to understand human reasoning can also enable robots to better understand their teammates' gestures, and, further, produce their own gestures in a way that helps their teammates grasp complex situations efficiently. This type of coupling between human cognition and physical gestures cannot emerge from computational-level theories; it requires an algorithmic account of relevant mental representations and their associated operations.

One downside of algorithmic approaches is that, for many of these models, the specific reasoning processes they capture are in isolation from another. For example, the above models can capture that you have retrieved an incorrect memory or made an incorrect inference, but not both. Additionally, unlike when capturing people's overall, observable behavior, the unobservable nature of people's thought processes and representations leads to difficulty modeling them explicitly and faithfully over long periods of time, or on tasks that are wide in scope. This typically leads to models at this level that are fairly narrow in scope and short in timescale. Although these limitations are practical ones, not theoretical ones, they are nonetheless difficult to overcome, and we view these shortcomings as the biggest downside of algorithmic models.

Implementational Models

Implementational models capture how algorithms are physically realized. This can be used to understand the processing and memory capabilities, and limitations, on algorithmic processes of human cognition (Stewart and Eliasmith 2008). Leabra (Local, Error-Drive and Associative, Biologically Realistic Algorithm), for

example, is a model of the brain's neocortex that captures many of the biological principles of neural processing (Kachergis et al. 2014); Spaun is another biologically-plausible model (Stewart and Eliasmith 2008).

In our view, the implementational level represents a promising area of research that can capture how people's cognition varies from internal and external influences. For example, the implementational level would include models of how stress might affect a worker's ability to focus, or how alcohol and caffeine can influence cognition. Models at this level, however, are very much still in the early stages of research, and so are difficult to use for human-robot collaboration in their current form. Currently, their useful predictions are better incorporated into approaches tailored at the algorithmic level, such as ACT-R/E (Vinokurov et al. 2012).

General Discussion

In this paper, we have discussed various techniques for modeling human cognition and behavior. Using Marr's levels of analysis, we have categorized different methodologies depending on what aspects of human cognition and behavior they capture. We have also discussed specific properties of the methodologies that affect their benefit to different scenarios in human-robot collaboration, such as how they handle human's noisy processing, and whether the models' structures are hand-coded or are learned from data.

When modeling human behavior and reasoning for human-robot interaction, it is important to select the methodology that will best further ones goals for the collaboration. These approaches have different knowledge to offer to human-robot collaboration. For example, assume that a researcher would like to use a robot to help a person stack boxes in a specific order. A computational-level model would be able to recognize what action the person was taking, or what object the person was reaching for. An algorithmic-level model would not be able to recognize these things, but, given them, but would be able to model why the person took that step, even if it was erroneous. A model at the implementational level could predict how the person's performance might degrade if they got little sleep the night before.

To help to better understand how the different levels and approaches address different modeling needs, we next discuss two different graphs that compare the approaches given the considerations that, as we have described, naturally stem from Marr's levels of analysis: how they handle human errors, how much data is required to train them in new domains, whether the model structures can be automatically learned or must be hand-specified, and the timescale that they typically operate over. We follow that with some discussion of how the approaches can be combined to make use of their complimentary strengths.

Comparison of Approaches

As we have discussed, computational approaches to modeling human behavior focus on recognizing what a human is doing, algorithmic approaches focus on how the human accomplishes his or her task, and implementational approaches focus on how those operations are physically realized. As we have shown, there are a variety of approaches that have been developed to model each of these levels, each with their own strengths and weaknesses.

Figure 1 compares the different approaches according to how they handle human error, and the timescale that the approaches generally operate in. It shows a clear correlation between the different levels and how the approaches handle the noisiness of human processing time and error. Computational approaches, as expected, generally assume perfect rationality and are sometimes robust to noisy observations; computational and algorithmic borderline approaches are typically robust to noisy observations and, in some limited cases, can account for human processing errors; and algorithmic approaches focus on accounting on errors and noise in human processing. Implementational approaches, in contrast, do not clearly fit into this trend. We expect, however, that as approaches at such levels

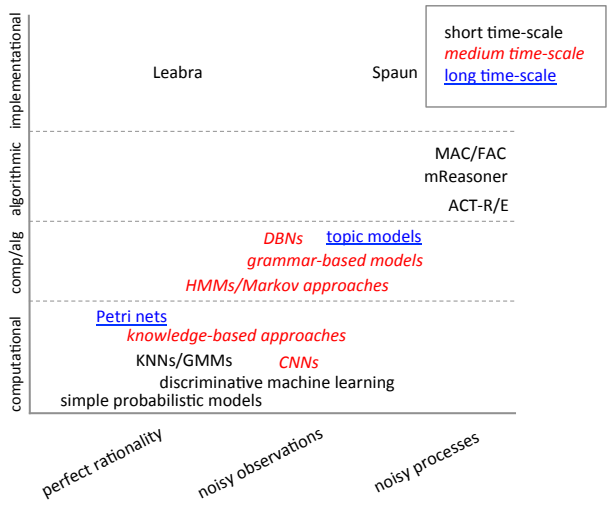

Figure 1. Plot of different approaches considering how they handle errors, as well as the timescale that the approaches generally operate in.

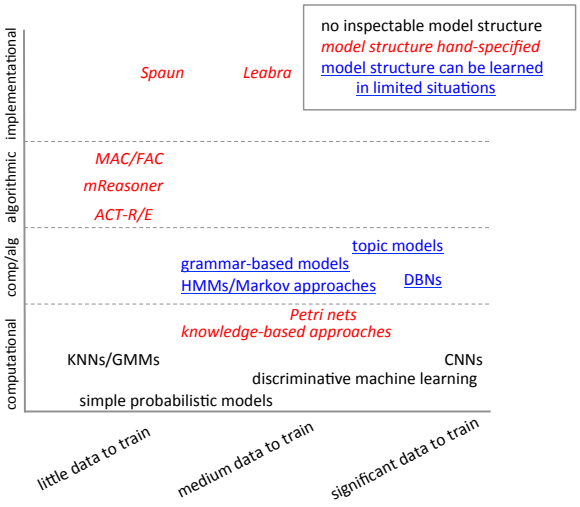

Figure 2. Plot of different approaches considering the amount of data required to train a model in a new domain, as well as whether a model's structure must be hand-specified or can be automatically learned.

become more established, they will become stronger at accounting for noisy human processes and errors.

This figure also suggests that the approaches that are robust to noisy observations, but that do not explicitly model them, are those that have the most success operating at long time. Intuitively, this makes sense. Over long periods of time, models that assume perfect rationality will likely diverge from the human's true state; in contrast, models that explicitly capture human processing noise will be left with a state space that is too large to reason over. In addition, the representations of these two groups of models does not lend themselves well to extended task operation: mathematical equations currently have limited use when talking about a days' worth of activities, as do low-level productions and memory processes. We do, however, believe that there is promising work to be done here to ameliorate these weaknesses. For example, one could apply a Kalman filter to an ACT-R/E model to prune and update its state space from observations of what a human partner is doing, allowing it to better operate over longer periods of time.

Figure 2 shows the approaches separated by how much data is required to train them, as well as whether the model structure can be learned or must be hand-specified. Again, we see the computational/algorithm borderline level stand out from the other two levels, and, again, this difference is an intuitive one. Research on models at that level has started focusing more, in general, on automatically learning models' structure; that, in turn, requires more data than learning the parameters of an existing model. As before, it also suggests a possible line of research for algorithmic models, where, if provided more data, they could also automatically learn their models' structure within the existing architecture.

Combining Approaches Across Levels

The above figures also suggest that the approaches at the different levels have complimentary strengths and weaknesses. This then raises the question of how approaches can be combined to take advantage of each of their strengths while minimizing their weaknesses. For example, we earlier discussed how one approach recognizes human actions by using HMMs or BNs, and then uses a grammar-based model to parse the sequence of actions and understand the structure of behavior. This work combines approaches that

target different timescales, increasing the scope of the combined model's power; we also believe that there can be an event greater benefit in combining approaches that target different Marr's levels.

For example, one could have an HMM where each latent state corresponds to a different model in ACT-R/E. Such a composite model would allow the robot to, in part, use its knowledge of how a person completes a task, given noisy human processing, to help it infer what task the person is completing. In other words, such a model would give the robot the combined power of both the computational and algorithmic levels. This combination of different levels of modeling is the best way, we believe, to model human-robot collaboration tasks: using both information about what the person is doing as well as how they are doing it.

References

Aarno D, Ekvall S and Kragić D (2005) Adaptive virtual fixtures for machine-assisted teleoperation tasks. In: *Proceedings of the International Conference on Robotics and Automation*.

Admoni H and Scassellati B (2014) Data-driven model of nonverbal behavior for socially assistive human-robot interactions. In: *Proceedings of the 16th International Conference on Multimodal Interaction*. ACM, pp. 196–199.

Albanese M, Chellappa R, Moscato V, Picariello A, Subrahmanian V, Turaga P and Udrea O (2008) A constrained probabilistic petri net framework for human activity detection in video. *IEEE Transactions on Multimedia* 10(6): 982–996.

Anderson J (1983) A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior* 22(3): 261–295.

Anderson J (2007) *How Can the Human Mind Occur in the Physical Universe?* Oxford University Press.

Anderson J, Bothell D, Lebiere C and Matessa M (1998) An integrated theory of list memory. *Journal of Memory and Language* 38(4): 341–380.

Andrist S, Mutlu B and Tapus A (2015) Look like me: Matching robot personality via gaze to increase motivation. In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, pp. 3603–3612.

Avrahami-Zilberbrand D, Kaminka G and Zarosim H (2005) Fast and complete symbolic plan recognition: Allowing for duration, interleaved execution, and lossy observations. In: *Proc. of the AAAI Workshop on Modeling Others from Observations*.

Baker C, Saxe R and Tenenbaum JB (2005) Bayesian models of human action understanding. In: *Advances in neural information processing systems*. pp. 99–106.

Baker CL, Saxe RR and Tenenbaum JB (2011) Bayesian theory of mind: Modeling joint belief-desire attribution. In: *Proceedings of the Thirty-Third Annual Meeting of the Cognitive Science Society*.

Baum LE (1972) An equality and associated maximization technique in statistical estimation for probabilistic functions of markov processes. *Inequalities* 3: 1–8.

Bennewitz M, Burgard W, Cielniak G and Thrun S (2005) Learning motion patterns of people for compliant robot motion. *The International Journal of Robotics Research* 24(1): 31–48.

Bergmann K and Kopp S (2010) Modeling the production of co-verbal iconic gestures by learning bayesian decision networks. *Applied Artificial Intelligence* 24(6): 530–551.

Billard A and Siegwart R (2004) Robot learning from demonstration. *Robotics and Autonomous Systems* 47(2): 65–67.

Blaylock N and Allen J (2014) Hierarchical goal recognition. In: Sukthankar G, Goldman RP, Geib C, Pynadath DV and Bui HH (eds.) *Plan, Activity, And Intent Recognition: Theory and Practice*. Morgan Kaufman.

Blei DM, Ng AY and Jordan MI (2003) Latent dirichlet allocation. *the Journal of machine Learning research* 3: 993–1022.

Blunsom P (2004) Hidden markov models. URL http://digital.cs.usu.edu/~cyan/CS7960/hmm-tutorial.pdf.

Breazeal C (2004) Social interactions in HRI: the robot view. *Systems, Man, and Cybernetics, Part C:*

Applications and Reviews, IEEE Transactions on 34(2): 181–186.

Breazeal C, Gray J and Berlin M (2009) An embodied cognition approach to mindreading skills for socially intelligent robots. *International Journal of Robotics Research* 28(5): 656–680.

Bucciarelli M, Mackiewicz R, Khemlani SS and Johnson-Laird PN (2016) Children's creation of algorithms: simulations and gestures. *Journal of Cognitive Psychology* : 1–22.

Bui HH (2003) A general model for online probabilistic plan recognition. In: *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*. pp. 1309–1315.

Bui HH, Phung DQ and Venkatesh S (2004) Hierarchical hidden markov models with general state hierarchy. In: *Proceedings of the National Conference on Artificial Intelligence*. pp. 324–329.

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

Calinon S, Guenter F and Billard A (2007) On learning, representing, and generalizing a task in a humanoid robot. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on* 37(2): 286–298.

Carberry S (2001) Techniques for plan recognition. *User Modeling and User-Adapted Interaction* 11: 31–48.

Castel C, Chaudron L and Tessier C (1996) What is going on? a high level interpretation of sequences of images. In: *ECCV Workshop on Conceptual Descriptions from Images*. pp. 13–27.

Castellano G, Leite I, Pereira A, Martinho C, Paiva A and McOwan PW (2012) Detecting engagement in HRI: An exploration of social and task-based context. In: *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*. IEEE, pp. 421–428.

Charniak E and Goldman RP (1993) A bayesian model of plan recognition. *Artificial Intelligence* 64(1): 53 – 79. DOI:http://dx.doi.org/10.1016/0004-3702(93)90060-O. URL http://www.sciencedirect.com/science/article/pii/000437029390060O.

Choi J and Kim KE (2011) Inverse reinforcement learning in partially observable environments. *Journal of Machine Learning Research* 12(Mar): 691–730.

Craik K (1943) *The nature of explanation*. Cambridge University Press.

Dagum P, Galper A and Horvitz E (1992) Dynamic network models for forecasting. In: *Proceedings of the eighth international conference on uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., pp. 41–48.

Dai W, Yang Q, Xue GR and Yu Y (2007) Boosting for transfer learning. In: *Proceedings of the 24th international conference on Machine learning*. ACM, pp. 193–200.

De La Higuera C (2000) Current trends in grammatical inference. In: *Advances in Pattern Recognition*. Springer, pp. 28–31.

Doucet A, De Freitas N, Murphy K and Russell S (2000) Rao-blackwellised particle filtering for dynamic bayesian networks. In: *Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., pp. 176–183.

Dragan AD, Lee KC and Srinivasa SS (2013) Legibility and predictability of robot motion. In: *Human-Robot Interaction (HRI), 2013 8th ACM/IEEE International Conference on*.

Dragan AD and Srinivasa SS (2012) Formalizing assistive teleoperation. In: *Proceedings of Robotics: Science and Systems*.

Earley J (1970) An efficient context-free parsing algorithm. *Communications of the ACM* 13(2): 94–102.

Eddy SR (2004) What is a hidden Markov model? In: 22 (ed.) *Nature Biotechnology*. pp. 1315–1316.

Ermes M, Parkka J, Mantyjarvi J and Korhonen I (2008) Detection of daily activities and sports with wearable sensors in controlled and uncontrolled conditions. *IEEE Transactions on Information Technology in Biomedicine* 12(1): 20–26.

Fagg AH, Rosenstein M, Platt, Jr R and Grupen RA (2004) Extracting user intent in mixed initiative teleoperator control. In: *American Institute of Aeronautics and*

Astronautics.

Fong T, Nourbakhsh I and Dautenhahn K (2003) A survey of socially interactive robots. *Robotics and autonomous systems* 42(3): 143–166.

Fong T, Thrope C and Baur C (2001) Collaboration, dialogue, and human-robot interaction. In: *Proceedings of the International Symposium on Robotics Research*.

Forbus KD, Gentner D and Law K (1995) MAC/FAC: A model of similarity-based retrieval. *Cognitive Science* 19: 141–205.

Freedman RG, Jung HT and Zilberstein S (2015) Temporal and object relations in unsupervised plan and activity recognition. In: *AAAI Fall Symposium on AI for Human-Robot Interaction (AI-HRI)*.

Freitag D and McCallum A (2000) Information extraction with HMM structures learned by stochastic optimization. *AAAI/IAAI* : 584–589.

Geib C (2012) Considering state in plan recognition with lexicalized grammars. In: *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*.

Gentner D (1983) Structure-mapping: A theoretical framework for analogy. *Cognitive Science* 7: 155–170.

Ghahramani Z (2001) An introduction to hidden markov models and bayesian networks. *International Journal of Pattern Recognition and Artificial Intelligence* 15(01): 9–42.

Ghanem N, DeMenthon D, Doermann D, and Davis L (2003) Representation and recognition of events in surveillance video using petri nets. In: *Computer Vision and Pattern Recognition Workshops (CVPRW)*.

Gillespie K, Molineaux M, Floyd MW, Vattam SS and Aha DW (2015) Goal reasoning for an autonomous squad member. In: *Annual Conference on Advances in Cognitive Systems: Workshop on Goal Reasoning*.

Gong S and Xiang T (2003) Recognition of group activities using dynamic probabilistic networks. In: *Ninth IEEE International Conference on Computer Vision*. pp. 742–749.

Griffiths T, Steyvers M, Blei D and Tenenbaum J (2004) Integrating topics and syntax. In: *NIPS*. pp. 537–544.

Griffiths TL, Chater N, Kemp C, Perfors A and Tenenbaum JB (2010) Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in cognitive sciences* 14(8): 357–364.

Grosz BJ and Kraus S (1999) The evolution of sharedplans. In: Wooldridge M and Rao A (eds.) *Foundations and Theories of Rational Agents*. Springer.

Hecht F, Azad P and Dillmann R (2009) Markerless human motion tracking with a flexible model and appearance learning. In: *IEEE International Conference on Robotics and Automation, 2009. ICRA '09*. pp. 3173–3179. DOI:10.1109/ROBOT.2009.5152494.

Hiatt LM, Harrison AM and Trafton JG (2011) Accommodating human variability in human-robot teams through theory of mind. In: *Proceedings of the International Joint Conference on Artificial Intelligence*.

Hiatt LM and Trafton JG (2010) A cognitive model of theory of mind. In: *Proceedings of the International Conference on Cognitive Modeling*.

Hostetter AB and Alibali MW (2008) Visible embodiment: Gestures as simulated action. *Psychonomic bulletin & review* 15(3): 495–514.

Hostetter AB and Alibali MW (2010) Language, gesture, action! a test of the gesture as simulated action framework. *Journal of Memory and Language* 63(2): 245–257.

Huynh T, Fritz M and Schiele B (2008) Discovery of activity patterns using topic models. In: *Proceedings of the 10th international conference on Ubiquitous computing*. ACM, pp. 10–19.

Iba S, Weghe JMV, Paredis CJJ and Khosla PK (1999) An architecture for gesture-based control of mobile robots. In: *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*.

Ivanov YA and Bobick AF (2000) Recognition of visual activities and interactions by stochastic parsing. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 22(8): 852–872.

Jenkins OC, González G and Loper MM (2007) Tracking human motion and actions for interactive robots. In: *Proceedings of the ACM/IEEE international conference on Human-robot interaction*. ACM, pp. 365–372.

Johnson-Laird PN (2006) *How we reason*. Oxford University Press, USA.

Kachergis G, Wyatte D, O'Reilly RC, De Kleijn R and Hommel B (2014) A continuous-time neural model for sequential action. *Phil. Trans. R. Soc. B* 369(1655): 20130623.

Kahneman D (2011) *Thinking, fast and slow*. Macmillan.

Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R and Fei-Fei L (2014) Large-scale video classification with convolutional neural networks. In: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*.

Kautz H and Allen J (1986) Generalized plan recognition. In: *Proceedings of the Fifth National Conference on Artificial Intelligence*. pp. 32–37.

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

Kennedy WG and Trafton JG (2007) Using simulations to model shared mental models. In: *Proceedings of the 8th International Conference on Cognitive Modeling*. pp. 253–254.

Khemlani S (2016) Automating human inference. In: *Proceedings of the 2nd IJCAI Workshop on Bridging the Gap between Human and Automated Reasoning*. pp. 1–4.

Khemlani S and Johnson-Laird PN (2013) The processes of inference. *Argument & Computation* 4(1): 4–20.

Khemlani S and Johnson-Laird PN (2016) How people differ in syllogistic reasoning. In: Papafragou A, Grodner D, Mirman D, and Trueswell J (eds.) *Proceedings of the 38th Annual Conference of the Cognitive Science Society*.

Khemlani S, Lotstein M, Trafton JG and Johnson-Laird P (2015) Immediate inferences from quantified assertions. *The Quarterly Journal of Experimental Psychology* 68(10): 2073–2096.

Kitani KM, Sato Y and Sugimoto A (2007) Recovering the basic structure of human activities from a video-based symbol string. In: *Motion and Video Computing, 2007. WMVC'07. IEEE Workshop on*. IEEE, pp. 9–9.

Krizhevsky A, Sutskever I and Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: *Advances in neural information processing systems*. pp. 1097–1105.

Kuehne H, Arslan A and Serre T (2014) The language of actions: Recovering the syntax and semantics of goal-directed human activities. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 780–787.

Kulić D, Takano W and Nakamura Y (2007) Representability of human motions by factorial hidden markov models. In: *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, pp. 2388–2393.

Lavee G, Rudzsky M, Rivlin E and Borzin A (2010) Video event modeling and recognition in generalized stochastic petri nets. *Circuits and Systems for Video Technology, IEEE Transactions on* 20(1): 102–118.

Lesh N, Rich C and Sidner CL (1999) Using plan recognition in human-computer collaboration. In: *Proceedings of the Seventh International Conference on User Modeling*. Springer.

Levine S and Williams B (2014) Concurrent plan recognition and execution for human-robot teams. In: *International Conference on Automated Planning and Scheduling*.

Li M and Okamura AM (2003) Recognition of operator motions for real-time assistance using virtual fixtures. In: *Proceedings of the 11th Sympsoium on Haptic Interfaces for Virtual Environment and Teleoperator Systems*.

Liang C and Forbus K (2014) Constructing hierarchical concepts via analogical generalization. In: *Proceedings of the Cognitive Science Society*.

Liu C, Conn K, Sarkar N and Stone W (2008) Online Affect Detection and Robot Behavior Adaptation for Intervention of Children With Autism. *IEEE Transactions on Robotics* 24(4): 883–896. DOI:10.1109/TRO.2008.2001362.

Lovett A, Sagi E, Gentner D and Forbus K (2009) Modeling perceptual similarity as analogy resolves the paradox of difference detection. In: *Proceedings of the 2nd International Analogy Conference*.

Makino T and Takeuchi J (2012) Apprenticeship learning for model parameters of partially observable environments. In: *Proceedings of the 29th International Conference on Machine Learning*.

Marr D (1982) *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco, CA: W. H. Freeman.

Maynord M, Vattam S and Aha DW (2015) Increasing the runtime speed of case-based plan recognition. In: *Proceedings of the Twenty-Eighth Florida Artificial Intelligence Research Society Conference*.

Minnen D, Essa I and Starner T (2003) Expectation grammars: Leveraging high-level expectations for activity recognition. In: *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 2. IEEE, pp. II–626.

Moore D and Essa I (2002) Recognizing multitasked activities from video using stochastic context-free grammar. In: *AAAI/IAAI*. pp. 770–776.

Morales Saiki LY, Satake S, Huq R, Glas D, Kanda T and Hagita N (2012) How do people walk side-by-side?: Using a computational model of human behavior for a social robot. In: *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*. ACM, pp. 301–308.

Mower E, Feil-Seifer DJ, Mataric MJ and Narayanan S (2007) Investigating implicit cues for user state estimation in human-robot interaction using physiological measurements. In: *Robot and Human interactive Communication, 2007. RO-MAN 2007. The 16th IEEE International Symposium on*. IEEE, pp. 1125–1130.

Murata M (1989) Petri nets: Properties, analysis and applications. *Proceedings of the IEEE* 77(4): 541–580.

Murphy KP (2002) *Dynamic bayesian networks: representation, inference and learning*. PhD Thesis, University of California, Berkeley.

Natarajan P and Nevatia R (2007) Coupled hidden semi markov models for activity recognition. In: *IEEE Workshop on Motion and Video Computing*. IEEE.

Nevatia R, Zhao T and Hongeng S (2003) Hierarchical language-based representation of events in video streams. In: *Computer Vision and Pattern Recognition Workshop, 2003. CVPRW'03. Conference on*, volume 4. IEEE, pp. 39–39.

Nguyen NT, Phung DQ, Venkatesh S and Bui H (2005) Learning and detecting activities from movement trajectories using the hierarchical hidden markov model. In: *Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 955–960.

Nikolaidis S, Gu K, Ramakrishnan R and Shah J (2014) Efficient model learning for human-robot collaborative tasks. *arXiv preprint arXiv:1405.6341* .

Oliver N and Horvitz E (2005) A comparison of hmms and dynamic bayesian networks for recognizing office activities. *User Modeling* : 199–209.

Oliver N, Horvitz E and Garg A (2002) Layered representations for human activity recognition. In: *Fourth IEEE International Conference on Multimodal Interfaces*.

Ou-Yang C and Winarjo H (2011) Petri-net integration: An approach to support multi-agent process mining. *Expert Systems with Applications* 38(4): 4039–4051.

Pavlović V, Rehg JM, Cham TJ and Murphy KP (1999) A dynamic bayesian network approach to figure tracking using learned dynamic models. In: *The Proceedings of the Seventh IEEE International Conference on Computer Vision*. pp. 94–101.

Pérez-D'Arpino C and Shah JA (2015) Fast target prediction of human reaching motion for cooperative human-robot manipulation tasks using time series classification. In: *IEEE International Conference on Robotics and Automation (ICRA)*.

Perrault R and Allen J (1980) A plan-based analysis of indirect speech acts. *American Journal of Computational Linguistics* 6(3-4): 167–182.

Pynadath DV and Wellman MP (2000) Probabilistic state-dependent grammars for plan recognition. In: *Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., pp. 507–514.

Rabiner LR (1989) A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE* 77(2): 257–286.

Ramırez M and Geffner H (2011) Goal recognition over pomdps: Inferring the intention of a pomdp agent. In: *IJCAI*. IJCAI/AAAI, pp. 2009–2014.

Ravi N, Dandekar N, Mysore P and Littman M (2005) Activity recognition from accelerometer data. In: *Association for the Advancement of Artificial Intelligence (AAAI)*.

Reason JT (1984) Absent-mindedness and cognitive control. *Everyday Memory, Actions and Absent-mindedness* .

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

Ristic B, Arulampalam S and Gordon N (2004) Beyond the kalman filter. *IEEE Aerospace and Electronic Systems Magazine* 19(7): 37–38.

Rota N and Thonnat M (2000) Activity recognition from video sequences using declarative models. In: *ECAI*. Citeseer, pp. 673–680.

Ryoo MS and Aggarwal JK (2006) Recognition of composite human activities through context-free grammar based representation. In: *Computer vision and pattern recognition, 2006 ieee computer society conference on*, volume 2. IEEE, pp. 1709–1718.

Ryoo MS and Aggarwal JK (2009) Semantic representation and recognition of continued and recursive human activities. *International journal of computer vision* 82(1): 1–24.

Scassellati B (2002) Theory of mind for a humanoid robot. *Autonomous Robots* 12(1): 13–24.

Schmidhuber J (2015) Deep learning in neural networks: An overview. *Neural Networks* 61: 85–117.

Schneider D and Anderson J (2011) A memory-based model of hick's law. *Cognitive Psychology* 62(3): 193–222.

Stewart TC and Eliasmith C (2008) Building production systems with realistic spiking neurons. In: *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Stolcke A and Omohundro S (1993) Hidden markov model induction by bayesian model merging. *Advances in neural information processing systems* : 11–11.

Stout RJ, Cannon-Bowers JA, Salas E and Milanovich DM (1999) Planning, shared mental models, and coordinated performance: An empirical link is established. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 41(1): 61–71.

Tang Y (2013) Deep learning using linear support vector machines. *arXiv preprint arXiv:1306.0239* .

Tapus A, Mataric MJ and Scasselati B (2007) Socially assistive robotics [grand challenges of robotics]. *Robotics & Automation Magazine, IEEE* 14(1): 35–42.

Tenenbaum JB (1999) Bayesian modeling of human concept learning. In: *Advances in Neural Information Processing Systems*.

Toshev A and Szegedy C (2014) DeepPose: Human pose estimation via deep neural networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Trafton JG, Altmann EM and Ratwani RM (2011) A memory for goals model of sequence errors. *Cognitive Systems Research* 12: 134–143.

Trafton JG, Hiatt LM, Harrison AM, Tamborello, II FP, Khemlani SS and Schultz AC (2013) ACT-R/E: An embodied cognitive architecture for human-robot interaction. *Journal of Human-Robot Interaction* 2(1): 30–55.

Trafton JG, Jacobs A and Harrison AM (2012) Building and verifying a predictive model of interruption resumption. *Proceedings of the IEEE* 100(3): 648–659.

Turaga P, Chellappa R, Subrahmanian VS and Udrea O (2008) Machine recognition of human activities: A survey. *Circuits and Systems for Video Technology, IEEE Transactions on* 18(11): 1473–1488.

Vasquez D, Fraichard T, Aycard O and Laugier C (2008) Intentional motion on-line learning and prediction. *Machine Vision and Applications* 19(5-6): 411–425.

Vilain MB (1990) Getting serious about parsing plans: A grammatical analysis of plan recognition. In: *AAAI*. pp. 190–197.

Vinokurov Y, Lebiere C, Wyatte D, Herd S and O'Reilly R (2012) Unsurpervised learning in hybrid cognitive architectures. In: *Neural-Symbolic Learning and*

Reasoning Workshop at AAAI.

Vu VT, Bremond F and Thonnat M (2003) Automatic video interpretation: A novel algorithm for temporal scenario recognition. In: *IJCAI*, volume 3. pp. 1295–1300.

Wagner SM, Nusbaum H and Goldin-Meadow S (2004) Probing the mental representation of gesture: Is handwaving spatial? *Journal of Memory and Language* 50(4): 395–407.

Wang S, Yang J, Chen N, Chen X and Zhang Q (2005) Human activity recognition with user-free accelerometers in the sensor networks. In: *International Conference on Neural Networks and Brain (ICNNB)*.

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

White BA, Blaylock N and Bölöni L (2009) Analyzing team actions with cascading HMM. In: *FLAIRS Conference*.

Wilensky R (1983) *Planning and Understanding*. Addison-Wesley.

Won KJ, Prügel-Bennett A and Krogh A (2004) Training HMM structure with genetic algorithm for biological sequence analysis. *Bioinformatics* 20(18): 3613–3619.

Yin J and Meng Y (2010) Human activity recognition in video using a hierarchical probabilistic latent model. In: *Computer Vision and Pattern Recognition Workshops (CVPRW)*. pp. 15–20.

Yu W, Alqasemi R, Dubey R and Pernalete N (2005) Telemanipulation assistance based on motion intention recognition. In: *International Conference on Robotics and Automation (ICRA)*.

Zhang H and Parker LE (2011) 4-dimensional local spatio-temporal features for human activity recognition. In: *IEEE/RSJ international conference on Intelligent robots and systems (IROS)*. pp. 2044–2049.