Human Behavior Modeling with Maximum Entropy Inverse Optimal Control

Brian D. Ziebart, Andrew Maas, J.Andrew Bagnell, and **Anind K. Dey**
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
bziebart@cs.cmu.edu, amaas@andrew.cmu.edu, dbagnell@ri.cmu.edu, anind@cs.cmu.edu

Abstract

In our research, we view human behavior as a structured sequence of context-sensitive decisions. We develop a conditional probabilistic model for predicting human decisions given the contextual situation. Our approach employs the principle of maximum entropy within the Markov Decision Process framework. Modeling human behavior is reduced to recovering a context-sensitive utility function that explains demonstrated behavior within the probabilistic model.

In this work, we review the development of our probabilistic model (Ziebart *et al.* 2008a) and the results of its application to modeling the context-sensitive route preferences of drivers (Ziebart *et al.* 2008b). We additionally expand the approach's applicability to domains with stochastic dynamics, present preliminary experiments on modeling time-usage, and discuss remaining challenges for applying our approach to other human behavior modeling problems.

Introduction

Accurate models of human behavior are an important component for realizing an improved symbiosis between humankind and technology across a number of different domains. These models enable intelligent computer interfaces that can anticipate user actions and intentions, ubiquitous computing environments that automatically adapt to the behaviors of their occupants, and robots with human-like behavior that complements our own actions and goals. We view human behavior as a structured sequence of context-sensitive decisions. For simple domains, the choices for each decision may be a deterministic function of a small set of variables, and possible to completely specify by hand. However, for sufficiently complex behavior, manual specification is too difficult, and instead the model should be automatically constructed from observed behavior.

Fortunately, most human behavior is purposeful – people take actions to efficiently accomplish objectives – rather than completely random. For example, when traversing the road network (Figure 1), drivers are trying to reach some destination, and choosing routes that have a low personalized *cost* (in terms of time, money, stress, etc.). The modeling problem then naturally decomposes into modeling a person's changing "purposes," and the context-efficient actions taken to achieve those objectives. When "purposes" have domain-level similarity, we'd expect the notion of efficiency to be similar for differing objectives, allowing behavior fulfilling different goals to be useful for modeling common notions of utility and efficiency. Markov decision processes provide a framework for representing these objectives and notions of efficiency.

Figure 1: The road network covering a portion of Pittsburgh.

Dealing with the uncertainty inherent in observed behavior represents a key challenge in the machine learning problem of constructing these models. There are many sources of this uncertainty. The observed agent may base its decision making on additional information that the learner may not be able to observe or that may differ from the observer's information due to noise. Another possibility is that the agent's behavior may be intrinsically random due to the nature of its decision selection process. Additionally, nature often imposes additional randomness on the outcome of observed actions that must be taken into account.

We take a thoroughly probabilistic approach to reasoning about uncertainty in behavior modeling (Ziebart *et al.* 2008a). Under the constraint of matching the reward value of demonstrated behavior within the Markov decision process (MDP) framework, we employ the principle of *maximum entropy* to resolve the ambiguity in choosing a distribution over decisions. We provide efficient algorithms for learning and inference within this setting. The resulting distribution is a probabilistic model that normalizes globally over behaviors and can be understood as an extension to chain conditional random fields that incorporates the dynamics of the planning system and extends to the infinite horizon. We view our approach as a bridge between optimal decision making frameworks, which provide strong performance guarantees, but ignore decision uncertainty, and probabilistic graphical models, which model decision uncertainty, but provide no performance guarantees.

Our research effort is motivated by the problem of modeling real human decision making. First, we apply our approach to modeling the context-dependent route preferences of taxi drivers (Ziebart *et al.* 2008b) using 100,000 miles of collected GPS data. Second, we model the daily activities of individuals collected from time-use surveys. Lastly, we discuss remaining challenges for applying our approach to other human behavior modeling problems.

Related Work

Our approach reconciles two disparate threads of research – inverse optimal control and probabilistic graphical models. We review each in this section.

Inverse Optimal Control and Imitation Learning

Inverse optimal control (IOC) (Boyd *et al.* 1994; Ng & Russell 2000), originally posed by Kalman, describes the problem of recovering an agent's reward function, R(s,a), given demonstrated sequence(s) of actions, $\{\tilde{\zeta_1} = \{a_1|s_1, a_2|s_2, ...\}, \tilde{\zeta_2}, ...\}$, when the remainder of the MDP is known. It is synonymously known as inverse reinforcement learning (IRL). Vectors of reward factors $\mathbf{f}_{s,a}$ describe each available action, and the reward function is assumed to be a linear function of those factors, $R(s,a) = \theta^\top \mathbf{f}_{s,a}$ parameterized by reward weights, θ. Ng & Russell (2000) formulate inverse optimal control as the recovery of reward weights, θ, that make demonstrated behavior optimal.

Unfortunately this formulation is ill-posed. Demonstrated behavior is optimal for many different reward weights, including degeneracies (e.g., all zeros). Abbeel & Ng (2004) propose recovering reward weights so that a planner based on those reward weights and the demonstrated trajectories have equal reward (in expectation). This formulation reduces to matching the planner and demonstrated trajectories' expected *feature counts*, $\mathbf{f}_\zeta = \sum_{s,a \in \zeta} \mathbf{f}_{s,a}$.

$$\sum_\zeta P_{\text{plan}}(\zeta|\theta)\mathbf{f}_\zeta = \mathbf{f}_{\tilde{\zeta}} \tag{1}$$

Abbeel & Ng (2004) employ a series of deterministic controls obtained from "solving" the optimal MDP for the distribution over trajectories. When sub-optimal behavior is demonstrated (due to the agent's imperfection or unobserved reward factors), mixtures of policies are required to match feature counts. Many different mixtures will match feature counts and no method is proposed to resolve this ambiguity.

Ratliff, Bagnell, & Zinkevich (2006) resolve this ambiguity by posing inverse optimal control as a maximum margin problem with loss-augmentation. While the approach yields a unique solution, it suffer from significant drawbacks when no single reward function makes demonstrated behavior both optimal and significantly better than any alternative behavior. This arises quite frequently when, for instance, the behavior demonstrated by the agent is imperfect, or the planning algorithm only captures a part of the relevant state-space and cannot perfectly describe the observed behavior.

An imitation learning approach to the problem, which still aims to obtain similar behavior, but without any performance guarantees, relaxes the MDP optimality assumption by employing the MDP "solution" policy's reward, $Q_\theta(a,s) = \max_{\zeta \in \Xi_{s,a}} \theta^\top \mathbf{f}_\zeta$, within a Boltzmann probability distribution.

$$P(\text{action } a|s) = \frac{e^{Q_\theta(a,s)}}{\sum_{\text{action } a'} e^{Q_\theta(a',s)}} \tag{2}$$

Neu & Szepesvári (2007) employ this distribution within a loss function penalizing the squared difference in probability between the model's action distribution and the demonstrated action distribution. Ramachandran & Amir (2007) utilize it within a Bayesian approach to obtain a posterior distribution over reward values using Markov Chain Monte Carlo simulation. The main weaknesses of the model are that maximum likelihood (and MAP) estimation of parameters is a non-convex optimization, and the learned model lacks performance guarantees with respect to the Markov decision process.

Our proposed approach is both probabilistic and convex. Unlike the mixture of optimal behaviors (Abbeel & Ng 2004), training behavior will always have non-zero probability in our model, and parameter choices are well-defined. Unlike maximum margin planning (Ratliff, Bagnell, & Zinkevich 2006), our method realistically assumes that demonstrated behavior may be sub-optimal (at least for the features observed by the learner). Finally, unlike the Boltzmann Q-value stochastic model (Neu & Szepesvári 2007; Ramachandran & Amir 2007), learning in our model is convex, cannot get "stuck" in local maxima, and provides performance guarantees.

Probabilistic Graphical Models

A great deal of research within the machine learning community has focused on developing probabilistic graphical models to address uncertainty in data. These models provide a framework for representing independence relationships between variables, learning probabilistic models of data, and inferring the values of latent variables. Two main variants are directed models (i.e., Bayesian networks) and undirected models (i.e., Markov random fields and conditional random fields).

Bayesian networks model the joint distribution of a set of variables by factoring the distribution into a product of conditional probabilities of each variable given its "parent" variables (Pearl 1985). A number of Bayesian network models for decision making have been proposed (Attias 2003;

Verma & Rao 2006). Unfortunately in many real world decision making problems, decisions are based not only on the current action's features, but the features of all subsequent actions as well. This leads to a very non-compact model that generalizes poorly when predicting withheld data. We investigate these deficiencies empirically in our experiments.

Markov random fields model the energy between combinations of variables using potential functions. In their generalization, conditional random fields (CRFs) (Lafferty, McCallum, & Pereira 2001), the potential functions can depend on an additional set of variables that are themselves not modeled. In a number of recognition tasks, these additional variables are observations, and the CRF is employed to recognize underlying structured properties from these observations. This approach has been applied to recognition problems for text (Lafferty, McCallum, & Pereira 2001), vision (Kumar & Hebert 2006), and activities (Liao, Fox, & Kautz 2007; Vail, Veloso, & Lafferty 2007). The maximum entropy inverse optimal control model we derive for Markov decision problems with deterministic action outcomes can be interpreted as a chain conditional random field where the entire sequence of decisions is conditioned on all state and action features. This is significantly different than how conditional random fields have been applied for recognition tasks, where labels for each variable in the sequence are conditioned on local observations from portion of the sequence.

Maximum Entropy IOC

We take a different approach to matching feature counts that allows us to deal with this ambiguity in a principled way, and results in a single stochastic policy. We employ the principle of maximum entropy (Jaynes 1957) to resolve ambiguities in choosing distributions. This principle leads us to the distribution over behaviors constrained to match feature expectations, while being no more committed to any particular path than this constraint requires.

Decision Sequence Distribution

Unlike previous work that reasons about policies, we consider a distribution over the entire class of possible behaviors. For deterministic MDPs, the class of behaviors consists of paths of (potentially) variable length. Similar to distributions of policies, many different distributions of paths match feature counts when any demonstrated behavior is sub-optimal. Any one distribution from among this set may exhibit a preference for some of the paths over others that is not implied by the path features. We employ the principle of maximum entropy, which resolves this ambiguity by choosing the distribution that does not exhibit any additional preferences beyond matching feature expectations (Equation 1). The resulting distribution over paths for deterministic MDPs is parameterized by reward weights θ (Equation 3). Under this model, plans with equivalent rewards have equal probabilities, and plans with higher rewards are exponentially more preferred.

$$P(\zeta_i|\theta) = \frac{1}{Z(\theta)} e^{\theta^\top \mathbf{f}_{\zeta_i}} = \frac{1}{Z(\theta)} e^{\sum_{s_j \in \zeta_i} \theta^\top \mathbf{f}_{s_j}} \qquad (3)$$

Given parameter weights, the *partition function*, $Z(\theta)$, always converges for finite horizon problems and infinite horizons problems with discounted reward weights. For infinite horizon problems with zero-reward absorbing states, the partition function can fail to converge even when the rewards of all states are negative. However, given demonstrated trajectories that are absorbed in a finite number of steps, the reward weights maximizing entropy must be convergent.

Stochastic Policies

This distribution over paths provides a stochastic policy (i.e., a distribution over the available actions of each state) when the partition function of Equation 3 converges. The probability of an action is weighted by the expected exponentiated rewards of all paths that begin with that action.

$$P(\text{action } a|\theta) \propto \sum_{\zeta : a \in \zeta_{t=0}} P(\zeta|\theta) \qquad (4)$$

Learning from Demonstrated Behavior

Maximizing the entropy of the distribution over paths subject to the feature constraints from observed data implies that we maximize the likelihood of the observed data under the maximum entropy (exponential family) distribution derived above (Jaynes 1957).

$$\theta^* = \operatorname*{argmax}_\theta L(\theta) = \operatorname*{argmax}_\theta \sum_{\text{examples}} \log P(\tilde{\zeta}|\theta)$$

Given the agent's expected feature counts, this function is convex and the optima can be obtained using gradient-based optimization methods. The gradient is the difference between expected empirical feature counts and the leaner's expected feature counts, which can be expressed in terms of expected state visitation frequencies, D_{s_i}.

$$\nabla L(\theta) = \tilde{\mathbf{f}} - \sum_\zeta P(\zeta|\theta)\mathbf{f}_\zeta = \tilde{\mathbf{f}} - \sum_{s_i} D_{s_i} \mathbf{f}_{s_i} \qquad (5)$$

At the maxima, the feature expectations match, guaranteeing that the learner performs equivalently to the agent's demonstrated behavior regardless of the actual reward weights the agent is attempting to optimize (Abbeel & Ng 2004).

Efficient State Frequency Calculations

Given the expected state frequencies, the gradient can easily be computed (Equation 5) for optimization. The most straight-forward approach for computing the expected state frequencies is based on enumerating each possible path. Unfortunately, the exponential growth of paths with the MDP's time horizon makes enumeration-based approaches computationally infeasible.

Instead, our algorithm computes the expected state occupancy frequencies efficiently using dynamic programming (forward-backward algorithm for Conditional Random Fields or value iteration in Markov decision problems). The key observation is that the partition function can be defined recursively.

$$Z(\theta, s) = \sum_{\zeta_s} e^{\theta^\top \mathbf{f}_{\zeta_s}} = \sum_{\text{action } a} \left[e^{\theta^\top \mathbf{f}_{s,a}} \sum_{\zeta_{s,a}} e^{\theta^\top \mathbf{f}_{\zeta_{s,a}}} \right]$$

Here we denote all paths starting at state s (and with action a) as ζ_s (and $\zeta_{s,a}$). We refer the reader to our previous paper for full details (Ziebart *et al.* 2008a).

Stochastic Dynamics

In general, the outcomes of a person's actions may be influenced by randomness. In this case, the next state given an action is a stochastic function of the current state according to the conditional probability distribution, $P(s_{t+1}|s_t, a_t)$. Our previous work (Ziebart *et al.* 2008a) provides an approximate approach for modeling behavior in this setting. We now present an exact approach based on the principle of maximum entropy.

First, we define an *uncontrolled* distribution over trajectories, $Q(\zeta) \propto \prod_{s_{t+1}, s_t, a_t \in \zeta} P(s_{t+1}|s_t, a_t)$. We maximize the entropy of our distribution over trajectories, $P(\zeta)$, relative to this uncontrolled distribution. Solving the dual of this optimization yields a new formula for recursively computing the partition function:

$$Z(\theta, s) = \sum_{\text{action } a} e^{\theta^\top \mathbf{f}_{s,a} + \sum_{s'} P(s'|s,a) Z(\theta, s')} \qquad (6)$$

From this recursive formula, action probabilities are obtained. State frequencies are then computed from the action probabilities for learning (Equation 5).

Driver Route Modeling

Our research effort on maximum entropy approaches to IOC was motivated by applications of imitation learning of driver route choices. We are interested in recovering a utility function useful for *predicting driving behavior* as well as for *route recommendation*.

Route Choice as an MDP

Road networks present a large planning space with known structure. We model this structure for the road network surrounding Pittsburgh, Pennsylvania, as a deterministic MDP with over 300,000 states (i.e., road segments) and 900,000 actions (i.e., transitions at intersections). We assume that drivers who are executing plans within the road network are attempting to reach some goal while efficiently optimizing some trade-off between time, safety, stress, fuel costs, maintenance costs, and other factors. We call this value a *cost* (i.e., a negative reward). We represent the destination within the MDP as an absorbing state where no additional costs are incurred. Different trips have different destinations and slightly different corresponding MDPs. We assume that the reward weight is independent of the goal state and therefore a single reward weight can be learned from many MDPs that differ only in goal state.

Collecting and Processing GPS Data

We collected over 100,000 miles of GPS trace data (Figure 2) from taxi drivers. We fit the GPS traces to the road network using a particle filter and applied our model to learn driver preferences as a function of road network features (e.g., segment distance, speed limits, road class, turn type) (Ziebart *et al.* 2008a). Our evaluations (Table 1) show significant improvements in most likely path estimation and path density estimation of our model over other Boltzmann Q-value models (Ramachandran & Amir 2007; Neu & Szepesvári 2007) and Maximum Margin Planning (Ratliff, Bagnell, & Zinkevich 2006).

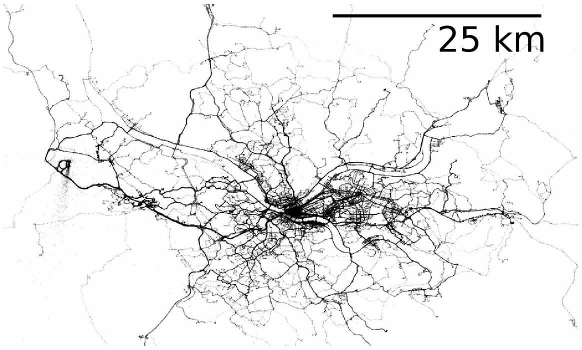

Figure 2: The collected GPS datapoints

	Matching	**90% Match**	**Log Prob**
Time-based	72.38%	43.12%	N/A
Max Margin	75.29%	46.56%	N/A
Action	77.30%	50.37%	-7.91
Action (costs)	77.74%	50.75%	N/A
MaxEnt paths	**78.79%**	**52.98%**	**-6.85**

Table 1: Evaluation results for optimal estimated travel time route, max margin route, Boltzmann Q-value distributions (Action) and Maximum Entropy

In extensions to this work (Ziebart *et al.* 2008b), we added contextual information (time of day, accidents, construction, congestion) to the model and compared it to other approaches previously applied to route prediction, turn prediction, and destination prediction. In Table 2 we compare against directed graphical models, which have been employed for transportation routine modeling (Liao *et al.* 2007). Since we are only concerned with single modal transportation, we compare against Markov models of decision at next intersection conditioned on the goal location (Simmons *et al.* 2006) and conditioned on the previous k road segments (Krumm 2008).

Model	**Dist. Match**	**90% Match**
Markov (1x1)	62.4%	30.1%
Markov (3x3)	62.5%	30.1%
Markov (5x5)	62.5%	29.9%
Markov (10x10)	62.4%	29.6%
Markov (30x30)	62.2%	29.4%
Travel Time	72.5%	44.0%
Our Approach	**82.6%**	**61.0%**

Table 2: Evaluation results for Markov Model with various grid sizes, time-based model, and our umodel

We also evaluated our model on the problem of predicting destination given partial trajectory by simply employing Bayes rule and incorporating a prior distribution over destinations. In Figure 3, we compare against the Predestination system (Krumm & Horvitz 2006) and a destination-based Markov model (Simmons *et al.* 2006). Predestination discretizes the world into grid cells and probabilistically predicts drivers' destinations based on a statistical model of driver efficiency. It assumes a fixed metric (travel time) and models efficiency (i.e., preference) given that metric, whereas our model assumes a fixed preference model and learn the driver's metric.

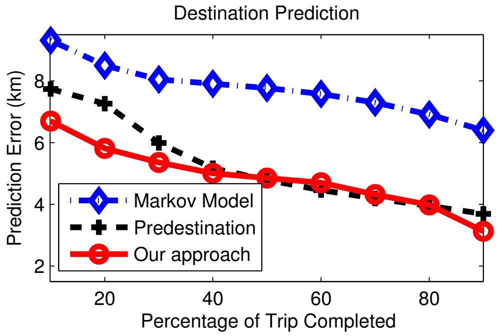

Figure 3: The best Markov Model, Predestination, and our approach's prediction errors

We find both our model and Predestination significantly outperform the Markov model, and our model performs somewhat better given large amount (90%) and small amount (10%–30%) of trip completed.

Time-Use Modeling

We now present preliminary experiments in applying our approach to modeling the time-use data collected from 12,248 individuals in the American Time-Use Survey (ATUS). As has been argued by Partridge & Golle (2008), this data serves as a valuable resource to the human behavior modeling community due to its broad scope. The data covers a broad corpus of activities collected from a diverse set of individuals over a long time period. We apply our approach to model one day of time-use for an individual based on demographic information (e.g., age, gender).

Model	Bits
Naïve Model	1200.9
Stationary Markov Model	122.6
Our approach	**118.3**

Table 3: Preliminary evaluation results for three models.

We evaluate three different models using 80% of the dataset for training and 20% for testing. We use the empirical average log probability of the test set, $\tilde{E}[\log_2 P(\mathbf{y}|\mathbf{x})]$ as our metric of evaluation. We choose base 2 so that the results can be interpreted as the number of bits required (on average) to represent a day's sequence of activities (from the 18 broadest activity classes defined in the corpus). In Table 3, we evaluate on three different models. The first is a naïve model that assumes a uniform distribution over activities and an independence between each consecutive timestep. It serves as a baseline. We next evaluate a stationary Markov model. Finally, we evaluate our approach, where we condition on demographic information (e.g., age, gender) and also incorporate time-of-day. We find that of the three models, our approach performs the best.

Future efforts in modeling this data will involve incorporating additional information into the predictive model, allowing additional flexibility in how time-of-day is incorporated, and conducting a more comprehensive evaluation.

Applicability Challenges

We now outline the challenges for applying our approach more generally to problems of human behavior modeling.

Modeling Behavior Duration

In the driving domain, the duration spent on any one road segment is influenced by many random external factors. Rather than explicitly model the duration as part of the state space, we abstracted expected duration as one component in our cost function, and instead considered the sequence of decisions as our state space. For many human behavior modeling applications, the duration of different behavior or activities, and not just the sequence of activities is important.

A well known shortcoming of Markovian models is that the duration of staying in any state through self-transition tends toward a decaying geometric distribution. For many human behaviors, [pseudo-]Gaussian distributions better model the duration of time spent on any particular behavior. For instance, depending on the exact person, he may brush his teeth for 90 seconds give or take roughly 15 seconds.

Semi-Markov decision processes (Sutton, Precup, & Singh 1999) and semi-Markov conditional random fields (Sarawagi & Cohen 2004) have been previously proposed. Extending their ideas to the feature-based utility function of our maximum entropy model, while still enabling efficient inference and learning, is an important challenge for modeling human behavior.

Leveraging Human Similarity

For many domains, behavior is very personalized, and personalized models are much more important to obtain accurate predictions. Obtaining models for an individual rather than a group is trivial in our approach – simply use an individual's training data rather than a group's training data to construct the model. However, in domains where data is not plentiful for each individual, a model trained on a small amount of an individual's data may not generalize as well as a model built for a group of people with more pooled data.

In domains where characteristics are known about each individual, we hope to leverage the data of similar individuals to build a more accurate model of behavior for a new individual. This will enable, for instance, a good prior model of behavior based on demographic information when no prior behavior has been observed for a user new to a system. Developing and evaluating alternative methods for accomplishing this goal remains as a future challenge.

Conclusions and Future Work

We have presented our recently developed novel approach for modeling human behavior by bridging two disparate lines of research – optimal decision modeling and probabilistic graphical models. The resulting model has a number of mathmetical niceties – it is compact, it can be reasoned about efficiently, and it can be trained efficiently as a convex optimization. Unlike optimal decision making, our maximum entropy approach yields a probabilistic model that can be easily incorporated within a Bayesian framework. Unlike directed graphical models, which degrade poorly to random walks in the absence of data, our model generalizes well and provides strong performance guarantees.

We applied our method to the problem of modeling route preferences, and evaluated the differences between our model and other imitation learning models using a small feature space, and comparing against other previous approaches to the problem with additional contextual information. We also employed our approach on modeling time-use based on demographic data. We found quite promising results in terms of comparative predictive accuracy.

Finally, we discussed a number of challenges for applying our approach more generally to other behavior modeling domains. We specifically outlined modeling durations of different behaviors and groups of similar people as important challenges for future research.

Acknowledgments

The authors thank Eric Oatneal and Jerry Campolongo of Yellow Cab Pittsburgh for their assistance and John Krumm for his help in improving portions of this paper and sharing GPS hardware hacks. Finally, we thank the Quality of Life Technology Center/National Science Foundation for supporting this research under Grant No. EEEC-0540865.

References

Abbeel, P., and Ng, A. Y. 2004. Apprenticeship learning via inverse reinforcement learning. In *Proc. ICML*, 1–8.

Attias, H. 2003. Planning by probabilistic inference. In *Proc. of the 9th Int. Workshop on Artificial Intelligence and Statistics*.

Boyd, S.; El Ghaoui, L.; Feron, E.; and Balakrishnan, V. 1994. Linear matrix inequalities in system and control theory. *SIAM* 15.

Jaynes, E. T. 1957. Information theory and statistical mechanics. *Physical Review* 106:620–630.

Krumm, J., and Horvitz, E. 2006. Predestination: Inferring destinations from partial trajectories. In *Proc. Ubicomp*, 243–260.

Krumm, J. 2008. A markov model for driver route prediction. *Society of Automative Engineers (SAE) World Congress*.

Kumar, S., and Hebert, M. 2006. Discriminative random fields. *Int. J. Comput. Vision* 68(2):179–201.

Lafferty, J.; McCallum, A.; and Pereira, F. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. ICML*, 282–289.

Liao, L.; Patterson, D. J.; Fox, D.; and Kautz, H. 2007. Learning and inferring transportation routines. *Artificial Intelligence* 171(5-6):311–331.

Liao, L.; Fox, D.; and Kautz, H. 2007. Extracting places and activities from gps traces using hierarchical conditional random fields. *Int. J. Rob. Res.* 26(1):119–134.

Neu, G., and Szepesvári, C. 2007. Apprenticeship learning using inverse reinforcement learning and gradient methods. In *Proc. UAI*, 295–302.

Ng, A. Y., and Russell, S. 2000. Algorithms for inverse reinforcement learning. In *Proc. ICML*, 663–670.

Partridge, K., and Golle, P. 2008. On using existing time-use study data for ubiquitous computing applications. In *Proceedings of UbiComp*, 144–153.

Pearl, J. 1985. Bayesian networks: A model of self-activated memory for evidential reasoning. In *Proceedings of the 7th Conference of the Cognitive Science Society, University of California, Irvine*, 329–334.

Ramachandran, D., and Amir, E. 2007. Bayesian inverse reinforcement learning. In *Proc. IJCAI*, 2586–2591.

Ratliff, N.; Bagnell, J. A.; and Zinkevich, M. 2006. Maximum margin planning. In *Proc. ICML*, 729–736.

Sarawagi, S., and Cohen, W. W. 2004. Semi-markov conditional random fields for information extraction. In *Proc. NIPS*, 1185–1192.

Simmons, R.; Browning, B.; Zhang, Y.; and Sadekar, V. 2006. Learning to predict driver route and destination intent. *Proc. Intelligent Transportation Systems Conference* 127–132.

Sutton, R. S.; Precup, D.; and Singh, S. 1999. Between mdps and semi-mdps: a framework for temporal abstraction in reinforcement learning. *Artificial Intelligence* 112(1-2):181–211.

Vail, D. L.; Veloso, M. M.; and Lafferty, J. D. 2007. Conditional random fields for activity recognition. In *Proc. AAMAS*, 1–8.

Verma, D., and Rao, R. 2006. Goal-based imitation as probabilistic inference over graphical models. In *Proc. NIPS*, 1393–1400.

Ziebart, B. D.; Maas, A.; Bagnell, J. A.; and Dey, A. K. 2008a. Maximum entropy inverse reinforcement learning. In *Proc. AAAI*.

Ziebart, B. D.; Maas, A.; Dey, A. K.; and Bagnell, J. A. 2008b. Navigate like a cabbie: Probabilistic reasoning from observed context-aware behavior. In *Proc. Ubicomp*.