Intelligent Robotic Systems: Papers from the AAAI 2013 Workshop

Towards Cooperative Bayesian Human-Robot Perception: Theory, Experiments, Opportunities

Nisar Ahmed, Eric Sample, Tsung-Lin Yang, Daniel Lee, Lucas de la Garza, Ahmed Elsamadisi, Arturo Sullivan, Kai Wang, Xinxiang Lao, Rina Tse and Mark Campbell

Autonomous System Laboratory, Cornell University, Ithaca, NY 14853
e-mail: nra6@cornell.edu

Abstract

Robust integration of robotic and human perception abilities can greatly enhance the execution of complex information-driven tasks like search and rescue. Our goal is to formally characterize and combine diverse information streams obtained from multiple autonomous robots and humans within a unified probabilistic framework that naturally supports autonomous perception, human situational awareness, and cooperative human-robot task execution under stochastic uncertainties. This approach requires well-designed human-robot interfaces, flexible and accurate probabilistic models for exploiting "human sensor" data, and sophisticated Bayesian inference methods for efficient learning and online dynamic state estimation. We review some of recent theoretical developments and insights from experiments using real human-robot teams, and discuss some open challenges for future research.

Introduction

Although advances in AI continue to improve robotic self-sufficiency in the face of complex real-world uncertainties, the physical and computational limitations of real robots often force human reasoning to enter in at some level to help ensure robustness in planning and perception, e.g. via manual tuning of heuristics and model parameters by expert human programmers; mixed-initiative/supervisory control by trained operators (Sheridan, 2002; Fong, et al. 2003); or natural interactions with untrained non-experts (Bauer, et al. 2009). As such, intelligent human-robot interaction will remain an important issue for the foreseeable future. Indeed, as sophisticated automatons become more pervasive, extensive work by the human factors community suggests they should be designed to treat human intelligence as a valuable (but constrained and imperfect) resource that complements their own (constrained and imperfect) machine intelligence (Sheridan, 2002). In other words, robots should not only acknowledge and support different facets of human intelligence (i.e. situational awareness, decision making, etc.), but also know how to exploit them for improving their own autonomous behavior. Such ideas have been explored by the robotics community using uncertainty-based AI, although the primary focus has been on interpretation of multimodal command inputs for human-assisted robot planning, e.g. via natural language speech, sketches or physical gestures (Huang, et al., 2011; Tellex, et al. 2011; Shah and Campbell, 2011; Bauer, et al. 2009).

In contrast, we are interested in exploring how humans provide *generalized sensor observations* that can be formally combined with robot sensor data to improve autonomous state estimation and model learning. Our main thrust is towards complex information-driven applications such as scientific exploration, environmental monitoring, search and rescue, and surveillance, where humans are often in a position to contribute useful information beyond robotic sensing horizons, especially in large, harsh, remote, and/or dynamic settings. As shown in (Lewis, et al. 2009), cooperative perception can also help improve overall human-robot team performance by allowing operators to better maintain high-level situational awareness, which in turn leads to more effective delegation of cognitively demanding low-level planning/navigation tasks. However, the problem of formally integrating human-robot sensing through uncertainty-based AI has not been well-studied.

Early work by (Kaupp, et al. 2007) and (Bourgault, et al. 2008) showed how certain human sensor inputs could be formally characterized and fused with robotic sensor data for augmented physical perception through the Bayesian paradigm. However, these works greatly limit the expected complexity and scope of human sensor inputs, largely for the sake of analytical and computational tractability. For instance, (Kaupp, et al. 2007) assume that humans provide numerical range and bearing measurement data for target localization ("The object is at range 10 m and bearing 45 degrees"), while (Bourgault, et al. 2008) assumes that humans provide binary "detection/no detection" visual observations for a 2D multi-target search problem. Humans are capable of providing a much broader range of information in such scenarios, e.g. using natural language semantic information (“The target is nearby in front of me” or “Nothing is behind the truck moving North”). Although more flexible and user-friendly, such “soft” human sensor data present significant challenges for cooperative Bayesian perception. Unlike robotic sensors, semantic human sensor data cannot be modeled solely from physical first principles, and they are quite sensitive to both contextual and cognitive factors (Hall and Jordan, 2010).

In this work, we summarize our recent and ongoing research toward a unified Bayesian framework that formally addresses these issues and naturally supports robust cooperative perception with general human-robot data sources. We seek flexible yet accurate joint probabilistic models of human-robot perception processes (e.g. via DBNs, MRFs, factor graphs, etc.), as well as sophisticated Bayesian inference methods that allow these models to be efficiently used for intelligent human-robot interaction in a wide variety of applications. We review some of recent modeling and algorithmic developments along with experimental results using real human-robot teams, and discuss some insights and open research problems that are relevant to AI and robotics.

Motivating Example: Bayesian Target Search

Futuristic robotic search and rescue operations will require teams of unmanned vehicles and human supervisors to gather and coordinate complex information in order to detect, localize and appropriately respond to interesting objects/events in uncertain environments (Lewis, et al. 2009; Goodrich, et al. 2009; Kruijff, et al. 2012). Figure 1 shows the experimental testbed we developed in (Ahmed, et al. 2013; Ponda, et al. 2011; Sample, et al. 2012) to study the fundamentally related problem of autonomous target search under conditions of limited, ambiguous, and possibly unreliable information.

The experimental scenario consists of an autonomous Pioneer ground robot conducting a visual search for multiple static targets (orange traffic cones) in a cluttered area using a known map. The robot is remotely supervised by a human operator seated at a linked computer, which has access to the robot’s video feed, environment map, and pose data. The robot localizes itself through a Vicon motion tracking system and can autonomously plan and safely navigate search paths. It can also detect targets visually up to a limited 1m range using simple blob detection, but cannot determine their identities (a number written on the cone). The operator’s primary role is to confirm detections reported by the robot and provide target IDs, but he/she cannot directly teleoperate or command the robot.

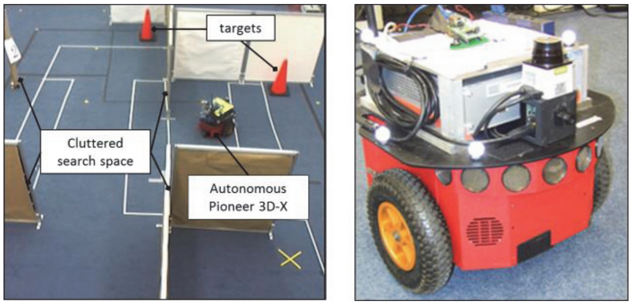

Figure 1 Experimental indoor search area and Pioneer 3DX search robot.

The human-robot team’s mission is to detect and localize all targets within a limited amount of time (7 or 15 mins). To make the search scenario more realistic and challenging, the operator is given a noisy video feed (presented in artificially blurred grayscale with small variable time delays) and false targets may be present (a blue cone or unlabeled cone that can only be identified through close inspection).

The robot determines its search path based on sensor information it fuses together via sequential Bayesian inference. Given a common prior belief $p(x)$ over each target location x, the robot uses observation likelihood $p(z_r|x,q_r)$ for binary visual detector reports z_r (“detection” or “no detection”) and robot poses q_r to arrive at a posterior belief,

$$p(x|z_r, q_r) \propto p(x) \prod_{r=1}^{N_r} p(z_r|x, q_r),$$

which summarizes the information from N_r independent sensor reports and can be calculated recursively. The robot uses a very simple (suboptimal) greedy search strategy that visits maximum a posteriori (MAP) target location estimates at different points in time, which accounts for the fact that $p(x|z_r,\ q_r)$ is generally a complex multi-modal probability distribution. D*Lite (Koenig and Likhachev, 1999) is used to find collision-free paths to MAP points.

Numerous trials on this testbed revealed that the Bayesian search strategy was quite sensitive to both the miss rate of the robot’s visual target detector and to the subtle coupling between the initial uncertainty $p(x)$ and the resulting locations for z_r. In general, targets can be successfully found given a diffuse prior that covers the whole search space, but most targets will never be found given a “bad” misinformed prior that forces the robot to search the space very slowly with its limited sensing field (see Figure 2). If the robot moves past a target in its field of view without detecting it, the robot will not return to the target’s location until the posterior pdf accumulates enough evidence there again from searching the remainder of the search space. If both of these “failure modes” are present, the robot’s search performance (in terms of number of

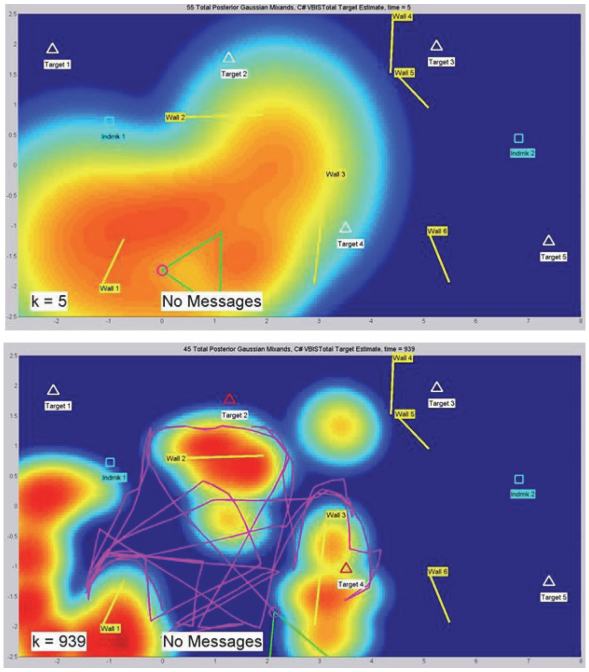
55 Total Posterior Gaussian Mixands, C#VBISTotal Target Estimate, time = 5
Target 1
Target 2
Target 3
Target 4
Target 5
Wall 1
Wall 2
Wall 3
Wall 4
Wall 5
Wall 6
lndmk 1
lndmk 2
k = 5
No Messages
45 Total Posterior Gaussian Mixands, C#VBISTotal Target Estimate, time = 939
k = 939
No Messages

Figure 2 Search map with bad prior (top) and resulting posterior (bottom) after 15 minutes. (blue =low probability, small triangles= target locations, purple line = robot path).

targets found, total distance traveled and MAP localization error) can be very poor.

Semantic Human Sensor Fusion

One strategy for improving performance in these situations is to augment the robot's sensor data with observations made by the human operator during the course of the mission. This not only helps to correct missed detections by the robot, but also allows the posterior to be modified in areas far away from the current robot pose at any given time, so that inconsistent prior beliefs can be naturally corrected with (noisy) data that are consistent with the true target locations. Furthermore, this allows the cognitively difficult tasks of robot planning and navigation to remain fully autonomous, so that the human can focus on interpreting the video feed for useful information.

To save time and cognitive effort, operators will be naturally inclined to report information via simple semantic expressions like "There is nothing behind that wall" or "Something is next to you". In our testbed, operators are restricted to reporting spatial data cues s_h of the fixed form "<Something/Nothing> is <*preposition*> <*anchor point*>", where each field member is selected from a pre-defined dictionary displayed on the operator's computer console. The prepositions included "next to", "nearby", "far from", "behind" and "in front of", while various landmarks in the search map and the robot's location served as spatial anchor points. The "Something / Nothing" field allows the operator to convey either positive or negative information in each s_h.

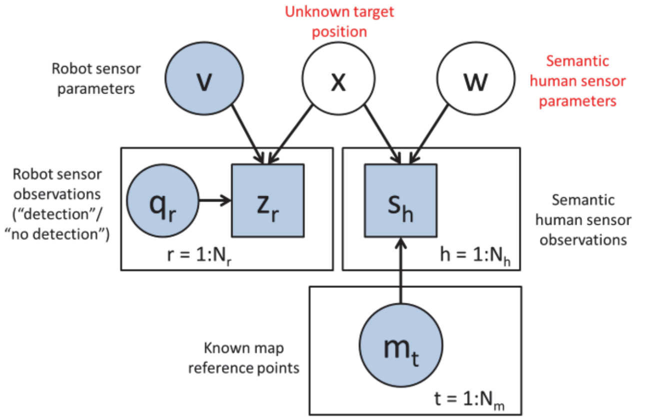
Unknown target position
Robot sensor parameters
v
x
w
Semantic human sensor parameters
Robot sensor observations ("detection"/ "no detection")
q_r
z_r
s_h
r = 1:N_r
h = 1:N_h
Semantic human sensor observations
Known map reference points
m_t
t = 1:N_m

Figure 3 Probabilistic graphical model for human-robot sensor fusion (square nodes are discrete random variables; shaded nodes are observed).

As shown by the directed graphical model in Figure 3, Bayesian inference for the target search problem now involves fusing both z_r and s_h observations given $p(x)$. This leads to two important technical challenges: (i) how to model $p(s_h|x,m_t)$? (ii) how to perform the required inference operations online and represent the posterior efficiently?

Probabilistic Semantic Human Sensor Models

Given a finite dictionary, prepositional phrases like "next to the robot", "nearby the robot", and "far from the robot" can be viewed as noisy discrete categorizations of continuous space by the human, so that s_h corresponds to a discrete multinomial random variable conditioned on continuous random variable x. As discussed in (Ahmed, et al. 2013), this implies that $p(s_h|x,m_t)$ can generally be modeled via probabilistic discriminative classifier models used for statistical pattern recognition (Bishop, 2006). Hence, given a predefined codebooks over s_h, parameters

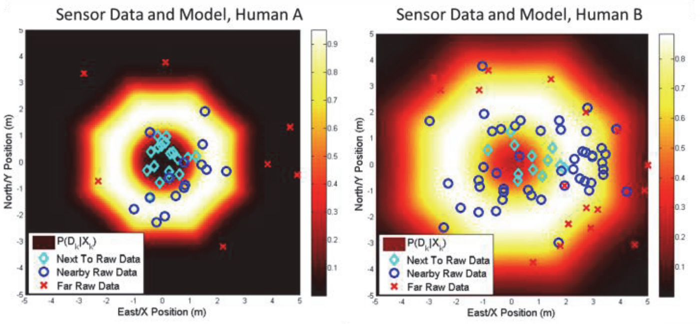

***Figure 4.** Labeled training data and maximum likelihood MMS models for two different human subjects providing semantic spatial range observations, e.g. "The target is nearby the robot."*

for $p(s_h|x,m_t)$ can be learned from labeled ground truth training data to "calibrate" semantic human sensors.

One particularly convenient parameterization for $p(s_h|x,m_t)$ is the multimodal softmax (MMS) model, which generalizes the well-known statistical softmax model and handles non-convex spatial class labels through the use of piecewise linear log-odds boundaries. As Figure 4 shows, MMS model parameters w can be rigorously estimated offline from training data using either maximum likelihood or approximate Bayesian methods, which can return either point estimates or estimated posterior probability distributions for w (Ahmed and Campbell, 2011).

Online Hybrid Bayesian Inference

Given s_h and an MMS model for $p(s_h|x,m_t)$, Bayes' rule can be applied to $p(x|z_r, q_r)$ to give the new posterior pdf

$$p(x|z_r, q_r, s_h, m_t) \propto p(x|z_r, q_r)p(s_h|x, m_t),$$

which represents an analytically intractable hybrid inference problem that must be approximated (Lerner, 2003). In (Ahmed, et al., 2013), we showed that fast and accurate finite Gaussian mixture (GM) model approximations of $p(x|z_r,\ q_{r,}s_h,\ m_t)$ can be obtained through a combination of variational Bayesian inference and Monte Carlo importance sampling methods whenever $p(x)$ and $p(x|z_r,\ q_r)$ are also given by GMs, which are widely used to model arbitrarily complex probability densities. Our resulting variational Bayesian importance sampling (VBIS) inference algorithm was shown to be much more robust to "bad" prior information than conventional Monte Carlo particle filtering methods based on likelihood weighted inference (Arulampalam, et al. 2002; Shachter and Peot, 1989) and leads to a recursive parallelizable "Kalman-like" filter for nonlinear/non-Gaussian state space systems that listens to robot and human sensor data at the same time.

Data Association Ambiguities

The inherent target assignment ambiguities of "Something is…" statements can be handled via the conservative probabilistic error modeling strategy discussed in (Maskell, 2008). This effectively dilutes the informativeness of $p(x|z_r,\ q_{r,}s_h,\ m_t)$ by marginalizing over the uncertain error hypotheses e_h that s_h either does (e_h=0) or does not (e_h=1) "belong" to a particular target. This approach is computationally cheap and conveniently leads to GM pdfs, but requires an estimate of $p(e_h)$.

In our studies, we assumed $p(e_{h)}$ was either "objectively" given by either a uniform distribution (justifiable via the maximum entropy principle) or subjectively set by the operator using a slider bar on his/her console to indicate a "confidence estimate", e.g. "I am 80% certain that something is behind the wall," so that $p(e_h)$ = 1 – 0.8 = 0.2. Although not shown here, e_h can also be modeled as another parent node to s_h in Fig. 3.

Experimental Validation and Results

Our proposed semantic human sensor fusion approach was validated in two separate experimental studies. In (Ahmed, et al. 2013), we studied search performance with a single human operator as a function of the following sensing modalities, using both diffuse and bad priors $p(x)$: robot-only (i.e. no human sensor input baseline), human-only (no robot sensor input), and human-with-robot. Both human-only and human-with-robot fusion led to significant improvements in terms of the number of targets found, the time to find targets, and target localization accuracy improved significantly with respect to baseline search conditions, for both diffuse and bad priors. While positive semantic data were useful for correcting missed target detection events, negative semantic information data were mainly useful for reducing the overall distance traveled by the robot, since the human could remove large probability masses from areas that were clearly free of targets early on. The results also showed that more s_h data were generated in human-only search trials than in human-with-robot trials, in order to compensate for the inherently limited and coarse localization precision afforded by the finite semantic codebook. This effect (and the increased cognitive load on the operator) was greatly mitigated by fusion of finer grained z_r data in human-with-robot trials.

In the experiments of (Sample, et al. 2012), we studied search performance with 16 different volunteer human operators to address the following questions: (i) is it necessary to train individual MMS models of semantic observations for a particular human operator, or do generic "one size fits all" models suffice? (ii) is there a noticeable performance difference between using the operator's own subjective confidence estimates for $p(e_h)$ instead of a fixed "objective" estimate? Each volunteer operator completed 30 minute training data collection sessions and then conducted four search missions according to a randomized 2x2 experimental design with the following axes: individual/generic MMS model (i.e. parameters estimated using individual or pooled training data) and confidence/no confidence bar (i.e. operator sets $p(e_h)$ or maximum entropy probability used).

Our results suggest that, provided operators are all trained to use the same semantic dictionary: (i) generic operator sensor models were just as useful as individual models, and (ii) search performance was generally better with subjective confidence measures, since operators could more easily convey their observations with fewer messages. As shown in Figure 5, the results also showed that targets beyond the robot's visual detection horizon were often localized to within 0.5 meters of their true locations in all test conditions using semantic human sensor data, a noticeable improvement over the baseline search results without human input.

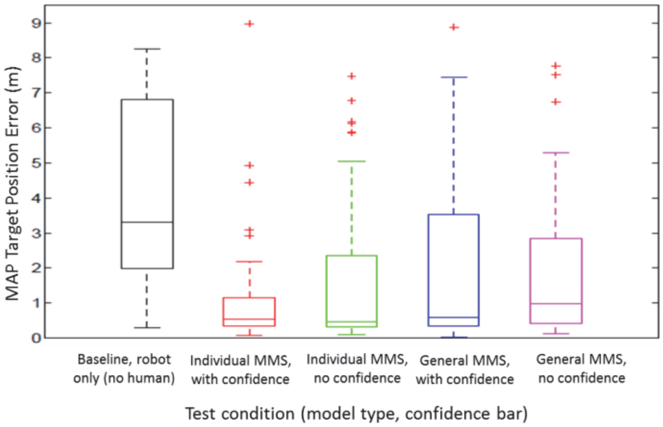

Figure 5. *MAP target position errors across different test conditions for 16 human operators.*

Ongoing and Future Work

We are currently working on several extensions of the methods from our recent work to tackle more generalized versions of the target search problem. Some particularly interesting variations that lead to challenging inference problems include using uncertain map reference points m_t (for maps built by robots in unknown environments), uncertain w (for coping with sparse human operator training data and/or online calibration), and/or hierarchical distributions over $p(e_h)$. We are exploring inference methods based on Rao-Blackwellization and sparse Gauss-Hermite quadrature to address these challenges.

We also are working on extending our methods to other cooperative human-robot sensing problems like dynamic target tracking of extended rigid objects, where semantic human inputs could be used to improve estimates of positions, attitudes and velocities for objects beyond robotic sensor ranges. These could also be coupled with direct semantic human observations of certain abstract discrete Markov-switching states used in hybrid dynamical system models. We have also begun to look at how human inputs can be fused with robotic lidar and vision data in Markov Random Field models to improve probabilistic terrain estimates (Tse, et al., 2012).

We have also started to explore new ways of sharing information between humans and robots, both in individual one-on-one teams and in large scale distributed networks. Although not presented here due to limited space, we have conducted human-only target search experiments using tablet smartphone-based sketch interfaces for probabilistic target localization. These sketch interfaces allow human operators to “draw” probabilistic target information directly onto a search map, thus overcoming some of the difficulties encountered with training and implementing semantic sensor interfaces with finite codebooks. We are also studying how different decentralized planning and distributed data fusion strategies can be exploited for efficient large scale applications involving ad hoc networks of humans and robots. A recent collaboration in (Ponda, et al., 2011) for a single human/multi-robot version of the target search problem showed that information-based planners (iRRT and consensus based bundle adjustment) can benefit greatly from human perceptual involvement. However, it was found that the rates at which humans provide data must be balanced against the rate at which autonomous agents decided to replan and reallocate tasks, in order to avoid undesirable “chattering” behavior, in which robots constantly replan at their maximum rate in response to new human information, which in turn is provided at a maximum rate in response to the robot’s replanning behavior. This has many interesting implications for developing intelligent planners for both single and multi-robot systems that are based on “active sensing” with human operators, whereby metrics of uncertainty with respect to states of interest (e.g. entropy) can be used as a means for dynamically assigning or suppressing different human sensors. Related ideas were explored in (Tellex, et al. 2012) and (Kaupp and Makarenko, 2009) in the context of active polling of human agents for uncertain robotic decision-aiding in path planning and navigation tasks, but these works did not consider the possible role of the human as a sensor and did not consider intelligent buffering, suppression, or rejection of human inputs to avoid unnecessary replanning. Thus, the manner in which robots both present perception-related questions to human sensors and handle data provided by human sensors present interesting avenues for future work.

References

Sheridan, T. 2002. *Humans and Automation*, Santa Monica, CA: Wiley.

Fong, T., Thorpe, C. and Baur, C. 2003. “Robot, Asker of Questions,” *Robotics and Autonomous Systems*, v. 42, pp. 235-243.

Bauer, A., Klasing, K., Lidoris, G., Muhlbauer, Q, Rohrmuller, F., Sosnowski, S., and Buss, M. 2002. "The Autonomous City Explorer: Towards Natural Human-robot Interaction in Urban Environments." *Intl. J. of Social Robotics*, v.1, no. 2, pp. 127-140.

Huang, A., Tellex, S., Bachrach, A., Kollar, T., Roy, D. and Roy, N. 2010. "Natural Language Command of an Autonomous Micro-Air Vehicle," in *IEEE/RSJ International Conference on Intelligent Robots and Systems, 2010 (IROS 2010)*, pp. 2663-2669.

Tellex, S., Kollar, T., Dickerson, S., Walter, M., Banerjee, A. Teller, S. and Roy, N. 2011. “Understanding Natural Language Commands for Robotic Navigation and Mobile Manipulation.” *Proc. of the Natl. Conf. on Artificial Intelligence (AAAI)*, San Francisco, CA.

Shah, D., Schneider, J. and Campbell, M. 2012. "A Sketch Interface for Robust and Natural Robot Control," *Proc. of the IEEE*, vol.100, no.3, pp.604-622

Lewis, M., Wang, H., Velgapudi, P., Scerri, P., and Sycara, K. 2009. "Using humans as sensors in robotic search," in *12th Intl. Conf. on Information Fusion*, Seattle, WA.

Kaupp, T., Douillard, B., Ramos, F., Makarenko, A. and Upcroft, B. 2007. "Shared Environment Representation for a Human-Robot Team Performing Information Fusion," *J. of Field Robotics*, vol. 24, pp. 911-942.

Bourgault, F., Chokshi, A., Wang, J., Shah, D., Schoenberg, J., Iyer, R., and Campbell, M. 2008. Scalable Bayesian human-robot cooperation in mobile sensor networks. in *IEEE/RSJ International Conference on Intelligent Robots and Systems, 2008 (IROS 2008),* pp. 2342-2349.

Hall, D. and Jordan, J. 2010. *Human-centered Information Fusion*, Artech House.

Goodrich, M. A., Morse, B. S., Engh, C., Cooper, J. L., and Adams, J. A. 2009. "Towards using Unmanned Aerial Vehicles (UAVs) in Wilderness Search and Rescue: Lessons from field trials." *Interaction Studies*, v. 10, no. 3, pp. 453-478

Kruijff, G. J. M., Colas, F., Svoboda, T., van Diggelen, J., Balmer, P., Pirri, F., and Worst, R. 2012. Designing Intelligent Robots for Human-Robot Teaming in Urban Search & Rescue. *In 2012 AAAI Spring Symposium Series: Designing Intelligent Robots: Reintegrating AI.*

Ahmed, N., Sample, E., and Campbell, M. 2013. "Bayesian Multicategorical Soft Data Fusion for Human–Robot Collaboration," *IEEE Trans. on Robotics*, v.29, no.1, pp.189-206.

Sample, E., Ahmed, N., and Campbell, M. 2012. "An experimental evaluation of Bayesian soft human sensor fusion in robotic systems," in *Proc.of AIAA GNC 2012*, Minneapolis, MN.

Koenig, S., and Likhachev, M. 1999. "D* Lite." in *Proceedings of the National Conference on Artificial Intelligence*. Menlo Park, CA: AAAI Press.

Bishop, C.M. 2006. Pattern.recognition and machine learning. Vol. 1. New York: Springer.

Ahmed, N., and Campbell, M. 2011. Variational Bayesian Learning of Probabilistic Discriminative Models with Latent Softmax Variables, *IEEE Trans.on Signal Proc.*, v.59, no. 7, pp. 3143-3154.

Lerner, U. 2002. Hybrid Bayesian Networks for Reasoning About Complex Systems. PhD Dissertation, Computer Science Department, Stanford University, Stanford, CA.

Arulampalam, M. S., Maskell, S., Gordon, N., and Clapp, T. 2002. "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking". *IEEE Trans. on Signal Proc.*, v. 50, no. 2, 174-188.

Shachter, R. D., and Peot, M. A. 1989. "Simulation approaches to general probabilistic inference on belief networks." in *Uncertainty in Artificial Intelligence*, vol. 5, pp. 221-231.

Maskell, S. 2008. "A Bayesian approach to fusing uncertain, imprecise and conflicting information". *Information Fusion*, v. 9, no. 2, pp. 259-277.

Tse, R., Ahmed, N. and Campbell, N. 2012."Unified mixture-model based terrain estimation with Markov Random Fields." in *IEEE Conference on Multisensor Fusion and Integration for Intelligent System 2012*, pp. 238-243.

Ponda, S., Ahmed, N., Luders, B., Sample, E., Hoossainy, T., Shah, D., and How, J. P. 2011. "Decentralized Information-Rich Planning and Hybrid Sensor Fusion for Uncertainty Reduction in Human-Robot Missions," in *Proc. of AIAA GNC 2011*, Portland, OR.

Tellex, S., Thaker, P., Deits, R., Kollar, T., & Roy, N. 2012. Toward information theoretic human-robot dialog. Proceedings of Robotics: Science and Systems, Sydney, Australia.

Kaupp, T., & Makarenko, A. 2008. Decision-theoretic human-robot communication. in *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction*, pp. 89-96.