Invited Paper

Collaborative Autonomous Sensing with Bayesians in the Loop

Nisar Ahmed

Aerospace Engineering Sciences, University of Colorado Boulder, Boulder, CO, USA

ABSTRACT

There is a strong push to develop intelligent unmanned autonomy that complements human reasoning for applications as diverse as wilderness search and rescue, military surveillance, and robotic space exploration. More than just replacing humans for 'dull, dirty and dangerous' work, autonomous agents are expected to cope with a whole host of uncertainties while working closely together with humans in new situations. The robotics revolution firmly established the primacy of Bayesian algorithms for tackling challenging perception, learning and decision-making problems. Since the next frontier of autonomy demands the ability to gather information across stretches of time and space that are beyond the reach of a single autonomous agent, the next generation of Bayesian algorithms must capitalize on opportunities to draw upon the sensing and perception abilities of humans-in/on-the-loop. This work summarizes our recent research toward harnessing 'human sensors' for information gathering tasks. The basic idea behind is to allow human end users (i.e. non-experts in robotics, statistics, machine learning, etc.) to directly 'talk to' the information fusion engine and perceptual processes aboard any autonomous agent. Our approach is grounded in rigorous Bayesian modeling and fusion of flexible semantic information derived from user-friendly interfaces, such as natural language chat and locative hand-drawn sketches. This naturally enables 'plug and play' human sensing with existing probabilistic algorithms for planning and perception, and has been successfully demonstrated with human-robot teams in target localization applications.

Keywords: Human-machine systems, collaborative sensing, Bayesian methods, machine learning, sensor fusion, estimation, robotics

1. INTRODUCTION

From self-driving cars[1] and storm-chasing aircraft,[2] to robots exploring icy Jovian moons[3] and patrolling large swaths of open space,[4] the future of unmanned autonomous systems looks very promising. Thanks to their ability to gather, process, share, and act on vast amounts of information, autonomous systems are transforming how society thinks about many complex activities that were once considered beyond machine reasoning. Alongside improvements in computing, communication, and sensing hardware, a key technological factor in this development has been the accelerated sophistication of perception, learning and decision making algorithms that can nearly match (or, in some cases, clearly surpass) human reasoning.[5] Machines are no longer confined to routine automation tasks focused on low-level control and signal processing – they are being given license to make sense of the larger world and make decisions on their own.

However, many hurdles stand in the way of 'set and forget' autonomy. Autonomous systems are products of imperfect human engineering, and thus will never operate perfectly out of the box, i.e. knowing everything they will ever need to know to behave exactly as needed. With the widespread adoption of non-deterministic perception, learning, and planning algorithms to cover these gaps, there is still no way yet to guarantee that autonomous systems will behave as intended in all circumstances.[6,7] The space exploration domain highlights these challenges: how should an autonomous robot explorer reason about what to do and expect on an icy Jovian moon, if it is gathering detailed data about conditions there for the very first time? Prior knowledge from coarse remote sensing data, human expertise, etc. will be vital to designing the autonomy beforehand, but are not enough to ensure that the robot will be completely self-sufficient. The robot will encounter unexpected situations during the mission that require reasoning beyond its designed capabilities, and which will be impractical to handle via the traditional means of teleoperation or extremely detailed planning from the ground.[8]

Further author information: send e-mail correspondence to: Nisar.Ahmed@colorado.edu

Unmanned/Unattended Sensors and Sensor Networks XII, edited by Edward M. Carapezza, Panos G. Datskos, Christos Tsamis, Proc. of SPIE Vol. 9986, 99860B · © 2016 SPIE
CCC code: 0277-786X/16/$18 · doi: 10.1117/12.2246705

Hence, the old adage 'no man is an island' applies to unmanned autonomy as well. Intelligent autonomy should not just be defined by the ability to gather, process, and act on information completely on its own. Rather, it should also include the ability to seek out and exploit other autonomous agents (including humans) for help when needed. This view naturally follows from the under-appreciated fact that 'autonomy' represents a *relationship*, in which a machine is *delegated by a user* to perform certain tasks.[9] As such, it should be kept in mind that an autonomous system (or, any intelligent reasoning machine) represents a deliberate complementary extension of human reasoning – not merely a wholesale replacement of it.

Human-machine interaction should therefore be considered an essential component of unmanned autonomy, alongside perception, planning, learning, etc. Autonomy should enable stakeholders and users (soldiers, pilots, scientists, astronauts, farmers, etc.) to stay in/on the loop to delegate, assess and help improve operations, while also keeping them at a safe distance from dull, dirty and dangerous tasks (especially ones they cannot perform well). This in turn raises issues of managing *trust* for human-machine interaction. If machines and humans are to trust (i.e. willingly depend on) each other, then each must be able to communicate and form useful mental models of the other's abilities, goals, percepts and actions.[10]

Yet, effective human-machine interaction is difficult to realize in practice, and is often only considered after systems are already designed. Indeed, human-machine interaction is sometimes viewed as a 'necessary evil': a post hoc band aid for corner cases where planning and perception haven't caught up yet. Such thinking opens the door to poorly designed human-machine interfaces, which can lead to many unintended (yet avoidable) consequences such as loss of situational awareness, user distrust, system misuse and abuse.[11] This also prematurely shuts out novel pathways to exploiting collaborative human and machine reasoning from the outset. Sophisticated strategies for integrated human-autonomy interaction have begun to develop along these lines. For instance, there is much research nowadays on human-assisted robot planning using multi-modal commands, e.g. natural language speech, sketches or physical gestures.[12–16] However, there are also many important implications for reliable sensing, data fusion and perception, which are still major choke points for unmanned systems.[17]

This work summarizes our recent and ongoing research toward harnessing 'human sensors' for autonomous information gathering tasks. The basic idea behind is to allow human end users of autonomous systems (i.e. non-experts in robotics, statistics, machine learning, etc.) to directly 'talk to' the information fusion engine and perceptual processes aboard any autonomous agent. Our approach is grounded in rigorous Bayesian modeling and fusion of flexible semantic information derived from user-friendly interfaces, such as natural language chat and locative hand-drawn sketches. This naturally enables 'plug and play' human sensing with existing probabilistic planning and perception algorithms; that is, human sensors can freely provide information to autonomy without undermining its ability to reason, or forcing undesirable dependencies on human inputs. We have successfully demonstrated our fusion methods and interfaces with real human-robot teams in target localization applications. Section 2 provides some background for probabilistic modeling and reasoning. Sections 3 and 4, respectively, describe our work on Bayesian fusion of soft data provided via semantic natural language and locative hand drawn sketching for target search.

2. PROBABILISTIC AND BAYESIAN REASONING FOR AUTONOMY

Perception and decision making architectures for unmanned autonomous systems can be described in terms of the classical closed-loop 'observe-orient-decide-act' (OODA) process.[18] The 'decide' and 'act' portions (planning and control) are typically designed to maximize some set of performance optimality criteria, whereas the 'observe' and 'orient' portions (sensing and perception) are designed to extract maximum information from environmental signals to support decision making and execution. Modern state space and optimal control theory underscores the importance of accounting for both *model and state uncertainties* in such closed-loop processes. Such uncertainties directly govern how an agent would best decide to gather more information (acting to observe) and what information it should prioritize gathering (observing for action).

Due to their highly flexible nature and ease of use for describing stochastic uncertainties, probabilistic models have been adopted as the lingua franca in modern robotics and autonomy.[19] The development of probabilistic graphical model (PGM) theory in particular provides a powerful unified formal framework for combining these techniques in a scalable way.[20] PGMs enable efficient embedding and reasoning over complex probabilistic

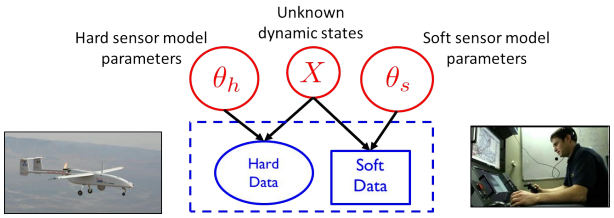


Figure 1. Generic PGM for soft-hard sensor fusion problem in a human-autonomous robot team.

dependencies, and thus make it trivial to model high-dimensional joint probability distributions. PGMs can also integrate deterministic information and constraints, e.g. based on causal and logical reasoning, and permit hierarchical reasoning on uncertain model structures.

With these properties, PGMs make it relatively easy to perform *Bayesian inference* to update probabilistic knowledge in light of new information. The basic premise of Bayesian inference is simple: given some unknown set of random variables X and some observed data in random variables Y that obey some joint probability distribution $P(X, Y)$, Bayes' rule computes an updated posterior probability distribution $P(X|Y)$ that reflects how much the probability of obtaining *any possible value* of X is affected by the evidence Y. Large 'forward models' for autonomous perception and planning problems can be represented by PGMs that decompose $P(X, Y)$ into an easily obtainable set of prior distribution $P(X)$ and likelihood (or evidence) functions $P(Y|X)$. This decomposition greatly simplifies the operations involved in Bayesian inference to find $P(X|Y)$. Then $P(X|Y)$ (the 'inverse' of the joint $P(X)$ and $P(Y|X)$ forward model) can be subsequently analyzed to get a single 'best' estimate value $\hat{X}$, e.g. as done sequentially in the Kalman filter, or in batch form for maximum a posteriori (MAP) estimates in robotic pose graphs.[21] The posterior $P(X|Y)$ can also passed directly to some other reasoning algorithm for decision making and planning, e.g. a policy function for a probabilistic planner.[19]

2.1 Bayesian Soft-Hard Data Fusion

Probabilistic models and Bayesian reasoning provide a powerful general framework for augmenting robotic perception systems with 'human sensors', which can provide soft data to complement 'hard data' from conventional sensors such as lidar, cameras, sonars, etc. in partially observable environments. 'Soft data' is any set of observations that originates from human sources.[22] For instance, human pilots and payload specialists in wilderness search and rescue (WiSAR) missions can interpret video feeds and electro-optical/IR data streams provided by small fixed-wing UAVs, and can spot important clues that help narrow down probable lost victim locations and movements.[23] Likewise, in large-scale surveillance for defense applications, dismounted soldiers can provide evidence on the whereabouts and behaviors of potential intruders moving across unsecure areas; it is desirable to directly fuse such soft data with hard data from UAV patrols to improve intruder detection and tracking performance. Combined hard-soft sensing can also help cope with design limitations in UXV systems, where autonomous vehicles are subject to hardware constraints that restrict onboard sensing, processing and communication abilities. Soft data integration also lets humans stay 'in the loop' without overloading them with cognitively demanding planning/navigation tasks.[24]

A key problem then is: how should soft sensor data be formally integrated with hard data to augment robotic estimation and perception algorithms? Figure 1 shows the corresponding PGM for the generic human-robot sensor fusion problem, where sensor model parameters Θ for both hard and soft sensors may also be unknown, and thus may have to be estimated along with the state of interest X. Soft data can be broadly related to either 'abstract' phenomena that cannot be measured by robotic sensors (e.g. labels for object categories and behaviors) or measurable dynamical physical states that must be monitored constantly (object position, velocity, attitude, temperature, size, mass, etc.)[22] Our work focuses on the latter, under the key assumption that *humans are* ***not*** *oracles*: as with any other sensor data, human observations are subject to errors, limitations and ambiguities that must be modeled properly. We aim to adapt widely used statistical sensor fusion and robotic state estimation algorithms, e.g. Bayes filters and the like, so that soft data can be exploited with minimal effort on the part of the robot or the human sensor.

3. BAYESIAN FUSION OF SOFT LANGUAGE OBSERVATIONS

Refs. 25–27 were among the first to develop Bayesian fusion techniques allowing human sensors to directly 'plug into' robotic state estimation and perception algorithms. However, these works assume that humans report data the same way robots do, and thus greatly limit the flexibility of human-robot communication. In the context of target tracking with extended Kalman filters, for instance, ref. 25 assumes that humans provide numerical range and bearing measurement reports ('The target is at range 10 m, bearing 45 degrees').

Ref. 28 showed how to model and fuse flexible semantic natural language soft data to provide a broad range of positive/negative information for Bayesian state estimation, e.g. 'The target is parked near the tree in front of you', 'Nothing is next to the truck heading North'. One nice theoretical property of the resulting fusion algorithm is its ability to directly plug into Gaussian mixture (GM) filters for state estimation. GM filters can accurately represent complex posterior pdfs, while avoiding the curse of dimensionality encountered by grid or particle filter methods.[25,29,30] We briefly summarize the cooperative human-robot state estimation method developed in ref. 28 and discuss several recent extensions to address the issues of semantic likelihood sensor modeling, natural language processing, and optimal querying for active semantic sensing.

3.1 Bayesian Fusion of Semantic Data

Let X be a continuous random vector representing the dynamic state of interest (e.g. target location, velocity, heading) with prior pdf $p(X)$ (which may already be conditioned on hard data), and D be a discrete random variable representing a human-generated semantic observation related to X (e.g. 'The target is on the bridge', 'The target is heading over the bridge and slowing down', etc.). Given the likelihood function $P(D|X)$, Bayes' rule gives the posterior pdf

$$p(X|D) = \frac{p(X)P(D|X)}{\int p(X)p(D|X)dx}. \tag{1}$$

where $P(D|X)$ models the human's 'semantic classification error' as a function of X. If $D = l$ corresponds to one of m exclusive semantic categories for a known dictionary, then a softmax function can be used to model $P(D = l|X)$,

$$P(D|X) = \frac{\exp(w_l^T x + b_l)}{\sum_{c=1}^{m} \exp(w_c^T x + b_c)}. \tag{2}$$

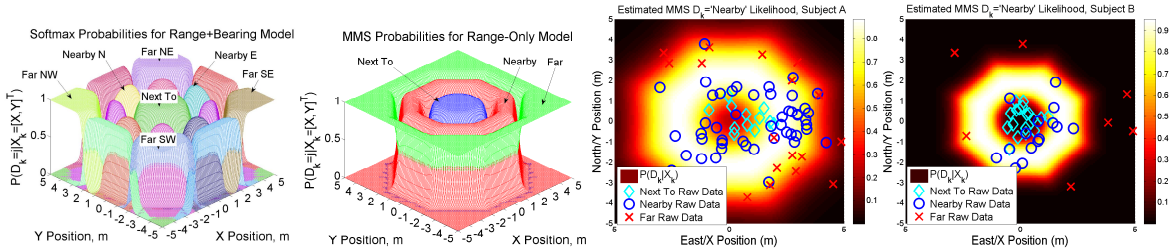
Fig. 2 (a) shows an example softmax model for semantic spatial range and bearing observations in 2D. An important feature of this likelihood model is that, for a given parameter set of class weights and biases $\Theta = \{w_c, b_c\}_{c=1}^{m}$, the state space X is divided into m convex 'ϵ-probability polytopes', i.e. convex subsets in X where certain semantic categories occur with a probability $P(D|X) \geq \epsilon$.[31] Label classes in a softmax model can also be internally grouped together as 'subclasses' within larger semantic classes. This yields the generalized multimodal softmax (MMS) likelihood model,[32] which can represent arbitrary non-convex probabilistic polytopes in X via piecewise-convex subclasses. For instance, as shown in Fig. 2 (b), a range-only semantic MMS model can be obtained from Fig. 2 (a) by grouping together range labels. The parameters Θ of both softmax and MMS models can be learned from human-generated calibration data using standard parameter estimation techniques, as shown in Fig. 2 (c)-(d) for two different human sensors.

To perform recursive Bayesian data fusion with softmax or MMS likelihoods, eq. (1) must be approximated, since the exact posterior pdf $p(X|D)$ cannot be obtained in closed-form for any prior $p(X)$. Ref. 28 showed that, if $P(D = i|X)$ is generally given by an MMS model with q_i subclasses for observation label i and the prior is given by a finite Gaussian mixture (GM) with m_p prior components,

$$p(X) = \sum_{p=1}^{m_p} w_p \mathcal{N}_X(\mu_p, \Sigma_p)$$

(where w_p, $\mu_p \in \mathbb{R}^n$, and $\Sigma_p \in \mathbb{R}^{n \times n}$ are the weights, mean vector and covariance matrix for mixand p), then $p(X|D = i)$ can be well-approximated by a $q_i m_p$ component GM,

$$p(X|D = i) \approx \sum_{q=1}^{q_i m_p} w_q \mathcal{N}_X(\mu_q, \Sigma_q). \tag{3}$$



(a) (b) (c) (d)

Figure 2. (a) Probability surfaces for example softmax likelihood model, where class labels take on a discrete range in 'Next To','Nearby','Far From' and a canonical bearing 'N','NE','E','SE',...,'NW'; (b) MMS model for semantic range derived from softmax model in (a), using 1 subclass for 'next to', 8 subclasses for 'nearby' and 8 subclasses for 'far'; (c)-(d) 'Nearby' label probabilities in estimated MMS range-only models for two different human sensors.

The weights, means and covariances of posterior component q can be determined by fast numerical quadrature techniques such as likelihood weighted importance sampling (LWIS) or variational Bayesian importance sampling (VBIS).[28] These techniques exploit the fact that the exact product of an MMS likelihood function and GM prior is a mixture of non-Gaussian component pdfs, where each component is guaranteed to be unimodal and thus can be well-approximated by a moment-matched Gaussian. To manage the growth of mixture terms from m_p to $q_i m_p$, mixture compression methods such as Runnalls' joining algorithm[33] can used to find a GM with $m_f < q_i m_p$ components that minimizes an information loss with respect to (3). This allows semantic human sensor data to plug seamlessly into existing GM Bayes filters for hard robot sensor data fusion.[25,30]

Figure 3 illustrates the semantic data fusion process for an indoor target localization application, discussed in refs. 28, 34 and 35. In this example, a ground robot and a remote human supervisor perform a time critical search for static targets (red traffic cones) using a Bayesian search algorithm with GM prior distributions on the target locations. The robot autonomously plans optimal search paths using negative information gathered from an onboard visual detector, which has a very limited range and field of view. The human supervisor can confirm possible target detections (confirming true detections or false alarms), but can also voluntarily provide semantic observations about the environment based on a live (grainy and delayed) video feed. In this application, the human's semantic observations are selected from a pre-defined dictionary, which filled in templated observations of the form '[Something/Nothing] is [preposition][reference object]'.

As shown in the bottom right of Fig. 3, soft semantic data fusion leads to a massive injection of information and hence a significant shift in the robot's beliefs about the target locations. This allows the robot to plan and execute more efficient search paths. The soft semantic data fusion process does not require the human to engage in cognitively demanding planning or control activities: the robot simply responds to new human sensor information by updating its beliefs about the target state and executing corresponding optimal search actions. Experiments with 16 different human participants acting as supervisors[35] showed that targets well outside the robot's field of view could be localized to sub-meter accuracy, using only semantic data fusion updates provided by the human (which included both positive and positive observations, i.e. 'Something is in front of the door' vs. 'Nothing is in front of the door'). In these experiments, human supervisors were also allowed to view a live heat map representing the robot's beliefs about the target locations. As such, soft data fusion can be particularly useful as a 'coarse sensing' input for both priming and correcting robotic perception in complex dynamic settings. For example, many human supervisors were able to use soft data to correctively/preepmtively 'nudge' the robot's beliefs whenever its detector encountered false negatives, thus preventing the robot from wandering away from the true target locations. The fusion of negative semantic information (e.g. 'There is nothing in this room') also allowed the robot to quickly narrow down search regions, thus saving considerable time and energy.

While promising, these initial results were obtained with significant design constraints on the semantic data fusion system and human-machine interface. For instance, a highly restrictive semantic dictionary and set of softmax/MMS likelihoods were used to model human sensor semantics in order to avoid the up front difficulties of natural language processing. This approach also assumes that it is possible to construct the entire dictionary

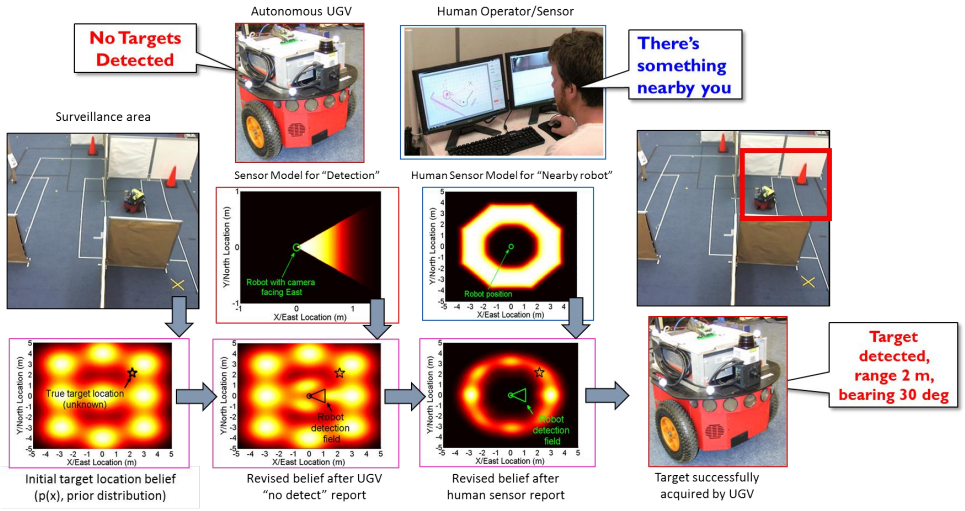


Figure 3. Static target localization application from refs. 28, 35. The surveillance area consist of mapped obstacles and landmarks, and targets (orange traffice cones) with unknown locations X given by a Gaussian mixture (GM) prior $p(X)$. An autonomous UGV robot performs Bayesian fusion of prior beliefs with detection/no detection reports from a vision sensor and semantic observations provided by a human supervisor, producing updated GM posteriors for X. The updated GMs are then used by autonomous path planning algorithms to navigate towards likely target positions, i.e. the human only provides sensor observations, and never directly commands or steers the robot.

and set of semantics ahead of time, which is prohibitive for many applications (e.g. mapping and tracking targets in unknown environments). Furthermore, a strictly passive data fusion process was used: the human supervisor would only voluntarily provide inputs if he/she deemed them necessary. The following subsections describe recent advances that address these issues.

3.2 Semantic Likelihood Synthesis and Compression

Softmax and MMS likelihood models are theoretically convenient for recursive data fusion via (3), but it is not immediately obvious how to physically interpret or manipulate these sensor models in different settings. Refs. 36 and 37 established several key properties of general softmax and MMS model geometry to address this issue, and provide the basis for novel solutions to two related problems: semantic likelihood model synthesis, and batch semantic likelihood compression.

Likelihood model synthesis: How should softmax/MMS models be constructed 'from scratch' and/or dynamically modified in general state spaces X? For instance, we may wish to model likelihood functions in complex 2D/3D spatial domains and beyond to include velocities, angles, etc. Or we may wish to build likelihood models 'on the fly' to exploit spatial semantics for newly perceived objects or environments that are mapped in real time. It is thus practically necessary to translate known geometric constraints in X directly into prior specifications/constraints for Θ. This is especially important for learning with sparse human sensor calibration data, and for adapting Θ online to enforce expected geometric properties of likelihood function polytopes in X space. For instance, meaningful spatial invariances can be enforced at different scales and for different reference geometries, e.g. 'near the table' vs. 'near the house'. In Fig. 2 (a), Θ was obtained via maximum liklihood optimization on manually generated 'prototypical' training data, which was tuned to produce the desired polytope geometries via trial and error. This brute force approach is computationally expensive: it requires non-convex optimization, and is not suitable for higher dimensional settings.

To solve these problems, we can recall and exploit the important fact that the softmax function (2) describes a set of probabilistic polytopes in X space. In particular, the set of log-odds functions between classes i and j define the linear hyperplane boundaries of their corresponding polytopes at different relative probability levels; for equal probabilities ϵ, we have

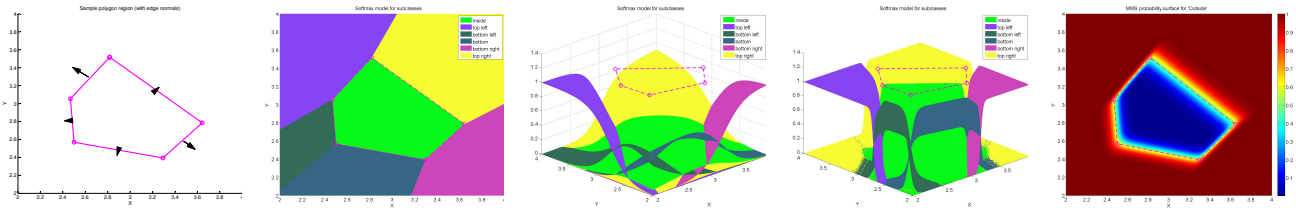
$$\log \frac{P(D=j|X)}{P(D=i|X)} = (w_j - w_i)^T x + (b_j - b_i) = 0. \tag{4}$$

Thus, if we are given *normal vectors* $\vec{\beta}$ that describe a desired set of polytopes that should exist between semantic classes (or constraints/invariants on those class boundaries), it is easy to show that eq. (4) leads to a system of linear difference equations for Θ,

$$M\vec{\theta} = \vec{\beta}, \tag{5}$$

where $\vec{\theta}$ represents a vectorized version of Θ and M is a relative difference operator that encodes the appropriate differencing operations on $\vec{\theta}$ via (4) to produce the neighboring class polytopes defined by $\vec{\beta}$. Hence, the set of Θ that produce a desired convex semantic decomposition of X (as encoded by M and $\vec{\beta}$) form the solution space of (5). Once established, these parameters Θ and their relationships can be further manipulated to alter the likelihood model's embedded probabilistic polytopes as needed. Note that (5) does not require training data, but also does not guarantee that a (unique) solution $\vec{\theta}$ will be found. Even then, (5) allows imposes useful constraints that greatly improve the efficiency of learning Θ from (highly sparse) calibration data.

Fig. 4 shows a simple example of synthesizing a semantic MMS likelihood model that describes the space 'inside' and 'outside' an arbitrary irregular 2D polygon. This specification of 5 polytope face boundaries for $m = 6$ softmax subclasses (5 softmax subclasses describe 'outside', while one describes 'inside') leads to a system of $N_S(n+1) = 15$ linear equations in $m(n+1) = 18$ unknown softmax parameters. Since one set of subclass parameters can always be set the zero vector, assuming $w_i = 0$ and $b_i = 0$ for $i =$'inside', gives 15 unknown softmax parameters in 15 difference equations. Thus, in eq. (5) , $M = I$ (the identity matrix), and so $\vec{\theta} = \vec{\beta}$, i.e. the softmax model parameters for the 5 'outside' subclasses come directly from the corresponding polygon edge specifications in $\vec{\beta}$. In general, such simple solutions will not always be obtained, since $\vec{\beta}$ can represent a more complex set of semantic constraints derived from given maps and natural language processing models that restrict the meaning of certain human observations. The key point here, however, is that such constraints can be embedded into the likelihood and exploited by the fusion process in a mathematically sound way.



(a) (b) (c) (d) (e)

Figure 4. (a) Polytope specification; (b) resulting subclass regions; (c) subclass probability surfaces with unit normals n_{ji}; (d) desired normals magnifed by 80; (e) non-convex likelihood for 'outside'.

Likelihood model compression: The GM fusion approximation assumes that $P(D = i|X)$ captures all information contained within a semantic human observation D. However, D can contain mixed information about different parts of X, e.g. target position and speed, but not heading. $P(D = i|X)$ could be decomposed into more basic likelihoods that model relevant semantic information after D is parsed (e.g. by a natural language processing front-end, as discussed later). For instance, if $D^a = i$ and $D^b = j$ correspond to the observations 'target near building' and 'target moving quickly away from building', then the likelihood for the joint observation $D = ([D^a = i] \wedge [D^b = j])$ can be modeled as the product of the corresponding softmax/MMS models (assuming both are conditionally independent given X),

$$P(D|X) = P(D^a|X)P(D^b|X).$$

For N_o observations D^o, $o \in \{1, ..., N_o\}$, sequential processing via repeated application of the GM fusion and merging approximations could be very expensive and inefficient. Instead, we seek to extend the GM fusion method of[28] to handle the general case of 'batch' semantic measurement updates for fast online data fusion,

$$p(X|D^{1:N_o}) \propto p(X) \prod_{i=1}^{N_o} P(D^o = i_o|X) \propto p(X)L(D^{1:N_o}|X). \tag{6}$$

where a single 'compressed' softmax/MMS likelihood $L(D^{1:N_o}|X)$ captures all information from the product $P(D^{1:N_o}|X) = \prod_{i=1}^{N_o} P(D^o = i_o|X)$, so that GM fusion and merging methods only need to be applied once. We

exploit the fact that a product of N_o softmax/MMS models can be expressed as another softmax/MMS model for N_o conditionally independent semantic measurements,

$$P(D^{1:N_o}|X) = \prod_{i=1}^{N_o} P(D^o = i_o|X) = \frac{e^{w_{\mathcal{I}}^T x + b_{\mathcal{I}}}}{\prod_{l=1}^{N_o} \sum_{c_l=1}^{m_l} e^{w_{c_l}^T x + b_{c_l}}} = \frac{e^{w_{\mathcal{I}}^T x + b_{\mathcal{I}}}}{\sum_{t=1}^{\bar{m}} e^{w_t^T x_k + b_t}} = L(D^{1:N_o}|X) \tag{7}$$

where $\mathcal{I} = \{i_1, i_2, \ldots, i_{N_o}\}$ is the set of N_o class observations taken from the N_o softmax models, where $i_o \in \{1, ..., m_o\}$ and m_o is the number of classes for the oth softmax model. Assuming that the class labels are ordered within each softmax model, the product model parameters are defined as $w_{\mathcal{I}} = \sum_{o=1}^{N_o} w_{i_o}$ and $b_{\mathcal{I}} = \sum_{o=1}^{N_o} b_{i_o}$. Each measurement $i_o \in \mathcal{I}$ comes from a different consituent softmax model of the product. Thus, $P(D^{1:N_o}|X)$ can be exactly described as a single softmax likelihood over $m_1 \times m_2 \times \cdots \times m_n = \bar{m}$ 'product classes', which appear in the denominator. The w_t and b_t terms cover all $\bar{m}$ combinations for the sum of the weights from the product of N_o softmax models.

Eq. (7) is computationally expensive to compute for online fusion for large N_o and $\bar{m}$. As with GM compression algorithms, this motivates approximation of (7) via parameter compression techniques,

$$L(D^{1:N_o}|X) \approx P(\tilde{D} = i^*|X) = \frac{e^{\tilde{w}_{i^*}^T x + \tilde{b}_{i^*}}}{\sum_{c=1}^{m_*} e^{\tilde{w}_c^T x + \tilde{b}_c}} \tag{8}$$

where $m_* \ll \bar{m}$, and the parameters $\tilde{\Theta} = \left\{\tilde{w}_c, \tilde{b}_c\right\}_{c=1}^{m_*}$, are based on some approximation technique. Ref. 37 presents two approximation techniques, geometric compression and neighborhood compression, that are based on the softmax model synthesis approach presented earlier. Geometric compression attempts to extract the relevant information in $L(D_k^{1:N_o}|x_k)$ according to minimal set of log-odds boundaries needed to specify the resulting 'product class' polytope that appears in the numerator of (7). Neighborhood compression extends geometric compression by retaining additional '2nd order' information about the polytopes for the neighboring classes of each observed class $i_o \in \mathcal{I}$ (where the neighboring polytopes of a given class in any softmax model can be determined offline). Both compression methods use linear programming to identify which class polytope boundaries should be kept in (8), and trade speed for accuracy in online GM fusion.

Fig. 5 shows results for compressing a product of three MMS models corresponding to the observation: 'The target is inside the front yard, near the garage and the front porch'. The likelihoods for the single 'inside' and two 'near' observations are shifted/scaled from a base MMS model describing the 'inside' and 'outside' of the gray rectangular region in the upper left of Fig. 5. These results show a great speedup in computation for geometric compression (0.3 secs) over either the exact product or neighborhood fusion method (20 mins and 10 mins for GM fusion, respectively), for a fairly small and acceptable sacrifice in fusion accuracy.

3.3 Natural Language Chat Interfaces

A human could report soft data by composing structured observations from a list of available statements, as shown upper left in Fig. 6. This 'direct selection' interface bypasses the difficulties of natural language processing (NLP), but leads to several major limitations. Firstly, it restricts the human to a rigid pre-defined dictionary and message structure, which may not provide an intuitive or sufficiently rich set of semantics to convey desired information. Secondly, it is very inconvenient to select items one-by-one from a list for structured messaging; especially as the dictionary size m grows, this quickly becomes infeasible and time-consuming enough to render data irrelevant in dynamic settings. Furthermore, this approach does not scale well with environment/problem

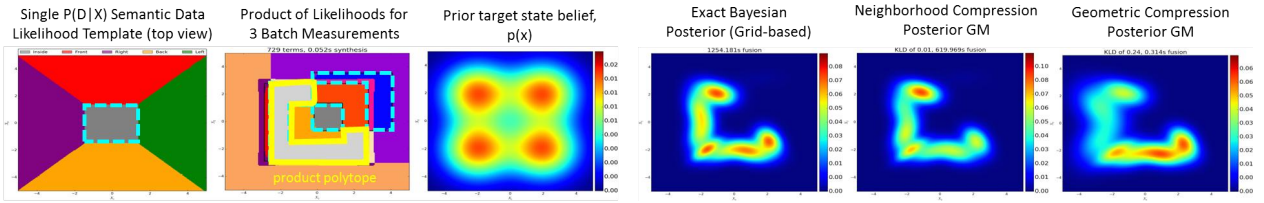


Figure 5. Comparison of MMS likelihood compression methods for three measurements: 'Near the porch', 'Near the garage', and 'Inside the front yard', shown by the grey area in the first plot.

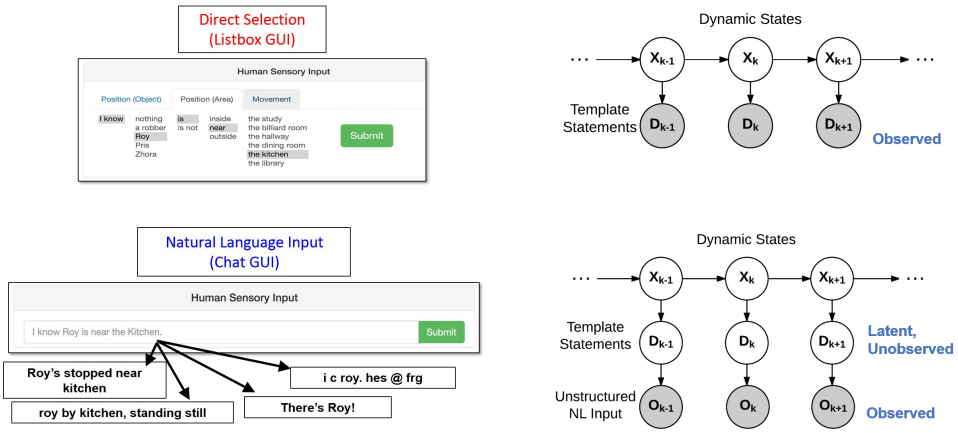


Figure 6. Soft data fusion PGMs for direct selection and natural language chat interfaces in target localization application.

complexity and lexical richness for soft data reporting. In particular, it is often desirable to provide observations that activate multiple semantic D terms, e.g. 'Roy is by the table, heading slowly to the kitchen' simultaneously provides location, orientation and velocity information.

We are developing an unstructured natural language chat interface to support fast and highly flexible 'free-form' soft data reporting. The chat interface should ideally support a wide range of semantics. However, it is highly non-trivial to deriving meaningful and contextually relevant dynamic state information from free-form chat observations O. Unlike structured messages, it is infeasible to explicitly construct likelihood functions $p(O|X)$ in advance to find the Bayes posterior $p(X|O)$. Many different chat messages can also convey similar kinds of information, leading to additional uncertainties in lexical meaning (i.e. possible translation errors) in addition to intrinsic semantic (state spatial meaning) uncertainty. For example, the phrases 'That guy's moving past the books', 'Roy next to the bookcase going to kitchen', and 'He's nearby shelf heading left' all overlap in the sense that they could all to essentially refer to a structured set of atomic phrases: 'Target is near the bookcase; and Target is moving toward the kitchen'. Our approach to handle free-form chat inputs separately accounts for lexical/translation uncertainties (using off-the-shelf NLP components, e.g. probabilistic syntactic parsers and sense-matchers) and semantic/meaning uncertainties (via state filtering) in a statistically consistent way that avoids joint reasoning over a large set of latent variables. The main idea is to translate a given O_k at time k into a reasonable 'on the fly' estimate of the likelihood $p(O_k|X_k)$ via a very large (possibly expandable) dictionary of latent D_k^i semantic observations, which have known generalized softmax likelihoods $p(D_k^i|X_k)$. As illustrated in the bottom of Figure 6, given the expansion

$$p(X_k|O_k) \propto p(X_k) \sum_{i=1}^{m} \sum_{j=1}^{m_i} p(O_k|D_k^i = j) p(D_k^i = j|X_k),$$

(where O_k is conditionally independent of X_k given D_k^i), we can generally approximate $p(O_k|D_k^i)$ as the second summation term on the RHS, where $p(O_k|D_k^i = j)$ accounts for the lexical uncertainty and $p(D_k^i = j|X_k)$ accounts for the semantic uncertainty. Since O_k may also point to multiple soft observations (i.e. target position and velocity), the D_k^i likelihoods on the RHS generally could correspond to unique products of independent dictionary terms, e.g. as in eq. (7). The general problem, then, is to identify how D^i likelihoods (or sets of soft observations) should be 'activated' for a given O input by identifying the scalars $p(O|D^i = j)$. Since $p(X)p(D|X)$ can generally be approximated as a GM, it follows then that the LHS $p(X|O)$ generally leads to a 'mixture of GMs'.

Ref. 38 details one approach for estimating $p(O_k|D_k^i)$ from chat inputs using off-the-shelf NLP tools. We focus here on the problems of *phrase parsing* and *sense matching*, i.e. matching human-generated synonyms and phrases from input chat messages O_k to corresponding words that lead to valid sets of semantic observations D_k in a fixed dictionary. Phrase parsing classifies raw input chat messages O_k into recognizable clusters of word tokens according into a set of predefined Target Description Clauses (TDCs) T_k. TDCs are conceptually similar

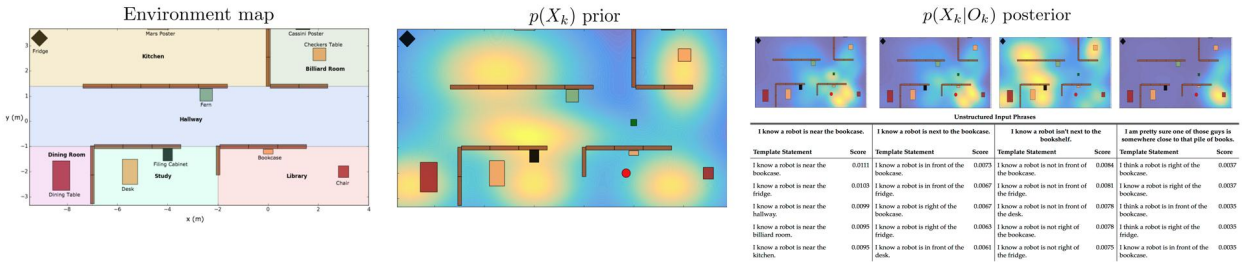


Figure 7. Labeled indoor map, target location GM prior, and resulting GM posteriors for sense matching, with unstructured input phrases and corresponding template dictionary scores.

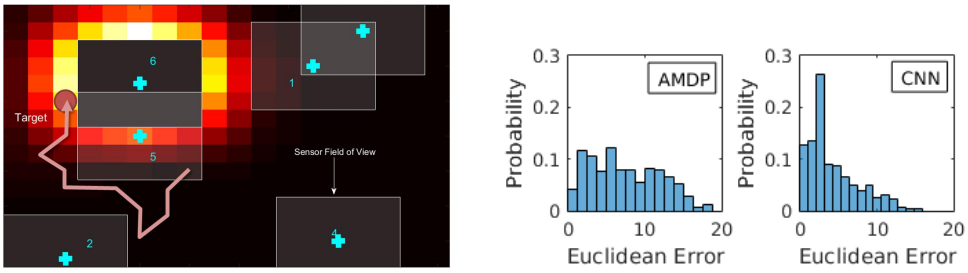
to Spatial Description Clauses (SDCs) used in[39] for natural language control of robots, and allow the fusion engine to construct base templates for acceptable message types in terms of expected parts of speech and topical content. This allows the fusion engine to reject unhelpful semantic observations, e.g. 'It is a nice day outside'. Then, the key issue for word sense matching (for each part of a given TDC) becomes deriving the relationship between the estimated 13 million tokens in the English language and the comparatively minuscule number of tokens in a set of predefined template semantic soft data observations. Recent efforts, particularly by Mikolov et al.,[40] introduced a negative sampling-based approach to efficiently learn multi-dimensional vector representations of all tokens in a vocabulary. Using their `word2vec` tool, we can develop a mapping between a set of tokens contained within an unstructured utterance and a set of tokens contained within a structured semantic template from the dictionary. Conditional probabilities from the parsing and word sense matching tools can be combined to form a score $s(D_k, T_k)$ for each possible latent template statement D_k and parsed TDC T_k.

Figure 7 shows initial proof-of-concept results, using $\arg\max_{D_k} s(D_k, T_k)$ to select template sensor statements that are most similar to the input tokenization for four correctly tokenized test phrases. In this example, 2682 possible template statements are considered, which covers 79.81% of the 208 input sentences collected in a pilot experiment with 12 human participants. From left-to-right, the four columns demonstrate the effects of increasing dissimilarity with template statements: the first input sentence is exactly a sensor statement template; the second input sentences replaces a spatial relation template token, 'near', with a non-template token, 'next to'; the third input sentence replaces the grounding and changes the positivity; and the fourth is an imprecise reformulation of the first sentence. The results are promising, as the top-scoring statements are all qualitatively similar to the original phrase and produce sensible fusion results.

3.4 Optimal Value of Information Querying for Active Soft Data Fusion

We have thus far only considered passive data fusion strategies, where human sensors voluntarily provide observations as they see fit. However, semantic data fusion can also be extended to incorporate *active soft sensing*, i.e. intelligent 'closed-loop' querying of human sensors to gather information that would be most beneficial for complex machine planning and perception tasks. Active sensing problems have a rich tradition in target tracking and controls communities, but have focused on hard data sources such as radar, lidar, cameras, etc. One particularly relevant issue is the *sensor scheduling problem*, which seeks optimal selection of sensing assets given constraints on how many can be tasked to deliver quality data at any given instant. This problem also applies to scheduling of interactions between human sensors and autonomous agents: what is the most valuable soft information to request from humans, and when/how should such soft information be obtained? These issues can be tackled within formal planning frameworks that seek to maximize the *value of information* (VOI) under uncertainty.[27]

Considering the target localization problem, suppose $D_k \in \left\{d_k^1, ..., d_k^{n_s}\right\}$, where $d_k^j \in [0, 1]$ denotes a specific kind of binary soft observation report at time k. For instance, d_k^j could represent a detection/no detection event for camera j, or a binary true/false response to a semantic query j from a large list of possible queries generated by a fixed map and dictionary (e.g. 'yes/no' queries such as 'Is something by the door?', 'Is something behind the table?',..., etc.). Given a utility function $U(D_k, X_k)$ representing the expected long-term benefit of taking some discrete action A_k while the target is in state X_k, the VOI for receiving a single noisy report d_k^j in response



(a) (b)

Figure 8. (a) Snapshot of grid world problem, showing $p(X_k|D_{1:k})$ (heat map) for target randomly walking through field of 6 human-accessible cameras with limited coverage (rectangles); (b) histogram of MAP target state estimation errors for AMDP and CNN querying.

to a soft data query is

$$\begin{aligned}\mathrm{VOI}(d_k^j) &= \mathbb{E}\left[\max_{A_k} U(A_k, X_k)\right]_{(d_k^j, X_k)} - \max_{A_k} \mathbb{E}\left[U(A_k, X_k)\right]_{(X_k)}, \\ &= \sum_{d_k^j} p(d_k^j|D_{1:k-1}) \left[\max_{A_k} \int_{x_k} p(X_k|D_{1:k-1}, D_k^j) U(A_k, X_k)\right] - \max_{A_k} \int_{x_k} p(X_k|D_{1:k-1}) U(A_k, X_k), \end{aligned} \quad (9)$$

where $\mathbb{E}[f]_{(v)}$ is the expected value of f over v. Assuming cost $c(d_k^j)$ for requesting d_k^j, the human sensor should be queried for d_k^j if $\mathrm{VOI}(d_k^j) > c(d_k^j)$. Thus, (9) gives a formal way to assess whether asking for d_k^j is worth the cost, regardless of the outcome. All n_s alternatives for d_k^j can thus be compared to select the one with the highest VOI at each time k. Such *myopic* querying strategies do not consider all possible combinations of d_k^j that could be taken together at time k, but are practically implementable and still capture the bulk of information to be gained from querying. For human sensors, $c(d_k^j)$ can be related to the expected cognitive cost of re-tasking human sensors.[27] For now, $c(d_k^j)$ is ignored for simplicity, so only the utility defined by *expected information gain for sensing actions* is considered. Here, A_k is related only to the choice of $j \in \{1, ..., n_s\}$ and we seek to minimize the entropy of $p(X_k|D_{1:k-1}, d_k^j)$, so that $U(d_k^j, X_k) = \log p(X_k|D_{1:k-1}, d_k^j)$. Thus, the VOI for d_k^j in (9) is the expected decrease in posterior entropy,

$$\mathrm{VOI}(d_k^j) = \mathbb{E}\left[\mathcal{H}[p(X_k|D_{1:k-1}, d_k^j)]\right]_{(d_k^j)} - \mathcal{H}[p(X_k|D_{1:k-1})], \quad \text{where } \mathcal{H}[p(X_k|D_{1:k})] = \mathbb{E}\left[-\log p(X_k|D_{1:k-1})\right]_{(X_k)}.$$

This means that soft data d_k^j will be (myopically) requested to minimize the entropy of $p(X_k|D_{1:k})$. Entropy minimization is also widely used for tasking of hard sensors in target localization applications,[41] and so provides a useful common objective for combined hard-soft sensor scheduling. However, VOI calculations are computationally expensive and lead to NP-hard Bayesian inference calculations for marginal observation likelihoods $p(d_k^j|D_{1:k-1})$. The comparison of VOI for various d_k^j reports also becomes expensive when n_s is large for large semantic dictionaries. Hence, even with simplifications such as myopic reasoning, optimal soft data querying remains challenging.

To address this, ref. 42 describes a novel querying policy approximation based on deep learning. This approach produces a multi-layer convolutional neural net (CNN) classification model to select among the n_s possible semantic queries d_k^j to maximize the VOI eq. (9) given $p(X_k|D_{1:k-1})$. The CNN learning process uses training data labels for pairs of $p(X_k|D_{1:k-1})$ and d_k^j, which are generated from brute force VOI optimizations obtained on simulated runs of the active sensing problem. The key advantage of this approach is that the major computational expense for generating a query is moved offline and can be implemented very quickly online once the CNN is learned. This is very useful for in problems where n_s is large (i.e. large semantic dictionaries and sets of possible semantic observations). Furthermore, the CNN deep learning approach can exploit non-obvious features of $p(X_k|D_{1:k-1})$ that lead to highly accurate predictions of VOI-optimal d_k^j in complex scenarios.

Figure 8 shows an application of the CNN query policy approximation approach to a dynamic version of the target localization problem. The target moves with random walk dynamics in a 2D grid world, which is

incompletely covered by 6 static cameras that can be accessed serially by a human supervisor. In this case, $n_s = 6$, and the problem here is to determine which camera j the human should look through at each time k for the best binary measurement ('detection' or 'no detection', with known false alarm and missed detection rates). The right plots show the resulting localization errors based on maximum a posteriori (MAP) estimates of the target location obtained from the post-fusion Bayesian posteriors $p(X_k|D_{1:k})$ at each time. The first histogram shows results using an alternative baseline sensor querying policy based on a partially observable Markov decision process (POMDP) model,[43] solved using the feature-based infinite horizon augmented MDP (AMDP) approximation.[19] The results here show that, despite the highly limited coverage area of the cameras, the CNN VOI approximation does a much better job of using querying sensors than AMDP, since it leads to much better tracking of the true target location. However, the CNN policy clearly rests on having adequate training data and an accurate model of the search environment; it is more brittle than the AMDP policy approximation to changes in camera layout or subtle changes in target dynamics (both of which are easily encoded as explicit POMDP parameter variations). A promising research direction to overcome such issues is to combine the strengths of AMDP and CNN policy approximations in a structured manner. To overcome the curse of dimensionality, we are also developing GM adaptations of these policy approximations.[44]

4. FUSION OF FREE-FORM LOCATIVE SKETCH DATA

Natural language data fusion can be easily implemented with predictable verbal communication patterns (e.g. for human operators in specialized operational settings such as soldiers, pilots, WiSAR incident commanders, first responders, etc.) and also tends to work well in structured environments for autonomous perception. However, it can be difficult to adequately capture the intended meaning of 'off nominal' or unexpected verbal human sensor observations. Language-based observations may also be difficult to interpret in large unstructured or featureless environments, e.g. outdoor spaces with many non-distinct features ('near those rocks', 'behind the tree'). Furthermore, the soft data fusion techniques discussed so far assume that human sensor models are perfectly known and obtainable via offline calibration, which can be very time-consuming and is not robust to unanticipated semantic context shifts. A priori calibration is also infeasible for large-scale spatial sensing operations with *ad hoc* human elements that can opportunistically join or leave a sensing network at any time, e.g. hikers and volunteers who provide sporadic reports during a WiSAR mission.

To address these issues in the context of outdoor target search, ref. 45 proposed a novel sketch-based interface and probabilistic model for fusing human-generated positive/negative target location observations. This sketch interface allows in situ human search agents to quickly convey binary observations in different regions of the search space X via free-form encirclements drawn directly on known map $\mathcal{M}$, as shown in Fig. 9.

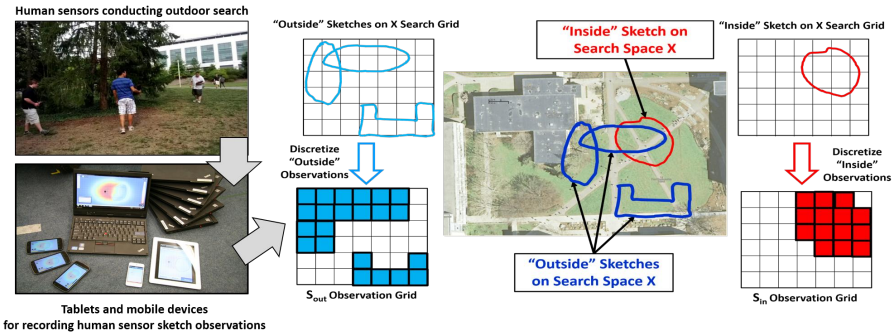


Figure 9. Example search map with 'inside' and 'outside' region free-form human sketches; (left and right) discretization of sketches on search space grid for X to form binary observation grids S_{in} and S_{out}, where filled red/blue cells corresponding to observations are assigned value '1' and empty/unobserved cells are assigned '0'.

A single human sketch sensor observation in this case specifies ad hoc spatial region boundaries in which the target could either be present/'inside' or absent/'outside'. This protocol enables coarse classification of the search space based on independent positive/negative evidence obtained during the search mission, e.g. from visual terrain sweeps or clues extracted from the search environment (footprints, disturbed features, etc.). For

instance, the blue sketches in Fig. 9(b) imply 'target not in these areas' (e.g. summarizing negative information obtained by a visual sweep of the areas), while the red sketch region imply 'target may be around here' (e.g. which might collectively summarize positive information gathered from clues in the search environment). Such sketch observations are qualitatively similar to 'belief' sketches used for priority searches in WiSAR,[46] and are intuitively simple to understand and implement on networked mobile devices (e.g. smartphones, tablets). A key difference is that the sketches here do not directly reflect subjective probabilistic beliefs, but instead indicate *possible constraints* on the true target state. Such sketch reports must be interpreted very carefully to account for various sources of 'human sensor noise'. For instance, sketches will not always be drawn precisely (especially in time critical situations) and thus might convey inconsistent observations from the same human at different times (cf. Fig. 9). Tendencies to report positive/'inside' or negative/'outside' information can also vary significantly across different humans. Given a suitable parametric conditional probability distribution model of a human sensor's sketch observation accuracy, consistent positive/negative information can be extracted from sketch data via Bayesian fusion while accounting for possible observation errors. As such, each sketched 'inside'/'outside' region must be parsed into an uncertain observation vector conditioned on the latent target state X, which is then fused with prior information for X. Ref. 45 describes two key technical contributions to this end: (i) specification of a likelihood function for each human that correctly accounts for spatially correlated information within each sketch; and (ii) a fully Bayesian hierarchical inference procedure for fusing sketch data and estimating target states simultaneously and online in the presence of multiple uncertain human sketch sensor likelihood parameters (which is especially useful if sparse or no training data are available for offline calibration).

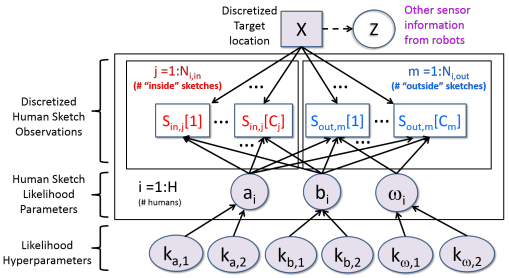
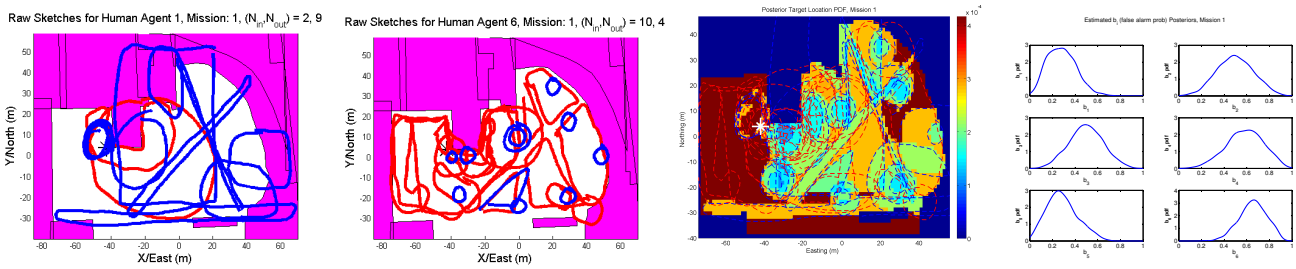


Figure 10. PGM for hierarchical Bayesian human sketch sensor data fusion: arrows denote conditional dependencies; shaded nodes denote unknown variables, unshaded nodes denote observations; continuous/discrete random variables indicated by round/square nodes. C_j and C_m denote number of marked cells in the jth 'inside' and mth 'outside' sketches for human i, respectively ($N_{i,in}$ and $N_{i,out}$ are total number of 'inside' and ''outside' sketches for human sensor i). Boxes around nodes denote repeated graph substructures.

Figure 10 shows the resulting PGM used for centralized fusion of sketch reports from multiple human sensors in a networked static target localization problem. The human sensor i sketch likelihood parameter variables $\Theta_i = (a_i, b_i, \omega_i)$ correspond to true detection rate, false alarm rate, and negative information spatial correlation, respectively. These can be estimated offline with sufficient training data, or estimated online alongside the unknown target state X. The variables labeled S_{in} and S_{out} denote parsed grid squares on the map $\mathcal{M}$ that take on values of 1 or 0, depending on whether or not human i's 'inside'/'outside' sketch contained those cells. The hyperparameters k control Gamma distribution hyperpriors that are assigned to the set of all human sensor sketch likelihood parameters. The hyperpriors effectively capture 'population statistics' for different human sensors, i.e. reflecting the idea that all human sensors tend to have similar values for Θ_i. This acts as a 'soft constraint', which allows information obtained during inference about human sensor i's parameters to indirectly restrict the values that human j's parameters can take on via conditional independence (in particular: if j's sketches are very similar to i's, then we can infer that Θ_j is probably very similar to Θ_i). As discussed in ref. 45, simultaneous parameter inference on each set of Θ_i and data fusion to estimate X can be carried out via a computationally efficient Gibbs Monte Carlo sampler. The Gibbs sampler in this case uses adaptive rejection techniques to draw samples of each unknown variable according to local posteriors that can be easily obtained analytically (up to an unknown normalizing constant) from the PGM itself.

Figure 11 (a)-(b) shows typical sketch inputs provided by 2 of 6 mobile human sensors performing an outdoor static target localization experiment (locating a small key chain buried somewhere on a campus quad, with a uniform prior $p(X)$ for the target location X). On the other hand, Fig. 11 (c) shows that a sensible target



(a) (b) (c) (d)

Figure 11. (a)-(b) Typical sketch inputs provided by 2 different mobile human sensors for a static target search mission (true target location shown as asterisk $*$); (c) Bayes posterior for X using hierarchical inference on PGM from Fig. 10; (d) posterior estimates for human sensor false alarm parameters b_i.

location posterior is obtained using the proposed simultaneous Bayesian parameter and state estimation approach based on the hierarchial PGM in Fig. 10. Fig. 11 (d) shows the resulting posterior false alarm rates obtained for each human sensor using all the corresponding sketch observations shown in Fig. 11 (c). Human sensors who report many false positive 'inside' sketches are estimated to have high false alarm rates b_i (e.g. agents 4 and 6), whereas humans who report more negative information via 'outside' sketches are estimated to be more reliable (e.g. sensors 1 and 5). The Bayesian inference process automatically accounts for these estimated discrepencies in human sensor quality, and fuses sketch information to update the posterior over X accordingly. This makes the proposed sketch fusion method especially useful for applications like WiSAR, where false alarm and missed detection rates are hard to obtain for human sensors.

5. CONCLUSIONS

Probabilistic models and Bayesian algorithms are firmly established cornerstones for tackling challenging autonomous robotic perception, learning and decision-making problems. Since the next frontier of autonomy demands the ability to gather information across stretches of time and space that are beyond the reach of a single autonomous agent, the next generation of Bayesian algorithms must capitalize on opportunities to draw upon the sensing and perception abilities of humans-in/on-the-loop. This work summarized our recent and ongoing research toward harnessing 'human sensors' for general information gathering tasks. The approach described here is grounded in rigorous Bayesian modeling and fusion of flexible semantic information derived from user-friendly interfaces, such as natural language chat and locative hand-drawn sketches. This allows human users (i.e. non-experts in robotics, statistics, machine learning, etc.) to directly 'talk to' the information fusion engine and perceptual processes aboard any autonomous agent, while also providing an formal framework for online adaptive human sensor modeling and optimal human sensor querying. This naturally enables 'plug and play' human sensing with existing probabilistic algorithms for planning and perception, which we have successfully demonstrated with real human-robot teams in target localization applications.

REFERENCES

[1] Miller, I., Campbell, M., Huttenlocher, D., Kline, F.-R., Nathan, A., Lupashin, S., Catlin, J., Schimpf, B., Moran, P., Zych, N., et al., "Team cornell's skynet: Robust perception and planning in an urban environment," *Journal of Field Robotics* **25**(8), 493–527 (2008).

[2] Argrow, B., Frew, E., Elston, J., Stachura, M., Roadman, J., Houston, A., and Lahowetz, J., "The tempest uas: The vortex2 supercell thunderstorm penetrator," *InfotechAerospace, American Institute of Aeronautics and Astronautics (AIAA)* (2011).

[3] Werner, D., "Ambition: Europa," *Aerospace America Magazine* **June**, 22–28 (2016).

[4] Kingston, D., "Intruder Tracking Using UAV Teams and Ground Sensor Networks," in [*German Aviation and Aerospace Conference 2012 (DLRK 2012)*], (2012).

[5] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al., "Mastering the game of go with deep neural networks and tree search," *Nature* **529**(7587), 484–489 (2016).

[6] Sweet, N., Ahmed, N., Kuter, U., and Miller, C., "Towards self-confidence in autonomous systems," in [*AIAA InfoTechAtAerospace 2016*], (2016).

[7] Aitken, M., Ahmed, N., Lawrence, D., Argrow, B., and Frew, E., "Assurances and machine self-confidence for enhanced trust in autonomous systems," in [*RSS 2016 Workshop on Social Trust in Autonomous Systems*], (2016).

[8] McGuire, S., Furlong, P., Heckman, C., Julier, S., Szafir, D., and Ahmed, N., "Teamwork across the stars: Machine learning to overcome the brittleness of autonomy," in [*IROS 2016 Workshop on Human-Robot Collaboration: Towards Co-Adaptive Learning Through Semi-Autonomy and Shared Control*], (2016).

[9] Miller, C., Goldman, R., Funk, H., Wu, P., and Pate, B., "A playbook approach to variable autonomy control: Application for control of multiple, heterogeneous unmanned air vehicles," in [*Proceedings of FORUM 60, the Annual Meeting of the American Helicopter Society*], 7–10 (2004).

[10] Lee, J. D. and See, K. A., "Trust in automation: Designing for appropriate reliance," *Human Factors: The Journal of the Human Factors and Ergonomics Society* **46**(1), 50–80 (2004).

[11] Sheridan, T., [*Humans and Automation: System Design and Research Issues*], Wiley, Santa Monica, CA (2002).

[12] Tellex, S., Kollar, T., Dickerson, S., Walter, M. R., Banerjee, A. G., Teller, S., and Roy, N., "Approaching the symbol grounding problem with probabilistic graphical models," in [*AAAI Conference on Artificial Intelligence*], (2011).

[13] Shah, D. and Campbell, M., "A robust qualitative path planner for mobile robot navigation using human-provided maps," in [*2011 Intl.Conf. on Robotics and Automation (ICRA 2011)*], 2580–2585 (2011).

[14] Tellex, S., Thaker, P., Deits, R., Simeonov, D., Kollar, T., and Roy, N., "Toward information theoretic human-robot dialog," in [*Robotics Science and Systems*], (2012).

[15] Arkin, J. and Howard, T. M., "Towards learning efficient models for natural language understanding of quantifiable spatial relationships," in [*RSS 2015 Workshop on Model Learning for Human-Robot Communication*], (2015).

[16] Howard, T. M., Tellex, S., and Roy, N., "A natural language planner interface for mobile manipulators," in [*Robotics and Automation (ICRA), 2014 IEEE International Conference on*], 6652–6659, IEEE (2014).

[17] Daftry, S., Zeng, S., Bagnell, J. A., and Hebert, M., "Introspective perception: Learning to predict failures in vision systems," *arXiv* **abs/1607.08665** (2016).

[18] Boyd, J. R., [*Destruction and creation*], US Army Comand and General Staff College (1987).

[19] Thrun, S., Burgard, W., and Fox, D., [*Probabilistic Robotics*], MIT Press, Cambridge, MA (2001).

[20] Bishop, C., "Pattern Recognition and Machine Learning," (2006).

[21] Morrison, J. G., Gavez-Lopez, D., and Sibley, G., "Scalable multirobot localization and mapping with relative maps: Introducing moarslam," *IEEE Control Systems* **36**, 75–85 (April 2016).

[22] Hall, D. L. and Jordan, J. M., [*Human-centered Information Fusion*], Artech House (2010).

[23] Goodrich, M. A., Morse, B. S., Engh, C., Cooper, J. L., and Adams, J. A., "Towards using Unmanned Aerial Vehicles (UAVs) in Wilderness Search and Rescue," *Interaction Studies* **10**(3), 453–478 (2009).

[24] Lewis, M., Wang, H., Velgapudi, P., Scerri, P., and Sycara, K., "Using humans as sensors in robotic search," in [*12th International Conference on Information Fusion (FUSION 2009)*], 1249–1256 (2009).

[25] Kaupp, T., Douillard, B., Ramos, F., Makarenko, A., and Upcroft, B., "Shared Environment Representation for a Human-Robot Team Performing Information Fusion," *Journal of Field Robotics* **24**(11), 911–942 (2007).

[26] Bourgault, F., Chokshi, A., Wang, J., Shah, D., Schoenberg, J., Iyer, R., Cedano, F., and Campbell, M., "Scalable Bayesian human-robot cooperation in mobile sensor networks," in [*International Conference on Intelligent Robots and Systems*], 2342–2349 (2008).

[27] Kaupp, T., Makarenko, A., and Durrant-Whyte, H., "Humanrobot communication for collaborative decision making A probabilistic approach," *Robotics and Autonomous Systems* **58**, 444–456 (May 2010).

[28] Ahmed, N., Sample, E., and Campbell, M., "Bayesian multi-categorical soft data fusion for human-robot collaboration," *IEEE Transactions on Robotics* **29**(1), 189–206 (2013).

[29] Alspach, D. and Sorenson, H. W., "Nonlinear Bayesian Estimation Using Gaussian Sum Approximations," *IEEE Transactions on Automatic Control* **AC-17**(4), 439–448 (1972).

[30] Schoenberg, J., Campbell, M., and Miller, I., "Localization with Multi-modal Vision Measurements in Limited GPS Environments Using Gauss-sum Filters," in [*2009 International Conference on Robotics and Automation (ICRA 2009)*], (2009).

[31] Taguchi, S., Suzuki, T., Hayakawa, S., and Inagaki, S., "Identification of Probability Weighted Multiple ARX Models and Its Application to Behavior Analysis," in [*48th IEEE Conf. on Decision and Control (CDC09)*], 3952–3957 (2009).

[32] Ahmed, N. and Campbell, M., "On estimating simple probabilistic discriminative models with subclasses," *Expert Systems with Applications* **39**, 6659–6664 (June 2012).

[33] Runnalls, A. R., "Kullback-leibler approach to Gaussian mixture reduction," *IEEE Transactions on Aerospace and Electronic Systems* **43**(3), 989–999 (2007).

[34] Ponda, S., Ahmed, N., Luders, B., Sample, E., Hoossainy, T., Shah, D., Campbell, M., and How, J., "Decentralized information-rich planning and hybrid sensor fusion for uncertainty reduction in human-robot missions," in [*2011 AIAA Guidance, Navigation and Control Conf.*], (August 2011).

[35] Sample, E., Ahmed, N., and Campbell, M., "An experimental evaluation of Bayesian soft human sensor fusion in robotic systems," in [*2012 AIAA Guidance, Navigation and Control Conf.*], (August 2012).

[36] Ahmed, N. and Sweet, N., "Softmax Modeling of Piecewise Semantics in Arbitrary State Spaces for Plug and Play Human-Robot Sensor Fusion," in [*Robotics: Science and Systems*], (2015).

[37] Sweet, N. and Ahmed, N., "Structured synthesis and compression of semantic human sensor models for bayesian estimation," in [*2016 American Control Conference (ACC)*], 5479–5485 (2016).

[38] Sweet, N. and Ahmed, N., "Towards natural language semantic sensing in dynamic spaces," in [*2016 RSS Workshop on Model Learning for Human-Robot Communication (MLHRC)*], (2016).

[39] Kollar, T., Tellex, S., Roy, D., and Roy, N., "Toward understanding natural language directions," in [*Proceeding of the 5th ACM/IEEE international conference on Human-robot interaction - HRI '10*], 259, ACM Press, New York, New York, USA (2010).

[40] Mikolov, T., Chen, K., Corrado, G., and Dean, J., "Distributed Representations of Words and Phrases and their Compositionality," in [*NIPS*], 1–9 (2013).

[41] Huber, M. F., Bailey, T., Durrant-Whyte, H., and Hanebeck, U. D., "On entropy approximation for gaussian mixture random vectors," in [*Multisensor Fusion and Integration for Intelligent Systems, 2008. MFI 2008. IEEE International Conference on*], 181–188, IEEE (2008).

[42] Lore, K. G., Sweet, N., Kumar, K., Ahmed, N., and Sarkar, S., "Deep value of information estimators for collaborative human-machine information gathering," in [*Proceedings of the ACM/IEEE Seventh Int'l Conf. on Cyber-Physical Systems (ICCPS 2016)*], 80–89, ACM (2015).

[43] Krishnamurthy, V. and Djonin, D. V., "Structured threshold policies for dynamic sensor scheduling: A partially observed markov decision process approach," *IEEE Transactions on Signal Processing* **55**, 4938–4957 (Oct. 2007).

[44] Porta, J. M., Vlassis, N., Spaan, M. T., and Poupart, P., "Point-based value iteration for continuous pomdps," *Journal of Machine Learning Research* **7**(Nov), 2329–2367 (2006).

[45] Ahmed, N., Campbell, M., Casbeer, D., Cao, Y., and Kingston, D., "Fully Bayesian learning and spatial reasoning with flexible human sensor networks," in [*Proceedings of the 2015 Int'l Conf. on Cyberphysical Systems (ICCPS 2015)*], accepted, to appear, ACM/IEEE (2015).

[46] Adams, J. A., Humphrey, C. M., Goodrich, M. A., Cooper, J. L., and Morse, B. S., "Cognitive Task Analysis for Developing Unmanned Aerial Vehicle Wilderness Search Support," *Journal of Cognitive Engineering and Decision Making* **3**(1), 1–26 (2009).