Application of Partially Observable Markov Decision Processes to Robot Navigation in a Minefield

Lihan He, Shihao Ji, and Lawrence Carin
Department of Electrical and Computer Engineering
Duke University
Durham, NC 27708-0291, USA
{lihan,shji,lcarin}@ee.duke.edu

Abstract

We consider the problem of a robotic sensing system navigating in a minefield, with the goal of detecting potential mines at low false alarm rates. Two types of sensors are used, namely, electromagnetic induction (EMI) and ground-penetrating radar (GPR). A partially observable Markov decision process (POMDP) is used as the decision framework for the minefield problem. The POMDP model is trained with physics-based features of various mines and clutters of interest. The training data are assumed sufficient to produce a reasonably good model. We give a detailed description of the POMDP formulation for the minefield problem and provide example results based on measured EMI and GPR data.

Introduction

In many sensing problems, a robotic platform is preferred to a humanly-operated platform, an important example being that of ground-based sensing of landmines (MacDonald 2003). The robotic platform navigates in a minefield in an autonomous fashion, with optimal decisions dynamically made for its position, orientation, and the deployment of multiple sensors. The decision optimization is based on minimizing two fundamental types of costs in landmine detection: the detection cost and the sensing cost.

The landmines and mine-like clutter vary considerably in their contents (metal, plastic, etc) and size (small, large, etc), therefore it is vital to build a unified model to represent the mines and clutter so as to make the decision making possible. There are several typical sensors used in landmine detection, including ground-penetrating radar (GPR) and electromagnetic induction (EMI) sensor, which we consider in the present paper.

The minefield problem may be cast in the form of an adaptive sensor-management problem (Kastella 1997; Abdel-Samad & Tewfik 1999) (here with two sensors, the GPR and EMI sensors), though the problem is complicated significantly by the variety of the landmine and clutter signatures. We here consider a partially observable Markov decision process (POMDP) formalism (Kaelbling, Littman, & Cassandra 1998). In the POMDP formulation the environment under test is assumed to reside within a particular state S_E, and this state is not observable directly; the state of the environment, defined by the presence/absence of a mine in the region being sensed, is unchanged by the sensing itself. The state S_E is partially observable, in the form of the measured sensor data. The agent has particular actions at its disposal, including "moving to a new location", "deploying one of the sensors", "declaring the presence or absence of landmines'. Each of these actions has an expected immediate cost, as well as an impact on the long-term sensing cost. The POMDP constitutes a framework that balances the (discounted) infinite-horizon performance of this multi-sensor problem, i.e., it accounts for the immediate expected cost, as well as discounted future costs, over an infinite horizon (Kaelbling, Littman, & Cassandra 1998).

The POMDP is employed to constitute a sensing policy, defining the optimal next action to take based upon the agent's current belief about the environment under test (Kaelbling, Littman, & Cassandra 1998). The belief is defined in terms of a belief state, a probability mass function (pmf) of the environmental states S_E, conditional on all previous actions and observations (Kaelbling, Littman, & Cassandra 1998). To compute the belief state one requires an underlying model of the environment under test (Kaelbling, Littman, & Cassandra 1998), characterized by a statistical representation of observations given a sequence of controlling actions. We assume that we have access to a sufficient ensemble of measured data collected by the GPR and EMI sensors of the mines and mine-like clutter, so that we can design the POMDP model and find the corresponding optimal policy. The target states S_T of the POMDP are defined by sensor positions relative to the target, and the sequence of target states visited is modeled as a Markov process, conditioned on the sensor-platform motion; since the target position is unknown (hidden), the state is partially observable. In this setting we must distinguish the overarching state of the environment under test S_E, which is to be inferred by the POMDP policy (via the belief state), *vis-a-vis* the states of the underlying target model S_T, which are visited when performing the adaptive sensing. Given a set of GPR and EMI data, measured at a sequence of spatial positions relative to the target, we must now develop the POMDP model.

In this paper we develop a POMDP formulation based on the assumption that *a priori* and adequate training data are available for model development. We here employ measured

GPR and EMI data, for real mines and realistic clutter. The measured data considered in this study are available upon request, and therefore it is hoped that it will evolve to a standard data set researchers may use to test different adaptive sensor-management algorithms.

Partially Observable Markov Decision Processes

A POMDP model is represented by a six-element tuple $\langle S, A, T, \Omega, O, R\rangle$, where S is a finite set of discrete states, A is a finite set of discrete actions, and Ω is a finite set of discrete observations. The state-transition probability

$$T(s, a, s') = \Pr(S_{t+1} = s' | S_t = s, A_t = a) \quad (1)$$

describes the probability of transitioning from state s to state s' when taking action a. The observation function

$$O(a, s', o) = \Pr(O_{t+1} = o | A_t = a, S_{t+1} = s') \quad (2)$$

describes the probability of sensing observation o after taking action a and transiting to state s'. Finally, the reward function $R(s, a)$ represents the immediate expected reward the agent receives by taking action a in state s.

Since the state is not observed directly, a belief state b is introduced. The belief state is a probability distribution over all states, representing the agent's probability of being in each of the states based on past actions and observations, assuming access to the correct underlying model. The belief state is updated by Bayes rule after each action and observation, based on the previous belief state:

$$b_t(s') = \frac{1}{c} O(a, s', o) \sum_{s \in S} T(s, a, s') b_{t-1}(s) \quad (3)$$

with the normalizing constant

$$c = \sum_{s' \in S} O(a, s', o) \sum_{s \in S} T(s, a, s') b_{t-1}(s) = \Pr(o | a, b) \quad (4)$$

A POMDP policy is a mapping from belief states to actions, telling the agent which action to take based on the current belief state. The goal of the POMDP is to find an optimal policy by maximizing the expected discounted reward

$$V = E[\sum_{t=0}^{k-1} \gamma^t R(s_t, a_t)] \quad (5)$$

which is accrued over a horizon of length k. The discount factor $\gamma \in (0, 1]$ describes the degree to which future rewards are discounted relative to immediate rewards. If k is finite the optimal action depends on the distance from the horizon, and therefore the policy is termed non-stationary. However, often an appropriate k is not known, so we may consider an infinite-horizon policy, i.e., k goes to infinity, for which we require $\gamma < 1$. An infinite horizon also implies a stationary policy, independent of the agent's temporal position.

When in belief state b, the maximum expected reward k steps from the horizon $V(k)$ is

$$V^{(k)}(b)$$

$$= \max_{a \in A} \left[\sum_s R(s, a) b(s) + \gamma \sum_o p(o | a, b) V^{(k-1)}(b_a^o) \right] \quad (6)$$

where b_a^o the belief state after the agent takes action a and observes o, as updated in (5). The $V^{(k)}(b)$ represents the maximum expected discounted reward the agent will receive if it is in belief state b and takes actions according to the optimal policy for future steps. In this paper policy design is performed using the PBVI algorithm, with details provided in (Pineau, Gordon, & Thrun 2003).

The POMDP Model for Landmine Detection

We consider a minefield as an area of land where mines of several known types and other mine-like objects (clutter) are buried underground. The positions of the mines and clutter are unknown. The task is to detect the mines at a low false alarm rate, with an economic use of sensors. This is a highly dangerous task and therefore a robot platform is designed to perform it. Below we specify the POMDP model for this problem.

Feature extraction

The EMI measurement in any position is the complex response of the magnetic field as a function of frequency. A typical EMI response when the sensor is above a metal mine is shown in Figure 1. The magnetic field induced by a target is represented by the formula (Gadar, Mystkowski, & Zhao 2001)

$$H(\omega) \propto a + \frac{b_1 \omega}{\omega - j\omega_1} + \frac{b_2 \omega}{\omega - j\omega_2} \quad (7)$$

where a, b_1, b_2 are related to the magnetic dipole moments of the target, and ω_1 and ω_2 represent the associated EMI resonant frequencies. Features can be extracted from an EMI observation by fitting the measured data to the model in (7), assuming additive noise n in the observation, i.e.,$Y(\omega) = H(\omega) + n$. The nonlinear fitting parameters $\{a, b_1, b_2, \omega_1, \omega_2\}$ are our EMI features.

The GPR observation for a given position is recorded as the radar signature as a function of time. The time dimension is associated with the depth of the soil: the signals reflected from deeper positions have larger time delays. Figure 2(a) shows a typical GPR observation when the sensor is above a plastic mine, and Figure 2(b) is a 2-dimensional scan of the landmine signature. Features extracted from a GPR observation include the raw moments (corresponding to energy features) and central moments (corresponding to variance features) of the time series.

Specification of S, A, Ω, and $R(s, a)$

The landmine detection problem can be viewed as a generalization of the tiger problem (Kaelbling, Littman, & Cassandra 1998). Each mine type represents a type of tiger, and each clutter type represents a type of non-tiger (reward). The robot can observe sensor readings (listening in the tiger problem) to gain information or make a declaration with regard to the presence or absence of a mine (opening the

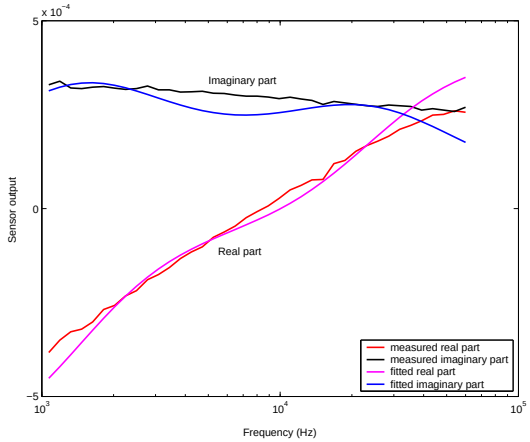
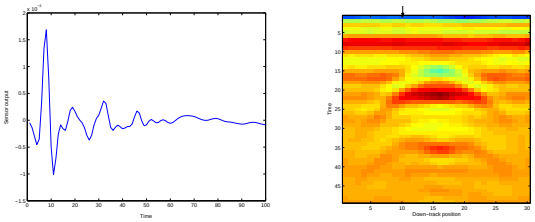


Figure 1: EMI response and model fit when the sensor is above a metal mine.



(a) (b)

Figure 2: The GPR response when the sensor is above a plastic mine. (a) Amplitude vs. time signal in one position. Units in time are 0.05 ns. The first peak corresponds to the reflection from the ground surface. (b) 2-dimensional scan of a plastic mine signature. Units in down-track position are 2 cm. The arrow indicates the position where a sensor measures the signal in (a).

door in the tiger problem) to complete the present detection phase. When learning the policy, the problem resets immediately after a declaration is made, and a mine or clutter is randomly presented to the robot. This corresponds to the robot randomly encountering a mine or clutter in the next detection phase.

Across all five types of mines and clutter considered, we define a total of 29 states, i.e. $S = \{1, 2, \cdots, 29\}$. The 29 states are divided into 5 disjoint subsets: $S = S_m \cup S_p \cup S_{c_1} \cup S_{c_2} \cup S_{c_0}$, denoting metal mine, plastic mine, Type-1 clutter (large-sized), Type-2 clutter (small-sized), and "clean", respectively. The number of states in each of the five subsets are 9, 9, 9, 1, and 1, respectively. The multiple states of metal mine, plastic mine, Type-1 clutter represent their respective 9 annulus sectors. Definition of the states is illustrated in Figure 3(a).

In most cases, a mine and clutter is cylindrically symmetric and is buried with its axis perpendicular to the ground surface. This implies that the robot will not distinguish states $1, 2, 3, 4$ (which are approximately equidistant to the metal mine) by observing a single sensor reading in each respective state. However, by remembering its past observations and actions, the robot will be able to tell apart these ambiguous states.

The robot has 15 possible actions, i.e., $A = \{1, 2, \cdots, 15\}$, of which the first 10 are sensing actions and the rest are declaration actions. Each sensing action has the format of "$move$ and then $sense$", where $move \in \{$stay, walk south, walk north, walk east, walk west$\}$ and $sense \in \{$sense with EMI, sense with GPR$\}$, with EMI representing an electromagnetic induction sensor and GPR a ground penetrating radar. Of the 5 declaration actions, one declares the present sub-area (where the robot currently is) to be "clean", and four respectively declare that there is a "metal mine", "plastic mine", "Type-1 clutter", or "Type-2 clutter" buried beneath the present sub-area.

The set of possible observation Ω is obtained as the codebook resulting from vector quantization (Gersho & Gray 1992) of the continuous sensor signatures. Each of the two sensors, EMI and GPR, generates its own codebook independently, resulting in two disjoint codebooks, which are taken a union over to produce Ω.

The reward function $R(s, a)$ is specified as follows. Denote by m any of the 9 states for a metal mine, by p any of the 9 states for a plastic mine, and by c_1 any of the 9 states for a Type-1 clutter. Denote by c_2 the Type-2 clutter and by c_0 the "clean" state. See Figure 3(a) for definition of the states. Denote by A_t the action of declaring the present sub-area to be the state of t. Then $R(s = t,\, a = A_t) = 10$, for $t = m$, p, c_1, c_2, or c_0; $R(s = m$ or $p,\, a = A_{c_1}$ or A_{c_2} or $A_{c_0}) = -100$; $R(s = c_1$ or c_2 or $c_0,\, a = A_m$ or $A_p) = -50$; $R(s = m,\, a = A_p) = 5$; $R(s = p,\, a = A_m) = 5$. All the remaining entries of $R(s, a)$ are zero.

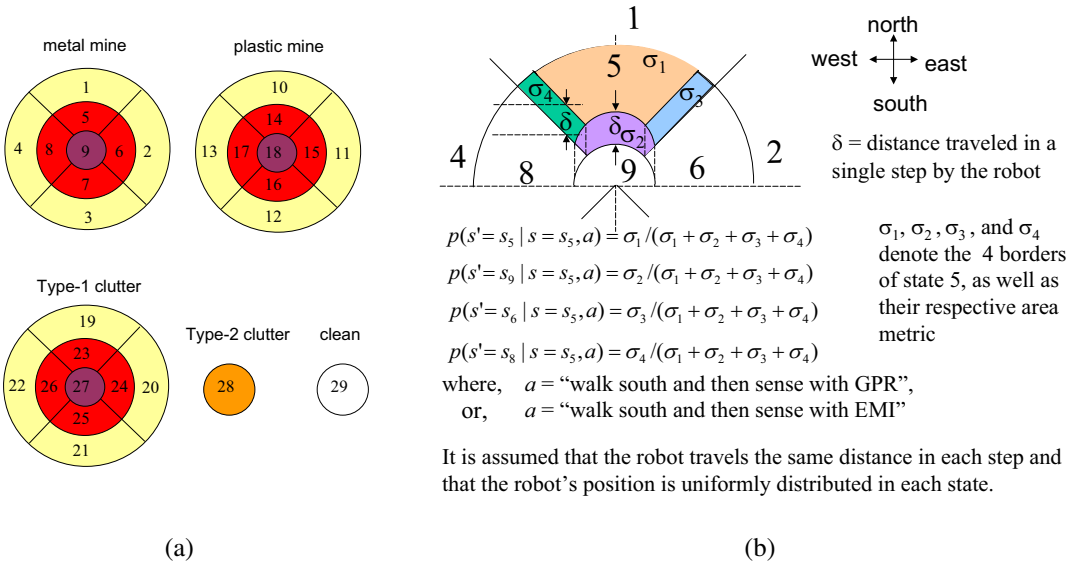
Estimation of $T(s, a, s')$ and $O(a, s', o)$

The two sensing actions involving "stay" do not cause state transitions, hence $T(s, a, s')$ is an identity matrix when a is "stay and sense with GPR" or "stay and sense with EMI". All remaining sensing actions can result in transitions from one state to another. Assuming that the robot travels the same distance in each step and that the robot's position is uniformly distributed in any given state, the probabilities of these transitions are easily determined by using an elementary geometric probability computation. Figure 3(b) illustrates how the transition probabilities for the two sensing actions involving "walk south" are computed.

Computing $T(s, a, s')$ and $O(a, s', o)$ requires prior knowledge of the possible mines and clutters. This poses no problem here, as we have the templates of the possible mines and clutter, which can be employed to compute $T(s, a, s')$ as well as collecting the training signatures for estimating $O(a, s', o)$.

Experimental Results

We consider a robot navigating in three simulated mine fields. The EMI and GPR data are pre-collected over a $1.6 \times 1.6\ m^2$ per simulated mine field, with sensor data collected at a 2 cm sample rate in two coordinate dimensions.



(a) (b)

Figure 3: (a) Definition of the states for the minefiel navigation problem. Metal mine, plastic mine, and Type-1 clutter (large-sized) are each modeled by 9 states, indexed 1 to 9, 10 to 18, and 19 to 27, respectively; Type-2 clutter (small-sized) is modelled by a single state (state 28); state 29 is used to indicate "clean" (i.e., there are no mine or mine-like objects buried underground). (b) Illustration of the geometric method in computing the state transition probabilities $T(s, a, s')$ when a is one of the two sensing actions involving "walk south". It is assumed that the robot travels the same distance in each step and that the robot's position is uniformly distributed in any given state.

The pre-collected data are used to simulate the data collected by an autonomous two-sensor agent, as it senses within the mine field. The first simulated mine field is shown in Figure 6.

Clearly, to avoid missing landmines the robot should search almost everywhere in a given mine field. However, we hope that the robot can actively decide where to sense as well as which sensor to use, to minimize the detection cost. Considering these two requirements together, we assign a basic path as shown in Figure 4 (dark blue curve with arrows). The basic path defines the lanes as indicated by light blue in the figure, and the robot is restricted to move along the lanes by taking actions within the lanes. The basic path restrains the robot from moving across the lanes, and the robot defines sectors along each lane as being characterized by one of the mines/clutter, including clean, while moving in an overall direction consistent with the arrows in Figure 4. The distance between two neighboring basic paths should be less than the diameter of a landmine signature.

It is possible that after many measurements in one local area, the agent still cannot make a declaration. For example, this can occur if the model we build does not fit the data in this area, possibly because our model does not include the current underground target. More measurements do not help to make a better decision. If this happens, it is better to say "I do not know" rather than continue sensing or make a reluctant declaration. We let the robot declare unknown in this situation, while in the lifelong learning algorithm the oracle is employed.

In the offline-learning approach the training data are given in advance, and the training phase and test phase are separate. We use Mine Field 1 (Figure 6) as the training data to learn the model and the policy, and then test our method on all three mine fields. The training data and test data match well in that the three mine fields contain almost the same types of metal mines, plastic mines and clutter. The clutter includes metal clutter (soda can, shell, nail, coin, screw, lead, rod, and ball bearing) and nonmetal clutter (rock, bag of wet sand, bag of dry sand, and a CD).

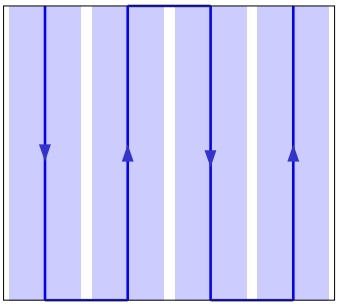
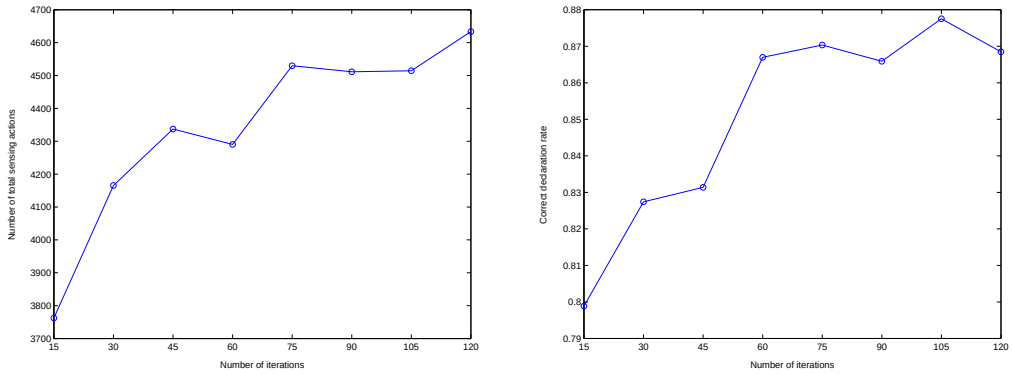


Figure 4: Robot navigation path in a mine field. The dark blue curve is the basic path, which defines the lanes as indicated by light blue. The robot is restricted to move along the lanes by taking actions within the lanes. The basic path restrains the robot from moving across the lanes.



(a) (b)

Figure 5: Detection performance as a function of number of iterations when learning the policy. (a) Number of total sensing actions. (b) Correct declaration rate.

Table 1: Detection results on three mine fields

		Mine Field 1	Mine Field 2	Mine Field 3
Ground truth	Number of mines (metal+plastic)	5 (3+2)	7 (4+3)	7 (4+3)
	Number of clutter (metal+nonmetal)	21 (18+3)	57 (34+23)	29 (23+6)
Detection result	Number of mines missed	1	1	2
	Number of mines missed	2	2	2

Model training and policy design

Using Mine Field 1 as the training data set the POMDP model is built and the policy is learned by PBVI. The number of sensing actions and the correct declaration rate as a function of iteration number when determining the policy are plotted in Figure 5. The correct declaration rate is defined as the ratio of the number of correct declarations relative to the number of all declarations. Note that the correct rate is not equivalent to probability of detection since one landmine could be declared multiple times, and the correct declaration of clutter or clean is also counted in the correct rate. However, it does reflect the detection performance by comparing declaration position and ground truth. From Figure 5, after 75 iterations and five belief expansion phases, the PBVI-learned policy becomes stable.

Landmine detection results

The stationary policy from the last subsection is then used to navigate the robot in three simulated mine fields. The ground truth and detection results are summarized in Table 1. As an example, the layout of Mine Field 1, the declaration result and a zoom-in of sensor choices are shown in Figure 6. Note that one target may be declared several times.

Missed landmines are usually caused by one of the following two reasons: the mine has very weak signal in both EMI and GPP responses, such as a small anti-personnel mine, which is a low-metal content mine; or the mine is very close to some large metal clutter, so that the clutters strong response hides the weak signal of the mine. From Figure 6(c), we see that the policy selects GPR sensors to interrogate plastic mines, while it prefers EMI sensors when metal mines are present. This verifies the policy to some degree since the EMI sensor is almost useless for detecting plastic mines, but is good for detecting metal mines. We also see that on the clean area or at the center of a landmine, a declaration is made only based on very few sensing actions, usually two or three, since it is relatively easy for the robot to estimate its current states. However, at the edge of a landmine, where there is an interface between two objects (the landmine and the clean), the robot usually requires many Number of total sensing actions sensing actions to make a declaration. The robot requires, on average, approximately 4500 sensing actions in one mine field; the correct declaration rate is about 0.87 (see Figure 5). As a comparison, if a myopic policy is applied, where the agent considers only one step ahead to select actions, a total of around 8000 sensing actions are needed, and a correct declaration rate of 0.82 is achieved. Note that if one senses on every grid point using both sensors, the total number of measurements is $2 \times 800^2 = 12800$.

Conclusions

We have addressed the problem of employing ground-penetrating radar (GPR) and electromagnetic induction (EMI) sensors placed on a single platform, with the objective of performing adaptive and autonomous sensing of landmines. The problem has been formulated in a partially observable Markov decision process (POMDP) setting, under the assumption that adequate and appropriate data are available for learning the underlying POMDP models, with which policy design can be effected. The algorithm has been tested, with encouraging performance, on measured EMI and GPR data from simulated mine fields.

The assumption that adequate training data are available

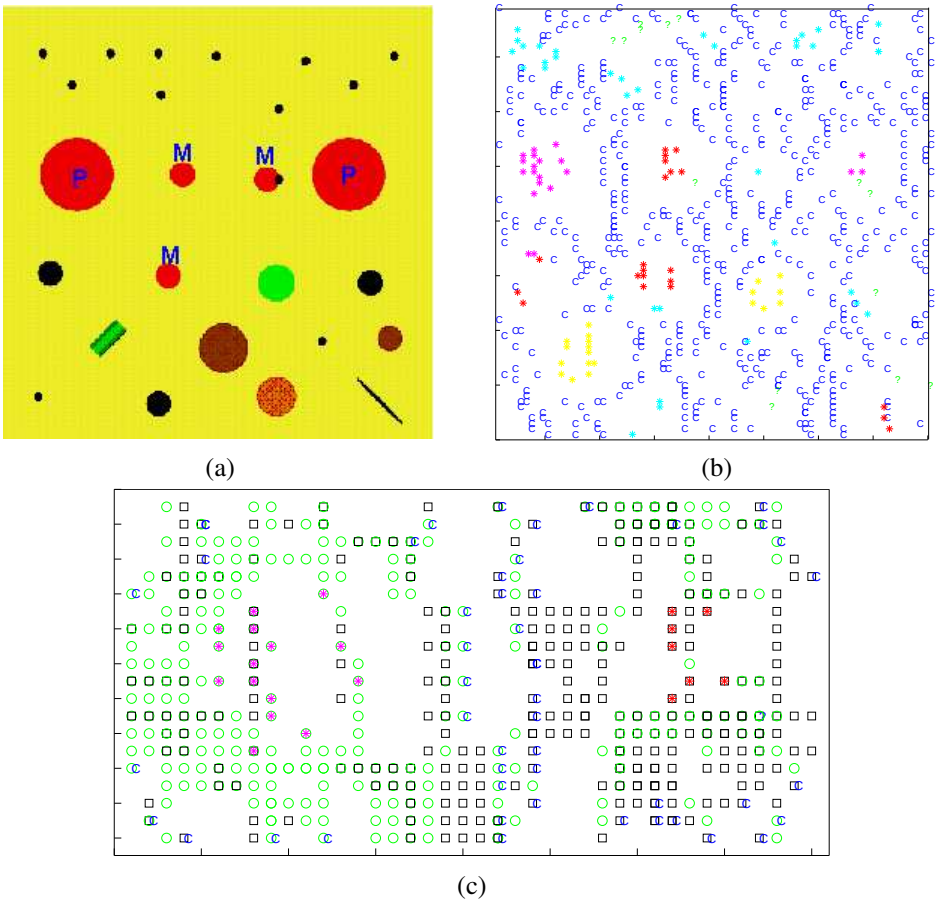


Figure 6: Ground truth and detection details in Mine Field 1. (a) Ground truth. The red circles are landmines, with "M" and "P" indicating metal mine and plastic mine, respectively; the other symbols represent clutter. Black dots are small metal segments and the rest are large-sized metal or nonmetal clutter. (b) Declaration result. The blue "C" means a declaration of "clean", the green "?" means "unknown", and the stars with various colors represent declarations of mines or clutter. Red star: metal mine; pink star: plastic mine; yellow star: Type-1 clutter; cyan star: Type-2 clutter. (c) Sensor choice in the broken-lined rectangular area shown in (b). The black square means sensing with EMI sensor and the green circle means GPR sensor. It can be seen that the policy prefers the GPR sensor for plastic mine (left half in (c)) and the EMI sensor for metal mine (right half in (c)).

is often inappropriate, and therefore in the next phase of this work we will consider a lifelong-learning algorithm in which little if any *a priori* information is assumed with regard to the mines, clutter and soil conditions. The formulation considered for this latter case will be based on the recently developed MEDUSA algorithm (Jaulmes, Pineau, & Precup 2005).

References

Abdel-Samad, A. A., and Tewfik, A. H. 1999. Search strategies for radar target localization. *Proc. International Conf. Image Proc.* 3:862–866.

Gadar, P. D.; Mystkowski, M.; and Zhao, Y. 2001. Landmine detection with ground penetrating radar using hidden markov models. *Geoscience and Remote Sensing* 39:1231–1244.

Gersho, A., and Gray, R. M. 1992. *Vector Quantization and Signal Compression*. Kluwer Academic Press/Springer.

Jaulmes, R.; Pineau, J.; and Precup, D. 2005. Active learning in partially observable markov decision processes. In *Proceedings of ECML*, 601–608.

Kaelbling, L.; Littman, M.; and Cassandra, A. 1998. Planning and acting in partially observable stochastic domains. *Artificial Intelligence* 101:99–134.

Kastella, K. 1997. Discrimination gain to optimize detection and classification. *IEEE Trans. Syst., Man, Cybernetics Part A: System and Humans* 27:112–116.

MacDonald, J. 2003. *Alternatives for Landmine Detection*. Rand Corporation.

Pineau, J.; Gordon, G.; and Thrun, S. 2003. Point-based value iteration: An anytime algorithm for POMDPs. In *IJCAI*, 1025 – 1032.