Intelligent Exploration of Unknown Environments with Vision Like Sensors

Suman Chakravorty, and John L. Junkins,
Department of Aerospace Engineering
Texas A&M University, College Station

***Abstract*— In this work we present a methodology for intelligent path planning in an uncertain environment using vision like sensors. We show that the problem of path planning can be posed as the adaptive control of an uncertain Markov decision process. The strategy for path planning then reduces to computing the control policy based on the current estimate of the environment, also known as the "certainty equivalence" principle in the adaptive control literature. We propose a Monte-Carlo based estimation scheme, incorporating non local sensors, for estimating the probabilities of the environment process, which significantly accelerates the convergence of the associated path planning algorithms.**

I. Introduction

In this work we present a methodology for "intelligent exploration" of an uncertain environment. The state space of any path planning problem can be expressed as the ordered pair $(s, q(s))$ where s represents the system state and $q(s)$ represents the state of the environment at the state s. For example, in the case of a robot exploring an unknown terrain, s corresponds to the (x, y) co-ordinates of the robot and $q(s)$ corresponds to the height of the terrain $z(x, y)$ at the point (x, y). This methodology is not limited to only robotic path planning and is equally applicable to UAV navigation and multi-spacecraft imaging problems [1], [2]. The goal of the path planning strategy is to use all available information about the environment, until the current time instant, in order to plan the "best possible" path. We show that the planning problem can be modeled as a Markov decision process whose transition probabilities are unknown. Then, we show that the "intelligent path planning" paradigm reduces to the adaptive optimal control of a Markov decision process. Our formulation allows the integration of vision-based or similar sensors, i.e., sensors that allow the sensing of an environment locally as well as non-locally.

There has been substantial research in the adaptive control of controlled Markov Chains, or Markov Decision Processes, in the past two decades and it is usual to distinguish the different methods into "direct" and "indirect" adaptive control. In indirect adaptive control, the transition probabilities of the underlying Markov chain are estimated and the control is applied based on the most recent estimate of the transition probabilities [3], [4], [5]. This is known as the so-called "certainty equivalence principle". The "direct" approach to stochastic adaptive control falls under the category of "reinforcement learning" methodologies wherein the optimal control is calculated directly without resorting to estimating the transition probabilities of the underlying Markov chain [6], [7], [8]. These methods can be further distinguished into "Q-learning" and "adaptive-critics", please refer to [6], [7] for more details. Underlying all these methods is Bellman's "principle of optimality" or Dynamic Programming, which is a methodology for sequential decision-making under uncertainty [9]. In this work, we show that the path planning problem can be reduced to the adaptive optimal control of a Markov decision process and thus the above methodologies can be applied to the same. However, our methodology allows the use of non-local sensor observations into the planning algorithms. In fact, we take the indirect approach to adaptive control since we are also interested in mapping the environment. Moreover, we show that the path planning problem has special structure which can be exploited to significantly reduce the dimensionality of the problem.

Over the past three decades robot motion planning has been an active area of research. Various different methodologies have been devised for the design of collision-free path planning of mobile robots in known environments [10]. In the past decade, there has been an increasing interest in the case when the environment in which the robots are operating is partially unknown. The uncertainty in the environment is treated on a deterministic worst case [11], [12] or in a probabilistic average case basis [19]. In "probabilistic robotics", there has been substantial research in the localization of a mobile robot while simultaneously mapping the environment [13], [14]. The expectation maximization algorithm and extended Kalman filters have been used to accomplish this objective and these methods are distinguished into the concurrent mapping and localization (CML) [15] and simultaneous localization and mapping (SLAM) [17], [18] algorithms respectively. However, the primary focus of this work [13]-[18] has been the estimation of the robot position while simultaneously mapping the operating environment and not much attention has been paid to the control aspect of this problem. In [19], a game-theoretic framework is proposed for robotic motion planning wherein the uncertainties inherent in the robotic system and the environment are modeled probabilistically. Then, the authors resort to Bellman's principle of optimality [9] in order to tackle the motion-planning problem. However, even a probabilistic model of a robotic system and its environment suffers from uncertainties, i.e., the transition probabilities of the underlying markov processes are unknown or uncertain. Thus, the problem of control of such a robotic system needs to take this uncertainty into account and as such can be posed as a stochastic adaptive control problem. We make this connection in the current work. We further note that the methodology presented in this paper remains valid for path planning problems in which the environment can be adequately modeled through a stationary random process. Moreover, we show that the intelligent path planning problem has very special structure which can be exploited to significantly reduce the dimensionality of the problem. In

[20], the problem of planning the motion of a robot in the presence of randomly appearing obstacles, in a manufacturing setting, is studied. In the above mentioned work, there is a probabilistic component to the cost function, dependent on the random obstacles. Assuming certain a priori distributions for the random obstacles, the expected value of the probabilistic cost component is sequentially estimated using a Bayesian estimator and the control is applied based on the Principle of optimality. This is the Bayesian approach to the control of Markov decision processes [3]. In our methodology, we adopt a non-Bayesian parameter estimation approach, based on the Monte-Carlo method, which reduces the dimensionality of the problem by a great deal.

The original contributions of the current work are as follows. We identify that the problem of "intelligent path planning" can be reduced to the adaptive optimal control of an uncertain Markov decision process. Then, we use the "certainty equivalence principle" [3] in order to adaptively/intelligently change the control policies for path planning in an unknown environment. We propose a Monte-Carlo based approach for the estimation of the probabilities governing the environment process for vision-like sensors and demonstrate the convergence of the algorithms, under certain assumptions. The rest of the paper is organized as follows. Section II contains the formulation of the exploration problem as a Markov decision problem and a heuristic analysis of the estimation and control schemes proposed to solve the decision problem. Section III contains the formal results regarding the convergence of the estimation scheme under a fixed exploration policy. In section IV, we provide formal results regarding the convergence of the exploration policies under the planning methodology introduced in section II. In section V, we extend the formulation to the case when we do not have perfect observations of the system state and section VI contains the conclusions.

II. Intelligent Exploration Systems

A. Preliminaries

Let the state of the exploration system be denoted by s, $s \in S$, where S denotes a finite state space. We denote the state of the environment at the system state s, i.e., the local state of the environment at system state s, by $q(s)$. For example:

- In the case of robotic exploration of unknown terrain, s corresponds to the (x, y) co-ordinates of the robot and $q(s)$ corresponds to the height of the terrain $z(x, y)$ at the point (x, y).
- In the case of a UAV navigating enemy territory while avoiding radar detection, the s variable corresponds to the position (x, y, z) of the UAV while $q(s)$ corresponds to the binary valued variable indicating the presence or lack of radar coverage at the point (x, y, z).
- In the case of a robot navigating on a shop floor, s represents the position of the robot on the shop floor and $q(s)$ is a binary valued variable which represents the presence or absence of an obstacle at the point s.
- In the case of multi-spacecraft imaging, the s variable represents the positions of the (u, v) coverage of the component spacecrafts in the imaging system while $q(s)$ represents the frequency content of the image at the point s ckh.

In fact, we note that if the environment at any system state can be modeled as a stationary random process, and if the environment process is independent for different system states, the methods developed in this paper can be applied to such a system. For example, if the presence of obstacles at a certain point on a manufacturing shop floor can be modeled as a stationary random process, then the method is applicable.

From hereon we shall assume that the system state is sensed perfectly and only the environment is sensed imperfectly. We shall extend the formulation to the imperfectly observed system state case in section 5.

For simplicity, we shall assume that the state space is finite. Let the number of system states be N and let the number of possible environment states, at any system state s, be D and we denote this set by Q. We denote the local state of the exploration system by the ordered pair $(s, q(s))$. Thus, the total number of states in the state space is ND. Let the set of control actions be denoted by U and let the number of control actions possible be denoted by M. We shall denote any particular control action by u. We make the following Markovian assumption about the system. Let $F^t = \{(s_0, q_0(s_0)), u_0, \cdots, (s_{t-1}, q_{t-1}(s_{t-1})), u_{t-1}\}$ represent the history of the process till time t.

$\mathcal{A}$ **2.1:** We assume that the current system state, s_t, is dependent only on the system state and control input at the previous time instant, i.e.,

$$p(s_t/F^t) = p(s_t/s_{t-1}, u_{t-1}). \quad (1)$$

The above assumption is satisfied if the system state is "controllable". Though this assumption is strictly not true for any choice of control, nevertheless it can be satisfied if there is enough control authority in the system or for suitably high level control (please refer to the numerical example). The environment is assumed to satisfy the "incoherence" assumption below.

$\mathcal{A}$ **2.2:** The environment process is "incoherent", i.e., the environment process is spatially uncorrelated and temporally stationary. In other words, if $\{q_t(s), s \in S\}$ denotes the environment process, $q_t(s)$ is a stationary process for all $s \in S$. Moreover, $q_t(s)$ is independent of $q_\tau(s')$ whenever $s \neq s'$, for all t, τ.

In statistical optics, the term "incoherent" is used to describe a light source which satisfies the above assumptions, i.e., when $q(s)$ corresponds to the electromagnetic field emitted by the light source at the position s. Hence, the term "incoherent" environment. Deterministic environments (like an unstructured terrain) automatically satisfy the above assumptions. Typically, a large component of any environment can be shown to satisfy the "incoherence" assumption.

Then, we have the following result for the system transition probabilities:

Proposition 2.1: Under assumptions $A2.1$,$A2.2$, the following holds:

$$p((s_t, q_t(s_t))/F^t) = p(s_t/s_{t-1}, u_{t-1})p(q_t(s_t)). \quad (2)$$

Proof: See Appendix.

The transition probilties $p(s_t/s_{t-1}, u_{t-1})$ quantify the control uncertainties inherent in the system and are assumed to be known beforehand, at least through a simulation model. This is a reasonable assumption if we have good knowledge of the the system or can simulate it on a computer. The terrain/

environmental uncertainty $p(q(s))$ is unknown and successive estimates are made of the uncertainty as the planning proceeds to completion. The question is then, "how to use this increasing information of the terrain in order to plan better paths?".

The goal of path planning can be framed as an infinite horizon discounted stochastic optimization problem, i.e., given the initial state $(s_0, q_0(s_0))$, the optimal control policy $\mu^*(s_0, q_0(s_0)) = \{\mu_1, \mu_2, ...\}$ corresponding to the optimal path is defined by

$$\mu^*(s_0, q_0(s_0)) = argmin_\mu E_\mu(\sum_{t=1}^{\infty} \beta^t c((s_t, q_t(s_t)), (s_{t-1}, q_{t-1}(s_{t-1})), \mu_{t-1})/(s_0, q_0(s_0))), \tag{3}$$

where $c((s_t, q_t(s_t)), (s_{t-1}, q_{t-1}(s_{t-1})), \mu_{t-1})$ is a positive pre-defined cost that the system incurs in making the transition from state $(s_{t-1}, q_{t-1}(s_{t-1}))$ to $(s_t, q_t(s_t))$ under the control action μ_{t-1}, $E_\mu(.)$ denotes the expectation operator with respect to the policy μ, and $\beta < 1$ is a given discount factor.

We adopt the following environment observation model:

- At every instant t, the system at state s_t, can observe the environment state $q_t(s)$, (i.e., the current environment state at the state s),if $s_t \in F(s) \subseteq S$, where $F(s)$ is assumed to be known beforehand. The set $F(s)$ constitutes a "footprint" of the sensor system.
- Associated with every observation-vantage point pair, $(q(s), s')$, $s' \in F(s)$, (i.e., we are observing the local environment at s, $q(s)$, from the state s'), there exists a known measurement error model, $p(\hat{q}(s)/q(s), s')$, $\hat{q}(s), q(s) \in Q$, and $s, s' \in S$, i.e., the probability that $\hat{q}(s)$ is observed when the environment is actually at the state $q(s)$, at system state s. If $s' \notin F(s)$, then $p(\hat{q}(s)/q(s), s') = 0$, for all $\hat{q}(s), q(s) \in Q; s, s' \in S$. This can be deduced from sensor calibration.

The above observation model facilitates the inclusion of visual sensors, or any other sensor system that allows the observation of the environment non-locally, into the problem formulation, i.e., the system does not need to be at a particular state in order to sense the environment at that state. In the next subsection, we shall discuss the particular estimation and control schemes that will be used in order to tackle the exploration problem, in a heuristic fashion. We shall formalize the results in the following sections.

B. Heuristic Analysis

1) Estimation: In the following, we shall present the estimation and control methodologies that we intend to use in order to solve the exploration problem. Consider the following relationship:

$$\pi(\hat{q}(s)) = \frac{1}{\pi(F(s))} \sum_{q(s)} p(\hat{q}(s)/q(s), s') p(q(s)) \pi(s'), \tag{4}$$

where

- $\pi(\hat{q}(s))$ denotes the probability of observing the noise corrupted environment state $\hat{q}(s)$ during the course of the exploration, i.e., the fraction of the time that the environment at state s is observed to be at $\hat{q}(s)$ during the course of the exploration.
- $p(q(s))$ denotes the true probability that the environment state is $q(s)$ at the state s.
- $\pi(s')$ denotes the probability that the system is at state s' during the course of the exploration, i.e., the fraction of the time that the system is at state s' and $\pi(F(s))$ represents fraction of time that the system spends in the set $F(s)$.

The above equation states that the frequency of observing a particular value of the environment during the course of the exploration process is related to the actual probability of the environment taking that value, the noise model and the frequency of visiting the states of the system. Since the noise model is known, and the values of $\pi(\hat{q}(s))$ and $\pi(s)$ can be estimated during the course of the exploration using the Monte-Carlo method, it is possible to obtain the true environment probabilities using equation 4. Mathematically, we have:

$$\pi_t(\hat{q}(s)) := \frac{1}{t} \sum_{n=1}^{t} 1(\hat{q}_n(s) = \hat{q}(s)), \tag{5}$$

$$\pi_t(s) := \frac{1}{t} \sum_{n=1}^{t} 1(s_n = s), \tag{6}$$

where $1(A)$ denotes the indicator function of the event A. Then, the true probabilities of the environment process $p(q(s))$ are obtained recursively as:

$$P_t(s) := arg \min_{P \in \bar{V}} ||\Pi_t(s) - \Gamma_t(s) P||^2, \tag{7}$$

where

$$P_t(s) = [p_t(q_1(s)), \cdots, p_t(q_D(s))]', \tag{8}$$

$$\Pi_t(s) = [\pi_t(q_1(s)), \cdots \pi_t(q_D(s))], \tag{9}$$

$$\Gamma_t(s) = \alpha(s)[\gamma_t^{ij}(s)], \tag{10}$$

$$\gamma_t^{ij}(s) = \sum_{s'} p(q_i(s)/q_j(s), s') \pi_t(s'), \tag{11}$$

$\alpha(s)$ is a normalizing factor that renders the matrix $\Gamma(s)$ stochastic, $||.||$ denotes the euclidean norm in $\Re^D$, and $\bar{V}$ represents the space of all probability vectors in $\Re^D$.

Note that $\alpha(s) = \frac{1}{\pi(F(s))}$. However, we do not need to keep track of $\pi(F(s))$, it is equal to the factor that renders the matrix $\Gamma(s)$ stochastic.

It is true that $P(s) = \Gamma(s)^{-1} \Pi(s)$, when the variables in the equation are the asymptotic values. However, it is not necessary that $P_t(s) = \Gamma_t^{-1}(s) \Pi_t(s)$ be a probability vector, hence, we solve the least squares problem posed in equation 7.

As mentioned previously, suppose that for observations at s, we trust sensor observations made from s' only if $s' \in F(s)$, and

$$\Gamma_t^{ij}(s) = \alpha(s) \sum_{s' \in F(s)} p(q_i(s)/q_j(s), s') \pi_t(s'), \tag{12}$$

where $\alpha(s)$ is a normalizing constant which renders $\Gamma_t(s)$ a stochastic matrix. Suppose further we have that

$$\sum_{j \neq i} p(q_j(s)/q_i(s), s') \leq \epsilon, \tag{13}$$

for all $s' \in F(s)$. Physically, the above assumption implies that if ϵ is small enough, then the sensor observations are trustworthy for observing the environment at s from the point s'. Then, it follows using the fact $\Gamma_t(s)$ is a stochastic matrix,

that each of the eigenvalues of $\Gamma_t(s)$ lies in the right half plane in a disk of radius ϵ centered at some point in the interval $[(1-\epsilon), 1]$ on the real axis HJ. Thus, if ϵ is small enough, then $\Gamma_t(s)$ is assured to be well-conditioned and hence, the estimates $P_t(s)$ are guaranteed to be bounded. These considerations need to be taken into account while specifying the footprint of the sensor at a point s, $F(s)$. Thus, assuming that the sets $F(s)$ have been defined keeping the above considerations in mind, i.e., we use only those sensor observations which are trustworthy for estimating the environment, then the estimates are guaranteed to remain bounded (in the numerical example, we found that an ϵ level of 0.45 was sufficient to guarantee boundedness of the iterates). In light of these arguments, we shall not concern ourselves with the boundedness of the sequence $P_t(s)$ in the rest of the paper. Thus, keeping a Monte Carlo account of the probabilities, $\pi(s)$ and $\pi(\hat{q}(s))$, we can recover the true probabilities of the environment process asymptotically. Note that due to the incoherence assumption on the environment, it follows that this allows us to fully characterize the environment process. However, the notion of the probabilities $\pi(s)$ and $\pi(\hat{q}(s))$ need to be formalized. In fact, they correspond to the invariant distributions of the underlying Markov chains and thus, if the chains were ergodic, the Monte Carlo estimation scheme would converge to the invariant distributions asymptotically. These notions shall be formalized in section III and section IV. Now, we turn to the control problem, i.e., the problem of planning an optimal path given the history of the exploration till the current instant.

2) Control: In this subsection, we shall study the control problem associated with the exploration problem. Consider the stochastic optimal problem posed in equation 3. Using the Bellman principle of optimality bertsi, it can be shown that the optimal policy is stationary, i.e., the optimal control is independent of time, and that the optimal control at the state $(s, q(s))$, $\mu^*(s, q(s))$, is given by the optimality equation bertsi:

$$u^*(s, q(s)) = argmin_u \sum_{(r,\bar{q}(r))} p(r/s, u)p(\bar{q}(r))[c((r, \bar{q}(r)), (s, q(s)), u) + \beta J^*(r, \bar{q}(r))], \quad (14)$$

where $J^*(r, \bar{q}(r))$ is the optimal cost-to-go from the state $(r, \bar{q}(r))$. Moreover, J^* satisfies the following fixed point equation:

$$J^*(s, q(s)) = \min_u \sum_{(r,\bar{q}(r))} p(r/s, u)p(\bar{q}(r))[c((r, \bar{q}(r)), (s, q(s)), u) + \beta J^*(r, \bar{q}(r))]. \quad (15)$$

Adaptive Control involves controlling an uncertain system while simultaneously obtaining better estimates of the system parameters. We envisage the problem of path planning as one of adaptive control of an uncertain Markov decision process, (i.e., the transition probabilities of the Markov decision process are not known). In such a scenario, the strategy of adaptive control is to use the policy that is optimal with respect to the current estimate of the system, since it corresponds to the current knowledge of the system that is being controlled and is referred to as the "certainty equivalence principle" in adaptive control var1.

Let $T = \{t_1, t_2, \cdots, t_k, \cdots\}$ denote the set of all times at which the control policy is updated during the path planning. Let the updated control policy at time instant t_k be denoted by $\mu_k(s, q)$. Let $p_t(q(s))$ denote the estimated environmental uncertainty at the time t, obtained from the estimation equation 7. Then, the control update at time $t_k \in T$, $\mu_k(.)$, using the principle of optimality and the "certainty equivalence principle", is given by

$$\mu_k(s, q(s)) = argmin_u \sum_{(r,\bar{q}(r))} p(r/s, u)p_{t_k}(\bar{q}(r))[c((r, \bar{q}(r)), (s, q(s)), u) + \beta J_k(r, \bar{q}(r))],$$

where

$$J_k(s, q(s)) = \min_u \sum_{(r,\bar{q}(r))} p(r/s, u)p_{t_k}(\bar{q}(r))[c((r, \bar{q}(r)), (s, q(s)), u) + \beta J_k(r, \bar{q}(r))].$$

In the development till this point, we have outlined the model for path planning in an unknown environment. We have shown that the problem can be modeled as the adaptive control of a Markov Decision Process consisting of a known control dependent system and an unknown "incoherent" environment. Then, using the principle of optimality and the "certainty equivalence principle", we have outlined a methodology for updating the path planning policies. Moreover, we have also shown that the transition probabilities of the underlying MDP have a special structure and outlined a Monte-Carlo based scheme to recursively estimate the environmental process using vision like sensors. Due to paucity of space we do not formalize the heuristic ideas that were presented in this section. In the next section, we present an example of a mobile rover navigating an unknown unstructured terrain using vision like sensors.

III. Illustrative Example: A Mobile Rover Navigating an Unknown Unstructured Terrain

In this section, we apply the methodology developed so far in this work to the problem of path planning of a mobile robotic rover exploring an unknown terrain. The terrain is unstructured and is estimated as the maneuvers proceed to completion. Also, it is assumed that the terrain does not change throughout the duration of the path planning maneuvers.

In this example the system state, s, is the (x, y) grid points and the environment variable q is the height of the terrain, z, at the grid point. We discretized the (x, y) plane into a 20x20 grid, i.e., the number of system states, N, is equal to 400. The height of the terrain was discretized into 10 equispaced intervals, i.e., in this case the number of environment states, D, is equal to 10. There were four allowable control actions, $N/E/S/W$, which corresponded to the rover going North/ East/ South/ West to the adjacent grid point. Note that the above (suitably high level) description of the system makes the system state "controllable". Thus, in this example, the number of controls, M, was equal to 4.

We assumed that there was no uncertainty in control or localization, i.e., $p(r/s, u)$ corresponds to a point measure. This is an idealized situation, however, the method remains valid even for situations when there is uncertainty $p(r/s, u)$.

We assumed that we had no prior knowledge of the terrain that was to be negotiated. Thus, the initial environment uncertainty, $p_0(q(s))$, was modeled as a uniform distribution. For the noise model $p(q(s')/q'(s'), s)$, we assumed that

the sensors observed the environment at state s' only if the system was at states s lying in the region $||s - s'|| \leq R$, i.e., $F(s) = \{s' : ||s - s'|| \leq R\}$. This constitutes the "footprint" of the sensor system at s. Moreover, the noise grows monotonically, with distance of the observed point from the vantage point/ sensor, and the height of the terrain at the observed point. For the sake of comparison, we did simulations where the rover made observations $(s, q(s))$ only when it was at the grid point s, i.e., only local observations were made.

The incremental cost function $c((r, \bar{q}(r)), (s, q(s)), u)$ penalized path length of the maneuver and also, the distance to the goal point, which was assumed to be the (1,1) grid point. The respective penalties were in the ratio 10:1, i.e., the rover was penalized ten times as much for climbing steep terrain with respect to minimizing the incremental path length. The control policies were updated after every trip to the goal point. The initial state was randomly varied. This is necessary in order that the rover be able to fully explore the terrain, i.e., the strong ergodicity assumption is satisfied for the underlying Markov chain. Another approach could be to have a suit of way points distributed uniformly across the terrain, in some order, and letting the rover find optimal routes from one way point to the next. This would also insure that the terrain is explored sufficiently. With vision sensors, the rover finds a near-optimal path by starting always from the point $(0, 0)$ but many parts of the environment remain unexplored. This is a manifestation of the exploration-exploitation trade-off. However, the environment can be identified in closed loop because it is control independent and thus, no randomization is required in the control input.

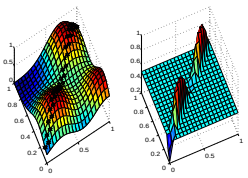

Fig. 1. Performance after 1 trial

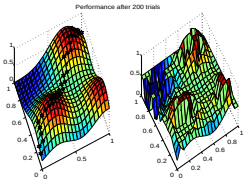

Fig. 2. Performance after 200 trials

The simulation results are shown in Fig.1- Fig.11. The left plot in each figure shows the actual terrain and the path of the rover along the terrain according to the current estimate of the terrain (which is shown in the right plot). Fig.1- Fig.5 represent the progress of the algorithm when only localized sensors are used while Fig. 6 - Fig. 11 show the progress of the learning when vision-like sensors are used. As can be seen from the plots, the rover is able to find a near optimal route after around 400 trials for the localized sensors and around 25 trials for the vision like sensors. Moreover, for the terrain estimates to converge, with vision sensors, the algorithm takes around 100 trials while it is the order of 800 trials for the localized sensors. This order of magnitude improvement is to be expected since the vision sensors sense around 10 environment states at every time step whereas the local sensor senses only one and the convergence of the respective learning histories reflect this phenomenon faithfully.

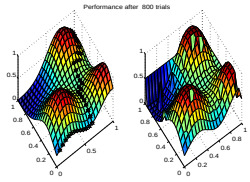

Fig. 3. Performance after 800 trials

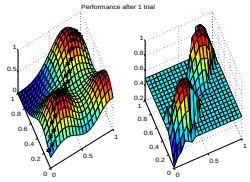

Fig. 4. Performance after 1 trial with vision sensors

Thus, in this section, we have presented a simple example where the methodology developed in this paper was applied to the problem of a mobile robotic rover negotiating an unknown and unstructured terrain. The results obtained from the numerical study are seen to validate the methodology for intelligent path planning using vision sensors as presented in this paper.

IV. Conclusion

In this work, we have presented a methodology for intelligent exploration of an unknown environment with vision like sensors. We have shown that any path planning problem, under certain assumptions, can be reduced to the adaptive control of an uncertain Markov decision process, consisting of a known control dependent system state and an unknown

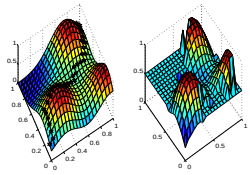

Fig. 5. Performance after 10 trials with vision sensors

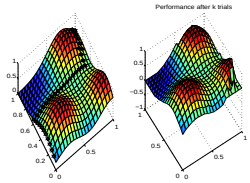

Fig. 6. Performance after 50 trials with vision sensors

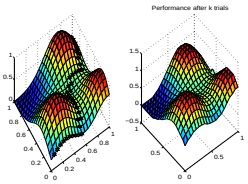

Fig. 7. Performance after 100 trials with vision sensors

control independent environment. We have used the "certainty-equivalence" principle in order to adaptively/intelligently update the control policies for path planning using a Monte-Carlo based estimation scheme which recursively estimates the probabilities of the environment process, based on non-local observations of the environment by vision-like sensors. We have illustrated through a numerical example that the inclusion of vision like sensors into the methodology significantly accelerates the convergence of the algorithms. Further, we have shown that the problem possesses special structure which can be exploited to significantly reduce the dimensionality of the problem. We have also suggested an extension of our methodology to the case when the system state is imperfectly observed.

The frequency of the control update is of considerable interest. At one end of the spectrum, we could update the control policies at every time instant (which might be computationally infeasible), while at the other end of the spectrum, we could wait until our estimates of the environment converged before we update the control policy. However, both these extremes are possibly not "optimal" and the best solution might be somewhere midway. We surmise that the policies need to be changed when the environment starts to look "significantly different" from the estimate according to which the current control has been planned. However, these are qualitative statements and need to be quantified. This will be one of the directions of our future research. For the case of imperfect state observations though we have proposed an intuitive extension of our methodology based on an estimate of the system state, the stability of the scheme needs to be analyzed. This is another avenue of research that we are currently pursuing. The use of approximate DP methods, i.e., DP with functional approximation can be pursued to make the methodology applicable to real systems, where it can act as a high level motion planner. Local robust controllers can be designed for tracking the high level path plans and the integration of the two can result in truly intelligent autonomous ssytems.

REFERENCES

[1] I. K. Nikolos, K. P. Valavanis, N. C. Tsourvelodis, N. A. Kostaras, "Evolutionary Algorithm Based Offline/Online Path Planner for UAV Navigation," *IEEE Transactions on Systems, Man and Cybernetics, Part B: Cybernetics*, vol. 33, no. 6, pp. 898-912, Dec. 2003

[2] S. Chakravorty, P. T. Kabamba and D. C. Hyland, "Modeling of Image Formation in Multi-Spacecraft Interferometric Imaging Systems," *AIAA paper 2004-5895*.

[3] P. R. Kumar and P. Varaiya, *Stochastic Systems: Estimation, Identification and AdaptiveControl*, Englewood Cliffs, NJ: Prentice Hall, 1986.

[4] V. Borkar and P. Varaiya, "Adaptive Control of Markov Chains, I: Finite parameter set," *IEEE Trans. Aut. Contr.*, vol. AC-24, pp953-958, 1979.

[5] P. Mandl, "Estimation and control in Markov chains," *Adv. Appl. Prob.*, vol. 6, pp. 40-60, 1974.

[6] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-Dynamic Programming*, Belmont, MA: Athena, 1996.

[7] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, Cambridge, MA: MIT Press, 1998.

[8] R. S. Sutton, A. G. Barto and R. J. Williams, "Reinforcement Learning is Direct Adaptive Optimal Control," *IEEE Control Systems Magazine*, vol. 12, no. 2, Apr. 1992, pp. 19-22.

[9] R. E. Bellman, *Dynamic Programming*, Princeton University Press, Princeton, NJ, 1957.

[10] J. C. Latombe, "Robot Motion Planning," Kluwer, Boston, MA, 1991.

[11] J. C. Latombe, A. Lazanas and S. Shekhar, "Robot Motion Planning with Uncertainty in Control and Sensing", *Artificial Intelligence*, 52:1-47, 1991.

[12] M. T. Mason, "Automatic Planning of Fine Motions: correctness and Completeness", *Proc. of IEEE Conf. Robotics. Automat.*, pp 484-489, 1989.

[13] S. Thrun, "A Probabilistic On-Line Mapping Algorithm for Teams of Mobile Robots," *The International Journal of Robotic Research*, vol. 20, no. 5, May 2001, pp. 335-363.

[14] S. Thrun, "Probailistic Algorithms in Robotics," *AI Magazine*, 21(4): 93-109

[15] W. Burgard, D. Fox, H. Jans, C. Matenar and S. Thrun, "Sonar-based Mapping of large-scale mobile robot environments using EM,"' *Proceedings of the International Conference on Machine Learning*, Bled., Slov., 1999.

[16] J. J. Leonard and H. J. S. Feder, "A computationally efficient method for large-scale concurrent mapping and localization," *Proceedings of the Ninth International Symposium on Robotics Research*, Salt Lake City, UT, 1999.

[17] J. A. Castellanos, J. M. M Montiel, J. Neira and J. D. Tardos, " The SP map: A Probabilistic Framework for Simultaneous Localization and Mapping," *IEEE Transactions on Robotics and Automation*, vol. 15, pp. 948-953.

[18] G. Dissanayake, H. Durant-White and T. Bailey, " A Computationally Efficient Solution to the Simultaneous Localization and Mapping Problem," *ICRA'2000 Workshop W4: Mobile Robot Navigation and Mapping*, April 2000.

[19] S. M. Lavalle, "Robot Motion Planning: A Game-Theoretic Foundation", *Algorithmica*, vol. 26, pp. 430-465, 2000.

[20] H. Hu and M. Brady, "Dynamic Global Path Planning with Uncertainty for Mobile Robots in Manufacturing," *IEEE Transactions on Robotics and Automation*, vol. 13, no. 5, pp. 760-767, October 1997.

[21] D. L. Isaacsen and R. W. Madsen, *Markov Chains Theory and Applications*, Wiley Series in probability and Mathematical Statistics, 1978

[22] O. Hernandez-Lerma and J. B. Lasserre, *Markov Chains and Invariant Probabilities*, Birkhaiuser Verlag, 2003

[23] L. P. Kaelbling, M. L. Littman and A. R. Cassandra, "Planning and Acting in Partially Observable Stochastic Domains", *Artificial Intelligence*, vol. 101, 1998, pp. 99-134

[24] W. Yang, "Convergence in the Cesaro sense and Strong Law of Large Numbers for Non-Homogeneous Markov Chains", *Linear Algebra and its Applications*, vol. 354, 2002, pp. 275-288