# Reinforcement Learning of Behaviors in Mobile Robots Using Noisy Infrared Sensing.

**R. A. Willgoss**
Department of Mechatronics,
School of Mechanical & Manufacturing Engineering
The University of New South Wales, Sydney 2052,
Australia.
Email: r.willgoss@unsw.edu.au

**J Iqbal**
Department of Mechatronics,
School of Mechanical & Manufacturing Engineering
The University of New South Wales, Sydney 2052,
Australia.
Email: p2219212@student.mech.unsw.edu.au

## Abstract

Multiple mobile robots are now being envisioned as being able to achieve a wide range of non-supervised tasks, from being hospital gofors [Evans, 1994] to organizing themselves in a factory situation [Wallace, 1998]. The tasks are characterized by requiring some learning ability in the robot that enables it to improve its performance or acquire local important information related to the task. Many researchers have used the challenges of making mobile robots play soccer as a test bed [Kim et al, 1997]. There are a variety of paradigms proposed for the learning phase such as fuzzy, neural, genetic [Zurada et al, 1994], reinforcement [Sutton & Barto, 1998] and methods including combinations such as fuzzy-neural systems [Brown & Harris, 1994]. Work reported to date has been largely theoretical but the balance is changing as a range of versatile platforms become available with trade names such as Rug Warrior, Numbat, Pioneer and Helpmate, all of which have www-based access. We report research into ways of making mobile robots learn behaviors appropriate to tasks allotted to them via the reinforcement paradigm but with noisy infrared sensors. This work is part of a more extensive exercise in comparing the practicality of various learning paradigms on the same platform. Of interest has been the learning rates and how they are affected by the richness of the environment, the need for some form of annealing to avoid local minimums, the overall specification of the behavior that is desired and the programming support offered by the platform. Results indicate that reinforcement learning for behaviors such as avoidance can be achieved in periods of a few minutes with noisy sensors at the 68HC11 computing power level chosen to work with.

## 1 Introduction

Machine learning has a theoretical history of some 50 years but has found significant practical uses in the last 10 years. Simulation [Habib et al, 1992; Willgoss, 1998] has enabled a variety of interacting scenarios to be predicted but real world experiments for multiple mobile robots are still few and far between. We address the reinforcement learning paradigm in connection with teaching mobile robots to behave correctly in multi-robot domains. Of interest is the learning behaviors such as avoidance and cooperation that may be needed for lifting and transporting heavy cumbersome objects by a fleet of robots or making sure a surface area was covered or traversed once by any one of the mobile robots in a workspace. Mataric [1997a, 1997b, 1997c, 1997d] and Michaud & Mataric [1998] have been prominent in introducing the concepts of multiple behaviors interacting to give rise to personality. Kelly & Keating [1998] have also investigated the reinforcement learning algorithm for a fleet of mobile robots and shared learning over radio links using ultrasonic sensing of proximity.

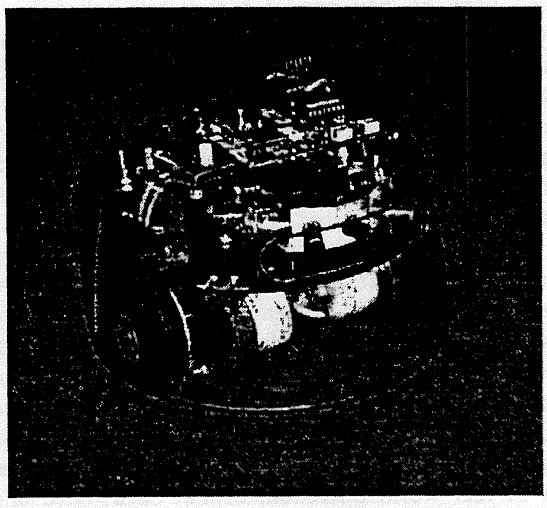
We report the use of infra red as the sensing medium relying on strength of return signals from robot on-board emitters to be a measure of the closeness to obstructions around the robot. Infrared is a noisy medium and subject to interference from a wide range of extraneous sources such as daylight, hot bodies and artificial lighting. It has been rejected for use by some on these grounds. However, sensor noisiness is an inherent feature of real world operation, cannot be avoided and makes the platform useful for testing the robustness of algorithms. It is to be noted that there are better sensing techniques that have been used for obstacle avoidance such as sonar and, more expensively, laser range-finding. However, the exercise reported here is part of a broad approach to evaluating learning algorithms with the same boundary conditions i.e. on the same platform with the same sensor-cpu-actuator combination.

## 2 Platform

To keep costs within reasonable limits, we worked with a small mobile robot, the Rug Warrior Pro, having modified it to give further sensor capability and radio communication between a fleet of five (see Figure 1). The on-board microprocessor had up to 16kb of memory available for programming with other supporting software

parked in the total of 64kb memory. Programs were often written in a subsumptive manner due to the underlying support from the software shell provided. Infrared return signals are sampled from either side of the forward aspect using pulsed emitters to distinguish noise from true reflections. Action and sensing was quantized similar to work reported by Kelly & Keating [1998]. A set of seven sensor categories (nil/nil, far left/nil, nil/far right, near left/nil, nil/near right, far left/far right, near left/near right) were matched to a set of eight actions (back left, left, forward left, forward, forward right, right, back right, back) via a probability matrix. The elements of the probability matrix were adjusted in the light of return signal changes. The sensor categories indicate the signal strength measured on the left and the right sensor simultaneously. They cover the sensor combinations possible that mainly determine movement. The actions are achieved by sending combinations of movement to the two drive motors. Hence back left is made up of a large reverse movement on the left motor and a small reverse movement on the right motor.

**Figure 1.** Modified Rug Warrior Pro.



## 3 Algorithm

The algorithm consisted of a modified version of Q learning as described in [Sutton & Barto, 1998].

$$Q(s,a) \leftarrow Q(s,a) + \alpha(r + \gamma \max_{a} Q(s',a) - Q(s,a))$$

where Q(s,a) is the value function for action a from state s and is non-Markovian. The modification dispensed with all but the reinforcement term r and substituted a normalizing process across possible actions once a preferred candidate had been chosen.

Random actions were made on the basis of the well-known roulette wheel choice mechanism followed by evaluation of the effect on the strength of both sensor inputs. The evaluation compared changes in sensor input to the overall intention. Probabilities linking decreases in sensor input arising from a particular action were reinforced and all others detracted from in the following manner:

- Pick an action i (1 of 9) based on roulette wheel probability p(i,j), j ranging over sensor input states.
- Sample the sensors i.e. measure the background input - turn on emitter - measure the input - record the difference results as Lb and Rb.
- Make a small movement in the action direction i chosen.
- Sample the sensors as before and get results
  La and Ra.
- Calculate the change in signals because of the action.
  L=Lb-La R=Rb-Ra.
  [ put together the duple (L,R) that determines the j input case ]
- If (nil/nil)
  then A = sign(motorvelL) + sign(motorvelR)
- If (near left/nil)
  if((sign(motorvelL) + sign(motorvelR)) >0)
  then forwardspeed = -1
  else forwardspeed = 1
  A = frontspeed + L
- If (nil/near right)
  if((sign(motorvelL) + sign(motorvelR)) >0)
  then forwardspeed = -1
  else forwardspeed = 1
  A = frontspeed + R
- If (near left/near right)
  if((sign(motorvelL) + sign(motorvelR)) >0)
  then forwardspeed = -1
  else forwardspeed = 1
  A = frontspeed + L + R
- If (far left/nil or nil/far right or far left/far right)
  if((sign(motorvelL) + sign(motorvelR)) >0)
  then forwardspeed=1
  else forwardspeed =-1
  A = frontspeed + L + R
- p(i,j)=p(i,j)+A
- Apply the sigmoidal function to all probabilities in the j input row
- Normalize all p(i,j) in the range 0.0 to 100.0
- Return to beginning

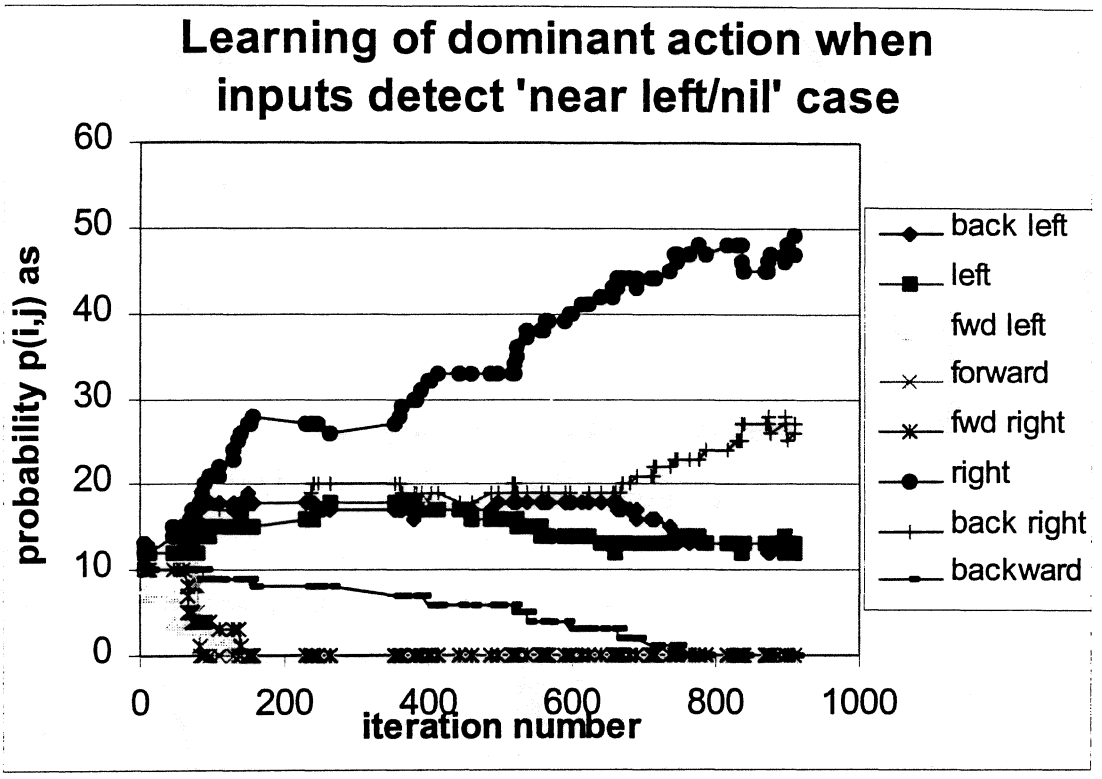
## 4 Experiments

The overall objective given to the robot was to attempt to minimize signal returns from the sensors as it moved around carrying out some other objective. The Rug Warrior Pro was loaded with software that essentially had two subsumptive modules running continuously. One module sampled and interpreted the signals from the left and right sensors then generated the motor command parameters. The other module sampled the motor

parameters and continually fed latest values to the motors thereby determining directions and speed.

The mobile robot was exposed to a variety of environments ranging from rich to sparse and the effect on learning convergence rates observed. Since the mobile robot had one small two-line LCD, all quantitative data could not be displayed on-board but was broadcast back to the PC and analyzed there. Time stamps of which actions were used and the full outcome of learned reinforcement was recorded for runs of upwards of 1000 action/reinforcement iterations. In such a set, it was observed that some actions were used more than others the frequency being linked to how rich an environment could be created around the robot. Richness was a measure of how frequently the robot could detect some return signal to work with and was enhanced by putting plenty of obstacles in the working area of the robot.

**Figure 2**




## 5 Results

Observations were made on the learning rates for one action to eventually dominate over all others and the frequency with which each input scenario was visited. The initial algorithm did not have any annealing that prevented a greedy policy from prevailing. Results showed a tendency to go for any local minimum, sometimes to the detriment of the final choice of action. A further modification included a sigmoidal transfer function on probabilities so that a greedy policy was significantly mitigated giving less preferred actions the chance to survive longer before being suppressed.

Figure 2 shows one set of probabilities changing for the case of learning to avoid an obstacle that is found 'near left/nil'. The learning has been interrupted at around 900 iterations and shows clearly the time stamps when reinforcement took place. Actions 'right' 'back right', 'left' and 'back left' all still exist as significant with a tendency for 'right' to become the greedy result given enough time.

Figure 3 shows the result for the sensor input 'nil/nil'. The learning has taken place relatively quickly to move exclusively 'forward' and is more or less complete after 300 iterations. Most of the action/reinforcement pairs chosen were monopolized by this input condition, giving limited scope for the opportunity to learn in other actions. The percentage probability actually exceeded 100% caused by the normalization process working on a sum figure of 103.0. The results were all processed in floating point on the 68HC11 microprocessor leading to longer but more accurate learning curves than for integer arithmetic.

Figure 4 shows all the learning results except for the input duple (nil/nil) expressed as net vectoral movement based on the learning that had taken place after 900 iterations, equivalent to approximately 7 minutes of running time.

Figure 3

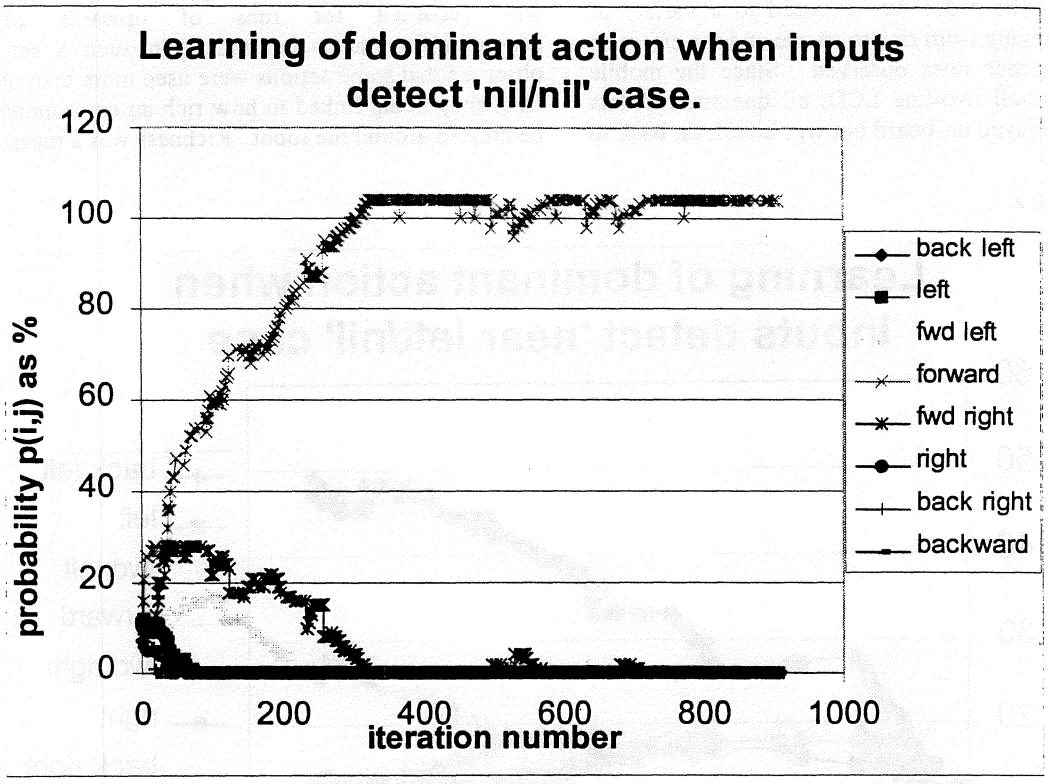
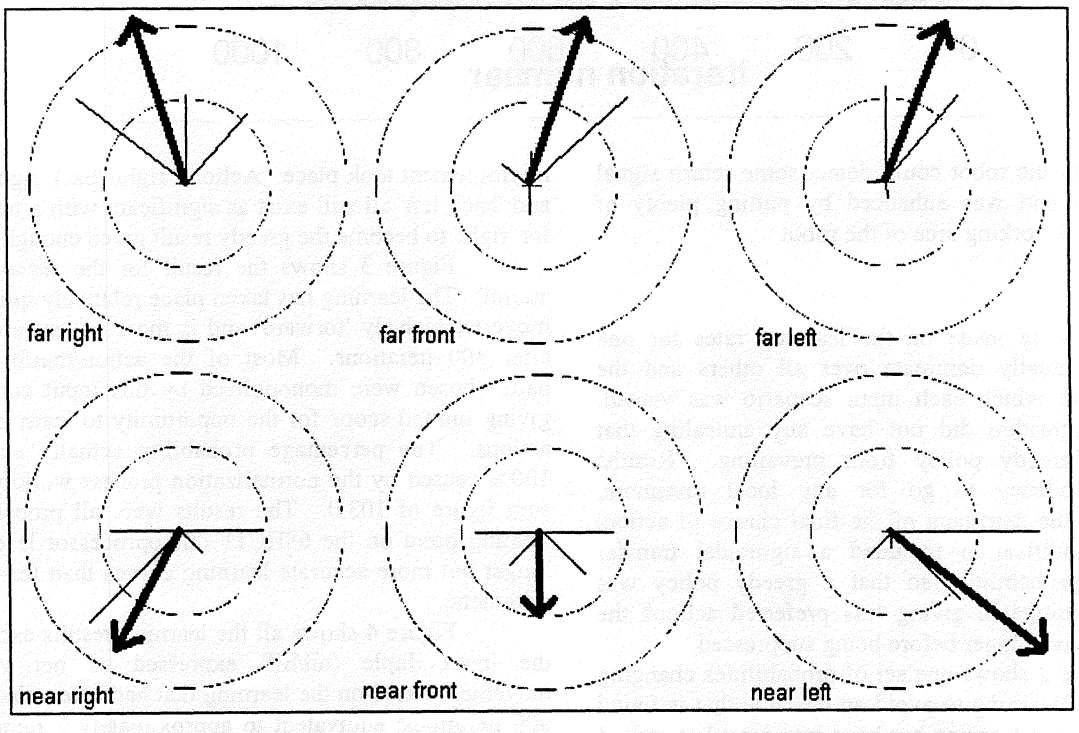



**Figure 4.** Net vectoral movement of mobile robot after most of the learning has taken place.



It can be seen that all results are moving in a sensible direction to make the robot avoid obstacles detected all around the forward-looking aspect of the robot. It is also possible to determine a learning time constant so that the length of time needed to learn can be specified. The learning time constant, expressed as a function of time, changes for each sensor input duple due to the change in frequency of use. However, for the example chosen in Figure 2, allowing for the fact that the policy will enable the dominant action to reach 1.0, the time constant is of order 1200 iterations. Such a prediction does not allow for vacillation between two or three remaining candidate actions.

## 6 Discussion

The nature of the reinforcement learning phase requires some optimization for it to work well. As mentioned previously, the results were calculated in floating point arithmetic that slowed learning down a little. However, the accuracy so gained offset the approximations and noisiness of using integer arithmetic. The main advantage in moving to floating point arithmetic came in being able to retain accuracy below 1.0 on a scale of 0 to 100. Unless this happens, the rounding errors can eliminate the less dominant contributions too quickly. With each choice of action starting at the level of approximately 11, integer arithmetic would have had an inherent error of 9% at the start of learning. The discretization of action space requires that movement was great enough to record a significant change in infrared return signal from an obstacle but small enough so that a true differential of signal strength is recorded. There is a significant problem in sensing the near field to the robot since emitter and detector cannot be continuously co-parallel in the limit of zero distance. As with all electromagnetic radiation sensing techniques, spurious sources, obliqueness of obstacles and different surface textures lead to misleading return signals. Despite these difficulties, the mobile robots using infrared (radar) adjust and cope with all of them to exhibit avoidance that is observed to be quite smooth.

## 7 Future Work

There are a number of ways the infrared sensing can be improved such as using more powerful emitters, modulation to reduce noise and employing highly directional emission. The cost benefit of these innovations is being evaluated. In addition, the Rug Warrior Pro platform also has been loaded with neurofuzzy, fuzzy and genetic based learning so that parametric exercises are possible. The next major step is to enable mobile robots to reliably identify and communicate with each other both by radio link and by infrared emission signal coding.

## 8 Conclusions

The mobile robot learning behavior in the real world has been characterized in such a way as to have confidence that learned behavior is sensible and robust even with noisy sensors. The avoidance behavior turns out to be a good vehicle for evaluating a number of issues concerning real world environmental uncertainty.

## References


[Brown & Harris, 1994] M Brown & Chris Harris, *Neurofuzzy Adaptive Modelling and Control.* Prentice Hall, UK, 1994.

[Evans, 1994] J M Evans "Helpmate: an autonomous mobile robot courier for hospitals." Proceedings IEEE/RSJ International Conference on Intelligent Robots & Systems, Munich, Germany, pp 1695-1700, 1994.

[Habib et al, 1992] M K Habib, H Asama, Y Ishida, A Matsumoto & I Endo "Simulation environment for an autonomous and decentralized multi-agent robotic system." Proceedings IEEE/RSJ International Conference on Intelligent Robots & Systems, Rayleigh, NC, USA, pp 1550-1557, 1992.

[Kelly & Keating, 1998] I D Kelly & Dave Keating "Increased learning rates through the sharing of experiences of multiple autonomous mobile robot agents." IEEE World Congress on Computational Intelligence. Alaska, USA, 1998.

[Kim et al, 1997] S H Kim, J S Choi, J K Kim & B K Kim "A cooperative micro robot system playing soccer: design and implementation." Robotics & Autonomous Systems, Vol 21, pp 177-189, 1997.

[Mataric, 1997a] Maja Mataric "Behavior based robotics as a tool for synthesis of artificial behavior and analysis of natural behavior." Journal of Experimental & Theoretical AI Vol 9, pp 323-336, 1997.

[Mataric, 1997b] Maja Mataric "Reinforcement learning in a multi-robot domain." Autonomous Robots, Vol 4, No 1, pp 73-83, 1997.

[Mataric, 1997c] Maja Mataric "Studying the role of embodiment in cognition." Cybernetics & Systems, Vol 28, No 6, pp 457-470, 1997.

[Mataric, 1997d] Maja Mataric "Learning social behaviors." Robotics & Autonomous Systems, Vol 20, pp 191-204, 1997.

[Michaud & Mataric, 1998] Francois Michaud & Maja Mataric, "Learning from history for behavior-based

mobile robots in non-stationary conditions." Autonomous Robots/Machine Learning (Joint Issue Feb 1998) pp 1-29, 1998.

[Sutton & Barto, 1998] Richard Sutton & Andrew Barto *Reinforcement Learning*. MIT Press, Bradford Books, USA, 1998.

[Wallace, 1998] A Wallace, "Flow control of mobile robots using agents." 29th International Symposium on Robotics 97, BRA, Birmingham UK, April 27-May 1, 1998.

[Willgoss, 1998] Richard Willgoss "The simulation of multiple mobile robot task assignments." 29th International Symposium on Robotics 97, BRA, Birmingham UK, April 27-May 1, pp 127-130, 1998.

[Zurada et al, 1994] J Zurada, R Marks II & C Robinson eds. *Computational Intelligence*. IEEE Press No. PC04580, 1994.