Learning and Shaping in Emergent Hierarchical Control Systems

Bruce L. Digney[1]
Defence Research Establishment Suffield
Box 4000, Medicine Hat, Alberta, CANADA, T1A 8K6

Abstract

The use of externally imposed hierarchical structures to reduce the complexity of learning control is common. However it is clear that the learning of the hierarchical structure by the machine itself is an important step towards more general and less bounded learning. Presented in this paper is a nested Q-learning technique that generates a hierarchical control structure as the robot interacts with its world. Furthermore, given the frailties of real machines and the long learning times required, it is becoming clear that fully unassisted learning for robots is unrealistic and when one considers the tremendous amount of information that novice humans/animals receive it is also unreasonable. Also, presented in this paper are methods for pre-training and supplying initial guidance to prepare robots for future situations.

1 Introduction

Much research is currently being pursued to allow autonomous agents or robots to learn from their environments [1] [2]. Researchers have realized that for robots to advance, beyond their current role of unintelligent drones working in highly structured and controlled environments, to intelligent self-directed and adaptive robots operating in unstructured and changing environments, they must be capable of learning. The reasoning for such a statement is clearly seen, when one considers how impractical it would be to attempt to quantify, solve and embed in the robot all the control information that is required to operate in an unstructured environment. Almost certainly, one critical problem would be overlooked and the robot would fail, notwithstanding the vast majority of information that it was burdened with, but never needed. Furthermore, situations such as space exploration often deny designers the information required to pre-specify robotic control systems. As well, many situations exist in which specific control would simply be impractical. For

[1]The author can be contacted via phone (403) 544-4854, FAX (403) 544-4704 or E_mail bdigney@dres.dnd.ca.

instance, if a designer has to foresee all the possible problems and then step in to augment the control of a domestic lawn mower robot whenever a situation arises that it was not explicitly programmed for, the difficulty would far outweigh any benefits gained. In the face of these problems, researchers have concluded that it is important to make robots more like humans/animals in so much as they learn what is needed and they adapt to changes as they occur.

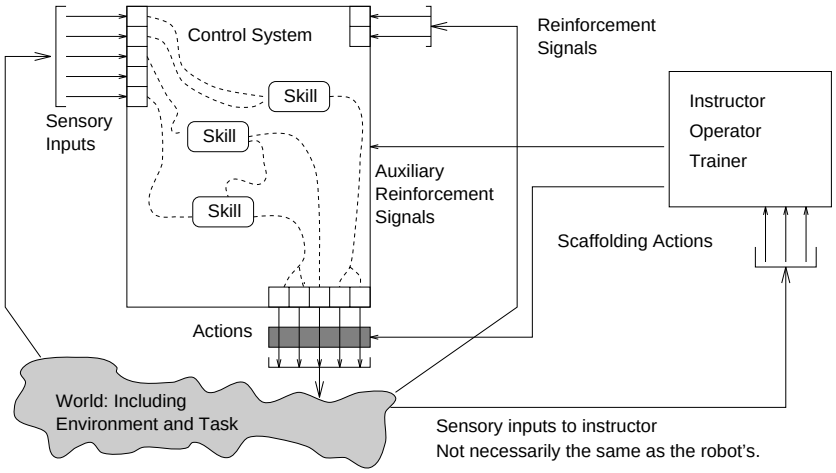


Figure 1: Schematic of sensory, action and reinforcement signal configuration for a learning control system. In addition, external guidance in the form of scaffolding actions and auxiliary reinforcement signals is being supplied from a more experienced operator, instructor or another robot.

2 Nested Q-learning

In ongoing research a nested Q-learning algorithm [3] is being developed that allows for the generation of hierarchical control systems. Due to the nature of this paper nested Q-learning will only be described in general terms, a more detailed derivation is available elsewhere [3]. In nested Q-learning any distinct recognizable sensory state becomes a *feature*. These features are discovered by the robot as it moves through its world. The control strategies that move the robot between features become known as *skills*. These skills are learned by the nested Q-learning algorithm and can invoke not only primitive actuator movements, but other skills as well. It is this nested and sometimes recursive action selection that allows for a hierarchical control system to emerge.

In addition to learning how to cascade skills into a hierarchical top-down (goal driven) control system, the nested Q-learning also learns the bottom-up (sensor driven) reactive/opportunistic responses. These bottom-up responses detect sensory conditions that represent reactive or opportunistic situations that require attention. These are usually surprising or unforeseeable events that occur outside of the normal learned routines. The bottom-up activations trigger a cascade of top-down control signals in an attempt to avert a reactive or benefit from an opportunistic situation. The overall result of the emergent hierarchy combined with the learned bottom-up activations is a flexible hierarchical control system. Within

this control system operation is generally hierarchical, but with the flow of control signals and activations not always in a top-down direction. This *tangled hierarchy* combines the benefits of both hierarchical abstraction and bottom-up flexibility.

The generation of hierarchical structures requires more computational (*mental*) effort which will payoff by transferring previously learned skills to future task/environment settings. The learning of a hierarchical control structure is analogous to a student who invests initial effort in discovering the underlying principles (the structure) of a problem, rather than simply memorizing a solution (monolithic). Once the underlying principles are understood they can be transferred to more difficult tasks. Such is not possible if the solutions are only memorized were the information gained, albeit at a lesser expense, would be useless in other situations.

3 Shaping Emergent Hierarchical Control Systems

Given the frailties and sluggishness of real robotic hardware (when compared to computer simulations) the learning of a monolithic solution, let alone the added initial difficulties of hierarchical structure, is possible only in toy-like domains. However, when one considers the tremendous amount of information that is imparted to humans through infancy and training, not to mention genetic predispositions, it is clearly unreasonable to expect fully unassisted learning in robots. It is proposed in this paper that, a robot controlled by a nested Q-learning algorithm can be prepared for future endeavours much the same way that humans are. That is, the robot is allowed to practice related tasks and what is learned will benefit it later. Thus, the concept is to prepare the robot as much as possible for its future tasks without explicitly specifying (because in most cases one cannot) what the robot is to do. The robot simply learns as much as it can and then will try and hopefully (if pretraining was accurate) be able to use part of what it learned in training to its benefit in learning other tasks and environments. Apart from regimented pretraining, as the robot moves from normally task to task and environment to environment it will have the accumulated information from its past experiences which are available to it in the useful form of skills. These transportable skills will allow the robot to learn progressively more complex tasks. Eventually this continual learning will allow the robot to learn tasks that would be impossible in a simple monolithic approach.

It has been proposed that in order to have learning robots perform in complex (human/animal like) situations, the robot must be able to benefit from similar pretraining and guidance (also referred to as shaping) received by biological agents and carry previously learned information to future endeavours. Figure 1 shows an emergent control system benefiting from two forms of external guidance. Scaffolding actions are shown subsuming the robot's own actions. Auxiliary reinforcement signals are shown providing favorable reinforcement signals that encourage the robot to perform, and hence, learn skills that it would not normally concentrate on. If the trainer is correct, these skills will later prove useful to the robot in the future. Both these external contributions, in addition to the cumulative effect of the robot's past experiences will influence the shape of the control structure that emerges. It is important to realize that just as one cannot reach into the head of a student or animal in training and directly shape the control systems within, one can only influence what develops in these emergent controls systems indirectly by

controlling the situations and rewards that are experienced.

3.1 Scaffolding Actions

From the development of nested Q-learning [3], the selection of a primitive action or a complex skill is a combination of an exploitive and an exploratory component. Once an action has been taken, be it exploratory or exploitive, state changes and reinforcement signals result from which the control system learns. Scaffolding actions subsume the actions chosen by the control system (initially poor and potentially damaging) and replace them with the actions of a more experienced robot or expert operator (perhaps through tele-operation). One can picture scaffolded actions by thinking of how one move the legs of an infant in an approximate stepping pattern while teaching the infant how to walk. The actions of the expert operator would give the novice robot a coarse approximation of the control strategies that the expert employs to accomplish a task. Through such an expert/novice interface, the novice may be able to learn strategies from the expert that the expert may not be able to directly articulate or present to the novice in any other, more direct, manner. The old adage *show me, don't tell me* applies. Furthermore, only the robot can see the world through its own eyes. Any attempt to teach it or prespecify control strategies to it can only be improved once the robot is in control and operates according to its own perceptions and actuator capabilities.

3.2 Auxiliary Reinforcement and Staged Learning

The most important factor in the emergence of a control structure is the robot's past experiences gained either by chance or regimented training scheme, just as the present actions of a human/animal are undeniably shaped by their past. As argued earlier, it is unreasonable to expect fully unaided learning of human-like tasks from an uninitiated robot when humans/animals have a lifetime of experiences to draw upon when learning a new task. In addition, many of these human experiences may have been selected as part of a pre-training scheme. Thus, many component control strategies useful in learning new complex tasks are already present when the animal/human progresses to new situations.

Whenever an external trainer is used to encourage the performance of potentially useful skills into an emergent hierarchical control system, it is called staged learning. It is accomplished by using auxiliary reinforcement signals, as shown in Figure 1, to reward the performance of potentially useful skills. Normally, these skills would not be directly rewarded from external reinforcements but would have to be learned by external reinforcement flowing downward through an emerging hierarchy. The greater the distance between the actuators and the final desired skill, that is the number of skill levels, the more difficult it will be to learn. Eventually, as a repertoire of useful skills is built up, the robot will be able to interact and understand its world on an increasingly abstract level. Through a combination of pre-trained and self discovered skills the robot will attain self-sufficient and productive operation. As always, it will continually discover and learn new skills on its own in further attempts to improve operation or recover from changes.

4 Simulation

To demonstrate the learning of hierarchical control structures and the effect of shaping using scaffolding actions and auxiliary reinforcement signals, a simple two dimensional robot and the world of Figure 2 was used. The robot's primitive actions were capable of moving it to one of four adjacent spatial locations. The robot's sensors sensed the color of the floor panel below it, the color of an overhead signalling light and the robot's spatial location within the world. The world is shown in Figure 2 (c). It had white colored floor panels except for one blue and one green floor panel at the locations indicated. It was possible for the location of these blue and green panels to change, moving to locations indicated by $World_1$ and $World_2$ in Figure 2 (c). The robot itself remained unaware of these changes as it could only sense what was happening locally at its current spatial location within the world.

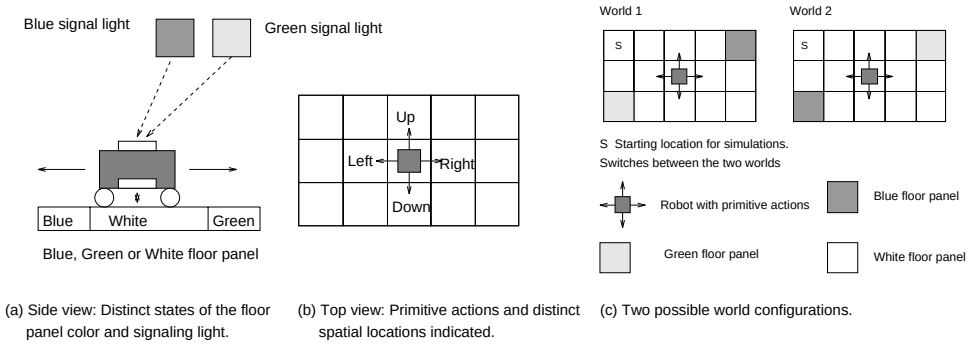


Figure 2: Simulated robot's sensory systems: (a) Floor color and signal light sensors, (b) distinct spatial locations and possible movements of the robot. Simulated world: (c) Blue and green floor panels in two possible configurations as indicted. The locations of blue and green floor panels can change to that of $World_1$ or that of $World_2$.

In these simulations situations were presented to the robot involving changes to the training environment, the tasks and external influences. In related work it has been shown that in an impoverished training environment the robot will not learn or grasp the complete structure of the task [4]. The robot was pre-trained with two tasks using auxiliary reinforcement signals. These two tasks proved useful in the robot's next task and demonstrated staged learning. Transfer of information between tasks has also been demonstrated by first training the robot on one task and then requesting a new but related task [4]. Figure 3 summarized the features found to be relevant by the robot. Although these features were subject to random discovery, they are presented in an orderly list for the readers benefit in the following analysis.

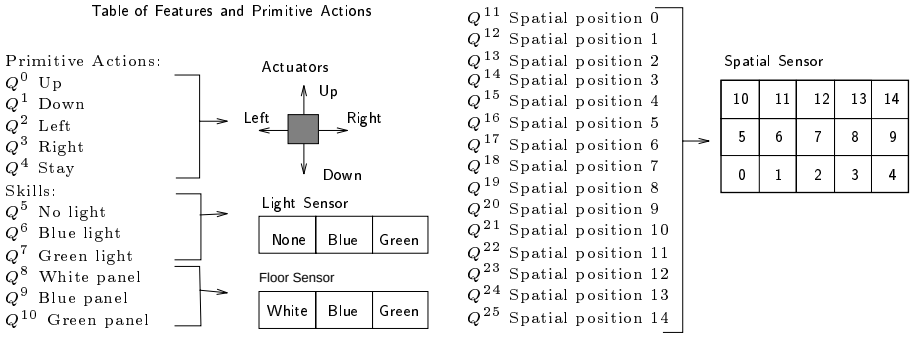


Figure 3: Summary of skills and features.

4.1 Generation of Control Hierarchies with Scaffolding Actions

To evaluate the capabilities of nested Q-learning to generate control hierarchies the agent was rewarded with the external reinforcement signal. r_{EXT} of Equation 1.

$$r_{\text{EXT}} = \begin{cases} +2 & \text{if light is blue and floor is blue.} \\ +4 & \text{if light is green and floor is green.} \\ -1 & \text{otherwise.} \end{cases} \quad (1)$$

In this simulation scaffolding actions were supplied to assist the robot in learning two skills, Q^{11} (spatial 0) and Q^{25} (spatial 14). The scaffolding actions used were purposely chosen to be poor, but better than those that would be initially taken by the inexperienced robot. After a period of time (around Time =70), the scaffolding actions were disabled and the robot was allowed to fully control itself.

The locations of the blue and green floor panels were set randomly switching between the two possible configurations as shown in Figure 2 (c). The overhead signal light was set alternating between blue and green. A robot with no prespecified information and scaffolding actions to assist only two skills was placed in this world and allowed to attempt to learn to maximize its rewards over time. The performance plots for all skills and primitive actions versus time are presented in Figure 4. The vertical axis shows the performance of the skill or primitive action. The performance is taken to be the total reinforcement signal that the skill responds with whenever it is invoked (the less negative the better the skill was performed). The axis that projects out of the page is the number of the skill or primitive action and the horizontal axis represents the expired time. For clarity, selected skills are shown in Figure 5. The vertical axis shows the performance of the skill or primitive action. The performance is taken to be the total reinforcement signal that the skill responds with whenever it is invoked. Figure 5 shows that the primitive actions responded consistently with a performance of -1.0, while all the adaptive skills change over time and usually improve. The first skills to be mastered were the ones defined by the spatial locations, skills Q^{11} through Q^{25}. What proved to be two higher level skills, Q^9, *find blue panel* and Q^{10}, *find green panel* were subsequently mastered through using the two skills of Q^{11} and Q^{25}. The initial scaffolding prevented the short period of excessive poor performance experienced by the other non-scaffolded skills and once the scaffolding actions were removed the performance of Q^{11} and Q^{25} improved dramatically,. It is also seen that the two higher level skills, Q^9 and Q^{10} improve once the scaffolding

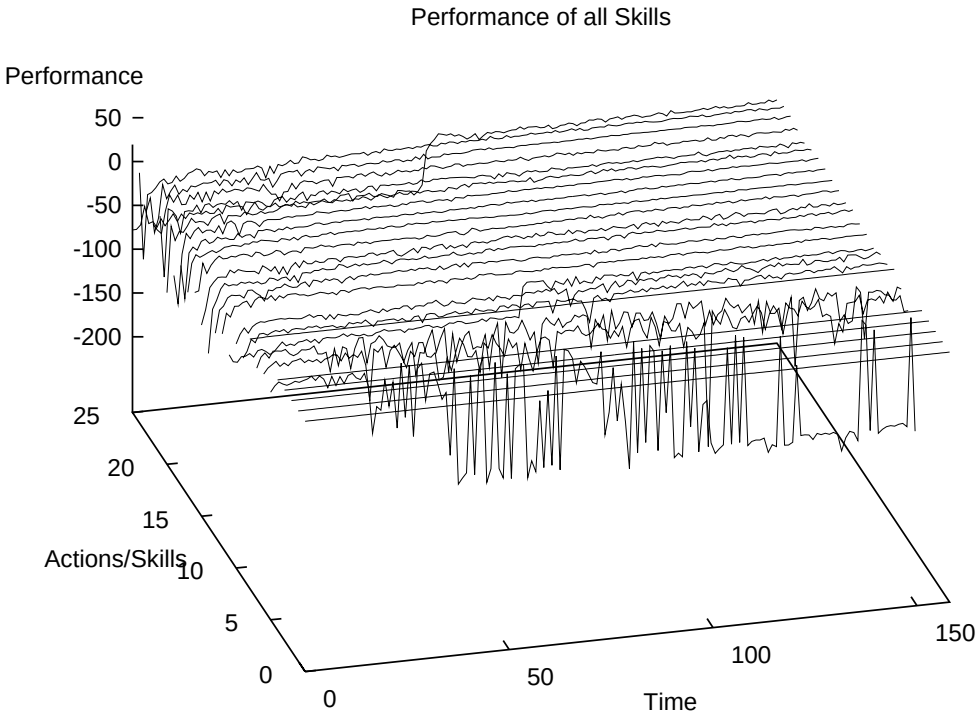
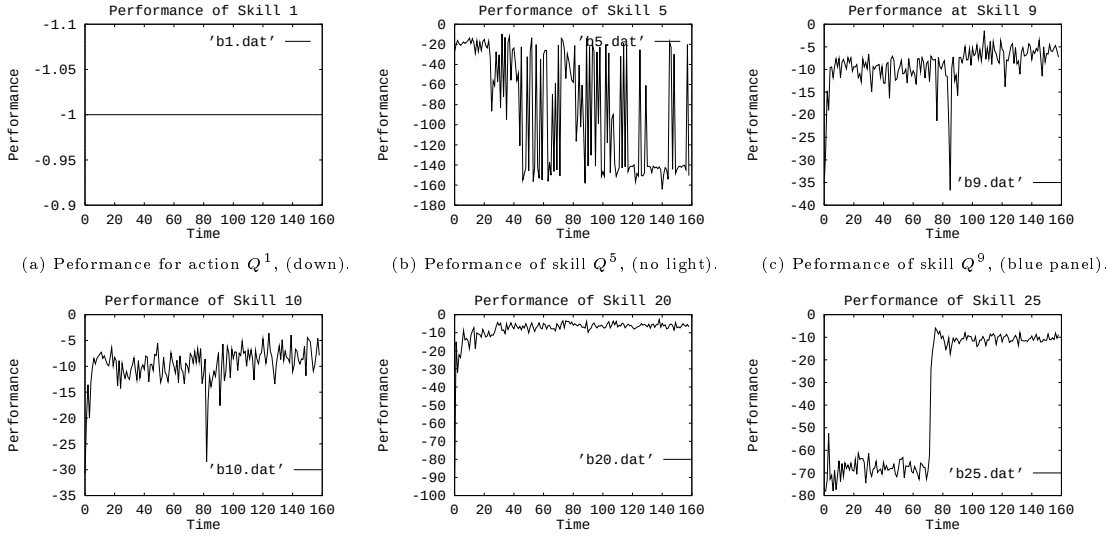


Figure 4: The performance of all skills.

actions were disabled and the two lower level skills, Q^{11} and Q^{25} were allowed to improve. Also shown is the difference in performance for a non-scaffolded skill, Q^{20}, and a scaffolded skill, Q^{25} (See Figure 5(e) and (f)). The schematic shown in Figure 6(a) shows the distinct two level architecture that emerged from the simulation. Figure 6(b) shows the movements of the robot for the skill, Q^9 for both world configurations. It is clear that a hierarchical control system emerged which, when invoked by a bottom-up opportunistic drive, began to search for the locations of the correctly colored panels.

4.2 Staged Learning

The world was set alternating between its two possible configurations. The robot was then requested through the auxiliary reinforcement signal of Equation 2 to find spatial position 0, requiring the robot to learn skill Q^{11}, and spatial position 14 requiring the robot to learn skill Q^{25}. The control system quickly learned these two skills as shown in Figure 6(c). The auxiliary reinforcements were then removed and the main task reinforcement signal of Equation 3 was applied. Having two of the relevant skills, (Q^{11} and Q^{25}) already present, the control system easily grew to its final structure and achieved successful operation as shown in Figure 6(d). This successful operation required the robot to learn skill Q^9 which fulfilled the task defined by the external reinforcement signal.



(a) Peformance for action Q^1, (down). (b) Peformance of skill Q^5, (no light). (c) Peformance of skill Q^9, (blue panel).

(d) Peformance of skill Q^{10}, (green panel). (e) Peformance of skill Q^{20}, (spatial 9). (f) Peformance of skill Q^{25}, (spatial 14).

Figure 5: Performance plots for selected skills.

$$r_{EXT} = \begin{cases} +4 & \text{if robot is at spatial position 0.} \\ +4 & \text{if robot is at spatial position 14.} \\ -1 & \text{otherwise.} \end{cases} \tag{2}$$

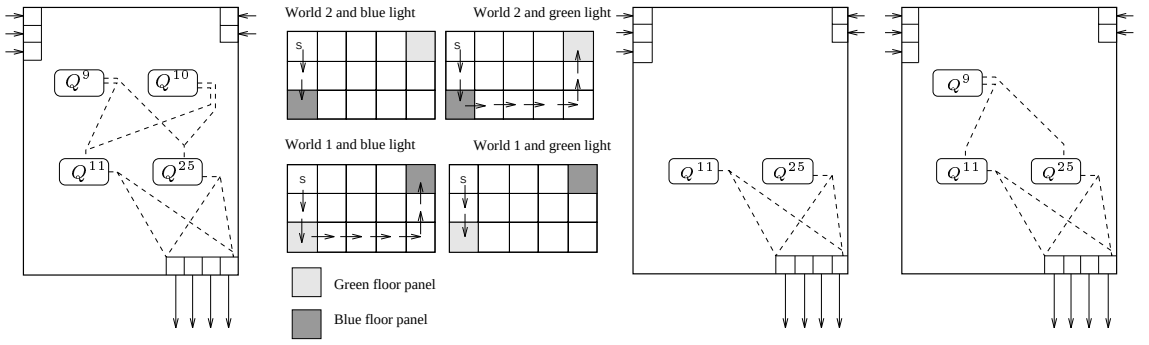
$$r_{EXT} = \begin{cases} +4 & \text{if light is blue and floor is blue.} \\ -1 & \text{otherwise.} \end{cases} \tag{3}$$

5 Conclusions

The nested Q-learning algorithm was successfully shown to generate control hierarchies of encapsulated skills. These structures resulted from having the control strategies represented as many nested evaluation functions, rather than a single monolithic structure. The skills encapsulated by these control strategies could be invoked by other skills (top-down) or invoke themselves (bottom-up). As a bottom-up reactive/opportunistic drive was triggered, it was fulfilled by a cascade of top-down invoked skills. The structure generation was assisted by various forms of pre-training and guidance. In general, previously learned skills were shown to be transferable between tasks. Since next generation robots will be expected to learn many tasks and operate in an autonomous self-directed manner, this re-use and the continual expansion of learned information will make it possible for them to learn difficult tasks quicker.

References

[1] Meyer, J.A. and Guillot, A. (1994), From SAB90 to SAB94, *Simulation of Adaptive Behavior SAB 94*, Brighton UK, August 1994, pp 2-11, MIT Press-Bradford Books, Massachussets.



(a) Emergent structure. (b) Actions taken by the robot. (c) Pre-training. (d) Second stage of learning.

Figure 6: Generation of a hierarchical control system. (a) the structure that emerged with two bottom up driven skills, Q^9 and Q^{10}, at the top and (b) actions taken by the robot. Staged learning. (c) learning low level skills Q^{11} and Q^{25} in response to the tasks conveyed with the auxiliary reinforcement signal, and (d) eventual learning of higher level skill Q^9 to fulfil the final task.

[2] Digney B.L. (1994), A Distributed Adaptive Control System for a Quadruped Mobile Robot, *Simulation of Adaptive Behavior SAB 94*, Brighton UK, August 1994, pp 344-354, MIT Press-Bradford Books, Massachussets.

[3] Digney, B.L. (1995), Emergent Control Structures: Bottom up/Top down Driven Generation of Control Structures, *The Department of National Defence and Canadian Space Agency Conference on Robotics*, October 1995, St. Hubert, PQ, CANADA.

[4] Digney, B.L. (1995), The Operator's Apprentice: Shaping Control Structures, *The Department of National Defence and Canadian Space Agency Conference on Robotics*, October 1995, St. Hubert, PQ., CANADA.