Multi-Dimensional Reinforcement Learning Using a Vector Q-Net Application to Mobile Robots

Kazuo Kiguchi, Thrishantha Nanayakkara, Keigo Watanabe, and Toshio Fukuda

Abstract: Reinforcement learning is considered as an important tool for robotic learning in unknown/uncertain environments. In this paper, we propose an evaluation function expressed in a vector form to realize multi-dimensional reinforcement learning. The novel feature of the proposed method is that learning one behavior induces parallel learning of other behaviors though the objectives of each behavior are different. In brief, all behaviors watch other behaviors from a critical point of view. Therefore, in the proposed method, there is cross-criticism and parallel learning that make the multi-dimensional learning process more efficient. By applying the proposed learning method, we carried out multi-dimensional evaluation (reward) and multi-dimensional learning simultaneously in one trial. A special neural network (Q-net), in which the weights and the output are represented by vectors, is proposed to realize a *critic* network for Q-learning. The proposed learning method is applied for behavior planning of mobile robots.

Keywords: Reinforcement learning, Q-learning, multi-dimensional evaluation, neural networks, intelligent robot.

1. INTRODUCTION

Learning algorithms based on evaluative feedback signals are generally referred to as reinforcement learning algorithms. In a reinforcement learning paradigm, a system called *agent* senses the environment and produces control actions. The environment responds to these control actions. Based on these responses, a *reward function* will evaluate the control actions. The agent tries to optimize the control policy to maximize the total expected *reward* over a finite time-span. Learning may occur using the prediction error of expected rewards. Such a learning mechanism can be found in the basal ganglia of the mammalian brain also [1]. In [1], it is experimentally shown that the activity of dopamine neurons in the ventral tegmental area and the substantia nigra of rats reflect the prediction of temporal difference or the prediction error of the expected rewards.

Reinforcement learning [2][3] plays an important role in robot learning under unknown/uncertain environments. In a reinforcement learning paradigm, the optimum control policy can be obtained based on interactive explorations in the environment. Therefore, reinforcement learning is effective for intelligent robots in making a game strategy [4] or skillful motions [5] based on their experience. Many studies on reinforcement learning have been performed to make the robots work intelligently in an unknown/uncertain environment [4]-[13]. In these studies, only one optimal or desired behavior of the robot is assumed, and evaluated with a single evaluation function or a weighted sum of evaluation functions. For some sophisticated systems such as intelligent robots, however, it is sometimes difficult to evaluate their performance with only one evaluation (*reward*). The desired behavior sometimes depends on the circumstances since contradicting objectives may have special importance in certain circumstances. For example the behavior of less energy consumption is usually preferred. However, time efficiency is more important than energy efficiency when the robot is in a rush Usually, the best behavior with respect to energy consumption is not the same as that with respect to time efficiency. Furthermore, safety is the most important when the robot carries out important tasks. Thus the desired behavior should be changed according to the situation. This kind of idea is similar to the idea of multiple reward criterion proposed by Uchibe and Asada [13].

Manuscript received January 20, 2000; accepted March 10, 2000.

Kazuo Kiguchi is with the Department of Advanced S ystems Control Engineering, Saga University, Japan. (e-mail: kiguchi@me.saga-u.ac.jp).

Thrishantha Nanayakkara is with the Department Biomedical Engineering, JohnsHopkins University, Japan. (e-mail: thrish@bme.jhu.edu).

Keigo Watanabe is with the Department Advanced Syst ems Control Engineering, Saga University, Japan. (e-mail: watanabe@me.saga-u.ac.jp).

Toshio Fukuda is with the Department Micro System E ngineering, Nagoya University, Japan. (e-mail: fukuda@mein.nagoya-u.ac.jp).

In this paper, we propose an evaluation function expressed in a vector form to realize multi-dimensional reinforcement learning. Q-learning [3], one of the basic reinforcement learning methods, has been applied in this study. A special neural network (Q-net), in which the weights and the output are represented by vectors, is proposed to realize *critic* networks for Q-learning. Each parallel network in the Q-net works as an element of the vector Q-net. The novel feature in the proposed learning algorithm is that learning occurs in all the networks while implementing any given behavior. This simultaneous learning is realized through cross-criticism by reward functions at any given time. When a certain behavior is performed, reward or punishment with respect to the performed behavior is evaluated by all the elements in the vector evaluation function. At the same time, all the networks in the Q-net try to predict the expected sum of future rewards from each network's point of view, even though the actual behavior corresponds to only one of the objectives in the vector of objectives. This kind of cross evaluations can be found in the learning process of human beings in social interaction as well. Sometimes, we observe the behavior of another person in a given situation and try to subconsciously predict future results based on a self-centered internal model. While observing we continuously criticize the internal model of prediction, and thus, performing cross learning. Therefore, we learn not only from our own behavior but also by observing the behaviors of others. The proposed learning method is based on a similar phenomenon.

In this study, we have assumed that there are obstacle regions, slippery regions, and danger regions in the working environment of the mobile robot. The robot is supposed to waste some energy and time for the slip in the slippery regions, and waste much energy and time in struggling to move in the danger regions. The dynamics of the mobile robot is taken into account. The energy minimum behavior, the hasty behavior, and the safe behavior are efficiently explored using the proposed reinforcement learning in this environment. Consequently, each weight vector and the output vector of the Q-net consist of three components in this case: 1st component for energy minimum behavior; 2nd component for hasty behavior; and 3rd component for safe behavior. The robot is able to change the optimal behavior according to the situation after the proposed learning. The effectiveness of the proposed reinforcement learning has been evaluated in simulation.

2. DYNAMIC MODEL OF THE MOBILE ROBOT

The schematic diagram of the mobile robot is shown on the left side of Fig. 1, where I_v is the moment of inertia around the c.g. of robot, v is the velocity of robot, ϕ is the azimuth of robot, and l is the distance between the left or right wheel and the c.g. of the robot.

Let

$$\boldsymbol{x}(t) = [v(t)\ \ \phi(t)\ \ \dot{\phi}(t)]^T$$

be the state variable vector and

$$\boldsymbol{u}(t) = [u_r\ \ u_l]^T$$

be the manipulated variable vector. Then the state space model for the mobile robot can be written as:

$$\dot{\boldsymbol{x}}(t) = A\boldsymbol{x}(t) + B\boldsymbol{u}(t) \tag{1}$$

with

$$A = \begin{bmatrix} -2c/(Mr^2 + 2I_w) & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & -2cl^2/(I_v r^2 + 2I_w l^2) \end{bmatrix}$$

$$B = \begin{bmatrix} -kr/(Mr^2 + 2I_w) & -kr/(Mr^2 + 2I_w) \\ 0 & 0 \\ krl/(I_v r^2 + 2I_w l^2) & krl/(I_v r^2 + 2I_w l^2) \end{bmatrix}$$

where M represents the mass of robot, I_w is the moment of inertia of wheel, c is the viscous friction factor of wheel, k is the driving gain factor, r represents the radius of wheel, and u_r and u_l are the right and left driving input torques, respectively.

The physical parameters of the mobile robot used in this study are given by $I_v = 0.6541$[kg m^2], $M = 25.5$ [kg], $l = 0.165$ [m], $r = 0.05$ [m], $I_w = 0.4419$*10-3 [kg m^2], $k = 90$, and $c = 0.0479$ [kg m^2/s].

3. MULTI-DIMENSIONAL REINFORCEMENT LEARNING

To clarify the basic concept of the proposed learning, the Q-learning method, one of the basic reinforcement learning methods, has been selected in this study. The proposed learning method is applied for behavior planning of the mobile robot. A special neural network (Q-net) is proposed to realize *critic* networks. In the proposed Q-net, the weights and the output are represented by vectors, although those are

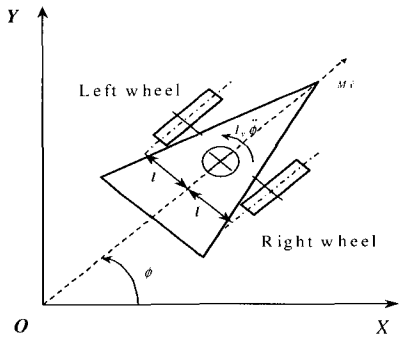


Fig. 1. Schematic diagram of the mobile robot.

usually represented by scalars. Each component of the vectors is in charge of each item of the evaluation (reward). In this study, evaluation is carried out with respect to energy consumption (energy minimum behavior), time efficiency (hasty behavior), and safety (safe behavior), assuming that there are obstacle regions, slippery regions, and danger regions in the working environment of the mobile robot. In this case, each weight vector and output vector of the Q-net consist of three components (i.e., 1st component: for energy consumption, 2nd component: for time efficiency, and 3rd component: for safety). After a certain behavior is performed, each component of the weight vectors, and the output vector of the Q-net is adjusted based on reward or punishment for energy minimum behavior, hasty behavior, and safe behavior.

3.1. Q-net architecture

The proposed Q-net consists of three layers (input layer, hidden layer, and output layer). There are 16 input variables (1: distance to target, 2: angle to target, 3: distance to obstacle, 4: angle to obstacle, 5: distance to the first slippery area, 6: angle to the first slippery area, 7: distance to the second slippery area, 8: angle to the second slippery area, 9: position of robot in x-direction, 10: position of robot in y-direction, 11: velocity of robot in x-direction, 12: velocity of robot in y-direction, 13: azimuth of robot, 14: azimuth change rate, 15: left wheel torque, 16: right wheel torque).

There are 50 neurons in the hidden layer. The activation function used in the neurons is written as:

$$y_i = \frac{1}{1+e^{-s_i}}, \quad i = 1,...,50 \tag{2}$$

$$s_i = w_{oi} + \sum_{j=1}^{16} w_{ij} x_j$$

$$w_{oi} = [w_{1oi} \;\; w_{2oi} \;\; w_{3oi}] \tag{3}$$

$$w_{ij} = [w_{1ij} \;\; w_{2ij} \;\; w_{3ij}]$$

where w_{oi} is the bias weight vector of the ith activation function, w_{ij} represents the connecting weight vectors between the ith activation function and the jth input given by x_j.

The output of the Q-net is the Q values of the current control input combination given the situation. The Q values are calculated by:

$$Q_v = \sum_{i=1}^{50} wo_i y_i \tag{4}$$

where wo_i is the output weight vectors of the Q-net that connect the activation function and the output node.

3.2. Definition

Let the right and left side control torque inputs to the mobile robot by a conventional controller based on a potential field method be denoted by u_{cr} and u_{cl}, respectively. Denote the right and left side control torque inputs given by the Q-net be $u_{qr} \in U_{qr}$ and $u_{ql} \in U_{ql}$, respectively, where U_{qr} and U_{ql} are real bounded spaces within which the right and left hand torques are defined.

3.3. External reward function

The external reward function is a vector of functions each rewarding distinct behaviors. In this case, the reward function vector consisted of three component functions for 1: hasty behavior, 2: Energy conscious behavior, 3: safety conscious behavior. Therefore the vector of functions were given by:

$$\boldsymbol{r}(t) = [r_1(t) \;\; r_2(t) \;\; r_3(t)]^T \tag{5}$$

$$r_1(t) = \frac{4}{1+100e^{(|u_r|+|u_l|)}} + r_{obs} + e^{-D} + P \tag{6}$$

where D is the distance to the target, P is a punishment given by $P = -10$ if ($|u_r| > 0.04$ or $|u_l| > 0.04$), and r_{obs} is the reward or penalty for avoiding or colliding with the obstacle, which is calculated by $r_{obs} = -100e^{-5|d_{obs}-0.5|}$ if close to the obstacle region, and $r_{obs} = 1$ if sufficient distance is kept; d_{obs} is the distance to the obstacle.

$$r_2(t) = 4(\dot{v}_{tar} + e^{-D}) + r_{obs} \tag{7}$$

where $\dot{v}_{tor}$ is the target reaching velocity.

$$r_3(t) = -100e^{-5|d_{da}-0.5|} + r_{obs} + e^{-D} \tag{8}$$

where d_{da} is the distance to the danger region. Inputs to the right and left wheels are given by:

$$u_r = u_{cr} + u_{qr}$$

$$u_l = u_{cl} + u_{ql} \tag{9}$$

Let the output of the Q-net for a given vector of environmental sensor information and a chosen control input be denoted by:

$$Q_v(t) = [Q_{1v}(t) \quad Q_{2v}(t) \quad Q_{3v}(t)]^{\mathrm{T}},$$

the maximum $Q_v(t)$ that can be obtained by changing the right and left wheel torques in U_{qr} and U_{ql} for a given environmental situation be denoted by $Q_{v\_max}(t)$ and the reward obtained from an external reward function be given by:

$$r(t) = [r_1(t) \quad r_2(t) \quad r_3(t)]^{\mathrm{T}}.$$

Then the following algorithm can be applied to obtain the optimum behaviors of the robot.

3.4. Reinforcement algorithm

The algorithm of the proposed reinforcement learning is expressed as follows:

Step 1: Initialize the weights of the Q-net, and set time t = 0.

Step 2: Sense the state of the robot and calculate u_r and u_l.

Step 3: Given the current control input and the environmental information, evaluate the Q-net and obtain a vector $Q_v(t)$.

Step 4: Run the robot for one sampling time duration and obtain a reward vector $r(t+1)$ from a set of external reward functions.

Step 5: For a given behavioral objective, i.e., energy optimization, hasty movement, or safe movement, Evaluate the Q-net and obtain $Q_{v\_max}(t+1)$, and the pair of control inputs u_{r_opt} and u_{l_opt} that renders $Q_{v\_max}(t+1)$.

Step 6: Calculate the temporal difference

$$\Delta(t+1) = [\Delta_1(t+1)\ \Delta_2(t+1)\ \Delta_3(t+1)]^T$$

given by

$$\Delta(t+1) = r(t+1) + \gamma Q_{v\_max}(t+1) - Q_v(t), \quad 0 < \gamma < 1 \tag{6}$$

Step 7: Use this $\Delta(t+1)$ vector to update the respective weight vectors of the Q-net.

Step 8: Set $u_r = u_{r\_opt} + N(0,\sigma)$ so that $= 1/(1 + e^{r_p(t+1)})$, where p is the counter of the behavior type that decides the control inputs at time $t+1$. Go to Step 3, and set time $t = t+1$;

Continue these steps until a predetermined level of performance is achieved by all the vectors of weights in the Q-net. Note, that a vector of Q values given by $Q_v(t)$ and reward values given by $r(t+1)$ are evaluated at any given time, eventhough only one behavior is executed at any given time. This ability of parallel learning while executing a single behavior is the main advantage of the proposed method. This results from the mechanism of cross-criticism found in the proposed method.

4. SIMULATION

To evaluate the effectiveness of the proposed learning method, computer simulation has been performed. In this simulation, the mobile robot is supposed to head toward the goal subjected to various performance criteria. There are one obstacle region, two slippery regions, and one danger region in the working environment as shown in Fig. 2. In this simulation, the robot is supposed to waste 20% of driving torque for the slip in the slippery regions, and waste 80% of driving torque for struggling to move in the danger regions. The dynamics of the mobile robot

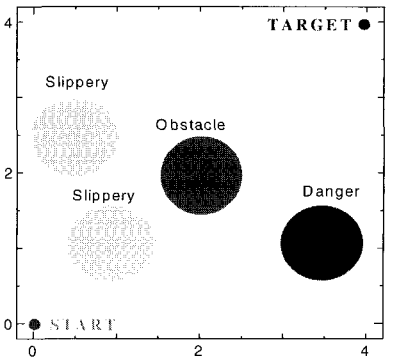


Fig. 2. Working environment of the mobile robot.

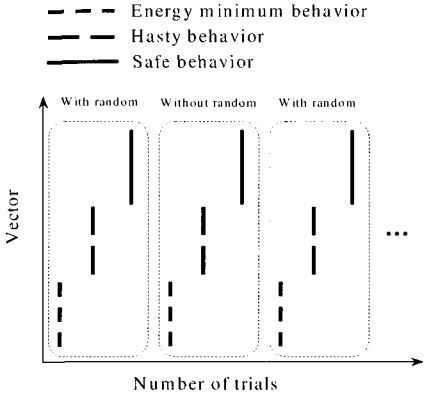


Fig. 3. Learning at each trial.

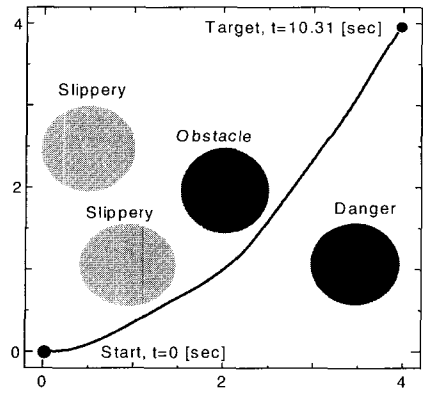
explained in Section 2 is taken into account. The energy minimum behavior, the hasty behavior, and the safe behavior are considered in this simulation, although other behaviors can be considered.

Although multi-dimensional learning is carried out in each trial, one representative behavior is chosen in turn from among the three evaluating behaviors (energy minimum behavior, hasty behavior, and safe behavior). The random behavior is generated in certain range during the learning at every other trial of each evaluating behavior as shown in Fig. 3.

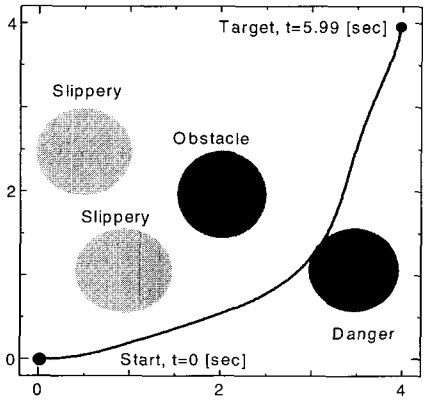
Fig. 4 and 5 show the simulation results after 1000 trials. The obtained energy minimum behavior, hasty behavior, and safe behavior are depicted in Fig. 4 (a), (b), and (c), respectively. The torque profiles of energy minimum behavior, hasty behavior, and safe behavior are shown in Fig. 5 (a), (b), and (c), respectively. One can see that the energy minimum behavior consumes less energy than the other behavior. In the hasty behavior, the robot quickly arrives at the target although a lot of energy is consumed. The safe behavior takes a lot of time to get to the target. These results show that the behavior of the robot can change depending on the situation.

5. CONCLUSIONS

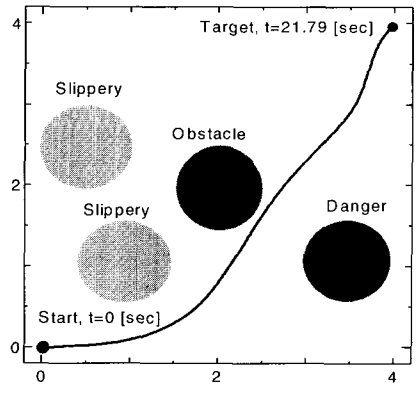
A novel multi-dimensional reinforcement learning method has been proposed and applied to Q-learning



(a) Energy minimum behavior

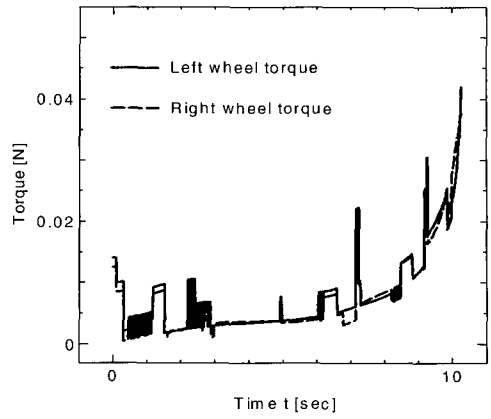


(b) Hasty behavior

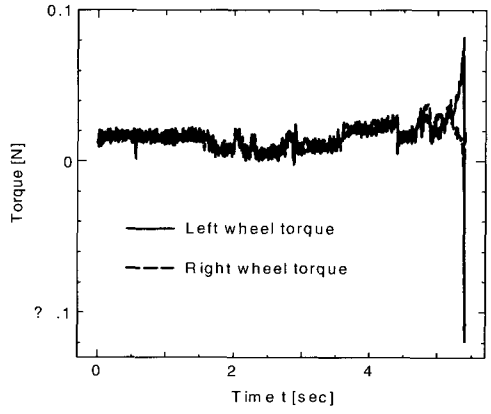


(c) Safe behavior

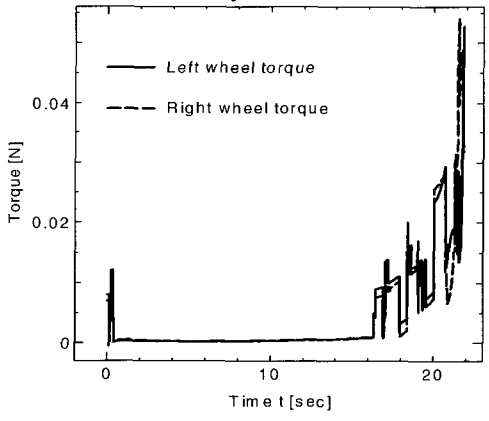
Fig. 4. Simulation results.



(a) Energy minimum behavior



(b) Hasty behavior



(c) Safe behavior

Fig. 5. Torque profiles.

in this study. A special neural network (Q-net) is proposed to realize *critic* networks. In the proposed Q-net, the weights and the output are represented by vectors, although these are usually represented by scalars. Each component of vectors is in charge of each item of the evaluation (*reward*). Consequently, each component of the weight vectors and the output vector of the Q-net is adjusted based on reward or punishment for each item of the evaluation after a certain behavior is performed. The novelty of the proposed method is that the algorithm facilitated parallel learning of all behaviors while executing a single behavior. This novel feature is expected to accelerate the learning speed of multi-dimensional reinforcement learning algorithms. This kind of a cross-criticism is expected to function in the human brain, though there is no biological evidence so far. Yet, this phenomenon is seen in human learning through social interaction, where one updates its internal models by observing the behaviors of others. In this method, the robot is able to change the optimal behavior according to the situation after the learning. Simulation results show the effectiveness of the proposed reinforcement learning.

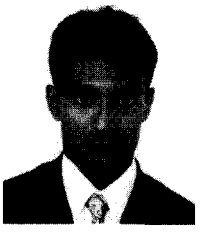
REFERENCES

[1] W. Schultz, P. Dayan, and P. R. Montague, "A

neural substrate of prediction and reward," *Science*, vol. 275, pp. 1593-1599, 1997.
[2] C. H. An, C. G. Atkeson, and J. M. Hollerbach, "Estimation of internal parameters of rigid body links of manipulators," *Artificial Intelligence Memo 887*, MIT Artificial Intelligence Laboratory, 1986.
[3] R. S. Sutton and A. G. Barto, *Reinforcement Learning*, MIT Press, 1998.
[4] C. J. C. H. Watkins, "Learning from delayed rewards," Ph. D. Dissertation, Cambridge University, 1989.
[5] M. Asada, S. Noda, S. Tawaratumida, and K. Hosoda, "Purposive behavior acquisition for a real robot by vision-based reinforcement learning," *Machine Learning*, vol. 23, pp. 279-303, 1996.
[6] F. Saito and T. Fukuda, "Learning architecture for real robotic systems Extension of connectionist q-learning for continuous robot control domain," *Proc. of IEEE International Conference on Robotics and Automation*, pp. 27-32, 1994.
[7] S. Mahadevan and J. Connell, "Automatic programming of behavior-based robots using reinforcement learning," *Proc. of 9th National Conf. on Artificial Intelligence*, pp. 768-773, 1991.
[8] L. J. Lin, "Reinforcement learning for robots using neural networks," Ph. D. Dissertation, Carnegie Mellon University, 1992.
[9] M. J. Mataric, "Interaction and intelligent behavior," Ph. D. Dissertation, MIT, 1994.
[10] V. Gullapalli, J. A. Franklin, and H. Benbrahim, "Acquiring robot skills via reinforcement learning," *IEEE Control Systems Magazine*, vol. 14, no. 1, pp. 13-24, 1994.
[11] H. K. Beom and H. S. Cho, "A sensor-based navigation for a mobile robot using fuzzy logic and reinforcement learning," *IEEE Trans. on Systems, Man, and Cybernetics*, vol. 25, pp. 464-477, 1995.
[12] Z. Kalmar, C. Szepesvari, and A. Lorincz, "Module-based reinforcement learning: experiments with a real robot," *Machine Learning*, vol 31, pp. 55-85, 1998.
[13] E. Uchibe and M. Asada, "Multiple reward criterion for cooperative behavior acquisition in a multiagent environment," *Proc. of IEEE International Conf. on Systems, Man, and Cybernetics*, pp. VI 710-VI 715, 1999.

Kazuo Kiguchi received the B.E. degree in mechanical engineering from Niigata University, Japan in 1986, the M.A.S. degree in mechanical engineering from the University of Ottawa, Canada in 1993, and the Ph.D. degree from Nagoya University, Japan in 1997.

He was a Research Engineer with Mazda Motor Co. between 1986-1989, and with MHI Aerospace Systems Co. between 1989-1991. He worked for the Dept. of Industrial and Systems Engineering, Niigata College of Technology, Japan between 1994-1999. He is currently an associate professor in the Dept. of Advanced Systems Control Engineering, Saga University, Japan. He received the J.F.Engelberger Best Paper Award at WAC2000. His research interests include biorobotics, intelligent robots, machine learning, soft computing, and application of robotics in medicine. He is a member of the Robotics Society of Japan, IEEE (SMC, R&A, EMB, IE, and CS Societies), the Japan Society of Mechanical Engineers, the Society of Instrument and Control Engineers, International Neural Network Society, Japan Neuroscience Society, and the Virtual Reality Society of Japan.



Thrishantha Nanayakkara was born in 1970, in Galle, Sri Lanka. He graduated with a first class honors B.Sc. degree in Electrical Engineering from the University of Moratuwa in 1996. In 1996 he won the Monbusho Fullbright scholarship to undertake postgraduate studies in Japan. He secured the M.Sc. degree in Electrical Engineering in 1998 and Ph.D. degree in Systems Control in 2001, both from Saga University, Japan. Since then, he is working as a postdoctoral research fellow in the Department of Biomedical Engineering, Johns Hopkins University, USA. His research interests are in, adaptive predictive control in the human motor system, machine learning, and skillful motion control of industrial robot manipulators based on biologically inspired systems control approaches. He is a member of IEEE.

Keigo Watanabe received the B.E. and M.E. degrees in Mechanical Engineering from the University of Tokushima in 1976 and 1978, respectively, and a D.E. degree in Aeronautical Engineering from Kyushu University in 1984. From 1980 to March 1985, he was a research associate in Kyushu University. From April 1985 to March 1990, he was an Associate Professor in the College of Engineering, Shizuoka University. From April 1990 to March 1993 he was an Associate Professor, and from April 1993 to March 1998 he was a full Professor in the Department of Mechanical Engineering at Saga University. From April 1998, he is now with the Department of Advanced Systems Control Engineering, Graduate School of Science and Engineering, Saga University. His research interests are in stochastic adaptive estimation and control, robust control, neural network control, fuzzy control, genetic algorithms and their applications to the machine intelligence and robotic control. He is a member of the Society of Instrument and Control Engineers, the Japan Society of Mechanical Engineers, the Japan Society for Precision Engineering, the Institute of Systems, Control and Information Engineers, the Japan Society for Aeronautical and Space Sciences, the Robotics Society of Japan, Japan Society for Fuzzy Theory and Systems, and IEEE.

Toshio Fukuda was graduated from Waseda University in 1971 and received the M.S and Dr. Eng. from the University of Tokyo in 1973 and 1977, respectively. Meanwhile, he studied at the graduate school of Yale University from 1973 to 1975. In 1977, he joined the National Mechanical Engineering Laboratory and became Visiting Research Fellow at the University of Stuttgart from 1979 to 1980. He joined the Science University of Tokyo in 1982, and then joined Nagoya University in 1989. Currently, he is Professor of Dept. of Micro System Engineering, Nagoya University, Japan, mainly engaging in the research fields of intelligent robotic system, cellular robotic system, mechatronics and micro robotics. He was awarded IEEE Fellow, SICE Fellow (1995), IEEE Eugene Mittlemann Award (1997), Banki Donat Medal form Polytechnic University of Budapest, Hungary (1997), Medal from City of Sartillo, Mexico (1998) and IEEE Millennium Medal (2000). He is the Vice President of IEEE IES (1990 - 1999), IEEE Neural Network Council Secretary (1992 -), IFSA Vice President (1997 -), IEEE Robotics and Automation Society President (1998 - 1999), current Editor-in-Chief, IEEE / ASME Transactions on Mechatronics (2000 -) and current IEEE Division X Director (2001-).