

Reinforcement Learning Algorithms In Humanoid Robotics

Duško Katić, Miomir Vukobratović
*Robotics Laboratory, Mihailo Pupin Institute
Belgrade, Serbia*

1. Introduction

Many aspects of modern life involve the use of intelligent machines capable of operating under dynamic interaction with their environment. In view of this, the field of biped locomotion is of special interest when human-like robots are concerned. Humanoid robots as anthropomorphic walking machines have been in operation for more than twenty years. Currently, research on the design and the humanoid robots are one of the most exciting and challenging topics in the field of robotics. The potential applications of this research area are very foremost in the middle and long term. Humanoid robots are expected to be servants and maintenance machines with the main task to assist human activities in our daily life and to replace humans in hazardous operations. It is as obvious as interesting that anthropomorphic biped robots are potentially capable to effectively move in all unstructured environments where humans do. There also raises strong anticipations that robots for the personal use will coexist with humans and provide supports such as the assistance for the housework, care of the aged and the physically handicapped. Consequently, humanoid robots have been treated as subjects of robotics researches such as a research tool for human science, an entertainment/mental-commit robot or an assistant/agent for humans in the human living environment.

Humanoid robot are autonomous systems capable of extracting information from their environments and using knowledge about the world and intelligence of their duties and proper governing capabilities. Intelligent humanoid robots should be autonomous to move safely in a meaningful and purposive manner, i.e. to accept high-level descriptions of tasks (specifying what the user wants to be done, rather than how to do it) and would execute them without further human intervention. Future humanoid robots are likely to have greater sensory capabilities, more intelligence for valid reasoning and decision making, higher levels of manual dexterity and adequate mobility as compared to humans. Naturally, the first approach to making humanoid robots more intelligent was the integration of sophisticated sensor systems as computer vision, tactile sensing, ultrasonic and sonar sensors, laser scanners and other smart sensors. However, today's sensor products are still very limited in interactivity and adaptability to changing environments. As the technology and algorithms for real-time 3D vision and tactile sensing improve, humanoid robots will be able to perform tasks that involve complex interaction with the environment (e.g. grasping and manipulating the objects). A major reason is that uncertainty and dynamic changes make the development of reliable artificial systems particularly challenging. On the other

hand, to design robots and systems that best adapt to their environment, the necessary research includes investigations in the field of mechanical robot design (intelligent mechanics), environment perception systems and embedded intelligent control that ought to cope with the task complexity, multi-objective decision making, large volume of perception data and substantial amount of heuristic information. Also, in the case when the robot performs in an unknown environment, the knowledge may not be sufficient. Hence, the robot has to adapt to the environment and to be capable of acquiring new knowledge through the process of learning. The robot learning is essentially concerned with equipping robots with the capacity of improving their behavior over time, based on their incoming experiences.

Although there has been a large number of the control methods used to solve the problem of humanoid robot walking, it is difficult to detect a specific trend. Classical robotics and also the more recent wave of humanoid and service robots still rely heavily on teleoperation or fixed behavior-based control with very little autonomous ability to react to the environment. Among the key missing elements is the ability to create control systems that can deal with a large movement repertoire, variable speeds, constraints and most importantly, uncertainty in the real-world environment in a fast, reactive manner. There are several intelligent paradigms that are capable of solving intelligent control problems in humanoid robotics. Connectionist theory (NN - neural networks), fuzzy logic (FL), and theory of evolutionary computation (GA - genetic algorithms), are of great importance in the development of intelligent humanoid robot control algorithms (Katić & Vukobratovic, 2003a; Katić & Vukobratovic, 2003b; Katić & Vukobratović, 2005). Due to their strong learning and cognitive abilities and good tolerance of uncertainty and imprecision, intelligent techniques have found wide applications in the area of advanced control of humanoid robots. Also, of great importance in the development of efficient algorithms are the hybrid techniques based on the integration of particular techniques such as neuro-fuzzy networks, neuro-genetic algorithms and fuzzy-genetic algorithms.

One approach of departing from teleoperation and manual 'hard coding' of behaviors is by learning from experience and creating appropriate adaptive control systems. A rather general approach to learning control is the framework of *Reinforcement Learning*, described in this chapter. Reinforcement learning offers one of the most general framework to take traditional robotics towards true autonomy and versatility.

Reinforcement learning typically requires an unambiguous representation of states and actions and the existence of a scalar reward function. For a given state, the most traditional of these implementations would take an action, observe a reward, update the value function, and select as the new control output the action with the highest expected value in each state (for a greedy policy evaluation). Updating of value function and controls is repeated until convergence of the value function and/or the policy. This procedure is usually summarized under "value update - policy improvement" iterations. The reinforcement learning paradigm described above has been successfully implemented for many well-defined, low dimensional and discrete problems (Sutton & Barto, 1998; Bertsekas & Tsitsiklis, 1996) and has also yielded a variety of impressive applications in rather complex domains in the last 20 years. Reinforcement learning is well suited to training mobile robots, in particular teaching a robot a new behavior (e.g. avoid obstacles) from scalar feedback. Robotics is a very challenging domain for reinforcement learning. However, various pitfalls have been encountered when trying to scale up these methods to high dimension, continuous control problems, as typically faced in the domain of humanoid robotics.

2. Control Problem in Humanoid Robotics

In spite of a significant progress and accomplishments achieved in the design of a hardware platform of humanoid robot and synthesis of advanced intelligent control of humanoid robots, a lot of work has still to be done in order to improve actuators, sensors, materials, energy accumulators, hardware, and control software that can be utilized to realize user-friendly humanoid robots. We are still in an initial stage when the understanding of the motor control principles and sensory integration subjacent to human walking is concerned. Having in mind the very high requirements to be met by humanoid robots, it is necessary to point out the need for increasing the number of degrees of freedom (DOFs) of their mechanical configuration and studying in detail some previously unconsidered phenomena pertaining to the stage of forming the corresponding dynamic models. Besides, one should emphasize the need for developing appropriate controller software that would be capable of meeting the most complex requirements of accurate trajectory tracking and maintaining dynamic balance during regular (stationary) gait in the presence of small perturbations, as well as preserving robot's posture in the case of large perturbations. Finally, one ought to point out that the problem of motion of humanoid robots is a very complex control task, especially when the real environment is taken into account, requiring as a minimum, its integration with the robot's dynamic model.

There are various sources of control problems and various tasks and criteria that must be solved and fulfilled in order to create valid walking and other functions of humanoid robots. Previous studies of biological nature, theoretical and computer simulation, have focussed on the structure and selection of control algorithms according to different criteria such as energy efficiency, energy distribution along the time cycle, stability, velocity, comfort, mobility, and environment impact. Nevertheless, in addition to these aspects, it is also necessary to consider some other issues: capability of mechanical implementation due to the physical limitations of joint actuators, coping with complex highly-nonlinear dynamics and uncertainties in the model-based approach, complex nature of periodic and rhythmic gait, inclusion of learning and adaptation capabilities, computation issues, etc.

The major problems associated with the analysis and control of bipedal systems are the high-order highly-coupled nonlinear dynamics and furthermore, the discrete changes in the dynamic phenomena due to the nature of the gait. Irrespective of the humanoid robot structure and complexity, the basic characteristic of all bipedal systems are: a) the DOF formed between the foot and the ground is unilateral and underactuated ; b) the gait repeatability (symmetry) and regular interchangeability of the number of legs that are simultaneously in contact with the ground. During the walk, two different situations arise in sequence: the statically stable double-support phase in which the mechanism is supported on both feet simultaneously, and statically unstable single-support phase when only one foot of the mechanism is in contact with the ground. Thus, the locomotion mechanism changes its structure during a single walking cycle from an open to a closed kinematic chain. Also, it is well known that through the process of running the robot can be most of the time in no-support phase. In this case, the control schemes that are successful for walking problem are not necessarily successful for the running problem. All the mentioned characteristics have to be taken into account in the synthesis of advanced control algorithms that accomplish stable, fast and reliable performance of humanoid robots.

The stability issues of humanoid robot walking are the crucial point in the process of control synthesis. In view of this humanoid walking robots can be classified in three different

categories. First category represents static walkers, whose motion is very slow so that the system's stability is completely described by the normal projection of the Centre of Gravity, which only depends on the joint's position. Second category represents dynamic walkers, biped robots with feet and actuated ankles. Postural stability of dynamic walkers depends on joint's velocities and acceleration too. These walkers are potentially able to move in a static way provided they have large enough feet and the motion is slow. The third category represents purely dynamic walkers, robots without feet. In this case the support polygon during the single-support phase is reduced to a point, so that static walking is not possible. In the walk with dynamic balance, the projected centre of mass is allowed outside of the area inscribed by the feet, and the walker may essentially fall during parts of the walking gait. The control problems of dynamic walking are more complicated than in walking with static balance, but dynamic walking patterns provide higher walking speed and greater efficiency, along with more versatile walking structures.

The rotational equilibrium of the foot is the major factor of postural instability with legged robots. The question has motivated the definition of several dynamic-based criteria for the evaluation and control of balance in biped locomotion. The most common criterion are the centre of pressure (CoP), the zero-moment point (ZMP) concept, that has gained widest acceptance and played a crucial role in solving the biped robot stability and periodic walking pattern synthesis (Vukobratović and Juričić, 1969). The ZMP is defined as the point on the ground about which the sum of all the moments of the active forces equals zero. If the ZMP is within the convex hull of all contact points between the foot and the ground, the biped robot can walk.

For a legged robot walking on complex terrain, such as a ground consisting of soft and hard uneven parts, a statically stable walking manner is recommended. However, in the cases of soft terrain, up and down slopes or unknown environment, the walking machine may lose its stability because of the position planning errors and unbalanced foot forces. Hence, position control alone is not sufficient for practical walking, position/force control being thus necessary. Foot force control can overcome these problems, so that foot force control is one of the ways to improve the terrain adaptability of walking robots. For example, in the direction normal to the ground, foot force has to be controlled to ensure firm foot support and uniform foot force distribution among all supporting legs; foot force in the tangential direction has to be monitored to avoid slippage.

A practical biped needs to be more like a human - capable of switching between different known gaits on familiar terrain and learning new gaits when presented with unknown terrain. In this sense, it seems essential to combine force control techniques with more advanced algorithms such as adaptive and learning strategies. Inherent walking patterns must be acquired through the development and refinement by repeated learning and practice as one of important properties of intelligent control of humanoid robots. Learning enables the robot to adapt to the changing conditions and is critical to achieving autonomous behaviour of the robot.

Many studies have given weight to biped walking which is based only on stability of the robot: steady-state walking, high-speed dynamic walking, jumping, and so on. A humanoid robot is however, a kind of integrated machine: a two-arm and two-leg mechanism. Hence, we must not only focus on the locomotion function but also on arm's function with this kind of machines; manipulation and handling being major functions of robot's arms.

When the ground conditions and stability constraint are satisfied, it is desirable to select a walking pattern that requires small torque and velocity of the joint actuators. Humanoid robots are inevitably restricted to a limited amount of energy supply. It would therefore be

advantageous to consider the minimum energy consumption, when cyclic movements like walking are involved. With this in mind, an important approach in research is to optimise simultaneously both the humanoid robot morphology and control, so that the walking behaviour is optimised instead of optimising walking behaviour for the given structure of humanoid robot. Optimum structures can be designed when the suitable components and locomotion for the robot are selected appropriately through evolution. It is well known that real-time generation of control algorithms based on highly-complex nonlinear model of humanoid robot commonly suffers from a large amount of computation. Hence, new time-efficient control methods need to be discovered to control humanoid robots in real time, to overcome the mentioned difficulty.

In summary, conventional control algorithms is based on a kinematics and dynamic modeling of the mechanism structure.(Vukobraović et al, 1990). This implies precise identification of intrinsic parameters of biped's robot mechanism, requires a high precise measurement of humanoid state variables and needs for precise evaluation of interaction forces between foot and ground. Moreover, these methods require a lot of computation together with some problems related to mathematical tractability, optimisation, limited extendability and limited biological plausibility. The second approach based on intelligent control techniques have a potential to overcome the mentioned constraints. In this case, it is not necessary to know perfectly the parameters and characteristics of humanoid mechanism. Also, these methods take advantage from off-line and on-line learning capabilities. This last point is very important because generally the learning ability allows increasing the autonomy of the bioed robot.

3. Reinforcement Learning Framework and Reinforcement Learning Algorithms in Humanoid Robotics

Recently, reinforcement learning has attracted attention as a learning method for studying movement planning and control (Sutton & Barto, 1998; Bertsekas & Tsitsiklis, 1996). Reinforcement learning is a kind of learning algorithm between supervised and unsupervised learning algorithms which is based on Markov decision process (MDP). Reinforcement learning concept is based on trial and error methodology and constant evaluation of performance in constant interaction with environment.

In many situations the success or failure of the controller is determined not only by one action but by a succession of actions. The learning algorithm must thus reward each action accordingly. This is referred to as the problem of delayed reward. There are two basic methods that are very successful in solving this problem, *TD learning* (Sutton & Barto, 1998) and *Q learning* (Watkins & Dayan, 1992). Both methods build a state space value function that determines how close each state is to success or failure. Whenever the controller outputs an action, the system moves from one state to another. The controller parameters are then updated in the direction that increases the state value function.

For the solution of large-scale MDPs or continuous state and action spaces, it's impossible for reinforcement learning agent to go through all the states and actions. In order to realize the optimal approximation for value functions of continuous states and actions respectively, therefore, learning agent must have generalization ability. In other words, such an agent should be able to utilize finite learning experience to acquire and express a good knowledge of a large-scale space effectively. How to design a function approximator with abilities of high generalization and computation efficiency has become a key problem for the research

field of reinforcement learning. Using prior knowledge about the desired motion can greatly simplify controller synthesis. Imitation-based learning or learning from demonstration allow for policy search to focus only on the areas of the search space that pertain to the task at hand. Both, model-based and model-free approaches exist to find optimal policies when agents are allowed to act for unlimited time. For physical agents, such as humanoid robots acting in the real world, it is much more difficult to gain experience. Hence, the exhaustive exploration of highdimensional state and action spaces is not feasible. For a physical robot, it is essential to learn from few trials in order to have some time left for exploitation.

The robot learning is essentially concerned with equipping robots with the capacity of improving their behavior over time, based on their incoming experiences. For instance, it could be advantageous to learn dynamics models, kinematic models, impact models, for model-based control techniques. Imitation learning could be employed for the teaching of gaits patterns, and reinforcement learning could help tuning parameters of the control policies in order to improve the performance with respect to given cost functions.

Dynamic bipedal walking is difficult to learn for a number of reasons. First, biped robots typically have many degrees of freedom, which can cause a combinatorial explosion for learning systems that attempt to optimize performance in every possible configuration of the robot. Second, details of the robot dynamics such as uncertainties in the ground contact and nonlinear friction in the joints must be only experimentally validated. Since it is only practical to run a small number of learning trials on the real robot, the learning algorithms must perform well after obtaining a very limited amount of data. Finally, learning algorithms for dynamic walking must deal with dynamic discontinuities caused by collisions with the ground and with the problem of delayed reward -torques applied at one time may have an effect on the performance many steps into the future.

In area of humanoid robotics, there are several approaches of reinforcement learning (Benbrahim & Franklin, 1997; Chew & Pratt, 2002; Li et al. , 1992; Mori et al., 2004; Morimoto et al., 2004; Nagasaka et al., 1999; Nakamura et al., 2003; Salatian et al., 1997; Zhou & Meng, 2000) with additional demands and requirements because high dimensionality of the control problem.

In the paper (Benbrahim & Franklin, 1997), it is shown how reinforcement learning is used within a modular control architecture to enable a biped robot to walk. The controller structure consists of central (CPG) and peripheral controllers. The learning architecture succeeded in dealing with the problems of large numbers of inputs, knowledge integration and task definition. The central controller controls the robot in nominal situations, and the peripheral controllers intervene only when they consider that the central controller's action contradicts their individual control policies (Figure 1). The action is generated by computing the average of the outputs of all controllers that intervene including the central controller. Each peripheral controller's role is to correct the central controller's mistakes and issue an evaluation of the general behaviour. The central controller then uses the average of all evaluations to learn a control policy that accommodates the requirements of as many peripheral controllers as possible. The central controller as well as some of the peripheral controllers in this study use adaptive CMAC neural networks. Because of modular nature, it is possible to use several neural networks with small numbers of inputs instead of one large neural network. This dramatically increases the learning speed and reduces the demand on memory and computing power. The architecture also allows easy incorporation of any knowledge by adding a peripheral controller that represents that knowledge. The CPG uses reinforcement learning in order to learn an optimal policy. The CMAC weights are updated using the reinforcement signals received from the peripheral controllers. Reinforcement learning is

well suited for this kind of application. The system can try random actions and choose those that yield good reinforcement. The reinforcement learning algorithm uses the actor-critic configuration. It searches the action space using a Stochastic Real Valued (SRV) unit at the output. The reinforcement signal is generated using TD Learning. The CMAC neural networks used in the biped's learning are pre-trained using a biped walking robot simulator and predefined simple walking gates. The size of the search domain is determined by the standard deviation of the Gaussian unit. If the standard deviation is too small, the system will have a very small search domain. This decreases the learning speed and increases the system's vulnerability to the local minima problem. If the factor is too large, the system's performance will not reach its maximum because there will always be a randomness even if the system has learned an optimal solution. It is in general safer to use a large factor than a small one. Even though this learning algorithms and architecture have successfully solved the problem of dynamic biped walking, there are many improvements that can be added to increase learning speed, robustness, and versatility. The performance must also be improved by dynamically setting the PID gains to deal with each specific situation.

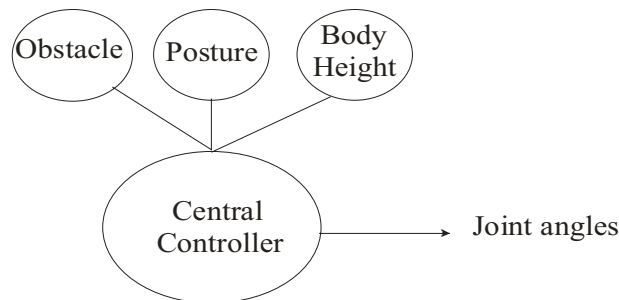


Fig.1. Controller Architecture Benbrahim & Franklin.

More recently, (Kun & Miller III, 1999) has developed a hierarchical controller that combines simple gait oscillators, classical feedback control techniques and neural network learning, and does not require detailed equations of the dynamics of walking. The emphasis is on the real-time control studies using an experimental ten-axis biped robot with foot force sensors. The neural network learning is achieved using CMAC controller, where CMAC neural networks were used essentially as context sensitive integral errors in the controller, and the control context being defined by the CMAC input vector. There are 3 different CMAC neural networks for humanoid posture control. The Front/Back Balance CMAC neural network was used to provide front/back balance during standing, swaying and walking. The training of this network is realized using data from foot sensors. The second CMAC neural network is used for Right/Left Balance, to predict the correct knee extension required to achieve sufficient lateral momentum for lifting the corresponding foot for the desired length of time. The training of this network is realized using temporal difference method based on the difference between the desired and real time of foot rising. The third CMAC network is used to learn kinematically consistent robot postures. In this case, training is also realized by data from foot sensors. The results indicated that the experimental biped was able to learn the closed-chain kinematics necessary to shift body weight side-to-side while maintaining good foot contact. Also, it was able to learn the quasi-static balance required to avoid falling forward or backward while shifting body weight side-to-side at different speeds. It was able to learn the dynamic balance in order to lift a foot off the floor for a desired length of time and different

initial conditions. There were, however, many limitations (limited step length, slow walking, no adaptation for left-right balance, no possibility of walking on sloping surfaces). The new dynamically balanced scheme for handling variable-speed gait was proposed based on the preplanned but adaptive motion sequences in combination with closed-loop reactive control. This allows the algorithm to improve the walking performance over consecutive steps using adaptation, and to react to small errors and disturbances using reactive control. New sensors (piezoresistive accelerometers and two solid-state rate gyroscopes) are mounted on the new UNH biped (Fig. 2).



Fig. 2. The UNH biped walking.

Training of neural networks is realized through the process of temporal difference learning using information about ZMP from robot foot sensors. The CMAC neural networks were first trained during repetitive foot-lift motion similar to marching in place. Then, training was carried out during the attempts at walking for increased step length and gait speeds.

The experimental results indicate that the UNH biped robot can walk with forward velocities in the range of 21 - 72 cm/min, with sideways leaning speed in the range of 3.6 - 12.5 cm/s. The main characteristic of this controller is the synthesis of the control signal without dynamic model of biped. The proposed controller could be used as a basis for similar controllers of more complex humanoid robots in the future research. However, this controller is not of a general nature, because it is suitable only for the proposed structure of biped robot and must be adapted for the bipeds with different structures. More research efforts are needed to simplify the controller structure, to increase the gait speed, and to ensure stability of dynamic walking.

The policy gradient method is one of the reinforcement learning methods successfully applied to learn biped walking on actual robots (Benbrahim & Franklin, 1997; Tedrake et al., 2004). However, this method requires hours to learn a walking controller, and approach in (Tedrake et al., 2004) requires a mechanically stable robot.

There are direct (model-free) and model-based reinforcement learning (Doya, 2000; Sutton & Barto, 1998). The direct approach to RL is to apply policy search directly without learning a model of the system. In principle, the model-based reinforcement learning is more data efficient than direct reinforcement learning. Also, it was concluded that model-based reinforcement learning finds better trajectories, plans and policies, and handles changing

goals more efficiently. On the other hand, reported that a model-based approach to reinforcement learning is able to accomplish given tasks much faster than without using knowledge of the environment.

The problem of biped gait synthesis using the reinforcement learning with fuzzy evaluative feedback is considered in paper (Zhoj & Meng, 2000). As first, initial gait from fuzzy rules is generated using human intuitive balancing scheme. Simulation studies showed that the fuzzy gait synthesizer can only roughly track the desired trajectory. A disadvantage of the proposed method is the lack of practical training data. In this case there are no numerical feedback teaching signal, only evaluative feedback signal exists (failure or success), exactly when the biped robot falls (or almost falls) down. Hence, it is a typical reinforcement learning problem. The dynamic balance knowledge is accumulated through reinforcement learning constantly improving the gait during walking. Exactly, it is fuzzy reinforcement learning that uses fuzzy critical signal. For human biped walk, it is typical to use linguistic critical signals such as "near-fall-down", "almost-success", "slower", "faster", etc. In this case, the gait synthesizer with reinforcement learning is based on a modified GARIC (Generalized Approximate Reasoning for Intelligent Control) method. This architecture of gait synthesizer consists of three components: action selection network (ASN), action evaluation network (AEN), and stochastic action modifier (SAM) (Fig. 3) The ASM maps a state vector into a recommended action using fuzzy inference. The training of ASN is achieved as with standard neural networks using error signal of external reinforcement. The AEN maps a state vector and a failure signal into a scalar score which indicates the state goodness. It is also used to produce internal reinforcement. The SAM uses both recommended action and internal reinforcement to produce a desired gait for the biped. The reinforcement signal is generated based on the difference between desired ZMP and real ZMP in the x-y plane. In all cases, this control structure includes on-line adaptation of gait synthesizer and local PID regulators. The approach is verified using simulation experiments. In the simulation studies, only even terrain for biped walking is considered, hence the approach should be verified for irregular and sloped terrain.

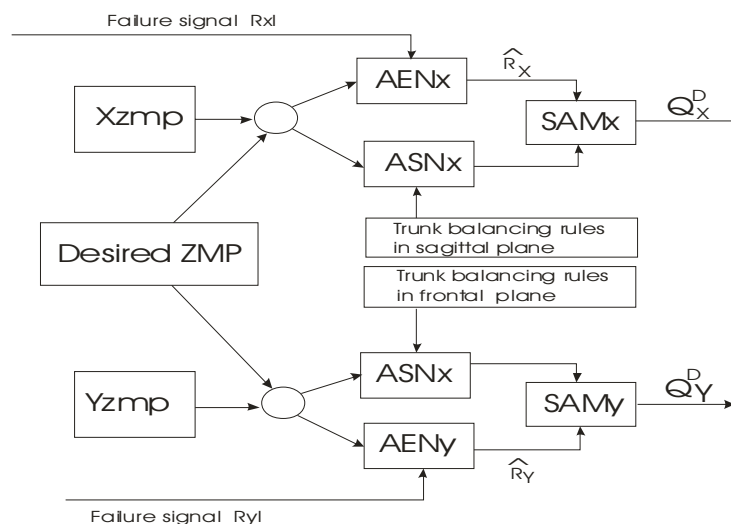


Fig. 3. The architecture of the reinforcement learning based gait synthesizer.

where X_{zmp}, Y_{zmp} are the ZMP coordinates; $\theta_{zmp}^d, \theta_{zmp}^d$ are the desired joint angles of the biped gait.

There are some research works that include the application of reinforcement learning control algorithms for passive or semi-passive biped walkers (Fig.4) (Tedrake et al., 2004; Morimoto et al., 2005; Schuitema et al., 2005).

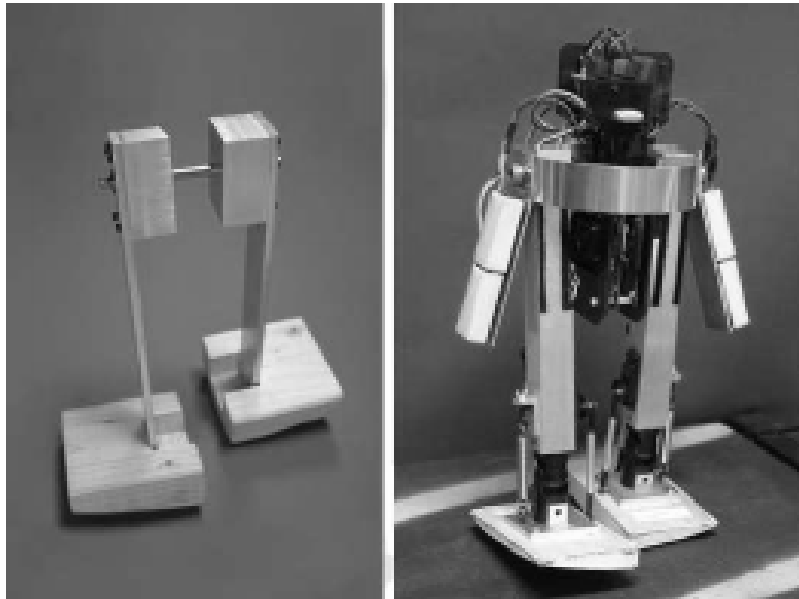


Fig. 4. Simple Passive Dynamic Walker.

In (Schuitema et al. , 2005), Reinforcement Learning passive walker is proposed as model free approach with fully driven optimization method. This approach is adaptive, in the sense that when the robot or its environment changes without notice, the controller can adapt until performance is again maximal. Optimization relatively easily towards several goals, such as: minimum cost of transport, largest average forward speed, or both. Statistical learning algorithm makes small changes in control parameters and uses correlations between control changes and system error changes. Stochasticity is added to deterministic feedback control policy. Gradient following algorithm based on temporal difference error was applied.

In paper (Morimoto et al., 2004) a model-based reinforcement learning algorithm for biped walking in which the robot learns to appropriately place the swing leg was proposed. This decision is based on a learned model of the Poincare map of the periodic walking pattern. The model maps from a state at the middle of a step and foot placement to a state at next middle of a step. Actor-Critic algorithms of reinforcement learning has a great potential in control of biped robots. For a example, a general algorithm for estimating the natural gradient, the Natural Actor-Critic algorithm, is introduced in paper (Peters et al., 2003). This algorithm converges to the nearest local minimum of the cost function with respect to the Fisher information metric under suitable conditions. It offers

a promising route for the development of reinforcement learning for truly high-dimensionally continuous state-action systems. In paper (Tedrake et al., 2004) a learning system which is able to quickly and reliably acquire a robust feedback control policy for 3D dynamic walking from a blank-slate using only trials implemented on physical robot. The robot begins walking within a minute and learning converges in approximately 20 minutes. This success can be attributed to the mechanics of our robot, which are modelled after a passive dynamic walker, and to a dramatic reduction in the dimensionality of the learning problem. The reduction of the dimensionality was realized by designing a robot with only 6 internal degrees of freedom and 4 actuators, by decomposing the control system in the frontal and sagittal planes, and by formulating the learning problem on the discrete return map dynamics. A stochastic policy gradient algorithm to this reduced problem was applied with decreasing the variance of the update using a state-based estimate of the expected cost. This optimized learning system works quickly enough that the robot is able to continually adapt to the terrain as it walks. The learning on robot is performed by a policy gradient reinforcement learning algorithm (Baxter & Bartlett, 2001; Kimura & Kobayashi, 1998; Sutton et al., 2000). Some researchers (Kamio & Iba, 2005) were efficiently applied hybrid version of reinforcement learning structures, integrating genetic programming and Q-Learning method on real humanoid robot.

4. Hybrid Reinforcement Learning Control Algorithms for Biped Walking

The new integrated hybrid dynamic control structure for the humanoid robots will be proposed, using the model of robot mechanism. Our approach consists in departing from complete conventional control techniques by using hybrid control strategy based on model-based approach and learning by experience and creating the appropriate adaptive control systems. Hence, the first part of control algorithm represents some kind of computed torque control method as basic dynamic control method, while the second part of algorithm is reinforcement learning architecture for dynamic compensation of ZMP (Zero-Moment-Point) error.

In the synthesis of reinforcement learning structure, two algorithms will be shown, that are very successful in solving biped walking problem: adaptive heuristic approach (AHC) approach, and approach based on Q learning. To solve reinforcement learning problem, the most popular approach is temporal difference (TD) method (Sutton & Barto, 1998). Two TD-based reinforcement learning approaches have been proposed: The adaptive heuristic critic (AHC) (Barto et al., 1983) and Q-learning (Watkins & Dayan, 1992). In AHC, there are two separate networks: An action network and an evaluation network. Based on the AHC, In (Berenji & Khedkar, 1992), a generalized approximate reasoning-based intelligent control (GARIC) is proposed, in which a two-layer feedforward neural network is used as an action evaluation network and a fuzzy inference network is used as an action selection network. The GARIC provides generalization ability in the input space and extends the AHC algorithm to include the prior control knowledge of human operators. One drawback of these actor-critic architectures is that they usually suffer from the local minimum problem in network learning due to the use of gradient descent learning method.

Besides the aforementioned AHC algorithm-based learning architecture, more and more advances are being dedicated to learning schemes based on Q-learning. Q-learning collapses the two measures used by actor/critic algorithms in AHC into one measure referred to as

the Q-value. It may be considered as a compact version of the AHC, and is simpler in implementation. Some Q-learning based reinforcement learning structures have also been proposed (Glorennec & Jouffe, 1997; Jouffe, 1998; Berenji, 1996).. In (Berenji & Jouffe, 1997), a dynamic fuzzy Q-learning is proposed for fuzzy inference system design. In this method, the consequent parts of fuzzy rules are randomly generated and the best rule set is selected based on its corresponding Q-value. The problem in this approach is that if the optimal solution is not present in the randomly generated set, then the performance may be poor. In (Jouffe, 1998), fuzzy Q-learning is applied to select the consequent action values of a fuzzy inference system. For these methods, the consequent value is selected from a predefined value set which is kept unchanged during learning, and if an improper value set is assigned, then the algorithm may fail. In (Berenji, 1996), a GARIC-Q method is proposed. This method works at two levels, the local and the top levels. At the local level, a society of agents (fuzzy networks) is created, with each learning and operating based on GARIC. While at the top level, fuzzy Q-learning is used to select the best agent at each particular time. In contrast to the aforementioned fuzzy Q-learning methods, in GARIC-Q, the consequent parts of each fuzzy network are tunable and are based on AHC algorithm. Since the learning is based on gradient descent algorithm, it may be slow and may suffer the local optimum problem.

4.1 Model of the robot's mechanism

The mechanism possesses 38 DOFs. Taking into account dynamic coupling between particular parts (branches) of the mechanism chain, a relation that describes the overall dynamic model of the locomotion mechanism in a vector form:

$$P + J^T(q)F = H(q)\ddot{q} + h(q, \dot{q}) \quad (1)$$

where: $P \in R^{n \times 1}$ is the vector of driving torques at the humanoid robot joints; $F \in R^{n \times 1}$ is the vector of external forces and moments acting at the particular points of the mechanism; $H \in R^{n \times n}$ is the square matrix that describes 'full' inertia matrix of the mechanism; $h \in R^{n \times 1}$ is the vector of gravitational, centrifugal and Coriolis moments acting at n mechanism joints; $J \in R^{n \times n}$ is the corresponding Jacobian matrix of the system; $n=38$, is the total number of DOFs; $q \in R^{n \times 1}$ is the vector of internal coordinates; $\dot{q} \in R^{n \times 1}$ is the vector of internal velocities.

4.2 Definition of control criteria

In the control synthesis for biped mechanism, it is necessary to satisfy certain natural principles. The control must to satisfy the following two most important criteria: (i) accuracy of tracking the desired trajectories of the mechanism joints (ii) maintenance of dynamic balance of the mechanism during the motion. Fulfillment of criterion (i) enables the realization of a desired mode of motion, walk repeatability and avoidance of potential obstacles. To satisfy criterion (ii) it means to have a dynamically balanced walk.

4.3. Gait phases and indicator of dynamic balance

The robot's bipedal gait consists of several phases that are periodically repeated. Hence, depending on whether the system is supported on one or both legs, two macro-phases can be distinguished: (i) single-support phase (SSP) and (ii) double-support phase (DSP). Double-support phase has two micro-phases: (i) weight acceptance phase (WAP) or heel strike, and (ii) weight support phase (WSP). Fig. 5 illustrates these gait phases, with the

projections of the contours of the right (RF) and left (LF) robot foot on the ground surface, whereby the shaded areas represent the zones of direct contact with the ground surface.

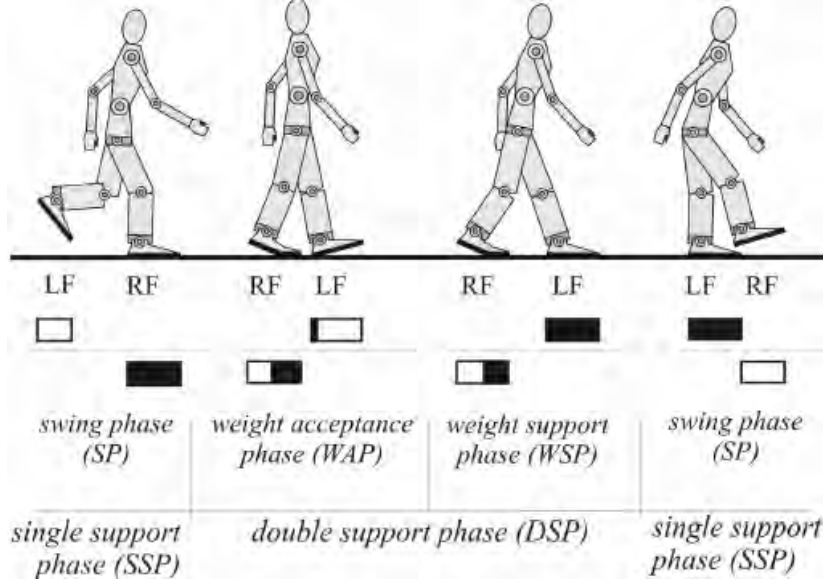


Fig. 5. Phases of biped gait.

The indicator of the degree of dynamic balance is the ZMP, i.e. its relative position with respect to the footprint of the supporting foot of the locomotion mechanism. The ZMP is defined (Vuobratović & Juričić, 1969) as the specific point under the robotic mechanism foot . at which the effect of all the forces acting on the mechanism chain can be replaced by a unique force and all the rotation moments about the x and y axes are equal zero. Figs 6a and 6b show details related to the determination of ZMP position and its motion in a dynamically balanced gait. The ZMP position is calculated based on measuring reaction forces $F_i, i=1, \dots, 4$ under the robot foot. Force sensors are usually placed on the foot sole in the arrangement shown in Fig. 6 a. Sensors' positions are defined by the geometric quantities l_1, l_2 and l_3 . If the point 0_{zmp} is assumed as the nominal ZMP position (Fig. 6a), then one can use the following equations to determine the relative ZMP position with respect to its nominal:

$$\Delta M_x^{(zmp)} = \frac{l_3}{2} \left[(F_2 + F_4) - (F_2^0 + F_4^0) \right] - \frac{l_3}{2} \left[(F_1 + F_3) - (F_1^0 + F_3^0) \right],$$

$$\Delta M_y^{(zmp)} = l_2 \left[(F_3 + F_4) - (F_3^0 + F_4^0) \right] - l_1 \left[(F_1 + F_2) - (F_1^0 + F_2^0) \right],$$

$$F_r^{(z)} = \sum_{i=1}^4 F_i, \quad \Delta x^{(zmp)} = \frac{-\Delta M_y^{(zmp)}}{F_r^{(z)}}, \quad \Delta y^{(zmp)} = \frac{\Delta M_x^{(zmp)}}{F_r^{(z)}}$$

where F_i and $F_i^0, i=1, \dots, 4$, are the measured and nominal values of the ground reaction force; $\Delta M_x^{(zmp)}$ and $\Delta M_y^{(zmp)}$ are deviations of the moments of ground reaction forces around

the axes passed through the 0_{zmp} ; $F_r^{(z)}$ is the resultant force of ground reaction in the vertical z-direction, while $\Delta x^{(zmp)}$ and $\Delta y^{(zmp)}$ are the displacements of ZMP position from its nominal 0_{zmp} . The deviations $\Delta x^{(zmp)}$ and $\Delta y^{(zmp)}$ of the ZMP position from its nominal position in x- and y-direction are calculated from the previous relation. The instantaneous position of ZMP is the best indicator of dynamic balance of the robot mechanism. In Fig. 6b are illustrated certain areas (Z_0, Z_1 and Z_2), the so-called safe zones of dynamic balance of the locomotion mechanism. The ZMP position inside these "safety areas" ensures a dynamically balanced gait of the mechanism whereas its position outside these zones indicates the state of losing the balance of the overall mechanism, and the possibility of its overturning. The quality of robot balance control can be measured by the success of keeping the ZMP trajectory within the mechanism support polygon (Fig. 6b).

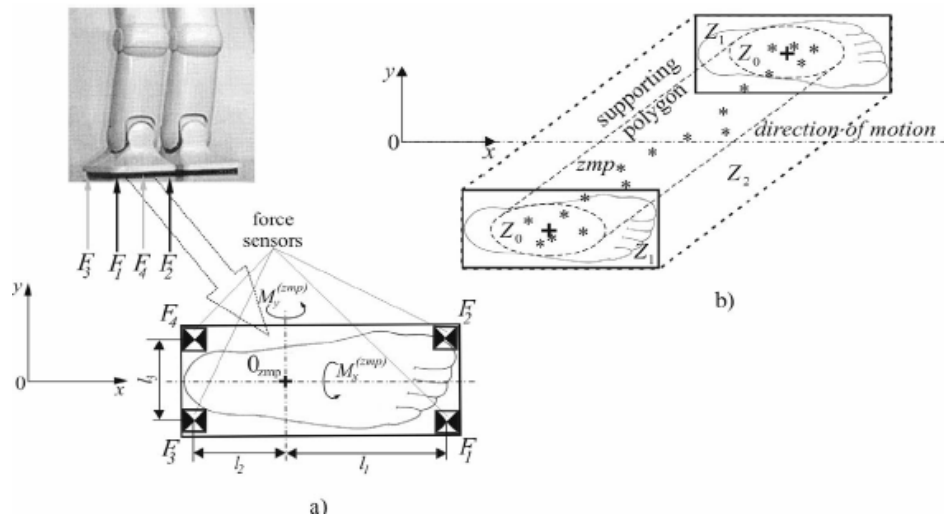


Fig. 6. Zero-Moment Point: a) Legs of "Toyota " humanoid robot ; General arrangement of force sensors in determining the ZMP position; b) Zones of possible positions of ZMP when the robot is in the state of dynamic balance.

4.4 Hybrid intelligent control algorithm with AHC reinforcement structure

Biped locomotion mechanism represents a nonlinear multivariable system with several inputs and several outputs. Having in mind the control criteria, it is necessary to control the following variables: positions and velocities of the robot joints and ZMP position. In accordance with the control task, we propose the application of the hybrid intelligent control algorithm based on the dynamic model of humanoid system. Here we assume the following: (i) the model (1) describes sufficiently well the behavior of the system; (ii) desired (nominal) trajectory of the mechanism performing a dynamically balanced gait is known. (iii) geometric and dynamic parameters of the mechanism and driving units are known and constant. These assumptions can be taken as conditionally valid, the rationale being as follows: As the system elements are rigid bodies of unchangeable geometrical shape, the parameters of the mechanism can be determined with a satisfactory accuracy.

Based on the above assumptions, in Fig. 7 a block-diagram of the intelligent controller for biped locomotion mechanism is proposed. It involves two feedback loops: (i) basic dynamic controller for trajectory tracking, (ii) intelligent reaction feedback at the ZMP based on AHC reinforcement learning structure. The synthesized dynamic controller was designed on the basis of the centralized model. The vector of driving moments \hat{P} represents the sum of the driving moments \hat{P}_1, \hat{P}_2 . The torques \hat{P}_1 are determined so to ensure precise tracking of the robot's position and velocity in the space of joints coordinates. The driving torques \hat{P}_2 are calculated with the aim of correcting the current ZMP position with respect to its nominal. The vector \hat{P} of driving torques represents the output control vector.

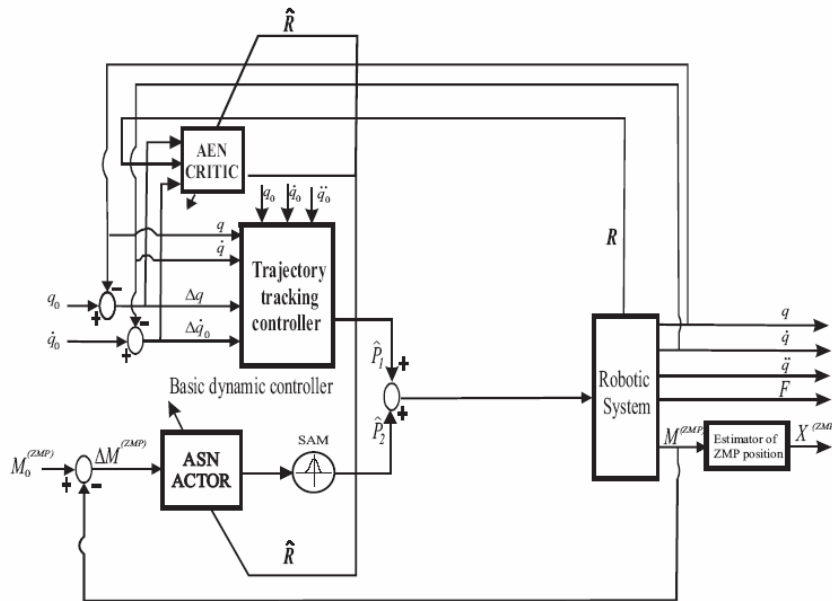


Fig. 7 Hybrid Controller based on Actor-Critic Method for trajectory tracking.

4.5 Basic Dynamic Controller

The proposed dynamic control law has the following form:

$$\hat{P} = \hat{H}(q)[\ddot{q}_0 + K_v(\dot{q} - \dot{q}_0) + K_p(q - q_0)] + \hat{h}(q, \dot{q}) \quad (2)$$

where \hat{H}, \hat{h} and \hat{J} are the corresponding estimated values of the inertia matrix, vector of gravitational, centrifugal and Coriolis forces and moments and Jacobian matrix from the model (1). The matrices $K_p \in R^{n \times n}$ and $K_v \in R^{n \times n}$ are the corresponding matrices of position and velocity gains of the controller. The gain matrices K_p and K_v can be chosen in the diagonal form by which the system is decoupled into n independent subsystems. This control model is based on centralized dynamic model of biped mechanism.

4.6 Compensator of dynamic reactions based on reinforcement learning structure

In the sense of mechanics, locomotion mechanism represents an inverted multi link pendulum. In the presence of elasticity in the system and external environment factors, the mechanism's motion causes dynamic reactions at the robot supporting foot. Thus, the state of dynamic balance of the locomotion mechanism changes accordingly. For this reason it is essential to introduce dynamic reaction feedback at ZMP in the control synthesis. There are relationship between the deviations of ZMP positions $(\Delta x^{(zmp)}, \Delta y^{(zmp)})$ from its nominal position 0_{zmp} in the motion directions x and y and the corresponding dynamic reactions $M_x^{(zmp)}$ and $M_y^{(zmp)}$ acting about the mutually orthogonal axes that pass through the point 0_{zmp} . $M_x^{(zmp)} \in R^{1 \times 1}$ and $M_y^{(zmp)} \in R^{1 \times 1}$ represent the moments that tend to overturn the robotic mechanism. $M_0^{(zmp)} \in R^{2 \times 1}$ and $M^{(zmp)} \in R^{2 \times 1}$ are the vectors of nominal and measured values of the moments of dynamic reaction around the axes that pass through the ZMP (Fig. 6a). Nominal values of dynamic reactions, for the nominal robot trajectory, are determined off-line from the mechanism model and the relation for calculation of ZMP; $\Delta M^{(zmp)} \in R^{2 \times 1}$ is the vector of deviation of the actual dynamic reactions from their nominal values; $P_{dr} \in R^{2 \times 1}$ is the vector of control torques, ensuring the state of dynamic balance.

The control torques P_{dr} has to be displaced to the some joints of the mechanism chain. Since the vector of deviation of dynamic reactions $\Delta M^{(zmp)}$ has two components about the mutually orthogonal axes x and y , at least two different active joints have to be used to compensate for these dynamic reactions. Considering the model of locomotion mechanism, the compensation was carried out using the following mechanism joints: 9, 14, 18, 21 and 25 to compensate for the dynamic reactions about the x -axis, and 7, 13, 17, 20 and 24 to compensate for the moments about the y -axis. Thus, the joints of ankle, hip and waist were taken into consideration. Finally, the vector of compensation torques \hat{P}_2 was calculated on the basis of the vector of the moments P_{dr} in the case when compensation of ground dynamic reactions is performed using all six proposed joints, using the following relation

$$\hat{P}_2(9) = \hat{P}_2(14) = \hat{P}_2(18) = \hat{P}_2(21) = \hat{P}_2(25) = 1/5 P_{dr} \quad (3)$$

$$\hat{P}_2(7) = \hat{P}_2(13) = \hat{P}_2(17) = \hat{P}_2(20) = \hat{P}_2(24) = 1/5 P_{dr} \quad (4)$$

In nature, biological systems use simultaneously a large number of joints for correcting their balance. In this work, for the purpose of verifying the control algorithm, the choice was restricted to the mentioned ten joints: 7, 9, 13, 14, 17, 18, 20, 21, 24, and 25. Compensation of ground dynamic reactions is always carried out at the supporting leg when the locomotion mechanism is in the swing phase, whereas in the double-support phase it is necessary to engage the corresponding pairs of joints (ankle, hip, waist) of both legs.

On the basis of the above the fuzzy reinforcement control algorithm is defined with respect to the dynamic reaction of the support at ZMP.

4.7. Reinforcement Actor-Critic Learning Structure

This subsection describes the learning architecture that was developed to enable biped walking. A powerful learning architecture should be able to take advantage of any available

knowledge. The proposed reinforcement learning structure is based on Actor-Critic Methods (Sutton & Barto, 1998).

Actor-Critic methods are *temporal difference (TD)* methods, that have a separate memory structure to explicitly represent the control policy independent of the value function. In this case, control policy represents policy structure known as **Actor** with aim to select the best control actions. Exactly, the control policy in this case, represents the set of control algorithms with different control parameters. The input to control policy is state of the system, while the output is control action (signal). It searches the action space using a Stochastic Real Valued (SRV) unit at the output. The unit's action uses a Gaussian random number generator. The estimated value function represents a **Critic**, because it criticizes the control actions made by the actor. Typically, the critic is a state-value function which takes the form of TD error necessary for learning. TD error depends also from reward signal, obtained from environment as result of control action. The TD Error can be scalar or fuzzy signal that drives all learning in both actor and critic.

Practically, in proposed humanoid robot control design, it is synthesized the new modified version of GARIC reinforcement learning structure (Berenji & Khedkar, 1992). The reinforcement control algorithm is defined with respect to the dynamic reaction of the support at ZMP, not with respect to the state of the system. In this case external reinforcement signal (reward) R is defined according to values of ZMP error.

Proposed learning structure consists from two networks: AEN(Action Evaluation Network) - CRITIC and ASN(Action Selection Network) - ACTOR. AEN network maps position and velocity tracking errors and external reinforcement signal R in scalar or fuzzy value which represent the quality of given control task. The output scalar value of AEN is important for calculation of internal reinforcement signal. \hat{R} AEN constantly estimate internal reinforcement based on tracking errors and value of reward. AEN is standard 2-layer feedforward neural network (perceptron) with one hidden layer. The activation function in hidden layer is sigmoid, while in the output layer there are only one neuron with linear function. The input layer has a bias neuron. The output scalar value v is calculated based on product of set C of weighting factors and values of neurons in hidden later plus product of set A of weighting factors and input values and bias member. There are also one more set of weighting factors B between input layer and hidden layer. The number of neurons on hidden later is determined as 5. Exactly, the output v can be represented by the following equation:

$$v = \sum_i B_i \Delta M_i^{(zmp)} + \sum_j C_j f\left(\sum_{ji} A_i \Delta M_i^{(zmp)}\right) \quad (5)$$

where f is sigmoid function.

The most important function of AEN is evaluation of TD error, exactly internal reinforcement. The internal reinforcement is defined as TD(0) error defined by the following equation:

$$\hat{R}(t+1) = 0, \quad \text{begining state} \quad (6)$$

$$\hat{R}(t+1) = R(t) - v(t), \quad \text{failure state} \quad (7)$$

$$\hat{R}(t+1) = R(t) + \gamma v(t+1) - v(t), \quad \text{otherwise} \quad (8)$$

where γ is a discount coefficient between 0 and 1 (in this case γ is set to 0.9).

ASN (action selection network) maps the deviation of dynamic reactions $\Delta M^{(zmp)} \in R^{2 \times 1}$ in recommended control torque. The structure of ASN is represented by The ANFIS - Sugeno-type adaptive neural fuzzy inference systems. There are five layers: input layer, antecedent part with fuzzification, rule layer, consequent layer, output layer with defuzzification. This system is based on fuzzy rule base generated by expert knowledge with 25 rules. The partition of input variables (deviation of dynamic reactions) are defined by 5 linguistic variables: *NEGATIVE BIG*, *NEGATIVE SMALL*, *ZERO*, *POSITIVE SMALL* and *POSITIVE BIG*. The member functions is chosen as triangular forms.

SAM (Stochastic action modifier) uses the recommended control torque from ASN and internal reinforcement signal to produce final commanded control torque P_{dr} . It is defined by Gaussian random function where recommended control torque is mean, while standard deviation is defined by following equation:

$$\sigma(\hat{R}(t+1)) = 1 - \exp(-|\hat{R}(t+1)|) \quad (9)$$

Once the system has learned an optimal policy, the standard deviation of the Gaussian converges toward zero, thus eliminating the randomness of the output.

The learning process for AEN (tuning of three set of weighting factors A, B, C) is accomplished by step changes calculated by products of internal reinforcement, learning constant and appropriate input values from previous layers, i.e. according to following equations:

$$B_i(t+1) = B_i(t) + \beta \hat{R}(t+1) \Delta M_i^{(zmp)}(t) \quad (10)$$

$$C_j(t+1) = C_j(t) + \beta \hat{R}(t+1) f\left(\sum_{ji} A_{ij}(t) \Delta M_i^{(zmp)}(t)\right) \quad (11)$$

$$A_{ij}(t+1) = A_{ij}(t) + \beta \hat{R}(t+1) f\left(\sum_{ji} A_{ij}(t) \Delta M_i^{(zmp)}(t)\right) (1 - f\left(\sum_{ji} A_{ij}(t) \Delta M_i^{(zmp)}(t)\right)) \text{sgn}(C_j \Delta M_j^{(zmp)}(t)) \quad (12)$$

where β is learning constant. The learning process for ASN (tuning of antecedent and consequent layers of ANFIS) is accomplished by gradient step changes (back propagation algorithms) defined by scalar output values of AEN, internal reinforcement signal, learning constants and current recommended control torques.

In our research, the precondition part of ANFIS is online constructed by special clustering approach. The general grid type partition algorithms perform either with training data collected in advance or cluster number assigned a priori. In the reinforcement learning problems, the data are generated only when online learning is performed. For this reason, a new clustering algorithm based on Euclidean Distance measure, with the abilities of online learning and automatic generation of number of rules is used.

4.8 Hybrid intelligent control algorithm with Q reinforcement structure

From the perspective of ANFIS Q-learning, we propose a method, as combination of automatic precondition part construction and automatic determination of the consequent

parts of a ANFIS system. In application, this method enables us to deal with continuous state and action spaces. It helps to solve *the curse of dimensionality* encountered in high-dimensional continuous state space and provides smooth control actions. Q-learning is a widely-used reinforcement learning method for an agent to acquire optimal policy. In this learning, an agent tries an action, $a(t)$, at a particular state, $x(t)$, and then evaluates its consequences in terms of the immediate reward $R(t)$. To estimate the discounted cumulative reinforcement for taking actions from given states, an evaluation function, the Q-function, is used. The Q-function is a mapping from state-action pairs to predict return and its output for state x and action a is denoted by the Q-value, $Q(x, a)$. Based on this Q-value, at time t , the agent selects an action $a(t)$. The action is applied to the environment, causing a state transition from $x(t)$ to $x(t+1)$, and a reward $R(t)$ is received. Then, the Q-function is learned through *incremental dynamic programming*. The Q-value of each state/action pair is updated by

$$Q(x(t), a(t)) = Q(x(t), a(t)) + \alpha(R(t) + \gamma Q^*(x(t+1)) - Q(x(t), a(t)))$$

$$Q^*(x(t+1)) = \max_{b \in A(x(t+1))} Q(x(t+1), b) \tag{13}$$

where $A(x(t+1))$ is the set of possible actions in state ; α is the learning rate; γ is the discount rate.

Based on the above facts, in Fig. 8 a block-diagram of the intelligent controller for biped locomotion mechanism is proposed. It involves two feedback loops: (i) basic dynamic controller for trajectory tracking, (ii) intelligent reaction feedback at the ZMP based on Q-reinforcement learning structure

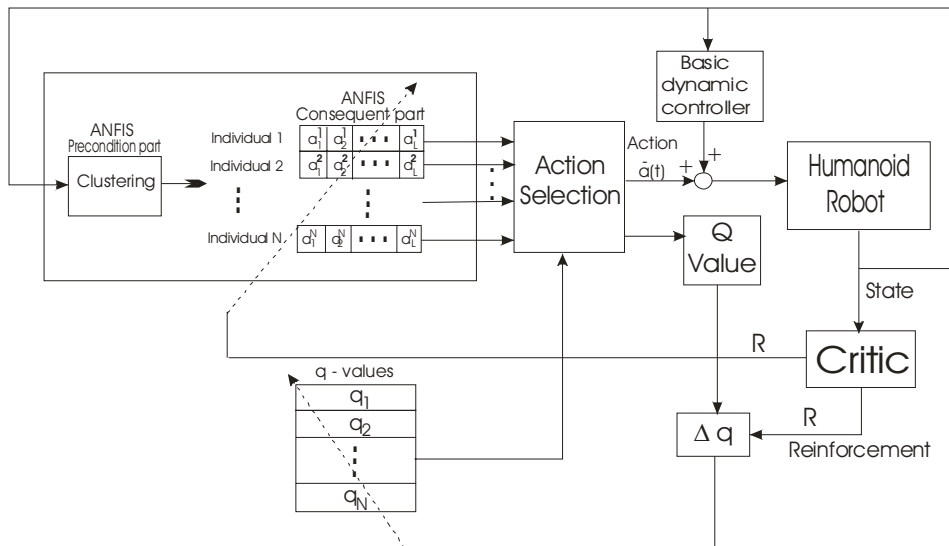


Fig. 8. Hybrid Controller based on Q-Learning Method for trajectory tracking.

4.9. Reinforcement Q-Learning Structure

The precondition part of the ANFIS system is constructed automatically by the clustering algorithm. Then, the consequent part of this newly generated rule is designed. In this methods, a population of candidate consequent parts is generated. Each individual in the population represents the consequent part of a fuzzy system. Since we want to solve reinforcement learning problems, a mechanism to evaluate the performance of each individual is required. To achieve this goal, each individual has a corresponding Q-value. The objective of the Q-value is to evaluate the action recommended by the individual. A higher Q-value means a higher reward that will be achieved. Based on the accompanying Q-value of each individual, at each time step, one of the individuals is selected. With the selected individual (consequent part), the fuzzy system evaluates an action and a corresponding system Q-value. This action is then applied to the humanoid robot as part of hybrid control algorithm with a reinforcement returned. Based on this reward, the Q-value of each individual is updated based on temporal difference algorithm. The parameters of consequent part of ANFIS is also updated based on back propagation algorithm and value of reinforcement. The previous process is repeatedly executed until success.

Each rule in the fuzzy system is presented in the following form:

$$\text{Rule: If } x_1(t) \text{ is } A_{i1} \text{ And } \dots x_n(t) \text{ is } A_{in} \text{ Then } a(t) \text{ is } a_i(t) \quad (14)$$

Where $x(t)$ is the input value, $\bar{a}(t)$ is the output action value, A is a fuzzy set and $a(t)$ is a recommended action is a fuzzy singleton. If we use a Gaussian membership function as fuzzy set, then for given an input data $x = (x_1, x_2, \dots, x_n)$, the firing strength $\Phi_i(x)$ of rule i is calculated by

$$\Phi_i(x) = \exp \left\{ -\sum_{j=1}^n \left(\frac{x_j - m_{ij}}{\sigma_{ij}} \right)^2 \right\} \quad (15)$$

where m_{ij} and σ_{ij} denote the mean and width of the fuzzy set.

Suppose a fuzzy system consists of L rules. By weighted average defuzzification method, the output of the system is calculated by

$$\bar{a} = \frac{\sum_{i=1}^L \Phi_i(x) a_i}{\sum_{i=1}^L \Phi_i(x)} \quad (16)$$

A population of recommended actions, involving individuals is created. Each individual in the population represents the consequent values, a_1, \dots, a_L of a fuzzy system. The Q-value used to predict the performance of individual i is denoted as q_i . An individual with a higher Q-value means a higher discounted cumulative reinforcement will be obtained by this individual. At each time step, one over these N individuals is selected as the consequent part of a fuzzy system based on their corresponding Q-values. This fuzzy system with competing consequences may be written as

If (**Precondition Part**) Then (**Consequence**) is

Individual 1 (a_1^1, \dots, a_L^1 with q_1)

Individual 2 (a_1^2, \dots, a_L^2 with q_2)

.....

Individual N (a_1^N, \dots, a_L^N with q_N)

To accomplish the selection task, we should find the individual i^* whose Q-value is the largest, i.e. We call this a greedy individual, and the corresponding actions for rules are called greedy actions. The greedy individual is selected with a large probability $1 - \varepsilon$. Otherwise, the previously selected individual is adopted again. Suppose at time t , the individual \hat{i} is selected, i.e., actions $a_1^{\hat{i}}(t), \dots, a_L^{\hat{i}}(t)$ are selected for rules 1, ..., L, respectively. Then, the final output action of the fuzzy system is

$$\bar{a}(t) = \frac{\sum_{i=1}^L \Phi_i(x(t)) a_i^{\hat{i}}(t)}{\sum_{i=1}^L \Phi_i(x(t))} \quad (17)$$

The Q-value of this final output action should be a weighted average of the Q-values corresponding to the actions $a_1^{\hat{i}}(t), \dots, a_L^{\hat{i}}(t)$ i.e.,

$$Q(x(t), \bar{a}(t)) = \frac{\sum_{i=1}^L \Phi_i(x(t)) q_i(t)}{\sum_{i=1}^L \Phi_i(x(t))} = q_{\hat{i}}(t) \quad (18)$$

From this equation, we see that the Q-value of the system output is simply equal to $q_{\hat{i}}(t)$, the Q-value of the selected individual \hat{i} . This means $q_{\hat{i}}$ that simultaneously reveals both the performance of the individual and the corresponding system output action. In contrast to traditional Q-learning, where the Q-values are usually stored in a look-up table, and can deal only with discrete state/action pairs, here both the input state and the output action are continuous. This can avoid the impractical memory requirement for large state-action spaces. The aforementioned selecting, acting, and updating process is repeatedly executed until the end of a trial.

Every time after the fuzzy system applies an action $\bar{a}(t)$ to the environment and a reinforcement $R(t)$, learning of the Q-values is performed. Then, we should update $q_{\hat{i}}(t)$ based on the immediate reward $R(t)$ and the estimated rewards from subsequent states.

Based on the Q-learning rule, we can update $q_{\hat{i}}$ as

$$\begin{aligned}
q_i(t) &= q_i(t) + \alpha(R(t) + \gamma Q^*(x(t+1)) - q_i(t)) \\
Q^*(x(t+1)) &= \max_{i=1, \dots, N} Q(x(t+1), \bar{a}^i) \\
&= \max_{i=1, \dots, N} q_i(t) = q_i(t)
\end{aligned} \tag{19}$$

That is

$$q_i(t) = q_i(t) + \alpha(R(t) + \gamma q_i(x(t+1)) - q_i(t)) = q_i(t) + \alpha \Delta q_i(t) \tag{20}$$

Where $\Delta q_i(t)$ is regarded as the temporal error.

To speed up the learning, the eligibility trace is combined with Q-learning. The eligibility trace for individual i at time t is denoted as $e_i(t)$. On each time step, the eligibility traces

for all individuals are decayed by λ , and eligibility trace for the selected individual \hat{i} on the current step increased by 1, that as

$$\begin{aligned}
e_i(t) &= \lambda e_i(t+1), & \text{if } i \neq \hat{i} \\
&= \lambda e_i(t-1) + 1. & \text{if } i = \hat{i}
\end{aligned}$$

λ is a trace-decay parameter. The value $e_i(t)$ can be regarded as an accumulating trace for each individual i since it accumulates whenever an individual is selected, then decays gradually when the individual is not selected. It indicates the degree to which each individual is eligible for undergoing learning changes. With eligibility trace, (20) is changed to

$$q_i(t) = q_i(t) + \alpha \Delta q_i(t) e_i(t) \tag{21}$$

for all $i=1, \dots, N$. Upon receiving a reinforcement signal, the Q-values of all individuals are updated by (21).

4.10 Fuzzy Reinforcement Signal

The detailed and precise training data for learning is often hard to obtain or may not be available in the process of biped control synthesis. Furthermore, a more challenging aspect of this problem is that the only available feedback signal (a failure or success signal) is obtained only when a failure (or near failure) occurs, that is, the biped robot falls down (or almost falls down). Since no exact teaching information is available, this is a typical reinforcement learning problem and the failure signal serves as the reinforcement signal.

For reinforcement learning problems, most of the existing learning methods for neural networks or fuzzy-neuro networks focus their attention on numerical evaluative information. But for human biped walking, we usually use linguistic critical signal, such as "near fall down", "almost success", "slower", "faster" and etc., to evaluate the walking gait. In this case, using fuzzy evaluation feedback is much closer to the learning environment in the real world. Therefore, there is a need to explore possibilities of the reinforcement learning with fuzzy evaluative feedback, as it was investigated in paper (Zhou & Meng, 2000). Fuzzy reinforcement learning generalizes reinforcement learning to fuzzy environment where only the fuzzy reward function is available.

The most important part of algorithm represent the choice of reward function - external reinforcement. It is possible to use scalar critic signal (Katić & Vukobratović, 2007), but as one of solution, the reinforcement signal was considered as a fuzzy number $R(t)$. We also assume that $R(t)$ is the fuzzy signal available at time step t and caused by the input and action chosen at time step $t-1$ or even affected by earlier inputs and actions. For more effective learning, a error signal that gives more detail balancing information should be given, instead of a simple "go -no go" scalar feedback signal. As an example in this paper, the following fuzzy rules can be used to evaluate the biped balancing according to following table.

$\Delta x^{(zmp)}$	SMALL	MEDIUM	HUGE
$\Delta y^{(zmp)}$			
SMALL	EXCELLENT	GOOD	BAD
MEDIUM	GOOD	GOOD	BAD
HUGE	BAD	BAD	BAD

Fuzzy rules for external reinforcement

The linguistic variables for ZMP deviations $\Delta x^{(zmp)}$ and $\Delta y^{(zmp)}$ and for external reinforcement R are defined using membership functions that are defined in Fig.9.

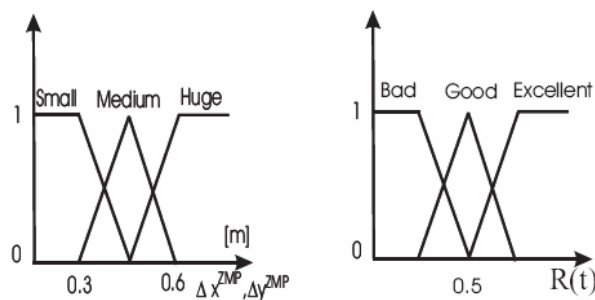


Fig. 9. The Membership functions for ZMP deviations and external reinforcement.

5. Experimental and Simulation Studies

With the aim of identifying a valid model of biped locomotion system of anthropomorphic structure, the corresponding experiments were carried out in a caption motion studio (Rodić et al., 2006). For this purpose, a middle-aged (43 years) male subject, 190 [cm] tall, weighing 84.0728 [kg], of normal physical constitution and functionality, played the role of an experimental anthropomorphic system whose model was to be identified. The subject's geometrical parameters (the lengths of the links, the distances between the neighboring joints and the particular significant points on the body) were determined by direct measurements or photometrically. The other kinematic parameters, as well as dynamic ones, were identified on the basis of the biometric tables, recommendations and empirical relations (Zatsiorsky et al., 1990). A summary of the geometric and dynamic parameters identified on the considered experimental bio-mechanical system is given in Tables 1 and 2. The selected subject, whose parameters were identified, performed a number of motion tests (walking, staircase climbing, jumping), whereby the measurements were made under

the appropriate laboratory conditions. Characteristic laboratory details are shown in Fig. 10. The VICON motion studio equipment was used with the corresponding software package for processing measurement data. To detect current positions of body links use was made of the special markers placed at the characteristic points of the body/limbs (Figs. 10a and 10b). Continual monitoring of the position markers during the motion was performed using six VICON high-accuracy infra-red cameras with the recording frequency of 200 [Hz] (Fig. 10c). Reactive forces of the foot impact/contact with the ground were measured on the force platform (Fig. 10d) with a recording frequency of 1.0 [kHz]. To mimic a rigid foot-ground contact, a 5 [mm] thick wooden plate was added to each foot (Fig. 10b).

Link	Length [m]	Mass [kg]	CM Position	
			Sagittal	Longitudinal
Head	0.2722	5.8347	0.0000	0.1361
Trunk	0.7667	36.5380	0.0144	0.3216
Thorax	0.2500	13.4180	0.0100	0.1167
Abdomen	0.3278	13.7291	0.0150	0.2223
Pelvis	0.1889	9.3909	0.0200	0.0345
Arm	0.3444	2.2784	0.0000	-0.1988
Forearm	0.3222	1.3620	0.0000	-0.1474
Hand	0.2111	0.5128	0.0000	-0.0779
Thigh	0.5556	11.9047	0.0000	-0.2275
Shank	0.4389	3.6404	0.0000	-0.1957
Foot	0.2800	1.1518	0.0420	-0.0684

Table 1. The anthropometric data used in modeling of human body (kinematic parameters and mass of links).

Link	Radii of gyration [m]	Moments of inertia
	Sagitt Trans. Longit.	Ix Iy Iz
Head	0.0825 0.0856 0.0711	0.0397 0.0428 0.0295
Trunk	0.2852 0.2660 0.1464	2.9720 2.5859 0.7835
Thorax	0.1790 0.1135 0.1647	0.4299 0.1729 0.3642
Abdomen	0.1580 0.1255 0.1534	0.3427 0.2164 0.3231
Pelvis	0.1162 0.1041 0.1109	0.1267 0.1017 0.1155
Arm	0.0982 0.0927 0.0544	0.0220 0.0196 0.0067
Forearm	0.0889 0.0854 0.0390	0.0108 0.0099 0.0021
Hand	0.1326 0.1083 0.0847	0.0090 0.0060 0.0037
Thigh	0.1828 0.1828 0.0828	0.3977 0.3977 0.0816
Shank	0.1119 0.1093 0.0452	0.0456 0.0435 0.0074
Foot	0.0720 0.0686 0.0347	0.0060 0.0054 0.0014

Table 2. The anthropometric data used in modeling of human body (dynamic parameters - inertia tensor and radii of gyration).

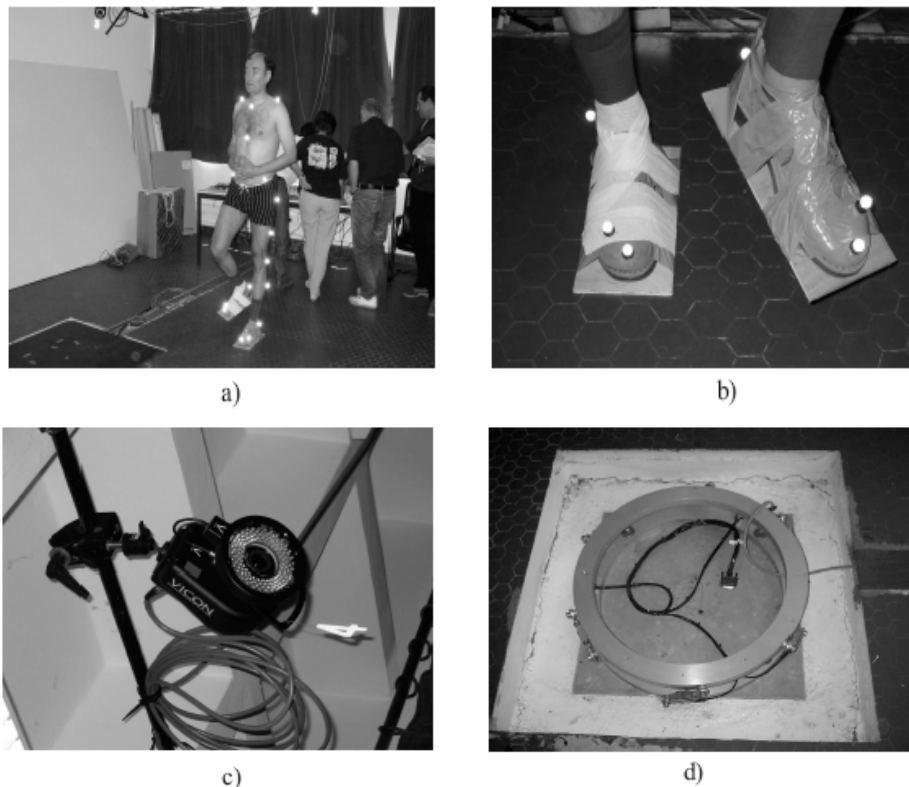


Fig. 10. Experimental capture motion studio in the Laboratory of Biomechanics (Univ. of La Reunion, CURAPS, Le Tampon, France): a) Measurements of human motion using fluorescent markers attached to human body; b) Wooden plates as feet-sole used in locomotion experiments; c) Vicon infra-red camera used to capture the human motion; d) 6-DOFs force sensing platform -sensor distribution at the back side of the plate.

A moderately fast walk ($v = 1.25$ [m/s]) was considered as a typical example of task which encompasses all the elements of the phenomenon of walking. Having in mind the experimental measurements on the biological system and, based on them further theoretical considerations, we assumed that it is possible to design a bipedal locomotion mechanism (humanoid robot) of a similar anthropomorphic structure and with defined (geometric and dynamic) parameters. In this sense, we have started from the assumption that the system parameters presented in Tables 1 and 2 were determined with relatively high accuracy and that they reflect faithfully characteristics of the considered system. Bearing in mind mechanical complexity of the structure of the human body, with its numerous DOFs, we adopted the corresponding kinematic structure (scheme) of the biped locomotion mechanism (Fig. 11) to be used in the simulation examples. We believe that the mechanism (humanoid) of the complexity shown in Fig. 11 would be capable of reproducing with a relatively high accuracy any anthropomorphic motion -rectilinear and curvilinear walk, running, climbing/descending the staircase, jumping, etc. The adopted structure has three

active mechanical DOFs at each of the joints -the hip, waist, shoulders and neck; two at the ankle and wrist, and one at the knee, elbow and toe. The fact is that not all available mechanical DOFs are needed in different anthropomorphic movements. In the example considered in this work we defined the nominal motion of the joints of the legs and of the trunk. At the same time, the joints of the arms, neck and toes remained immobilized. On the basis of the measured values of positions (coordinates) of special markers in the course of motion (Figs. 10a, 10b) it was possible to identify angular trajectories of the particular joints of the bipedal locomotion system. These joints trajectories represent the nominal, i.e. the reference trajectories of the system. The graphs of these identified/reference trajectories are shown in Figs. 12 and 13. The nominal trajectories of the system's joints were differentiated with respect to time, with a sampling period of $\Delta t = 0.001$ [ms]. In this way, the corresponding vectors of angular joint velocities and angular joint accelerations of the system illustrated in Fig. 11 were determined. Animation of the biped gait of the considered locomotion system, for the given joint trajectories (Figs. 12 and 13), is presented in Fig. 14 through several characteristic positions. The motion simulation shown in Fig. 14 was determined using kinematic model of the system. The biped starts from the state of rest and then makes four half-steps stepping with the right foot once on the platform for force measurement. Simulation of the kinematic and dynamic models was performed using Matlab/Simulink R13 and Robotics toolbox for Matlab/Simulink. Mechanism feet track their own trajectories (Figs. 12 and 13) by passing from the state of contact with the ground (having zero position) to free motion state.

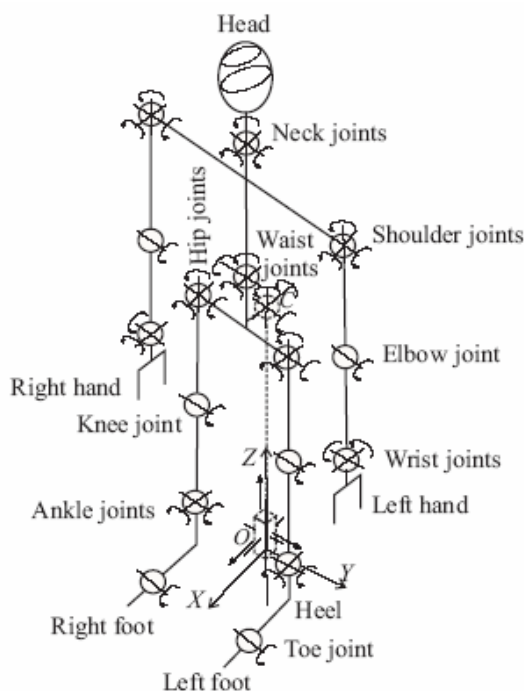


Fig. 11. Kinematic scheme of a 38-DOFs biped locomotion system used in simulation as the kinematic model of the human body referred to in the experiments.

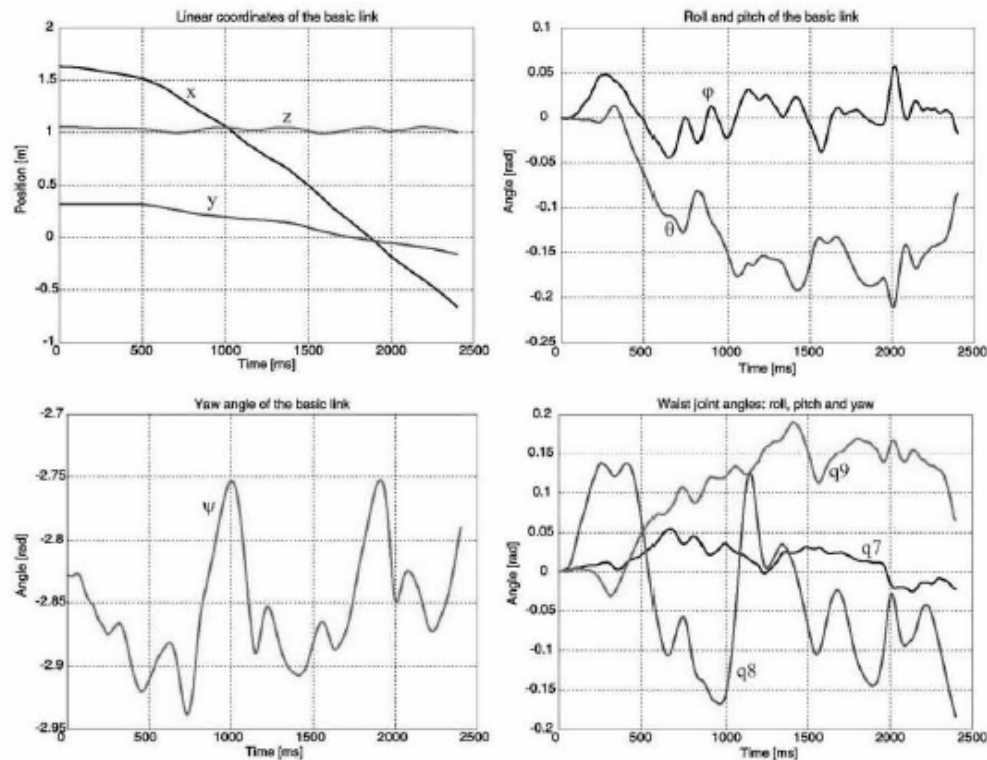


Fig. 12. Nominal trajectories of the basic link: x-longitudinal, y-lateral, z-vertical, ϕ -roll, θ -pitch, ψ -yaw; Nominal waist joint angles: q7-roll, q8-yaw, q9-pitch.

Some special simulation experiments were performed in order to validate the proposed reinforcement learning control approach. Initial (starting) conditions of the simulation examples (initial deviations of joints' angles) were imposed. Simulation results were analyzed on the time interval 0.1[s]. In the simulation example, two control algorithms were analyzed: (i) basic dynamic controller described by computed torque method (without learning) and (ii) hybrid reinforcement learning control algorithm. (with learning). The results obtained by applying the controllers (i) (without learning) and (ii) (with learning) are shown on Figs. 15 and Fig.16. It is evident, that better results were achieved with using reinforcement learning control structure.

The corresponding position and velocity tracking errors in the case of application reinforcement learning structure are presented on Figs. 17 and 18. The tracking errors converge to zero values in the given time interval. It means that the controller ensures good tracking of the desired trajectory. Also, the application of reinforcement learning structure ensures a dynamic balance of the locomotion mechanism.

In Fig. 19 value of internal reinforcement through process of walking is presented. It is clear that task of walking within desired ZMP tracking error limits is achieved in a good fashion.

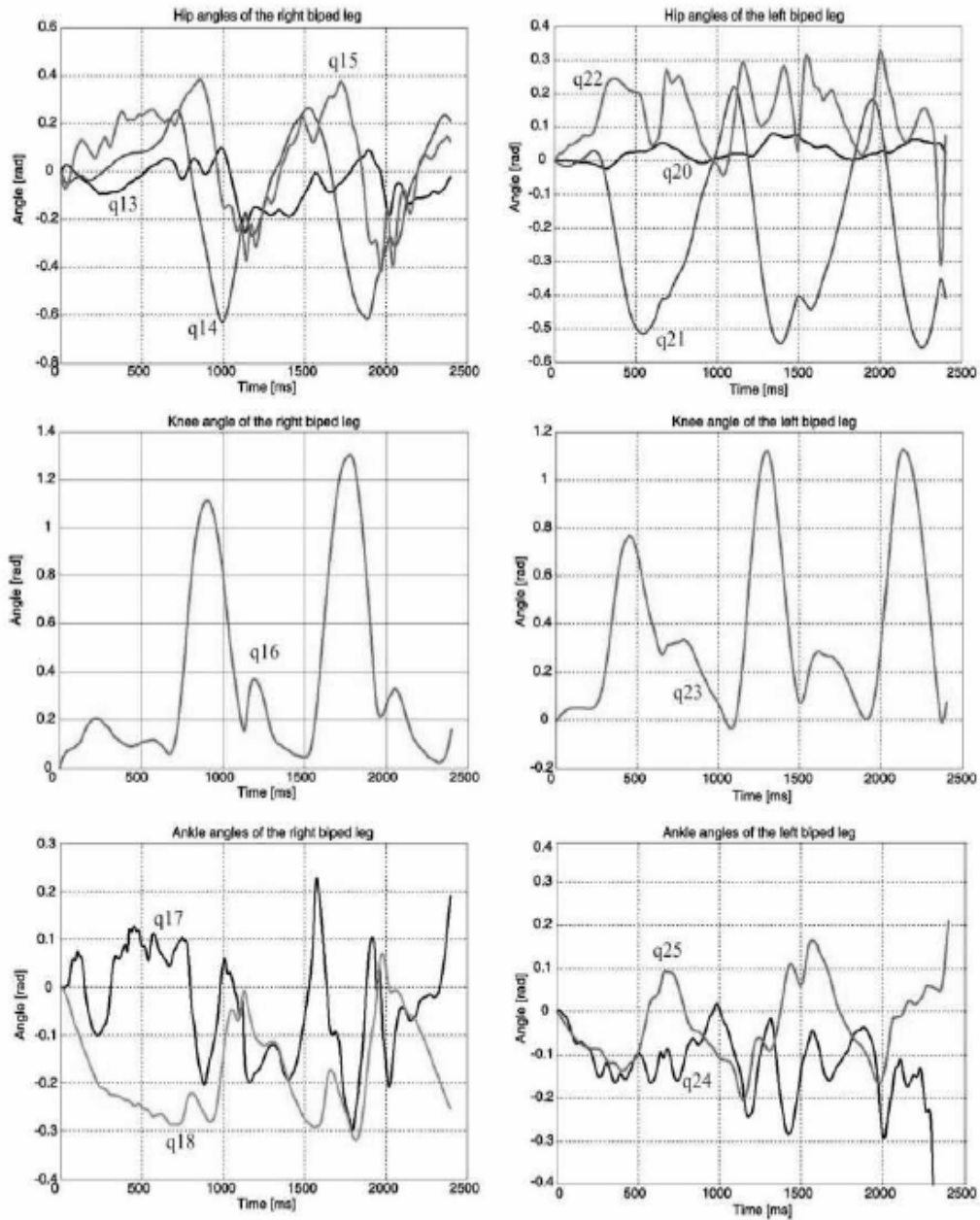


Fig. 13. Nominal joint angles of the right and left leg: q13, q17, q20, q24-roll, q14, q21, q16, q18, q23, q25-pitch, q15, q22-yaw.

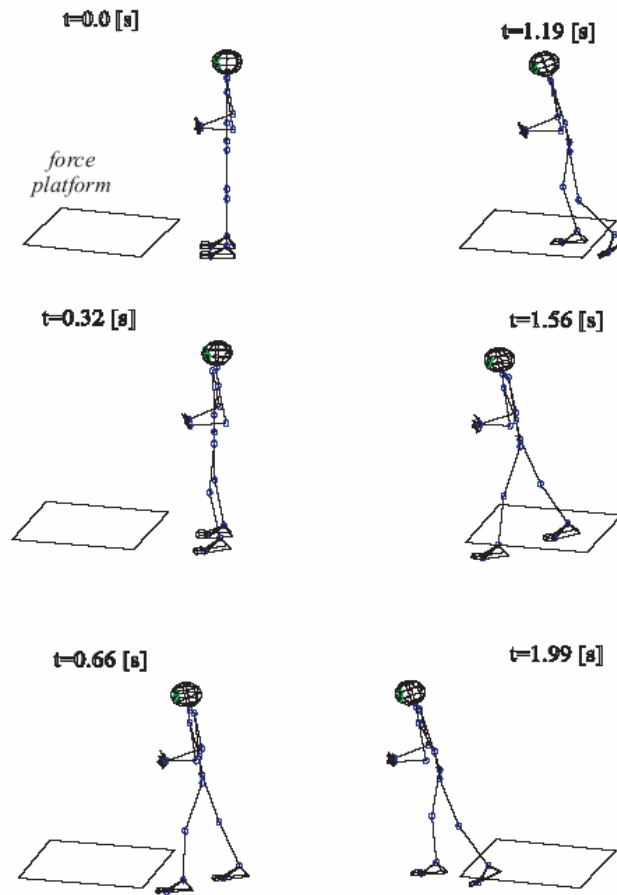


Fig.14. Model-based animation of biped locomotion in several characteristic instants for the experimentally determined joint trajectories.

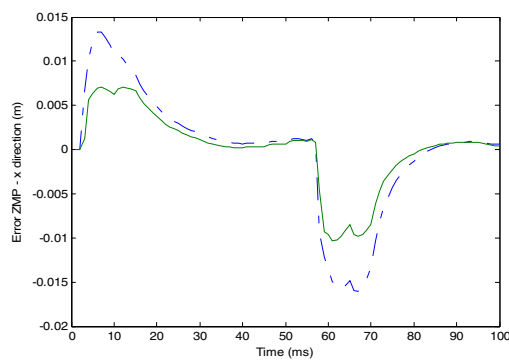


Fig. 15. Error of ZMP in x-direction (- with learning - - without learning).

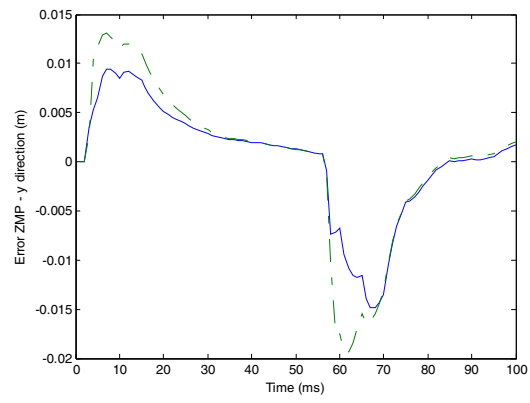


Fig. 16. Error of ZMP in y-direction (- with learning - - without learning).

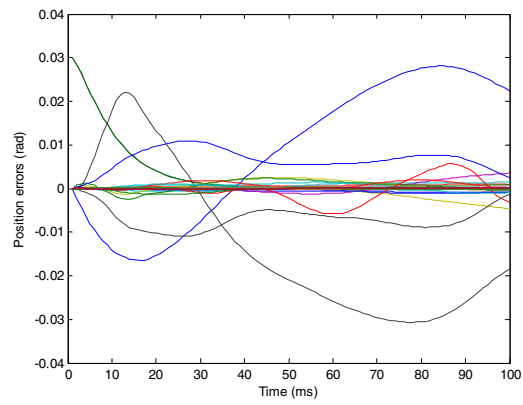


Fig. 17. Position tracking errors.

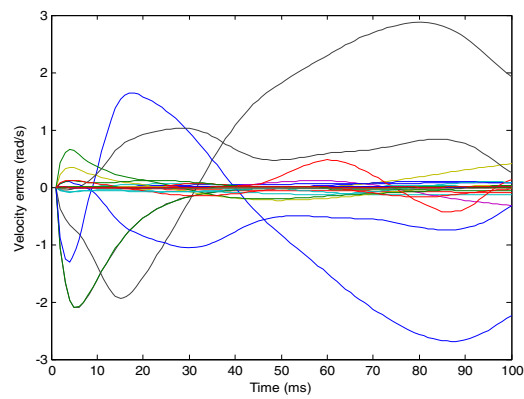


Fig. 18. Velocity tracking errors.

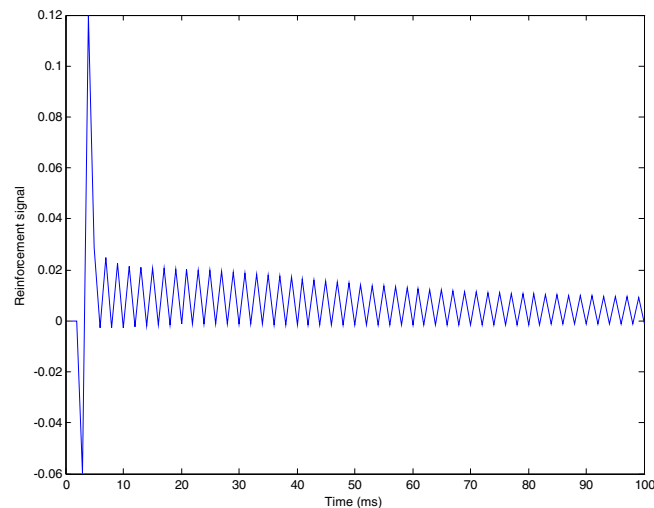


Fig. 19. Reinforcement through process of walking.

6. Conclusions

This study considers a optimal solutions for application of reinforcement learning in humanoid robotics Humanoid Robotics is a very challenging domain for reinforcement learning, Reinforcement learning control algorithms represents general framework to take traditional robotics towards true autonomy and versatility. The reinforcement learning paradigm described above has been successfully implemented for some special type of humanoid robots in the last 10 years. Reinforcement learning is well suited to training biped walk in particular teaching a robot a new behavior from scalar or fuzzy feedback. The general goal in synthesis of reinforcement learning control algorithms is the development of methods which scale into the dimensionality of humanoid robots and can generate actions for biped with many degrees of freedom. In this study, control of walking of active and passive dynamic walkers by using of reinforcement learning was analyzed.

Various straightforward and hybrid intelligent control algorithms based RL for active and passive biped locomotion is presented. The proposed RL algorithms use the learning elements that consists of various types of neural networks, fuzzy logic nets or fuzzy-neuro networks with focus on fast convergence properties and small number of learning trials.

Special part of study represents synthesis of hybrid intelligent controllers for biped walking. The hybrid aspect is connected with application of model-based and model free approaches as well as with combination of different paradigms of computational intelligence. These algorithms includes combination of a dynamic controller based on dynamic model and special compensators based on reinforcement structures. Two different reinforcement learning structures were proposed based on actor-critic approach and Q-learning. The algorithms is based on fuzzy evaluative feedback that are obtained from human intuitive balancing knowledge. The reinforcement learning with fuzzy evaluation feedback is much closer to the human biped walking evaluation than the original one with scalar feedback.

The proposed hybrid intelligent control scheme fulfills the preset control criteria. Its application ensures the desired precision of robot's motion, maintaining dynamic balance of the locomotion mechanism during a motion

The developed hybrid intelligent dynamic controller can be potentially applied in combination with robotic vision, to control biped locomotion mechanisms in the course of fast walking, running, and even in the phases of jumping, as it possesses both the conventional position-velocity feedback and dynamic reaction feedback. Performance of the control system was analyzed in a number of simulation experiments in the presence of different types of external and internal perturbations acting on the system. In this paper, we only consider the flat terrain for biped walking. Because the real terrain is usually very complex, more studies need to be conducted on the proposed gait synthesis method for irregular and sloped terrain.

Dynamic bipedal walking is difficult to learn because combinatorial explosion in order to optimize performance in every possible configuration of the robot., uncertainties of the robot dynamics that must be only experimentally validated, and because coping with dynamic discontinuities caused by collisions with the ground and with the problem of delayed reward -torques applied at one time may have an effect on the performance many steps into the future. Hence, for a physical robot, it is essential to learn from few trials in order to have some time left for exploitation. It is thus necessary to speed the learning up by using different methods (hierarchical learning, subtask decomposition, imitation,...).

7. Acknowledgments

The work described in this conducted was conducted within the national research project "Dynamic and Control of High-Performance Humanoid Robots: Theory and Application". and was funded by the Ministry of Science and Environmental Protection of the Republic of Serbia. The authors thank to Dr. Ing. Aleksandar Rodić for generation of experimental data and realization of humanoid robot modeling and trajectory generation software.

8. References

- Barto, A.G., Sutton, R.S & Anderson, C.W., (1983), Neuron like adaptive elements that can solve difficult learning control problems, *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-13, 5, September 1983, 834-846.
- Baxter, J. and Bartlett, P., (2001). Infinite-horizon policy-gradient estimation , *Journal of Artificial Intelligence Research*, 319-350.
- Benbrahim, H. & Franklin, J.A. (1997), Biped Dynamic Walking using Reinforcement Learning, *Robotics and Autonomous Systems*, 22, 283-302.
- Berenji, H.R. & Khedkar, P., (1992), Learning and Tuning Fuzzy Logic controllers through Reinforcements, *IEEE Transactions on Neural Networks*, 724-740
- Berenji, H.R., (1996), Fuzzy Q-learning for generalization of reinforcement," in *Proc. IEEE Int. Conf. Fuzzy Systems*, 2208-2214.
- Bertsekas, D.P. and Tsitsiklis, J.N. (1996), *Neuro-Dynamic Programming*, Athena Scientific, Belmont, USA.
- Chew, C. & Pratt, G.A., (2002), Dynamic bipedal walking assisted by learning, *Robotica*, 477-491
- Doya, K., (2000), Reinforcement Learning in Continuous Time and Space, *Neural Computation*, 219-245.

- Glorennec, P.Y. & ouffe, L. (1997), Fuzzy Q - Learning, in *Proceedings of FUZZ-IEEE'97*, Barcelona, July 1997.
- Jouffe, L., (1998), Fuzzy inference system learning by reinforcement methods, *IEEE Transactions on Systems, Man, Cybernetics:Part C, Appl. Rev.*, 28, 3, August 1998, 338-355.
- Kamio, S., & Iba, H., (2005), Adaptation Technique for Integrating Genetic Programming and Reinforcement Learning fir Real Robot, *IEEE Transactions on Evolutionary Computation*, 9, 3, June 2005, 318 - 333.
- Katić, D. & Vukobratović, M., (2003a), Survey of Intelligent Control Techniques for Humanoid Robots, *Journal of Intelligent and Robotic Systems*, 37, 2003, 117 -141.
- Katić, D. & Vukobratović, M. (2003b), *Intelligent Control of Robotic Systems*, Kluwer Academic Publishers, Dordrecht, Netherlands
- Katić, D. & Vukobratović, M.: (2005), Survey of Intelligent Control Algorithms For Humanoid Robots", in *Proceedings of the 16th IFAC World Congress*, Prague, Czech Republic, July 2005.
- Katić, D. & Vukobratović, M., (2007), Hybrid Dynamic Control Algorithm for Humanoid Robots Based on Reinforcement Learning, accepted for publication in *Journal of Intelligent and Robotic Systems* .
- Kimura, H. & Kobayashi, S. (1998), An analysis of actor/critic algorithms using eligibility traces: Reinforcement learning with imperfect value functions ,in *Proceedings of the International Conference on Machine Learning (ICML '98)*, 278-286.
- Kun , A.L. and Miller,III, W.T., (1999), Control of Variable -Speed Gaits for a Biped Robot, *IEEE Robotics and Automation Magazine*, 6, September 1999, 19-29.
- Li, Q., Takanishi, A. & Kato, I., (1992), Learning control of compensative trunk motion for biped walking robot based on ZMP, in *Proceedings of the 1992 IEEE/RSJ Intl.Conference on Intelligent Robot and Systems*, 597-603.
- Mori, T., Nakamura, Y., Sato, M. and Ishii, S., (2004), Reinforcement learning for a cpg-driven biped robot, in *Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI)*, 623-630.
- Morimoto, J., Cheng, G., Atkeson, C.G. & Zeglin,G., (2004) A Simple Reinforcement Learning Algorithm For Biped Walking, in *Proceedings of the 2004 IEEE International Conference on Robotics & Automation*, New Orleans, USA.
- Nagasaka, K., Inoe, H. and Inaba, M., (1999), Dynamic walking pattern generation for a humanoid robot based on optimal gradient method, in *Proceedings of the International Conference on Systems, Man. and Cybernetics*, 908-913.
- Nakamura, Y., Sato, M. & Ishii, S., (2003), Reinforcement learning for biped robot, in *Proceedings of the 2nd International Symposium on Adaptive Motion of Animals and Machines*.
- Peters, J., Vijayakumar, S., and Schaal, S., (2003), Reinforcement Learning for Humanoid Robots, in *Proceedings of the Third IEEE-RAS International Conference on Humanoid Robots*, Karlsruhe & Munich.
- Rodić, A., Vukobratović, M., Addi, K. & Dalleau, G. (2006), Contribution to the Modelling of Non-Smooth, Multi-Point Contact Dynamics of Biped Locomotion - Theory and Experiments, submitted to journal *Robotica*.
- Salatian, A.W., Yi, K.Y. and Zheng, Y.F., (1997) , Reinforcement Learning for a Biped Robot to Climb Sloping Surfaces, *Journal of Robotic Systems*, 283-296.
- Schuitema, E., Hobbelen, D.G.E, Jonker, P.P., Wisse, M. & Karssen, J.G.D., (2005), Using a controller based on reinforcement learning for a passive dynamic walking robot, *IEEE International conference on Humanoid Robots 2005*, Tsukuba, Japan.

- Sutton, R.S. and Barto, A.G. (1998) *Reinforcement Learning: An Introduction*, The MIT Press, Cambridge, USA.
- Sutton, R.S., McAllester, D., & Singh, S., (2000), Policy Gradient Methods for Reinforcement learning with Function Approximation, in *Advances in Neural Information Processing Systems*, 12, MIT Press, Cambridge, USA, 1057-1063.
- Tedrake, R., Zhang, T.W. & Seung, H.S., (2004), Stochastic policy gradient reinforcement learning on a simple 3d biped, in *Proceedings of the 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems*.
- Vukobratović, M., & Juričić, D. (1969), Contribution to the Synthesis of Biped Gait, *IEEE Transactions on Biomedical Engineering*, BME-16, 1, 1-6.
- Vukobratović, M., Borovac, B., Surla, D., & Stokić, D., (1990), *Biped Locomotion - Dynamics, Stability, Control and Application*, Springer Verlag, Berlin, Germany.
- Watkins, C.J.C.H. & Dayan, P., (1992), Q Learning, *Machine Learning*, 279-292.
- Zatsiorsky, V., Seluyanov, V. & Chugunova, L. (1990) ,Methods of Determining Mass - Inertial Characteristics of Human Body Segments, *Contemporary Problems of Biomechanics*, 272-291, CRC Press.
- Zhou, C. & Meng, D. (2000), Reinforcement Learning with Fuzzy Evaluative Feedback for a Biped Robot. In: *Proceedings of the IEEE International Conference on Robotics and Automation*, San Francisco, USA, 3829-3834.