Control Algorithm For Biped Walking Using Reinforcement Learning

Dusko KATIĆ Miomir VUKOBRATOVIĆ

Robotics Laboratory, Mihailo Pupin Institute
P.O.Box 15, Volgina 15, 11060 Belgrade, Serbia & Montenegro
fax: +381 - 11 - 775 - 870; e-mail:dusko,vuk@robot.imp.bg.ac.yu

Abstract: The work is concerned with the integrated dynamic control of humanoid locomotion mechanisms based on the spatial dynamic model of humanoid mechanism. The control scheme was synthesized using the centralized model with proposed structure of dynamic controller that involves two feedback loops: position-velocity feedback of the robotic mechanism joints and reinforcement learning feedback around Zero-Moment Point. The proposed reinforcement learning is based on modified version GARIC architecture for dynamic reactive compensation. Simulation experiments were carried out in order to validate the proposed control approach.

Keywords: Humanoid robots, Biped locomotion, Dynamically balanced gait, Reinforcement learning.

1 Introduction

Having in mind very high requirements to be satisfied by humanoid robots it should be pointed out the need to increase the number of degrees of freedom (DOFs) of their mechanical configuration, to study in more depth some previously unconsidered phenomenons in the stage of forming the corresponding dynamic models of humanoids, as well as the need to make appropriate controller software that would be capable of meeting the most complex requirements of stable trajectory tracking and maintaining dynamic balance of both the regular (stationary) gait in the presence of small perturbations and of the robot posture in the case of large perturbations. It should be also pointed out that the problem of motion of humanoid robots is a very complex control task, especially when the real environment (scene) is taken into account, which, as minimum, requires its integration with the robot's dynamic model.

In this paper, a novel, integrated dynamic control structure for the humanoid robots is proposed, using the overall model of robot mechanism. The first control algorithm represents some kind of computed torque control method as basic dynamic control method, while the second part of algorithm is modified GARIC reinforcement learning architecture for dynamic compensation of ZMP (Zero-Moment-Point) error.

Recently, reinforcement learning has attracted attention as a learning method for studying movement planning and control [3], [4],[5]. Reinforcement learning concept that is based on trial and error methodology and constant evaluation of performance in constant interaction with environment. Reinforcement learning typically requires an unambiguous representation of states and actions and the existence of a scalar reward function.

The goal of this paper is to propose the usage of reinforcement learning for humanoid robotics. Initially, there are several approaches [7],[6], [8], [9] with additional demands and requirements because high dimensionality of the control problem. Furthermore, Benbrahim and Franklin showed the potential of these methods to scale into the domain of humanoid robotics [6].

The basic reinforcement learning method is based on the Actor-Critic architecture. Actor-Critic methods are the natural extension of the idea of reinforcement comparison methods to Temporal Difference (TD) learning [5]. The Actor network can be thought of as the control agent, because it implements a policy. The Actor network is part of the dynamic system as it interacts directly with the system by providing control signals for the plant. The Critic network implements the reinforcement learning part of the control system as it provides policy evaluation and can be used to perform policy improvement. This learning agent architecture has the advantage of implementing both a reinforcement learning mechanism as well as a control mechanism. For the Actor, we selected the two-layer, feedforward neural network with sigmoid hidden units and linear output units. For the Critic, neuro-fuzzy network is proposed. The critic is trained to produce the expected sum of future reinforcement that will be observed given the current values of deviation of dynamic reactions and action. The Actor network receives the position and velocity tracking error from the biped system . It is trained via Back propagation (gradient descent) algorithm and training example provided by Critic net. The implemented algorithm was base on modified version of GARIC approach presented in paper [10]. In this paper, the external reinforcement signal was simply defined to be measure of ZMP error. Internal reinforcement signal is generated using external reinforcement signal and appropriate policy,

2 Model of the system

2.1 Model of the robot's mechanism

Biped locomotion mechanisms represent generally branched kinematic chains interconnected with spherical or cylindrical joints [1]. During the motion, some

kinematic chains in their interaction with the environment transform from open to closed type of chain [2]. In Fig. 1 is shown the kinematic scheme of the biped locomotion mechanism [2] whose spatial model will be considered in this work. The model will be used to synthesize dynamic control of the locomotion mechanism and to verify the research results obtained in simulation experiments. The mechanism possesses 18 powered DOFs, designated by the numbers *1-18*, and two unpowered DOFs (*1'* and *2'*) for the footpad rotation about the axes passing through the instantaneous ZMP position. Thus, the considered mechanism has in total *n=20* DOFs of motion.

The mechanism dynamic model presented in Fig. 1 has been formed using the relations known from Newton's rigid body dynamics. There are several approaches to forming the model of locomotion mechanisms, depending on which of the kinematic chain links is taken as the 'basic' one. In this paper, the mechanism model is defined solely in the state space of robotic internal coordinates [2]. For this purpose, the first link in the branched chain, representing the supporting foot, is adopted as the basic link of the mechanism.

Bearing in mind the selected basic link of the mechanism, recursive numerical relations are formed [1] that successively determine angular and translational velocities and accelerations of particular links of the robotic mechanism. Taking into account the dynamic coupling between particular parts (branches) of the mechanism chain one can derive the relation that describes the overall dynamic model of the locomotion mechanism in a vector form [2]:

$$P = H(q) + h(q, \dot{q}) \tag{1}$$

where: $P \in R^{n\times 1}$ is the vector of driving moments at the humanoid robot joints; $F \in R^{6\times 1}$ 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 shown in Fig. 1; $h \in R^{n\times 1}$ is the vector of gravitational, centrifugal and Coriolis moments acting at n mechanism joints; $n = 20$ is the total number of DOFs (Fig. 1). Of special importance in the calculation of the model (1) is the force F, which represents the vector of forces and moments of ground reaction at the moment of contact of the foot of free (unconstrained) leg and ground surface, i.e. at the moment when the weight is transferred from one foot to the other. The pertinent terminology distingushes between the so-called supporting or constrained foot and unconstrained foot, which in the moment of contact with the ground is transformed into the constrained one. In this paper, our primary concern to consider the contact of rigid foot with ground and walking on slightly horizontal plane.

2.2 Definition of control criteria

In the synthesis of control for biped mechanism gait it is necessary to satisfy certain natural principles. The control ought to satisfy the following criteria: (i) accuracy of tracking the desired trajectories at the mechanism joints (ii) maintaining dynamic balance of the mechanism during the motion, (iii) minimization

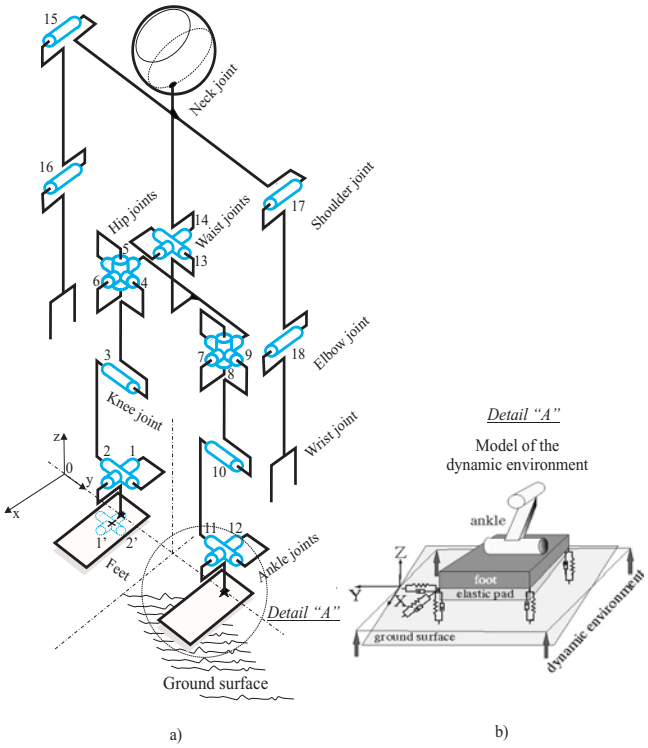

Figure 1: Model of the humanoid locomotion mechanism with 18 active and 2 passive DOFs: (a) kinematic scheme of the mechanism, (b) dynamic model of the environment

of the impact arising at the moment of contact of the free foot and the ground during the gait, (iv) minimization of dynamic loads at the robot joints, and (v) realization of anthropomorphic characteristics of the gait.

Fulfillment of criterion (i) enables realization of the desired mode of motion, walk repeatability and avoiding of potential obstacles in the way. To satisfy criterion (ii) it means to have a stable balanced walk. Fulfillment of criterion (iii) ensures a higher degree of stability of the overall system in respect of the impact appearing at the moment when the unconstrained leg foot strikes the ground. Fulfillment of criterion (iv) is needed for the purpose of minimizing dynamic loads at the robotic joints, which is especially important for the joints bearing the highest load during the walk, e.g. the hip. Criterion (v) is related to the quality of walk realization.

2.3 Gait phases and indicator of dynamic balance

The robot's bipedal gait consists of several phases that are periodically repeated [2]. At that, depending on whether the system is supported on one or two 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. 2 illustrates these gait phases of biped robot locomotion, 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 the direct contact with the support. While walking, the biped is constantly in the state of a certain dynamic balance. 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 [1],[2] 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 at which all the rotation moments about the x and y axes are equal to zero. Instantaneous position of the ZMP is the best indicator of the dynamic bal-

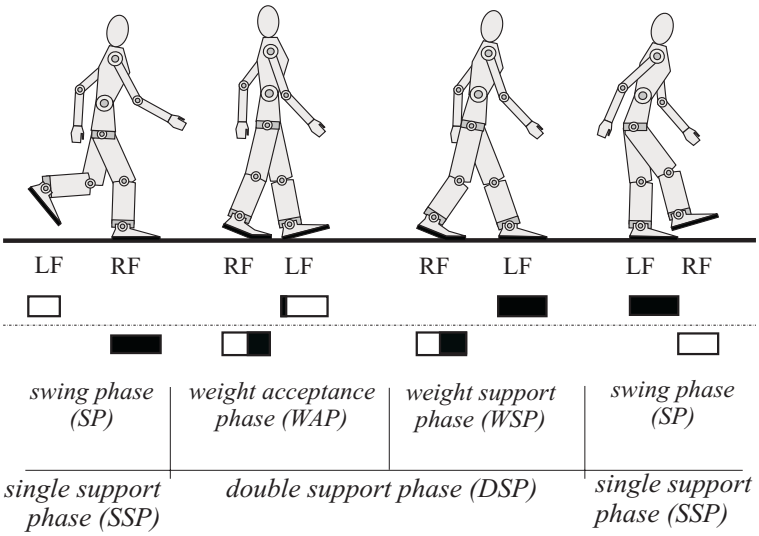

Figure 2: Phases of biped gai

ance of the biped robot. The ZMP position inside these stability areas ensures dynamically balanced gait of the mechanism [1], whereas its position outside these zones indicates the state of instability of the overall mechanism and the possibility of its overturning,. The quality of controling the robot balance can be measured by the success in tracking the ZMP trajectory within the support polygon of the mechanism. The ZMP position is determined by the calculation based on measuring reaction forces under the robot foot. Force sensors are

usually placed on the foot sole.

3 Dynamic Integrated Control algorithm

In accordance with the control task, we propose the application of the algorithm of the so-called integrated dynamic control, based on the knowing of the overall dynamic model of the system. At that, it is assumed that the following assumptions hold: (i) the model (1) describes sufficiently well the behavior of the system presented in Fig. 1; (ii) Desired (nominal) trajectory of the mechanism performing a dynamically balanced gait is known during motion. It is determined off-line (by some of the known mathematical methods) or calculated in real time on some of higher robot control levels; (iii) Geometric and dynamic parameters of the mechanism are known and constant.

Based on the above assumptions, in Fig. 3 is presented the block-diagram of the dynamic controller for biped locomotion mechanism, proposed in this work. It involves two feedback loops: (i) position-velocity feedback, (ii) dynamic reaction feedback at the ZMP based on GARIC reinforcement learning structure, The synthesized dynamic controller (Fig. 3) was designed on the basis of the centralized dynamic model.The vector of driving moments $\hat{P}$ represents the sum of the driving moments $\hat{P}_1$ and $\hat{P}_2$. The moments $\hat{P}_1$ are determined so to ensure precise tracking of the robot's position and velocity in the space of joints coordinates. The driving moments $\hat{P}_2$ are calculated with the aim of correcting the current ZMP position with respect to its nominal.

3.1 Controller of trajectory tracking

The controller of tracking nominal trajectory of the locomotion mechanism has to ensure the realization of a desired motion of the humanoid robot and avoiding fixed obstacles on its way. In [2], it has been demonstrated how local PD or PID controllers of biped locomotion robots are being designed. In this work, the controller for robotic trajectory tracking was synthesized using the computing torque method in the space of internal coordinates of the mechanism joints. For this purpose use was made of the robot dynamic model defined by the relation (1). The control law can be expressed in the known form:

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

where $\hat{H}$,$\hat{h}$ are the corresponding estimated values of the inertia matrix, vector of gravitational, centrifugal and Coriolis forces and moments 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.

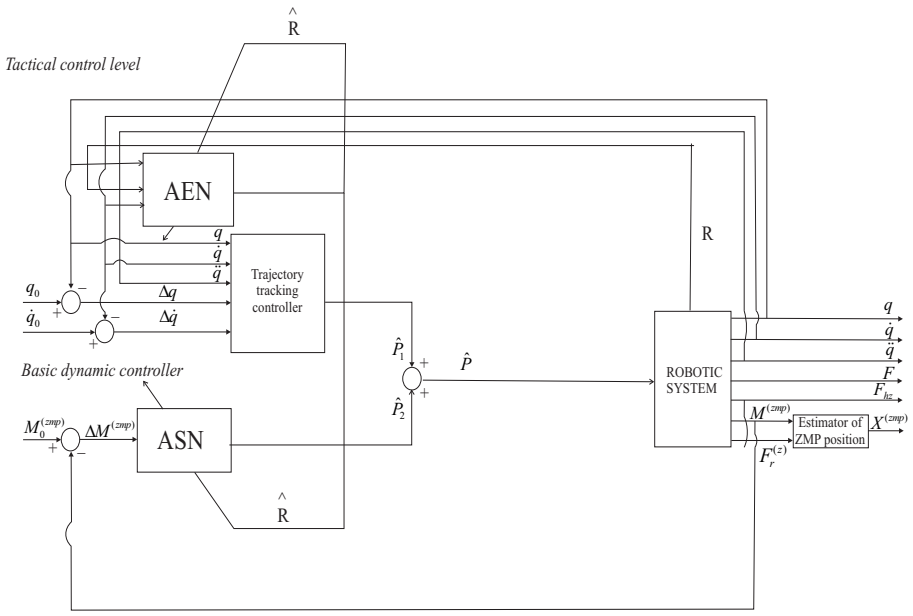

Figure 3: Block-scheme of the integrated dynamic control of biped with two feedback loops

3.2 GARIC Compensator of dynamic reactions

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, i.e. to produce its rotation about the mentioned rotation axes (axes of the joints *1'* and *2'* in Fig. 1). On the basis of the above the reinforcement control algorithm is defined with respect to the dynamic reaction of the support at ZMP. In this case external reinforcement signal R is defined according to values of ZMP error. If ZMP error is greater then chosen limit, external reinforcement signal is set to value 1. Hence, AEN network (action evaluation network) maps position and velocity tracking errors and external reinforcement signal R in scalar value (internal reinforcement $\hat{R}$) which represent the quality of given control task defined by

the following policy:

$$\hat{R}(t+1) = R(t) + \gamma v(t+1) - v(t) \tag{3}$$

where $v(t)$ is output of AEN; γ is a coefficient between 0 and 1. ASN (action selection network) maps the deviation of dynamic reactions in recommended control torque. Exactly, by using SAM(Stochastic action modifier), based on recommended control torque and internal reinforcement $\hat{R}$, control torque P_{dr} is generated. Learning process of AEN (tuning of network weighting factors) is realized by modified version of back propagation algorithm where error is defined by internal reinforcement signal $\hat{R}$. In the same way, using gradient method and internal reinforcement signal, learning proces of ASN is realized. $\Delta M^{(zmp)} \in R^{2\times 1}$ 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 moments at the joints *1'* and *2'* (Fig. 1) that ensures the state of dynamic balance. The control moments P_{dr} calculated from GARIC reinforcement learning structure can not be generated at the joints *1'* and *2'* because these are underactuated, i.e. passive joints. Because of that the control action is 'displaced' to the other, powered 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 presented in Fig. 1, the compensation was carried out using the following mechanism joints: *1*, *6* and *14* to compensate for the dynamic reactions about the x-axis and *2*, *4* and *13* to compensate for the moments about the y-axis. Thus, the ankle joints, hip joints and. waist joints are taken into consideration. Complete control $\hat{P}$ (Fig. 4), is calculated on the basis of the vector of the moments P_{dr} (after distribution it is $\hat{P}_2$ calculated using the GARIC structure , whereby it is borne in mind how many 'compensational joints' are really engaged. In the case when compensation of the ground dynamic reactions is performed using all six proposed joints the compensation moments P_{dr} are uniformly distributed over all of the selected joints, to load uniformly the . In nature, biological systems use simultaneously a large number of joints for correcting their balance. However, for the purpose of verifying the control algorithm, in this work the choice was restricted only to the mentioned six joints: 1, 2, 4, 6, 13 and 14 (Fig. 1).

4 Simulation experiments

Theoretical results presented previously were analyzed on the basis of numerical data obtained by simulation of the closed-loop model of the locomotion mechanism shown in Fig. 1. Total mass of the mechanism was $m = 70$ $[kg]$ and its geometric and dynamic parameters were taken from (Vukobratović, 1990). Simulation examples are concerned with the characteristic pattern of artificial gait in which the mechanism makes a half-step of the length $l = 0.40$ $[m]$in the time period of $t = 0.75$ $[s]$. Nominal trajectories at robot joints are synthesized for the gait in the horizontal plane. The simulation results were analyzed on the

time interval corresponding to the duration of one half-step of the locomotion mechanism in the swing phase (Fig. 2). In the analysis of the efficiency of the developed dynamic controller (Fig. 3) in realizing dynamically balanced motion the most delicate is the single-support phase (swing phase), as well as the moment when the so-called free foot touches/strikes the ground. For this reason of special importance for control is the analysis of dynamic robot behavior in these time intervals, so that the simulation examples were selected to encompass these critical phases. In the first simulation example the assigned initial deviations of particular angles at mechanism joints did not exceed $\Delta q_i \leq 10^\circ$ Constant inclinations of the ground surface in the sagittal plane $\gamma_1 = 3^\circ$ and frontal plane $\gamma_2 = 2^\circ$ were introduced as an additional disturbance. Thus the simulation dealt with the real case of walking on a quasi-horizontal support. Of concern was the robot's behavior in the swing phase (Fig. 2) when the robot by its rigid foot relies on the ground while the other (free) foot is above the ground surface. At that, two cases of control were analyzed: (i) applying only the controller of tracking the given trajectory with position-velocity feedback (Fig. 3) and (ii) applying the combined control with the controller of trajectory tracking and compensator of dynamic reactions of the ground around the ZMP. In the case (ii) use was made of the control structure called 'Basic dynamic controller' (see Fig. 3). In Figs. 4,5, and 6 are presented the results of applying the controller in the case (ii). On analyzing the results presented in Figs. 4 and 5 one can see that the we have better results for error of ZMP when algorithm with training of ASN neuro-fuzzy network is used. It can be concluded that without the feedback with respect to the ground reactions around the ZMP it is not generally possible to ensure dynamic balance of the locomotion mechanism in its motion. This comes out from the fact that the nominal trajectory was synthesized without taking into account the possible deviations of the surface on which biped walks from an ideally horizontal plane. Therefore, the ground surface inclination influences the system's balance as an external stochastic disturbance.

In Fig. 6 are presented the corresponding deviations (errors) Δq_i of the real values of angles at the robot joints from their nominal values when the controller of tracking desired trajectory was applied. The deviations of the variables converge to a zero value on the given time interval, which means that the controller employed ensured good tracking of the desired trajectory.

In Fig. 7 value of internal reinforcment throyh proces of walkkinh is presented. It is clear that task of walking within desired ZMP tracking error limits is achieved.

Conclusions

The control scheme of an integrated dynamic controller of locomotion mechanism was synthesized. Control level consists of the so-callled 'basic dynamic controller' was synthesized, consisting of a dynamic controller for tracking robot's

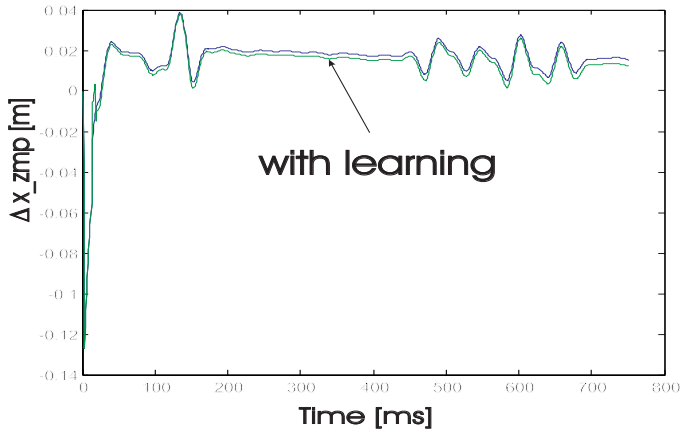

Figure 4: Error of ZMP in x-direction

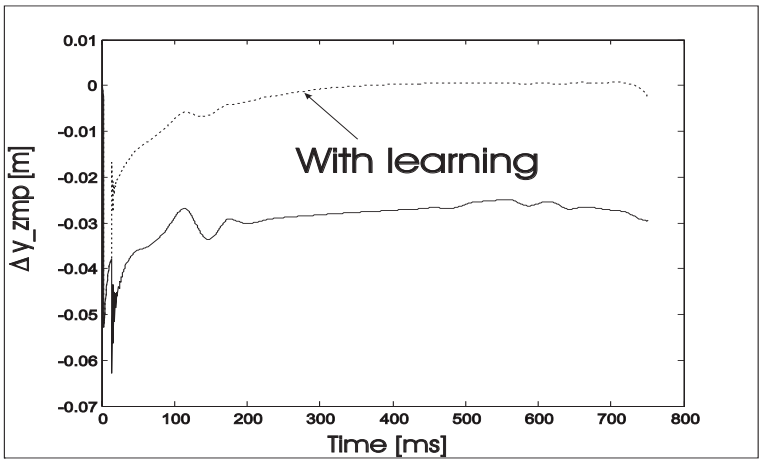

Figure 5: Error of ZMP in x-direction

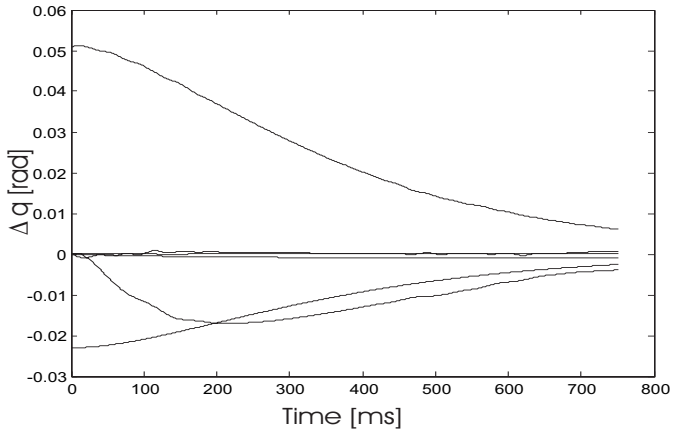

Figure 6: Position tracking errors for compensation joints

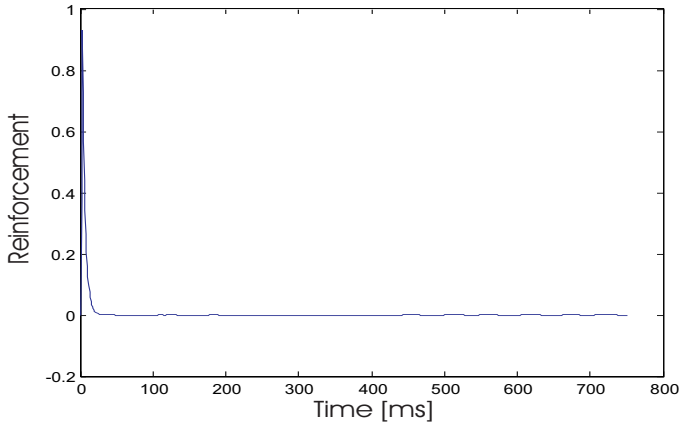

Figure 7: Internal reinforcement through process of walking

nominal trajectory and a compensator of dynamic reactions of the ground around the ZMP based on GARIC reinforcement learning architecture. At that, feedback loops were formed with respect to position and velocity of the mechanism joints, as well as with respect to dynamic ground reactions. Basic dynamic controller was designed with the aim of ensuring precise tracking of the given motion and maintaining dynamic balance of the humanoid mechanism. The proposed control scheme fulfills the preset control criteria.

References

[1] D. Juričić, and M. Vukobratović, " Mathematical Modeling of Biped Walking Systems", ASME Publ. 72-WAlBHF-13, 1972.

[2] M.Vukobratović, B.Borovac, B. Surla and D. Stokić, Biped Locomotion: Dynamics, Stability,Control and Application. Springer-Verlag. Berlin, 1990.

[3] V.Gullapalli," A Stochastic Reinforcement Learning Algorithm for Learning Real-Valued Functions", Neural Networks, Vol.3, pp.671-692",1990.

[4] V.Gullapalli and J,A,Franklin and H.Benbrahim, Acquiring Robot Skills via Reinforcement Learning, IEEE Control Systems Magazine,pp.13-24, February 1994.

[5] R.S.Sutton and A.G.Barto, eds.,Reinforcement Learning, MIT Press, Cambridge,1998.

[6] H.Benbrahim and J.A.Franklin","Biped Dynamic Walking using Reinforcement Learning", Robotics and Autonomous Systems", Vol.22, pp.283-302, Decemeber 1997.

[7] A.W.Salatian and K.Y.Yi and Y.F.Zheng", "Reinforcement Learning for a Biped Robot to Climb Sloping Surfaces", Journal of Robotic Systems", Vol.14, No.4, pp,283-296, April,199.

[8] C.Zhou and Q.Meng,"Reinforcement Learning and Fuzzy Evaluative Feedback for a Biped Robot",Proceedings of the 2000 IEEE International Conference on Robotics and Automation", San Francisko,pp. 3829-3834, April 2000.

[9] J.Peters and S.Vijayakumar and S.Schaal," Reinforcement Learning for Humanoid Robots", Proceedings of the Third IEEE-RAS International Conference on Humanoid Robots, Karlsruhe & Munich", October 2003.

[10] H.R.Berenji and P.Khedkar,"Learning and Tuning Fuzzy Logic controllers through Reinforcements",IEEE Transactions on Neural Networks", Vol.3,No.5",pp.724-740, September 1992.