UDK: 007.52
COSATI: 06-04, 14-07

Reinforcement learning control algorithm for humanoid robot walking

Duško Katić, PhD (Eng)[1)]

The integrated dynamic control of humanoid locomotion mechanisms based on the spatial dynamic model of humanoid mechanism is presented in this paper. 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 the modified version of *GARIC* architecture for dynamic reactive compensation. Simulation experiments were carried out in order to validate the proposed control approach.

Key words: **robotics, humanoid robot, locomotion system, dynamically balanced gait, reinforcement learning.**

Introduction

HAVING in mind that humanoid robot must satisfy very high requirements , the need to increase the number of degrees of freedom (DOFs) of their mechanical configuration should be pointed out, to study in more depth some previously unconsidered phenomenona 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 in the case of regular (stationary) gait with the presence of small perturbations and in the case of robot's posture with the presence of large perturbations. It should also be pointed out that the problem of motion of humanoid robots is a very complex control task, especially when the real environment is taken into account, which, as a 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 comprehensive 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 the algorithm is the 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 [3-5]. Reinforcement learning concept is based on trial and error methodology with 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. To begin with there are several approaches [6-9] with additional demands and requirements due to the 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 methods using Temporal Difference (T*D*) learning [5]. The Actor network can be considered 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, the two-layer feedforward neural network with sigmoid hidden units and linear output units is selected. 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. The network is trained via Back propagation (gradient descent) algorithm and using training example provided by Critic net. The implemented algorithm was based on modified version of GARIC approach, that was presented in paper [10]. In this paper, the external reinforcement signal was simply defined as measure of ZMP error. Internal reinforcement signal is generated using external reinforcement signal and appropriate policy.

Model of the system

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

[1)] Institute Mihailo Pupin, Robotics Laboratory, Volgina 15, 11060 Belgrade

chains in their interaction with the environment transform from open to closed type of chain [2]. In Fig.1, the kinematic scheme of the biped locomotion mechanism [2] is shown, 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 mechanism has in total $n = 20$ DOFs of motion.

The dynamic model of the mechanism 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 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 that successively determine angular and translational velocities and accelerations of particular links of the robotic mechanism are formed [1]. Taking into account the dynamic coupling between particular parts (branches) of the mechanism chain, can be derive the relation that describes the dynamic model of the locomotion mechanism in the form [2]:

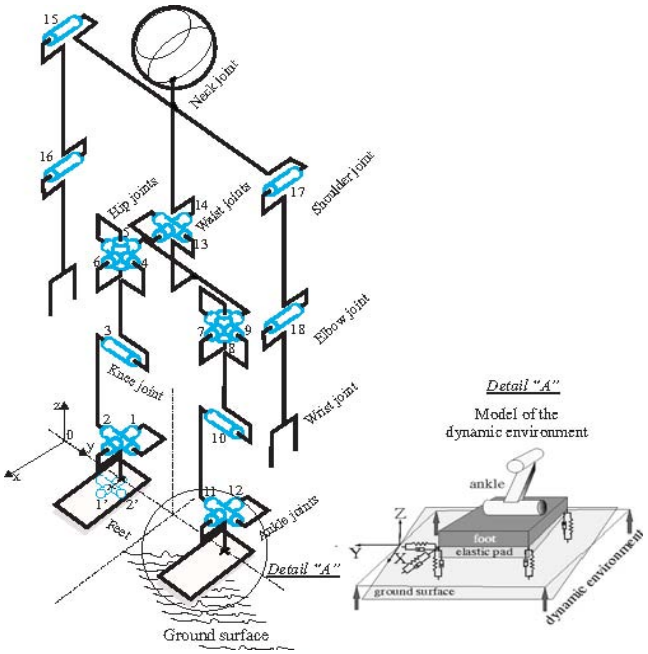
$$P = H(q)\ddot{q} + h(q,\dot{q}) \qquad (1)$$

where: $P \in R^{nx1}$ is the vector of driving moments at the humanoid robot joints; $F \in R^{6X1}$ s the vector of external forces and moments acting at the particular points of the mechanism; $H \in R^{nxn}$ is the square matrix that describes inertia matrix of the mechanism shown in Fig.1; $h \in R^{nx1}$ is the vector of gravitational, centrifugal and Coriolis moments acting at n mechanism joints; $n = 20$ is the total number of DOFs (Fig.1). The force F has a special importance for calculation the model (1), representing 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 distinguishes between the so-called supporting or constrained foot and unconstrained foot, which in the moment of contact with the ground, is transformed into the con-strained one. In this paper, the primary concern is to consider the contact of rigid foot with ground and walking on slightly horizontal plane.

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

When criterion is met it (i) enables the realization of the desired mode of motion, walk repeatability and avoiding the potential obstacles in the way. To satisfy criterion (ii) implies having a stable balanced walk. Fulfilling criterion (iii) ensures a higher degree of stability of the overall system in respect of the impact appearing at the moment when the unconstrained foot strikes the ground. 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.



a) b)

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

Gait phases and indicator of dynamic balance

The robot biped 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 balance 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 control 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.

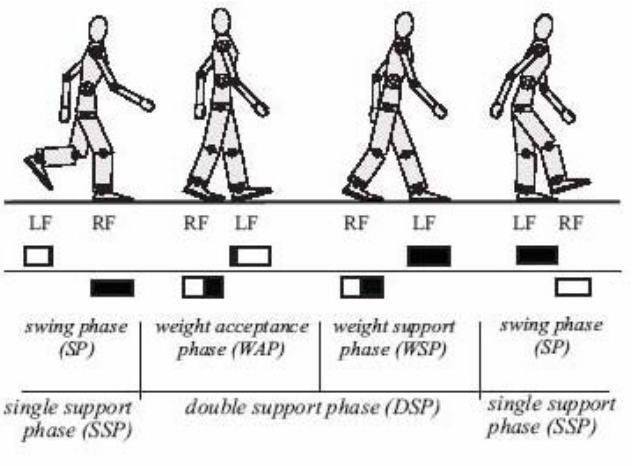


Figure 2. Phases of biped gait

Dynamic Integrated Control algorithm

In accordance with the control task, the application of the algorithm of the so-called integrated dynamic control, based on the knowing of the overall dynamic model of the system is proposed. At that, it is assumed that the following assumptions hold: (i) the model (1) describes sufficiently well the behaviour 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 the block-diagram of the dynamic controller for biped locomotion mechanism is presented. 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 the driving moment $\hat{P}$ represents the sum of the driving moments $\hat{P}_1$ and $\hat{P}_2$. The moment $\hat{P}_1$ are determined so to ensure the 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.

Controller of trajectory tracking

The controller for tracking of nominal trajectory has to ensure the realization of a desired motion of the humanoid robot and to avoid 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 computed torque method in the space of internal coordinates of the mechanism joints. For this purpose, the robot dynamic model defined by the relation (1) was used. The control law can be expressed in the known form:

$$P = H(q)\left[\ddot{q}^0 + K_p\left(q - q^0\right) + K_v\left(\dot{q} - \dot{q}^0\right)\right] + h(q,\dot{q}) \quad (2)$$

where *H*, *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 $K_p \in R^{nxn}$ and $K_v \in R^{nxn}$ 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.

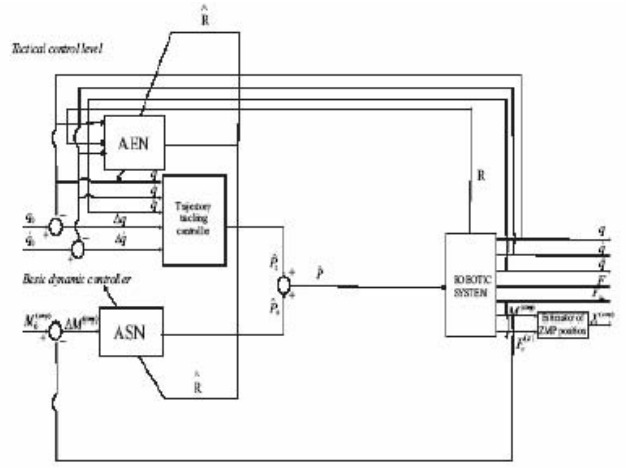


Figure 3. Block-scheme of the integrated dynamic control of biped with two feedback loops

GARIC Compensator of dynamic reactions

From the view point 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 the introduce dynamic reaction feedback at ZMP in the control synthesis. There is a relationship between the deviations of ZMP positions ($\Delta x(zmp)$, $\Delta y(zmp)$) from its nominal position 0zmp 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}. $M_x^{(zmp)} \in R1\times1$ and $M_y^{(zmp)} \in R1\times1$; they 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 is defined according to values of ZMP error. If ZMP error is greater than the 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 the scalar value (internal reinforcement $\hat{R}$) which represent the quality of the given control task is defined by the following policy:

$$\hat{R}(t+1) = R(t) + yv(t+1) - v(t) \quad (3)$$

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 the recommended control torque. Accordingly, by using SAM (Stochastic action modifier), based on the 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 via modified version of back propagation algorithm where the error is defined by internal reinforcement signal $\hat{R}$. In the same way, using gradient method and internal reinforcement signal, learning process of ASN is realized. $\Delta M^{(zmp)} \in R^{2\times1}$ is the vector of deviation of the actual dynamic reactions from their nominal value. $P_{dr} \in R^{2\times1}$ 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. Therefore, the control action is displaced to the other, powered joints of the mechanism chain. Since the vector of deviation of the dynamic reactions $\Delta M^{(zmp)}$ has two components around 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 around the *x*-axis and 2, 4 and 13 to compensate for the moments around 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 $\hat{P}_2$ is calculated using the GARIC structure, beaving in mind how many compensational joints are actually 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 the selected joints. In nature, biological systems use simultaneously a large number of joints for correcting their balance. However, for the purpose of verifying the control algorithm, this work has restricted the choice to the mentioned six joints only: 1, 2, 4, 6, 13 and 14 (Fig.1).

Simulation experiments

Theoretical results presented previously were analysed 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 $m = 70\,[\mathrm{kg}]$ and its geometric and dynamic parameters were taken from paper [2]. Simulation examples are concerned with the characteristic pattern of artificial gait in which the mechanism makes a half-step of the length $l = 0.40[\mathrm{m}]$ in the time period of $t = 0.75\,[\mathrm{s}]$. Nominal trajectories at robot joints are synthesized for the gait in the horizontal plane. The simulation results were analysed in the time interval corresponding 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. Therefore the analysis of dynamic robot behaviour in these time intervals is especially important for control, so that the simulation examples were selected to encompass these critical phases. In this simulation example the assigned initial deviations of particular angles at mechanism joints did not exceed $\Delta q_i \le 10^\circ$. 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. The robot's behaviour in the swing phase was observed (Fig.2), when the robot relies on the ground with its rigid foot while the other (free) foot is above the ground surface. At that, two cases of control were analysed: (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 Figures 4-6 the results of applying the controller in case (ii) are presented. Analyzing the results presented in Figures 4-5, it can be seen that better results for error of ZMP are obtained 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 the dynamic balance of the locomotion mechanism in its motion. This results from the fact that the nominal trajectory was synthesized without taking into account the possible deviations of the surface on which a biped walks on an ideally horizontal plane. Therefore, the ground surface inclination influences the system's balance as an external stochastic dis-turbance.

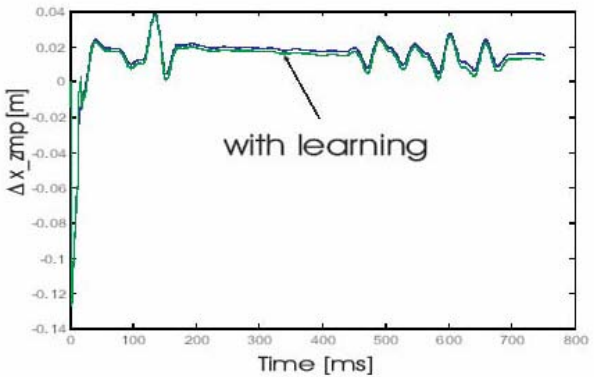


Figure 4. Error of ZMP in *x*-direction

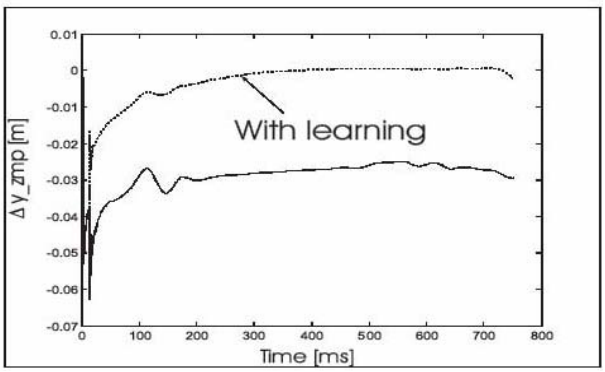


Figure 5. Error of ZMP in *y*-direction

In Fig.6 the corresponding deviations (errors) Δqi from of the real values of angles their nominal values at the robot's joints are presented, when tracking of desired trajectory was applied. The deviations of the variables converge to a zero value in the given time interval, i.e. the controller employed ensures good tracking of the desired trajectory. In Fig.7 the value of internal reinforcement through the process of walking is presented. It is clear that the task of walking within the desired ZMP tracking error limits is achieved.

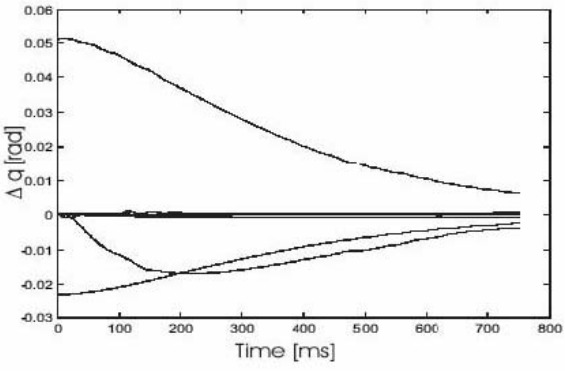


Figure 6. Position tracking errors for compensation joints

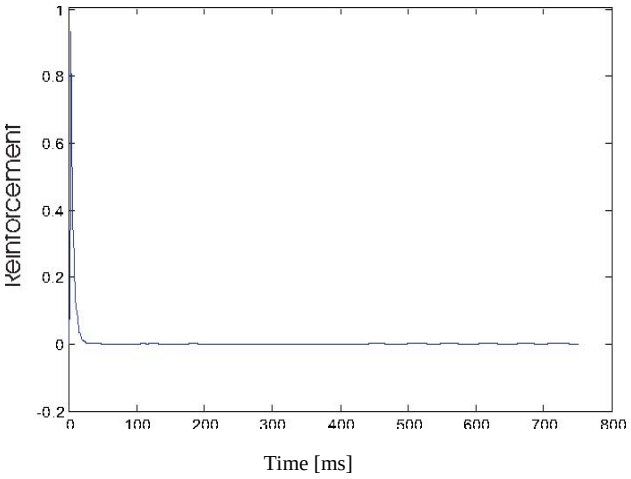


Figure 7. Internal reinforcement through the process of walking

Conclusions

The control scheme of an integrated dynamic controller of locomotion mechanism was synthesized. Control level consists of a dynamic controller for tracking robot's nominal trajectory and a compensator of dynamic reactions of the ground around the ZMP based on the GARIC reinforcement learning architecture. 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. The implemented algorithm was based on the modified version of the GARIC approach. The external reinforcement signal was defined simply to be the measure of ZMP error. Internal reinforcement signal is generated using external reinforcement signal and the appropriate policy. The presented shows that better results in tracking errors are obtained when algorithm with training of ASN neuro-fuzzy network is used.

References

[1] JURIČIĆ,D., VUKOBRATOVIĆ, M.: *Mathematical Modeling of Biped Walking Systems*, ASME Publication 72 –*WAlBHF–13*, 1972.

[2] VUKOBRATOVIĆ,M., BOROVAC,B., SURLA,D., STOKIĆ,D., Biped Locomotion: *Dynamics,Stability,Control and Application*,. Springer-Verlag. Berlin, 1990.

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

[4] GULLAPALI,V., FRANKLIN,J.A., BENBRAHIM,H.: *Acquiring Robot Skills via Reinforcement Learning*, IEEE Control Systems Magazine, February 1994, pp.13-24.

[5] SUTTON,R.S., BARTO,A.G.,eds.: *Reinforcement Learning*, MIT Press, Cambridge, 1998.

[6] BENBRAHIM,H., FRANKLIN,J.A.: *Biped Dynamic Walking using Reinforcement Learning*, Robotics and Autonomous Systems, December 1997, Vol.22, pp.283-302.

[7] SALATIAN,A.W., YI,K.Y., ZHENG,Y.F.: *Reinforcement Learning for a Biped Robot to Climb Sloping Surfaces*, Journal of Robotic Systems, April,1999, Vol.14, No.4, pp,283-296.

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

[9] PETERS,J., VIJAYAKUMAR,S., SCHAAL,S.: *Reinforcement Learning for Humanoid Robots*, Proceedings of the Third IEEE-RAS International Conference on Humanoid Robots, Karlsruhe & Munich", October 2003.

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

Received: 15.07.2005.

Upravljački algoritam zasnovan na primeni reinforcement učenja za kretanje humanoidnog robota

U radu se razmatra integrisano dinamičko upravljanje humanoidnim lokomocionim mehanizmima zasnovano ma prostornom dinamičkom modelu humanoidnog mehanizma.- Upravljačka šema je sintetizovana koristeći centralizovani model sa pretpostavljenom strukturom dinamičkog kontrolera koji uključuje dve povratne sprege: povratne sprege po poziciji i brzini zglobova robotskog mehanizma povratnu spregu zasnovanu na "reinforcement" u;enju oko Tačle Nula Momenta. Predloženi algoritam "reinforcement" učenja se zasniva na modifikovanoj verziji GARIC arhitekture za sinamičku teaktivnu kompenzaciju. Izvršeni su simulacioni eksperimenti u cilju verifikacije predložene upravljačke šeme.

Ključne reči: **robotika, humanoidni robot, lokomotorni sistem, dinamika kretanja, algoritam upravljanja, dinamičko upravljanje, inteligentno upravljanje, povratna veza, blok šema.**

Управляющий алгорифм обоснован на применении "Реинфорцемент" учения для движения интеллектуального робота

В настоящей работе рассматривается динамическое управление интеллектуальными локодвигательными механизмами, обосновано на просторной динамической модели интеллектуального механизма. Схема управления синтезирована пользованием централизованной модели с предположенной структурой динамического контроллера, включающего две обратные связи: обратные связи по позиции и по скорости зглобов (шарниров) механизма робота и обратную связь, обоснованную на "Реинфорцемент" учению около "Зеро Момент Поинт" (ЗМП). Предложенный алгорифм "Реинфорцемент" учения обосновывается на модифицированной версии "ГАРИЦ" архитектуры для динамической реактивной компенсации. Тоже выполнены эксперименты моделирования с целью подтверждения правильности предложенной схемы управления.

Ключевые слова: **робототехника, интеллектуальный робот, локодвигательная система, динамика движения, алгорифм управления, динамическое управление, интеллектуальное управление, обратная связь, блок-схема.**

Algorithme de commande basé sur l’application du pricipe “Reinforcement” pour la marche du robot humanoide

Ce papier traite la commande dynamique intégrée chez la mécanisme locomoteur humanoide, basée sur le modèle spatial dynamique du mécanisme humanoide. Le schéma de commande est synthétisé à l’aide du modèle centralisé avec la structure supposée du contrôleur dynamique qui implique deux réactions: réactions quant à la position et à la vitesse des articulations du mécanisme de robot et réactions basée sur le principe reinforcement autour du point zéro-moment. L’algorithme proposé du principe reinforcement est basé sur la vérsion modifiée de l’architecture GARIC pour la compensation dynamique réactive. On a effectué les essais de simulation dans le but de vérifier les schémas proposées de commande.

Mots clés: **robotique, robot humanoide, système locomoteur, dynamique du mouvement, algorithme de commande, commande dynamique, commande intelligente, réactions, le schéma bloc.**