Control Algorithm for Humanoid Walking Based on Fuzzy Reinforcement Learning

Duško Katić, Miomir Vukobratović

Mihajlo Pupin Institute, Robotics Laboratory, Volgina 15, Belgrade, 11060,
Serbia, E-mail: dusko,vuk@robot.imp.bg.ac.yu

Abstract: In this paper hybrid intelligent control algorithm for biped locomotion is presented. The proposed structure of controller involves two feedback loops: model-based dynamic controller and fuzzy reinforcement learning feedback around Zero-Moment Point. The proposed new reinforcement learning algorithm is based on modified version of actor-critic architecture for dynamic reactive compensation. The reinforcement learning architecture as external reinforcement use fuzzy evaluative feedback. Simulation experiments were carried out in order to validate the proposed control approach.The obtained numerical results served as the basis for a critical evaluation of the controller performance.

Keywords: Humanoid robots; biped locomotion; Intelligent Contrl; Reinforcement learning; Actor -Critic method, Fuzzy Logic

1 Introduction

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. 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 porssibilities of the reinforcement learning with fuzzy evaluative feedback, as it was investigated in paper [1]. Fuzzy reinforcement learning generalizes reinforcement learning to fuzzy environment where only the fuzzy reward function is available.

In this paper, a novel, integrated hybrid dynamic control structure for the humanoid robots is proposed, using the off-line line and on-line calculated complete 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 modified GARIC reinforcement learning architecture [2], [3], [4] for dynamic compensation of ZMP (Zero-Moment-Point) error. The goal of this paper is to propose the usage of fuzzy reinforcement learning for humanoid robotics. The reinforcement learning method proposed in this paper is based on the Actor-Critic architecture. In this paper, the external reinforcement signal was defined to be measure of ZMP error based on fuzzy linguistic variables. Internal reinforcement signal is generated using external reinforcement signal and appropriate stochastic gradient policy. The fuzzy evaluative feedback has aim to evaluate the degree of success for the biped dynamic walking by means of ZMP (zero moment point).

2 The Model of the System

2.1 Model of the Robot's Mechanism

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. Taking into account dynamic coupling between particular parts (branches) of the mechanism chain, one can derive a relation that describes the overall dynamic model of the locomotion mechanism in a vector form [5]:

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

where: P is the vector of driving torques at the humanoid robot joints; F is the vector of external forces and moments acting at the particular points of the mechanism; H is the square matrix that describes 'full' inertia matrix of the mechanism; h is the vector of gravitational, centrifugal and Coriolis moments acting at n mechanism joints; J is the corresponding Jacobian matrix of the system; n =20, is the total number of DOFs; q is the vector of internal coordinates.

2.2 Gait Phases and Indicator of Dynamic Balance

The robot's bipedal gait consists of several phases that are periodically repeated [5]. Hence, depending on whether the system is supported on one or both legs, two macro-phases can be distinguished, viz.: (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).

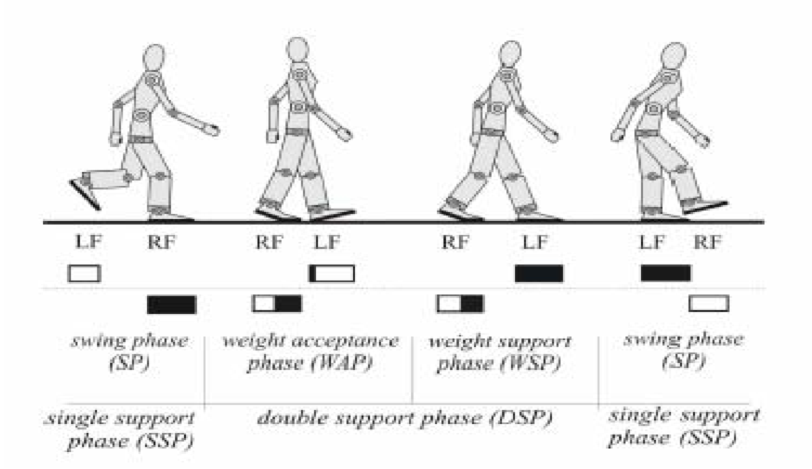

LF
RF
RF
LF
RF
LF
LF
RF
swing phase (SP)
weight acceptance phase (WAP)
weight support phase (WSP)
swing phase (SP)
single support phase (SSP)
double support phase (DSP)
single support phase (SSP)

Figure 1
Phases of biped gait

Fig. 1 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. 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 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. The deviations $\Delta x^{(zmp)}$ and $\Delta y^{(zmp)}$ of the ZMP position from its nominal position in x-and y-direction have a great influence in control synthesis. The instantaneous position of ZMP is the best indicator of dynamic balance of the robot mechanism. The ZMP position inside these 'safety areas' ensures a dynamically balanced gait of the mechanism [5] whereas its position outside these zones indicates the state of loosing 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.

3 Hybrid Intelligent Control Algorithm with Fuzzy Reinforcement Learning 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, ZMP position. In accordance with the control task, we propose the application of the algorithm of the so-called intelligent control based on the dynamic model of the complete system.

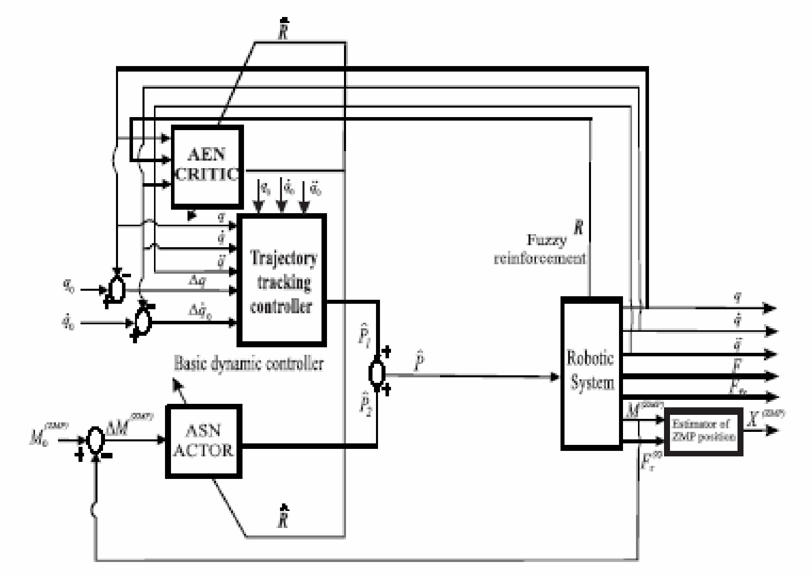


Figure 2
Block-scheme of the hybrid intelligent control of biped

In Fig. 2 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 fuzzy reinforcement learning structure. The synthesized dynamic controller was designed on the basis of the centralized model. The vector of driving moments $\hat{P}\hat{}$ 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 of driving torques $\hat{P}$ represents the output control vector.

3.1 Dynamic 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 [28], it has been demonstrated how local PD or PID controllers of biped locomotion robots are being designed. The proposed dynamic control law ha the following form:

$$\begin{aligned}\hat{P} = \hat{H}(q,x_{cf})[\ddot{q}_0 + K_V(\dot{q}-\dot{q}_0) + K_p(q-q_0)] + \hat{h}(q,\dot{q},x_{cf},\dot{x}_{crf},\ddot{x}_{cf}) \\ -J^T(q,x_{cf})\end{aligned} \tag{2}$$

where $\hat{H}, \hat{h}$ and $\hat{j}\hat{}$ 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 and K_V 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.

3.2 Compensator of Dynamic Reactions based on Fuzzy 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. 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}. $M_X^{(zmp)}$ and $M_Y^{(zmp)}$ represent the moments that tend to overturn the robotic mechanism, i.e. to produce its rotation about the mentioned rotation axes. Nominal values of dynamic reactions, for the nominal robot trajectory, are determined off-line from the mechanism model (1) and the relation for calculation of ZMP; $\Delta M^{(zmp)}$ is the vector of deviation of the actual dynamic reactions from their nominal values; P_{dr} is the vector of con trol torques at the joints *1'* and *2'*, ensuring the state of dynamic balance; On the basis of the above the fuzzy reinforcement control algorithm is defined with respect to the dynamic reaction of the support at ZMP.

3.2.1 Fuzzy 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 [3], [6]. 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 is scalar signal that drives all learning in both actor and critic (Fig. 3).

Practically, in proposed humanoid robot control design, it is synthesized the new modified version of GARIC reinforcement learning structure [3]. 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 using fuzzy inference algorithm. Proposed learning structure is based on 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 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.

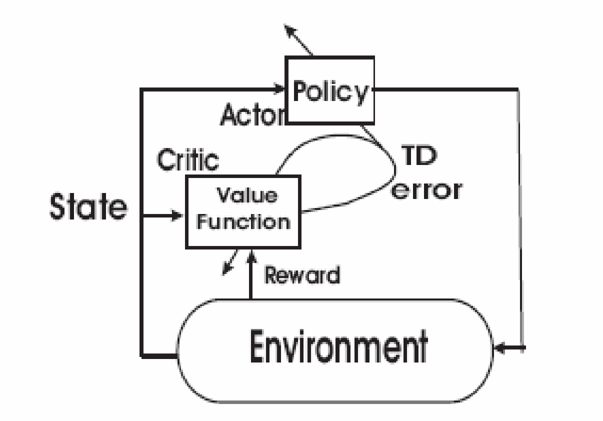


Figure 3
The Actor-Critic Architecture

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:

$$v = \sum_{ii} B_I \Delta M_i^{(zmp)} + \sum C_j f(\sum_{ji} A_I \Delta M_i^{(zmp)}) \tag{3}$$

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:

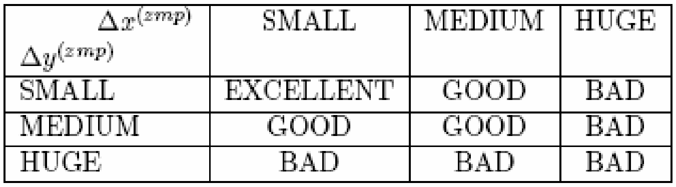
$$\hat{R}\,(\mathrm{t} + 1) = 0, \qquad \text{begining state} \tag{4}$$

$$\hat{R}\,(\mathrm{t} + 1) = \mathrm{R(t)} - \mathrm{v(t)}, \qquad \text{failure state} \tag{5}$$

$$\hat{R}\,(\mathrm{t} + 1) = \mathrm{R(t)} + \gamma \mathrm{v(t + 1)} - \mathrm{v(t)}, \qquad \text{otherwise} \tag{6}$$

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

The most important part of algorithm represent the choice of reward function - external reinforcement. It is possible to use scalar critic signal [11], but in this paper 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.



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

Table 1
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. 4.

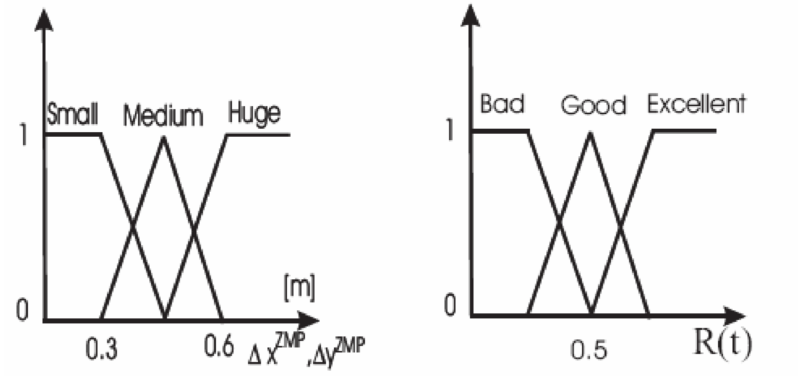


Figure 4
The Membership functions for ZMP deviations and external reinforcement

ASN (action selection network) maps the deviation of dynamic reactions $\Delta M^{(zmp)}$ in recommended control torque. The structure of ASN is represented by The ANFIS -Sugenotype adaptive neural fuzzy inference systems. There are five layers: input layer. antecedent part with fuzzification, rule layer, consequent layer, output layer wit defuzzification. This system is based on fuzzy rule base generated by expert kno0wledge 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 *Pdr*. It is defined by Gaussian random function where recommended control torque is mean, while standard deviation is defined by following equation:

$$\sigma(\hat{R}(t+1)) = 1 - \underset{\_}{\ } \exp(\left|\hat{R}(t+1)\right|) \tag{7}$$

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 The learning process for ASN (tuning of antedecent 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.

The control torques *Pdr* obtained as output of actor structure cannot be generated at the joints *1'* and *2'* since they are unpowered (passive) joints. Hence, the control action has to be '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, 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 joints of ankle, hip and waist were taken into consideration. Finally, the vector of compensation torques $\hat{P}2$ (Fig. 2) was calculated on the basis of the vector of the moments *Pdr* whereby it has to be borne in mind how many 'compensational joints' have really been engaged.

4 Simulation Studies

The proposed hybrid intelligent control method, presented in previous section, were analyzed on the basis of numerical data obtained by simulation of the closed-loop model of the locomotion mechanism. Total mass of the mechanism was *m* = 70 [*kg*]. Its geometric and dynamic parameters were taken from the literature [5]. Simulation examples concerned characteristic patterns of artificial gait in which the mechanism makes a half-step of the length l = 0.40 [m] in the time period T = 0.75 [s]. Nominal motion of the robot mechanism represents a walking on the ideally flat, horizontal surface during a half-step phase of the gait. Nominal trajectories at robot joints were synthesized for the gait on the horizontal ground surface. Nominal angles at the mechanism joints and the corresponding angular velocities and accelerations were determined by the semi-inverse method [28]. Initial conditions of the simulation examples (initial deviations of joints' angles) were imposed by a definition of the following deviation vectors:

$$\Delta q = [0\ 0\ 0\ 0.051\ 0\ \ -0.023\ 0.034\ 0\ \ -0.02\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0]\ [rad].$$

In the simulation example, the assigned initial deviations of particular angles Δq at the mechanism joints were assumed to be as large as it was previously emphasized. Constant, small inclinations of the ground surface in the sagittal plane $\gamma 1 = 3_{\circ}$ as well as in the frontal plane $\gamma 2 = 2_{\circ}$ were introduced as an additional perturbation, too. Thus the simulation experiment dealt with the real case of walking on a quasi-horizontal ground. The issue of interest was the robot's behavior in the swing phase (Fig. 1), when the robot relies upon the ground by its rigid foot. The other one (free or swinging foot) was above the ground. Here, the control was realized in using control algorithm consisting of the basic dynamic trajectory tracking controller and reinforcement learning compensator of dynamic reactions of the ground at the ZMP. In Figs 5 and 6, the errors of ZMP position in x and y direction are shown. Thus, it can be concluded that errors of ZMP position are in required polygons, and in the case of absence of the reinforcement learning feedback with respect to the ground reactions at the ZMP it is not generally possible to ensure (guarantee) dynamic balance of a locomotion mechanism during its motion. This comes out from the fact that the pre-designed (nominal) trajectory was synthesized without taking into account possible deviations of the surface inclination on which biped walks from an ideally horizontal plane. Therefore, the ground surface inclination influences the system's balance as an external stochastic perturbation. The corresponding deviations (errors) Δq of the actual angles at the robot joints from their reference values, in the case when the controller of tracking desired trajectory was applied, are presented in Fig. 7. The deviations Δq converge to zero values in the given time interval T=0.75 [s].

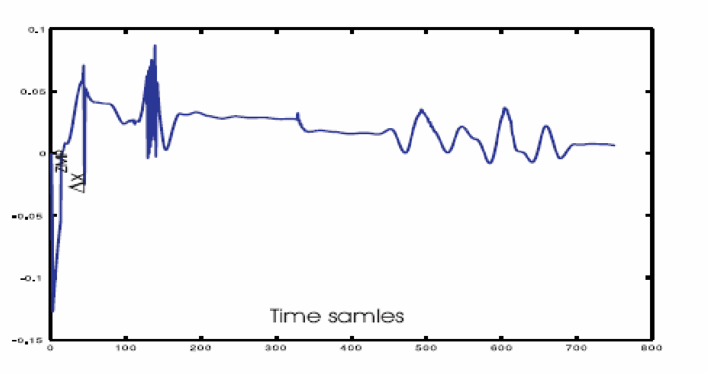


Figure 5
Error of ZMP in x-direction

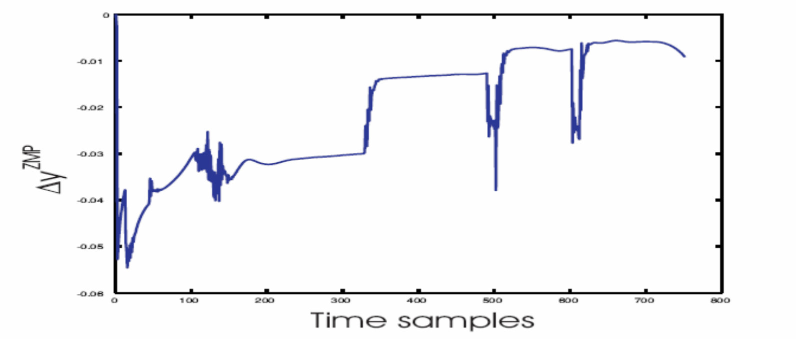


Figure 6
Error of ZMP in y-direction

It means that the controller employed ensures good tracking of the desired trajectory. well as a dynamic balance of the locomotion mechanism as it is illustrated in Fig. 7.

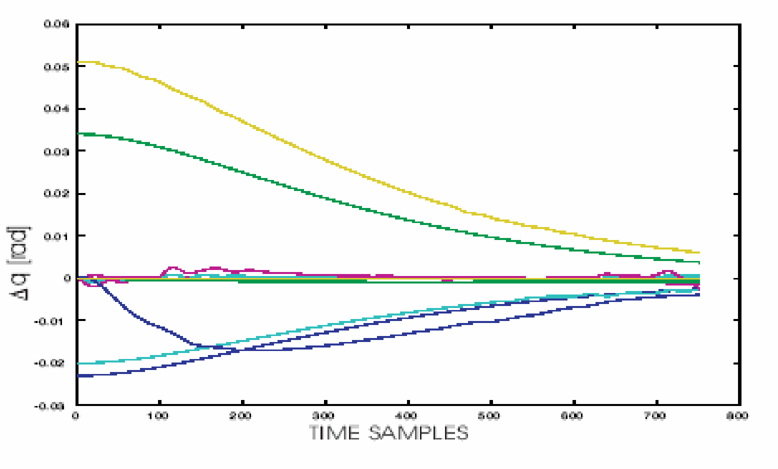


Figure 7
Convergence of the errors of tracking nominal angles

In Fig. 8 value of reinforcement signal through process of walking is presented. It is clear that task of walking within desired ZMP tracking error limits is achieved.

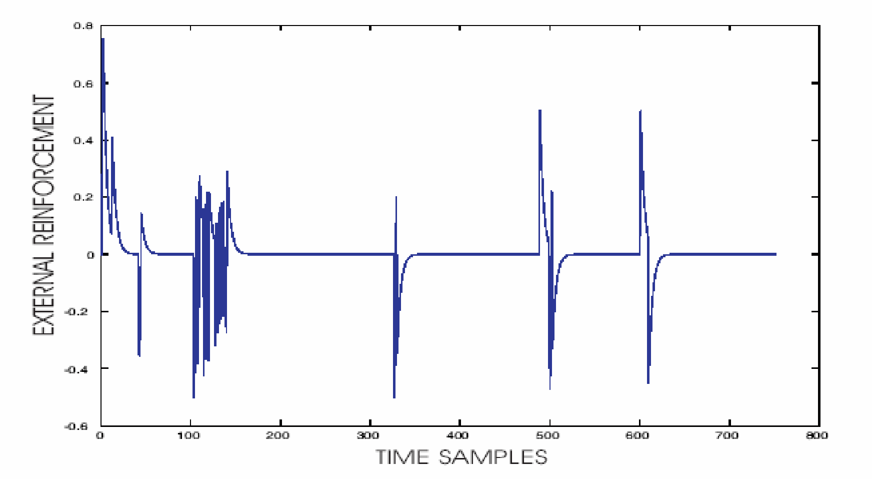


Figure 8
Reinforcement through process of walking

Conclusions

In this study, an hybrid intelligent controller for biped locomotion was proposed. The proposed scheme was designed by using the centralized dynamic model. 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 new fuzzy actor-critic reinforcement learning architecture. The proposed reinforcement learning structure based on AEN (neural network) and ASN (neuro-fuzzy network), represents the efficient learning tool for compensation of ZMP reactions. The algorithm is based on fuzzy evaluative feedback that are obtained from human intuitive balancing knowledge. On this way, biped then accumulate dynamic balancing knowledge through reinforcement learning, and thus constantly improve its gait during walking. The reinforcement learning with fuzzy evaluation feedback is much closer to the human biped walking evaluation than the original one with numerical feedback.

References

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

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

[3] Berenji, H. R., Khedkar, P.: Learning and Tuning Fuzzy Logic Controllers through Reinforcements, IEEE Transactions on Neural Networks, 3, 1992, pp. 724-740

[4] Salatian, A. W., Yi, K. Y., Zheng, Y. F.: Reinforcement Learning for a Biped Robot to Climb Sloping Surfaces, Journal of Robotic Systems, 14, 1997, pp. 283-296

[5] Vukobratovi´c, M., Borovac, B., Surla, D., Stoki´c, D.: Biped Locomotion -Dynamics, Stability, Control and Application, Springer Verlag, Berlin, Germany, 1990

[6] Sutton, R. S., Barto, A. G.: Reinforcement Learning: An Introduction, The MIT Press, Cambridge, USA, 1998