Proceedings of the 17th World Congress
The International Federation of Automatic Control
Seoul, Korea, July 6-11, 2008

Dynamic Control Algorithm for Biped Walking Based on Policy Gradient Fuzzy Reinforcement Learning ⋆

Duško M. Katić * **Aleksandar D. Rodić** **

* *Robotics Laboratory, Mihailo Pupin Institute, Volgina 15, 11060 Belgrade, Serbia (Tel: +381-11-2776174; e-mail: dusko@robot.imp.bg.ac.yu).*
** *Robotics Laboratory, Mihailo Pupin Institute, Volgina 15, 11060 Belgrade, Serbia (Tel: +381-11-2776174; e-mail: roda@robot.imp.bg.ac.yu).*

Abstract: This paper presents a novel dynamic control approach to acquire biped walking of humanoid robots focussed on policy gradient reinforcement learning with fuzzy evaluative feedback . The proposed structure of controller involves two feedback loops: conventional computed torque controller including impact-force controller and reinforcement learning computed torque controller. Reinforcement learning part includes fuzzy information about Zero-Moment Point errors. To demonstrate the effectiveness of our method, we apply it in simulation to the learning of a biped walking.

1. INTRODUCTION

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 case, even if stable trajectories are used, the existence of impulse disturbances on foots sole can make robot to tumble. 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 behavior of the robot. However, it is very difficult to control robots with human generated, preprogrammed, learned behavior. Learned behavior should be acquired by the robots themselves in a human-like way, not programmed manually. Humans learn actions by trial and error procedure or by emulating someone else's actions. Hence, therefore reinforcement learning could be applied for the control of humanoid robots because this process resembles a human's trial and error learning process through constant evaluation of performance in constant interaction with environment.

In area of humanoid robotics, there are several approaches of reinforcement learning (RL) (Benbrahim and Franklin [1997], Mori et al. [2004], Nakamura et al. [2003], Peters et al. [2003], Tedrake et al. [2004]) 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 (Benbrahim and Franklin [1997]).

In this paper, we use a policy-gradient method for learning efficient biped motion. The policy-gradient method is a kind of reinforcement learning method which maximizes the average reward with respect to parameters controlling action rules known as the policy (Shibata et al. [2007], Tedrake et al. [2004], Peters et al. [2003]). In comparison with most standard value function-based reinforcement learning methods, this type of method has particular features suited to robotic applications. Firstly, the policy-gradient method is applicable to Partially Observable Markov Decision Processes]. It is almost impossible to consider all possible states of the robot because even if it has a complete set of sensors there will be a degree of noise. The using of gradient-policy enables smooth change of parameters, stability of algorithm, incorporation of prior and incomplete information in control process.

The new integrated dynamic control structure for the humanoid robots is proposed, based on model of robot mechanism. Our approach consists in inclusion of reinforcement learning part only for compensation joints. The basic part of control algorithm represents computed torque control method. The external reinforcement signal was simply defined as fuzzy measure of Zero-Moment-Point (ZMP) error). Internal reinforcement signal is generated using external reinforcement signal and appropriate Critic network. The Critic network provides policy evaluation and can be used to perform policy improvement. For the Critic network, the two layer neural 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.

⋆ This work was supported by Ministry of Science and Environmental Protection of the Republic of Serbia under the national research project from area of robotics.

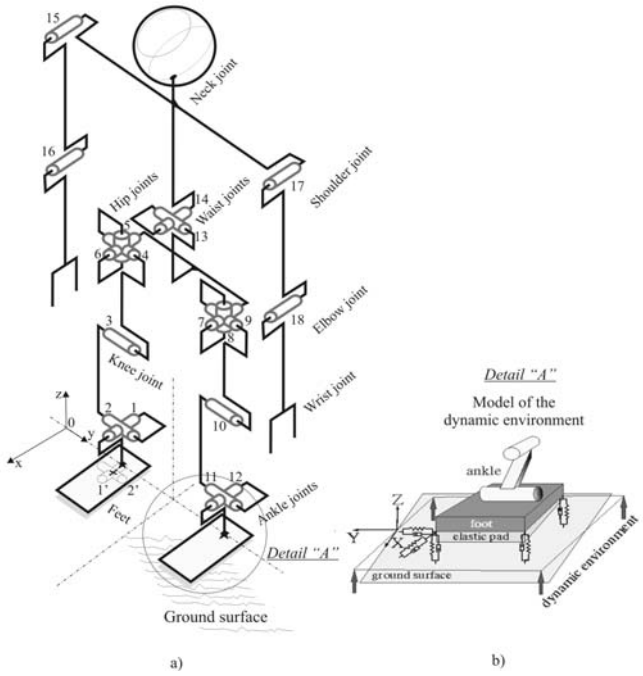

Fig. 1. Model of the biped locomotion mechanism

2. DYNAMIC MODEL OF THE SYSTEM AND CONTROL REQUIREMENTS

2.1 Model of the Robot's Mechanism

The kinematic scheme of the biped locomotion mechanism whose spatial model will be considered in this paper, is shown in Fig. 1a. The mechanism possesses 18 powered DOFs, designated by the numbers *1-18*, and two underactuated DOFs (*1'* and *2'*) for the footpad rotation about the axes passing through the instantaneous ZMP position. The overall dynamic model of the locomotion mechanism is represented in the following vector form:

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

where: $P \in R^{n\times 1}$ is the vector of driving torques 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; $J \in R^{6\times n}$ is the corresponding Jacobian matrix of the system; $n = 20$, is the total number of DOFs (Fig. 1); $q \in R^{n\times 1}$ is the vector of internal coordinates; $\dot{q} \in R^{n\times 1}$ is the vector of internal velocities. Exactly, the relation (1) represents the model of a biped mechanism relying on the absolutely rigid environment. In a general case, the constrained foot of biped locomotion mechanism has corresponding linear and rotational relative motions with respect to the fixed coordinate system attached to the ground. As example, it is case of robot's walking on an immobile support (plain surface, staircases, etc.) involving passive compliance elements in the form of elastic pads on the feet (Fig.1b).In Fig.1b is illustrated an example of spatial dynamic environment model suitable for describing the contact elastodynamics. Linear and angular deformations/displacements of the elastic pad attached to the foot sole influence the dynamic behavior of the biped mechanism. The velocity $\dot{x}_{cf}$ and acceleration $\ddot{x}_{cf}$ of the constrained foot (due to the elastic properties of the pad and ground surface) are transferred to all of the links of the robotic mechanism. Taking into account the considered transitive motion and relative position of the constrained foot x_{cf} with respect to the immobile fundament, as well as using the vector of measured external forces and moments F , equation (1) can be re-written in a similar form:

$$P + J^T(q, x_{cf})F = H(q, x_{cf})\ddot{q} + h(q, \dot{q}, x_{cf}, \dot{x}_{cf}, \ddot{x}_{cf}) \tag{2}$$

In that sense, equation (2) represents the model of a robot mechanism supported to the *dynamic environment*.

2.2 Gait phases and indicator of dynamic balance

The robot's bipedal gait consists of several phases that are periodically repeated (Vukobratović et.al [1990]). 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). 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 (Vukobratović et.al [1990]) 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 2a and 2b 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 $F_i,\ i = 1, ..., 4$ under the robot foot. Force sensors are usually placed on the foot sole in the arrangement shown in Fig. 2a. 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. 2a), then the following equations to determine the relative ZMP position with respect to its nominal:

$$\begin{aligned}
\Delta M_x^{(zmp)} &= \frac{l_3}{2}\left[(F_2 + F_4) - (F_2^0 + F_4^0)\right] - \\
&\quad \frac{l_3}{2}\left[(F_1 + F_3) - (F_1^0 + F_3^0)\right] \\
\Delta M_y^{(zmp)} &= l_2\left[(F_3 + F_4) - (F_3^0 + F_4^0)\right] - \\
&\quad l_1\left[(F_1 + F_2) - (F_1^0 + F_2^0)\right] \\
F_r^{(z)} &= \sum_{i=1}^{4} F_i \quad \Delta x^{(zmp)} = \frac{-\Delta M_y^{(zmp)}}{F_r^{(z)}} \\
\Delta y^{(zmp)} &= \frac{\Delta M_x^{(zmp)}}{F_r^{(z)}}
\end{aligned} \tag{3}$$

where F_i and F_i^0 , $i = 1, \ldots, 4$, 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

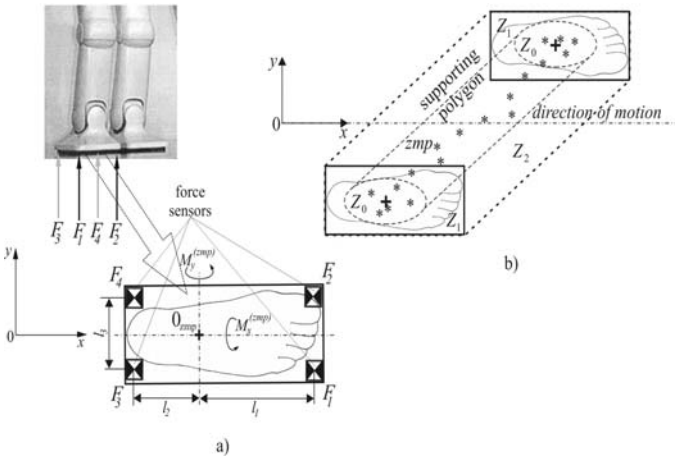

Fig. 2. Zero-Moment Point

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. 2b are illustrated certain areas (Z_0 , Z_1 and Z_2), the so-called *safe zones of dynamic balance* of the locomotion mechanism. The quality of robot balance control can be measured by the success of keeping the ZMP trajectory within the mechanism support polygon as explained above.

3. DYNAMIC CONTROL ALGORITHM WITH POLICY GRADIENT REINFORCEMENT STRUCTURE

In order to enable a balancing controller, the application of the so-called integrated dynamic control was proposed. Based on the above assumptions, the control algorithm involves three parts: (i) basic dynamic controller for trajectory tracking P_1, (ii) dynamic controller tuned by reinforcement learning structure for compensation joints P_2, (iii) impact-force feedback at the foot of the free (unconstrained) leg P_3.

3.1 Dynamic Controller of trajectory tracking

The controller of trajectory tracking of the locomotion mechanism has to ensure the realization of a desired motion of the humanoid robot and avoiding fixed obstacles on its way. As our applied approach, the controller for robotic trajectory tracking was adopted using the computed torque method in the space of internal coordinates of the mechanism joints based of the robot dynamic model. The proposed dynamic control law has the following form:

$$P_1 = \hat{H}(q, x_{cf})[\ddot{q}_0 + K_v(\dot{q} - \dot{q}_0) + K_p(q - q_0)] + \hat{h}(q, \dot{q}, x_{cf}, \dot{x}_{cf}, \ddot{x}_{cf}) - \hat{J}^T(q, x_{cf})F \quad (4)$$

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. 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 $K_p = diag\{k_p^i\}, K_v = diag\{k_v^i\}, i = 1, .., n.$ can be chosen in the diagonal form by which the system is decoupled into $n$independent subsystems. The proposed controller is valid for both the single-support and double-support gait phase, whereby it is assumed that in each instant, at least one foot is in contact with the support surface.

3.2 Compensator of dynamic reactions based on RL structure

Hence in this paper, main intention and idea is to include learning control component based on constant qualitative evaluation of biped walking performance. The reinforcement learning control as kind of unsupervised learning environment (evaluation of control action based on ZMP error rather than numerical error of state variables) can be very suitable for searching of optimal and balanced biped walking.

The main idea is to use chosen control policy (computed torque controller) but with tuning of policy control parameters by appropriate policy-gradient procedure. This reinforcement control part P_2 is realized only for special compensation joints. P_2 is the vector of compensation control torques at the selected compensation joints The control torques P_2 has to be 'displaced' to the other (powered) joints of the mechanism chain. 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.

The proposed Reinforcement learning structure is based on policy gradient Methods (Peters et al. [2003], Shibata et al. [2007], Tedrake et al. [2004]).The policy-gradient method is a stochastic gradient-descent method. The policy can therefore be improved upon every update. In this case, control policy represents computed torque controller structure with aim to select/tune the best control parameters. 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). The general aim of policy optimization in reinforcement learning is to optimize the control parameters policy κ in this way that the expected return

$$J(\kappa) = E\{\sum_{i=0}^{L} \gamma^i r_i \quad (5)$$

is optimized (where $\gamma^i \in [0, 1]$ is a discount factor; r_i is reward or reinforcement signal. It is important to notice that for biped motion, drastic change of control parameter is not valid and and smooth parameter change is required. Hence, policy gradient method based on steepest descent is chosen. The control parameter policy is updated according to the following rule:

$$\kappa_{t+1} = \kappa_k + \alpha \nabla_\kappa J(\kappa) \quad (6)$$

where α is learning rate; $\kappa = (0, 1, 2, ...)$. In this case, policy parameter vector is is defined as $\kappa = [K_p K_v \sigma]^T$, while control policy is defined as

$$\pi = \frac{1}{\sqrt{2\pi}\sigma}(\frac{-(P_2 - \psi(x,\kappa))^2}{2\sigma^2}) \quad (7)$$

where $\psi(x,\kappa) = \kappa^T\phi(x)$;$X$ is the state of the system; $\phi(x)$ is the Gaussian basis function.

There are various methods for gradient estimation Peters et al. [2003], but following algorithms is chosen:

$$\kappa_{t+1} = \kappa_t + \alpha B_t \delta_t \quad (8)$$

$$B_t = \beta B_{t-1} + \nabla log\pi \quad (9)$$

where β is a constant factor; defined by state value function.

The estimated value function represents a *Critic*, because it criticizes the control actions made by the basic controller. Critic 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 Critic is important for calculation of internal reinforcement signal$\hat{R}$. Critic constantly estimate internal reinforcement based on tracking errors and value of reward. Critic 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 A between input layer and hidden layer. The number of neurons on hidden later is determined as 5.

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

$$e_t = 1 + \gamma\lambda e_{t-1} \qquad if \quad x = x_t \quad (10)$$

$$e_t = \gamma\lambda e_{t-1} \qquad otherwise \quad (11)$$

$$\hat{R}_{t+1} = \delta_t = (R_t + \gamma v_{t+1} - v_t)e_t \quad (12)$$

where γ is a discount coefficient between 0 and 1 (in this case γ is set to 0.9). $\lambda, 0 \leq \lambda \leq 1$ is a new parameter.

The learning process for value function is accomplished by step changes calculated by products of internal reinforcement, learning constant and appropriate input values from previous layers.

3.3 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). 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 (Zhou and Meng [2000]). It

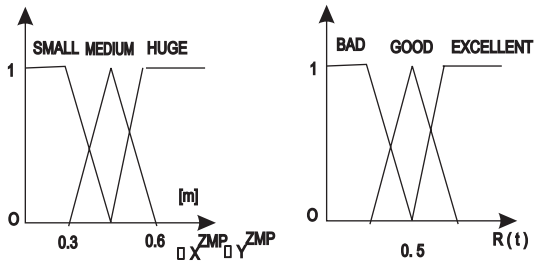

Fig. 3. The Membership functions for ZMP deviations and external reinforcement

is possible to use scalar critic signal, 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.

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

Fuzzy rules for external reinforcement

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.3.

3.4 Impact-force controller

The role of impact-force controller in the proposed control structure is to counteract the ground reaction force that appears when the locomotion mechanism foot strikes the ground (heel strike). The control law can be defined in the form of a PI-regulator as:

$$F_c = F_0 - \int_0^t Q(\Delta F)\cdot dt, \quad t \in (0,T], \quad \Delta F = F - F_0 \quad (13)$$

$$F_0 = \left[F_{0x}\ F_{0y}\ F_{0z}\ M_{0x}\ M_{0y}\ M_{0z}\right]^T,$$
$$F = \left[F_x\ F_y\ F_z\ M_x\ M_y\ M_z\right]^T, \quad (14)$$

$$Q(\Delta F) = K_{PF}\Delta F + K_{IF}\int_0^t \Delta F\cdot dt, \quad t \in (0,T] \quad (15)$$

where $F_c \in R^{6\times 1}$ is a vector of the so called generalized forces. $F_0 \in R^{6\times 1}$ and $F \in R^{6\times 1}$ represent the respective vectors of nominal and measured values of forces and moments of the ground reaction along, i.e. about three coordinate axes (x, y and z); $\Delta F \in R^{6\times 1}$ is the vector of deviation of forces (moments) F of the ground reactions from their nominal values F_0 calculated for the nominal robot's motionq_0 ; $K_{PF} \in R^{6\times 6}$ and $K_{IF} \in R^{6\times 6}$ are the

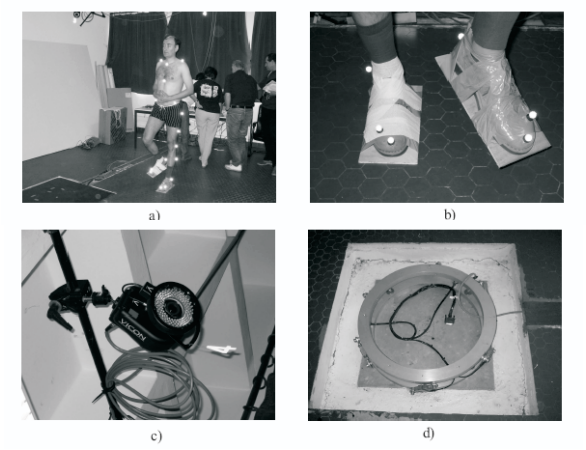

Fig. 4. Experimental capture motion studio

square matrices of proportional and integral gains of the designed PI-regulator. Finally, since the generalized forces F_c cannot be realized in a direct way, the control torques P are determined from the relation:

$$P_3 = -\hat{J}^T(q, x_{cf})F_c \tag{16}$$

where $\hat{J}(q, x_{cf})$ is the Jacobian matrix; x_{cf} is position vector of constrained foot

4. EXPERIMENTAL AND SIMULATION STUDIES

The corresponding experiments were carried out in a caption motion studio (Rodic et al. [2007]). 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 photometricaly. The other kinematic parameters, as well as dynamic ones, were identified on the basis of the biometric tables, recommendations and empirical relations. 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.4. To detect current positions of body links use was made of the special markers placed at the characteristic points of the body/limbs. 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]. To mimic a rigid foot-ground contact, a 5 [mm] thick wooden plate was added to each foot.

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. We assumed that it is possible to design a bipedal locomotion mechanism (humanoid robot) with defined parameters (same as in Fig. 1) . On the basis of the measured values of positions (coordinates) of special markers in the course of motion it was possible to identify angular trajectories of the particular joints of the bipedal locomotion system. Animation

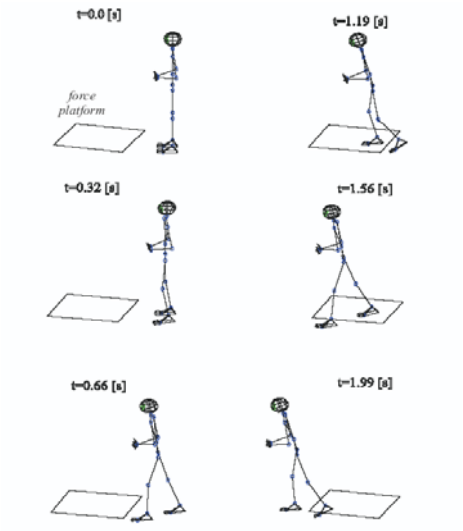

Fig. 5. Animation of biped locomotion

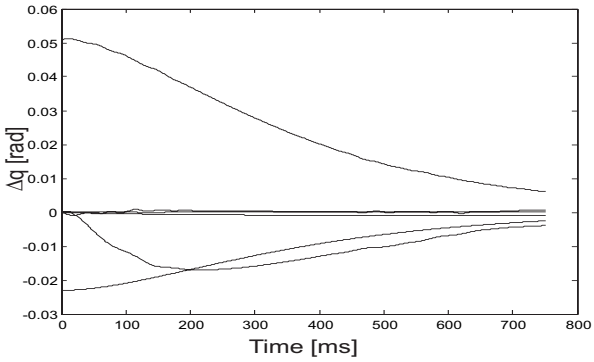

Fig. 6. Position errors

of the biped gait of the considered locomotion system, for the given joint trajectories, is presented in Fig.5 through several characteristic positions.

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. The results obtained by applying the controllers (position and velocity errors) are shown on Figs. 6 and 7. 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 the simulation example, it was shown how the basic dynamic controller together with reinforcement learning control structure is able to compensate the deviations of dynamic reactions even in the presence of uncertainty of the ground surface inclination. The simulation experiment considered in this example deals with a real case of robot walking on a quasi-horizontal ground surface. In that sense, small inclinations of the supporting surface in the longitudinal $\gamma_1 = 3^\circ$ and lateral direction $\gamma_2 = 2^\circ$ were introduced in the scope of the considered simulation experiment. 100 basis functions $\phi(x)$ are allocated to represent mean of the policy $\phi(x)$. In Figs.8 and 9 the comparison of the simulation results for

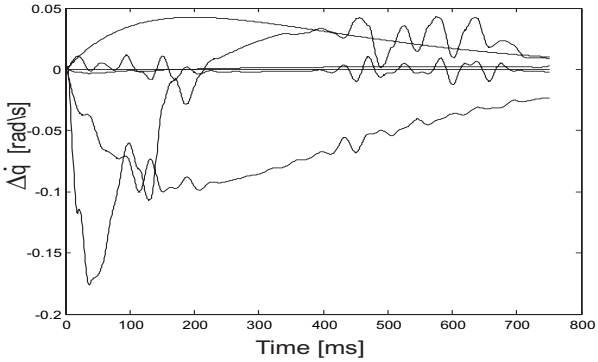

Fig. 7. Velocity errors

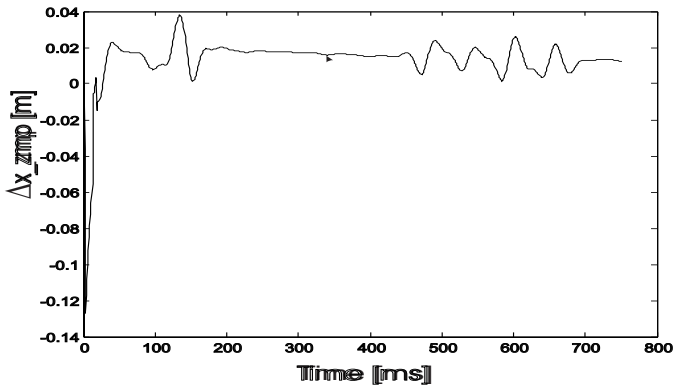

Fig. 8. ZMP error in x-direction

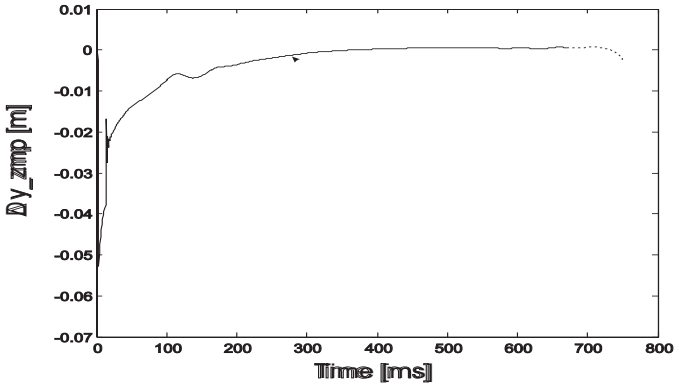

Fig. 9. ZMP error in y-direction

ZMP errors in coordinate directions are shown. In Fig. 10 value of reward or 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.

5. CONCLUSION

This paper presented an approach for acquiring biped motion focussed on learning a control policy. As a result, we demonstrated that it is possible to acquire dynamic motions though reinforcement learning using policy gradient method. 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

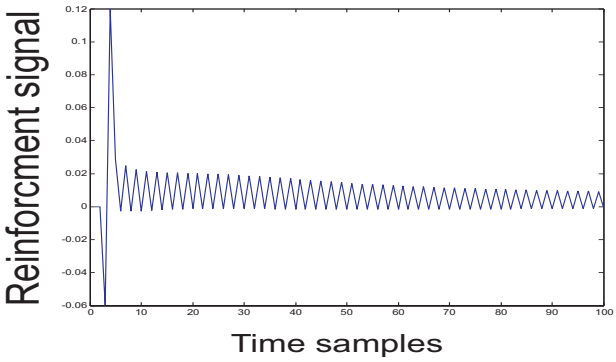

Fig. 10. The acquired reward

proposed 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 intelligent dynamic controller possesses both the conventional position-velocity feedback and dynamic reaction feedback.

REFERENCES

H.Benbrahim,H. and J.A Franklin. Biped Dynamic Walking using Reinforcement Learning. *Robotics and Autonomous Systems*, volume 22, pages.283–302, 1997.

T.Mori, Y. Nakamura, M. Sato, and S. Ishii. Reinforcement learning for a cpg-driven biped robot. in *Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI)*,pages.623-630, 2004.

Y. Nakamura, M.Sato, and S. Ishii. Reinforcement learning for biped robot. in *Proceedings of the 2nd International Symposium on Adaptive Motion of Animals and Machines*, 2003.

J.Peters, S.Vijayakumar, and S.Schaal. Reinforcement learning for humanoid robotics. in *Proceedings of the 3rd IEEE-RAS International Conference on Humanoid Robots Humanoids 2003*, Karlsruhe, Germany, September 2003.

A.Rodic, M.Vukobratovic, k.Addi, and G.Dalleau. Contribution to the Modelling of Non-Smooth, Multi-Point Contact Dznamcs of Biped Locomotion - Theory and Experiments. accepted for publication in journal *Robotica.,2007*

T. Shibata, R. Hitomoi, Y. Nakamura, and S.Ishii. Reinforcement Learning of Stable Trajectory for Quasi-Passive Dynamic Walking of an Unstable Biped Robot. In M.Hackel, editor, *Humanoid Robots:Human-like Machines*, Itech, Vienna, 2007.

R.Tedrake, T,W Zhang, H.S. Seung. Stochastic policy gradient reinforcement learning on a simple 3d biped. in *Proceedings of the 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2004.

M. Vukobratović, B.Borovac, D. Surla, and D. Stokić. *Biped Locomotion - Dynamics, Stability, Control and Application*. Springer Verlag, Berlin, Germany, 1990.

C. Zhou and Q.Meng. 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*, pages. 3829–3834, 2000.