**SERBIAN JOURNAL OF ELECTRICAL ENGINEERING**
**Vol. 9, No. 2, June 2012, 231-245**

**UDK: 007.52** **DOI: 10.2298/SJEE1202231K**

# Episodic Reinforcement Learning Control Approach for Biped Walking

**Duško Katić**[1]

**Abstract:** This paper presents a hybrid dynamic control approach to the realisation of humanoid biped robotic walk, focusing on the policy gradient episodic reinforcement learning with fuzzy evaluative feedback. The proposed structure of controller involves two feedback loops: a conventional computed torque controller and an episodic reinforcement learning controller. The reinforcement learning part includes fuzzy information about Zero-Moment-Point errors. Simulation tests using a medium-size 36-DOF humanoid robot MEXONE were performed to demonstrate the effectiveness of our method.

**Keywords:** Humanoid robot, Motion Control, Reinforcement Learning, Fuzzy evaluative feedback.

## 1 Introduction

The contemporary humanoid robots are becoming more complex and more intelligent because of the need to be able to function autonomously in changing task environments. Therefore, there is an increasing need for robots to adapt their parameters and control performance autonomously. On the other hand, practical biped needs to be more like a human, i.e. capable of switching between different known gaits on familiar terrain and learning new gaits when presented with an unknown terrain. In this case, even if stable trajectories are used, the existence of impulse disturbances on foot's sole can make the robot unstable. In some particular case of applications, there is often some degree of uncertainty regarding the state of the system, hence it is difficult to derive an accurate model of the humanoid robot and its interaction with the environment, to enable its more efficient control.

Hence, the inherent walking patterns have to 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 it is critical to achieving autonomous behavior of the robot. Learned behavior should be acquired by the robots themselves in a human-like way, not programmed manually. Humans

[1] Robotics Laboratory, Mihajlo Pupin Institute, University of Belgrade, Volgina 15, 11000 Belgrade, Serbia, E-mail: dusko.katic@pupin.rs

learn actions by repetitive trial and error procedure or by emulating someone else's actions. Thus, because of its resemblance to this human's way of learning, reinforcement learning (RL) could be applied for the control of humanoid robots, based on the experience gained in their interactions with the environment. Another approach to learning control of biped gait involves an open/closed-loop learning control algorithm [1].

RL methods have been applied to a variety of robot learning problems, as well as in complex learning tasks involving many degrees of freedom (DOFs), such as learning of locomotion [2 – 10]. Traditional representations of motor behavior in robotics are mostly based on the synthesis of desired trajectories. The resulting control policies, generated from a tracking controller of desired trajectories, are not robust towards unforeseen disturbances, and they do not easily generalize to new behavioral situations without complete recomputing of desired trajectories. Hence, there are two different application of reinforcement learning: first for traditional representations, and the second for dynamic motor primitives (DMP). The second approach to learnable dynamical systems originated from the desire to model elementary motor behaviors, called motor primitives, in humanoid robots as attractor systems [11]. The recent algorithm related to the DMP framework, called Policy Improvement with Path Integrals (PI2) [10], appeared to work well in robotic applications. This method takes a simple form with no open tuning parameters besides the exploration noise, and performs numerically robustly in high-dimensional learning problems, as it case in humanoid robotics.

The objective of this paper is to present a new integrated hybrid control strategy (model-based control together with learning control) for bipedal walking, using traditional representations. The biped trajectory tracking problem is considered as a repetitive control task by using non-learning parametric rigid body model-based dynamic control along with non-parametric episodic reinforcement learning from long-term rewards. The basic non-learning part of the control algorithm represents computed torque control method. The second control part consists of the inclusion of reinforcement learning part, but only for the compensation joints. The hybrid control strategy is chosen because the available prior knowledge from robot modeling can entail valuable information for robot learning that may result in a faster learning speed, higher accuracy, and better generalization. On the other hand, it is known that semiparametric learning methods outperform with high accuracy and better generalization, parametric rigid body dynamics methods [12].

The proposed reinforcement learning part is based on the application of Episodic Natural Actor Critic Method [13] as a well-known policy gradient method. This approach is different from similar methods based on real-time learning from immediate rewards [9]. 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. In comparison with most standard value function-based reinforcement learning methods, this type of method has particular features suited to robotic applications. The use of gradient-policy enables smooth changes of parameters, stability of algorithm, incorporation of prior and incomplete information in the control process. The autonomous tuning of control parameters was modeled as an episodic reinforcement learning task, with parameter tuning after each walking epoch. The reinforcement signal was simply defined as a fuzzy measure of Zero-Moment-Point (ZMP) error.

## 2 Dynamic Model of the System and Control Requirements

### 2.1 Model of the robot's mechanism

The studies were performed on a typical biped locomotion mechanism, MEXONE Humanoid Robot [14, 15] , whose spatial model is shown in Fig. 1.

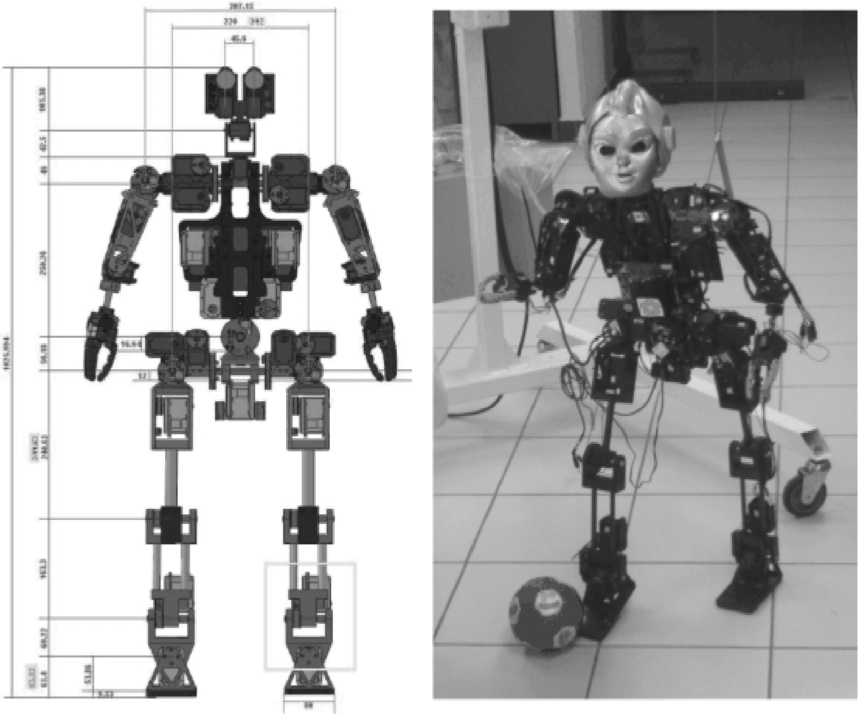


**Fig. 1** – MEXONE *humanoid robot.*

The mechanism possesses 36 DOFs. The overall dynamic model of the locomotion mechanism can be 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; $F \in R^{6X1}$ is the vector of external forces and moments acting at the particular points of the mechanism; $H \in R^{n\times n}$ is the square matrix that describes 'full' inertia matrix of the mechanism: $h \in R^{n\times 1}$ is the vector of gravitational, centrifugal and Coriolis moments acting at *n* mechanism joints; $J \in R^{6\times n}$ is the corresponding Jacobian matrix of the system; $q \in R^{n\times 1}$ 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.

## 2.2 Gait phases and indicator of dynamic balance

The robot's bipedal gait consists of several phases (see Fig. 2) that are periodically repeated [16]. 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). The DSP has two micro-phases: (i) weight acceptance phase (WAP) or heel strike, and (ii) weight support phase (WSP).

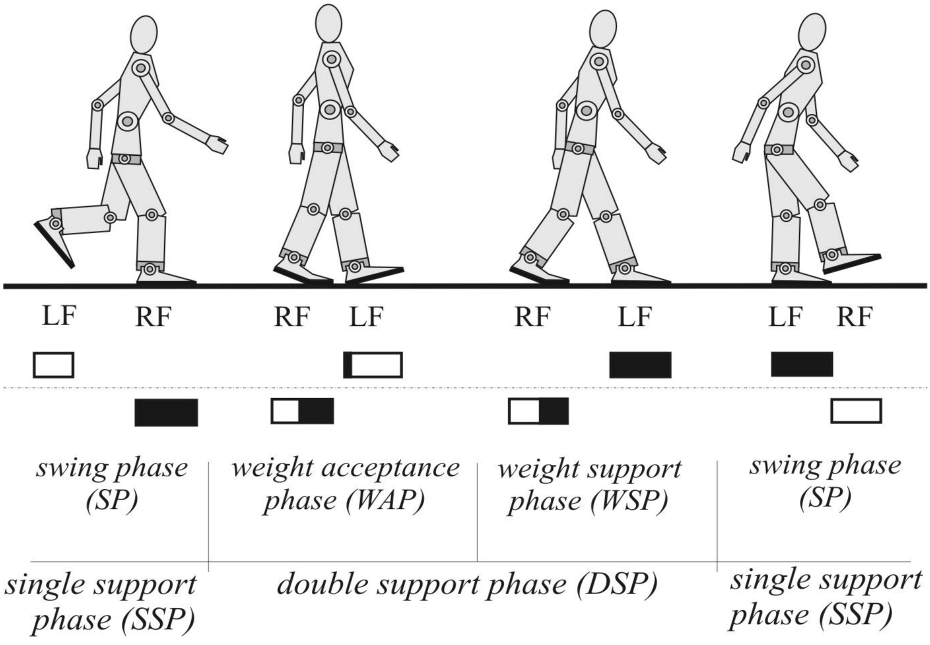
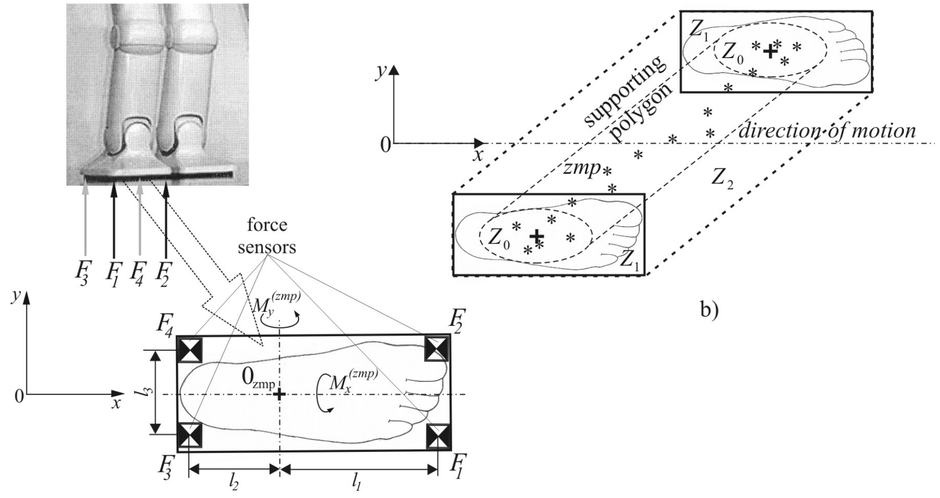

LF
RF
RF
LF
RF
LF
LF
RF
swing phase (SP)
weight acceptance phase (WAP)
weight support phase (WSP)
swing phase (SP)
single support phase (SSP)
double support phase (DSP)
single support phase (SSP)

**Fig. 2** – *Phases of biped gait.*



a)

b)

**Fig. 3** – *Zero-Moment Point.*

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 [16] 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. 3a and 3b 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 $F_i$, $i=1,2,3,4$, under the robot foot. Force sensors are usually placed on the foot sole in the polygonal framework. Sensors' positions are defined by the geometric quantities $l_1$, $l_2$ and $l_3$. If the point $0_{zmp}$ at the center of the foot is assumed to be the nominal ZMP position (Fig. 3a), then the following equations determine the relative ZMP position with respect to its nominal:

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

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

$$F_r^{(z)} = \sum_{i=1}^{4} F_i, \quad \Delta x^{(zmp)} = \frac{-\Delta M_y^{(zmp)}}{F_r^{(z)}}, \qquad \Delta y^{(zmp)} = \frac{\Delta M_x^{(zmp)}}{F_r^{(z)}}, \tag{2}$$

where $F_i$ and $F_i^0$, $i=1,2,3,4$, are the measured and nominal values of the ground reaction force; $\Delta M_x^{(zmp)}$ and $\Delta M_y^{(zmp)}$ are the deviations of the moments of ground reaction forces around the axes passing through the $0_{zmp}$; $F_r^{(z)}$ is the resultant force of ground reaction in the vertical $z$-direction, while $\Delta x^{(zmp)}$ and $\Delta y^{(zmp)}$ are the displacements of the 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 the $x$ and $y$ directions are calculated from the previous relation. The instantaneous position of ZMP is the best indicator of dynamic balance of the robot mechanism. The quality of the robot's 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 the balancing controller we proposed the application of the so-called integrated hybrid dynamic control. The control algorithm involves the summation of two parts: (i) basic dynamic controller $P_1$ for trajectory tracking that acts on all joints, and (ii) dynamic controller $P_2$ tuned by episodic reinforcement learning structure, but acting only on the chosen compensation joints.

### 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. There are various control techniques as in paper [17], while in our approach, the controller for robotic trajectory tracking was adopted using the well-known 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:

$$P_1 = \hat{H}(q)[\ddot{q}_0 + K_v(\dot{q} - \dot{q}_0) + K_p(q - q_0)] + \hat{h}(q,\dot{q}) - \hat{J}^T(q)F, \qquad (3)$$

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 $K_p = diag\{k_p^i\}$, $K_v = diag\{k_v^i\}$, $i = 1,2,\ldots,n$, 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 reinforcement learning structure

The main idea pursued in this paper is to include the reinforcement learning control component based on a constant qualitative evaluation of the biped walking performance. The evaluation of the control action based on the ZMP error, rather than on the numerical error of state variables, can be very convenient for searching of optimal and balanced biped walking. This reinforcement control part $P_2$ is realized only for the six special compensation joints. The quantity $P_2$ is the vector of compensation control torques at the selected compensation joints (ankle, knee and hip joints).

The proposed reinforcement learning structure is based on policy gradient method called Episodic Natural Actor Critic (ENAC) algorithm [13]. It is a stochastic gradient-descent method, in which the, parameters of control policy are improved upon each episode $e$ (exactly, an episode corresponds to every two-phase humanoid step). The control signal in every time instant $k - P_2 = u_k \in U = R^M$ is defined by the parameterized stochastic control policy $u_k : \pi_\theta(u_k \mid x_k)$ with the parameters $\theta \in R^K$ , while the input of the control policy is the state variable $x \in X = R^N$ with the transition probability distribution $x_{k+1} : p(x_{k+1} \mid x_k, u_k)$. The practical implementation of parameterized control policy is realized through an ACTOR fuzzy-neural network, with the aim to select/tune the best network parameters. It searches the action space using a Stochastic Real Valued unit at the output. The unit's action uses a Gaussian random number generator. There are five layers: input layer. antecedent part with fuzzification, rule layer, consequent layer, output layer with defuzzification. This system is based on the fuzzy rule base generated by the expert knowledge with 25 rules. The partition of input variables is defined by five linguistic variables: *NEGATIVE BIG, NEGATIVE SMALL, ZERO, POSITIVE SMALL* and *POSITIVE BIG*. The member functions are chosen in triangular form.

SAM (**S**tochastic **A**ction **M**odifier) uses the recommended control torque from ACTOR and reinforcement signal $r$ to produce the final commanded control torque $P_2$. It is defined by a Gaussian random function in which the recommended control torque is the mean, while the standard deviation is defined by the following equation:

$$\sigma(\hat{r}(t+1)) = 1 - \exp(-|\hat{r}(t+1)|) . \tag{4}$$

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 (tuning of the antecedent and consequent layers of the ACTOR) is accomplished by natural gradient changes (back propation defined by the reinforcement signal, learning constants and current recommended control torques). The total number of tuned parameters of the ACTOR is 62. The general aim of policy optimization in reinforcement learning is to optimize the control parameters policy in such a way that the expected return

$$J(\theta) = E\{\sum_{k=0}^{H} \gamma^i r_k\} \tag{5}$$

is optimized, where $\gamma^i \in [0,1]$ is a discount factor; $r_i$ is the reward or reinforcement signal; $H$ is the number of time instants during one episode. It is important to notice that for biped motion, abrupt changes of the control parameter are not acceptable, but smooth parameter changes are required. Hence, the policy gradient method based on episodic natural actor critic gradient [13] is chosen.

In order to estimate natural gradient, the functional approximator is defined by:

$$f_w^{\pi}(x,u) = \nabla_\theta \log \pi(u \mid x)^T w. \tag{6}$$

The practical realization of the ENAC algorithm includes the calculation of the values $P = \psi$ and $R$ according to the following equations:

$$\psi = \sum_{k=0}^{H} \gamma^k \nabla_\theta \log \pi(u_k \mid x_k), \tag{7}$$

$$R = \sum_{k=0}^{H} \gamma^k r(x_k, u_k), \tag{8}$$

$$w = P \cdot R. \tag{9}$$

$$J(\theta) = R. \tag{10}$$

**Algorithm ENAC**

```
Input: Parametrized policy $\pi(u \mid x) = p(u \mid x, \theta)$ with the initial parameters
    $\theta = \theta_0$, policy derivatives $\nabla_\theta \log \pi(u \mid x)$, and the function
    approximators $f_w^{\pi}(x_k, u_k)$.
for $e = 1,2,3,...$
        Execute Rollout:
        Draw initial state $x_0 : p(x_0)$.
        for $k = 0,1,2,...,H$
            Draw action $u_k : \pi(u_k \mid x_k)$, observe next state
              $x_{k+1} : p(x_{k+1} \mid x_k, u_k)$, and reward $r_t = r(x_k, u_k)$.
        end
        **Critic evaluation:** Determine the criterion function $J(\theta)$ and
                               gradient estimate $w$.
        Determine $\psi = \sum_{k=0}^{H} \gamma^k \nabla_\theta \log \pi(u_k \mid x_k)$ and reward statistics
        $R = \sum_{k=0}^{H} \gamma^k r(x_k, u_k)$.
        Form the matrix $P$ and vector $R$.
        Calculate the gradient estimate $w$ and performance $J(\theta)$.
        **Actor-Update:**
            When the gradient is converged, $(w_{e+1}, w_{e-\tau}) \leq \varepsilon$,
            update the policy parameters: $\theta_{e+1} = \theta_e + \alpha_e w_{e+1}$.
**end**

**Output:** Trained policy parameters $\theta$.
```

### 3.3 Fuzzy reinforcement signal

The detailed and precise training data for learning are often hard to obtain or may not be available at all 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 signals, such as “near fall down”, “almost success”, “slower”, “faster” and etc., to evaluate the walking gait. In this case, the use of fuzzy evaluation feedback is much closer to the learning environment in the real world [18]. It is possible to use scalar critic signal, but as one of solutions, the reinforcement signal was considered as a fuzzy number $R(t)$. We also assume that $R(t)$ is the fuzzy signal available at the time step $t$ and caused by the input and action chosen at the time step $t-1$, or even affected by earlier inputs and actions. For more effective learning, an error signal that gives more detailed balancing information should be given instead of a simple “go/no-go” scalar feedback signal. As an example we give the following fuzzy rules that can be used to evaluate the biped balancing according to the **Table 1**.

**Table 1**

*Fuzzy rules for reinforcement.*

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

The linguistic variables for the ZMP deviations $\Delta x^{(zmp)}$ and $\Delta y^{(zmp)}$ and for the reinforcement $R$ are defined using the membership functions that are defined in Fig. 4.

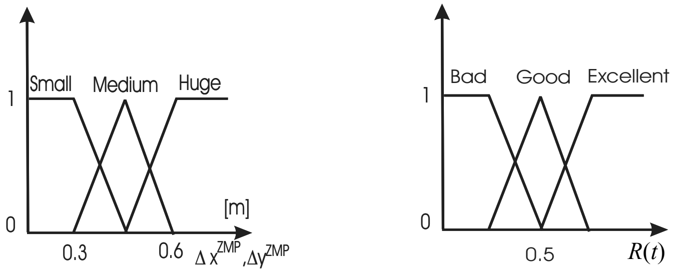


**Fig. 4** – *The membership functions for* ZMP *deviations and reinforcement.*

## 4 Simulation Results

The proposed control algorithm for biped walking was verified by the corresponding simulation experiments. For this purpose, the parameters of the 36-DOF (4 at the binocular head, 4 for each hand, 5 for each arm, 2 at pelvis, and 6 at each leg) MEXONE biped robot [14, 15] of 1.026 m height and 8.36 kg weight were assumed. The corresponding Matlab/Simulink HRSP software toolbox [19] was applied to simulate robot's kinematics and dynamics. In the simulation experiments we assumed the robot's planar motion in the sagital direction with a forward speed of 0.60 m/s, 0.40 m step size, and 0.075 m height of the swing leg.

Some special simulation experiments were performed in order to validate the proposed hybrid dynamic learning control approach. The simulation results were analyzed corresponding to the duration of one two-phase step of the locomotion mechanism in the swing phase, including the free (landing) foot strike instant. Initial conditions of the simulation examples (initial deviations of joints' angles and joints' angular velocities) were chosen to be the same in all simulation experiments. The process of learning was realized in more than 100 learning epochs (episodes).

Fig. 5 shows the values of return or reinforcement through the episodic process of the walk. It is clear that the task of walking within desired ZMP tracking error limits is achieved in a good fashion.

A comparison of the simulation results for ZMP errors in the coordinate directions during the learning episodes is given in Figs. 6 and 7.

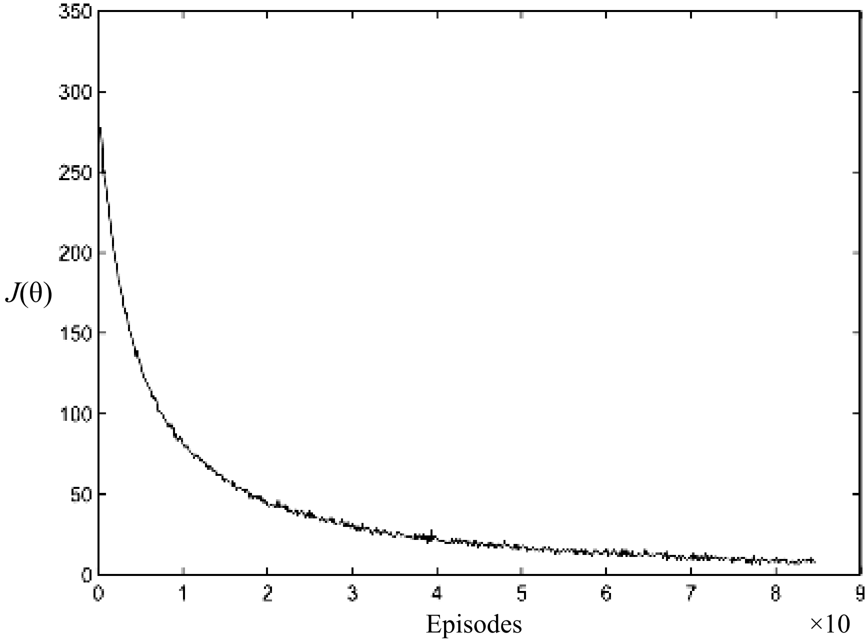


**Fig. 5** – *The acquired return during learning episodes.*

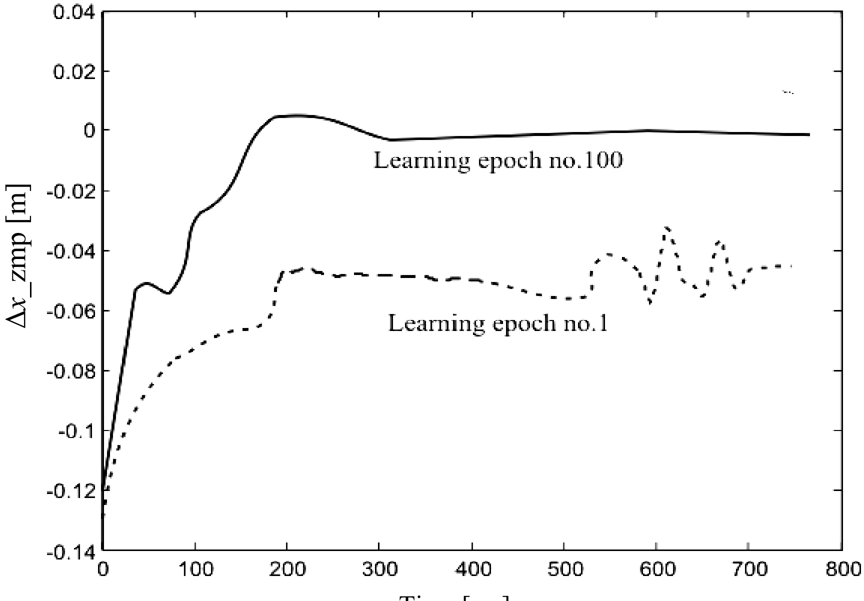


**Fig. 6** – ZMP *error in the x-direction during the learning episodes.*

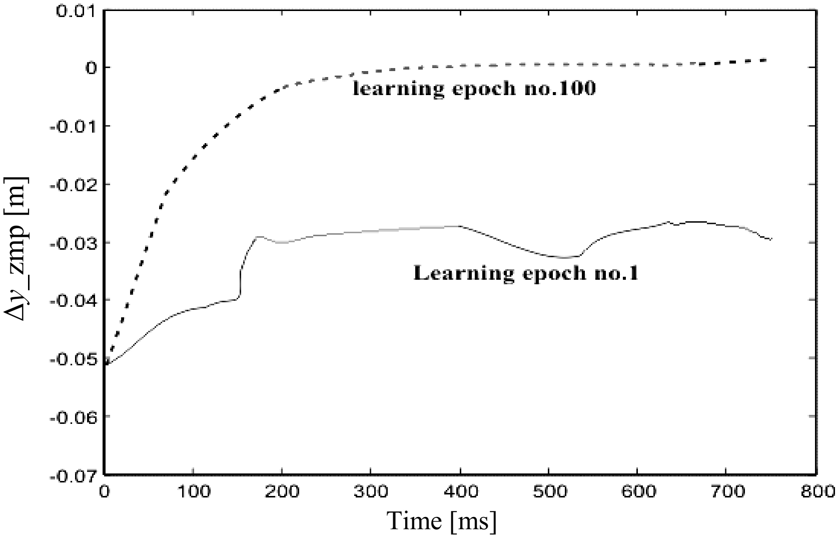


**Fig. 7** – ZMP *error in the y-direction during the learning episodes.*

These figures show how the basic dynamic controller, together with the reinforcement learning control structure, is able to compensate for the deviations of dynamic reactions in the presence of the system uncertainties. Finally, the joint and velocity tracking errors converge to zero values in the given time interval (Figs. 8 and 9). This means that the controller ensures good tracking of the desired trajectory after a sufficient number of learning episodes.

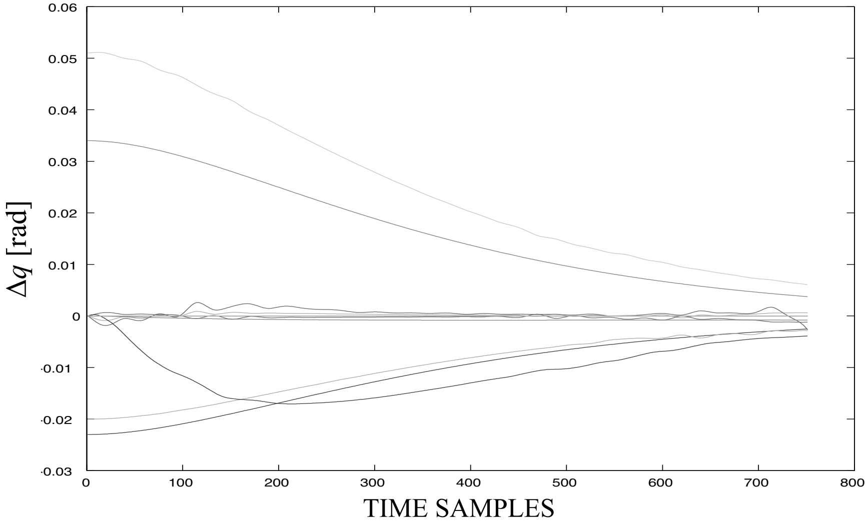


**Fig. 8** – *Convergence of the joint angles errors for position joints.*

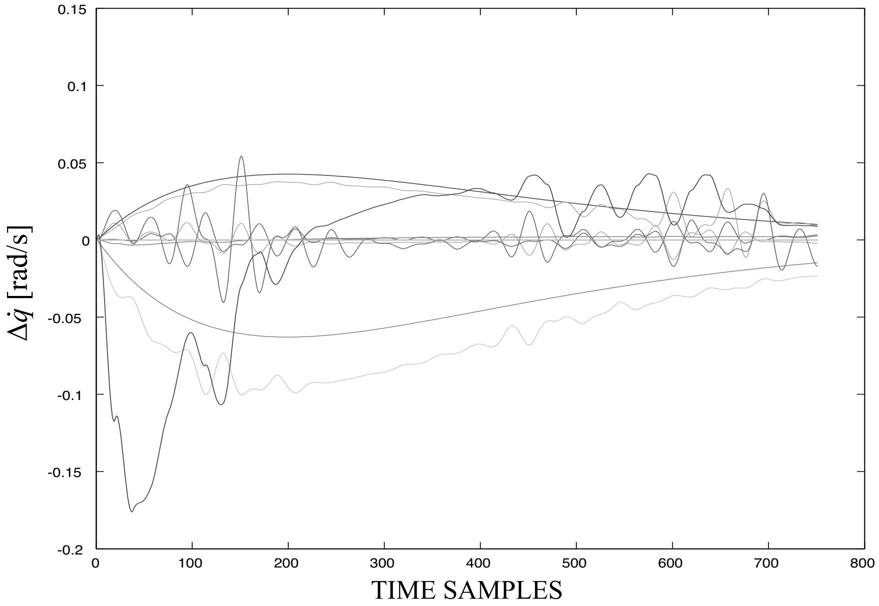


**Fig. 9** – *Convergence of the joint velocities errors for the compensation joints.*

## 5 Conclusion

The paper presented a hybrid approach for acquiring biped motion focussed on conventional control algorithm together with learning a control policy. It demonstrated the possibility of acquiring dynamic motions through reinforcement learning using ENAC policy gradient method. The algorithm is based on the fuzzy evaluative feedback that is obtained from human intuitive balancing knowledge. The reinforcement learning with fuzzy evaluation feedback is much closer to the human biped walking evaluation than the original one with scalar feedback. The proposed intelligent control scheme fulfills the preset control criteria. Its application ensures the desired precision of robot's motion and maintaining of dynamic balance of the locomotion mechanism during the motion. The developed intelligent dynamic controller can be potentially applied in combination with robotic vision, to control biped locomotion mechanisms in the course of fast walking, running, and even in the phases of jumping,

The proposed hybrid control algorithm is planned to be experimentally implemented using the MEXONE biped robot. In the future research, the alternative ideas will be the representation for control policy based on Gaussian Regression Kernel, Pi2 algorithm, and Covariance matrix adaptation evolution strategies.

## 6 Acknowledgments

This research was supported by the Republic of Serbia Ministry of Science, Grant TR-35003 "Ambient Intelligent Service Robots of Anthropomorphic Characteristics". The author would like to thank colleague Milan Vukov for his assistance in the software development.

## 7 References

[1] B. Wang, H. Xie, D. Cong, X. Xu: Biped Robot Control Strategy and Open-closed-loop Iterative Learning Control, Frontiers of Electrical and Electronic Engineering in China, Vol. 2, No. 1, March 2007, pp.104 – 107.

[2] H. Benbrahim, J.A. Franklin: Biped Dynamic Walking using Reinforcement Learning, Robotics and Autonomous Systems, Vol. 22, No. 3-4, Dec. 1997, pp. 283 – 302.

[3] Y. Nakamura, M. Sato, S. Ishii: Reinforcement Learning for Biped Robot, 2nd International Symposium on Adaptive Motion of Animals and Machines, Kyoto, Japan, 4 – 8 March 2003, p. ThP-II-5.

[4] J. Peters, S. Vijayakumar, S. Schaal: Reinforcement Learning for Humanoid Robotics, 3rd IEEE-RAS International Conference on Humanoid Robots Humanoids, Karlsruhe, Germany, 29 – 30 Sept. 2003.

[5] T. Mori, Y. Nakamura, M. Sato, S. Ishii: Reinforcement Learning for a CPG-driven Biped Robot, 19th National Conference on Artificial Intelligence, San Jose, CA, USA, 25 – 29 July 2004, pp. 623-630.

[6] R. Tedrake, T.W. Zhang, H.S. Seung: Stochastic Policy Gradient Reinforcement Learning on a Simple 3D Biped, IEEE/RSJ International Conference on Intelligent Robots and Systems, Sendai, Japan, 28 Sept. – 2 Oct. 2004, Vol. 3, pp. 2849 – 2854.

[7] M. Hackel: Humanoid Robots: Human-like Machines, Itech Education and Publishing, Vienna, Austria, 2007.

[8] G. Endo, J. Morimoto, T. Matsubara, J. Nakanishi, G. Cheng: Learning CPG-based Biped Locomotion with a Policy Gradient Method: Application to a Humanoid Robot, International Journal of Robotics Research, Vol. 27, No. 2, Feb. 2008, pp. 213 – 228.

[9] D. Katić, A. Rodić, M.Vukobratović: Hybrid Dynamic Control Algorithm for Humanoid Robots based on Reinforcement Learning, Journal of Intelligent and Robotic Systems, Vol. 51, No.1, Jan. 2008, pp. 3 – 30.

[10] F. Stulp, J. Buchli, E. Theodorou, S. Schaal: Reinforcement Learning of Full-body Humanoid Motor Skills, 10th IEEE-RAS International Conference on Humanoid Robots (Humanoids), Nashville, TN, USA, 6 – 8 Dec. 2010, pp. 405 – 410.

[11] J. Nakanishi, J. Morimoto, G. Endo, G. Cheng, S. Schaal, M. Kawato: A Framework for Learning Biped Locomotion with Dynamic Movement Primitives, IEEE-RAS/RSJ International Conference on Humanoid Robots (Humanoids 2004), Los Angeles, CA, USA, 10 – 12 Nov. 2004.

[12] D. Nguyen-Tuong, J. Peters: Using Model Knowledge for Learning Inverse Dynamics, IEEE International Conference Roboticss and Automation, Anchorage, AK, USA, 3 – 7 May 2010, pp. 2677 – 2682.

[13] J. Peters: Machine Learning for Robotics: Learning Methods for Robot Motor Skills, VDM Verlag Dr. Muller, Saarbrucken, Germany, 2008.

[14] http://www.gdl. cinvestav.mx/edb

[15] http://www.robai.com

[16] M. Vukobratović, D. Juričić: Contribution to the Synthesis of Biped Gait, IEEE Transaction on Biomedical Engineering, Vol. 16, No. 1, Jan. 1969, pp. 1 – 6.

[17] B. Svetozarević, K. Jovanović: Control of Compliant Anthropomimetic Robot Joint, Serbian Journal of Electrical Engineering, Vol. 8, No.1, Feb. 2011, 85 – 95.

[18] C. Zhou, Q. Meng: Reinforcement Learning with Fuzzy Evaluative Feedback for a Biped Robot, IEEE International Conference on Robotics and Automation, San Francisco, CA, USA, 24 – 28 April 2000, Vol. 4, pp. 3829 – 3834.

[19] http://www.pupin.rs/RnDProfile/robotics/hrsp.html