J Intell Robot Syst (2008) 51:3–30
DOI 10.1007/s10846-007-9174-5

Hybrid Dynamic Control Algorithm for Humanoid Robots Based on Reinforcement Learning

Duśko M. Katić · Aleksandar D. Rodić · Miomir K. Vukobratović

Received: 1 September 2006 / Accepted: 31 July 2007 /
Published online: 25 September 2007

Abstract In this paper, hybrid integrated dynamic control algorithm for humanoid locomotion mechanism is presented. The proposed structure of controller involves two feedback loops: model-based dynamic controller including impart-force controller and reinforcement learning feedback controller around zero-moment point. The proposed new reinforcement learning algorithm is based on modified version of actor-critic architecture for dynamic reactive compensation. Simulation experiments were carried out in order to validate the proposed control approach.The obtained simulation results served as the basis for a critical evaluation of the controller performance.

Keywords Humanoid robots · Biped locomotion · Integrated dynamic control · Reinforcement learning · Actor–critic method

1 Introduction

The contemporary humanoid robots are expected to be servants and maintenance machines with the main task to assist human activities in our daily life and to replace humans in hazardous operations. It is as obvious that anthropomorphic biped robots are potentially capable to effectively move in all unstructured environments where humans do. There are also strong anticipations that robots for the personal use will coexist with humans and provide supports such as the assistance for the housework,

D. M. Katić (✉) · A. D. Rodić · M. K. Vukobratović
Robotics Department, Mihajlo Pupin Institute,
Volgina 15, Belgrade 11060, Serbia
e-mail: dusko@robot.imp.bg.ac.yu

A. D. Rodić
e-mail: roda@robot.imp.bg.ac.yu

M. K. Vukobratović
e-mail: vuk@robot.imp.bg.ac.yu

care of the aged and the physically handicapped persons.It should be emphasized that many humanoid projects (http://www.androidworld.com) are currently being in progress worldwide, so that the number of representative humanoids will certainly grow in the near future.

Humanoid robot are autonomous systems capable of extracting information from their environments and using knowledge about the world and intelligence of their duties and proper governing capabilities. Intelligent humanoid robots should be autonomous to move safely in a meaningful and purposive manner, i.e. to accept high-level descriptions of tasks (specifying what the user wants to be done, rather than how to do it and would execute them without further human intervention). Naturally, the first approach to making humanoid robots more intelligent was the integration of sophisticated sensor systems as computer vision, tactile sensing, ultrasonic and sonar sensors, laser scanners and other smart sensors. However, todays sensor products are still very limited in interactivity and adaptability to changing environments. As the technology and algorithms for real-time 3D vision and tactile sensing improve, humanoid robots will be able to perform tasks that involve complex interaction with the environment (e.g. grasping and manipulating the objects). A major reason is that uncertainty and dynamic changes make the development of reliable artificial systems particularly challenging. On the other hand, to design robots and systems that best adapt to their environment, the necessary research includes investigations in the field of mechanical robot design (intelligent mechanics), environment perception systems and embedded intelligent control that ought to cope with the task complexity, multi-objective decision making, large volume of perception data and substantial amount of heuristic information. Also, in the case when the robot performs in an unknown environment, the current knowledge may not be sufficient. Hence, the robot has to adapt to the environment and to be capable of acquiring new knowledge through the process of learning. The robot learning is essentially concerned with equipping robots with the capacity of improving their behavior over time, based on their incoming experiences.

Although there has been a large number of the control methods used to solve the problem of humanoid robot walking, it is difficult to detect a specific trend. Classical robotics and also the more recent wave of humanoid and service robots still rely heavily on teleoperation or fixed behavior-based control with very little autonomous ability to react to the environment. Among the key missing elements is the ability to create control systems that can deal with a large movement repertoire, variable speeds, constraints and most importantly, uncertainty in the real-world environment in a fast, reactive manner. There are several intelligent paradigms that are capable of solving intelligent control problems in humanoid robotics [12–14].

One approach of departing from teleoperation and manual hard coding of behaviors is by learning from experience and creating appropriate adaptive control systems. A rather general approach to learning control is the framework of *reinforcement learning*. Reinforcement learning offers one of the most general framework to take traditional robotics towards true autonomy and versatility. Robotics is a very challenging domain for reinforcement learning, However, various pitfalls have been encountered when trying to scale up these methods to high dimension, continuous control problems, as typically faced in the domain of humanoid robotics.

The goal of this paper is to propose the usage of reinforcement learning for humanoid robotics. The new integrated hybrid dynamic control structure for the humanoid robots is proposed, based on 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 [3, 22] for dynamic compensation of ZMP (zero-moment-point) error. The GARIC reinforcement learning method proposed in this paper is based on the actor–critic architecture. Actor–critic methods are the natural extension of the idea of reinforcement comparison methods to temporal difference (TD) learning [4, 24]. The actor network represents the control agent, because it implements a control policy. The actor network is part of the dynamic system as it interacts directly with the system by providing control signals for the plant. The critic network implements the reinforcement learning part of the control system as it provides policy evaluation and can be used to perform policy improvement. This learning agent architecture has the advantage of implementing both a reinforcement learning mechanism as well as a control mechanism. For the actor, we selected the two-layer, feedforward neural network with sigmoid hidden units and linear output units. For the critic, the neuro-fuzzy network is proposed. The critic is trained to produce the expected sum of future reinforcement that will be observed given the current values of deviation of dynamic reactions and action. The actor network receives the position and velocity tracking error from the biped system. It is trained via *back propagation* (gradient descent) algorithm and training example provided by critic net. The external reinforcement signal was simply defined to be measure of ZMP error. Internal reinforcement signal is generated using external reinforcement signal and appropriate policy,

In Section 2, definition of biped control and advanced control methods for humanoid robots are introduced. After this, the basic system modelling, control criteria and gait analysis are described in Section 3. In Section 4, hybrid control algorithm based on dynamic controller including impact force controller and new reinforcement actor–critic control structure for biped walking is presented. Trough simulation studies (Section 5), proposed control methods that acquire a successful walking pattern are demonstrated.

2 Advanced Control Algorithms for Humanoid Robots

In spite of a significant progress and accomplishments achieved in the design of a hardware platform of humanoid robot and synthesis of advanced intelligent control of humanoid robots, a lot of work has still to be done in order to improve actuators, sensors, materials, energy accumulators, hardware, and control software that can be utilized to realize user-friendly humanoid robots.

There are various sources of control problems and various tasks and criteria that must be solved and fulfilled in order to create valid walking, balance and other functions of humanoid robots. Previous studies of biological nature, theoretical and computer simulation, have focussed on the structure and selection of control algorithms according to different criteria such as energy efficiency, energy distribution along the time cycle, stability, velocity, comfort, mobility, and environment impact.

The major problems associated with the analysis and control of bipedal systems are the high-order highly-coupled nonlinear dynamics and furthermore, the discrete changes in the dynamic phenomena due to the nature of the gait. Irrespective of the humanoid robot structure and complexity, the basic characteristic of all bipedal systems are: a) the DOF formed between the foot and the ground is unilateral and underactuated; b) the gait repeatability (symmetry) and regular interchangeability of the number of legs that are simultaneously in contact with the ground.

The stability issues of humanoid robot walking are the crucial point in the process of control synthesis. The control problems of dynamic walking are more complicated than in walking with static balance, but dynamic walking patterns provide higher walking speed and greater efficiency, along with more versatile walking structures. For static and dynamic walking robots, the issue of stable and reliable bipedal walk is the most fundamental and yet unsolved with a high degree of reliability.

The rotational equilibrium of the foot is the major factor of postural instability with legged robots. The question has motivated the definition of dynamic-based criterion for the evaluation and control of balance in biped locomotion. The most common criterion is the zero-moment point (ZMP) [26, 28, 29]. The ZMP concept has gained widest acceptance and played a crucial role in solving the biped robot stability and periodic walking pattern synthesis [26]. The ZMP is defined as the point on the ground about which the sum of all the moments of the active forces equals zero. If the ZMP is within the convex hull of all contact points between the foot and the ground, the biped robot can walk. Even if the stability based on ZMP only describes contact condition between foot and the ground, ZMP based controller is mainly used in humanoid robot communities because it is known to work well experimentally. In most of cases, the trajectories of humanoid robot are designed off-line, and then controller is designed to track these trajectories. A walking pattern is generated to ensure the robot's ZMP is all the time within the supporting foot projection. This is necessary for the biped robot to maintain dynamic balance during the walk [26]. In order to compensate for the ZMP error, it is necessary to implement the balance control using force/torque (FIT) sensor or inclination sensor. Several methods of on-line balance control have been proposed. The above research group has successfully realized dynamic walking experiment using their own balance control scheme [7, 10, 31]. Most research works dealing with balance control have focused on the compensation of ZMP error because the ZMP is the essential criterion of dynamic balance. In many cases, it has been assumed that the main source of errors represent the unexpected ground condition or model inaccuracies. As a matter of course, the balance controller should be robust against the inclination and model inaccuracy.

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 foot's sole can make robot to tumble. In this sense, it seems essential to combine force control techniques with more advanced algorithms such as adaptive and learning strategies. 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.

In summary, conventional control algorithms for humanoid robots can run into some problems related to mathematical tractability, optimization, limited extendability and limited biological plausibility. The presented intelligent control techniques have a potential to overcome the mentioned constraints.There are several intelligent paradigms that are capable of solving intelligent control problems in humanoid robotics [13]. Connectionist theory (NN – neural networks), fuzzy logic (FL), and theory of evolutionary computation (GA – genetic algorithms), are of great importance in the development of intelligent humanoid robot control algorithms.

Connectionist approach makes possible the learning of new gaits which are not weighted combinations of predefined biped gaits. Various types of neural networks are used for gait synthesis and control design of humanoid robots such as multilayer perceptrons, CMAC (cerebellar model arithmetic controller) networks, recurrent neural network, RBF (radial basis function) networks or Hopfield networks, which are trained by supervised or unsupervised learning methods. The majority of the proposed control algorithms have been verified by simulation, while there were few experimental verifications on real biped and humanoid robots. Neural networks have been used as efficient tools for the synthesis and off-line and on-line adaptation of biped gait. Another important role of connectionist systems in control of humanoid robots has been the solving of static and dynamic balance during the process of walking and running on terrain with different environment characteristics.

Some researchers used the fuzzy logic [13] as the methodology for biped gait synthesis and control of biped walking. Fuzzy logic was used mainly as part of control systems on the executive control level, for generating and tuning PID gains, fuzzy control supervising, direct fuzzy control by supervised and reinforcement error signals.

It is considered that GA can be efficiently applied for trajectory generation of the biped natural motion on the basis of energy optimization, as well as for walking control of biped robots and for generation of behavior-based control of these systems [13]. The main problem of GA application in humanoid robotics represents the coping with the reduction of GA optimization process in real time.

Because their complementary capabilities hybrid intelligent methods have also found their place in the research of gait synthesis and control of humanoid robots [13].

Recently, reinforcement learning (RL) has attracted attention as a learning method for studying movement planning and control [6, 8, 9]. Reinforcement learning is a kind of learning algorithm between supervised and unsupervised learning algorithms which is based on Markov decision process (MDP). Reinforcement learning concept that is based on trial and error methodology and constant evaluation of performance in constant interaction with environment.

There are two basic methods of RL that are very successful in solving this problem, TD learning [24] and Q learning [30]. Both methods build a state space value function that determines how close each state is to success or failure. Whenever the controller outputs an action, the system moves from one state to an other. The controller parameters are then updated in the direction that increases the state value function.

In area of humanoid robotics, there are several approaches of reinforcement learning [2, 5, 11, 15–18, 21, 22, 25, 32] 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 [2].

There are direct (model-free) and model-based reinforcement learning [6, 24]. The direct approach to RL is to apply policy search directly without learning a model of the system. In special tests [1], it were shown that model-based reinforcement learning is more data efficient than direct reinforcement learning. Also, it was concluded that model-based reinforcement learning finds better trajectories, plans and policies, and handles changing goals more efficiently. On the other hand, reported that a model-based approach to reinforcement learning is able to accomplish given tasks much faster than without using knowledge of the environment. In paper [17] a model-based reinforcement learning algorithm for biped walking in which the robot learns to appropriately place the swing leg was proposed. This decision is based on a learned model of the Poincare map of the periodic walking pattern. The model maps from a state at the middle of a step and foot placement to a state at next middle of a step. Actor–critic algorithms of reinforcement learning has a great potential in control of biped robots. For a example, a general algorithm for estimating the natural gradient, the Natural actor–critic algorithm, is introduced in paper [21]. This algorithm converges to the nearest local minimum of the cost function with respect to the Fisher information metric under suitable conditions. It offers a promising route for the development of reinforcement learning for truly high-dimensionally continuous state-action systems.

3 Dynamic Model of the System and Control Requirements

3.1 Model of the Robot's Mechanism

Biped locomotion mechanisms represent generally branched kinematic chains interconnected with spherical or cylindrical joints. Some kinematic chains in their interaction with the environment transform from open to closed type of chain 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. Thus, the mechanism considered has in total $n = 20$ DOFs of motion.

Dynamic model of the mechanism 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. For this purpose, the first link in the branched chain, representing the supporting foot, is adopted as the basic link of the mechanism. Taking into account dynamic coupling between particular parts (branches) of the mechanism chain, the overall dynamic model of the locomotion mechanism is represented in the following vector form [26]:

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

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

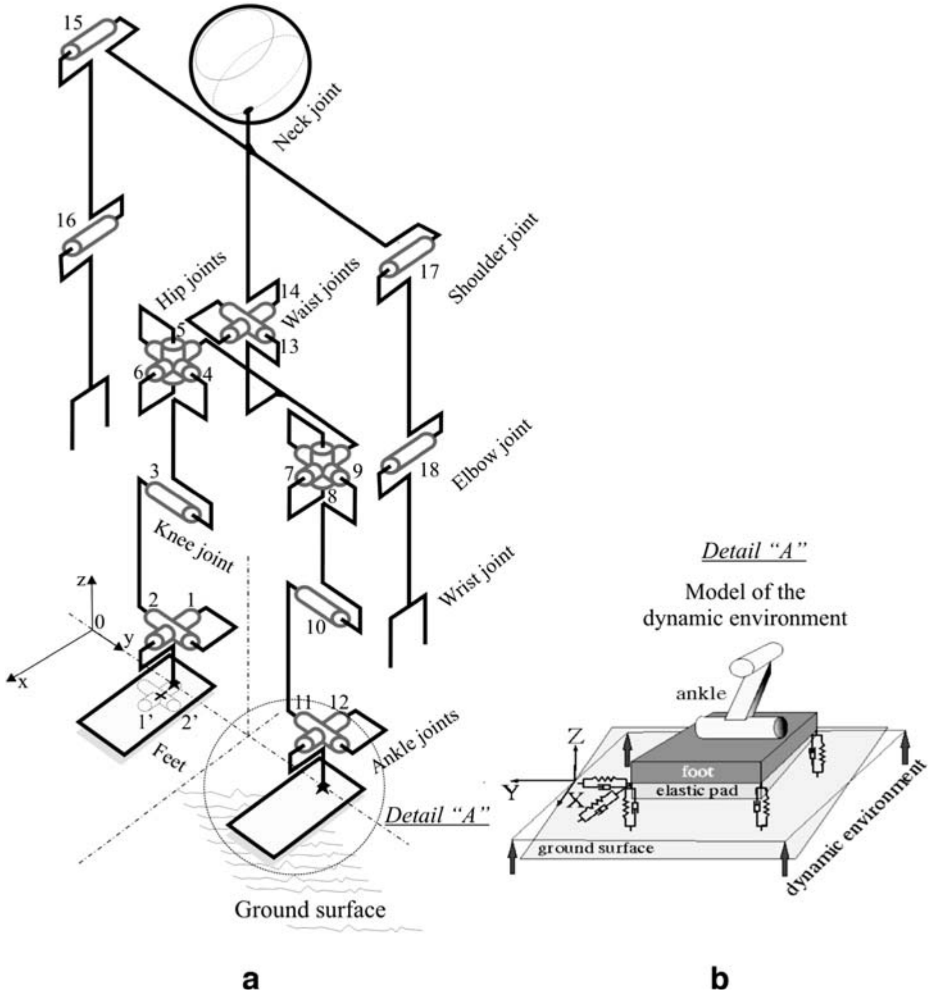


Fig. 1 Model of the biped locomotion mechanism with 18 active and 2 passive DOFs **a** kinematical scheme of the mechanism, **b** dynamic model of the environment

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. The vector F has a special importance in the calculation of the model (1). It represents the vector of forces and moments of ground reaction at the moment when the foot of the free (unconstrained) leg is forming contact with the ground, i.e. at the moment when the weight is transferred from one foot to the other. It is necessary to make difference between the so-called supporting (constrained) and free (unconstrained) foot, the latter becoming constrained at the moment of making contact with the ground.

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 elasto-dynamics. In the case considered, the micro-inductive transducers would be embedded into the elastic footpad in the way that would allow measurement of the corresponding micro-displacements in the directions defined by the orientation of the small springs in the environment model (Fig. 1b). 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, Eq. 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, Eq. 2 represents the model of a robot mechanism supported to the *dynamic environment*. Generally, under the notion *robotic dynamic environment* it can be assumed both the ground support on which the locomotion mechanism moves and elastic footpad of the supporting leg. During the walk, the interactive forces and moments of dynamic environment reaction are transferred onto the entire mechanism structure. It can be considered as a spring-mass-damper mechanic system, exactly as empirical model, consisting of four vertical units of a linear spring and a non-linear damper at the corners [20]. The two pairs of horizontal units consisting of a linear spring and a linear damper at the heel and toes, aligned along the walking and lateral directions, are also included in the model (Fig. 1b). The nonlinear vertical springs prevent the elastic pad from being squeezed more than is its thickness. The nonlinear damper model is used to describe accurately the impacts at the foot landing, as suggested in [20]. The inertia effects, taking into account the engaged mass of the footpad, are included in the model, too. Note also that this model of elastic pad also allows its rotational deformation due to the moment applied at the foot, as well as its asymmetric vertical deformation. As a result, it is possible for the foot in contact with the ground to move and rotate in any direction.

3.2 Definition of Control Criteria

The control synthesized for biped mechanism gait has to satisfy the following criteria: (1) accuracy of tracking the desired trajectories of the mechanism joints (2) maintenance of dynamic balance of the mechanism during the motion, (3) minimization of the impact arising at the moment of contact of the free foot and the ground during the gait, (4) minimization of dynamic loads at the robot joints, and (5) realization of anthropomorphic characteristics of the gait. Fulfillment of requirement (1) enables the realization of a desired mode of motion, walk repeatability and avoidance of potential obstacles. To satisfy requirement (2) it means to maintain balance of the dynamic reactions acting upon the robotic system during the motion. Fulfillment of requirement (3) decreases the impact effects on the overall system at the moment when the unconstrained leg foot strikes the ground. Fulfillment of requirement (4) 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. Requirement (5) is related to the quality of artificial (human-like) gait. Walk of a physically healthy human represents a balanced and harmonious sequence of movements, with minimal displacements of the position of the mass center about an imaginary central position corresponding to the human's posture at rest.

3.3 Gait Phases and Indicator of Dynamic Balance

The robot's bipedal gait consists of several phases that are periodically repeated [26]. Hence, depending on whether the system is supported on one or both legs, two macro-phases can be distinguished, viz.: (1) single-support phase (SSP) and (2) double-support phase (DSP). Double-support phase has two micro-phases: (1) weight acceptance phase (WAP) or heel strike, and (2) weight support phase (WSP). Figure 2 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. While walking, the biped is constantly in the state of dynamic balance. In the literature from this branch the terms "dynamically balanced walk" and "gait stability" are used very often in

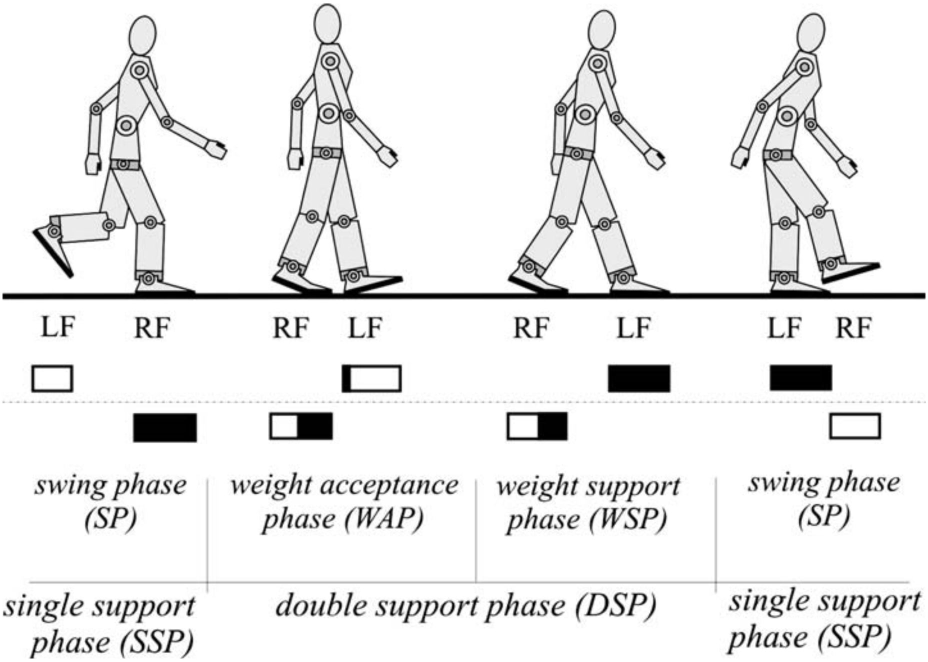

LF
RF
RF
LF
RF
LF
LF
RF
swing phase (SP)
weight acceptance phase (WAP)
weight support phase (WSP)
swing phase (SP)
single support phase (SSP)
double support phase (DSP)
single support phase (SSP)

Fig. 2 Phases of biped gait

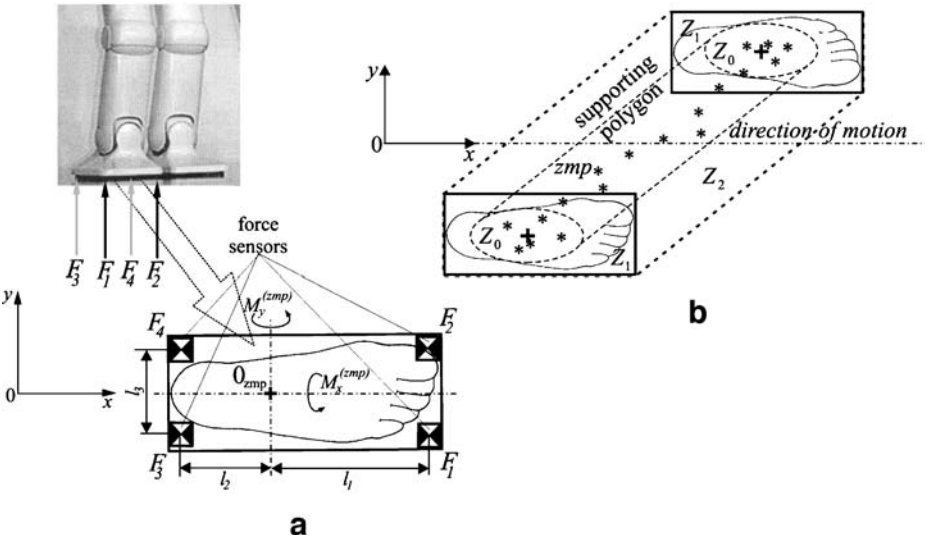


Fig. 3 Zero-moment point: **a** legs of "Toyota" humanoid robot (www.toyota.humanoids.co.jp); general arrangement of force sensors in determining the ZMP position; **b** zones of possible positions of ZMP when the robot is in the state of dynamic balance

the same context. In this paper, we have used the term "dynamic balance" of robot mechanism in the sense that does not mean the stable system. Dynamic balance of biped mechanism represents its dynamic state in which there is no overturning around the edge of the constrained foot during the walk. Because of that, the term "stable gait" is used unjustified very often for the gait in which the system is attempting to avoid overturning by preserving its 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 [28, 29] 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. Figure 3a and b 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 [26]. Force sensors are usually placed on the foot sole in the arrangement shown in Fig. 3a. 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. 3a), then the following equations to determine the relative ZMP position with respect to its nominal:

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

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

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

where F_i and F_i^0, $i = 1, \ldots, 4$, are the measured and nominal values of the ground reaction force; $\Delta M_x^{(\mathrm{zmp})}$ and $\Delta M_y^{(\mathrm{zmp})}$ are deviations of the moments of ground reaction forces around the axes passed through the 0_{zmp}; $F_r^{(z)}$ is the resultant force of ground reaction in the vertical z-direction, while $\Delta x^{(\mathrm{zmp})}$ and $\Delta y^{(\mathrm{zmp})}$ are the displacements of ZMP position from its nominal 0_{zmp}. The deviations $\Delta x^{(\mathrm{zmp})}$ and $\Delta y^{(\mathrm{zmp})}$ 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. 3b are illustrated certain areas (Z_0, Z_1 and Z_2), the so-called *safe zones of dynamic balance* of the locomotion mechanism. In the single-support phase the support area coincides with the area of the foot in contact with the ground, whereas in the double-support phase the support area is a convex surface determined by the areas of the feet and the ground and common tangents, so that the encompassed area is maximized (see Fig. 3b). In the case when an external disturbance force of smaller intensity acts upon the robotic system, the ideal tracking of joint trajectories is disturbed, but the ZMP position still remains within a safety zone (area in Fig. 3b). Appropriate control actions can return the system to the reference trajectory. The safety zone can be defined as a specific area within the support polygon, and it does not meet the margin of the support polygon. If the control system keeps the ZMP position inside the margins of the safety zone, then the biped mechanism dynamic balance will never be jeopardized. Therefore, it can be said that in the case of minor disturbances (when the control system cannot respond promptly in a satisfactory way, and the ZMP leaves the safety zone), a sufficiently large distance between the margins of the support polygon and safety zone (Fig. 3b) prevents the system from losing balance and overturning, as the ZMP still remains within the support polygon. 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.

4 Hybrid Integrated Dynamic Control Algorithm with Reinforcement Structure

It is well known from control point of view that overall human walking balance is preserved not only through feedback control, but the feedforward motion is also changed according to the current state of the body. These two balance compensation strategies vastly increase the flexibility and robustness of the human gait. In order to enable biped humanoids to approach the same level of performance and stability as humans, a balancing controller that is capable of *retuning* the upcoming balancing motion in real-time is needed.

Considering the mentioned control criteria, it is necessary to control the following variables: positions and velocities of the robot joints, ZMP position [26], contact force at the instant of foot striking the ground (i.e. at the moment of weight acceptance by the other leg). In accordance with the control task, the application of the so-called hybrid integrated dynamic control was proposed. Here we assume the following: (1) the models (1, 2) describes sufficiently well the behavior of the system presented in Fig. 1; (2) desired (nominal) trajectory of the mechanism performing a dynamically balanced gait is known; (3) geometric and dynamic parameters of the mechanism and driving units are known and constant. The ground reaction forces can be measured by implementing three-axis force sensors.

Based on the above assumptions, the block diagram of the hybrid dynamic controller for biped locomotion is presented in Fig. 4. It involves three feedback loops: (1) basic dynamic controller for trajectory tracking, (2) dynamic reaction feedback at the ZMP based on reinforcement learning structure, (3) impact-force feedback at the foot of the free (unconstrained) leg. Structurally, the controller consists of the so-called *basic dynamic controller*, *compensator of dynamic reactions*, and *force compensator* which is activated by the switch K depending on the need of the specified control task.

The vector of driving moments $\hat{P}$ represents the sum of the driving moments $\hat{P}_1$, $\hat{P}_2$ and $\hat{P}_3$. 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 torques $\hat{P}_3$ are determined with the objective of diminishing the amplitude of the force F that arises at the instant when the free foot touches the

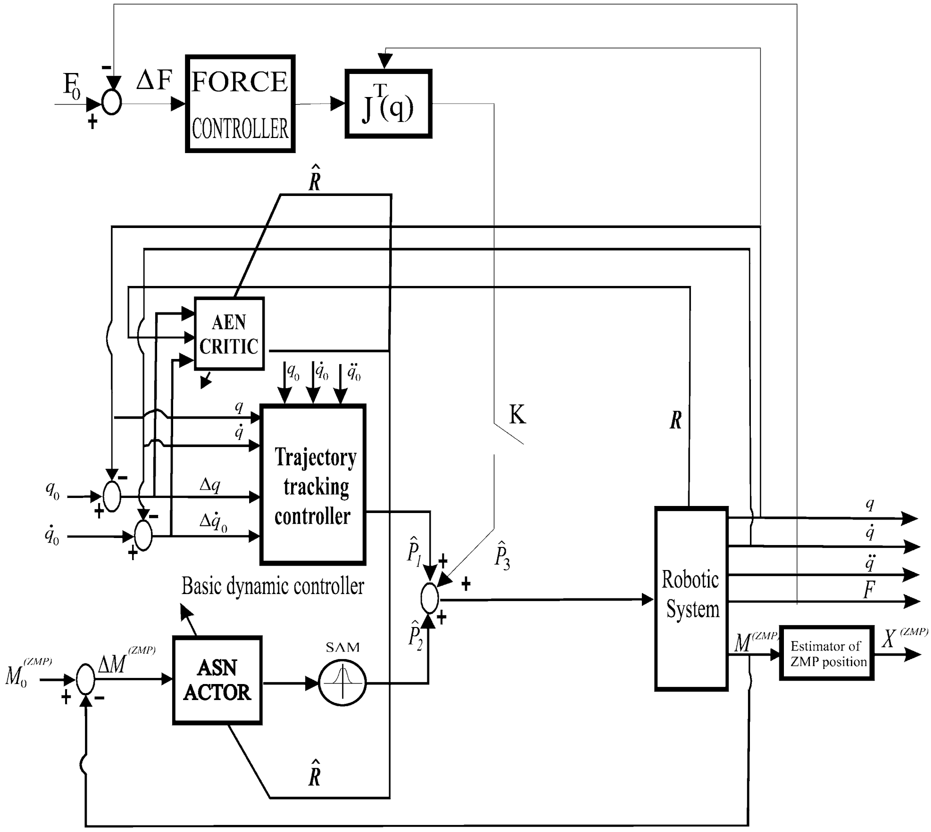


Fig. 4 Block-scheme of the hybrid integrated dynamic control of biped with two feedback loops

ground. The vector of driving torques $\hat{P}$ represents the output control vector. In the text below we describe in detail the control algorithms based on the mentioned feedbacks (1)–(3).

4.1 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. In [26], it has been demonstrated how local PD or PID controllers of biped locomotion robots are being designed. On the other hand, in [19, 20] it was demonstrated how to synthesize a trajectory tracking controller by combining the computed torque method and impedance control method, or by applying the hybrid position-force control. 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 defined by the relation (2). Since the estimated values of mechanism parameters generally differ from the real ones, The proposed dynamic control law has the following form:

$$\hat{P}_1 = \hat{H}(q, x_{cf})[\ddot{q}_0 + K_v(\dot{q} - \dot{q}_0) + K_p(q - q_0)] + \hat{h}(q, \dot{q}, x_{cf}, \dot{x}_{cf}, \ddot{x}_{cf}) - \hat{J}^T(q, x_{cf})F \tag{3}$$

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 $K_p = diag\{k_p^i\}$, $K_v = diag\left\{k_v^i\right\}, i = 1, .., n.$ can be chosen in the diagonal form by which the system is decoupled into n independent subsystems. Assuming that the corresponding matrices and vectors are identical (the case when the system parameters can be identified with a high accuracy), n characteristic polynomials of the subsystems are obtained in form $f_i(s) = s_i^2 + k_v^i s_i + k_p^i$. The gains k_p^i and k_v^i are scalar positive quantities that are determined by the root locus method or by assigning the characteristics in the frequency domain. In the case of large perturbations influencing the robotic mechanism motion, the conflict between the particular controllers (Fig. 4) is possible.

The proposed controller is valid for both the single-support and double-support gait phase (Fig. 3), whereby it is assumed that in each instant, at least one foot is in contact with the support surface. A prerequisite for a proper work of the controller is that in each time instant is known which leg is ‘supporting’ and which is conditionally ‘free’ (Fig. 3). A foot becomes ‘supporting’ at the instant when the other foot ceases to act as support, i.e. at the instant when it deploys from the support and the physical contact between the foot and the ground terminates (at the end of the weight-support phase, Fig. 3). At that moment, on the basis of the signals from the contact-force sensors at the feet, the controller switches the origin of the coordinate system from the base of the one foot to the other. In the double-support phase, the vector of ground reaction forces/torques is determined by the direct measurement on the foot.

4.2 Compensator of Dynamic Reactions Based on Reinforcement Learning Structure

In the sense of mechanics, a bipedal locomotion mechanism represents an inverted multilink pendulum. In the presence of elasticity in the system and external environment factors, the mechanism's motion causes dynamic reactions at the robot supporting foot. Thus, the state of dynamic balance of the locomotion mechanism changes accordingly. For this reason it is essential to introduce in the control synthesis the dynamic reaction feedback at ZMP.

There are relationship between the deviations of ZMP positions ($\Delta x^{(\mathrm{zmp})}$, $\Delta y^{(\mathrm{zmp})}$) from its nominal position 0_{zmp} in the motion directions x and y and the corresponding dynamic reactions $M_x^{(\mathrm{zmp})}$ and $M_y^{(\mathrm{zmp})}$ acting about the mutually orthogonal axes that pass through the point 0_{zmp}. $M_x^{(\mathrm{zmp})} \in R^{1\times 1}$ and $M_y^{(\mathrm{zmp})} \in R^{1\times 1}$ represent the moments that tend to overturn the robotic mechanism, i.e. to produce its rotation about the mentioned rotation axes (axes of the joints 1' and 2' in Fig. 1). $M_0^{(\mathrm{zmp})} \in R^{2\times 1}$ and $M^{(\mathrm{zmp})} \in R^{2\times 1}$ are the vectors of nominal and measured values of the moments of dynamic reaction around the axes that pass through the ZMP (Fig. 3a). 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; $\Delta M^{(\mathrm{zmp})} \in R^{2\times 1}$ is the vector of deviation of the actual dynamic reactions from their nominal values; $P_{dr} \in R^{2\times 1}$ is the vector of control torques at the joints 1' and 2' (Fig. 1), ensuring the state of dynamic balance.

On the basis of the above, the dynamic compensator can be realized using some conventional control structures (PI-regulator) [27]. But in this case, when moments of dynamic reactions are not state variables of the system,complex nonlinear relations together with parameter and model uncertainties and high level of system disturbances. this type of conventional control can not give optimal performance of biped controller.

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.

4.2.1 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, 24]. 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.

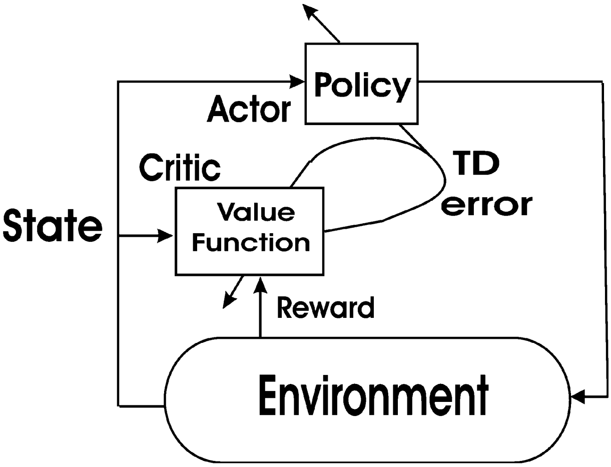
The input to control policy is state of the system, while the output is control action (signal). It searches the action space using a stochastic real valued (SRV) unit at the output. The unit's action uses a Gaussian random number generator. The estimated value function represents a *critic*, because 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. 5).

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. If ZMP error is greater then chosen limit, external reinforcement signal is set to value 1, otherwise this signal is set to zero.

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 two-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.

Fig. 5 The actor–critic architecture



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:

$$\hat{R}(t+1) = R(t) + \gamma v(t+1) - v(t) \tag{4}$$

where γ is a discount coefficient between 0 and 1 (in this case γ is set to 0.9). The previous equation defines TD(0) algorithm, where only new states are taken into account.Sutton and Barto [24] spreads the evaluation for all states, according to their *eligibility*, i.e. the degree of visit of a state in the recent past. The eligibility is defined by:

$$e_t = 1 + \gamma\lambda e_{t-1} \qquad \text{if} \quad x = x_t \tag{5}$$

$$e_t = \gamma\lambda e_{t-1} \qquad \text{otherwise} \tag{6}$$

In this case, TD(λ) algorithm is represented by following equation for internal reinforcement:

$$\hat{R}(t+1) = (R(t) + \gamma v(t+1) - v(t))e_t \tag{7}$$

where $\lambda, 0 \leq \lambda \leq 1$ is new parameter.

ASN (action selection network) maps the deviation of dynamic reactions $\Delta M^{(\text{zmp})} \in R^{2\times 1}$ in recommended control torque. The structure of ASN is represented by The ANFIS – Sugeno-type 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 knowledge with 25 rules. The partition of input variables (deviation of dynamic reactions) are defined by five 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 P_{dr}. It is defined by Gaussian random function where recommended control torque is mean, while standard deviation is defined by following equation:

$$\sigma\left(\hat{R}(t+1)\right) = \left|\left(1 - \exp^{-|\hat{R}(t+1)|}\right)\right| \tag{8}$$

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 appropriate input values from previous layers. The learning process for ASN (tuning of antecedent 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 P_{dr} obtained as output of actor structure cannot be generated at the joints 1' and 2' since they are empowered(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 $\Delta M^{(\text{zmp})}$ has two components about the mutually orthogonal axes x and y, at least two different active joints have to be used to compensate for these dynamic reactions. Considering the model of locomotion mechanism presented in Fig. 1, the compensation was carried out using the following mechanism joints: 1, 6 and 14 to compensate for the dynamic reactions about the x-axis, and 2, 4 and 13 to compensate for the moments about the y-axis. Thus, the joints of ankle, hip and waist were taken into consideration. Finally, the vector of compensation torques $\hat{P}_2$ (Fig. 4) was calculated on the basis of the vector of the moments P_{dr} whereby it has to be borne in mind how many 'compensational joints' have really been engaged. If only two joints (e.g. the ankle joints 1 and 2) are engaged, then:

$$\hat{P}_2(1) = SAM(P_{dr}(1)),\ \hat{P}_2(2) = SAM(P_{dr}(2)) \tag{9}$$

In the case when compensation of ground dynamic reactions is performed using all six proposed joints, the compensation torques P_{dr} are uniformly distributed over all of the selected joints, to load uniformly the actuators. Then the following relations hold:

$$\begin{aligned}\hat{P}_2(1) = \hat{P}_2(6) = \hat{P}_2(14) = 1/3 \quad SAM(P_{dr}(1)),\\ \hat{P}_2(2) = \hat{P}_2(4) = \hat{P}_2(13) = 1/3 \quad SAM(P_{dr}(2))\end{aligned} \tag{10}$$

In nature, biological systems use simultaneously a large number of joints for correcting their balance. In this work, for the purpose of verifying the control algorithm, the choice was restricted to the mentioned six joints: 1, 2, 4, 6, 13 and 14 (Fig. 1). Compensation of ground dynamic reactions is always carried out at the supporting leg when the locomotion mechanism is in the swing phase, whereas in the double-support phase it is necessary to engage the corresponding pairs of joints (ankle, knee, hip) of both legs. It is clear that the Eqs. 9 and 10 are valid only in the case when the kinematic scheme of the biped mechanism is chosen in the way illustrated in Fig. 1. Otherwise, if the humanoid robot configuration differs from the one presented in Fig. 1, the relations (9) and (10) should be adapted to the new kinematic scheme. In principle, the compensating joints ought to be always selected in accordance with the chosen mechanism configuration.

4.3 Impact-Force Controller

The role of impact-force controller in the proposed control structure (Fig. 4) is to counteract the ground reaction force that appears when the locomotion mechanism foot strikes the ground (heel strike). At that moment, the forces (and moments) that destabilize the system are transferred onto all of its links. Hence, a common practice is to introduce passive and/or active compliance into the system. Elastic footpad attached to the rigid foot sole is commonly applied as a passive compliance. Active compliance assumes the involvement of an active force or moment at the mechanism joint, generated by the contact-force controller. With humans, the load forces arising at the joints of the legs and hip during walk are absorbed by the appropriate contraction of leg muscles. With robots, the role of muscles is played by the actuators at the mechanism joints. In [19, 20], a scheme was proposed for impedance control, whereby contact force was controlled by adaptive change of the mechanism legs impedance. In this work we propose the introduction of a force controller to control in a direct way the contact forces and moments at the landing (free-leg) foot. The magnitude of the ground reaction force is equal to the magnitude of the strike force. In the sequel we will assume that the vector $F \in R^{6\times 1}$ represents both the forces and moments of ground reaction in the case when the foot is in contact with the environment. The control law can be defined in the form of a PI-regulator as:

$$F_c = F_0 - \int\limits_0^t Q(\Delta F)\cdot dt, \quad t \in (0, T], \quad \Delta F = F - F_0 \tag{11}$$

$$F_0 = \left[F_{0x}\ F_{0y}\ F_{0z}\ M_{0x}\ M_{0y}\ M_{0z}\right]^T, \quad F = \left[F_x\ F_y\ F_z\ M_x\ M_y\ M_z\right]^T, \tag{12}$$

$$Q(\Delta F) = K_{PF}\Delta F + K_{IF}\int\limits_0^t \Delta F\cdot dt, \quad t \in (0, T] \tag{13}$$

where $F_c \in R^{6\times 1}$ is a vector of the so called generalized forces. In order to compensate deviations of ground reactions ΔF the generalized forcesF_c should to act at the mass center of the last link of the swinging leg (i.e. at the mass center of the landing foot). $F_0 \in R^{6\times 1}$ 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 motion q_0; $K_{PF} \in R^{6\times 6}$ and $K_{IF} \in R^{6\times 6}$ are the square matrices of proportional and integral gains of the designed PI-regulator. If the gain matrices are adopted in the diagonal form $K_{PF} = diag\ k^j_{pf},\ \ K_{IF} = diag\ k^j_{if},\ \ j = 1, 6, k^j_{pf},\ \ k^j_{if} > 0$ the characteristic polynomial is obtained in the form $f_j(s) = s^2 + k^j_{pf}s + k^j_{if}$. The control gains are determined (see [27]) using the methods known from the theory of linear control systems. Finally, since the generalized forces F_c cannot be realized in a direct way, the control torques $\vec{P}$ (Fig. 4) are determined from the relation:

$$\hat{P}_3 = -\hat{J}^T(q, x_{cf})F_c \tag{14}$$

where $\hat{J}(q, x_{cf})$ is the Jacobian matrix; x_{cf} is position vector of constrained foot It should be pointed out that the contact force feedback serves to correct the dynamic performance of the robotic system, ie. to ensure a smoother walking and minimize the effect of the ground reaction of a pulsed character.

4.4 Conflict Between Controllers

The conflict between the compensation actions at the particular joints as a control problem with biped robotic systems has been known for many years (see for example [26]), whereby the control task considered was divided into two subtasks: (1) tracking nominal trajectories of the mechanism joints and (2) maintaining the mechanism dynamic balance. The subtask (1) performed all the mechanism joints, while subtask

(2) was allocated to only one (e.g. ankle) joint In the instance of endangered dynamic balance, the joint reacted so to return the ZMP to its nominal position. It is obvious that the additional requirement imposed on the ankle joint spoiled the quality of its nominal trajectory tracking [26], but the maintaining of dynamic balance (preventing mechanism's falling) is a priority task. The solving of this problem will be briefly explained in this section.

The proposed controllers (Fig. 4), operate independently from each other. These controllers generate the corresponding compensation torques $\hat{P}_1$, $\hat{P}_2$, $\hat{P}_3$, to be realized by the actuators at the particular joints of the mechanism. In a general case, in the presence of large disturbances (unpredictable changes of the environment, sudden impacts, model inaccuracies, etc.) acting upon the system, a conflict can arise between the calculated compensation torques. This conflict exerts a negative effect on the state of the system.

There are various methods for overcoming the problem of possible conflict between the particular controllers implemented with robotic biped systems. It is especially worth mentioning the novel whole-body cooperative balancing method with modification of predesigned motion trajectories [23]. By implementing this advance method, the authors showed through a simulation example the way how to control simultaneously the joints' positions and dynamic balance of a system in real-time with no conflict between the applied controllers. The essence of the applied method is in determining and manipulating the position of the center of gravity (COG). The manipulation was performed so to ensure the system's balance in an anthropomorphic way, by calculating new reference positions q_{ref} of the mechanism's joint angles instead of their pre-defined values. The important factors to 'balance' legged motions are [23]: (1) modification of the pre-designed posture trajectory in order to conserve the contact condition, and (2) robust compensation of the deviation of the real posture from the pre-defined one. The considered method can also be used to calculate the control torques determined by relation 3 in this paper. In the case of large perturbations acting on the robotic system, the proposed compensator of dynamic reactions P_{dr} can produce significant magnitudes of the compensating torques $\hat{P}_2$, to ensure desired dynamic reactions on the biped locomotion mechanism. As a consequence, these torques can spoil the accuracy of the joints' positions. In order to overcome this control conflict, the modified values of the reference positions $q_|ref$ should be used in Eq. 3 instead of the pre-designed values of the nominal trajectory q_0. By the real-time modification of the pre-defined trajectory in the course of the mechanism's motion, a possible conflict between the controllers $\hat{P}_1$ and $\hat{P}_2$ is avoided. The way of a suitable modification was explained in detail in the paper by Sugihara and Nakamura [23]. Using the reference trajectory q_{ref}, instead of the predefined nominal one q_0, less energy is consumed by the robot actuators to realize a desired motion, whereby the dynamic performance of the system is ensured.

5 Simulation Studies

The proposed control methods were analyzed on the basis of the numerical data obtained by simulation of the closed-loop model of the locomotion mechanism. Lump

Table 1 The parameters of the biped robot used in simulation examples

Link	Length (m)	Mass (kg)	Tensor of inertia (kg m^2)		
			Ix	Iy	Iz
Foot	0.250	1.15	0.0060	0.0054	0.0014
Ankle	0.100	1.53	0.0006	0.0055	0.0045
Shin	0.420	3.21	0.0393	0.0393	0.0038
Thigh	0.440	8.41	0.1120	0.1200	0.0300
Lower trunk–hip	0.270	6.96	0.0700	0.0565	0.0627
Upper trunk	0.400	30.85	1.5140	1.3700	0.2830
Head	0.220	5.83	0.0397	0.0428	0.0295
Upper arm	0.308	2.07	0.0200	0.0200	0.0022
Forearm	0.264	1.14	0.0250	0.0425	0.0014
Hand	0.120	0	0	0	0

mass of the mechanism is m = 70 [kg]. Basic geometric and dynamic parameters of the considered biped mechanism (Fig. 1) are given in Table 1.

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 (defined by its kinematic scheme in Fig. 1) walking on the ideally flat, horizontal surface during a half-step phase of the gait is shown in Fig. 6. The half-step is repeated with this period T, whereby the gait phases presented in Fig. 2 alternate regularly. Nominal trajectories at robot joints were synthesized for the gait on the ideally horizontal ground surface. Nominal angles at the mechanism joints (Fig. 7) and the corresponding angular velocities and accelerations were determined by the semi-inverse method

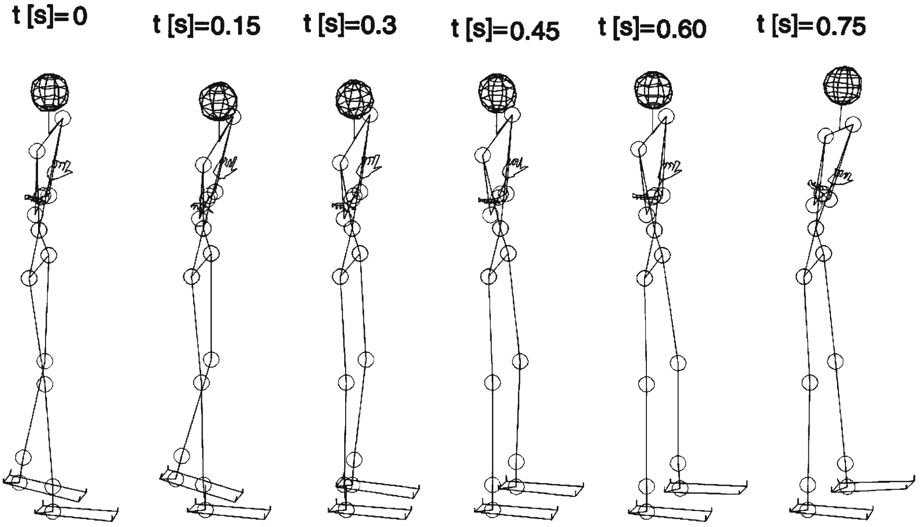


Fig. 6 Computer animation of the stick-model of biped locomotion mechanism during the half-step

[26]. 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. They were imposed by the definition of the following deviation vectors: $\Delta q = [0\ 0\ 0\ 0.051\ 0\ -0.023\ 0.034\ 0\ -0.02\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0]^T$ (rad). Simulation results were analyzed on the time interval T corresponding to the duration of one half-step of the locomotion mechanism in the swing phase, including the free (landing) foot strike instant.

The analysis of the efficiency of the proposed controllers (Fig. 4) in the realization of dynamically balanced motion shows that the most delicate instances are the single-support phase (swing phase) and the free foot striking the ground (beginning of the double-support phase). As the robot's dynamic behavior in these time intervals is of special importance for control, simulation examples were selected so to encompass these critical phases of biped gait. Since the single-support and double-support phases of the robotic gait (Fig. 2) are alternately repeated regularly, the physical phenomena existing within one half-step are repeated in the course of the gait without larger variations. Hence the research attention has been focused on the half-step domain, encompassing also the first instants of the double-support phase.

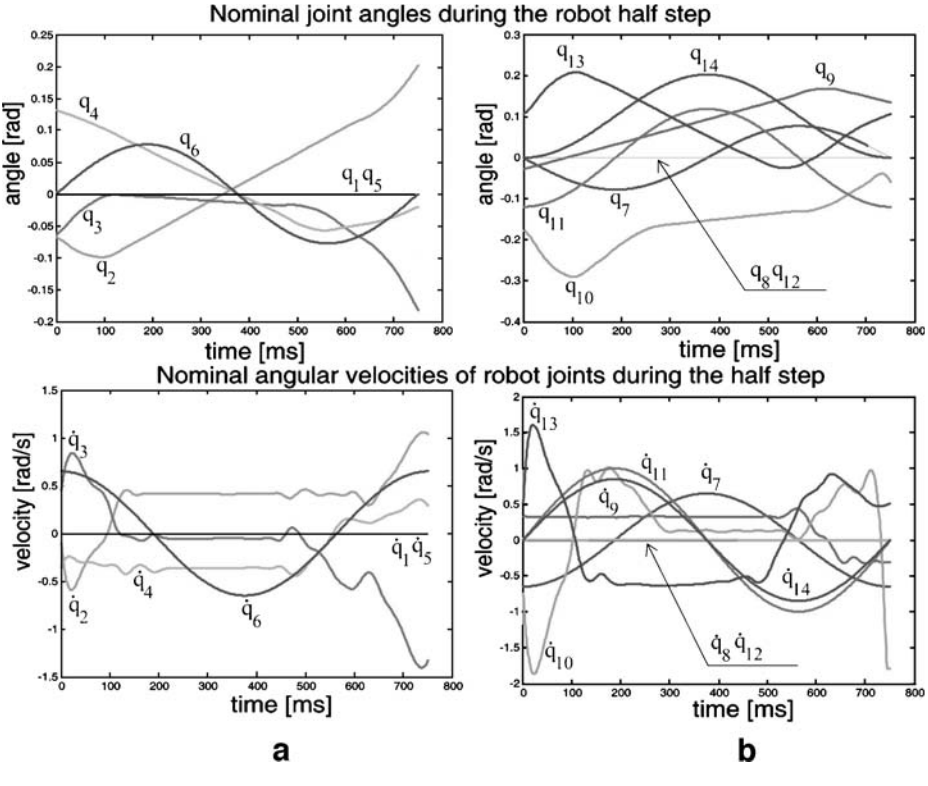


Fig. 7 Nominal trajectory of robot mechanism – joint angles and joint angular velocities **a** small constrained leg joints 1–6, **b** swing leg joints 7–12 and robot trunk joints 13–14

In the first 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. The nominal joint trajectories of the robotic mechanism were synthesized assuming that the biped walks on an ideally horizontal supporting surface. However, the supporting surface position can generally deviate from the ideal case a horizontal plane. Such geometry deviation represents a perturbation to the robotic system. Because of that, the robot controller has to adjust the mechanism to these unknown perturbations, to prevent its overturning and ensure a dynamically balanced walk. 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.

In this simulation, the robot's behavior in the swing phase is considered, when the robot relies upon the ground by its rigid foot. The other one (free or swinging foot) is above the ground. Here, the control algorithms were analyzed using two cases: (1) only trajectory tracking controller described by the relation 3 (without learning) and (2) hybrid control, consisting of the trajectory tracking controller 3 and reinforcement learning compensator of dynamic reactions of the ground at the ZMP (with learning), In Figs. 8 and 9 the comparison of the simulation results for ZMP errors in coordinate directions by applying the nonlearning (1) controller and hybrid learning (2) controller are shown.

Thus, it can be concluded that with inclusion of the reinforcement learning feedback structure, it is possible to ensure (guarantee) dynamic balance of a locomotion mechanism during its motion and good tracking capabilities. This comes out from the fact that the pre-designed (nominal) trajectory was synthesized without taking into account possible inclination deviations of the surface on which biped walks from

Fig. 8 Error of ZMP in x-direction

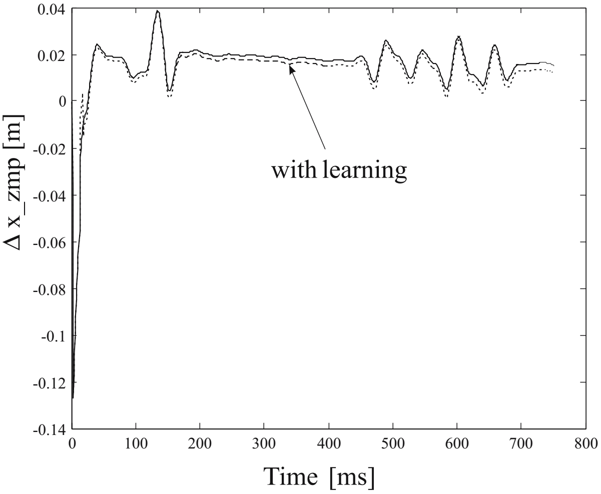
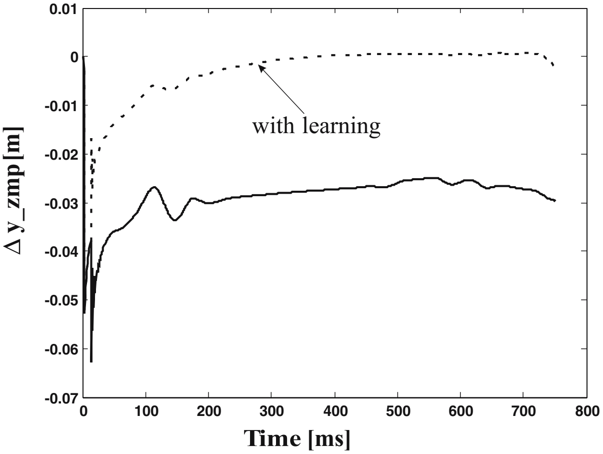


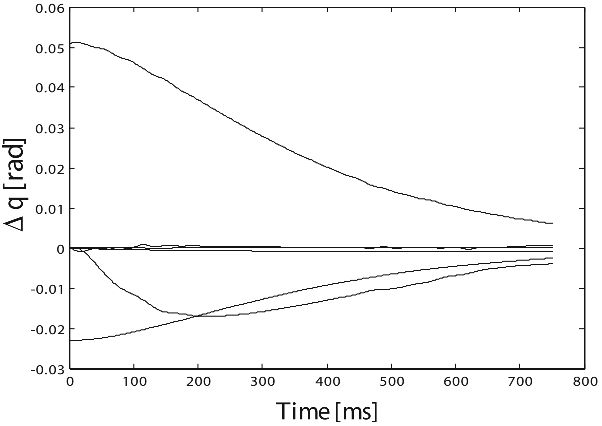
Fig. 9 Error of ZMP in y-direction



an ideally horizontal plane. Therefore, the ground surface inclination influences the system's balance as an external stochastic perturbation.

The corresponding deviations (errors) Δq_i and $\Delta \dot{q}$ of the actual angles and angular velocities at the robot joints from their reference values, in the case when the hybrid learning controller was applied, are presented in Figs. 10 and 11. The deviations Δq_i and $\Delta \dot{q}$ converge to zero values in the given time interval $T = 0.75$ (s). It means that the controller employed ensures good tracking of the desired trajectory. Besides, the corresponding control torques at the robot joints $\hat{P}_1$ and $\hat{P}_2$, ensures a desired convergence of joint positions and ZMP position to the corresponding reference positions as well as a dynamic balance of the locomotion mechanism as it is illustrated in Figs. 8 and 9. Simulation results shown in Figs. 8–11 confirm the efficiency of the proposed hybrid dynamic controller applied in controlling robot's gait on the ground surface whose inclination deviates from an ideal horizontal plane.

Fig. 10 Convergence of the errors of tracking nominal angles



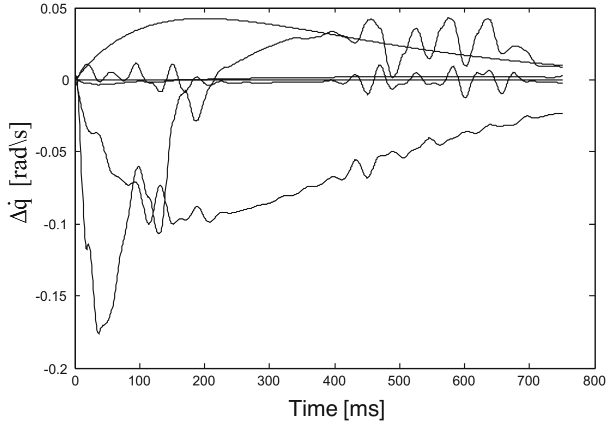


Fig. 11 Convergence of the errors of tracking nominal angular velocities

In Fig. 12, value of internal reinforcement through process of walking is presented. In this case, fast learning rate is achieved, because control algorithm have possibility to realize on fast way, task of walking within desired ZMP tracking error limits.

Finally, the simulation example proves that the impact force controller together with the reinforcement learning compensator of dynamic reactions successfully controls the impact/contact force at the landing foot as well as keeps well the dynamic balance of the robotic system.

This simulation example was concerned with the quality of controlling impact forces in the course of robot motion. The attention was focused on the instant when the mechanism's free (landing) foot comes in contact with its dynamic environment (weight-acceptance phase of the gait), whereby it is necessary to minimize the effect of the impact forces on the overall mechanism. It is the instant when the load is transferred from the supporting leg to the other leg. To realize control

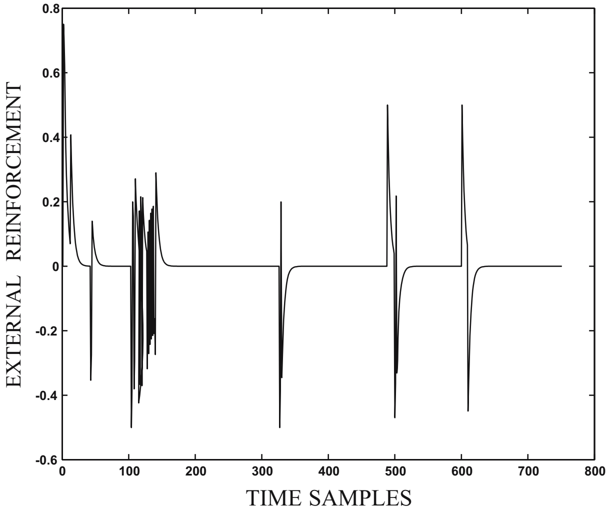


Fig. 12 Internal reinforcement through process of walking

action, additional contact-force feedback was introduced. Having in mind that both feet are equipped with force sensors for measuring force components in all three directions of robot motion (*x, y, z*), it is also possible to calculate the corresponding reaction moments acting on the foot about these axes. The additional controller was synthesized according to the relations 11–14. At the end of the swing phase, the free foot strikes the ground at a certain rate and acceleration, which is accompanied by the vertical impact presented in Fig. 13. By using the contact-force controller it was possible to reduce significantly the amplitude of this force. In Fig. 14 are also

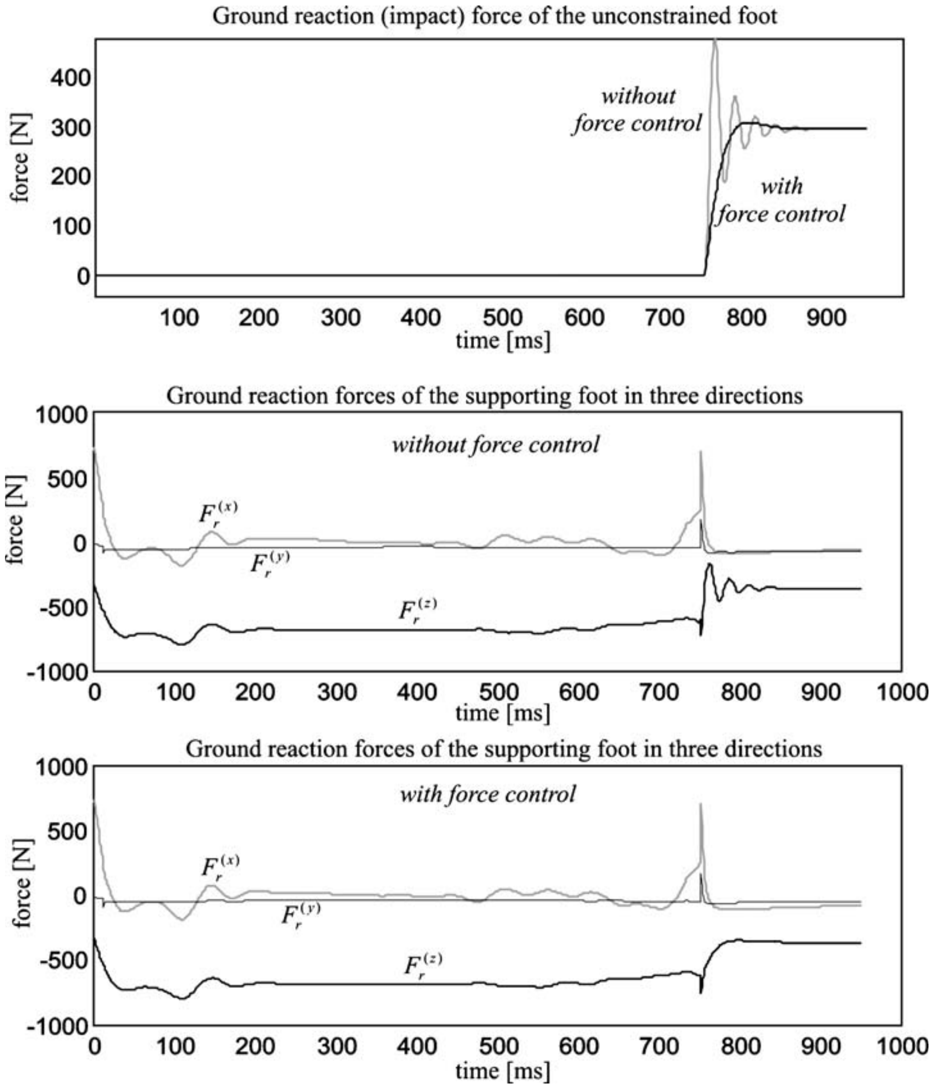


Fig. 13 Comparative indicators of controlling the contact force between the free foot and the ground. Forces acting between the ground and foot of the mechanism's supporting leg with as well without contact-force control in three coordinate directions

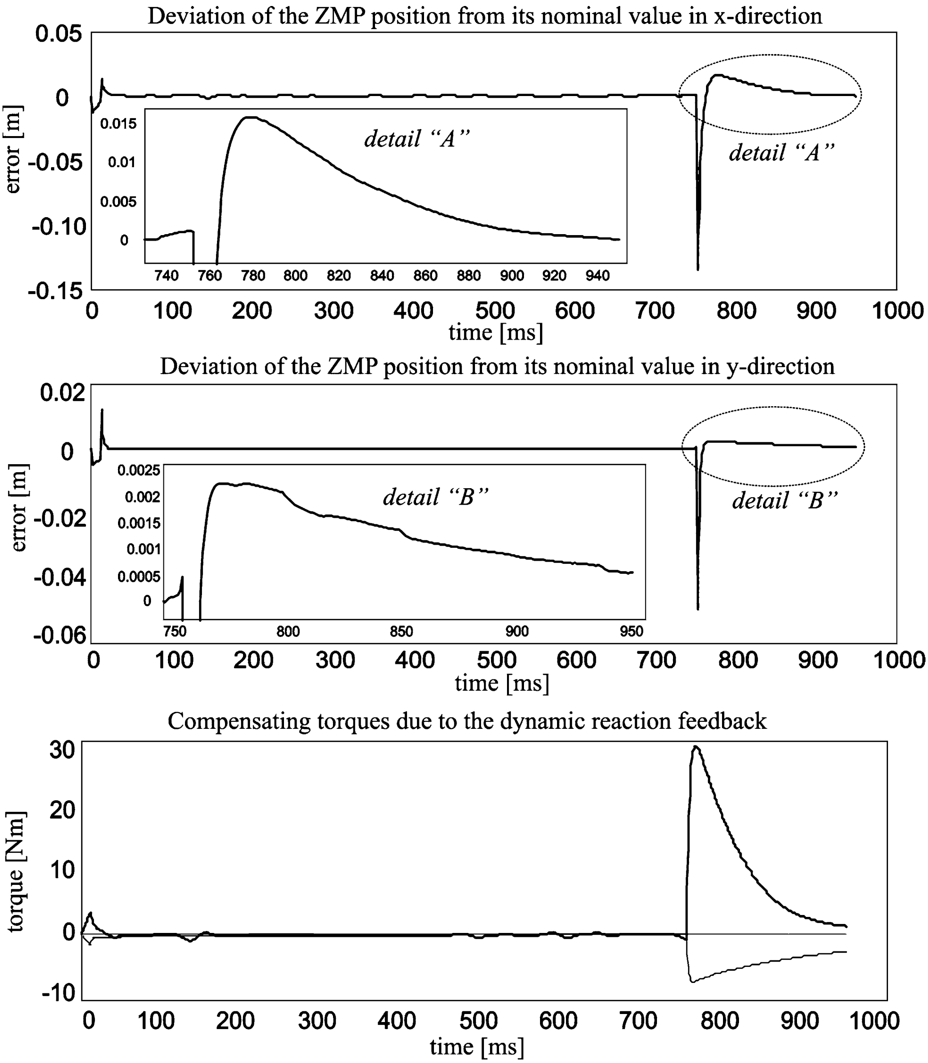


Fig. 14 Deviations of the ZMP position caused by the contact force arising at the moment of free-foot striking the ground (case of existence of the impact-force feedback). Corresponding driving torques

presented the considered pressure forces at the supporting foot on the ground, with and without contact-force control.

Figure 14 shows the indicators of the state of the mechanism's dynamic balance before and after the moment when the free foot strikes the ground, using the contact-force control. At the moment of strike, there appears a very pronounced discontinuity in the ZMP position. However, the ZMP comes back very quickly, and after only several tens of millisecond it is at the border of the safety zone. By an appropriate choice of nominal motion it is possible to reduce significantly the nominal of the ground impact. Thanks to the action of the hybrid dynamic controller, the mechanism will preserve dynamic balance at the moment of foot striking the ground.

6 Conclusions

In this paper, an hybrid integrated dynamic controller for biped locomotion was proposed. The 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 actor–critic reinforcement learning architecture. The proposed reinforcement learning structure is based on two networks: AEN (neural network) and ASN (neuro-fuzzy network) and represent the efficient learning tool for compensation of ZMP reactions. Hybrid dynamic controller was designed with the aim of ensuring precise tracking of the given motion and maintaining dynamic balance of the humanoid mechanism.The primary control structure was supplemented with additional impact-force controller which aim is to stabilize the forces around the nominal values, determined for the nominal conditions of motion.

The developed hybrid 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, as it possesses both the conventional position-velocity feedback and dynamic reaction feedback. Performance of the control system was analyzed and successfully validated in a number of simulation experiments in the presence of different types of external and internal perturbations acting on the system.

The future research includes improvements of reinforcement learning structures by using more efficient Q-learning algorithm, as well as possibility of using external fuzzy reinforcement signal.

Acknowledgements The described work was conducted within the research project *Dynamic and Control of High-Performance Humanoid Robots: Theory and Application*. that was funded by the Ministry of Science and Environmental Protection of the Republic of Serbia.

References

1. Atkeson, C.G., Santamaria, J.C.: A comparison of direct and model-based reinforcement learning. In: Proceedings of the 1997 IEEE International Conference on Robotics and Automation, pp. 3557–3564. Albuquerque, USA (1997)
2. Benbrahim, H., Franklin, J.A.: Biped dynamic walking using reinforcement learning. Robot. Auton. Syst. **22**, 283–302 (1997)
3. Berenji, H.R., Khedkar, P.: Learning and tuning fuzzy logic controllers through reinforcements. IEEE Trans. Neural Netw. **3**, 724–740 (1992)
4. Bertsekas, D.P., Tsitsiklis, J.N.: Neuro-Dynamic Programming. Athena Scientific, Belmont, USA (1996)
5. Chew, C., Pratt, G.A.: Dynamic bipedal walking assisted by learning. Robotica **20**, 477–491 (2002)
6. Doya, K.: Reinforcement learning in continuous time and space. Neural Comput. **12**, 219–245 (2000)
7. Gienger, M., Löffler, K., Pfeiffer, F.: Walking control of a biped robot based on inertial measurement. In: Proceedings of IARP International Workshop on Humanoid and Human Friendly Roboticsm, pp. 22–29. Tsukuba, Japan (2002)

8. Gullapalli, V.: A stochastic reinforcement learning algorithm for learning real-valued functions. Neural Netw. **3**, 671–692 (1990)
9. Gullapalli, V., Franklin, J.A., Benbrahim, H.: Acquiring robot skills via reinforcement learning. IEEE Control Systems Magazine, pp. 13–24 (1994)
10. Hirai, K., Hirose, M., Haikawa, Y., Takenaka, T.: The development of honda humanoid robot. In: Proceedings of the 1998 IEEE Int. Conference on Robotics and Automation, pp. 1321–1326 (1998)
11. Kamio, S., Iba, H.: Adaptation technique for integrating genetic programming and reinforcement learning for real robot. IEEE Trans. Evol. Comput. **9**(3), 318–333 (June 2005)
12. Katić, D., Vukobratović, M.: Survay of intelligent control techniques for humanoid robots. J. Intell. Robot. Syst. **37**, 117–141 (2003)
13. Katić, D., Vukobratović, M.: Intelligent Control of Robotic Systems. Kluwer, Dordrecht, The Netherlands (2003)
14. Katic, D., Vukobratovic, M.: Survey of intelligent control algorithms for humanoid robots. In: Proceedings of the 16th IFAC World Congress, Prague, Czech Republic, July 2005
15. Li, Q., Takanishi, A., Kato, I.: Learning control of compensative trunk motion for biped walking robot based on ZMP. In: Proceedings of the 1992 IEEE/RSJ Intl. Conference on Intelligent Robot and Systems, pp. 597–603 (1992)
16. Mori, T., Nakamura, Y., Sato, M., Ishii, S.: Reinforcement learning for a cpg-driven biped robot. In: Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI), pp. 623–630 (2004)
17. Morimoto, J., Cheng, G., Atkeson, C.G., Zeglin,G.: A simple reinforcement learning algorithm for biped walking. In: Proceedings of the 2004 IEEE International Conference on Robotics & Automation. New Orleans, USA (2004)
18. Nakamura, Y., Sato, M., Ishii, S.: Reinforcement learning for biped robot. In: Proceedings of the 2nd International Symposium on Adaptive Motion of Animals and Machines (2003)
19. Park H.J., Chung H.: Hybrid control for biped robots using impedance control and computed-torque control. In: Proceedings of the 1999 IEEE International Conference on Robotics and Automation, pp. 1365–1370. Detroit, USA (1999)
20. Park, H.J.: Impedance control for biped robot locomotion. IEEE Trans. Robot. Autom. **17**(6), 10–6 (2001)
21. Peters, J., Vijayakumar, S., Schaal, S.: Reinforcement learning for humanoid robots. In: Proceedings of the Third IEEE-RAS International Conference on Humanoid Robots, Karlsruhe & Munich (2003)
22. Salatian, A.W., Yi, K.Y., Zheng, Y.F.: Reinforcement learning for a biped robot to climb sloping surfaces. J. Robot. Syst. **14**, 283–296 (1997)
23. Sugihara, T., Nakamura, Y.: Whole-body cooperative balancing of humanoid robot using COG Jacobian. In: Proceedings of the 2002 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 2575–2580, Lausanne (2002)
24. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. The MIT Press, Cambridge, USA (1998)
25. Tedrake, R., Zhang, T.W., Seung, H.S.: 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)
26. Vukobratović, M., Borovac, B., Surla, D., Stokić, D.: Biped Locomotion - Dynamics, Stability, Control and Application. Springer, Berlin, Germany (1990)
27. Vukobratovic, M., Ekalo, Y.: New approach to control of robotic manipulators interacting with dynamic environment. Robotica **14**, 31–39 (1996)
28. Vukobratović, M., Borovac, B.: Zero-moment point – thirty five years of its life. Int. J. Humanoid Robot **1**, 157–173 (2004)
29. Vukobratović, M., Borovac, B.: Note on the article "zero-moment point – thirty five years of its life". Int. J. Humanoid Robot. **2**, 225–227 (2005)
30. Watkins, C.J.C.H., Dayan, P.: Q learning. Mach. Learn. **8**, 279–292 (1992)
31. Yokoi, F., Kanehiro, F., Kaneko, K., Fujiwara, K., Kajita, S., Hirukawa, H.: Experimental study of biped locomotion of humanoid robot HRP-1S. In: Siciliano, B., Dario, P. (eds.) Experimental Robotics VIII, pp. 75–84. Springer, Berlin, Germany (2003)
32. 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, pp. 3829–3834. San Francisco, USA (2000)