Policy Gradient Fuzzy Posture Learning for Frontal Plane Balance of a Biped Walker

Chun-Hang Hsu*, Tsan-Yu Chen, Ying-Hao Yu, and Quang Ha

***Abstract*— In robotic control, learning the biped gait of a robot is often a challenging issue due to the complexity of the robot's inverted pendulum motion. The difficulty of biped walking is not only one that is associating with highly nonlinear dynamics but also with the ability to adapt the gait to different speeds or terrain. More recently, to take advantage of the model-free approach, many researchers have developed biped gait using some variants of reinforcement learning. Due to a large number of tumbles, these robots generally rehearse by computer animation in advance. To lower the computational cost, here we adopt the Mamdani fuzzy inference method for economical use of parameters and then fine-tune the output membership functions by using policy gradient-based reinforcement learning. The proposed method, applied directly to an adult-sized biped walker, is demonstrated by balancing its frontal plane to see how it works on walking. The robot can then perform from a short-step length to longer ones. According to our experimental results, a good balance of the frontal plane can efficiently stabilize the walking gait as humans do.**

I. Introduction

Building a biped robot to mimic the life-like behavior of humans has been the long dreamed-of-goal of artificial intelligence. In such a case, the gait of the robot is not dependent on heavy-computational-load control functions but learning from interactions between environments (e.g., walking on any slope, based on predefined knowledge [1]). This motive makes reinforcement learning (RL) an alternative while the robot explores a new environment autonomously.

Unlike the supervised learning approach [2], the RL uses trial and error to learn from interaction with the environment. Therefore, the following action of a current environment status is as a result of this interaction [3]. A typical methodology of RL is Q-learning. It employs a Q-table to record the highest interacting reward as the basis of the agent's action. Such a mechanism is feasible to implement gait learning on biped robots. By receiving the data from sensors, the detected pose of the robot will be a state in Q-table. The robot's joint angles are then determined by the triggered action [1], [4].

However, Q-learning has two drawbacks. Firstly, the environment state needs to be defined in advance. That means that a lack of environmental status in the Q-table will directly degrade the performance of gait balance. Even though the designer has rehearsed the robot's gait with computer animation [1], the robot may still tumble due to exceeding tolerances or operating in an unknown region. Unfortunately, increasing the number of statuses will grow proportionally to an increase in the learning time [4]. Secondly, the data set in a Q-table is discrete, making Q-learning unsuitable for continuous motion (e.g., biped walking [5]). Although interpolating discrete data [1] or designing a multiple-layer neural network to replace states of Q-learning [5] can contribute to overcoming this kind of drawback, the introduction of policy gradient-based reinforcement learning (PGRL) may be promising for continuous motion [6]. Compared to Q-learning, PGRL provides the advantage of being able to fine-tune learning policies without the use of action-valued functions [7]. Such an advantage enables the biped robot to be controlled as a model-free system [6].

For the walking design of an adult-sized biped robot, the moment of inertia of the robot trunk is much larger than that of a miniature robot. Unless the biped robot is designed for the running gait, servomotors at the ankle will not be able to pull the trunk back while the center of mass (COM) is outside the base of support (the foot) [1], [8]. An alternative is to reduce the weight and increase the dimension of the foot sole to guarantee that the COM will not run off the base of support [9]. It can be seen that the walking gait designed by simply replaying simulated actions [4] is unrealistic for adult-sized biped robots in that the use of closed-loop control is critically important to keep COM within the base of support for single-leg stance. Rai et al. also intended to deal with uncertainties by optimizing the robot's gait on hardware directly [10]. However, due to a large number of training iterations, we can see that the learning methodologies such as supervised learning, Q-learning, and PGRL can all hardly take on the role of the controller to be trained directly on hardware [1], [5], [11].

Nevertheless, we can manipulate a limited set of control parameters adapting to a new gait using PGRL. For this reason, we consider fuzzy logic as a candidate for the biped robot's posture controller (PC). Fuzzy logic control (FLC) is a common approach that has been well studied in adaptive biped gait control. It is often used as an auxiliary in an adaptive control system to fine-tune the parameters of a reference model-based controller [12] or reinforcement learning system [6]. Another typical application is the complementary feedback for neurons in neural networks [13]. Unfortunately, these kinds of designs are still complicated, and an enormous rule base is usually required. The time for training and tuning these controllers is therefore prohibitive, and these control systems are difficult to be further implemented on embedded systems in the future.

For designing low computational control, a direct implementation of the fuzzy logic controller is necessary. The simplest is the Mamdani method, but the performance without tuning is quite limited. This problem can be overcome by direct adaptive fuzzy control to regulate the center of each

Chun-Hang Hsu, Tsan-Yu Chen, Ying-Hao Yu are with the National Chung Cheng University, Taiwan (e-mail: hati841123@gmail.com, appieqswa@gmail.com, yinhaun@gmail.com).

Quang Ha is with the University of Technology Sydney, Australia (e-mail: Quang.Ha@uts.edu.au).

output via a Lyapunov function-based variable [14] or tuning with the Takagi–Sugeno method [15]. Phan and Gale improved the nonlinear SISO system by replacing initial membership functions with new ones [16]. Alternatively, a simple approach is to fine-tune the lower, center, and upper bounds automatically of input triangular membership functions by adding a constant tracking error. Although these FPCs can control the robot's gait, the tuning methodology is inaccurate or intuitive. This drawback has motivated us to overcome the tradeoff between control performance and low computing load. Here by adopting the Mamdani method as our fuzzy PC, we change only the core position of the output membership function to achieve a low computing load and then fine-tune it by using PGRL. We will demonstrate that our adult-sized biped walker only needs ten membership functions to learn the balance of its frontal plane for different step lengths.

This paper is organized as follows. In Section II, we first briefly introduce the biped walker's structure and gait trajectory. The design of fuzzy posture controller (FPC) by using PGRL is shown in Section III, followed by experimental results and discussions given in Section IV. Finally, Section V contains our conclusion.

II. Overview of Walker Design

The proposed posture learning control is implemented by our laboratory biped walker. To avoid sensor installation on the walker, the walker's pose is estimated by zero moment point (ZMP) on foot soles. As infants learning how to walk, we focus on online hardware learning for stride balance from a short-step length to a longer one.

A. Biped walker structure

The adult-sized biped walker, shown in Fig. 1, is about 90 cm tall and weighed 10 kg with a size of foot 20.5 (L)×12 (W) cm. Each leg is designed with six degrees of freedom (DOF), and the distance between the two legs is 21 cm. The walker's foot (110 g) is made by 3D-printed PLA (Polylactic Acid) to bear over 70 kg.

Meanwhile, the robot's ZMP is detected by using force sensor resistors (FSRs). There are four FSRs attached on the four corners of each foot sole, and each FSR can maximally measure a payload of 11.3 kg. We tested each FSR in the range of force from 0 to 8 kg and achieved 10 g in resistance resolution. The ZMP based on FSR sensors is then determined by

$$ZMP_x = \frac{\sum_{x=1}^{8} F_x \times a_x}{\sum_{x=1}^{8} F_x}, \quad (1)$$

$$ZMP_y = \frac{\sum_{y=1}^{8} F_y \times b_y}{\sum_{y=1}^{8} F_y}. \quad (2)$$

where a_x and b_y are distances from the FSRs to the origin of ZMP(see Fig. 1) in the directions of x- and y-axis, and F_x and F_y are the corresponding forces detected by FSRs.

B. Walking gait

The walking gait is divided into two phases: the path planning and stride phases. In the path planning phase, we first determine a circular arc for the trajectory of the striding leg and then obtain joint angles by inverse kinematics. The circular arc is designed by cubic spline interpolation to find the start, vertex, and goal points [17].

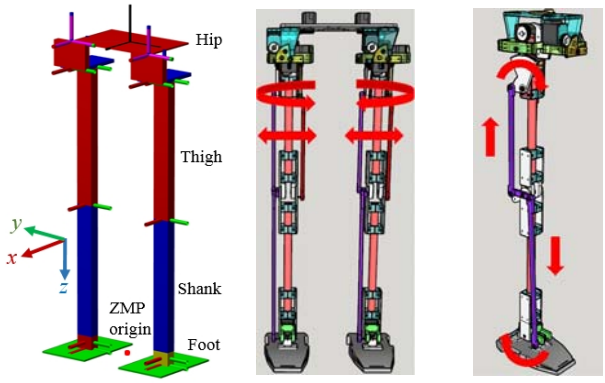

Figure 1. The DOF and mechanical structure of biped walker.

For the stride phase, a gait cycle involves eight poses (four poses on each leg) for the motions of leg lifts, striding with circular arcs, and transiting ZMP between two legs. The walker first needs to incline the trunk to the left-hand side owing to a lack of balancing by the upper body (pose 1). Next, it lifts the right leg in the direction of the y-axis (pose 2) and then swings its right leg forward after a short time (pose 3). At the end of the stride (pose 4), the walker transits ZMP from one leg to the other and then begins another half gait cycle from pose 5 to pose 8.

In our work, a semi-closed-loop control is designed for stride balance learning. That is, the pose of the inclined trunk (pose 1) can be preset while walking on a flat floor, and the motion of swinging the leg forward (pose 3) is performed in a curve trajectory. The rest of the problem is how to keep the balance for a single-leg stance at poses 2 and 6. Lacking such a capability, the walker will turn over in the direction of the y-axis (pose 2 or 6) or tumble in the phase of loading response (pose 4 or 8).

III. Policy Gradient Fuzzy Posture Learning

To incorporate PGRL in an adaptive fuzzy posture controller, we use the Mamdani with inputs ZMP error e_{ZMP} and the rate of change Δe_{ZMP} in the x- and y-axis of the reference trajectory. In each input universe of discourse, there are three triangular membership functions defined as negative (N), zero (Z), and positive (P). The triangular membership functions of each angular displacement of the servomotors as the system's output are fuzzified in terms of negative big (NB), negative small (NS), zero (Z), positive small (PS), and positive big (PB). These membership functions compose the rules as in TABLE I. With these simple rules, we can determine the output distribution by using the center of gravity method.

TABLE I. The Rules of Posture Controller

		e_{ZMPx} / e_{ZMPy}		
		N	Z	P
Δe_{ZMPx} /	N	NB	NS	Z
Δe_{ZMPy}	Z	NS	Z	PS
	P	Z	PS	PB

1.Initialize $C = \{c_1, c_2, \ldots\ldots, c_{10}\}$
2.**while** !done **do**
3. generate $(\{P_1, P_2, \ldots, P_{16}\})$
4. run $(\{P_1, P_2, \ldots, P_{16}\})$
5. get error $(\{P_1, P_2, \ldots, P_{16}\})$
6. **for** n =1 to 10 **do**
7. error $(Avg_{+\varepsilon,n}, Avg_{0,n}, Avg_{-\varepsilon,n})$
8. **if** $Avg_{0,n} < Avg_{+\varepsilon,n}$ and $Avg_{0,n} < Avg_{-\varepsilon,n}$
9. **then** $\Delta c_i \leftarrow 0$
10. **else** $Avg_{+\varepsilon,n} < Avg_{0,n}$and $Avg_{+\varepsilon,n} < Avg_{-\varepsilon,n}$
11. $\Delta c_i \leftarrow +\varepsilon$
12. **else** $Avg_{-,n} < Avg_{+\varepsilon,n}$ and $Avg_{-\varepsilon,n} < Avg_{0,n}$
13. $\Delta c_i \leftarrow -\varepsilon$
14. **end if**
15. **end for**
16. $c_i \leftarrow c_i + \Delta c_i$
17. $\varepsilon \leftarrow \varepsilon \times \alpha, \alpha < 1$
18.**end while**

Fig. 2 Pseudocode for the proposed PGRL algorithm.

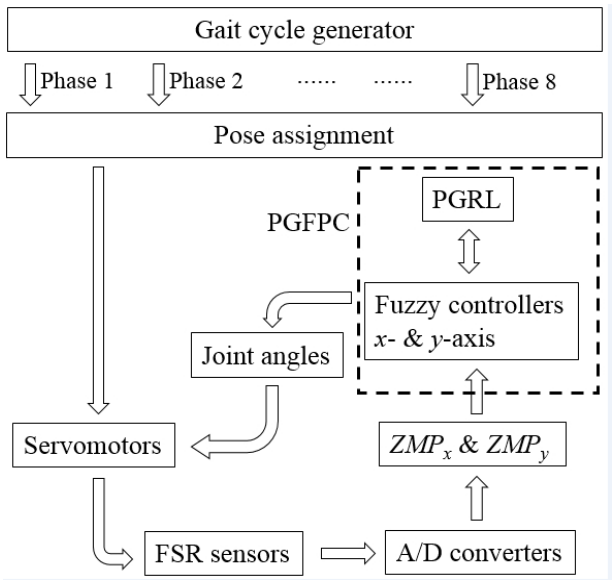

Figure 3. Block diagram of gait control by PGFPC.

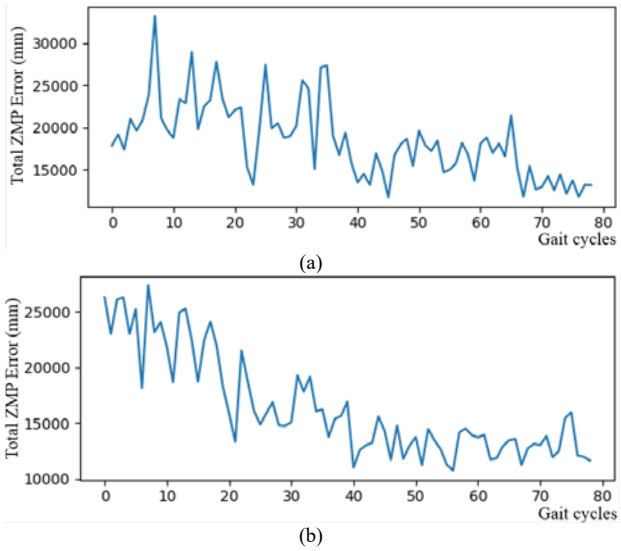

Figure 4. The cumulative ZMP errors of each gait cycle by PGFPC and step length: (a) 1 cm, (b) 9 cm.

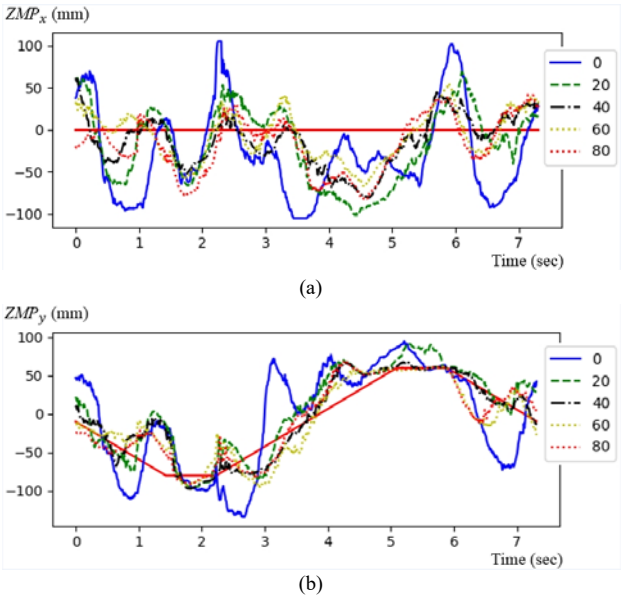

Figure 5. The improvement of ZMP errors by twenty gait cycles increments in: (a) *x*-axis, (b) *y*-axis, with 9-cm step length.

For the operation of PGRL, it has been designed to update the output membership functions. The learning process is designed with sixteen policies $P = [p_1, p_2, \ldots, p_{16}]^T$ for both legs, and every policy has to be tested by a stride. Furthermore, each policy is composed of initial parameter vector $C = [c_1+\Delta c_1, c_2+\Delta c_2, \ldots, c_{10}+\Delta c_{10}]^T$ which represents ten cores c_i from output membership functions and core translation variable Δc_i, i=1,2,..,10. When testing with a policy, the system randomly chooses Δc_i by prescribing parameters $+\varepsilon$, 0, and $-\varepsilon$ for each c_i and then collects the cumulative ZMP errors at the end of the stride for picking the best policy. Base on such a mechanism, the pseudocode of PGRL is developed as presented in Fig. 2.

A block diagram of the proposed gait control system is as shown in Fig. 3. The process allocates eight phases of a gait cycle first; each walking phase corresponds to a pose of the walking gait. The angles of joints are previously determined by inverse kinematics and fine-tuned for different poses. Next, the robot starts to walk with these poses. Poses 2 and 6 are controlled by the proposed posture learning control strategy.

IV. Results and Discussions

Experimental results are shown in Figs. 4 and 5. The effectiveness of gait learning can be observed therein, where the cumulative ZMP errors of each gait cycle were gradually improved with step increments and were convergent after the 70th gait cycle, Fig. 4(a), prompting us to terminate the learning process at the end of the 80th gait cycle.

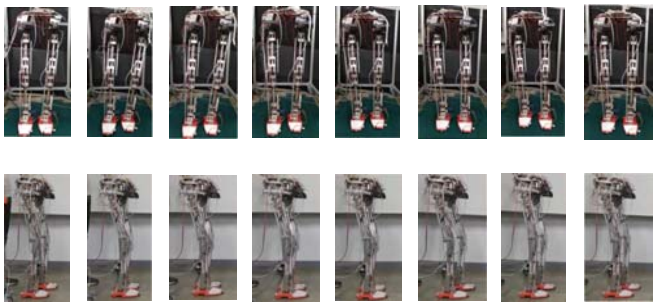

Figure 6. The walker strides forward with eight poses from the left- to the right-hand side for frontal and sagittal view.

The walker initially learned a step length of 1 cm. After output membership functions had been updated with the parameters for 1-cm step length, the walker explored a new gait by increasing the step length for 2 cm and learned with another eighty gait cycles again. Such a learning process repeated until the walker adapted to the step length of 9 cm. For the step length longer than 9 cm, the walker's foot sole is not parallel to the floor anymore. Figure 4(b) indicates that the robot learned the 9-cm walk after the 40th gait cycle and the ZMP error was averagely smaller than that of the 1-cm step length walking.

The gait balance performance of the proposed controller can also be observed in the example of Fig. 5, which illustrates the improvement of ZMP errors by eighty gait cycles and a 9-cm step length. Our test started with the 0th gait cycle (blue-solid line) by using just FLC. Although the basic FLC can track the desired ZMP trajectory, the test result in this figure exhibits some degraded tracking performance in either direction. This figure indicates the walker swung (back and forth) noticeably in the *x*-axis, and there was also some large wobble in the *y*-axis. When applying posture learning, the learning progress reached the 80th gait cycle, the walker's gait became stable. The ZMP mostly converged to the desired trajectories (red-solid line) as PGFPC significantly suppressed the worst wobble.

Finally, the snapshot of gait is depicted in Fig. 6 for the eight poses. Based on the walking posture captured, this figure indicates that the walker first inclined its trunk to the left-hand side, and then it lifted and swung its right leg forward to perform a stride. After the first stride, it leaned to the right-hand side and continued the second stride by the same gait pattern.

V. Conclusion

In this paper, an adaptive fuzzy controller using policy gradient-based reinforcement learning (PGRL) for the frontal plane balance of an adult-sized biped walker is proposed. The fuzzy-controller input is from the error of trunk ZMP detected by FSR pressure sensors on foot soles. The policy gradient-based reinforcement learning evaluates the performance of gait stability in each episode and then fine-tunes the positions of the output cores. Compared with the biped gait learning techniques, our approach is model-free and does not require heavy computational latency. The experimental results obtained demonstrate the feasibility effectiveness of the proposed approach. Our future work will involve gait control with an upper trunk design and different terrain.

Acknowledgment

Supports from the Ministry of Science and Technology, under Grant MOST108-2221-E-194-041-MY2, Advanced Institute of Manufacturing with High-Tech Innovation (AIM-HI) and Center for Innovative Research on Aging Society (CIRAS) from The Featured Areas Research Center Program within the framework of Higher Education Sprout Project by the Ministry of Education (MoE) in Taiwan are gratefully acknowledged.

References

[1] J. L. Lin, K. S. Hwang, W. C. Jiang, Y. J. Chen, "Gait balance and acceleration of a biped robot based on Q-learning," *IEEE Access*, https://doi.org.10.1109/ACCESS.2016.2570255, 2016.

[2] A, Zaidi, N. Rokbani, A. M. Alimi, "A Hierarchical fuzzy controller for a biped robot, " in *Proc. Int. Conf. Individual and Collective Behaviors in Robotics (ICBR)*, Tunisia, pp. 126-129, 2013.

[3] R. S. Sutton, G. B. Andrew,*Introduction to reinforcement learning,* 1st ed. Cambridge, MA, USA: MIT Press, 1998.

[4] K. S. Hwang, J. L. Lin, J. S. Li, "Biped balance control by reinforcement learning," *Journal of Information Science and Engineering*, vol. 32, pp. 1041-1060, 2016.

[5] X. Wu, S. Liu, T. Zhang, L. Yang, Y. Li, T. Wang, "Motion control for biped robot via DDPG-based deep reinforcement learning," in *Proc. 1st WRC Symposium. Advanced Robotics and Automation*, Beijing, China, August, pp. 40-45, 2018.

[6] Y. T. Su, K. Y. Chong, T. H. S. Li, "Design and implementation of fuzzy policy gradient gait learning method for walking pattern generation of humanoid robots," *Int. J. Fuzzy Systems*, vol. 13, no. 4, pp. 369-382, 2011.

[7] T. H. S. Li, Y. T. Su, S. W. Lai, J. J. Hu,"Walking motion generation, synthesis, and control for biped robot by using PGRL, LPI, and fuzzy logic," *IEEE Trans. Systems, Man, and Cybernetics*—Part B: Cybernetics, vol. 41, no. 3, pp. 736-748, 2011.

[8] S. M. Bruijn and J. H. V. Dieën, "Control of human gait stability through foot placement," J. R. Soc. Interface, vol. 15, 2018.

[9] Y. Mao, J. Wang, P. Jia, S. Li, Z. Qiu, L. Zhang, Z. Han," A Reinforcement learning based dynamic walking control. in *Proc, Intl. Conf. Robotics and Automation (ICRA)*, Italy, pp. 3609-3614, 2007.

[10] A. Rai, R, Antonova, S. Song, W. Martin, H. Geyer, C. Atkeson," Bayesian optimization using domain knowledge on the atrias biped," *IEEE Intl. Conf. Robotics and Automation (ICRA)*, Australia, pp. 1771-1778, 2018.

[11] J. P. Ferreira, M. M. Crisóstomo, A. P. Coimbra, "Svr versus neural-fuzzy network controllers for the sagittal balance of a biped robot," *IEEE Trans. Neural Networks*, vol. 20, no. 12, pp. 1885-1897, 2009.

[12] M. Talebi and M. Farrokhi," Adaptive fuzzy-PD controller for 3D walking of biped robots," *Intl. Conf. Artificial Intelligence and Robotics*, Iran, 2015 (7 pages).

[13] S. L. Cardenas-Maciel, O. Castillo, L. T. Aguilar. L. T, J. R. Castro," A T-S fuzzy logic controller for biped robot walking based on adaptive network fuzzy inference system," *Intl. Conf. Neural Networks (IJCNN)*, Spain, 2010 (8 pages).

[14] C. C. Wong, B. C. Huang, J. Y. Chen," Direct center adaptive fuzzy controller design. in *Proc. Intl. Conf. Fuzzy System*, USA, pp. 390-393, 1998.

[15] A. M. Zaki, M. El-Bardini, F. A. S. Soliman, M. M. Sharaf," Embedded two level direct adaptive fuzzy controller dc motor speed control. Ain Shams Engineering Journal, vol. 9, pp. 65-67, 2015.

[16] P. A. Phan, J. T. Gale," Direct adaptive fuzzy control with a self-structuring algorithm," *Fuzzy Sets and Systems*, doi: 10.1016/j.fss.2007.09.012, 2007.

[17] A. D. C. P. Filho, "Biped robots," China: IntechOpen, 2011.