Machine Learning for Biped Robot Locomotion

José Rosado

Resumo - **O presente artigo apresenta o estado da arte na aplicação de conceitos relacionados com o tema da aprendizagem automática e o seu potencial na aplicação em locomoção de robôs bípedes.Usando uma abordagem bottom-up, o artigo começa por introduzir conceitos sobre aprendizagem automática, o seu uso e o seu enquadramento na área da robótica. Seguidamente, é feito um estudo sobre métodos aplicados com sucesso na robótica. Na parte final focamo-nos na problemática da locomoção bípede em robôs humanóides e possíveis soluções usando aprendizagem (*e.g.*, aprendizagem por reforço). São apresentados alguns casos de utilização deste método e quais os resultados obtidos.**

Abstract - **This paper presents an overview of the state-of-the art methods in machine learning and the potential of application for biped robot locomotion. Adopting a bottom-up structure, the paper starts by introducing the state of machine learning, the fundamental questionsit addresses and the current research topics. Second, machine learning methods that have been successfully applied in robotic systems are briefly discussed. Finally, the paper focuses on problems of biped locomotion that can be solved by using learning(*e.g.*, reinforcement learning). Case-studies highlighting the implemented solutions help to understand how a locomotion robot system can improveits control strategy through experience.**

I. INTRODUCTION

Over the past 60 years, computers have played an important role in helping mankind in many tasks. Today, we reached a point where computers are present in our liveswithout we even notice it. Computers outperformhumans in various tasks, like for example:

- Complex calculations: they can do it faster and without mistakes.
- Repetitive tasks: humans have a trend to get bored and make mistakes in repetitive tasks.
- Unfriendly environment tasks: tasks that require working with dangerous chemicals ortasks where the environment is hostile to humans can easily be made by machines controlled by computers.

However, there are many situations where humans are far ahead in terms of their performance capabilities. For example, when something occurs in the environment that was not expected, machines donot know how to react or react very poorly. Instead, humans can react much faster in a way that can avoid disaster. This is because humans are able to do something much better that machines: learning. Through their lives,humans evolve and learn from different situations occurring in everyday life and they are able topredict the effects oftheir actionsin the environment and,inthis way,react accordingly. Inspired by biology, several researcher projects areunderway to develop machines that can learn the same way we do, endowedwithcognitive and learning abilities.

The remainder of the paper is organised as follows: Section II presents an overview of current state of machine learning. Section III explores the application of learning in several robotic domains. Section IV describes the main advances of machine learning paradigms for robot biped locomotion. Finally, Section V concludes the paper and outline the perspectives of future research.

II. STATE OF MACHINE LEARNING

The concept of machine learning refers, usually, to the changes in systems that automatically learn to recognize complex patterns, to link perception, reasoning and action processes, to make intelligent decisions, to predict situations that may encounter, etc. Machine learning can be achieved at different levels of complexity, much like different scientific fields investigate learning processes in biological systems [1-3]. In simple words, we may say that machinesandcomputer programs "learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E" [1].

In general, three learning paradigms are considered in literature: supervised, unsupervised and reinforcement learning. Supervised learning is the taks of inferring a function from a supervised set of training examples consisting of an input object and a desired output value. It has initially been successfully applied in classification and prediction tasks, but is not brain-like. Unsupervised learning is about understanding the world by mapping or clustering given data according to some principles, beingassociated with the cortex in the brain. Reinforcement learning (RL) is a powerful method to develop goal-directed action strategies where the system learns behavioural reactions controlled by reward (trial-and-error process). In fact, the mathematical model of RL reflects the brain's dopamine-based system by encoding reward aspects of environment stimuli. Just like reinforcement learning, many other mathematical models induce various forms of learning with parallels in biology.

Table I - Machine Learning Methods

Method name	Method Description
Decision tree learning	Uses a decision tree as predictive model to map observations about an item to conclusions about the item's target value.
Association rule learning	A method for discovering associations in large databases.
Artificial Neural Networks	is a mathematical or computational model that tries to simulate the structure and functional aspects of a biological neural network. It's composed by a group of interconnected artificial neurons.
Genetic programming	A evolutionary based algorithm that is inspired on the biological evolution and uses genetic algorithms.
Support vector machines (SVMs)	A set of related supervised learning methods used for classification and regression. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that predicts whether a new example falls into one category or the other.
Inductive logic program	An approach to rule learning using logic programming as a uniform representation for examples, background knowledge, and hypotheses. Given an encoding of the known background knowledge and a set of examples represented as a logical database of facts, an ILP system will derive a hypothesized logic program which entails all the positive and none of the negative examples.
Clustering	A set of observations are assigned into a subset/cluster so that observations that are similar belong to the same cluster.
Bayesian network	A probabilistic graphical model that represents a set of random variables and their conditional independencies via a directed acyclic graph (DAG). For example, a Bayesian network could represent the probabilistic relationships between diseases and symptoms. Given symptoms, the network can be used to compute the probabilities of the presence of various diseases.
Reinforcement learning	Concerned in how an agent ought to take actions in a environment in order to maximize an reward it receives. For a right decision, the agent receives an positive reward;for a negative decision the agent receives a negative reward.

Several methods and algorithms for machine learning have been developed over time. Table I lists some of the most used methods giving a brief description of each one. A more complete overview can be found elsewhere [1-3]. From the point of view of application, there are many examples where machine learning has been applied with success:

- Speech recognition: the speech recognition accuracy is greater if one trains the system, that trying to program it by hand. A learning speech recognition system will be able to adapt itself to the user, rather the opposite.
- Computer vision: many current vision systems, from face recognition systems to systems that automatically classify microscope images of cells use machine learning algorithms. Again, accuracy is much better than a fixed programmed system.
- Robot control: machine learning methods have been successfully applied in the robotic area. Machine learning can create better control methods for complex robots with dynamics that can change over time or that are nonlinear.

III. MACHINE LEARNING IN ROBOTICS

One of the areas where machine learning is most used is in robotics. In this section, various examples will be presented.

The first example comes from [4] where the authors compare the efficiency of four machine learning algorithms used to classify robotic soccer formations in the Robocup contest. The Robocup contest is divided into two main groups. The first group uses real robots with different sizes and rules, while the second group uses only computer simulations with the purpose to improve the research in AI and other aspects. The simulation contest is also divided into three groups: 2D simulation, 3D simulation and Mixed Reality. The paper compares four machine learning algorithms applied to the players of the FC Portugal in the 2D simulation league, namely, Artificial Neural Networks, Kernel Naïve Bayes, K-Nearest Neighbor and Support Vector Machine (SVM). The authors proceed to the simulation of several situations, using different databases. The obtained results are presented in TABLE II.

Table II - Accuracy And Time Of Experience

	Classifier	SVM	NN	3-NN	KNB
Data Base A	Accuracy(%)	95,77	80,99	99,78	78,0
	Time	48'21''	5h18'16''	1h7'3''	1'14''
Data Base B	Accuracy(%)	96,07	84,72	99,79	77,04
	Time	28'03''	3h25'17''	9'26''	1'13''
Data Base C	Accuray(%)	10,25	46,69	95,91	50,21
	Time	1h52'23	51'49''	3'54''	11''

From this results, the K-Nearest Neighbor presents the best results. However, in different situations of games and strategies, the best was SVM followed by K-NN, as the results in TABLE IIIshow.

Table III - Accuracy and Time

Classifier	SVM	NN	3-NN	KNB
Accuracy(%)	51,65	45,77	46,67	26,47
Time	4'58''	21'11''	3'55''	29''

Another interesting example is given in [5]. In this article the author analyses the applicability of several ML methods in various robots discovery tasks, in the light of the experience on the XPERO project (www.xpero.org). As the author says, the scientific goal of XPERO is to investigate the mechanisms of autonomous discovery through experiments in an agent's environment. In XPERO, the experimental domain is the robot's physical world, and the subject of discovery are various quantitative or qualitative laws in this world. Fig. 1 shows the experiment and the parameters of the experiment made. The experiment consisted of a mobile robot moving in a plane with a simple object (a red block or ball). The robot is equipped with a stereo vision which enables it to detect the area in the image belonging to the object, the distance to the object and the angle between the current orientation of the robot and the angle at which the object was observed. Further, the robot is aware of its actions, that is, ismoves, expressed as the step distance in the forward direction and the step angle, that is the change of the robot current orientation. Several machine learning algorithms were used, such as:

- Rule Learning using CN2;
- Induction of classification and regression trees;
- Equation Induction with GoldHorn;
- Learning qualitative models with QUIN;
- Learning qualitative models with Padé;

The most interesting results were obtained with the learning of qualitative models, both with QUIN and Padé.

In [6], Ryo Saegusa*et al.* propose a method to speed up the sensorimotor learning and coordination of autonomous robots. In a complex autonomous robotic system, motor-babbling-based sensorimotor learning is considered an effective method to develop an internal model of the self-body and the environment. However, this process requires much time for computation and exploration.The authors propose a model characterized by a function they call "confidence" that is a measure of the reliability of the state control. Fig. 2 shows the proposed learning system used to predict and control the state at the next time. To better understand the variables, TABLE IV explains their meaning.

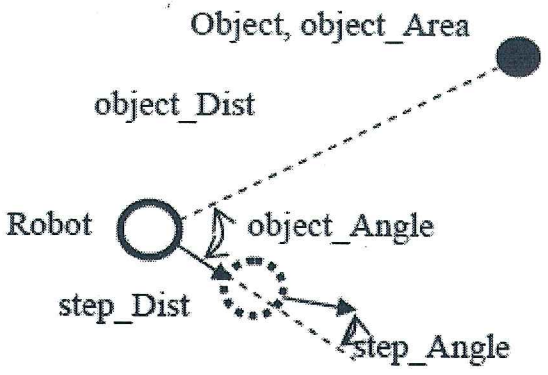

Fig. 1 - XPERO simple experiment[5]

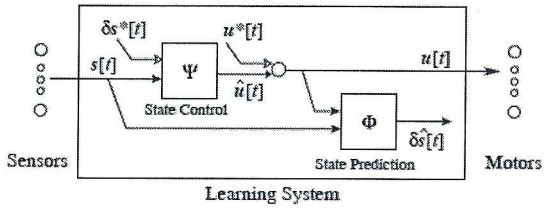

Fig. 2 - Proposed learning system[6]

Table IV - Variable Notation

notation	variable
s	measured sensory input
δs*	desired sensory input change
u*	desired motor control
û	estimated motor control
u	actuated motor control
$\delta\hat{s}$	estimated sensory input change

The "confidence" is based on the state control error given by the following equation:

$$e_u[t] = |\Psi(s[t], \delta s[t]) - u[t]| (1)$$

The learning procedure is divided into 2 stages called: exploration and learning (Fig. 3). In the exploration stage, the robot will generate joint movements in order to collect learning samples and evaluates mapping functions optimized in previous learning stages. In the learning stage, the robot will optimize the mapping functions off-line with the collected learning samples in the previous stages. Motor behavior of the robot in the exploration stage is generated using the "confidence" value as a probability to chose the motor command. The learning system is composed by a Multiple Layer Perceptron (neural network).

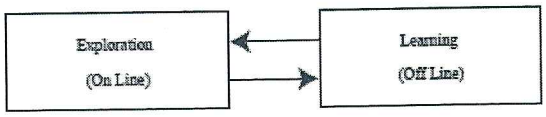

Fig. 3 - Learning strategy proposed by Ryo Saegusa*et al.*[6]

Experimental results were made with an humanoid robot called James. This is a fixed upper-body robotic platform dedicated to vision-based manipulation studies. It is composed by a 7 degree-of-freedom (DOF) arm with a dexterous 9-DOF hand and a 7-DOF head (Fig. 4).

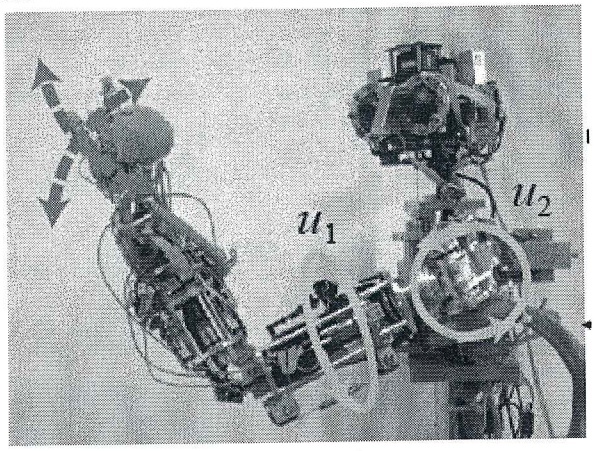

Fig. 4 - James robot[6]

Both active and passive sensorimotor learning were performed for comparison. Fig. 5shows the evolution of the state space confidence quantized as 8x8 pixel regions. Light intensity in each region indicates the local confidence value. From left to right the columns correspond to the confidence maps of state prediction in active learning, state control in active learning, state prediction in passive mode and state control in passive mode. From top to bottom, the number of cycles (0, 5, 10, 15 and 20). From the picture is evident that the active learning is faster.

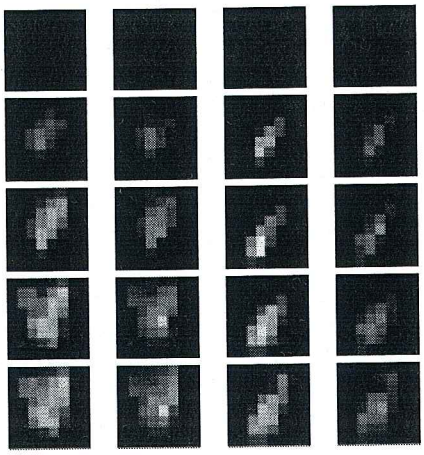

Fig. 5 - Temporal evolution of the space state confidence[6]

IV. BIPED LOCOMOTION AND MACHINE LEARNING

Biological systems are the result of an evolutionary process related with competition, survival and learning inside a specific environment. Biped locomotion presents several key problems, like for example[7]:

- Non linear dynamics
- Multi variable dynamics
- Unstable nature of the dynamics
- Limited foot/ground interaction
- Discrete changes on the dynamics

It is the first three that make much more difficult to implement a controller of a robotic biped system using classical control theory. The mathematical model of the system is very complex and it is described by nonlinear high order differential equations. Several strategies can be used to solve these potential problems, such asto simplify the dynamical model, to ignore the effects of friction and flexibility, and to minimize the impacts with the ground. The interaction between the foot and the ground is one of the key aspects in legged robots that distinguish them from manipulators.

The degree of freedom established between the foot and the ground is unilateral and, at the same time, the moment applied around the foot must be limited to avoid the complete rotation around the heel or toes. Another characteristic is the dynamic change that occurs along the walking cycle: during this process the system is supported by one foot or by both, meaning there is a change on the system's dynamics. This is an advantages that allow biped robots to walk in environments not accessible to wheel-based mobile platforms, such as climbing stairs. However, the problems mentioned before contribute to make difficult the development of a simple and robust control system for biped robots.

Being biped robots so hard to control, due to the complexity of the system dynamics and non linearity of the system, one solution is to introduce machine learning methods. These methods do not simply may allow to control a biped robot locomotion, but can also be used to make the robot perform other tasks or walk in untrained environments. In fact, for a robot to perform in natural real-world environments, it must be able to adapt its behaviour autonomously to unexpected challenges, something too complex to be formulated or programmed explicitly.

One successful method applied to biped locomotion is reinforcement learning. Salatian*et al.* [8] use reinforcement learning together with a neural network mechanism to modify the gait of a biped robot that must walk on a sloping surface, without prior knowledge of its inclination.

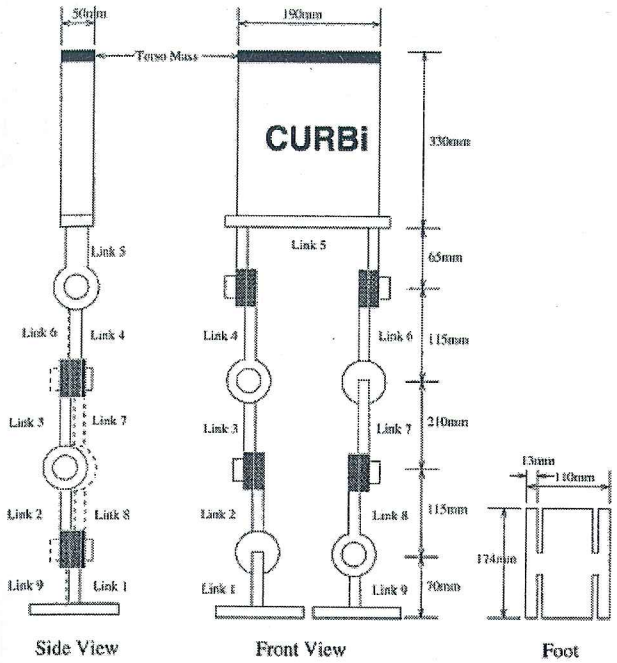

Fig. 6 - The SD-2 robot structure[8]

Fig. 6 shows the structure of the robot used, called SD-2. The robot has a total of nine links and eight joints, that allow 4 DOF by leg. This robot has a particular curiosity, that is the fact it has no knees. A pre-defined statically stable gait for this robot is shown inFig. 7. Each step is divide into 8 static configurations, called primitive points(PP), and each PP is decomposed into a large number of setpoints, with duration of 28ms, being the total step duration 2000ms. The robot has 2 force sensors in eachfoot(one in the toes, other in the heal), that allow to compute the Center Of Gravity (COG). The dotted squares in Fig. 7.represent the swinging foot and the big dots are the projection of the COG.

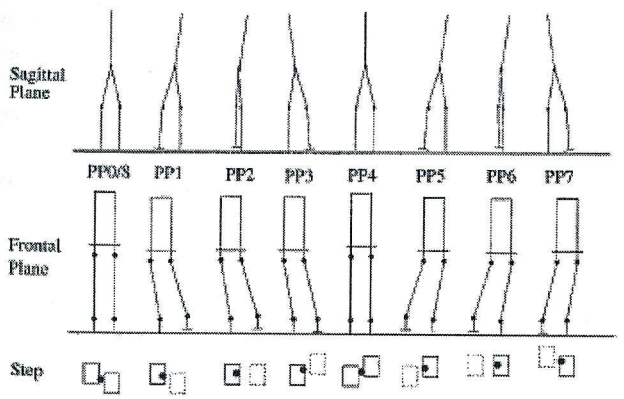

Fig. 7 - Gait on a level surface for the SD-2[8]

The system is controlled by a neural controller represented in Fig. 8. It's composed by a memory that stores previous learned gaits, an adaptive unit (AU) responsible for modifying the joint trajectories and a sensor unit. The AU is composed by a set of 4 neurons for each joint, giving a total of 24 neurons (both top hip joints are controlled by the same signal).

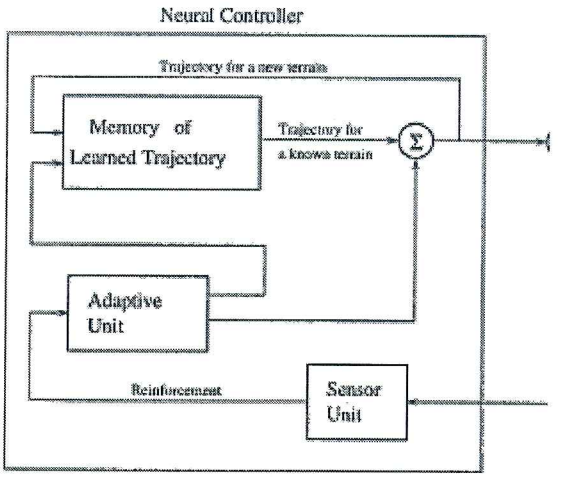

Fig. 8 - Neural Controller for the SD-2[8]

The difference between the forces exerted at the toe and the heel generate the reinforcement signal that trains the neural network (Eq. (2)). The robot is stable, as long as $\Delta f = \Delta f_{bal}$, where Δf_{bal} is the ideal force balance, obtained by recording it when the robot walks on a level surface and the gait is optimal.

$$\Delta f = f_{heel} - f_{toe} \quad (2)$$

The authors conducted several experiments and trainings with several unknown slopes, where they prove that is possible for a biped robot to walk adaptively on unknown terrains using the neural network approach with an unsupervised reinforcement learning.

Masa-aki Sato, Yutaka Nakamura and Shin Ishii propose a Reinforcement Learning method for Central Pattern Generators (CPG) in [9]. Neurobiological studies have revealed that rhythmic motor patterns are controlled by neural oscillators referred to as CPG [10]. The main purpose of their paper is to study the use of reinforcement learning for a CPG controller that generates stable rhythmic movements. However, RL for biped locomotion is very difficult, because the biped robot is very unstable and the system has continuous state and action spaces with a high degree of freedom. Standard RL methods, such as temporal difference learning, Q-leaning and actor critic methods are not suited for training the CPG and in order to deal with this, the authors propose a new RL method which they call the CPG-actor-critic method. Their method consists in dividing the CPG into 2 modules, *i.e.*, the basic CPG and the actor (Fig. 9). The method was applied to a robot with a structure illustrated in Fig. 10 (top) and after about 5800 trials, the robot started to walk Fig. 10 (bottom). Although the method workedproperly, the learning process was rather unstable and it was necessary to fine tune the weights of the mutual connections among the CPG neurons that compose the CPG controller.

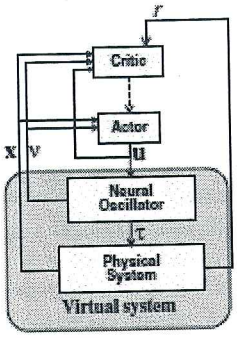

Fig. 9 - The CPG-actor-critic-method[9]

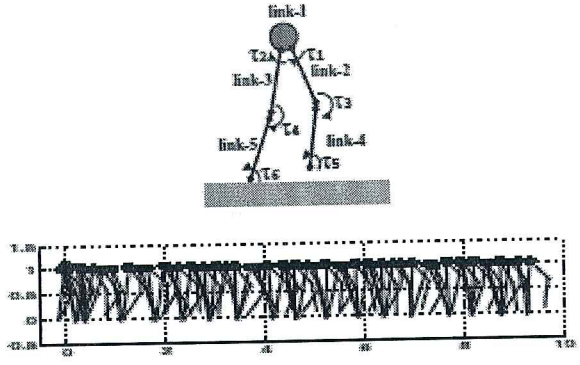

Fig. 10 - Structure of the robot (top) and results after 7000 episodes (bottom)[9]

Later, Nakamura *et al.* [11] reinforce the idea of the problems encountered in their previous article [9]. They also assume that although the use of RL with the called SARSAalgorithm[12] can be successful applied to various Markov Decision Process (MDP), with finite state and action spaces, they suffer from the "curse of dimensionality", called like this because when the number of state and action spaces increases, it becomes very difficult to use this method. So, they considered the use of another method, called policy gradient RL. In this method, the objective of the RL is to obtain the policy parameter that maximizes the expected reward accumulation defined by $\rho(\theta) \equiv E_\theta[\sum_t \gamma^{t-1} r(s(t),u(t))]$, where $\gamma \in [0,1]$ is a discount factor. The partial differential of $\rho(\theta)$ with respect to the policy parameter θ_i is calculated by:

$$\frac{\partial \rho(\theta)}{\partial \theta_i} = \int_{s,u} dsduD_\theta(s,u)\psi_i(s,u)Q_\theta(s,u) \quad (3)$$

where $\psi_i(s,u) \equiv \frac{\partial}{\partial \theta_i} ln\pi_\theta(u|s)$ and $Q_\theta(s,u)$ denotes the action-value function (Q-function)[12].

To estimate the policy gradient (3), a linear approximator of the Q-function: $f_\theta^w(s,u) = \sum_i \psi_i(s,u)w_i$, where **w** is the parameter vector, is used. If $\tilde{w} = \arg\min_w \left(Q_\theta(s,u) - f_\theta^w(s,u)\right)^2$, using $f_\theta^w(s,u)$ instead of the true Q-function is achieved then $Q_\theta(s,u)$ does not introduce any bias to the calculation of the policy gradient[12] and the parameter $\tilde{w}$ provides the natural policy gradient.

The authors then conducted 2 experiments. The first one had as objective to see if their method was able to obtain a CPG controller that could make the robot walk stably. For, that weight parameters of the neural network were set to random values, in which the robot could not walk (Fig. 11). The learning curve is show in Fig. 12. After about 7000 trials, the robot was less likely to fall down and Fig. 13 shows the gait pattern of the biped robot after 10000 learning trials.

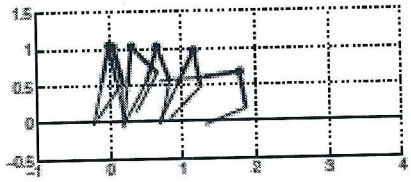

Fig. 11 - Before learning[11]

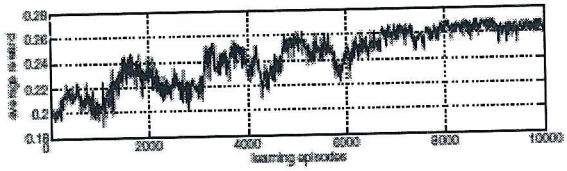

Fig. 12 - Learning curve[11]

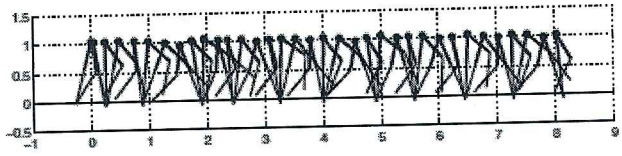

Fig. 13 - After learning[11]

In the second experiment, the CPG controller was tested to see if it was able to acquire a policy that produced a stable walk in various ground conditions. In this case, the weight parameters were fine hand-tuned. Several simulations were done, in which the ground surface was set to be piece-wise linear and the gradient of each linear piece was set randomly within a specific result. The results of the learning are shown in Fig. 14. After about 2000 learning episodes, a good control was acquired. Although the simulation results showed that the robot was able to walk stable, some factors introduced bias to the estimator $\tilde{w}$.

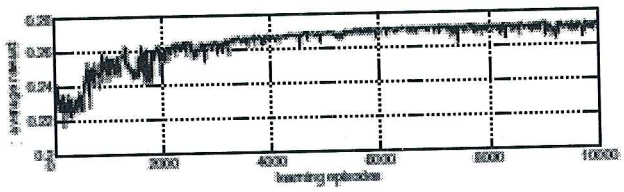

Fig. 14 - Learning curve for rough ground[11]

Jun Morimoto e*t al.* also present the use of reinforcement learning together with a Poincare map in [13] and [14].The articles use a five link robot (Fig. 15) in the

simulations, based on a real robot with the specifications of Table V.

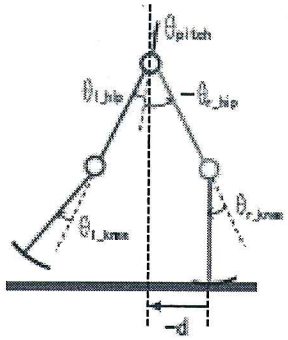

Fig. 15 - Five link biped robot[13-14]

Table V - Physical Parameters of the five link robot model

	trunk	thigh	shin
mass (kg)	2.0	0.64	0.15
length (m)	0.01	0.2	0.2

The authors divided the learning process in five stages:

1) A model that predicts the state of the biped a half cycle ahead based on the current state and the foot placement at touchdown. The model predicts the state at Poincaré section in phase $\phi = \frac{3\pi}{2}$(Fig. 16)based on the system's location at $\phi = \frac{\pi}{2}$. The same model is used to predict the location at state $\phi = \frac{\pi}{2}$ using the location at phase $\phi = \frac{3\pi}{2}$. This is done, because the state of the robot drastically changes at the foot touchdown. The Poincaré map is approximated by (4), with a parameter vector w^m,

$$\hat{x}_{\frac{3\pi}{2}} = \hat{f}\left(x_{\frac{\pi}{2}}, u_{\frac{\pi}{2}}; w^m\right)(4)$$

Where the input state is denoted as $x = (d, \dot{d})$, where d denotes the horizontal distance between the stance foot position and the body position (Fig. 17, left). The action of the robot $u = \theta_{act}$is the target knee joint angle of the swinging leg (Fig. 17).

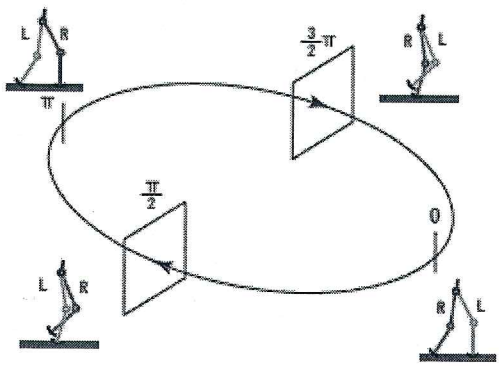

Fig. 16 - Biped walking cycle[13-14]

2) Representation of biped walking trajectories: One cycle of the biped walking is represented by 4 via points for each joint (Fig. 16). Zero desired velocity and acceleration are specified at each via point.

3) Reward: the robot gets a reward if it successfully continues on walking and gets punishment if it falls down.

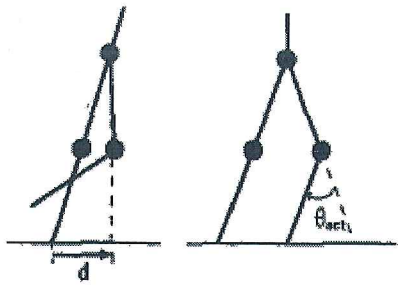

Fig. 17 - Input stance and output of the controller[13-14]

4) Learning the value function: in reinforcement learning, the learner tries to create a controller which maximizes the expected total return. The authors denoted the value function for the policy μ

$V^{\mu}(x(t)) = E[r(t+1) + \gamma r(t+2) + \gamma^2 r(t+3) + \ldots](5)$

where$r(t)$ is the reward at time t and $\gamma(0 \leq \gamma \leq 1)$ is the discount factor.

5) Learning policy for biped locomotion: A stochastic policy is used:

$$\mu(u(t)|x(t)) = \frac{1}{\sqrt{2\pi\sigma}} e^{\left(-\frac{(u(t)-A(x(t);w^a))^2}{2\sigma^2}\right)}(6)$$

where$A(x(t); w^a)$ denotes the mean of the model, represented by a function aproximator using w^a as a parameter vector. The variance σ is changed according to the trial as $\sigma = \frac{100 - N_{trial}}{100} + 0.1$ for $N_{trial} \leq 100$ and $\sigma = 0.1$ for $N_{trial} > 100$.

The proposed method was applied to the 5 link simulated robot with the results of Fig. 18.

In several articles presented later, where Morimoto was a co-author [15-17] the subject of policy gradient method combined with CPGs is brought again. This time, they go beyond simulation and the real robot that served as model in Fig. 15 is used with some interesting results.Fig. 19shows a photo of the real robot and Fig. 20 shows the learning system used.

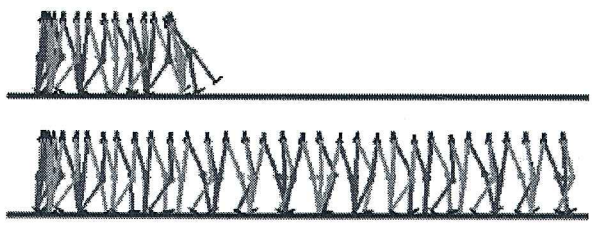

Fig. 18 - Results of the simulation: top - before learning; botom: after learning.[13-14]

It is interesting to note that the CPG controller only controls the hip joints and the knee joints are controlled by

a state machine. This process allows the simplification of the RL process, since the number of states and action spaces remains lower.

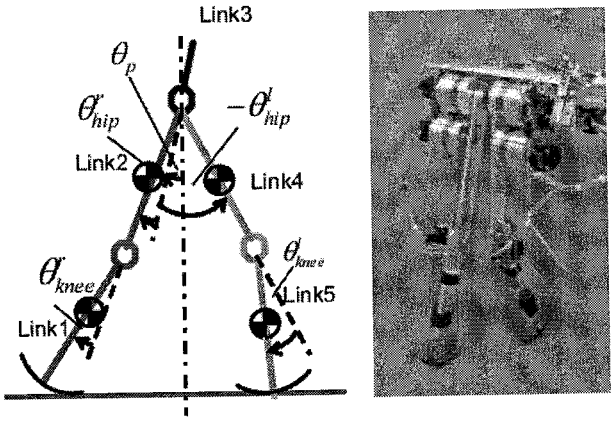

Fig. 19 - Five link biped robot[15-17]

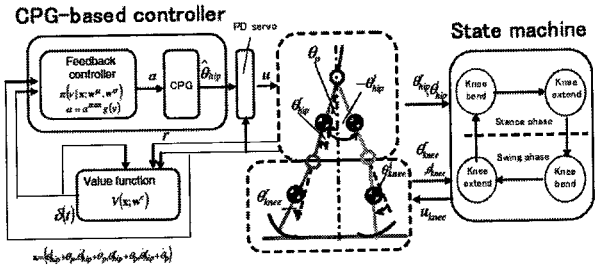

Fig. 20 - Learning system for the five link biped robot[15-17]

V. CONCLUSIONS

Machine learning has been a subject of study over the last 50 years and has been applied in many fields. Robotics are one of this fields, and in many areas it has been very successful. Today, many high risk or repetitive tasks are made by robots. Even on fun areas, robots are each day more present: robots that play football against each other(Robocup) and kids toys. One researcharea gaining increased interest is humanoid robotics. Biped locomotion robots are very interesting, because sooner or later they will be used to replace or cooperate with humans in various tasks, with improved performance. Several techniques have been applied with some successful to biped locomotion, being reinforcement learning the most promising. However, there is still much problems to solve since even the walking with success has been reached in controlled environments and limited conditions: walk a previous pre-planned path, follow a line or walk in the direction of an object. These means that we are still far from a biped robot that can walk among us like other humans do.Another task still hard to implement in robots is fine manipulation involving interaction with the environment, such as touch or grasp. At the same time, tasks like grabbing and pushing an object on a table, like for example a cup with water, without breaking it or spilling the water are not yet implemented.

ACKNOWLEDGMENTS

The author is in debt to Prof. Filipe Silva for all is help and ideas to build this article.

REFERENCES

[1] Tom M. Mitchel, "Machine Learning", p. 2, Mc-Graw Hill, 1997, ISBN 978-0071154673.

[2] Nils J. Nilsson, "Introduction to Machine Learning – an early draft proposed textbook", 1996.

[3] http://en.wikipedia.org/wiki/Machine_learning.

[4] BrígidaMónicaFaria,Luís Paulo Reis,Nuno Lau, Gladys Castillo, "Machine Learning Algorithms applied to the Classification of Robotic Soccer Formations and Opponent Teams, 2010 IEEE Conference on Cybernetics and Intelligent Systems, pp. 344-349, Singapore.

[5] Ivan Bratko; "An Assessment of Machine Learning Methods for Robotic Discovery; Proceedings of the 30th International Conference on Information Technology Interfaces, pp. 53-60, Croatia, 2008.

[6] Ryo Saesuga, Giorgio Metta, GiulioSandini, "Active Learning for Sensorimotor Coordinations of Autonomous Robots", 2nd Conference in Human Systems Interactions, pp. 701-706, Catania, Italy, 2009.

[7] Filipe Teixeira, "Análise Dinâmica e Controlo de Sistemas Robóticos de Locomoção Bípede", phD. Thesis, University of Porto, July 2001.

[8] A. W. Salatian, Keon Young Yi, Yuan F. Zheng, "Reinforcement Learning for a Biped Robot to Climb Sloping Surfaces", Journal of Robotic Systems, Volume 14, issue 4, April 1997, pp. 283-296.

[9] Masa-aki Sato, Yutaka Nakamura and Shin Ishii, "Reinforcement Learning for Biped Locomotion", Lecture Notes in Computer Science Artificial Neural Networks ICANN, 2002, pp. 777-782.

[10] Grillner, S., Wallen, P., Brodin, L., Lansner, A., "Neural Network Generating LocomotorBehavior in Lamprey: Circuitry, Transmitters, Membrane Properties and Simulations.", Annual Review of Neuroscience 14 (1991), pp. 169-199.

[11] Yutaka Nakamura, Takeshi Mori and Shin Ishii, "Natural Policy Gradient Reinforcement Learning for a CPG Control of a Biped Robot", Lecture Notes in Computer Science Parallel Problem Solving from Nature - PPSN VIII, 2004, pp. 972-981.

[12] Sutton, R.S., Barto, A.G, "Reinforcement Learning: An Introduction", MIT Press, 1998, ISBN: 978-0262193986.

[13] Jun Morimoto, Gordon Cheng, Christopher G. Atkeson and Garth Zeglin, "A Simple Reinforcement Learning Algorithm for Biped Walking", proceedings of the 2004 IEEE International Conference of Robotics & Automation, New Orleans, pp. 3030-3035.

[14] Jun Morimoto, Jun Nakanishi, Gen Endo, Gordon Cheng, Christopher Atkeson and Garth Zeglin, "Poincaré-Map-Based Reinforcement Learning for Biped Walking", Procedings of the 2005 IEEE International Conference on Robotics and Automation, Barcelona, Spain, pp. 2381-2386.

[15] Takamitsu Matsubara, Jun Morimoto, Jun Nakanishi, Masa-aki Sato and Kenji Doya, "Learning Sensory Feedback to CPG with Policy Gradient for Biped Locomotion", Proceedings of the 2005 IEEE International Conference on Robotics and Automation, Barcelona, Spain, pp. 4164-4169.

[16] Takamitsu Matsubara, Jun Morimoto, Jun Nakanishi, Masa-aki Sato and Kenji Doya, "Learning CPG-based Biped Locomotion with a Policy Gradient Method", Proceedings of the 2005 5th IEEE-RAS International Conference on Humanoid Robots, pp. 208-213

[17] Takamitsu Matsubara, Jun Morimoto, Jun Nakanishi, Masa-aki Sato and Kenji Doya, "Learning CPG-based Biped Locomotion with a Policy Gradient Method", Robotics & Autonomous Systems Dec. 2006, Vol. 54 Issue 12, pp. 982-988.