Robotics and Autonomous Systems 58 (2010) 1266–1272

Contents lists available at ScienceDirect

Robotics and Autonomous Systems

journal homepage: www.elsevier.com/locate/robot

Analysis and solution of a predator–protector–prey multi-robot system by a high-level reinforcement learning architecture and the adaptive systems theory

José Antonio Martín H. [a,*], Javier de Lope [b], Darío Maravall [c]

[a] Dep. Sistemas Informáticos y Computación, Universidad Complutense de Madrid, Spain
[b] Dept. Applied Intelligent Systems, Universidad Politécnica de Madrid, Spain
[c] Dept. Artificial Intelligence, Universidad Politécnica de Madrid, Spain

ARTICLE INFO

Article history:
Available online 6 September 2010

Keywords:
Multi-robot systems
Goal coordination
Reinforcement learning
Adaptation
Autonomous robot navigation

ABSTRACT

The area of competitive robotic systems usually leads to highly complicated strategies that must be achieved by complex learning architectures since analytic solutions are unpractical or completely unfeasible. In this work we design an experiment in order to study and validate a model about the complex phenomena of adaptation. In particular, we study a reinforcement learning problem that comprises a complex predator–protector–prey system composed by three different robots: a pure bio-mimetic reactive (in Brook's sense, i.e. without reasoning and representation) predator-like robot, a protector-like robot with reinforcement learning capabilities and a pure bio-mimetic reactive prey-like robot. From the high-level point of view, we are interested in studying whether the Law of Adaptation is useful enough to model and explain the whole learning process occurring in this multi-robot system. From a low-level point of view, our interest is in the design of a learning system capable of solving such a complex competitive predator–protector–prey system optimally. We show how this learning problem can be addressed and solved effectively by means of a reinforcement learning setup that uses abstract actions to select a goal or target towards which a pure bio-mimetic reactive robot must navigate. The experimental results clearly show how the Law of Adaptation fits this complex learning system and that the proposed Reinforcement Learning setup is able to find an optimal policy to control the defender robot in its role of protecting the prey against the predator robot.

1. Introduction

The area of competitive robotic systems, such as models of predator–prey, usually leads to highly complicated strategies that must be achieved by complex learning architectures since analytic solutions are unpractical or completely unfeasible.

The closed loop control paradigm defines a control scheme that could be explained by means of the interaction between just two elements: the environment and the control system. The objective in this control paradigm is to maintain or guide the environment to a desired state by means of the control actions emitted by the control system. The interaction between these two components is represented by the circular flow of information between the environment's state and the control actions emitted by the control system.

The classic model of an adaptive homeostatic system is expressed by a classical equation that follows the principle of negative feedback:

$$x' = -\mu \frac{\partial J}{\partial x}, \tag{1}$$

where x is the control action, J is the objective to be minimized and the parameter μ modulates the amplitude of the system's response or control action x. Along this line, we have previously proposed a framework [1,2], from which the term bio-mimetic robot navigation or reactive utilitarian navigation gets its meaning in this paper, that formalizes this methodology as an effective tool to solve complex robot navigation problems.

However, under this control architecture, all the system's performance depends on precise, constant and immediate information received through the negative feedback cycle and cannot operate properly when such information is delayed, noisy, non-constant and, especially, when the final result of its behavior is only known after an unpredictable (maybe long) period of time (i.e. when a relevant event shows that the system's performance has improved or not) which is the case of delayed rewards [3,4] in sequential decision problems. This fact is an inherent limitation

* Corresponding address: Facultad de Informática, Universidad Complutense de Madrid, C. Prof. José García Santesmases, s/n., 28040, Madrid, Spain. Tel.: +34 91 394 764; fax: +34 91394 7510.
E-mail addresses: jamartinh@fdi.ucm.es (J.A. Martín H.), javier.delope@upm.es (J. de Lope), dmaravall@fi.upm.es (D. Maravall).

0921-8890/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.robot.2010.08.005

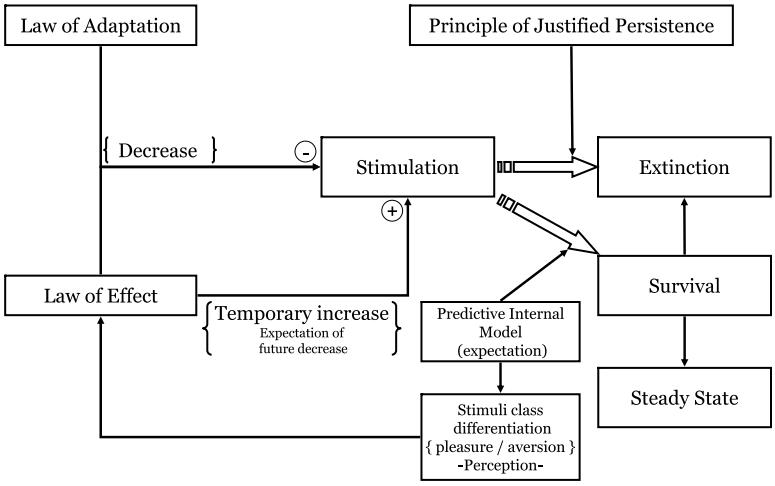

Fig. 1. Scheme of the principles acting over an anticipatory system [6].

of this kind of control architectures mainly guided by the *adaptation* phenomenon. This limitation could be reduced or indeed completely overcome when the system is able to *anticipate* the consequences of its actions and, in some sense, predict the future. In that case the system would have ensured the evaluation of the consequence of its actions at each instant and thus it could close the Wienner's [5] loop control paradigm, i.e. the homeostatic control loop.

In order to address such kind of problems we have presented previously [6] a model about the phenomena of Adaptation, Anticipation and Rationality as well as a series of fundamental principles and hypothesis. The main contributions of this model are the "*Law of Adaptation*" that states that "*every adaptive system converges to a state in which all kind of stimulation ceases*" and the "*Principle of the Justified Persistence*" that states that "*if an organism or system exists in a specific state of its environment, then, a priori, the maximum probability for surviving (avoiding the extinction) is obtained when the environment's conditions are constant or the change in the environment is highly smooth*" where proposed as a way to explain some complex adaptive phenomena.

A formal definition of the Law of Adaptation is as follows:

Given a system S, we say that a physical event E is a stimulus for the system S if and only if the probability $P(S \to S'|E)$ that the system suffers a change or is perturbed (in its elements or in its processes) when the event E occurs is strictly greater than the prior probability that S suffers a change independently of E:

$$P(S \to S'|E) > P(S \to S'). \tag{2}$$

Let S be an arbitrary system subject to changes in time t and let E be an arbitrary event that is a stimulus for the system S: we say that S is an adaptive system **if and only if** when t tends to infinity ($t \to \infty$) the probability that the system S changes its behavior ($S \to S'$) in a time step t_0 given the event E is equal to the probability that the system changes its behavior independently of the occurrence of the event E. In mathematical terms:

$$P_{t_0}(S \to S'|E) > P_{t_0}(S \to S') > 0 \tag{3}$$

$$\lim_{t\to\infty} P_t(S \to S'|E) = P_t(S \to S'). \tag{4}$$

Thus, for each instant t will exist a temporal interval h such that:

$$P_{t+h}(S \to S'|E) - P_{t+h}(S \to S') < P_t(S \to S'|E) - P_t(S \to S'). \tag{5}$$

One of the main hypotheses of the referred work was that it is possible to control the behavior of an adaptive system by means of only *external* control of its stimulation without having to deal with its internal structure. To overcome the inherent limitations of pure adaptive (reactive) systems, a framework to describe the behavior of anticipatory systems was developed [6] introducing a scheme (Fig. 1) of the principles acting over an anticipatory system.

Another concept that is evaluated under the presented experimental setup is the concept of *Goal Oriented Active Behavior*. As is known from the Information Theory concepts [7,8], the measurement of the information quantity (self-information) contained in a message can be interpreted as a measurement of how *unexpected*, *unpredictable* and in some sense *original* the contents of a message are. In this sense, if the *message* (m) is interpreted as the *action* (a) (behavior) taken by a system, then we can measure the amount of information in a, that is, how unexpected, unpredictable or original is the action/behavior of that system. This should lead us to a measurement of creativity, intelligence and finally a measure of active behavior as a term opposed to inert behavior observed in common "natural phenomena".

We have also previously defined a measurement [6] about the goal oriented active behavior of a system as a natural extension of such idea that measures the adaptive/creative capabilities of an agent's behavior:

The Quantity of Active Behavior of a system, to fulfil a goal (J) in a given situation s, is: the amount of information in its actions weighted by the effectiveness of its goal oriented actions, that is: how unexpected, unpredictable and original is its behavior, weighted by the degree of attainment of its goals, e.g. in maximizing the expected value of the cumulative sum of a reward signal—the reinforcement learning hypothesis [9].

Eq. (6) shows the proposed formulation of the Quantity of Active Behavior:

$$\chi(s, a) = -\log p(a) \times J(s, a) \tag{6}$$

where $p(a) = Pr(\mathcal{A} = a)$ is the probability that action a be chosen from all the possible choices in the action space $\mathcal{A}$ and $J(a)$ is the effectiveness; a positive performance measure to be maximized.

In this work we study a reinforcement learning problem that comprises a complex predator–protector–prey system composed by three different robots: a pure bio-mimetic reactive predator-like robot, a protector-like robot with reinforcement learning capabilities and a pure bio-mimetic reactive prey-like robot. From the high-level point of view, we are interested in studying whether the Law of Adaptation is useful enough to model and explain the whole learning process occurring in this multi-robot system. From the low-level point of view, our interest is the design of a learning system capable of solving such a complex competitive predator–protector–prey system optimally.

We show how this learning problem can be addressed and solved effectively by means of a high-level Reinforcement Learning