A bio-inspired solution for a local autonomous, reflex, obstacle avoiding behavior

Dobrea Monica-Claudia, Dobrea Dan Marius
Faculty of Electronics, Telecommunications and Information Technology, “Gh. Asachi” Technical University, Iasi, Romania
serbanm@etti.tuiasi.ro, mdobrea@etti.tuiasi.ro

***Abstract*—The main goal of our research consists in finding a simple, straightforward online solution for obstacle avoiding problem encountered in mobile robots. The solution allows the robot to develop a local autonomous obstacle avoiding behavior every time when a higher-level motor command that is driving the robot (e.g. go forward/backward) put it in imminent danger to collide. The solution we proposed for a robot with 36 evenly distributed infrared (IR) sensors is a very simple one, based only on a minimal artificial neural network (ANN) trained with a backpropagation-like algorithm. Computationally cheap, the on-line learning algorithm we implemented proved to be very successful in both, static and dynamic clustered environment. The results reported here were obtained in MobotSim 1.0.03 – a configurable 2D simulator of differential drive mobile robots.**

I. Introduction

For mobile robots a fundamental required ability is their local autonomous capacity to avoid obstacles. Until now, lots of methods [1], [2] were proposed to implement this behavior. While part of them was considered in the context of a simple wandering behavior, the others of them were implemented and discussed within some larger, goal-directed frameworks like those imposing either to reach a target, to plan the path or to follow the walls, a moving target or a line on a floor.

In what follows the local obstacle avoidance issue is addressed as part of a larger project whose final aim is to build an intelligent wheelchair mentally guided by impaired people. The intelligent wheelchair will assist the users in navigation such that every time, as soon as the users will give a command (e.g. go ahead), it will execute the command but also will take care to avoid/circumvent properly and autonomously the obstacles lying on the path.

Unlike most of the existing approaches our approach takes into consideration not the obstacle avoiding planning strategies but the defense reflex behavior that, in humans, is triggered when a sudden, unexpected, environmental change appears in the so-called peripersonal space[1] (PPS) [3]. The key role of this fast-adapting mechanism, implemented at the lowest motor execution level, consists in placing with priority the robot out of any danger of colliding with both, static and dynamic obstacles. According to a bottom-up reactive architecture, the decision of the lowest obstacle-avoidance level is further taken into account by higher motor control levels with increasingly abstract goals. For us, the next higher abstract goal is circumventing the encountered obstacle and re-gaining the movement direction and the overall goal consists in moving in a particular given direction. In this paper only the lowest behavior integrated with the highest, overall goal will be considered (particularly, by looking for an efficient bio-inspired solution), following that in the future the obstacle circumventing problem to be also addressed.

II. Some Biological and technical considerations

A. Biological considerations

In order to design a bio-inspired reflex obstacle-avoiding behavior some biological and psychological evidences were reviewed. Among them, the key elements exploited in our robotic implementation are summarized as follows.

1) Most of the central nervous system’s (CNS) cells are formed before birth, but most of the connections among cells are made during infancy and early childhood. The way these connections are made is *essentially shaped* (*a*) *through constant child interaction with the environment* and, more, (*b*) *by child’s chronologically acquired experiences* (i.e., early experience and interaction with the environment are most critical in a child's CNS development).

2) Reflexes *– as the main protective, motor involuntary motor responses to sensory inputs – may be (1)* inborn *or (2)* learned behaviors *that allow the automatic performance of some more complex activity. The last are (*i*) taught or* learned by each individual*, (*ii*) come from experience (through* trial and error*,* memories of past experiences *and observations of others) and (*iii*)* can be refined through practice *in order to adapt to suit changing environmental conditions.*

3) In the defense of the body and maintenance of a margin of safety two important cortical areas, with short latency neuron responses, are involved: *a*) the *ventral intra-parietal area* (VIP) and *b*) the *frontal polysensory area* (PZ) [3]. In VIP and PZ areas (whose neurons respond to similar types of sensory stimuli and whose stimulation leads to similar defensive-like outputs[2] [4]): *i*) most neurons are bimodal and trimodal, responding simultaneously to visual, tactile, and

This work was supported by CNCSIS – UEFISCSU, project number PN II – IDEI 1552/2008.

[1] A protective space like an invisible bubble surrounding the body; whenever this margin of safety is violated, the individual steps away to restore it.

[2] The evoked movements are consistent with avoiding, withdrawing, or protecting the part of the body on which the tactile receptive field is located.

auditory stimuli; *ii*) independently of the size of the stimulus, about half of VIP cells respond best to visual stimuli within 30 cm of the body, and many respond only within a few centimeters; similarly, in PZ about 46% of the neurons give a strong, sustained response only for visual stimuli within 5 cm while another 40% give a response for visual stimuli within 20 cm of the body surface; *iii*) multimodal cell responses show a gradient of firing rate as a function of distance to the stimulus located in the peripersonal space (i.e., responses decrease nonlinearly as a function of stimulus distance); *iv*) some VIP multisensory neurons receive vestibular input and they are thought to detect the direction of subject self-generated motion; *v*) PZ neurons projects directly at the motor spinal level, being thus involved in defensive motor output.

4) Descending inputs from the brain are directly responsible for the initiation and generation of voluntary movements. The complex reflexes (e.g. like the obstacle-avoiding reflex that usually sub-serves other voluntary goal-directed movements) arise from the projection of sensory inputs, through spinal inter-neurons, on to spinal neuronal circuits like central pattern generators (CGPs) that execute stereotyped motor sequences. In this wiring diagram, the role of the spinal inter-neurons is to integrate descending inputs from the brain with primary afferent inputs and, thus, to adapt reflexes and the activity of spinal motor neurons to different functional requirements.

5) The *Mowrer's two-factor theory of avoidance* postulates that avoidance learning involves two stages: (1) *a first stage*, in which the learner experiences *classical/fear conditioning* (e.g. an warning stimulus, like a small distance to an obstacle, is paired with an unpleasant situation, like collision; thus, a neutral stimulus becomes a conditioned stimulus (CS) capable of producing a strong conditioned fear response, and (2) *a second stage*, in which the learner experiences *operant conditioning* (i.e., the subject takes action response to the aversive CS and, thus, eliminates through negative reinforcement the aversive event). The last stage corresponds to avoidance learning process itself during which *the avoidance behavior is not reinforced by avoidance of the unpleasant situation but by termination of the aversive and fear-evoking CS*. In other words, the aversive stimuli reinforce the responses that remove them.

6) The *Gibson's ecological theory of visual perception* [5] claims that in order to generate the proper actions we need perceptual (not sensorial!) information related to the environmental factors that are important to the motor task. In our case we used an average-depth perception along with the perceptions of gait direction and speed.

B. Technical and practical considerations

In what follows our proposed obstacle avoidance learning implementation should be regarded only as a very reductionist model of its human analogue without loosing, however, of its relevance. Among all the bio-inspired models reported in the robotic literature as solutions for avoidance learning, the reinforcement learning (RL) models and the operant behavior (OB) models seem to be the most appealing. As in the human's escape-avoidance learning these learning models: (1) does not require a complete knowledge of the environment or the knowledge of the action to be taken according to each particular subject-environment context (anyway, there is not a unique solution for the action to be taken); (2) above all, they require as many and diverse as possible robot-environment interactions, being, thus, *online learning algorithms*; (3) their learning technique is based on *trial and error interaction* with a dynamic environment; (4) they provide the autonomous mobile robot with suitable evaluation of its performance in terms of punishment and rewards. However, while the results obtained with each of the above mentioned learning models look very promising they are still missing some details or are facing some implementing problems that, finally, lead the solution to be faraway from expressing behavior in a human-like manner. Particularly, when speaking of reinforcement learning and its widely used implementation (i.e., Q-learning algorithm [6], which requires estimating for each possible [state, action] pair an expected discounted numerical signal reward), at least some downsides ought to be mentioned. **1**) First, a huge Q-table needs to be manipulated when updating the Q-values (e.g., for a robot with 8 sensors, with 5 actions to choose and with an input range of 0-1022 for each sensor, a 1.1995 x 10^{24} x 5 Q-table[3] is needed). Recently, instead of using the look-up tables to store the Q-values, the neural networks (NNs) are used for their both capacities – to give a more compact representation of Q values and to interpolate the Q values for the state-action pairs that had never been visited. However, while using different NN paradigms new problems [1] arise like, for example, the instability reported for multilayer perceptron architectures trained with the backpropagation algorithm. Exactly, it becomes very hard to ensure that learning of new patterns does not erode the previously learned knowledge. **2**) A second important drawback is the very limited number of possible actions, entailed mainly by the computational constraints. **3**) Another major drawback is the large amount of unknown parameters of both the Q-learning algorithm (i.e. discount rate, initial temperature parameter, reward function etc.) and of the neural network implementation; for more drawbacks see [2].

Unlike the RL methods, where in a first stage of exploration (*learning phase*), independently of the current state, the agent explore the environment by selecting non-greedy actions (using the Boltzmann probability distribution), in operant conditioning [2] the stimuli received by the robot are used to learn what actions to perform more (i.e. the rewarded ones) and what actions to perform less (i.e. the punished ones). Consequently, while both types of methods are online methods, the operation in real time is a characteristic of only the OB methods. More, it was reported in [2] that neither of the above methods achieves a perfect performance; one of the possible reasons for this could be the fact that in the training phase the reward or punishment term is used instead of the sensory input signal which is more appropriate given its higher potential information.

[3] The size of Q table is bounded by (number of all possible states * number of all possible actions), having in total 1023^8 x 5 entries.

III. BIO-INSPIRED OBSTACLE AVOIDANCE LEARNING

In what follows, a new bio-inspired avoidance learning method is presented and particularly implemented for a differential wheeled robot endowed with 36 evenly distributed IR sensors, Fig. 1.a. A behavior controller based on a *minimal multilayer perceptron network* and trained with a new introduced *biological-like backpropagation* (BBP) algorithm was considered in order to meet the online and in real time learning criteria.

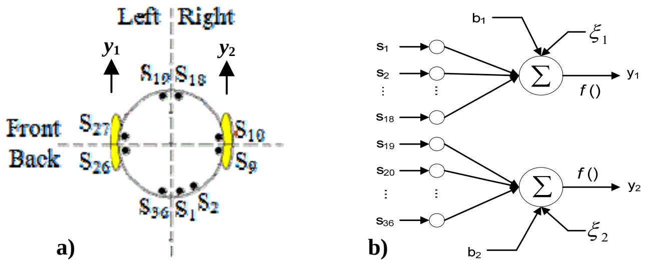

Figure 1. a) Arrangement of IR sensors; b) MLP's configuration.

The minimal architecture of the MLP network, Fig. 1.b, included 36 inputs (i.e. the 36 IR sensors readings, s_i) and 2 outputs (y_1, y_2 providing commands for the 2 wheel actuators). The short range of IR sensors of only 30 cm made them appropriate to use for environment fast-adapting mechanisms. The sensors values were first linearized and, then, normalized in [0, 0.9] range, with 0 denoting no obstacle and 0.9 denoting collision. For the two output processing elements (PEs) the *tanh* activation function was used that limits the output values in the range [-1,+1]. The overall movement commands, coded as follows – forward (1, 1), backward (-1, -1), left (0, 0.5), right (0.5, 0) and stop (0, 0), were exerted through the biases (b_1, b_2) of the output PEs. In our implementation the forward and backward wheel directions were indicated by positive and negative values, respectively, while the right and left driving were obtained providing proper differential commands to the two engines. Speed information, y_j, was provided as a proportional value between 0 (stop) and 1/-1 (maximum forward/backward movement). When calculating the activation of the output PEs we also added a small noise term, ξ, which prevented robot to get stick in a particular environment condition. To make a rough analogy with the human system, the NN's input layer corresponds to the PZ zone, the output PEs are the inter-neurons that integrate the primary sensory information (passed through the PZ zone) with the cortical command, and the wheel actuators are the spinal CGPs.

In its standard form (1) the BP algorithm needs a teacher that knows or can calculate the desired output for any given input. However, this is not our case since we cannot define the desired values for the engines commands, the latter depending on multiple (frequently unpredictable) factors.

$$
\begin{aligned}
&BP\!:\quad \Delta w_{ij} = -\eta \frac{\delta E}{\delta w_{ij}} = -\eta(d_j - y_j)s_i(1-y_j^2),\quad \Delta b_j = -\eta(d_j - y_j)(1-y_j^2) \qquad (1)\\
&\qquad E = \frac{1}{2}\sum_{j=1}^{2}\left(d_j - y_j\right)^2
\end{aligned}
$$

Here, w_{ij} are the ANN's weights, E is the error and η stands for the learning rate (in our case, 0.7).

Yet, to be still entitled to use the BP algorithm (which best describes the way the infants are learning by doing, in a trial and error manner) we took advantage of three of the psychological theories: ***i***) the *ecological theory of visual perception*, ***ii***) the *theory of internal motivation* [7] and ***iii***) the *Mowrer's two-factor theory of avoidance*. Finally, we came to a paradigm shift, namely, we substituted the teacher-centered paradigm, BP, with a new, learner-centered paradigm, BBP (2).

$$
\begin{aligned}
&Forward:\quad y_j = \tanh\left(\sum_i w_{ij}s_i + b_j + \xi_j\right), i=\overline{1,18}\ for\ j=1\ and\ i=\overline{19,36}\ for\ j=2\\
&BBP:\quad \Delta w_{ij} = -\eta T_j s_i(1-y_j^2),\qquad \Delta b_j = 0\\
&\qquad T_j = E_j\ Crit_j\\
&\qquad E_j = A_j - A_j^d = A_j,\quad A_j = \frac{\sum_i s_i}{n_j}, i=\overline{1,18} for j=1, i=\overline{19,36} for j=2 \qquad (2)\\
&\qquad Crit_j = sign(\Delta E_j)sign(y_j) = sign(\Delta A_j)sign(y_j)
\end{aligned}
$$

The weights' new updating rule (2) differs from rule (1) only in the composite term T_j for which the equivalent term in (1) is (d_j - y_j); the latter is a measure of how much of the error value is due to the *j*th ANN output. Unlike the standard BP (1) where this error term has a clear analytic form, depending on the desired value, d_j, and on the ANN's actual output, y_j, in BBP it is calculated in the particular manner the learner's brain perceives the environment stimuli (i.e E_j) and manipulates these perceptions (i.e. $Crit_j$). These last new introduced terms will be explained later in this paper.

Normally, the brain uses the difference between the way the world is perceived and the way it should be perceived (according to our goals) as an error information and tries to correct the movement. In order to generate the proper actions, humans use perceptual, not sensorial information related to the environmental factors that are important to the motor task. In our case, the *perception of the aversive stimuli*, A_j – that is related to the perception of average depth[4] to the obstacles lying within the contralateral right/left side of the robot – was used in the learning phase (2). For each output PE a different A_j was computed in order to facilitate the learning of both side reflexes (this ensures avoiding obstacles by the appropriate side). In (2), n_j represents the number of the sensors from the contra-lateral part that are of non-zero value. In the BBP learning algorithm the aversive CS, given by the perception A_j, becomes larger as the average depth to obstacles lying in the right/left PPS becomes smaller and ceases (becomes zero) when all corresponding sensor readings indicate a zero value. The last case corresponds to the desired perception of the environment, A_j^d = 0, when the robot is ideally located at a least 30 cm distance from any surrounding obstacle. Accordingly, the learner-derived error, E_j, is finally given by the perception A_j which is nothing else but the aversive stimuli that should be removed. The learner-derived error, assessed in [0, 0.9] range, complies with the fundamental paradigm that defines the BP error: when the adaptive system successfully solves the problem

[4] See the meaning given to the sensors' readings.

the error is zero; otherwise, the error measures the distance between the desired results and the current outputs of the adaptive system.

Further, according to the same operant conditioning stage, the learner reinforces those actions that remove the aversive stimuli. For this he evaluates the consequences of its own actions, as response to the environmental stimuli, and *internally generates a scalar reward*, *sign*($A_j[n]$ $-A_j[n\text{-}1]$), that accounts for the *learning progress*. This qualitative measure along with the perception of movement's direction of wheels make up the critic term, ***Crit***, which controls the way the ANN is updating its parameters. Namely, the updating in the same last movement direction is encouraged whenever the learner is moving such as to remove/diminish the aversive stimuli, $A_j[n] \leq A_j[n\text{-}1]$ or the updating takes place in the opposite direction in the other case.

In the BBP paradigm, in order to capture the fact that learning ceases when avoidance behavior is well practiced, we imposed that after the first 300 steps, during each trial the weights' updating to be further done only if at least one of the ANN input was more than 0.8.

The maximum wheel speed was 0.3 m/s and the delay between each system action (movement) and its subsequent sensors reading was 400 ms. The initial values of the weights were not randomly generated but, instead, they were assessed to zero in order to mimic the progressive neural wiring encountered in babies during their neural control development process.

IV. Results and Discussions

We tested our learning algorithm in both, static and dynamic environment (with two and, respectively, with five mobile robots), in MobotSim 1.0.03 simulator. The avoidance behavior learned with the new algorithm is presented in Fig. 2.a for the first 2527 steps and, respectively, in Fig. 2.b for the first 70517 steps.

Analyzing the way the robot learned the avoidance behavior the following main ideas can be extracted. *As in humans*: (1) The proposed method allowed the online and the in real time learning implementation (being computationally cheap). (2) The learning occurred quickly (after only one or two collisions) and it was very durable, without any other collisions. (2) For learning process the most important were the first interactions with the environment, these mainly shaping the later overall avoidance behavior. (3) One important consequence of our proposed solution was the construction of a margin of safety around the robot. This agrees with the recognized fact that human "personal space" is the result of a defensive mechanism that monitors potentially threatening objects near the body. (4) The robot reacted promptly whatever the approaching stimulus (second robot) direction was. (5) The way the movement is generated through the two differential wheels, and the way the ANN outputs are calculated allow for a wide range of movements, from here resulting the high movement flexibility. (6) For each of the five "cortical" motor commands a different ANN was implemented and trained and each time the robot succeeded to learn the avoidance behavior indifferently of the selected movement direction. (7) The "cortical" command and the autonomous behavior successfully switched control each time the environment context required.

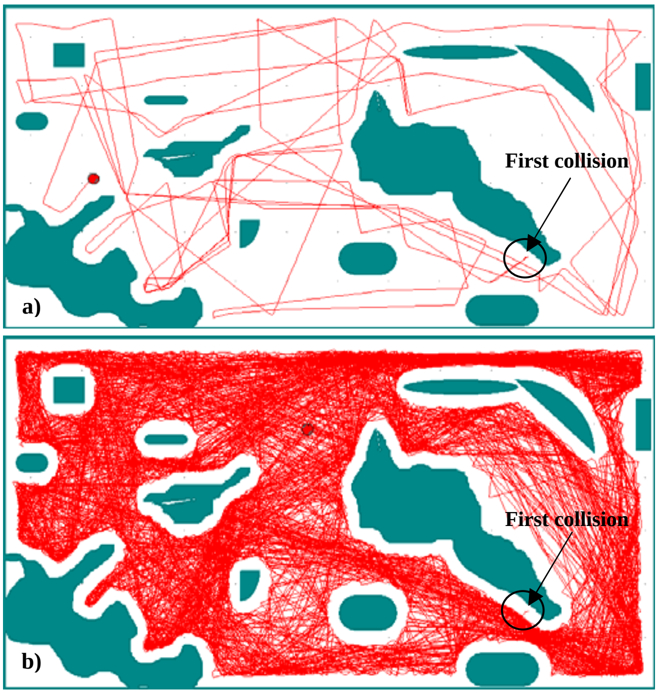

Figure 2. The avoidance behavior learned by the robot: a) the first 2527 steps and b) the first 70517 steps of the algorithm

V. Conclusions

In this paper, the concept of a robotic bio-inspired *reactive behavior* control is briefly presented. The biological plausibility, alongside the high simplicity of learning technique and learning paradigm, makes the system very effective: i) it learns fast and in a consistent manner; ii) the autonomous behavior is achieved online, in real time and in an unstructured and dynamic environment, iii) the "cortical" command and the autonomous behavior successfully switch control each time the environment context requires.

References

[1] A. H. Tan, N. Lu, and D. Xiao, "Integrating temporal difference methods and self-organizing neural networks for reinforcement learning with delayed evaluative feedback," in IEEE Trans Neural Netw., vol. 19, pp. 230-244, 2008.

[2] D. A. Gutnisky, B. S. Zanutto, "Learning obstacle avoidance with an operant behavior model," Artif. Life, vol. 10, pp. 65-81, 2004.

[3] M. S. Graziano, and D. F. Cooke, "Parieto-frontal interactions, personal space, and defensive behavior," Neuropsychologia, vol. 44, pp. 845–859, 2006.

[4] M. S. A. Graziano, C. S. R. Taylor, T. and Moore, "Complex movements evoked by microstimulation of precentral cortex," Neuron, vol. 34, pp. 841–851, 2002.

[5] A. Shumway-Cook, and M. H. Woollacott, Motor control: translating research into practice, 3rd ed., Lippincott Williams & Wilkins, U.S.A., 2007, pp. 16.

[6] C. Watkins, and P. Dayan, "Q-Learning", Mach. Learn., vol. 8, pp. 279-292, 1992.

[7] P.-Y. Oudeyer, and F. Kaplan, 'What is intrinsic motivation? A typology of computational approaches," Front. Neurorobotics, vol. 1(6), pp. 1-14, 2007.