Journal of Physiology - Paris 107 (2013) 107–115

Contents lists available at SciVerse ScienceDirect

Journal of Physiology - Paris

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

Reinforcement active learning in the vibrissae system: Optimal object localization

Goren Gordon [a,*], Nimrod Dorfman [b], Ehud Ahissar [a]

[a] Department of Neurobiology, Weizmann Institute of Science, Rehovot 76100, Israel
[b] Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel

ARTICLE INFO

Article history:
Available online 9 July 2012

Keywords:
Intrinsic reward
Whiskers
Curiosity loop
Perception
Palpation

ABSTRACT

Rats move their whiskers to acquire information about their environment. It has been observed that they palpate novel objects and objects they are required to localize in space. We analyze whisker-based object localization using two complementary paradigms, namely, active learning and intrinsic-reward reinforcement learning. Active learning algorithms select the next training samples according to the hypothesized solution in order to better discriminate between correct and incorrect labels. Intrinsic-reward reinforcement learning uses prediction errors as the reward to an actor-critic design, such that behavior converges to the one that optimizes the learning process. We show that in the context of object localization, the two paradigms result in palpation whisking as their respective optimal solution. These results suggest that rats may employ principles of active learning and/or intrinsic reward in tactile exploration and can guide future research to seek the underlying neuronal mechanisms that implement them. Furthermore, these paradigms are easily transferable to biomimetic whisker-based artificial sensors and can improve the active exploration of their environment.

1. Introduction

Rats are curious animals that use their vibrissae (whiskers) to explore their environment. Several stereotypical behaviors have been observed, such as periodic whisking (Gao et al., 2001; Berg and Kleinfeld, 2003) and touch-induced palpation (Grant et al., 2009). Recently, whisking behavior has been implemented in robotic whiskers in order to discriminate textures and ascertain three-dimensional shapes (Solomon and Hartmann, 2006; Evans et al., 2010; Sullivan et al., 2012). Palpation of novel objects, which is the focus of the current work, is observed when rats encounter such an object and can be characterized as a high-frequency small-amplitude whisker motion, always remaining in the vicinity of the object. It has received very little attention from the analytical and robotics-implementation fields (Gordon and Ahissar, 2011; Gordon and Ahissar, 2012).

Here we show that two seemingly unrelated paradigms, namely, active learning (Kolodziejski et al., 2009; Bhatnagar et al., 2007; Govindhasamy et al., 2005) and intrinsic-reward reinforcement learning (Barto et al., 2004; Weng, 2004; Oudeyer et al., 2007; Schmidhuber, 2010), predict that touch-induced palpation is the optimal behavior for whisker-based object localization. We then show that in the context of object localization, the two paradigms are tightly related and suggest neuronal mechanisms that may implement each.

Rats' vibrissae system serves as a unique model for neuroscience research due to its relative simplicity. Although its dynamics becomes more complex as investigations progress (Knutsen and Ahissar, 2009; Simony et al., 2010), it can be approximated as a one-dimensional process, controlling a single positional variable, the whisker's azimuth angle using a single motor variable, whisker velocity. Then whisker-based object localization can be defined as learning the forward model (Jordan, 1992; Shadmehr and Krakauer, 2008) of touch, i.e. the ability to predict at what angle and velocity a touch signal, due to contact between the whisker and object, will occur. The question we address is "how should a single-whisker rat move its whisker in order to optimally localize an object?" In other words, what is the rat's policy that optimizes learning of the forward model of touch, where optimization is performed with respect to the learned function (see below).

This scenario can be formulated using the active learning jargon in the following way (Adejumo and Engelbrecht, 1999; Dasgupta and Hsu, 2008). The rat *samples* the sensory-motor space (angle and velocity) and wishes to correctly *label* each point as touch or no-touch. We show that object localization is equivalent to learning a two-dimensional linear separator (albeit in bounded space due to angle and velocity limitations). Hence, the goal is to find the sampling policy that minimizes the error between the predicted linear separator and the correct one.

In reinforcement learning (RL) notations (Kolodziejski et al., 2009; Bhatnagar et al., 2007; Govindhasamy et al., 2005), the *states*

* Corresponding author. Tel.: +972 525374429; fax: +972 775374424.
E-mail addresses: goren@gorengordon.com (G. Gordon), Ehud.Ahissar@weizmann.ac.il (E. Ahissar).

0928-4257/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jphysparis.2012.06.004

are the angle of the whisker and the touch information, and the *action* is whisker velocity. Hence in an actor-critic setup (Bhatnagar et al., 2007), the critic learns the values of each angle/touch point, whereas the actor adjusts the probabilities of choosing a specific whisker velocity given an angle/touch state. In conventional RL, the reward is given by an *extrinsic* function that is adjusted to the desired goal, e.g. maximal reward for arriving at a specific location. However, in the current implementation of RL, the object localization component, i.e. a learner that learns the forward model of touch, provides *intrinsic* reward (Barto et al., 2004; Weng, 2004; Oudeyer et al., 2007; Schmidhuber, 2010), here taken to be the prediction error. Thus, the goal is to find the actor that optimizes learning object localization, i.e. minimizes the generalization error of the forward model of touch. In other words, the intrinsic-reward RL converges to a behavior that results in fast increase in accurate prediction of touch events.

The paper provides a unifying formalism for both approaches, with respect to object localization via whiskers. This allows the direct comparison of the two approaches, which exhibit remarkably similar results, namely, palpation behavior. It further enables a formulation of the connection between the two paradigms, also explored here. A biologically-plausible neuronal network that implements the proposed models is also presented and discussed. Finally, the synergistic analysis presented here can facilitate the application of either techniques in robotic whisker-base sensors (Solomon and Hartmann, 2006; Evans et al., 2010; Sullivan et al., 2012).

2. Materials and methods

2.1. Whisker model

We use a simplistic model, in which the rat can *control the velocity* of the whisker (Simony et al., 2010). Furthermore, the whisker itself is rigid, i.e. it cannot bend and hence its azimuth angle cannot pass the "object's angle". Thus, the whisker *angle is bounded* by the object position and depends on the initial whisker angle, i.e. if it is initially more retracted (smaller angle) or more protracted (larger angle) than the object (angular) position. For simplicity, we assume the whisker to be always more retracted than the object; hence the object is touched only upon protraction of the whisker. This assumption is validated by numerous videos of freely moving rats, in which they encounter novel objects upon protraction in the vast majority of cases. We also assume that the velocity is bounded, due to physical constraints.

2.2. Learning a linear separator in sensory-motor space

We formulate the whisker-based object localization setup mathematically: $\theta \in \left[\theta^{\min}, b\right]$ is the angle of the whisker, $\theta^{\min}$ is the fully retracted angle, and b is the (angular) position of the object, with $b \in \left[\theta^{\min}, \theta^{\max}\right]$, $\theta^{\max}$ being the fully protracted angle. This means that the object can appear anywhere inside the whisker field. We assume that the whisker is always more retracted than the angular position of the object, and hence bounded by it. $a \in [a^{\min}, a^{\max}]$ is the bounded velocity of the whisker.

The dynamics of the system are given by

$$\theta'_{t+1} = \theta_t + a_t \tag{1}$$

$$\theta_{t+1} = \max \theta_{\min}, \min(b, \theta'_{t+1})) \tag{2}$$

where θ'_{t+1} is the attempted angle and Eq. (1) guarantees that the angle stays within the bounds. The velocity a_t is the action that should be optimized (see below). The touch signal is then given by

$$B_{t+1} = \begin{cases} 1 & \theta_t \leqslant b \text{ and } \theta'_{t+1} > b \\ -1 & \text{otherwise} \end{cases} \tag{3}$$

This means that if the whisker tried to move from one side of the object to the other side of the object, there is a touch signal of 1, otherwise $B = -1$. One can then define a linear separator of touch, $u = \{u_\theta, u_a, u_b\}$, such that

$$u_\theta \theta_t + u_a a - u_b = u^T x_t = 0 \tag{4}$$

where $x_t = \{\theta_t, a_t, -1\}$ is a point in 2-dimensional $(x^1 = \theta, x^2 = a)$ space, where $x^3 = -1$ is a constant added to accommodate for the linear separator's threshold u_b. The linear separator, u, delineates the boundary between the labeled touch and no-touch regions in the two-dimensional (θ, a) space.

The setup can then be re-formulated as follows: (i) The agent's policy determines, based on past knowledge, the action a_t; (ii) the dynamics are determined via Eqs. (1) and (3); (iii) the agent receives $\{\theta_{t+1}, B_{t+1}\}$; (iv) based on the action, angle and touch signal, the agent updates its approximation of the linear separator. (v) $t \to t+1$, return to (i). The goal is then restated as: find a policy such that the touch-signal linear separator, u, is learned optimally.

2.3. Perceptron-based active learning

The setup described in the previous section can be modeled by a perceptron, which is a mathematical construct that receives many inputs and has a single output. The perceptron output is the result of applying a (usually) non-linear or threshold function on the weighted sum of its inputs. In the object localization scenario, the perceptron inputs and output are the two-dimensional point (θ, a) and touch signal, respectively.

The problem is also related to selective sampling, a branch of active learning (Settles, 2009), in which one can select whether to label the sample or not. Since the labeling is usually costly, the aim is to select which samples to label. We briefly describe a perceptron-based active learning algorithm taken from Dasgupta et al. (2009), which actively selects which samples to label and exhibits an exponential speedup compared to random selections.

Let x be a point on the N-dimensional unit hypersphere, $\sum_{i=1}^{N} x_i^2 = 1$. Let u be a vector on the same sphere, such that $y = \text{sign}(u^T x)$ is the label of each point x on the sphere. In each time-step, t, there is a hypothesis vector, v_t. The goal of active learning is to find u, i.e. change the hypothesis such that $v_t \to u$.

In selective sampling, one is presented with random samples from the unit sphere, x_t. The algorithm presented in Dasgupta et al. (2009) only labels samples obeying:

$$|v_t^T x_t| < q_t \tag{5}$$

where x_t is the sample at time t, v_t is the current hypothesis/classifier and q_t is an adaptive threshold that decreases as learning progresses. It was shown that the update rule of the hypothesis, v_t, given by

$$v_{t+1} = v_t - 2(v_t^T x_t)x_t \tag{6}$$

results in a number of required labels that is exponentially smaller for a given error, compared to random labeling. The crux of the algorithm in Dasgupta et al. (2009) is the adaptive threshold q_t, which adapts according to the following rule: if predictions were correct on R consecutive labeled examples, then set $q_{t+1} = q_t/2$, else $q_{t+1} = q_t$. This means that the adaptive threshold decreases as the error in the prediction decreases.

2.4. Reinforcement active learning

Reinforcement learning (RL) deals with the question of finding an actor that maximizes (future) accumulated rewards. In our