Neurocomputing 268 (2017) 142–152

Contents lists available at ScienceDirect

Neurocomputing

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

Grounding the experience of a visual field through sensorimotor contingencies

Alban Laflaquière

AI Lab, SoftBank Robotics Europe, 43 rue du Col. Pierre Avia, Paris 75015, France

ARTICLE INFO

Article history:
Received 5 July 2016
Revised 3 October 2016
Accepted 18 November 2016
Available online 2 May 2017

Keywords:
Autonomous systems
Developmental robotics
Sensorimotor contingencies
Predictive processing
Sensorimotor learning
Human-like vision

ABSTRACT

Artificial perception is traditionally handled by hand-designing task specific algorithms. However, a truly autonomous robot should develop perceptive abilities on its own, by interacting with its environment, and adapting to new situations. The sensorimotor contingencies theory proposes to ground the development of those perceptive abilities in the way the agent can actively transform its sensory inputs. We propose a sensorimotor approach, inspired by this theory, in which the agent explores the world and discovers its properties by capturing the sensorimotor regularities they induce. This work presents an application of this approach to the discovery of a so-called visual field as the set of regularities that a visual sensor imposes on a naive agent's experience. A formalism is proposed to describe how those regularities can be captured in a sensorimotor predictive model. Finally, the approach is evaluated on a simulated system coarsely inspired from the human retina.

1. Introduction

Autonomy in robotics relies on sensory data processing to capture information about the world and adapt to it. Although the influence of machine learning has been growing more important in the last decades, traditional approaches to this problem of data processing involve significant manual design from engineers that build the robot. Consequently the resulting techniques for artificial perception appear rigid and constrained for tractability. Each of these specialized algorithms is applicable to only a small set of tasks, with potentially limiting inbuilt biases from the designer. While acceptable for well-defined processes, such as industrial manufacturing, the potential need for a large degree of human involvement makes such methods inadequate as a source of long-term autonomy in a robot. Instead, an autonomous robot must be able to cope with the complexity of its world, build its own way to perceive it and adapt to its variations.

To address this issue, the field of developmental robotics takes inspiration from biological and cognitive development in children [4]. It proposes that an agent learns to interact with its environment, autonomously and on an ontogenic timescale. Without prior knowledge, a naive robot must learn the structure of its own body, of its environment, and how the two interact. In this context, *perception* is a prerequisite to developing more advanced cognitive abilities that allow a rich interaction with the environment. Yet, the emergence of this fundamental capacity, traditionally hand-coded in the system, poses a challenge: What is *perception* for a naive agent in which manually pre-defined features and labels are replaced by a flow of uninterpreted sensorimotor data?

Sensorimotor contingencies theory (SMCT) attempts to answer this question [24] by fundamentally re-defining perception: *perception is the mastery of regularities in the way actions transform sensory inputs*. It suggests that a naive agent can actively explore its environment, extract regularities that the world imposes on its sensorimotor flow, and later identify those regularities when interacting with the environment in order to perceive it. Those regularities, or *contingencies*, are the ground on which the agent can build perceptive abilities. Moreover this active account of perception naturally links actions to perception, meaning that the agent intrinsically knows what it could *do* with any perceived feature [27]. Despite its philosophical aspect, the SMCT is based on experimental results. Among other things, it elegantly accounts for instance for sensory substitution. Those are experiences in which a subject is provided with information from one modality (e.g. vision) through another modality's pathway (e.g. skin or ears) [1,2]. The theory naturally encompasses such a phenomenon as it defines perception as based on structure in the sensorimotor flow instead of properties of the pathway it takes. Such a possibility leads us to consider artificial "plug-and-play" agents that could be equipped with new sensors, and would discover how to perceive with them by learning to master the associated sensorimotor flows.

E-mail addresses: alaflaquiere@aldebaran.com, alban.laflaquiere@gmail.com

http://dx.doi.org/10.1016/j.neucom.2016.11.085
0925-2312/

The SMCT has been relatively slow to spread in the robotics community, partly because of the complete overhaul it induces in the field of artificial perception. To date, the approach has been applied to model the acquisition of perceptive concepts such as space [18,29], colors [25,30], environments [17], and objects [16]. Primarily, these works characterize properties of the external world explored by the agent. However, a naive agent's body is also part of the unknown world it has to discover. It contributes, like the structure of the environment, to shaping the regularities the agent experiences in its sensorimotor flow. As such, properties of the agent's body should also be captured through sensorimotor contingencies. In this paper, we address the problem of capturing properties of sensors plugged on a naive agent, and in particular properties of the *visual field* generated by visual sensors. The experience of visual field encapsulates the set of regularities describing how visual features are encoded differently by various parts of the sensor, as well as how they shift on it due to motion. This fact is particularly striking when considering heterogeneous visual sensors, like the human retina, for which visual features are encoded by significantly different cell patterns depending on where they land on the retina. This discrepancy between our stable subjective experience of visual features and their actual variable sensory encoding has already been brought forward in the paper introducing the SMCT [24]. Yet, only recently has it led to further inquiries with the development of psycho-motor experiments [8]. Their results suggest that the brain learns the correspondence between the different sensory patterns that encode the same visual feature on different parts of the retina, and the motor commands (ocular saccades) that transform one into the other. By exploring artificial visual setups where two distinct visual features are consistently associated before and after a saccade, it is possible to alter previously learned correspondences. This artificial interaction with the world leads to a modification of the subjective perceptive experience of visual features, even in adult subjects.

The work presented in this paper proposes a computational model inspired by this perceptive phenomenon. Nonetheless it also fits into a more general endeavor to develop a computational model for the autonomous learning of sensorimotor regularities [17,21,27], the lack of which has been the second reason of the slow spread of SMCT. The formalism converges toward the hierarchical building of a predictive model of sensorimotor experiences [17]. This approach is in line with recent developments in neuroscience, which describe the brain as a predictive machine [5,6,9]. By learning to predict future sensory outcomes of its actions, the agent estimates latent causes of its experience and progressively extends the control from its motor component to its sensory component. The work presented in this paper will focus on letting a naive agent discover the sensorimotor regularities that define the *visual field* associated with a visual sensor. The next section presents a formalization of the problem and describes a computational model to address it. A simulation is then introduced in Section 3 to illustrate the approach. The results are analyzed in Section 4 in light of previous works in the sensorimotor approach of perception. Finally, limitations and potential future extensions of the model are discussed in the last section.

2. Problem formulation

In this section we present the problem a naive agent is facing when discovering the sensorimotor structure induced by its visual sensor. We describe the regularities that underlie the experience of a *visual field*. Then we propose a computational formalism to process the agent's sensorimotor flow and detail how it can capture those regularities.

2.1. Experiencing a visual field

This work focuses on agents equipped with a visual-like sensor: an array of sensels collecting information from a part of the environment, where a sensel is the basic element of a sensor array (e.g. pixels in a camera, or rods and cones in our retina). In this work, we use the term *visual feature* to refer to the visual information received from a small part of the environment. Contrarily to computer vision literature where visual features are the internal outcome of some sensory processing, the term here describes the (partial) state of the external environment. Conversely, we use the term *sensory inputs* to refer to the information generated when visual features are projected on the sensor and transformed into an encoded signal accessible to the agent (see Fig. 1).

Depending on where it is present in the visual field, a visual feature can be projected onto different parts of the sensor array. It can thus be encoded by different sensory inputs. Such a claim does not appear obvious when considering a camera because it is usually assumed that the sensory encoding is translation-invariant: physical properties of the sensor are such that a visual feature generates the same sensory input regardless of where it is encoded in the array. This is for instance an implicit hypothesis in Convolutional Neural Networks, a class of algorithms that prove to be very efficient in visual scene analysis [14]. It also indirectly assumes that the later unit that processes sensory inputs does know the spatial organization of sensels and can switch on the fly between different groups of neighboring sensels.

Yet, such a property is far from evident for a biological system like our visual cortex. This fact appears even less realistic when taking into account the heterogeneity of the human retina [26]. As underlined in [24], the way visual features are encoded changes significantly across the retina, due to its physiological properties. Yet, our subjective experience of visual features is that they are stable across the whole field of view. The sensorimotor point of view on such a phenomenon claims that the brain learns to associate the different sensory inputs corresponding to the same visual feature by actively exploring visual scenes. This hypothesis has been recently strengthened by psycho-motor experiments in which those associations were artificially altered [8].

According to SMCT, the very mastery of those sensorimotor associations participates in the experience of seeing. More precisely, by focusing on regularities induced by the physical structure of the sensor, one can describe those that give rise the experience of having a *visual field*:

- Sets of different sensory inputs encode the same visual features on different parts of the sensor.
- Motor commands can transform sensory inputs into another one encoding the same visual feature.

Those two statements describe that visual features shift on the retina and that their encoding changes when the agent moves its sensor (see Fig. 1). It is important to notice that the only way for a naive agent to discover such properties is to actively explore visual scenes. They could not be extracted through a passive sensory processing, like the ones usually proposed in unsupervised contexts [3]. For instance, the different sensory inputs related to a single visual feature would not necessarily share the same statistical properties or lie close to one another in the sensory space, especially if the sensor is significantly heterogeneous. Additionally, passively extracted knowledge would not be directly useful to a naive agent as it would not know how to actively transform its sensory state to eventually reach a goal state (for instance, move the sensor to bring a visual feature in a given part of the visual field). Yet, the association of sensory inputs seems intimately linked to the ability to perform visual tasks such as search and recognition, as demonstrated in [8].