ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence
and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

Grounding the Experience of a Visual Field through Sensorimotor Contingencies

Alban Laflaquière, Michaël Garcia Ortiz, Ahmed Faraz Khan

AI Lab, Aldebaran Robotics, 43 rue du Colonel Pierre Avia, Paris, 75015 France

Abstract. Artificial perception is traditionally handled by hand-designing specific algorithms. However, a truly autonomous robot should develop perceptive abilities on its own by interacting with its environment. The sensorimotor contingencies theory proposes to ground those abilities in the way the agent can actively transform its sensory inputs. This work presents an application of this approach to the discovery of a visual field. It shows how an agent can capture regularities induced by its visual sensor in a sensorimotor predictive model. A formalism is proposed to address this problem and tested on a simulated system.

1 Introduction

Autonomy in robotics necessitates sensory data processing to capture information about the world. It is traditionally designed by engineers who write specialized algorithms for the extraction of specific sensory features, and for the detection of predefined entities in the world, to solve a task. Although very powerful in constrained environments, such an approach is too rigid and inadequate as a source of long-term autonomy in a robot. Instead, the latter must learn on its own how to interact with the world, and in its most basic form, how to perceive the world. The Sensorimotor contingencies theory (SMCT) [1] offers a re-definition of perception that can account for the autonomous acquisition of perceptive abilities: *perceiving means mastering regularities in the transformations that actions induce on sensory inputs*. By exploring its environment, an autonomous agent can discover those regularities, or *contingencies*, that constitute the material of perception. To date, SMCT has been applied to account for perceptive notions like space, colors, environments, and objects. This work focuses on a perceptive ability related not to the environment but to the agent itself: namely, the experience of having a *visual field* associated with an retina-like sensor array. This experience encapsulates the set of regularities describing how visual features, encoded differently in various parts of the retina, shift and transform during saccades. As shown in recent psycho-motor experiments, these regularities are learned in humans and can be altered even in adulthood [2]. This work proposes a computational model illustrating how regularities associated with a *visual field* can be captured by a naive agent.

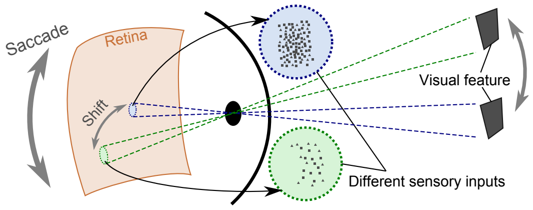

Fig. 1: Visual features are encoded as different sensory inputs depending on where they are projected on the heterogeneous retina. The eye can experience these different sensory inputs by saccading.

2 Problem Formulation

Taking inspiration from the human retina, this work focuses on agents equipped with a visual-like sensor: an array of sensels[1] collecting information from a part of the environment. We use the term *visual feature* to refer to the visual information received from only a small part of the environment and that is projected on the sensor array to generate sensory inputs. It is thus used differently than in computer vision literature where it refers to the outcome of sensory processing. For an agent with no a priori knowledge about the properties of its sensor, experiencing a *visual field* has two aspects:

- The way the visual features are encoded into sensory inputs may depend on the part of the array on which they are projected,
- Moving the sensor shifts the visual features on the array.

As discussed in [1], these phenomena are evident in the human vision system due to the retina's high heterogeneity (see Fig. 1). Yet, the brain learns those transformations, enabling feature search and recognition, as demonstrated in [2]. The mastering of those transformations also contributes to the subjective experience of uniform acuity in our field of view, even when peripheral vision is actually very coarse.

We propose a model of how a naive agent can capture those transformations, or *regularities*, in a predictive model [3]. Taking inspiration from human retinal structure, we treat the array of sensels as multiple *receptive fields*. Each receptive field includes numerous sensels, but covers only a limited part of the whole array, encoding like this visual features. No other constraint is assumed as the different receptive fields may have different properties (e.g. the number of sensels or the size in the array). For the naive agent, each receptive field initially appears to be an independent sensor generating its own sensory input. Formally, we define the *sensory state* vector $\mathbf{s}^a$ generated in receptive field a as:

$$\mathbf{s}^a = [s_1, \ldots, s_{d^a}] \tag{1}$$

[1] A sensel is a basic element of a sensor array (e.g. pixels in a camera, or rods and cones in our eye).

where s_i is the individual sensation provided by the i^{th} sensel in receptive field a, and d^a is the number of sensels in receptive field a. The agent is able to move its visual sensor using saccades, analogous to human eye movements. Formally, the saccadic motor command sent by the agent is denoted:

$$\mathbf{m} = [m_1, \ldots, m_M] \tag{2}$$

with m_i individual commands sent to the motors moving the sensor. No specific superscript is needed for the motor command as all receptive fields move together rigidly.

While randomly exploring the world, the agent builds a predictive model of sensorimotor transitions:

$$P(\mathbf{s}_j^b(t+1)|\mathbf{s}_i^a(t), \mathbf{m}_l(t)) \tag{3}$$

corresponding to the conditional probability of experiencing a sensory state $\mathbf{s}_j^b$ in receptive field b, given that a sensory state $\mathbf{s}_i^a$ was experienced in receptive field a before the execution of the saccadic motor command $\mathbf{m}_l$. Note that the temporal index is discarded in further notations for the sake of simplicity. The regular transformations associated with the agent's visual field should appear as highly predictable sensorimotor transitions $P(\mathbf{s}_j^b|\mathbf{s}_i^a, \mathbf{m}_l)$.

3 Simulation

A simple system is simulated in order to illustrate the approach (see Fig. 2). It intends to coarsely capture the interaction a moving eye has with its environment. The agent is a translatable camera viewing a two-dimensional visual environment. Its field of view is limited to a narrow 30 × 30 pixels square. Mimicking a retina, the field of view is divided into 9 receptive fields a of size 10 × 10 pixels. In order to limit the simulation complexity, the potentially continuous sensory experience in each receptive field a is discretized into a set of sensory states $\mathbf{S}_k^a$. They correspond to a set of N^a clusters in which sensory inputs $\mathbf{s}_i^a$ can be categorized. Those clusters are generated in each receptive field by applying a simple K-means algorithm to 100000 sensory inputs $\mathbf{s}_i^a$ collected by randomly exploring 100 successive environments. New incoming data are then encoded using a winner-takes-all strategy:

$$\mathbf{s}_i^a \rightarrow \mathbf{S}_k^a \text{ such that } k = \underset{j}{\operatorname{argmin}}(||\mathbf{s}_i^a - \mathbf{S}_j^a||) \tag{4}$$

Taking inspiration from the human retina, the resolution of each of the 8 peripheral receptive field is artificially lowered to imitate the coarser sensory encoding in periphery compared to the central fovea. This is done by averaging groups of 4 neighboring pixels into (meta-)pixels, leading to peripheral receptive fields of size 5 × 5 pixels. Sensory inputs are thus respectively vectors $\mathbf{s}^a$ of size $d^a = 100$ and $d^a = 25$ for the central receptive field and for peripheral ones. The number

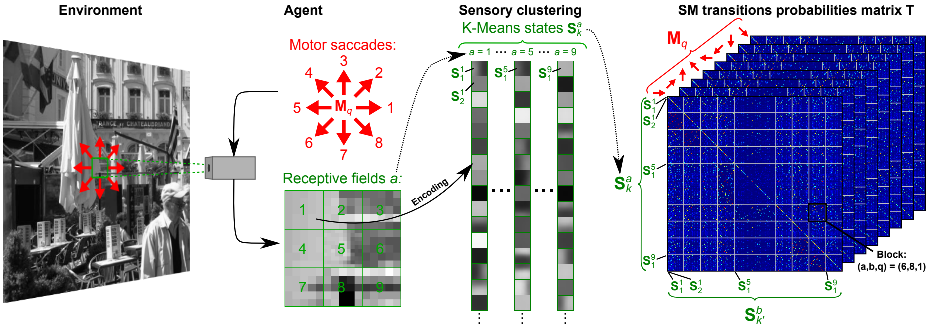

Fig. 2: The agent observes a small part of its environment. Its field of view is divided into 9 receptive fields: a central one (#5) of resolution 10×10 pixels, and eight peripheral ones (#1,2,3,4,6,7,8,9) of 5×5 (meta-)pixels. Sensory inputs are classified using K-Means clusters $\mathbf{S}_k^a$ built independently in each receptive field, and displayed as successive patches in columns. The agent can saccade in 8 directions $\mathbf{M}_q$ such that some pre- and post-saccadic receptive fields superimpose. The probabilities of transitions $P(\mathbf{S}_{k'}^b | \mathbf{S}_k^a, \mathbf{M}_q)$ are estimated by randomly exploring successive environments, and stored in a 3D transition matrix T.

of K-means clusters is arbitrarily set according to the dimension d^a of the associated sensory space: $N^a = 50$ for the central (foveal) one, and $N^a = 20$ for the peripheral ones.

The agent can translate in the plane to sample different parts of its environment. Similarly to sensory input, movements are discretized into a set of $Q = 8$ saccades $\mathbf{M}_q$. They correspond to all translations of the retina such that the central receptive field is shifted to the former location of a peripheral receptive field (see Fig. 2). They have been purposely chosen so that visual features entirely shift between receptive fields during saccades, which reduces the simulation complexity.

The environments explored by the agent are images from a standard RGB image database [4], converted to gray-scale. Although the agent does explore different successive environments, it is considered static during saccades; which is a reasonable assumption considering the speed of a human saccade.

To estimate the predictive model $P(\mathbf{S}_{k'}^b | \mathbf{S}_k^a, \mathbf{M}_q)$, the agent collects 10^6 experiences of sensorimotor transitions by executing 1000 random motor commands in 100 successive environments. This random policy, analogous to motor babbling, is natural for a naive agent with no a priori exploratory strategy. Each saccade generates an elementary sensorimotor transition experience $(\mathbf{S}_k^a, \mathbf{M}_q) \rightarrow \mathbf{S}_{k'}^b$ for each a and b (81 in total, given the 9 receptive fields). The probability $P(\mathbf{S}_{k'}^b | \mathbf{S}_k^a, \mathbf{M}_q)$ is finally estimated by building a normalized histogram of the outputs k' for each triplet (a, b, q). The overall predictive model is stored in a 3D matrix T gathering the estimated distributions for all triplets (a, b, q). As illustrated in Fig. 2, rows of T correspond to the pre-saccade sensory states k of all receptive fields. Its columns correspond to the same set of sensory states k'

after the saccade. Finally, its pages correspond to the saccadic motor commands q. This way, the matrix T can be interpreted by "blocks", corresponding to the predictive structure between pairs of receptive fields (a, b) for a given saccade q. Each row in those blocks correspond to the distribution $P(\mathbf{S}^b_{k'}|\mathbf{S}^a_k, \mathbf{M}_q)$ for the corresponding triplet (a, b, q).

In order to facilitate the later analysis of T, the normalized conditional entropy $H(a, b, q)$ is also computed for each block (a, b, q) of the matrix:

$$H(a, b, q) = -\sum_{k,k'} \frac{P(\mathbf{S}^b_{k'}, \mathbf{S}^a_k|\mathbf{M}_q)}{\log N^b} \log \frac{P(\mathbf{S}^b_{k'}, \mathbf{S}^a_k|\mathbf{M}_q)}{P(\mathbf{S}^a_k|\mathbf{M}_q)} \tag{5}$$

Entropy is a measure of uncertainty in the model. Intuitively, $H(a, b, q)$ measures the statistical predictability of the sensory input in receptive field b, given the input of the field a and the saccade q. As such, our hypothesis is that the structure implied by the existence of the visual field should manifest as a reduced entropy for particular triplets (a, b, q).

4 Results

Confirming our hypothesis, the physical existence of the visual field appeared as a highly predictable structure in the matrix T, presented in Fig. 3. For each saccade $\mathbf{M}_q$, a few blocks (a, b) exhibit a structure of higher predictability then others - they also exhibit lower values than other blocks in the entropy matrix. They correspond to pairs of receptive fields (a, b) for which the pre-saccade sensory state $\mathbf{S}^a_k$ predicts the post-saccade sensory state $\mathbf{S}^b_{k'}$ with significant accuracy, given $\mathbf{M}_q$. From an external point of view, this predictive structure corresponds to the shift of visual features between receptive fields induced by saccades; it capture the experience of a *visual field.*

Additional predictive structure also appears in blocks where the visual sensor's structure does not impose it. It is due to regularities in the environment captured by the predictive model, and disappear when the agent is set to explore randomly generated images (results not shown in this paper).

5 Conclusions

This work addresses the problem of autonomously discovering an agent's visual field according to the SMCT and predictive coding frameworks. It manifests as a set of regularities describing how different sensory inputs in different receptive fields correspond to the same visual feature, and how to move those features between receptive fields. Those regularities can be discovered while exploring the environment and captured in a predictive model. The approach has been applied to a simple simulated system coarsely inspired by the human retina. The physical structure of the sensor appears in the predictive model as highly probable sensorimotor transitions linking sensory inputs from different receptive fields and saccadic motor commands. The agent could for instance use this predictive model to perform visual search and foveation tasks [2].

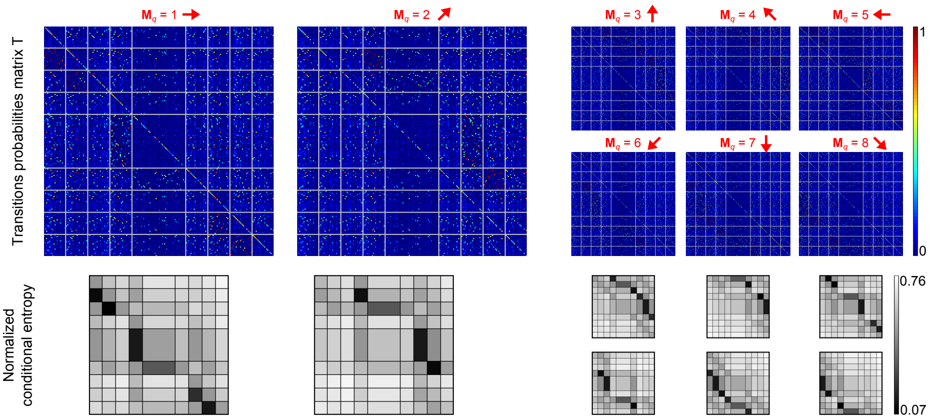

Fig. 3: Transitions probabilities matrix T and block normalized conditional entropy. For each saccade q, a few blocks in T display higher predictability than other blocks, corresponding to lower values in the entropy matrix. They correspond to pairs of receptive fields between which visual features shift during the corresponding saccade. The matrix T also contains additional structure corresponding to environmental structure captured by the predictive model.

By having two kinds of receptive fields (foveal and peripheral) with different properties, the simulation highlights the SMCT's claim that the sensory encoding of visual features is less relevant for perception than the way sensory inputs can be actively transformed. The predictive structure can indeed be captured regardless of the way visual features are encoded in each receptive field; it captures properties of the physical interaction of the agent with its environment.

Future work will focus on the development of a more sophisticated model, using neural networks to estimate the mapping between different receptive fields instead of the simplistic K-means/winner-takes-all combination. It should also be extended to continuous sensory inputs and motor outputs. Finally, a more detailed analysis of the environmental structure captured in matrix T will also be proposed. It will show how it can be used to define classes of visual features.

References

[1] J Kevin O'Regan and Alva Noë. A sensorimotor account of vision and visual consciousness. *Behavioral and brain sciences*, 24(05):939–973, 2001.

[2] Arvid Herwig and Werner X Schneider. Predicting object features across saccades: Evidence from object recognition and visual search. *Journal of Experimental Psychology: General*, 143(5):1903, 2014.

[3] Anil K Seth. A predictive processing theory of sensorimotor contingencies: explaining the puzzle of perceptual presence and its absence in synesthesia. *Cognitive neuroscience*, 5(2):97–118, 2014.

[4] Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva, and Antonio Torralba. Mit saliency benchmark.