Yielding self-perception in robots through sensorimotor contingencies

Pablo Lanillos, *Member, IEEE,* Emmanuel Dean-Leon, *Member, IEEE,* and Gordon Cheng, *Senior Member, IEEE*

***Abstract*—We address self-perception in robots as the key for world understanding and causality interpretation. We present a self-perception mechanism that enables a humanoid robot to understand certain sensory changes caused by naive actions during interaction with objects. Visual, proprioceptive and tactile cues are combined via artificial attention and probabilistic reasoning to permit the robot to discern between inbody and outbody sources in the scene. With that support and exploiting inter-modal sensory contingencies, the robot can infer simple concepts such as discovering potential "usable" objects. Theoretically and through experimentation with a real humanoid robot, we show how self-perception is a backdrop ability for high order cognitive skills. Moreover, we present a novel model for self-detection, which does not need to track the body parts. Furthermore, results show that the proposed approach successfully discovers objects in the reaching space, improving scene understanding by discriminating real objects from visual artefacts.**

***Index Terms*—Self-perception, Self-detection, Conceptual inference, Sensorimotor contingencies, Multi-modal integration, Embodied cognition.**

I. Introduction

During the last few years, roboticists have been looking for building machines that, whenever they are turned on for the first time, learn how to interact with the environment by means of their sensorimotor experience [1], [2], [3], [4]. We envisage that, as in humans, this mechanism is the key for adaptability, since they will be able to relearn when unexpected changes appear using the same machinery [5]. However, robots that learn from scratch are still a chimera. It is still unknown and even controversial how to get from sensor information to self-awareness, causality, semantic interpretation and agency attribution. In this sense, the ability of self-perception and the capacity to learn the body schema seems to be one of the core processes involved [4], [6].

Actually, recent evidence from psychology and neuroscience supports that self-perception enables self/other distinction and agency: sensorimotor temporal contingency is a key for discriminating inbody and outbody sources in four potential forms (contiguity, correlation, conditional probability and causal implication) [7]; the sensory consequences observed are tightly involved in the agency attribution of the actions [8]; sensorimotor understanding is a process learnt by interacting with ourselves and the environment [9]; and self and other's representation connects the sensorimotor map with more complex cognitive skills [10]. Accordingly, if we want to enable causality and semantic inference from sensory information, robots need to deploy multisensory binding based on contingency, self representation and agency attribution, while interacting with the environment. In this work, we present a grounding schema to go from sensors to abstract concepts. A representative example is the following: a robot sends the action to move the arm; proprioceptive and visual sensors measure changes due to action execution; then the robot can state: *this is my arm not only because I am sending the command to move it but also because I sense the consequences of moving it*.

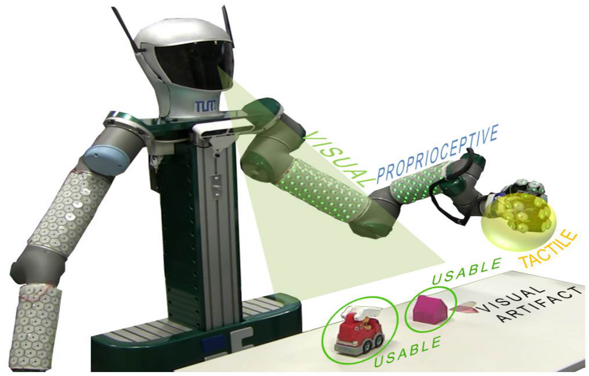


Fig. 1. The robot differentiates in/out body cues understanding the proprioceptive (artificial skin) and visual sensory contingencies and discovers objects by interacting with them.

Sensing the movement implies knowing what is the relationship between the action and the sensory effect. Hence, actions cannot be uncoupled from the sensing devices. This perceptual embodiment or self-perception depends on the nature of the sensors available in the robot. In this sense, the theory of SensoriMotor Contingencies (SMCs) [9] developed for human perception gives some insights about how this knowledge could be learnt.

Extending this idea for interacting with the environment, when the robot pushes an object in the scene the following argumentation can be formulated: the robot moves the arm and it makes contact with an object producing a change in the visual and tactile input; then the robot can interpret: *this object can be usable not only because I am moving the arm and the object is moving at the same time but also because I sense myself moving and touching and I understand the sensory consequence of my action.*

In the robotics literature, three main lines of research have

Institute for Cognitive Systems (ICS), Technische Universität München, Institute for Cognitive Systems, Arcisstrae 21 80333 München, Germany e-mail: p.lanillos@tum.de, dean@tum.de, gordon@tum.de

This work was supported by the Technische Universität München Foundation. Video to this paper: http://web.ics.ei.tum.de/~pablo/tcds2016.mp4

 DOI: http://dx.doi.org/10.1109/TCDS.2016.2627820

pursued self-perception. The first line is related to self and other recognition, where the robot objective is to distinguish between sensory cues promoted by itself or by other entity (human or another robot) [11], [12], [13]. Alternatively, other works have focused on tool use, where the robot performs a self extension of the body when learning how to use a tool [14], [15]. Finally, over the last few years, new approaches have been presented for adaptive exploration [16], active object learning [17] and object categorization [18]. The first two works include self-perceptual development as an important skill in their cognitive architectures. Conversely, the last one uses sensorimotor contingencies as the core to understand functional categories.

This paper formally presents the necessity of self-perception in robots, in the form of understanding the sensory consequences of their actions, in order to interpret high-level concepts while interacting with the environment. Specifically, we show that the robot perceives its own body and discovers potentially "usable"[1] objects, observing the causal effects that appear when the robot exerts actions.

Hence, we propose a novel hierarchical Bayesian computational model [19] that integrates proprioceptive and tactile cues from an artificial skin [20] with visual cues through bottom-up attention [21]. It is comprised of three layers of abstraction. The first two deal with self-detection by means of inter-modal contingencies, extending the works from [22], [11], [12] to avoid visual assumptions such as using markers. The third layer employs self-detection to enable conceptual interpretation such as objects discovery. To test the model we design an object discovery experiment based on the tapping or pushing setup proposed in [2] and [16] (Fig. 1).

Section II presents a general overview of self-perception in the literature, and a detailed comparison between self-detection methods and its relation with attention. Sec. III presents the in/out body perception problem, depicts the overall proposed solution and motivates its implementation in robots. Sec. IV presents how to transform sensor signals into meaningful cues (e.g., from accelerometers to arm moving). Sec. V describes the hierarchical probabilistic model to compute in/out body distinction and to find "usable" objects in the scene. Finally, sec. VI shows the experimental validation and results, sec. VII prompts the discussion and sec. VIII summarizes the work.

II. Self-perception in the literature

A. From embodied systems to self-perception

In 1991, the neurologist Antonio Damasio published his seminal work [23]. This was a major paradigm shift to understand humans' brain. He presented the *embodied mind*. In his own words: *"There is no such thing as a disembodied mind. The mind is implanted in the brain, and the brain is implanted in the body"*. Analogously, the principle of embodiment [3] in robotics states that there is no such a thing as a disembodied robot. The sensors and actuators are implanted in the body and the body is implanted in the robotic mind. This was originally investigated by Braitenberg in [24]. Afterwards, Brooks proposed a robot that used the sensors to move around without any internal representation of the environment [25]. Self-perception here is explained as sensor specific responses to stimuli. On the other hand, in developmental robotics [1], the robot is provided with a set of skills that will promote the *emergence* of more complex behaviours. For instance, the robot learns from simple exploration activities to complex manipulations [26]. Here, self-perception is presented as the mapping between motor actions and sensory inputs: *sensorimotor* approach [27]. The recent theory of SMC, originally presented by O'Regan and Noë [9] shows how this sensorimotor mapping explains human behaviours and why it is important for the emergence of awareness. It was developed to explain the conscious act of perception in humans as the mastering of the sensory consequences when performing an action. Sensorimotor contingencies are defined as the laws that govern the sensory changes according to the actions executed [28]. For instance, vision should be seen as a *"mode of exploration of the environment mediated by the knowledge of the sensorimotor contingencies"* [9, p. 943]. For an example of a computational model of the SMCs and its relation to previous psychological theories such as constructivism [29] please refer to [27]. This points towards a different kind of perception, where understanding the sensory changes promoted by the actions is the core for self-perception. It focuses on the modality-related changes when interacting, instead of the sensory input itself. Recent works in robotics that exploit SMCs are: [18] where the robot learns objects by their functionalities or [30] where a naive agent learns the notion of space in one dimension.

In this work, we present a mechanism inspired by the SMCs theory that follows the principle of embodiment. The robot starts with some knowledge about how to process the stimuli (e.g., visual artificial attention) and observing the multimodal sensory consequences of performing actions it changes its belief about the world, promoting the emergence of simple causality interpretation. To clarify, we define self-perception in robots as,

Definition 1. *Artificial self-perception is the machine ability to perceive its own body, i.e., the mastery of modal and intermodal contingencies of performing an action with a specific sensors/actuators body configuration.*

B. Own body distinction through self-perception

> *"The developmental sequence begins with learning a model of the robot's body"...[then]..."the robot can learn that certain behaviors can reliably cause an environmental object to move in the same way as some part of the robot's body"* [3, p. 129].

Some developmental roboticists have proclaimed that giving the robot the capacity to distinguish its own body is a key factor for interacting with the environment [6], [34], [17], [16]. To enable body distinction, we can either 1) learn the forward model (i.e., sensor output given an action, commonly $\hat{S}^{k+1}$ given S^k and A^k) by self-exploration and then compute the error between expected sensory outcome and the predicted one [13], [34] or 2) master the spatiotemporal contingencies

[1] In this paper, a usable object is considered as an object that can be moved by robot interaction - see sec. III.

TABLE I
OVERVIEW OF RELEVANT PERCEPTION SOLUTIONS FOR SELF-DETECTION AND INTERACTION WITH EMBEDDED ATTENTION. CUES ARE CODED AS: VISUAL(V), PROPRIOCEPTIVE(P), TACTILE(T) AND MOTOR COMMAND(M).

Work	cues	features	needs obj. identification/method	visual attention	self-detection/other	top-down	approach
Michel2004[22]	V, M	bounding box	*yes* / region similarity	motion by image differencing	✓/✗	saliency weights	Efferent-afferent initial delay classifier
Hikita2008[15]	V, P, T	10×10 attention map	*no* / -	Center-surround receptive fields [31]	✗/✗	-	Hebbian learning
Pitti2009[32]	V,P	saliency map	*no* / -	motion,intensity,colour	✗/✗	✗	Spikes neural network
Gold2009[11]	V, M	blobs	*yes* / area overlap	background subtraction and connected components	✓/✓	✗	Markov-process and Bayes classifier
Stoytchev2011[12]	V, M	coloured markers	*yes* / colour	colour-based segmentation	✓/✓	✗	Efferent-afferent initial delay discriminant thresholds
Nagai2011[13]	V, M	custom from optical flow	*no* / -	optical flow	✓/✓	✗	Hebbian learning
Rolf2014[33]	V, P	optical flow	assumed to be known	difference-of-gaussians	-	goal (implicit)	Reinforcement learning
Proposed	V, P, T	(level 1) $N \times N$ attention map (level 2) protoobjects	*no* / velocities estimation	protoobjects [21]	✓/✗	embodied model / saliency weights	Hierarchical Bayesian model

($S \times A \times t$ relation [35]) by observation and then use them to infer if the sensor stimuli have been produced by the robot body [11], [36]. In this sense, self-detection, a term borrowed from psychology [7], is defined as the prior process for the conscious act of self-recognition [12]. We review some of the most relevant works related to self-detection for object passive interaction[2], highlighting its connection with artificial attention and the sensory cues used. Besides, we have omitted all works dealing just with motor kinematics learning since we are more interested in analysing the sensory consequence of the action[3].

Table I describes some approaches for self-detection and object interaction. Although it is difficult to make a fair comparison due to the heterogeneity of the final addressed problems, we have tried to briefly summarise some interesting aspects. The approaches introduced by Michel et al. [22], Stoychev [12] and Pitti et al. [32] are based on temporal contingency although they use different methodologies. These works have psychology foundations on Watson theories [7] and studies on the visual cortex [32]. The idea is that causality, in form of motor-visual cues and temporal coherence, is the base for self-detection. However, despite the consistent temporal response to similar stimuli of the visual neurons, observation uncertainties should be treated. On the other hand, Gold et al. [11] have approached self-detection via probabilistic reasoning of the observed cues. We argue that causality, seen as the relation between the cause and the effect $A \rightarrow B$, cannot be uncoupled from the perception of the process (if A is observed then B becomes more plausible [37]). In practice, in robotic applications, visual segmentation algorithms usually have spatial-temporal incoherence of the output at different instants due to changing conditions (e.g., light changes).

Artificial attention must contribute to self-detection and object interaction processes [38]. Further information on attentional systems can be found in [39]. This is something that has been simplified using colour markers [12] or by means of connected components [11]. In both works, object tracking is crucial for the success of the method. Simplified models of attention such as difference-of-Gaussians or image-differencing have been used in [22], [33]. It is worth mentioning the work in [13] where the robot is able to learn the sensorimotor mapping to distinguish self and other using features extracted from optical flow. However, this mapping does not tackle objects interaction. A more interesting approach for sensory integration has been performed by Hikita et al. [15] where a biologically inspired attention system processes the visual information. Although that work is the most similar to the one proposed here, in terms of multimodal cues integration and attentional map approach, they only deal with tool extension and do not tackle causal implications of passive interaction with outbody objects.

Full cognitive architectures, which are not included in Table I, that include self-detection are presented in [17], [16]. The former uses appearance models to track the robot and the human hand. The visual segmentation that they propose converts the image into protoobjects using motion as the differentiating cue. Then visual features are extracted in order to further classify the robot arm. They assume that the space restrictions of the robot are known. Thus, not reachable regions are ignored. The later exploits visual and proprioceptive cues to discriminate inbody and outbody visual blobs. However, when performing interaction they assume a perfect tracking of the robot end-effector and the object.

III. PROPOSED MULTISENSORY MODEL

We enable self-perception and some simple causality inference in a robot that counts with visual and tactile sensing (Fig. 2(a)) by observing the inter-modal contingencies when an action is performed. For that purpose, a hierarchical probabilistic model that performs several abstractions is designed. In this section, the model is explained conceptually (sec. III-A) followed by an overview of the system in detail (sec. III-B) and the experimental design to show its expected behaviour (sec. III-C).

The overall process is as follows: first of all sensor signals are converted into meaningful cues represented by random

[2]First, there is a passive interaction where the robot arm is differentiated from the object even if there is contact [33] and afterwards, when the object is learnt as a tool, there is a body-extension where the object becomes a part of the self [15].

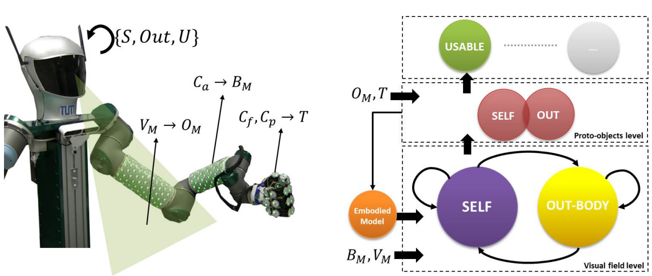
[3]This is a simplification of the self-perception and the proper relation between the action and the sensory change should be learned. However, in this paper for the sake of clarity we have restricted the study to analyse inter-modality sensory changes promoted by predefined simple actions.

TABLE II
FROM SENSORS TO CONCEPTS

Sensor signals	Meaningful cues	Inferred concepts
C_a skin accelerometer	B_M arm moving	S belongs to itself
V_M visual motion	O_M protoobject moving	Out is outbody
C_f, C_p skin force and proximity	T arm touching	U is "usable"

variables, e.g., accelerometer values are transformed into body movement (sec. IV); afterwards, those meaningful visual and proprioceptive cues are bound to enable in/out body discrimination and finally tactile cues are employed to infer "usability" of the objects in the scene (sec. V). All the processing is computed while the robot interacts with the environment.

A. Probabilistic model design



(a) Cues processed (b) Abstract model

Fig. 2. Meaningful signals and the proposed abstract model. Two layers of abstraction use the meaningful signals to discriminate between in/out body sources. Visual receptive field layer binds proprioceptive and visual cues to discern the robot arm in the field of view. Prior spatial or appearance information about the robot body also feeds this self-detection. The protoobject layer treats the information provided by the previous layer to visually classify attended objects. With this information more complex concepts such as discovering potentially usable objects can be inferred (top layer).

Instead of only studying the SMCs within one sensor modality we analyse the contingencies and relations of different sensor modalities in order to understand simple causality. Modality-related contingencies [18] are the ones that the robot uses to interpret that is moving when there is a specific change on the proprioceptive sensing (e.g., from the accelerometer information supplied by the skin C_a, the robot knows that the arm is moving B_M). Inter-modality contingencies are the ones that enable the robot to refine interpretation (i.e., visual moving objects O_M belong to the robot if there is a correlated change at the proprioceptive C_a and visual senses V_M). Note that in this work, we are more interested in the sensory consequences of the action than on the action itself. Figure 2(a) shows the robot sensors signals and the meaningful cues extracted from them, and Table II summarizes the meaning of each variable and the concepts extracted. While the sensor signals are represented by real values, the cues and the concepts are modelled as Bernoulli random variables.

The hierarchical probabilistic model is composed of three layers of inference (Fig. 2(b)):

- *Visual field self-detection*. The robot, binding visual (saliency map with motion) and proprioceptive (accelerometers) sensory contingencies, detects whether a pixel belongs to itself or not. This layer combines probabilistic inference grids with attentional maps [40]. To avoid tracking of robot parts 1st order dynamics (velocities) are learnt online.
- *Protoobject in/out body discrimination*. Bottom-up attention provides the most relevant proto-objects[4] in the scene. These attentional units are stored in the working memory. Using the visual field layer information the robot is able to classify whether the protoobject belongs to itself or it is an outbody source.
- *Object interaction*. It defines properties of the object based on the self-detection model and on the sensory consequences of the interaction. In this work, we have focused on discovering potentially usable objects. We define a "usable" object when the following causality is present (see sec. III-C): the robot moves the arm, it touches an object and the object visually moves.

This model can be seen as several layers that disambiguate sensory cues into concepts. Although the layered methodology has been also proposed in [17], they just present a segmentation pipeline. We face the problem with a more general approach using bottom-up artificial attention (see sec. III-B).

B. Overall system design

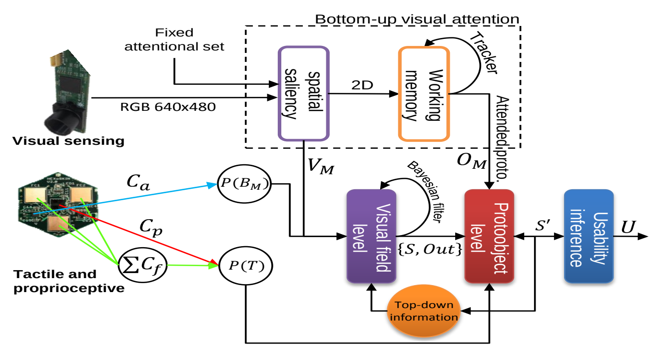


Fig. 3. System design and processing flow. A single camera and the body skin cells provide the necessary sensory information for the system.

Figure 3 details the sensors used and the information flow. On one hand, the image stream provided by the camera is processed by a visual attention system [21], that contributes with two main outputs: the saliency map with the protoobject relevance encoded in a 2D image and a list of attended protoobjects in the working memory. The spatial saliency is computed using several conspicuity maps that represent different features of the protoobjects [40]: color and intensity contrast, optical flow and color bias. These feature maps (2D images) are combined by weighted average using a fixed attentional set (weights), which in the case of having contextual information is used for top-down modulation. The visual moving cue V_M is obtained from the optical flow map. Furthermore, the working memory tracks a fixed number of protoobjects in the scene (humans are able to track around 5

[4] Protoobjects are pre-attentive units obtained by grouping structures (e.g., pixels) that have common characteristics [41].

objects simultaneously [21]). These protoobjects are potential objects to use. Their moving cue O_M is then obtained from the working memory.

On the other hand, the proprioceptive and tactile signals are extracted from each cell of the skin. The accelerometer information is transformed into the probability of moving the arm and the force and proximity values are transformed into the probability of touching. The position of each cell in the body is known. Thus, we know where the accelerometer changes are produced and obtain the arm moving cue B_M and touching cue T. Finally, all meaningful cues feed the probabilistic model that uses Bayesian filtering to update the new observations and to compute in/out body distinction and the "usability" of the objects.

Particularly, the robot used in this work counts with a vision system with one single See3CAMCU50 camera with 640x480 pixels definition working at 30fps. Although this configuration makes the robot 3D blind, objects discovering disambiguation is interestingly solved by interaction (see sec. VI-C). It also counts with artificial skin [20], [42] that provides proprioceptive and tactile information. Each cell of the skin has one 3-axis linear accelerometer, 3 capacitive force sensors and 1 infrared proximity sensor.

C. Experimental design

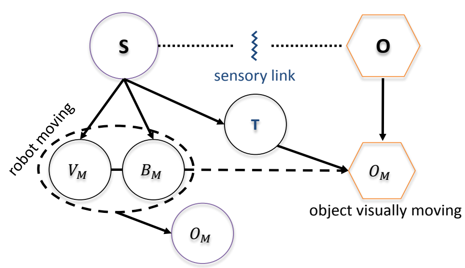


Fig. 4. System expected behaviour with in/out body distinction. The tactile link (T) promotes the causality between the body (B_M) and the moving object (O_M).

We formulate a theoretical experiment to explain how self-perception is involved in the robot interaction with the environment. The goal is to discover "usable" (U) objects by the following hypothesis: *an object is potentially usable if it is not part of myself and when I move my arm while touching, it also moves*[5]. The causality is the following: (1) the robot moves the arm (promoted action or cause), (2) it is touching (sensory link) and (3) the object moves (sensory consequence or effect). We assume that the robot already understands when something is moving in the visual field (V_M), when its own arm is moving (B_M) and when it is touching (T). Moreover, the visual attention system is able to provide relevant objects in the scene and when they are moving (O_M). We analyse the system behaviour in two different configurations.

- With self-detection skill, Fig. 4. It is able to discern the self S from the rest O. To do that it employs two variables: visual (V_M) and proprioceptive (B_M) movement. By observing the temporal sensory contingencies between both variables the robot makes in/out body distinction. Then, it can separate self and outbody moving objects (in Fig. 4, circular and hexagonal shapes respectively). Thus, when the robot moves the arm while touching the object, causality is generated: body B_M and object movement O_M are linked. Hence, the object is "usable". Finally, when the robot stops touching, the sensory link is no longer valid. If there has been enough interaction the outbody entity should remain as usable. In the case that the robot touches a part of its own body a double tactile link will appear. This produces a clear difference between self-touching and object interactions.
- Without self-detection skill. The robot is not able to distinguish if an object belongs to itself. Thus, we have only one type of object movement O_M (in this case, circles and hexagons belong to the same class; Fig. 4). Whenever the robot moves the arm and tactile interaction occurs, the sensory consequence of object visually moving appears. The defined causality will make all visual objects to be potentially usable, even body parts. However, when this tactile link disappears, causality vanishes and the objects do not have entity support to be maintained as usable.

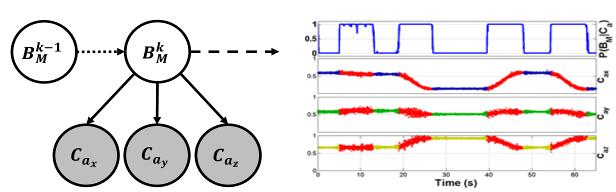
Intuitively, grounding self-distinction aids to discover objects by binding multimodal sensory consequences. This scenario will be discussed again using the real robot in sec. VI.

[5]A similar formulation using the object movement to interpret the effect promoted by the action has been expressed in [16, p. 13].

IV. Extracting meaningful cues from visual, proprioceptive and tactile sensors

To simplify the proposed approach we introduce in the robot enough knowledge to extract meaningful information from the sensor cues. For instance, we develop an algorithm to convert proprioceptive information (accelerometers) into moving cues. This means that the robot has mastery in modality-related sensorimotor contingencies for movement. Then we can focus on inter-modality contingencies for self-detection and object discovery. The levels of SMCs complexity proposed in [18] do not fit in our schema as the robot actually has to infer what an object is and what is not. Thus, here we only consider modality-related contingencies (changes in the signal depending on the agent action) and inter-modality contingencies (relation between changes in the signals from different sensors).

A. Proprioception information



(a) Model (b) Probability of moving

Fig. 5. Arm moving estimation using accelerometer information. The plot presents 70 seconds of the robot performing random movements. $P(B_M^k|C_a^k)$ is the blue line and the corresponding C_a signal movement is depicted in red.

We extract the moving information from the arm skin proprioception. Modelling it as a first order Markov process (Fig. 5) we get the probability of the arm being moving given the accelerometer observation $P(B_M|C_a)$. We only need to learn the likelihood distribution $P(C_a|B_M)$, which are the values that the accelerometer measures when the arm is moving. We bypass the classical solution of grabbing data and then learn the distribution by redefining the problem with a signal change detection function. Thus, we have the probability of the arm being moving given a change in the signal $P(B_M|change(C_a))$, where $change(C_a) = 1$ if signal changes and 0 otherwise.

This methodology helps to detect value changes while being robust to oscillations. It also simplifies the problem to a binary variable. We assume that the natural behaviour of the system is to maintain the current state (moving or static). Thus, it only depends on the likelihood of the observations[6]. Moreover, the three axes accelerometer variables (C_{a_i}) are assumed to be independent of each other. Then, the probability of the arm moving is,

$$P(B_M^k|change(C_a)) \propto \prod_i P(change(C_{a_i})|B_M^k)P(B_M^{k-1}) \quad (1)$$

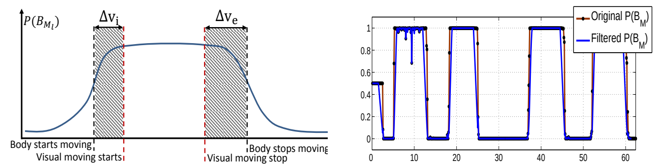
In order to calculate when the signal changes, we use an adapted online CUMSUM both-sides detector algorithm [43]. The method starts with an initial estimation of the signal value $\hat{\mu}, \hat{\sigma}^2$ computed from an initial set of samples with fixed size (window). Whenever the algorithm detects a change, the mean and the variance are updated using the new window samples[7].

B. Visual cues

Bottom-up artificial visual attention [21], [41] is used to extract salient protoobjects. First, it groups pixels that have similar characteristics (colour, intensity) and then a set of features are extracted and weighted (colour and intensity contrast, colour bias and optical flow) in order to evaluate their relevance. These salient regions are already meaningful representations of the scene. Thus, we have a set of visual objects, which contains the movement information $\{O_{M_1}, \cdots, O_{M_n}\}$. In order to enable self-detection before a body schema has been learnt, the protoobjects must be maintained through time. Assuming that all objects in the scene can be tracked over time is currently impracticable. We argue that attention should remain as a middleware process that manages objects in the scene and helps self-detection. Therefore, the protoobject saliency map is used as the visual input for the self-detection model.

C. Proprioceptive and visual temporal coherence

There is a time mismatch between the proprioceptive cue and the visual cue, as shown in Fig. 6(a). The delays found between cues are: the visual moving cue appears after the proprioceptive sensation of starting the movement and, contrary to [22], [12], visual movement cues stop before the current movement has been finished. This is happening because slow movements are totally inappreciable at the visual level. We



(a) Visual and proprioceptive delays (b) Pre and post average filter

Fig. 6. Efferent-afferent delays. The body moving meaningful signal is filtered to synchronize it with the visual moving stimuli, obtaining the blue line instead of the brown-dotted one.

have to learn the efferent-afferent delay as in [12]. As the proprioceptive rate computation ($60Hz$) is higher compared to the visual processing ($15fps$), the solution is implemented by means of a buffer that stores accelerometer measurements before outputting them. Then a filter delays (from static to moving) or anticipates (from moving to static) the sensor values to the self-detection model. Figure 6(b) shows the filtered signal (blue line) overlaid on the original signal (brown dotted line).

D. Tactile cues

Force and proximity sensors in each cell of the skin [42] provide information about the relative location (which part of the body) and the amount of force that the robot is performing. When touching an object the force sensor increases its value and we can extract the probability of touching something. However, in practice, we need to fuse proximity sensing to cope with very light objects by exploiting the saturation value of the sensor when touching. Defining C_{p_i} as proximity and C_{f_i} as force of each cell i, the probabilistic model of a set of cells to infer touching T is the following:

$$P(T|C_p \cup C_f) \propto 1 - \prod_i P(C_{p_i}|\overline{T})P(C_{f_i}|\overline{T})(1 - P(T)) \quad (2)$$

where $\overline{T}$ is a Bernoulli random variable that expresses no-touching. Force and proximity sensors contribute independently to obtain the probability of touching.

V. Hierarchical Probabilistic Model

The model presented in this section is designed taking into account the theoretical facts (sec. III) and implements the abstract scheme depicted in Fig. 2(b). In order to make the in/out body distinction, we need to keep tracking the robot arm and the objects in the scene. Within bioinspired approaches, the working memory is widely argued to be in charge of this process [21]. However, human studies show that we are able to store approximately five objects concurrently [44]. On the other hand, when dealing with self-detection we face other unsolved problem: which features of the body should be tracked allowing appearance changes? In essence, the attention system should be general enough to deal with the objects of the scene but also to help in the self-detection process. Preattentive stages cannot actually overcome with the hard task of self-detection, but in this rationale, several abstraction layers

[6]With slow dynamics we can assume $P(B_{M_l}^k|B_{M_l}^{k-1}) = diag(\mathbf{1})$.

[7]The window of input samples is maintained by means of a double linked queue and the new estimation is computed as follows: $\hat{\mu} = \text{mean(window)}$, $\hat{\sigma}^2 = \max(variance(\text{window}), \text{MIN\_VARIANCE})$.

and top-down body understanding (e.g., appearance and spatial sensorimotor model) could solve this issue. Here we provide a self-detection general solution when the robot parts cannot be tracked using two layers of abstraction: visual field self-detection and protoobjects in/out discrimination. Afterwards, an objects discovery model, which works on the top of the other layers, is presented.

A. *Visual field self-detection*

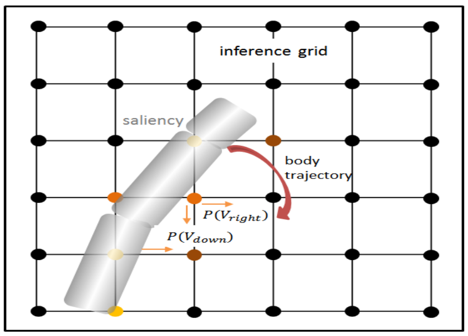


Fig. 7. Inference grid for self-detection. The posterior probability is computed by means of the estimated velocity and the observations generated by the artificial attention system (saliency). Then 1st order dynamics are estimated according to the posterior computation of the self.

In the case of not being able to track all objects coherently along the time and before having a self-representation of the body, we have to solve which parts of the scene belong to the robot by visual V_M and proprioceptive observations C_a. The method proposed here agrees with the fact that the body understanding should be constructed with the latest information available [3, p. 124]. We define the visual receptive field as a grid where we want to infer which node (i.e., the decimation of the pixel-wise image) belongs to the self and which does not (Fig. 7). We adapt Bayesian inference grids [19], [45] to estimate the probability of being self along the time . The prediction step is computed using the velocity in four directions (i.e., up, down, left, right). The probabilistic equation that governs self-detection is the following (see [45] for a detailed explanation),

$$P(S|\mathbf{V}, V_M, B_M) = \frac{\sum_{\mathbf{V}} P(V_M, B_M|S=1)\alpha_{self}}{\sum_{\mathbf{V}} [P(V_M, B_M|S=1)\alpha_{self} + P(V_M, B_M|S=0)\alpha_{out}]} \quad (3)$$

where $\mathbf{V}$ is the velocity in four directions k and α_{self} is computed as,

$$\alpha_{self} = (1-\epsilon)P(A_k)P(\mathbf{V}_k)\mathbf{T}_k P(S) + \epsilon P(A_k)\left[1 - P(\mathbf{V}_k)\mathbf{T}_k P(S)\right] \quad (4)$$

To compute α_{out} we set $\epsilon = 1 - \epsilon$. In order to make the equation clearer, we have defined $\mathbf{T}_k$ as the transition matrix that shifts all probabilities towards k direction and $P(A_k)$ as the prior probability of moving in the direction k. The term ϵ controls the amount of non-constant velocity in the visual input (with higher values the system becomes more reactive). Finally, we have to compute the posterior probabilities of the velocities:

$$P(\mathbf{V}_k) = P(V_M, B_M|S=1)\alpha_{self} + P(V_M, B_M|S=0)\alpha_{out} \quad (5)$$

With this method, we do not need tailor-made body objects tracking and we can still use the working memory to store relevant protoobjects given by the attention system, something important for discovering potentially usable objects.

B. *Protoobjects in/out discrimination*

We compute the probability of a protoobject being self depending on the visual field self-detection $P(S'|S)$. The protoobject pixels region can be $z = S$, when it is self, or $z = \overline{S}$, when it is outbody. The Bayesian update step under the 1st order Markov assumption is then,

$$P(S'|S) = \frac{P(z|S')P(S')}{P(z=S)P(S') + P(z=\overline{S})P(\overline{S'})} \quad (6)$$

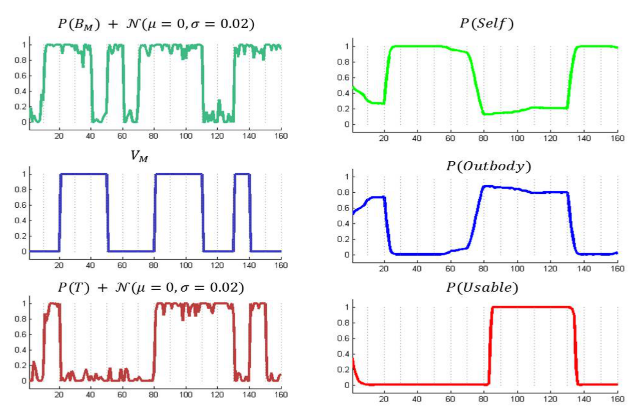
C. *Objects interaction: usable model*

The usable model is computed using the output of protoobjects in/out body discrimination model. The probability of being a usable object is then defined as:

$$P(U|B_M, T, O_M, S) \propto \underbrace{\cancel{P(B_M|U)}}_{\text{indep.}} \underbrace{P(T|B_M, U)}_{\text{uniform}} \underbrace{P(O_M|B_M, T, U)}_{\text{table}} \underbrace{P(S'|B_M, T, O_M, U)}_{\text{in/out proto discrim.}}$$
$$= \frac{1}{\eta} P(O_M|B_M, T, U)P(S'|U)P(S'|S) \quad (7)$$

where η is a normalization factor. Thus, the probability of being "usable" depends on the protoobject being self $P(S'|S)$, computed by Eq. 6, the likelihood of being self when "usable" $P(S'|U)$ and the likelihood of the being moving when the arm is moving and touching at the same time $P(O_M|B_M, T, U)$.

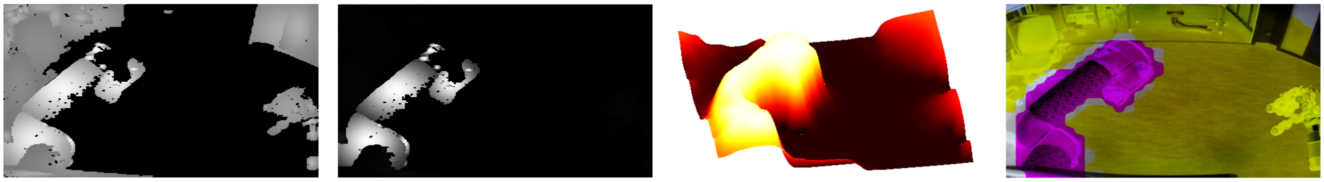
D. *Synthetic example of the model behaviour*



(a) Generated synthetic signals (b) Model output

Fig. 8. Example of the model behaviour *assuming a single correctly tracked object*. Synthetic signals are generated for $P(B_M)$, V_M and $P(T)$. $P(S)$ depends on both moving signals and $P(U)$ depends on the touch signal and on the probability of being self or outbody.

Figure 8 shows an example of the model for one object where the input signals are generated synthetically. Here we assume that we can coherently track the object O along the time. The probability of usability only rises when there is a sensory link (touching) and causality (arm moves $\rightarrow$ object moves) [4]. Furthermore, the probability of being usable decreases when the object belongs to the robot.



(a) saliency (b) moving protoobjects (c) $P(S)$ (d) self-detection

Fig. 9. Visual field self-detection combining visual attention and proprioceptive cues. From left to right, (a) saliency map (brighter represents more salient); (b) protoobjects moving; (c) probability of being self (whiter colour represents higher probability); and (d) self detection. Self (magenta), outbody (yellow) and unknown (greyscale).

E. Top-down influence

We define the most probable regions for the arm to appear $P(E)$ at the visual field level using any embodiment information. Here we have included two types of information: attended objects that are classified as self $P(O_i, S=1) > \kappa$ (protoobjects level) and a prior model defined by a smoothed mixture of Gaussians over a straight line that connects the left-bottom corner with the end-effector estimated location. This can be also used to include appearance or more complex prior spatial models of the robot body. The combined self-detection becomes $P(S) = wP(S) + (1-w)P(E)$, where $w \in (0, 1)$.

VI. Results

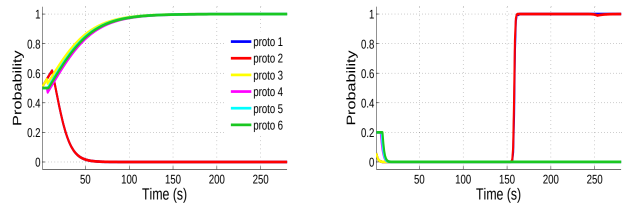
The self-perception mechanism is evaluated. First, we analyse self-detection and then we evaluate the integration of tactile cues for objects discovering. The experimental setup and some examples can be further explored in this video http://web.ics.ei.tum.de/~pablo/tcds2016.mp4. The parameters values, summarised in Table III, are fixed for all experiments and obtained by experimentation to satisfy the trade-off between computational costs and behaviour coherence.

TABLE III
DEFINED PARAMETER VALUES FOR THE EXPERIMENTS

Parameter	Notation	Value
grid decimation	-	5×5 pixels/node
Object moving when usable	$P(O_M \mid B_M, T, U=1)$	$(P_u, 0.08, P_u, Po, P_u, P_u, P_u, 1-Po)$
Object moving when not usable	$P(O_M \mid B_M, T, U=0)$	$(P_u, P_u, P_u, 1-Po, P_u, P_u, P_u, Po)$
uniform / outbody movement	P_u, P_o	$1/8, 0.15$
self being usable	$P(S \mid U)$	$(0.5, 0.5; 0.53, 0.47)$
velocity prior probability	P_A	$(0.1, 0.9/4, 0.9/4, 0.9/4, 0.9/4)$
non-constant velocity, top-down self thr.	ϵ, κ	$0.0001, 0.8$

A. Self-detection

An example of the self-detection inference is shown in Figure 9. The saliency and the protoobjects moving, outputted by the visual attention, are described in the first two columns. Fig. 9(c) shows the probability of each pixel belonging to the robot. Finally, Fig. 9(d) exhibits the moving arm being visually detected as own (magenta) and as outbody (yellow). As the arm is moving to the left, velocity estimation aids to spread the probabilities towards that direction. Note that larger area of unknown (grey) and self appear on the left side of the arm. Moreover, without velocity estimation self regions will appear on the arm trajectory.



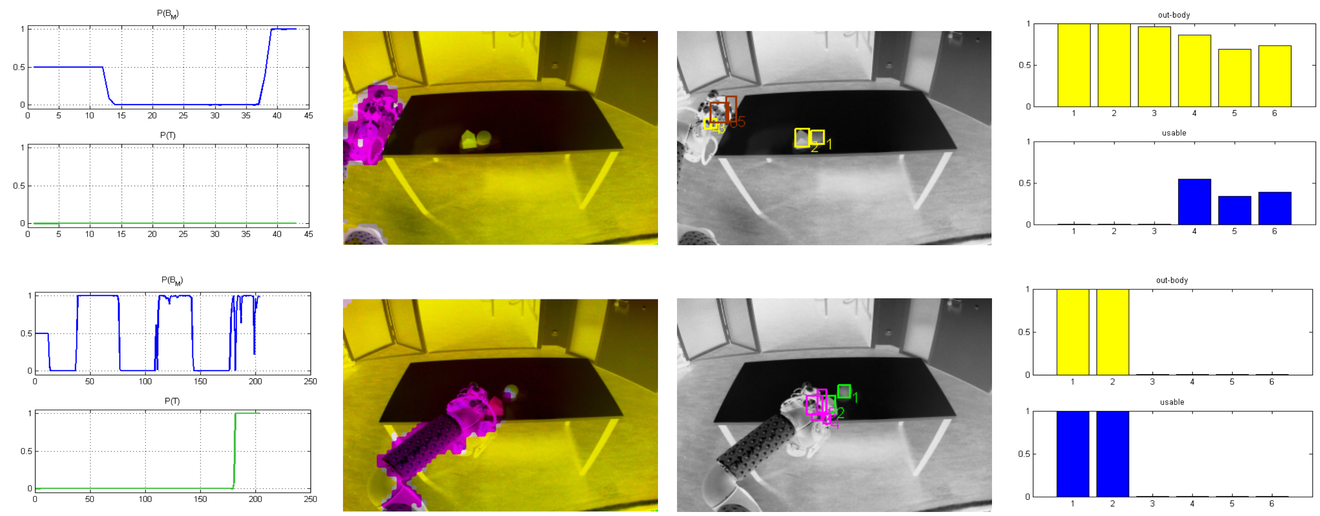
(a) Self (b) Usability

Fig. 11. Experimental model behaviour for the protoobjects inference. This plot corresponds to the two objects experiment of Fig. 10. (a) Protoobjects 1 and 2 are progressively differentiated as outbody and (b) when the tactile interaction occurs their probability of being usable increases drastically.

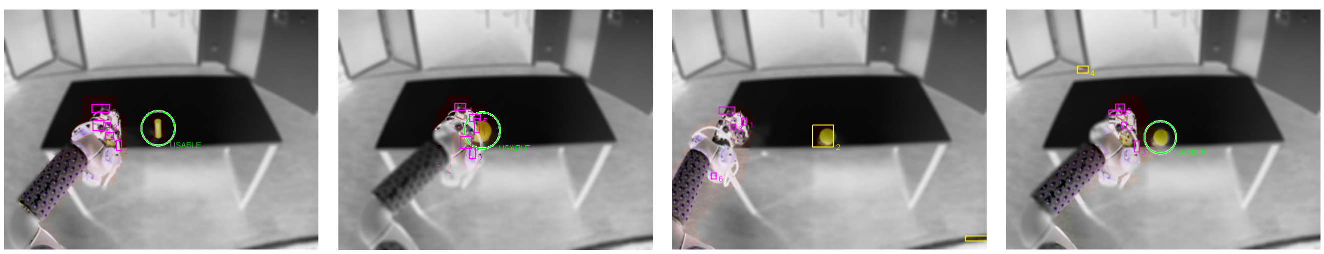
B. Object discovery

To evaluate the task of discovering potentially usable objects we design 10 experiments with the same initial configuration of the scene but with different objects (shapes and colours). The robotic arm has preprogrammed naive motion and it is not goal-directed. Figure 10 shows an example where the robot is able to distinguish two potential usable objects by interacting just with one of them. This happens because when it pushes one object, the other also moves. First, row represents the system 7 seconds after starting the experiment and the second row shows when the robot is pushing the object. After interaction (Fig. 10(h)), the robot has interpreted that object 1 and 2 are outbody and potentially usable. Furthermore, we analyse the distinct system behaviour with the same object but with different interaction time: (1) Fig. 10(k) short period of touching and failure, and (2) Fig. 10(l) longer interaction and success. The density of the interaction expressed by the number of meaningful events (e.g., inter-modal contingencies) determines how successful the grounding is. The failure case presents fewer events where tactile interaction, body moving and outbody object moving occurs at the same time.

Finally, the probabilities for each protoobject being self and usable during the experiment are shown in Fig. 11. Protoobjects 1 and 2 correspond to the objects in the table, which are gradually identified as outbody as the arm moves. When the robot touches the object and there is a movement sensory consequence on the visual sensing, the probability of being usable rises rapidly. This certitude of usability is maintained after the touch link disappears showing that there was enough interaction and the grounding was successful. This agrees with the simulated results previously presented.



(e) $P(B_M), P(T)$ (f) self-detection (g) working memory (h) $P(Out), P(U)$



(i) Cylinder (j) Big ball (k) Small ball, short interaction (l) Small ball, intense interaction

Fig. 10. Object discovery via visual attention and self-detection. First row describes the first stage of the system where the arm has started moving but self-features are still counted as outbody parts. Second row shows the robot pushing an object (which at the same time pushes another object). The system infers that both visual objects are potentially usable. First column displays the meaningful signals (arm moving and touching). Second shows self-detection at the visual low level. Third shows the attended protoobjects in the working memory (yellow - outbody, magenta - self, green - usable). Last column describes the probability of each protoobject (six potential objects) being outbody (yellow) and usable (blue). Last row depicts different experiments and the final result of the object discovery inference.

(a) Object illusion (b) Before touching (c) While touching (d) After touching

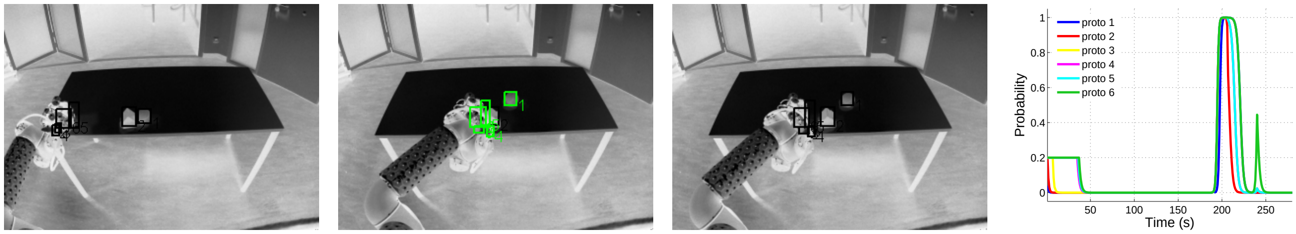
Fig. 12. Disambiguating usability in visual artefacts. One printed object (illusion) and a truck toy are placed over the table. Coloured pixels represent self-detection or protoobjects in the working memory. Colour code: self (magenta), inanimate (yellow), green (usable).

C. Scene disambiguation: illusion experiment

We show how tactile and visual sensory contingencies understanding disambiguates objects usability. The robot cannot know in advance which objects in the scene can be manipulable [16]. For that purpose, we print on a sticker an object that looks three-dimensional from the robot visual perspective (Fig. 12(a)). Then we put it on the table along with a real object (toy truck) that can be moved. The system is able to infer that the truck is usable and the illusion is not longer valid. Figure 12 shows different instants of the robot interacting. When tactile interaction begins some protoobjects are lost (tracking) and one of the new selections is also classified for a small period of time as a usable object. Afterwards, it converges towards self again and the robot correctly detects one real usable object. The visual artefact is finally established as an outbody non-usable region.

D. Disabling self-detection

We analyse how the system behaves without enabling self-detection. This experiment also supports the theoretical example introduced in sec. III. In the case of disabling self-detection, when the robot interact by touch, it infers that all



(a) Before touching (b) While touching (c) After touching (d) Protoobjects usability

Fig. 13. Discovering objects without self-detection. By disabling self-detection skill, the system is not able to correctly infer which objects are potentially usable. (b) When there is contact, all visual attended protoobjects, even the ones that belong to the robot, become usable for some instants (green rectangles). (c) After contact all protoobjects are labelled again as non-usable (black rectangles). (d) Shows the probability of each protoobject being usable during the experiment.

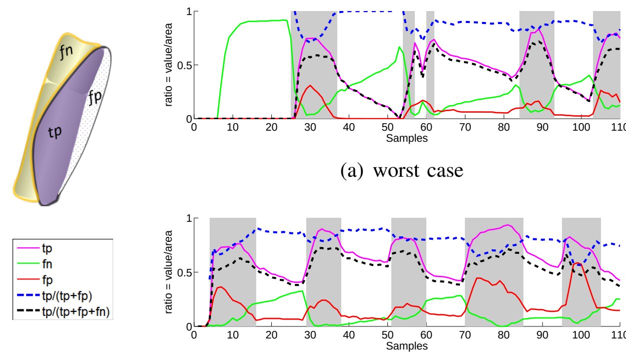
potential objects in the working memory are usable due to the sensory contingency of promoting movement (Fig. 13(b)). This means that, by construction, there is a sensory link with all objects in the scene and it cannot distinguish the difference of touching itself or an outbody object. However, when there is no contact all objects are labelled again as non-usable (Figs. 13(c) and 13(d)). Furthermore, in the case of disabling in/out body discrimination as well as the tactile sensory link, the objects in the visual field are inferred as unknown and non-usable during the whole experiment. This contributes with more evidence to support that in/out body distinction through self-perception is significant for scene understanding.

E. Quantitative study

TABLE IV
QUANTITATIVE ANALYSIS: SELF-DETECTION AND OBJECT DISCOVERY

Layer	Confusion Matrix				$\mathcal{E}$ (mov/¬mov)	Discovery
	expected \ detected (μ%)				$\mu \pm \sigma$	success/total%
visual field level		self	out	unknown	$0.595 \pm 0.107/0.483 \pm 0.083$	-
	self	60.59	12.96	26.45		
	out	**94.26**	1.10	4.64		
with top-down		self	out	unknown	$\mathbf{0.615} \pm 0.096/0.513 \pm 0.116$	-
	self	**74.87**	**2.11**	23.02		
	out	92.0	2.12	5.88		
proto-object level		self	out	unknown	-	92.31 %
	self	81.9	11.9	6.1		
	out	74.3	20.8	4.8		

We have also performed a quantitative analysis of self-detection and object discovery to show the accuracy of the proposed model. In order to perform the evaluation we have stored one image per second and then manually segmented the self region into a mask. This is then used as ground truth. Table IV shows the mean values for all experiments and Fig. 14 depicts two experiments in detail. The measure used to evaluate self-detection is the confusion matrix, where we show the percentage of correct pixel-wise classification in mean values. We also use a matching metric ($\mathcal{E} = tp/(tp{+}fp{+}fn)$), which is explained in the left side of Fig. 14. $\mathcal{E}$ is a conservative measure since we are computing the ratio of the correct detected area and all mismatches. True positives (tp) is the number of correct self pixels. False positives (fp) are pixels wrongly detected as self. False negatives (fn) are pixels wrongly classified as outbody. True negatives are not used since the area of outbody is too big and it does not represent an important indicator.



(a) worst case

(b) best case

Fig. 14. System evaluation example. The success ratio (measure/self_region_area) is shown along the time comparing the visual self-detection with the ground truth. Two measures are provided: precision ($tp/(tp + fp)$) and the matching metric ($\mathcal{E} = tp/(tp + fp + fn)$). Shaded grey regions determine when the robot is moving.

Confusion matrices in Table IV show that the visual layer is able to detect the 60% of the robot arm on average for all experiments. Using top-down influence we improve self-detection around 14% with low impact on the outbody inference. Finally, at the proto-object level, the out-body detection is failing 20% due to the first instants where the system is unable to induce the current class of the tracked object. The matching metric ($\mathcal{E}$) has been analysed when the robot is static and when it is moving. There is an increment in the performance when the robot moves. We can see this in detail in Fig. 14, where at shaded regions (when the robot is moving) the number of true positives rises. However, false positives also rise due to the delayed reaction of the Bayesian filter and the larger self detected area. Worse cases scenarios are produced when the observations (movement binding between the visual and proprioceptive cues) are not certain enough to maintain the region when the arm stops. This is perfectly shown in Fig. 14(a), where the ratio of tp drops when the robot stops (non-shaded regions). On the contrary Fig. 14(b) shows how to reduce this effect by means of top-down modulation.

The object discovery task success is presented in the last column of Table IV. 10 experiments with a total of 13 objects to be discovered exhibit one failure. In conclusion, the robot is able to sufficiently discover the object just by touching and analysing the posterior moving causality when it can discern outbody regions.

VII. Discussion

We have experimentally shown the advantages of self-perception in robots when interacting with the environment. On one hand, self-detection has been performed by correlating sensory consequences of visual and proprioceptive cues. On the other hand, objects discovery has been presented as the combination of the self-detection ability and the inter-modal sensory contingencies understanding. Despite the simple proposed model, the robot is able to successfully discover the objects in the scene by means of visual artificial attention and tactile cues. Thus, it shows the potential of building robots with self-perception understanding. This involves the ability to transform sensor information into meaningful cues, self-detection skill and the mastery of the multisensory binding. In this sense, low-level perception and top-down embodied representation seem to be the first stage for self-recognition and agency attribution.

By removing the ability to distinguish inbody from outbody sources (sec. VI-D) we have shown that the robot is not able to interpret correctly which objects are potentially usable. This means that even understanding modality-related contingencies the sensory tactile link makes incorrect causality inference. On the other hand, the mastery of inter-modal contingencies (multisensory fusion) disambiguates scene understanding. Visual artefacts are differentiated from real objects (sec. VI-C).

However, by developing a computational model, according to the theoretical facts, we take the risk of biasing the emergent behaviour. Thus, more robust evidence of what we are postulating here can be obtained by learning modal and inter-modal contingencies and analysing if they have embedded causality. Hence, we will study biologically inspired learning structures to obtain the same performance but improve generalization.

One interesting aspect that SMC theory tries to explain is the sensory substitution [9]. For instance, remote tactile sensing, where the sense of touch is substituted by other sensor modality such as sound echo. In this work, we have presented the idea of the tactile sensory link between the self and outbody objects, which generates causality. If we can replace this sensory link, by letting the robot learn the sensorimotor contingencies for that new sensor, the conceptual interpretation machinery will remain the same.

Interaction as an active process [46], [17], [47] has been insufficiently addressed. The robot should refine its knowledge or infer more complex causality by continuous interaction. This incremental development needs the generation of actions according to the sensory consequence, meaning that the robot does also have to learn the actions promoted by those changes in the sensor (e.g, within attention, the sensory changes trigger fixation actions).

VIII. Conclusion

We have presented a robotic self-perception mechanism that exploits multimodal contingencies in order to interpret simple causality. Three needed skills have been identified: (1) meaningful cues extraction, where the system transforms signal changes associated with each sensor into informative cues such as movement; (2) self-detection, where in/out body discrimination takes place; and (3) object interaction, where the robot employs low-level inferred knowledge and inter-modal contingencies to deploy conceptual interpretation of the scene.

The robot has successfully discovered potentially usable objects in the scene with a 92% of accuracy and it has been able to disambiguate a visual artefact from real objects. The experiments where the object has not been correctly identified indicate that the level of interaction, quantified as the density of the meaningful signals during the interaction time, is essential for the grounding of the knowledge. Thus, improvements can come by increasing the interaction time or by augmenting the number of meaningful signals or modalities.

Our self-detection method uses, as input, bottom-up protoobject artificial attention and proprioceptive cues. This approach is more general than other tailor algorithms that need prior knowledge of the body parts. The advantage of using visual artificial attention is that the same subsystem can be used for aiding self-detection and for object interaction. The quantitative analysis has shown that when the robot is moving, it is sufficiently capable of performing in/out discrimination. However, in absence of movement, the performance decreases considerably (12% on average) suggesting the necessity of including other features or top-down information. In this regard, self-detection with top-down has obtained the best accuracy at the pixel level (74.87%). This performance increases up to 81.9% when classifying the attended protoobjects.

We have shown, and validated in a real humanoid robot, that self-perception, taken as the understanding of the sensory consequences of performing an action, can improve the capabilities of the robot to interact in unknown environments. Then, the robot could speculate: ≪I understand the world because I understand my perception≫.

References

[1] M. Asada, K. F. MacDorman, H. Ishiguro, and Y. Kuniyoshi, "Cognitive developmental robotics as a new paradigm for the design of humanoid robots," *Robotics and Autonomous Systems*, vol. 37, no. 2, pp. 185–193, 2001.

[2] P. Fitzpatrick and G. Metta, "Grounding vision through experimental manipulation," *Phil. Trans. R. Soc. A: Mathematical, Physical and Engineering Sciences*, vol. 361, no. 1811, pp. 2165–2185, 2003.

[3] A. Stoytchev, "Some basic principles of developmental robotics," *IEEE Trans. Auton. Mental Develop.*, vol. 1, no. 2, pp. 122–130, 2009.

[4] P. Lanillos, E. Dean-Leon, and G. Cheng, "Multisensory object discovery via self-detection and artificial attention," in *Develop. Learning and Epigenetic Robotics (ICDL-Epirob), 2016 Joint IEEE Int. Conf. on*. IEEE, 2016.

[5] G. Cheng, *Humanoid Robotics and Neuroscience: Science, Engineering and Society*. CRC Press, 2014.

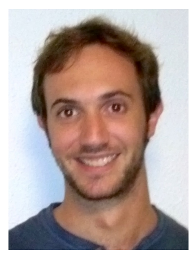
[6] M. Hoffmann, H. Marques, A. Arieta, H. Sumioka, M. Lungarella, and R. Pfeifer, "Body schema in robotics: a review," *IEEE Tran. on Auton. Mental Develop.*, vol. 2, no. 4, pp. 304–324, 2010.

[7] J. S. Watson, "Detection of self: The perfect algorithm," *Self-awareness in animals and humans: Developmental perspectives*, pp. 131–148, 1994.

[8] S.-J. Blakemore and C. Frith, "Self-awareness and action," *Current opinion in neurobiology*, vol. 13, no. 2, pp. 219–224, 2003.

[9] J. K. O'Regan and A. Noë, "A sensorimotor account of vision and visual consciousness," *Behavioral and brain sciences*, vol. 24, no. 05, pp. 939–973, 2001.

[10] A. Murata, W. Wen, and H. Asama, "The body and objects represented in the ventral stream of the parieto-premotor network," *Neuroscience research*, 2015.
[11] K. Gold and B. Scassellati, "Using probabilistic reasoning over time to self-recognize," *Robotics and autonomous systems*, vol. 57, no. 4, pp. 384–392, 2009.
[12] A. Stoytchev, "Self-detection in robots: a method based on detecting temporal contingencies," *Robotica*, vol. 29, no. 01, pp. 1–21, 2011.
[13] Y. Nagai, Y. Kawai, and M. Asada, "Emergence of mirror neuron system: Immature vision leads to self-other correspondence," in *Development and Learning (ICDL), IEEE Int. Conf. on*, vol. 2, 2011, pp. 1–6.
[14] A. Stoytchev, "Robot tool behavior: A developmental approach to autonomous tool use," Ph.D. dissertation, Georgia Institute of Technology, 2007.
[15] M. Hikita, S. Fuke, M. Ogino, T. Minato, and M. Asada, "Visual attention by saliency leads cross-modal body representation," in *Development and Learning, ICDL 7th IEEE Int. Conf. on*. IEEE, 2008, pp. 157–162.
[16] R. Saegusa, G. Metta, G. Sandini, and L. Natale, "Developmental perception of the self and action," *Neural Networks and Learning Systems, IEEE Tran. on*, vol. 25, no. 1, pp. 183–202, 2014.
[17] S. Ivaldi, N. Lyubova, A. Droniou, V. Padois, D. Filliat, P.-Y. Oudeyer, O. Sigaud *et al.*, "Object learning through active exploration," *Auton. Mental Develop., IEEE Tran. on*, vol. 6, no. 1, pp. 56–72, 2014.
[18] V. Hogman, M. Bjorkman, and D. Kragic, "Interactive object classification using sensorimotor contingencies," in *Intell. Robots and Systems (IROS), 2013 IEEE/RSJ Int. Conf. on*. IEEE, 2013, pp. 2799–2805.
[19] J. F. Ferreira and J. Dias, *Probabilistic approaches to robotic perception*. Springer, 2014.
[20] P. Mittendorfer and G. Cheng, "Humanoid multimodal tactile-sensing modules," *Robotics, IEEE Tran. on*, vol. 27, no. 3, pp. 401–410, 2011.
[21] P. Lanillos, J. F. Ferreira, and J. Dias, "Designing an artificial attention system for social robots," in *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*. IEEE, 2015, pp. 4171–4178.
[22] P. Michel, K. Gold, and B. Scassellati, "Motion-based robotic self-recognition," in *Intell. Robots and Systems, 2004.(IROS 2004). Proceedings. 2004 IEEE/RSJ Int. Conf. on*, vol. 3. IEEE, 2004, pp. 2763–2768.
[23] A. R. Damasio, "Descartes error: Emotion, rationality and the human brain," *New York: Putnam*, vol. 352, pp. 1061–1070, 1994.
[24] V. Braitenberg, *Vehicles: Experiments in synthetic psychology*. MIT press, 1986.
[25] R. A. Brooks, "Intelligence without representation," *Artificial intelligence*, vol. 47, no. 1, pp. 139–159, 1991.
[26] F. Guerin, N. Kruger, and D. Kraft, "A survey of the ontogeny of tool use: from sensorimotor experience to planning," *Auton. Mental Develop., IEEE Trans. on*, vol. 5, no. 1, pp. 18–45, 2013.
[27] T. Buhrmann, E. A. Di Paolo, and X. Barandiaran, "A dynamical systems account of sensorimotor contingencies," *Front. Psychol*, vol. 4, no. 285, pp. 10–3389, 2013.
[28] J. K. ORegan, "The explanatory status of the sensorimotor approach to phenomenal consciousness, and its appeal to cognition," in *Contemporary Sensorimotor Theory*. Springer, 2014, pp. 23–35.
[29] J. Piaget, M. Cook, and W. Norton, *The origins of intelligence in children*. International Universities Press New York, 1952, vol. 8, no. 5.
[30] A. V. Terekhov and J. K. O'Regan, "Learning abstract perceptual notions: The example of space," in *Develop. Learning and Epigenetic Robotics (ICDL-Epirob), IEEE Int. Conf. on*, 2014, pp. 368–373.
[31] L. Itti, N. Dhavale, and F. Pighin, "Realistic avatar eye and head animation using a neurobiological model of visual attention," in *Optical science and technology, SPIE's 48th annual meeting*. International Society for Optics and Photonics, 2004, pp. 64–78.
[32] A. Pitti, H. Mori, S. Kouzuma, and Y. Kuniyoshi, "Contingency perception and agency measure in visuo-motor spiking neural networks," *Auton. Mental Develop., IEEE Trans. on*, vol. 1, no. 1, pp. 86–97, 2009.
[33] M. Rolf and M. Asada, "Autonomous development of goals: From generic rewards to goal and self detection," in *Develop. and Learning and Epigenetic Robotics (ICDL-Epirob), 2014 Joint IEEE Int. Conf. on*. IEEE, 2014, pp. 187–194.
[34] G. Schillaci, V. V. Hafner, B. Lara, and M. Grosjean, "Is that me?: sensorimotor learning and self-other distinction in robotics," in *Proceedings of the 8th ACM/IEEE Int. Conf. on Human-robot interaction*. IEEE Press, 2013, pp. 223–224.
[35] R. Martın-Martın, A. Sieverling, and O. Brock, "Estimating the relation of perception and action during interaction," in *International Workshop on Robotics in the 21st century: Challenges and Promises*, 2016.
[36] P. Lanillos and G. Cheng, "Robots with self-perception: objects discovery and scene disambiguation using visual, proprioceptive and tactile cues correlation during interaction," in *International Workshop on Robotics in the 21st century: Challenges and Promises*, 2016.
[37] E. T. Jaynes, *Probability theory: the logic of science*. Cambridge university press, 2003.
[38] J. H. Reynolds and R. Desimone, "The role of neural mechanisms of attention in solving the binding problem," *Neuron*, vol. 24, no. 1, pp. 19–29, 1999.
[39] J. F. Ferreira and J. Dias, "Attentional mechanisms for socially interactive robots–a survey," *Autonomous Mental Development, IEEE Transactions on*, vol. 6, no. 2, pp. 110–125, 2014.
[40] P. Lanillos, J. F. Ferreira, and J. Dias, "Multisensory 3d saliency for artificial attention systems," in *3rd Workshop on Recognition and Action for Scene Understanding (REACTS), 16th International Conference of Computer Analysis of Images and Patterns (CAIP)*, 2015, pp. 1–6.
[41] R. Marfil, A. J. Palomino, and A. Bandera, "Combining segmentation and attention: a new foveal attention model," *Frontiers in computational neuroscience*, vol. 8, 2014.
[42] F. Bergner, E. Dean-Leon, and G. Cheng, "Event-based signaling for large-scale artificial robotic skin - realization and performance evaluation," in *Intell. Robots and Syst. (IROS), IEEE/RSJ Int. Conf. on*, 2016.
[43] E. Page, "Continuous inspection schemes," *Biometrika*, vol. 41, no. 1/2, pp. 100–115, 1954.
[44] C. Bundesen, T. Habekost, and S. Kyllingsbæk, "A neural theory of visual attention and short-term memory (ntva)," *Neuropsychologia*, vol. 49, no. 6, pp. 1446–1457, 2011.
[45] P. Bessière, C. Laugier, and R. Siegwart, *Probabilistic reasoning and decision making in sensory-motor systems*. Springer, 2008, vol. 46.
[46] A. Ude, D. Omrčen, and G. Cheng, "Making object learning and recognition an active process," *International Journal of Humanoid Robotics*, vol. 5, no. 02, pp. 267–286, 2008.
[47] M. Kaboli, R. Walker, and G. Cheng, "Re-using prior tactile experience by robotic hands to discriminate in-hand objects via texture properties," in *2016 IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2016.



Pablo Lanillos is a postdoctoral researcher at the Institute for Cognitive Systems granted with a TUM Foundation Fellowship. He received his PhD degree in computer science engineering (2013) from Complutense University of Madrid, Department of Computer Engineering and Automatic Control, Spain. As a Ph.D. candidate, he has been a visiting researcher at the Massachusetts Institute of Technology, the École Polytechnique Fédérale de Lausanne and the Australian Centre for Field Robotics. Prior to joining TUM, he was involved in CASIR project at the Institute of Systems and Robotics, University of Coimbra. His research interests include probabilistic decision making in multisensory robots, artificial attention, self-awareness and active exploration.

Emanuel Dean-Leon holds a position as senior researcher at the Institute for Cognitive Systems at the Technical University of Munich (TUM) since 2013. He studied Mechatronics at the CINVESTAV in Mexico, where he received his Ph.D. in 2006. In 2009 He performed a postdoctoral research in the Robotics and Embedded System Department at TUM. His research interests include robot modeling, low-level control design/implementation, sensor fusion, human-robot interaction/collaboration and cognitive systems.

Gordon Cheng is the Professor and Director of the Chair for Cognitive Systems, Technical University Munich. Formerly (2002,2008), he was the Head of the Department of Humanoids Robotics and Computational Neuroscience, ATR Computational Neuroscience Laboratories, Kyoto. He received a Ph.D. in Systems Engineering (2001) from the Australian National University. he was also the Managing Director of the company G.T.I. Computing in Australia. His research interests include humanoid robotics, cognitive systems, brain-machine interfaces, active and human bio-mimetic vision, human-robot interaction and robot navigation.