

Multisensory Object Discovery via Self-detection and Artificial Attention*

Pablo Lanillos, Emmanuel Dean-Leon and Gordon Cheng

Abstract

We address self-perception and object discovery by integrating multimodal tactile, proprioceptive and visual cues. Considering sensory signals as the only way to obtain relevant information about the environment, we enable a humanoid robot to infer potential usable objects relating visual self-detection with tactile cues. Hierarchical Bayesian models are combined with signal processing and protoobject artificial attention to tackle the problem. Results show that the robot is able to: (1) discern between inbody and outbody sources without using markers or simplified segmentation; (2) accurately discover objects in the reaching space; and (3) discriminate real objects from visual artefacts, aiding scene understanding. Furthermore, this approach reveals the need for several layers of abstraction for achieving agency and causality due to the inherent ambiguity of the sensory cues.

I. INTRODUCTION

In the last twenty years roboticist are seeking to build 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]. We envisage that, as in humans, this is the key for adaptability as they will be able to relearn when unexpected changes appear using the same machinery [4]. However, robots that learn from scratch are still a chimera. The difficulties do not only arise from the computational models and their limitations but also from the nature of the sensory cues. Moreover, it is still unknown and even controversial how to get from sensor information to self-awareness, causality, semantic interpretation and agency attribution.

Recent works related to self-perception in psychology and neuroscience give some insight about the potential paths to follow. Sensorimotor temporal contingency is a key for discriminating inbody and outbody sources in four potential forms (contiguity, correlation, conditional probability and causal implication) [5]; the sensory consequences observed are tightly involved in the agency attribution of the actions [6]; sensorimotor understanding is a process learnt by interacting with ourselves and the environment [7]; self and other's representation connects the sensorimotor map with more complex cognitive skills [8]. 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 show a novel robotic approach to go from sensors to abstract concepts. Instead of using the motor commands as the cue, we exploit the multimodal sensory consequence of the action. A representative example is the following: a robot sends the action to move the arm; body and visual sensors measure changes due to new stimuli; then the robot can state: *this is my arm not only because I am sending the command to move but also because I sense the consequences of moving it.*

One of the biggest challenges in self-perception is to arrive from bottom-up attention to coherently identify the self body parts in the scene as the time passes. Within bioinspired approaches the working memory is argued to be in charge of the objects tracking [9]. However, when dealing with self-detection we do not know which features of the body should be tracked. Hence, the attention system should be general enough to deal with the objects of the scene but also to help in the self-detection process.

This paper first addresses self-detection from the multisensory perception point of view, extending the works from [10], [11] to avoid visual assumptions such as placing markers or objects tracking. Secondly, we employ self-detection to enable object discovery. For that purpose, we replicate the taping experiment (Fig. 1(a)) proposed in [2]. The only considered way to get information is through sensor signal processing. Thus, we integrate proprioceptive and tactile cues from an artificial skin [12] with visual cues through bottom-up attention [9] to provide the robot self-detection and simple objects interaction causality skills. We show that without any prior knowledge of the scene the robot perceives its own body and discovers potentially usable objects by simple causal effects that it can promote to those objects. We propose a hierarchical Bayesian computational model [13] comprised of three layers of abstraction. This model is capable of transforming sensory information into "concepts" taking into account observation uncertainty.

*(c) 2016 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other users, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works for resale or redistribution to servers or lists, or reuse of any copyrighted components of this work in other works. Pre-print version of the submitted paper to the sixth joint IEEE International Conference Developmental Learning and Epigenetic Robotics (ICDL-Epirob 2016). The final paper can be found at the proceedings: <http://ieeexplore.ieee.org/servlet/opac?punumber=1001919>.

All authors are with the Institute for Cognitive Systems (ICS), Technische Universität München, Institute for Cognitive Systems, Arcisstrae 21 80333 München, Germany {p.lanillos, dean, gordon}@tum.de.

This work was supported by the Technische Universität München Foundation. Video to this paper: <http://web.ics.ei.tum.de/~pablo/sawicdl2016.mp4>

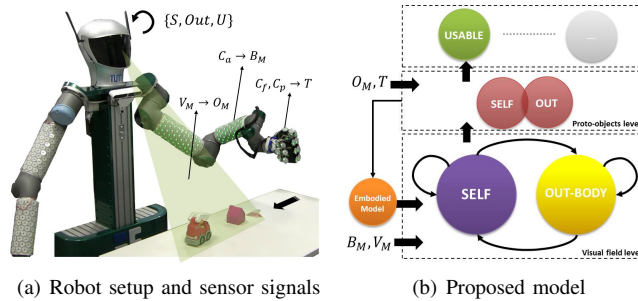


Fig. 1. The robot differentiates in/out body cues using proprioceptive and visual sensing and discovers objects by moving them. It counts with vision system and artificial skin with accelerometers, force and proximity sensors.

A. Multisensory-based approach

The robot counts with artificial skin [12] that provides proprioceptive (accelerometers, C_a) and tactile (force C_f and proximity C_p sensors) information, as well as a vision system (i.e., one See3CAMCU50 camera). The robot signals and concepts extracted, and the proposed approach in abstract form is depicted in Fig. 1. It is composed of three layers of inference: visual field self-detection, protoobject self-detection and object interaction. The system can be seen as several layers that disambiguate sensory cues into concepts. First, meaningful visual and proprioceptive signals are bound to enable in/out body cues discrimination. Then, the protoobjects provided by visual attention and stored in the working memory are classified as self or outbody using lower layer information. Finally, the robot interacts with the potential outbody objects to infer “usability” using tactile cues. Besides, this model also provides an easy way to include top-down modulation as a prior knowledge (e.g., an embodied or appearance model).

- *Visual field self-detection.* The robot, binding visual (saliency map with motion) and proprioceptive (accelerometers) cues, detects whether a pixel belongs to itself or not. This layer combines probabilistic inference grids with attentional maps [14]. To avoid the tracking of the robot parts 1st order dynamics (velocities) are learnt online.
- *Protoobject in/out body discrimination.* Bottom-up attention provides the most relevant regions of the scene. These attentional units, called protoobjects, 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 the sensory consequences of the interaction. We have focused on discovering potential usable objects. We define a usable object when the following causality appears: (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).

II. REVISITING SELF-DETECTION AND ATTENTION

Several authors have expressed robotic self-embodiment as the ground skill for higher level interaction with the environment [3], [15]. The first stage in the developmental process is learning the body model and then, the robot can learn behaviours that it can promote to objects [3]. However, where does perception take part? Causality inference, as a high order skill, arrives when we are able to attribute the agency of the actions [6] and this definitely undergirds on the understanding of the sensorimotor response [7]. The mechanism behind the construction of the own representation in humans is still unknown. Evidence point towards the parieto-premotor network where own and other’s body representation connect low level sensorimotor skills with high-order cognitive functions [8]. In this configuration there are self-dominant and other-dominant neurons [8]. This shows that although the self-other representation shares the same machinery there are some regions dedicated to each aspect.

Approaches introduced by Stoychev [11] and Pitti et Al. [16] are supported on temporal contingency although they use different methodologies. This has psychology foundations on Watson theories [5] and studies on the visual cortex [16]. The idea is that causality, in the form of motor-visual cues and their temporal coherence, is the base for self-detection. However, despite the consistent spatio-temporal response to similar stimulus of the visual neurons there should be an implicit treatment of the observation uncertainties. On the other hand, Gold et al. [10] 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 [17]). 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. This is something that has been simplified using colour markers [11] or by means of connected components [10]. In both works, object tracking is crucial for the success of the method. Other models of attention that have been used are difference-of-gaussians or image-differencing.

It is worth mentioning the work in [18] where the robot is able to learn the sensorimotor mapping to distinguish self and other using features extracted from optical flow. However, this mapping do not tackle objects interaction. A more interesting approach for sensory integration has been performed by Hikita et al. [19] where a biologically inspired attention system processes the visual information. Although [19] 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 they do not tackle causal implications of passive interaction with outbody objects.

III. EXTRACTING MEANINGFUL CUES FROM VISUAL, PROPRIOCEPTIVE AND TACTILE CUES

TABLE I
FROM SENSORS TO CONCEPTS

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

We extract meaningful signals from the sensor information for later use inside the models (Table I). Sensor signals C_a and C_f are real measurements from the sensors and the meaningful cues are modelled as Bernoulli random variables¹.

A. Proprioception information

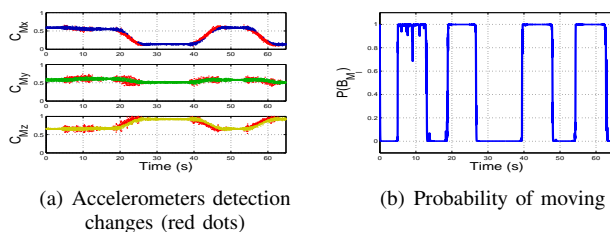


Fig. 2. Left arm moving estimation using accelerometer information $P(B_M^k|C_a^k)$. The arm is moving depending on the previous knowledge and the new observation from the sensors. The data plotted correspond to 70 seconds of the robot performing random movements.

First we extract the moving information from the left arm skin proprioception. Modelling it as a first order Markov process we get the probability of the left arm being moving (Fig. 2) 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. Alternatively, 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 left arm being moving given a change on the signal $P(B_M|change(C_a))$ where, $change(C_a) = 1$ if the signal changes.

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 observations². Moreover, accelerometer variables (C_{a_i}) are assumed to be independent to each other. Then the probability of the left 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 [20]. 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 samples³.

¹ $P(B_M = 1) = p$ where $\{p \in (0, 1) \in \mathbb{R}\}$.

²With slow dynamics we can assume $P(B_{M_i}^k|B_{M_i}^{k-1}) = diag(\mathbf{1})$.

³The window of input samples is maintained by means of a double linked queue and the new estimation is computed as follows: $\hat{\mu} = \text{mean}(\text{window})$, $\hat{\sigma}^2 = \max(\text{variance}(\text{window}), \text{MIN.VARIANCE})$.

B. Visual cues

Bottom-up artificial visual attention [21], [9] is used as the preattentive stage to extract salient protoobjects (see section V). 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 $\{O_{M_1}, \dots, O_{M_n}\}$. For enabling self-detection before any body schema has been learnt, the protoobjects must be maintained over time. However, tracking all objects in the scene is impracticable. We argue that attention should remain as a middleware process [9] 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. Tactile cues

Force and proximity sensors in each cell of the skin [22] 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:

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

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

IV. HIERARCHICAL BAYESIAN MATHEMATICAL MODEL

First we describe the mathematical model assuming a perfect tracking of the object to show its correctness. Afterwards, we provide the solution when the robot parts cannot be tracked due to segmentation failures.

A. Assuming features tracking

This section studies the theoretical model when the object can be correctly identified through time. We define a set of classes that should be inferred by the *self-detection model* $\zeta = \{S, Out\}$ ⁴, where self (S) is defined as the parts of the scene that belong to the robot and outbody (Out) defines any region of the scene that do not correlate with the robot movement. The probability of being S given the meaningful cues is (protoobject moving O_M , body moving B_M and touch T):

$$P(\zeta = S|O_M, B_M, T) \propto (1 - P(T)) \sum_{B_M} P(O_M|B_M, S)P(S) + P(T)P(S) \quad (3)$$

The object class i is obtained by normalization and computing the maximum *a posteriori* [10]: $\arg \max_i P(\zeta = i) = \arg \max_i P(\zeta_i) / \sum_{\zeta} P(\zeta)$.

The *usable model* is computed using the output of self-detection model. The joint probability is defined as:

$$P(U, B_M, T, O_M, S) = \underbrace{P(B_M|U)}_{\text{indep.}} \underbrace{P(T|B_M, U)}_{\text{uniform}} \underbrace{P(O_M|B_M, T, U)}_{\text{table}} \underbrace{P(S|B_M, T, O_M, U)}_{\text{low level}} \quad (4)$$

by knowing $P(S|B_M, T, O_M, U)$ using the visual-field layer the probability of being a usable object at the proto-object level is then simplified to:

$$P(U|B_M, T, O_M, S) = \frac{1}{\eta} P(O_M|B_M, T, U)P(S|U)P(S|B_M, T, O_M) \quad (5)$$

where η is a normalization factor.

Figure 3 shows an example of the theoretical model for one object where the input signals are generated synthetically. The probability of usability only rises when there is a sensory link (touching) and causality (arm moves \rightarrow object moves). Furthermore, the probability of being usable decreases when the object belongs to the robot.

⁴Note that in this case $P(S) = 1 - P(Out)$.

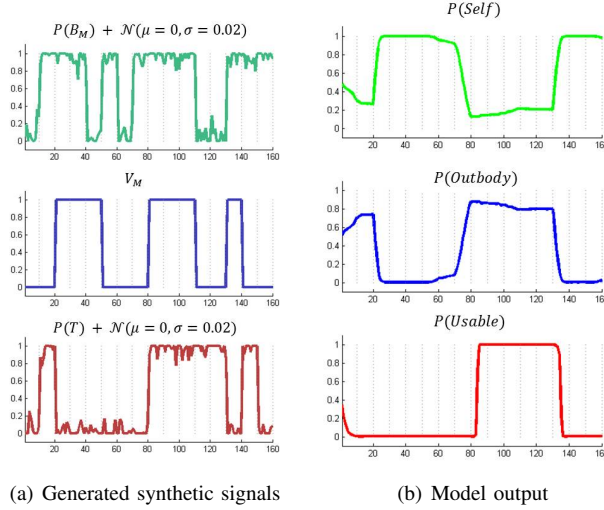


Fig. 3. Example of the model behaviour assuming a single correctly tracked object. Synthetic signals are generated for $P(B_M)$, V_M and $P(T)$. $P(S)$ depends on both moving signals and $P(U)$ depends on the touch signal and on the probability of being self or outbody.

B. Self-detection without features tracking: inference grid and velocities estimation

In the case of not being able to track all objects coherently over 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 . We define the visual receptive field as a grid where we want to infer which node (i.e., decimation of the pixel-wise image) belongs to the self and which does not. We adapt Bayesian inference grids [13], [23] to estimate the probability of being self along the time. The prediction step is computed using the velocities in four directions $W_{d \in (0,3)}$ (i.e., up, down, left, right). The probabilistic equation that governs self-detection is the following (see [23] for a detailed explanation),

$$P(S|W, V_M, B_M) = \frac{\sum_{d=0}^{d=3} P(V_M, B_M|S=1)\alpha_{self(d)}}{\sum_{d=0}^{d=3} P(V_M, B_M|S=1)\alpha_{self(d)} + P(V_M, B_M|S=0)\alpha_{out(d)}} \quad (6)$$

where $\alpha_{self(d)}$ is computed in every velocity direction d as,

$$\alpha_{self(d)} = (1 - \epsilon)P(A_d)P(W_d)\mathbf{T}_dP(S) + \epsilon P(A_d)[1 - P(W_d)\mathbf{T}_dP(S)] \quad (7)$$

Analogously, to compute $\alpha_{out(d)}$, we set $\epsilon = 1 - \epsilon$ in Eq. 7. We have defined \mathbf{T}_d as the transition matrix that shifts all probabilities towards d direction and $P(A_d)$ as the prior probability of moving in that direction. The term ϵ controls the amount of non-constant velocity in the visual input (with higher values the system becomes more reactive). Afterwards we compute the posterior probabilities of the velocities to be used in the next instant:

$$P(W_d) = P(V_M, B_M|S=1)\alpha_{self(d)} + P(V_M, B_M|S=0)\alpha_{out(d)} \quad (8)$$

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 potential usable objects.

C. 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 $w \in (0, 1)$.

V. RESULTS

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 the following video <http://web.ics.ei.tum.de/~pablo/sawicd12016.mp4>. The parameters value, obtained empirically, for all executions are summarized in Table II.

TABLE II
DEFINED PARAMETERS VALUE FOR THE EXPERIMENTS

Parameter	Notation	Value
grid decimation	-	5×5 pixels/node
Object moving when usable	$P(O_M B_M, T, U = 1)$	$(P_u, 0.08, P_u, P_o, P_u, P_u, P_u, 1 - P_o)$;
Object moving when not usable	$P(O_M B_M, T, U = 0)$	$(P_u, P_u, P_u, 1 - P_o, P_u, P_u, P_u, P_o)$
uniform, outbody mov	P_u, P_o	1/8, 0.15
self being usable	$P(S U)$	(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 thr.	ϵ, κ	0.0001, 0.8

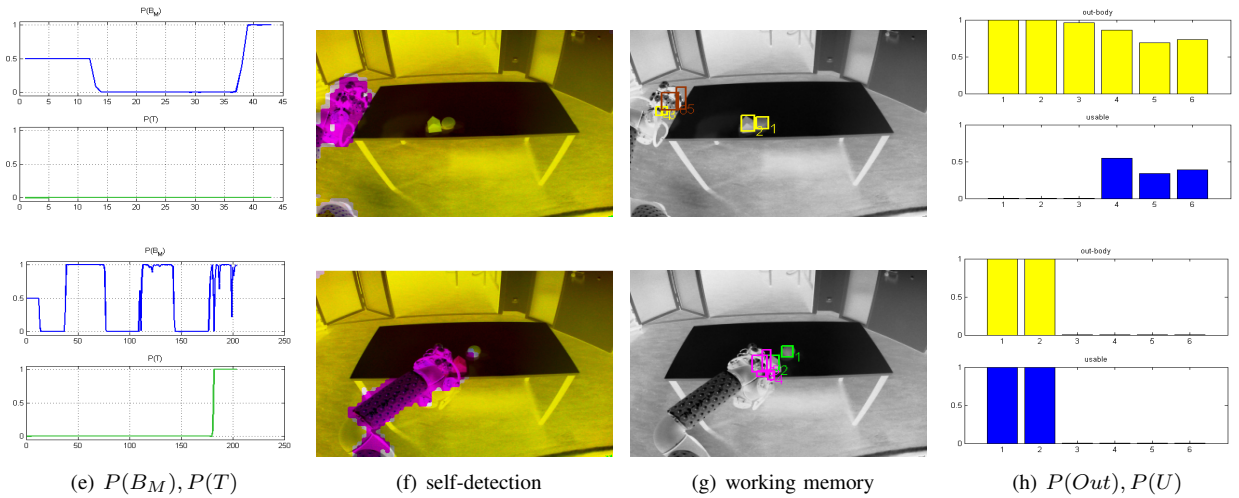


Fig. 4. Object discovery via visual attention and self-detection. First row describes the first stage of the system where the arm has started moving but self-features are still counted as outbody parts. Second row shows the robot pushing an object (due to the pushing another object has been also moved). The system infers that those visual objects are actually potential usable ones. First column displays the meaningful signals (left 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 are being tracked) being outbody (yellow) and usable (blue).

A. Self-detection

An example of the self-detection inference is shown in Figure 5. The saliency and the protoobjects moving, outputted by the visual attention, are described the first two columns. Fig. 5(c) shows the probability of each pixel belonging to the robot. Finally, Fig. 5(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. This can be seen in the unknown areas (grey) and the self larger in the left side of the arm. Moreover, without velocity estimation self regions will appear on the arm trajectory.

B. Object discovery

To evaluate the task of discovering potential 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 4 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. The first row represents the system 7 seconds after starting the experiment and the second row shows when the robot is pushing the object. Note that, after interaction, the robot is certain about object 1 and 2: they are outbody and potentially usable.

C. Scene disambiguation: illusion experiment

We show how tactile and visual cues fusion disambiguates objects usability. We print on a sticker an object that looks three-dimensional from the robot perspective (Fig. 6(a)). Then we put it on the table along with a real object (toy truck)

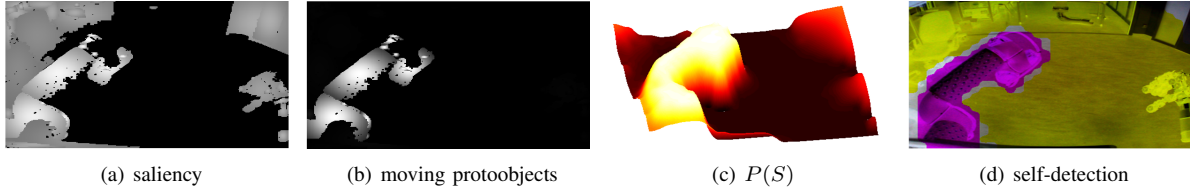


Fig. 5. Self-detection combining visual attention and proprioceptive cues. (a) Saliency map (brighter represents more salient); (c) probability of being self (whiter colour represents higher probability); and (d) self-detection, self (magenta), outbody (yellow) and unknown (greyscale).

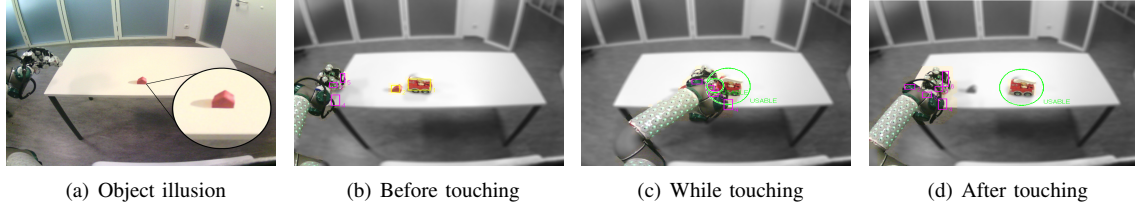
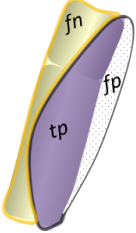


Fig. 6. Disambiguating usability in visual artefacts. One printed object (illusion) and a truck toy are placed on the table. Coloured pixels represent self-detection or protoobjects in the working memory (self (magenta), inanimate (yellow), green (usable)).

that can be moved. The system is able to infer that the truck is usable and the illusion is no longer valid. Figure 6 shows different instants of the robot interacting. When the touching 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 only detects the real usable object. The visual artefact is finally established as an outbody non-usable region.

D. Quantitative study



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

Fig. 7. Quantitative analysis: self-detection and object discovery.

We have also performed a quantitative analysis of self-detection and object discovery. In order to perform the evaluation we have stored one RGB image per second and then segmented the self region by hand into a mask. This is then used as the ground truth. Figure 7 shows the mean values for all experiments. 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. 7. \mathcal{E} is a conservative measure as 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 as the area of outbody is too big and it does not represent an important indicator.

The statistical analysis shows that the visual layer is able to detect the 60% of the robot arm in average during 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. Moreover, the matching metric (\mathcal{E}) describes a 10% performance decrement when the robot is static.

The object discovery task success is summarized in the last column of Fig. 7. 10 experiments with a total of 13 objects to be discovered exhibit one failure. The error is due to poor response of the tactile sensor. The robot is able to discover

the object by touching and analysing the posterior moving causality when it can discern outbody regions.

VI. CONCLUSION

We have presented a perception method for interpreting visual, proprioceptive and tactile signals and to enable self-detection and object discovery. Results shows that self-detection of the arm is 60% on average and 74% when using some prior top-down information. By differentiating self from outbody sources the robot has been able to discover objects in the scene (92% accuracy) and disambiguate visual cues. Thus, by providing in/out body discrimination abilities, more complex types of interpretation activities, such as finding usable objects, is simplified.

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*, vol. 361, no. 1811, pp. 2165–2185, 2003.
- [3] A. Stoytchev, "Some basic principles of developmental robotics," *Auton. Ment. Dev., IEEE Trans. on*, vol. 1, no. 2, pp. 122–130, 2009.
- [4] G. Cheng, *Humanoid Robotics and Neuroscience: Science, Engineering and Society*. CRC Press, 2014.
- [5] J. S. Watson, "Detection of self: The perfect algorithm," *Self-awareness in animals and humans: Developmental perspectives*, pp. 131–148, 1994.
- [6] S.-J. Blakemore and C. Frith, "Self-awareness and action," *Current opinion in neurobiology*, vol. 13, no. 2, pp. 219–224, 2003.
- [7] 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.
- [8] A. Murata, W. Wen, and H. Asama, "The body and objects represented in the ventral stream of the parieto-premotor network," *Neuroscience research*, 2015.
- [9] P. Lanillos, J. F. Ferreira, and J. Dias, "Designing an artificial attention system for social robots," in *Intelligent Robots and Systems (IROS), IEEE/RSJ Int. Conf. on*, 2015, pp. 4171–4178.
- [10] 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.
- [11] A. Stoytchev, "Self-detection in robots: a method based on detecting temporal contingencies," *Robotica*, vol. 29, no. 01, pp. 1–21, 2011.
- [12] P. Mittendorf and G. Cheng, "Humanoid multimodal tactile-sensing modules," *Robotics, IEEE Trans. on*, vol. 27, no. 3, pp. 401–410, 2011.
- [13] J. F. Ferreira and J. Dias, *Probabilistic approaches to robotic perception*. Springer, 2014.
- [14] P. Lanillos, J. F. Ferreira, and J. Dias, "Multisensory 3d saliency for artificial attention systems," in *REACTS Workshop, Int. Conf. of Computer Analysis of Images and Patterns (CAIP)*, 2015, pp. 1–6.
- [15] G. Schillaci, V. V. Hafner, B. Lara, and M. Grosjean, "Is that me?: sensorimotor learning and self-other distinction in robotics," in *ACM/IEEE Int. Conf. on Human-robot interaction (HRI)*, 2013, pp. 223–224.
- [16] A. Pitti, H. Mori, S. Kouzuma, and Y. Kuniyoshi, "Contingency perception and agency measure in visuo-motor spiking neural networks," *Autonomous Mental Development, IEEE Trans. on*, vol. 1, no. 1, pp. 86–97, 2009.
- [17] E. T. Jaynes, *Probability theory: the logic of science*. Cambridge university press, 2003.
- [18] 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.
- [19] 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), IEEE Int. Conf. on*, 2008, pp. 157–162.
- [20] E. Page, "Continuous inspection schemes," *Biometrika*, vol. 41, no. 1/2, pp. 100–115, 1954.
- [21] R. Marfil, A. J. Palomino, and A. Bandera, "Combining segmentation and attention: a new foveal attention model," *Frontiers in computational neuroscience*, vol. 8, 2014.
- [22] F. Bergner, E. Dean-Leon, and G. Cheng, "Event-based signaling for large-scale artificial robotic skin - realization and performance evaluation," in *Intelligent Robots and Systems (IROS), IEEE/RSJ Int. Conf. on*, 2016, p. to Appear.
- [23] P. Bessière, C. Laugier, and R. Siegwart, *Probabilistic reasoning and decision making in sensory-motor systems*. Springer, 2008, vol. 46.