Adaptive robot body learning and estimation through predictive coding

Pablo Lanillos*
Institute for Cognitive Systems
Technische Universität München, Germany
p.lanillos@tum.de

Gordon Cheng
Institute for Cognitive Systems
Technische Universität München, Germany
gordon@tum.de

Abstract

The predictive functions that permit humans to infer their body state by sensorimotor integration are critical to deploy safe interaction in complex environments. These functions are adaptive and robust to non-linear actuators and noisy sensory information. This paper presents a powerful and scalable computational perceptual model based on predictive processing that enables any multisensory robot to learn, infer and update its body configuration when using arbitrary sensors with Gaussian additive noise. The proposed method integrates different sources of information (tactile, visual and proprioceptive) to drive the robot belief to its current body configuration. The motivation is to enable robots with the embodied perception needed for self-calibration and safe physical human-robot interaction.

We formulate body learning obtaining the function that encodes the sensory consequences of the body configuration and its partial derivative with respect to body variables, and we solve it by Gaussian process regression. We model body estimation as minimizing the discrepancy between the robot body configuration belief and the observed posterior. We minimize the variational free energy using the sensory prediction errors (sensed vs expected).

In order to evaluate the model we test it on a real multisensory robotic arm. We show how different sensor modalities contributions, included as additive errors, improve the refinement of the body estimation and how the system adapts itself to provide the most plausible solution even when injecting strong sensory visuo-tactile perturbations. We further analyse the reliability of the model when different sensor modalities are disabled. This provides grounded evidence about the correctness of the perceptual model and shows how the robot estimates and adjusts its body configuration just by means of sensory information.

Index Terms

Bio-inspired perception, body-schema, predictive processing, embodied artificial intelligence, learning and adaptive systems, humanoid robotics.

I. Introduction

Providing the robot with the predictive functions of the body, environment and others is a critical aspect for complex interaction. Appropriately, in order to generate safe and robust interaction the artificial agent must take into account uncertainties related to the sensory input as well as the unexpected events that can occur. Unfortunately, a perfect model of the body, environment and others is almost impossible to design. We present an adaptive robot body learning and estimation algorithm able to deal with noisy sensory inputs and to integrate multiple sources of information (touch, visual and proprioceptive cues). This model is framed on the predictive processing theory proposed by Friston [1] and biologically grounded on predictive coding evidence about the brain as observed by Rao and Ballard in the visual cortex [2]. This approach has been extensively studied in computational biology and psychology but it has not been properly tested in robotics [3].

The main idea behind this embodied approach of robot body perception is that the only available information is the sensory input. By learning the predictor of the sensor outcome given an action and its current body variables (sensory consequences forward model) [4], the robot is able to properly infer its real body configuration. The error between the expected sensory signal and the real input contributes to refine the most plausible hypothesis that the robot has about its body configuration. This simplifies the complexity of online estimation of the body internal variables and increases the ability of the robot to adapt to uncertain situations. The proposed model is described in Fig. 1.

A. Previous works

As discussed in the reviews on the enactive self [5] and body schema [6], there are several methods for learning the body model of the robot without prior knowledge by exploration. For instance, locally weighted projection regression [7], local Gaussian process [8] or infinite experts algorithm [9]. These methods are able to compute the mapping between the sensory input and the configuration of the body and are used to learn forward and inverse kinematics and dynamics. Feed-forward and recurrent nets can also be assimilated for learning body schemas and the needed predictors but relay on supervised information and hundreds of parameters optimization [10], [11]. Unsupervised and self-exploration learning of the body has also been addressed in works like [12] using temporal contingencies or in [13] by means of goal-babbling. Moreover, biologically plausible sensorimotor learning has been investigated in works like [14] by means of Hebbian-based methods

*This work was supported by SELFCEPTION project (www.selfception.eu) European Union Horizon 2020 Programme (MSCA-IF-2016) under grant agreement no. 741941. All authors are affiliated to the Institute for Cognitive Systems (ICS), Technische Universität München, Arcisstrae 21 80333 München, Germany {p.lanillos, gordon}@tum.de.

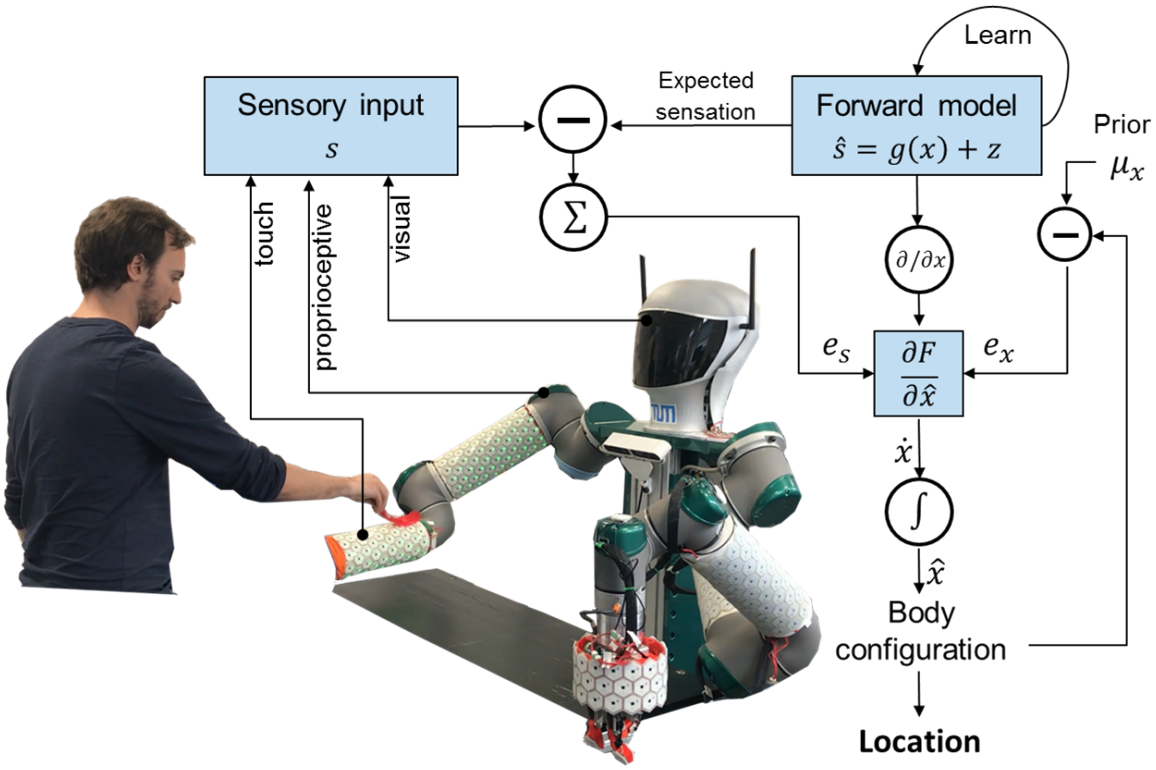


Fig. 1. Proposed adaptive robot body learning and estimation using prediction errors: expected sensation minus sensory input. Visual, tactile and proprioceptive sensing contribute to obtain the most plausible body configuration, and hence the end-effector location.

where body calibration can be learnt through sensorimotor mapping. Dynamic Hebbian learning has also been proposed for obtaining intermodal forward models in [15]. Body model free visual detection [16] has been approached as an intermodal inference problem but it is restricted to the camera view of the robot.

On the other hand, predictive coding [2] and predictive processing [1], [17], has mainly been studied for human perception and control. Just a few works have applied it to robots. For instance, in [3] the viability of predictive processing for robot control on a simulated robotic arm is discussed. However, the generative functions and the parameters were known in advance. Besides, implementing predictive coding with neural networks has gained popularity for modelling multisensory perception [18]. Finally, this approach shares some background with sensorimotor contingencies [4], [19], and predictive learning [20], where the robot develops its perception as infants do.

B. Motivation and method

To produce safe interaction the robot should robustly predict its body and other agents in every instant using all sensory information available. We redefine the body estimation problem as computing the posterior probability of body internal variables x depending on the sensory information s: $p(x|s)$[1]. The robot does not have prior knowledge about its model neither access to its internal variables. Hence, to infer the body configuration from the sensory input we apply Bayes rule:

$$p(x|s) = \frac{p(s|x)p(x)}{p(s)} \tag{1}$$

where $p(s|x)$ is the sensory consequence of being in state x and $p(x)$ is the prior belief of the internal variables.

Unlikely, the robot has incomplete knowledge about the real generative process of the body. According to predictive processing theory [21], the generative process $\mathbf{f}(x)$ governs the dynamics of the environment and the generative process $\mathbf{g}(x)$ governs the sensory information. On the other hand, the processes $f(x)$, $g(x)$ that the agent knows might be different from the real processes as they are just an approximation of them. The agent is continuously adapting its belief about itself and the world just with the sensory input. This is performed by dynamically updating its internal variables x by means of the error between the expected and the real sensory input: the prediction error. We have adopted the predictive processing mathematical framework for the following reasons: it provides indirect minimization of the Kullback-Leibler divergence

[1]We have intentionally left the action out to be coherent with the active inference perceptual theory from [17]. At the end of the paper we remark the role of the action within the proposed scheme.

[21]; it supports multisensory integration; it is scalable in its variational free-energy form [22]; it permits unsupervised Hebbian-like parameters tuning [23]; and it is biologically plausible [2].

Defining the belief of the agent as $q(x)$ and its difference with the posterior distribution, which includes the observations, as the Kullback-Leibler divergence $D_{KL}(q(x)||p(x|s))$, we can reduce the discrepancy by minimizing the free-energy F that is expressed as [17], [22]:

$$D_{KL}(q(x)||p(x|s)) = F + \ln p(s) \tag{2}$$

When F converge to $\emptyset$, only the sensory surprise differs from the belief and the posterior is properly approximated. Hence, in theory, the inference of the body internal configuration depending on the sensory input can be computed by minimizing the variational free energy through a gradient descent scheme: $\dot{x} = \frac{\partial F}{\partial x}$.

C. Contribution and organization

This work introduces an alternative approach to robot body perception based on predictive processing [21], where the robot first learns the sensors or features forward generative model and then it is able to dynamically provide the most plausible body configuration and the location of the end-effector, incorporating in a scalable way several noisy sources of sensory information. In this paper, we face free model body learning and estimation, where the sensory consequences generative model is learnt using a Gaussian process regression. The body configuration and end-effector location is obtained by means of on-line free-energy minimization using the prediction error.

The computational model is presented in Sec. II. We provide a way to learn the generative sensor model $g(x)$ and its derivative by exploration (sampling), as well as the differential equations to solve body estimation by free-energy minimization. The experimental set-up on a real multisensory robot is presented in Sec. III. We show at Sec. IV the performance of the proposed approach, evaluating the body estimation with different sensor modalities and inducing visuo-tactile perturbations. Finally, in Sec. V and VI we discuss the advantages and drawbacks of the proposed approach, and enforce the applicability of the method to improve self localization and interaction.

II. Mathematical model

Notation

$\mathbf{x}, x$	Distribution and value of the body vars.
$\hat{x}$	Most plausible hypothesis of body vars.
μ_x	Prior belief of the body variables
$s : s_v, s_p, s_t$	Sensor value: visual, proprioceptive, tactile
$e : e_v, e_p, e_t$	Error value: visual, proprioceptive, tactile
$\mathbf{g}(.), g(.)$	Real and agent sensor generative process
$\mathbf{f}(.), f(.)$	Real and agent world generative process
$p(.)$	Probability
F	Free-energy
$\frac{\partial F}{\partial x}$	Derivative of free-energy w.r.t internal state
$h(s; g(.), \sigma)$	Normal with $g(.)$ mean and σ variance

We first describe the proposed mathematical model for visual and proprioceptive information and then we extend it with a more complex visuo-tactile input. The model is based on works from predictive processing [21], [17] and free-energy approaches to perception [23], [22]. For this model and without loss of generality, we restrict that the robot cannot perceive the gradient of the sensor signal and that the robot generative model is static. In Sec. IV we discuss the drawbacks of these simplifications. For the sake of clarity, we adopted the free-energy framework presented in [23].

We define the robot for body perception as a set of sensors s and body internal variables x. The proprioceptive sensors s_p outputs a value depending on the body configuration that follows a Normal distribution with linear or non-linear mean $g_p(x)$: $s_p = h(x; g_p(x), \sigma_p)$. The visual sensor s_v provides the location of the end-effector in the visual field also following a Normal distribution with linear or non-linear mean $g_v(x)$: $s_v = h(x; g_v(x), \sigma_v)$. Finally, the robot counts with artificial skin sensors on the end-effector limb and is able to detect other's hand in the visual field - See Fig. 1.

A. Perception model for visual and proprioceptive cues

The body configuration x is inferred via visual and proprioceptive cues through a common Bayes rule. Assuming that the visual and proprioceptive sensing are independent, the distribution of x is:

$$p(\mathbf{x}|s_p, s_v) = \frac{p(s_p|\mathbf{x})p(s_v|\mathbf{x})p(\mathbf{x})}{p(s_v, s_p)} \tag{3}$$

The denominator $p(s_v, s_p)$ has integrals that make intractable exact computation for large distributions. However, instead of working with the whole distribution $\mathbf{x}$, we use the most plausible value [22]: $\hat{x}$. According to the predictive processing

theory, the brain works with the most plausible model of the world to perform predictions [17]. The denominator does not depend on $\hat{x}$ anymore [23] and hence we get:

$$p(\hat{x}|s_p, s_v) = p(s_p|\hat{x})p(s_v|\hat{x})p(\hat{x}) \tag{4}$$

Applying logarithms we obtain the negative free-energy formulation:

$$F = \ln p(s_p|\hat{x}) + \ln p(s_v|\hat{x}) + \ln p(\hat{x}) \tag{5}$$

Substituting the probability distributions by their functions $f(.;.)$, and under the Laplace approximation [17], [22] and assuming normally distributed noise, we can compute the negative free energy as:

$$\begin{aligned} F &= \ln h(s_p; g_p(\hat{x}), \sigma_p) + \ln h(s_p; g_p(\hat{x}), \sigma_p) + \ln h(x; \mu_x, \sigma_x) \\ &= -\frac{(\hat{x} - \mu_x)^2}{2\sigma_x} - \frac{(s_p - g_p(\hat{x}))^2}{2\sigma_p} - \frac{(s_v - g_v(\hat{x}))^2}{2\sigma_v} + \\ &+ \frac{1}{2}\left[-\ln \sigma_x - \ln \sigma_{s_p} - \ln \sigma_{s_v}\right] + C. \end{aligned} \tag{6}$$

To compute the posterior distribution we minimize the free-energy. In a gradient-descent scheme we minimize F through:

$$\dot{x} = \frac{\partial F}{\partial \hat{x}} \tag{7}$$

Computing the partial derivative of Eq. 6 we obtain:

$$\dot{x} = \underbrace{-\frac{\hat{x} - \mu_x}{\sigma_x}}_{\text{Error prior}} + \underbrace{\frac{s_p - g_p(\hat{x})}{\sigma_p}}_{\text{Error expected proprio.}} g_p'(\hat{x}) + \underbrace{\frac{s_v - g_v(\hat{x})}{\sigma_v}}_{\text{Error visual}} g_v'(\hat{x}) \tag{8}$$

Note that the first term is the error between the most plausible value of the body configuration and its prior belief (e_x), the second term is the error between the observed proprioceptive value and the expected one (e_p) and the third term is the prediction error between the visual obtained position of the end-effector and the expected location (e_v).

Nevertheless, Eq. 8 is only possible to use if we know in advance the sensor generative functions. Furthermore, in our body configuration estimation problem the robot does not have access to x. Therefore the state has to be directly encoded in the proprioceptive sensing space. Thus, we substitute $g_p(\hat{x})$ by $\hat{x}$ and its partial derivative $g_p'(\hat{x})$ is set to 1. For instance, if the body configuration is defined by the joint angles, the state x becomes set of values that the joints sensors (encoders) output. For notation convenience we maintain x as the body configuration but it represents $\hat{s}_p$. By defining prediction errors as:

$$e_x = \frac{1}{\sigma_x}(\hat{x} - \mu_x) \tag{9}$$

$$e_p = \frac{1}{\sigma_p}(s_p - \hat{x}) \tag{10}$$

$$e_v = \frac{1}{\sigma_v}(s_v - g_v(\hat{x})) \tag{11}$$

We have the final differential equation that infers the expected displacement (internal state $\hat{x}$):

$$\dot{x} = -e_x + e_p + e_v g_v'(\hat{x}) \tag{12}$$

According to Eq. 12 the update of the internal state is driven by the observed and the expected value of the state and the error prediction. The gradient of the sensory forward model provides the contribution of each sensor modality to each body configuration variable.

It is important to highlight that the computation of e_x is a simplification of the predictive processing approach for passive static perception [23] as we are omitting the generative model of the world $f(x)$. Accordingly, to certainly reduce the difference between the believed distribution and the observed one, e_x should describe the error between the world generative function and the internal belief: $f(x) - \mu_x$. We leave this extension for further works and in Sec. VI we point out the challenges to obtain the full construct without knowing $f(x)$.

Generalizing the free energy minimization for i sensors the body configuration is driven by:

$$\dot{x} = -e_x + e_p + \sum_i \frac{\partial g_i(\hat{x})^T}{\partial \hat{x}} e_i \tag{13}$$

The full dynamics of body estimation is then given by:

$$\begin{aligned}\dot{x} &= -e_x + e_p + e_v g_v'(\hat{x}) \\ \dot{e}_x &= s_x - \mu_x - \sigma_x e_x \\ \dot{e}_p &= s_p - \hat{x} - \sigma_p e_p \\ \dot{e}_v &= s_v - g_v(\hat{x}) - \sigma_v e_v \\ \dot{\mu}_x &= \mu_x + \lambda e_x \end{aligned} \tag{14}$$

where λ is the learning ratio parameter that specifies how fast the prior of body configuration μ_x is adjusted to the prediction error.

B. Body learning – learning the sensory states caused by the body configuration

The body learning is based on obtaining the unknown $s = g(x)$ and $g'(x)$ functions that relate the sensory consequences with the body state. To learn the function we use Gaussian process regression with collected data generated by body exploration. We obtain sensor samples $\overline{s}$ from the robot in several body configurations $\overline{x}$. For instance, for the visual generative process x is the proprioceptive state and s is the visual information.

The *training* is performed by computing the covariance matrix $K(X, X)$ on the collected data with noise σ_n^2, where the covariance function $k(x_i, x_j)$ is defined as:

$$k_{ij} = \sigma_n^2 \mathbf{I} + k(x_i, x_j) \quad |\forall i, j \in \overline{x} \tag{15}$$

The *prediction* of the sensory outcome s given x is then computed as [24]:

$$g(\hat{x}) = k(\hat{x}, X) K(X, X)^{-1} \overline{s} = k(\hat{x}, X)\boldsymbol{\alpha} \tag{16}$$

where $\boldsymbol{\alpha} = \text{choleski}(K)^T \backslash (\text{choleski}(K) \backslash \overline{s})$ for numerical stability.

Finally, in order to compute the gradient of the posterior $g(x)'$ we differentiate the kernel [25], and obtain its prediction analogously as Eq. 16:

$$g(\hat{x})' = \frac{\partial k(\hat{x}, X)}{\partial \hat{x}} K(X, X)^{-1} \overline{s} = \frac{\partial k(\hat{x}, X)}{\partial \hat{x}} \boldsymbol{\alpha} \tag{17}$$

Using the squared exponential kernel with the Mahalanobis distance covariance function, the derivative becomes:

$$g(\hat{x})' = -\Lambda^{-1}(\hat{x} - X)^T (k(\hat{x}, X)^T \cdot \boldsymbol{\alpha}) \tag{18}$$

where Λ is a matrix where the diagonal is populated with the length scale for each dimension ($\text{diag}(1/l^2)$) and $\cdot$ is element-wise multiplication.

C. Adding tactile feedback and other's interaction

We exploit the artificial skin of the robot to refine the body-configuration estimation. For that purpose, we model a specific visuo-tactile intermodal contingency [4]. When other agent is touching the robot end-effector, it should adjust its body configuration to fit the end-effector location in the visual field where the other agent is touching. In other words, other agent touching the robot end-effector in o_v location in the visual field $\rightarrow$ robot end-effector s_v is there $\rightarrow$ body configuration is adjusted.

First, we consider that the robot is able to discern that its end-effector limb is being touched, and that it knows the relation between the touch signal and the location on the body. Secondly, inspired by the causal reasoning approach presented in [26], we define the likelihood function of being touch by other $p(\tau, o_v|x)$ by means of spatial f_s and temporal f_t coherence. We can learn this function by touching the limb in different end-effector locations. Alternatively, in this paper we reuse learnt model from the visual field g_v to compute the expected end-effector location and define the visuo-tactile sensory forward model as:

$$g_t(\hat{x}) = f_s \cdot f_t = a_1 e^{-b_1 \sum (g_v(x) - o_v)^2} \cdot a_2 e^{-b_2 \delta^2} \tag{19}$$

where a_1, b_1, a_2, b_2 are parameters that shape the likelihood and have been tuned in concordance with the data acquired in [26] from human participants; δ is the level of synchrony of the event (e.g., time difference between the visual and the tactile event); and o_v is the other agent end-effector location in the visual field.

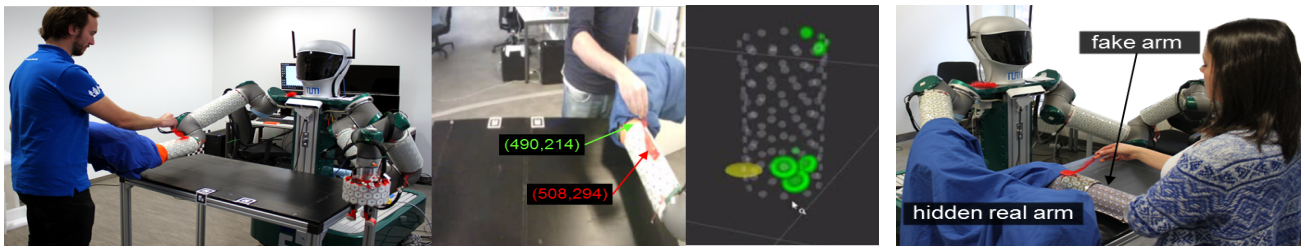
We directly introduce this generative function into the free-energy scheme as follows[2]:

[2]Under the predictive processing framework we might include another internal variable that defines being touched and a second layer of hierarchy that is able to infer similarity (temporal and spatial) between the patterns generated in the visual field by the other agent and the patterns perceived in the skin.

$$\dot{x} = -\frac{\hat{x} - \mu_x}{\sigma_x} + \left[\frac{s_p - \hat{x}}{\sigma_p}, \frac{s_v - g_v(\hat{x})}{\sigma_v}, \frac{s_t - g_t(\hat{x})}{\sigma_t}\right] \begin{bmatrix} \mathbf{1} \\ g'_v(\hat{x}) \\ g'_t(\hat{x}) \end{bmatrix} \tag{20}$$

When a synchronous touching and visual pattern occurs the body configuration is adjusted depending on the expected end-effector visual location $g_v(\hat{x})$ and the other's visual location o_v.

D. *Adaptive body estimation and learning through predictive processing*



(a) Sensory data acquisition (b) Adaptation test

Fig. 2. Experimental setup. (a) Gathering proprioceptive (joints angles), visual robot (green) and other (red) end-effector pixel coordinates) and tactile sensory data from the robot (proximity values) with different participants and positions. Green circles represent the likelihood of being touched. (b) Adaptation test with the body illusion experiment with the real arm covered and a fake arm being touched.

Algorithm 1 summarizes the learning and estimation stages to dynamically compute the internal body configuration based on the sensory error prediction, using for multiple independent sources $i \in N$ of sensory information or features, and body internal variables $j \in M$. The learning stage is using GP regression described in [24] for each sensor modality contribution to the body configuration. Using the $\boldsymbol{\alpha}$ GP solution we also reduce the complexity of the prediction. The estimation stage computes the prediction error for every sensor and solves the differential equations by variational free-energy minimization. Note that we are applying 1st order Euler integration method. More accurate approaches are out of the scope of this paper.

Algorithm 1 Multisensory body learning and estimation

Forward sensor model learning using GP

$X, \overline{s}_i$ ▷ samples $\overline{s} = g(X)$
$K = \text{choleski}(\sigma_n^2 \mathbf{I} + k(X, X))$ ▷ Covariance
for i=1:N **do** ▷ For every sensor/feature modality
  $\boldsymbol{\alpha}_i = K^T \backslash (K \backslash \overline{s})$
end for
return $X, \boldsymbol{\alpha}_i$

Predictive processing body estimation

$X, \boldsymbol{\alpha}_i$ ▷ GP training
$\mu_x \in \mathbb{R}^M$ ▷ Body configuration prior
$\hat{x} \in \mathbb{R}^M$ ▷ Initial body estimation
$e \in \mathbb{R}^N$ ▷ Initial predition error
$s_i \leftarrow \mathbf{g}(x)$ ▷ Input sensor information
for i=1:N **do** ▷ Compute N predictions
  $g_i(\hat{x}) = k(\hat{x}, X)\boldsymbol{\alpha}_i$ ▷ Eq. 16
  $g'_i(\hat{x}) = -\Lambda^{-1}(\hat{x} - X)^T (k(\hat{x}, X)^T \cdot \boldsymbol{\alpha}_i)$ ▷ Eq. 18
end for
for i=1:N **do**
  $\dot{e}_i = s_i - g_i(\hat{x}) - \sigma_i e_i$ ▷ Prediction errors Eq. 20
end for
$\dot{x} = -e_x + \sum_i e_i g'_i(\hat{x})$ ▷ Free-energy minimization dynamics
$\dot{e}_x = \hat{x} - \mu_x - \sigma_x e_x$
$\dot{\mu}_x = \mu_x + \lambda\ e_x$
$\mathcal{X} \leftarrow [\hat{x}, e_{1:N}, e_x, \mu_x]^T$
$\mathcal{X} = \mathcal{X} + \Delta_t \dot{\mathcal{X}}$ ▷ Integration
return $\mathcal{X}$

III. Experimental setup on a robotic arm

We test the model on the multisensory UR-5 arm of robot TOMM [27] as depicted in Fig. 2. The proprioceptive input data is three joint angles with added noise (shoulder$_1$, shoulder$_2$ and elbow - Fig. 3(a)). The visual input is a rgb camera mounted on the head of the robot with 640×480 pixels definition. The tactile input is generated by multimodal skin cells distributed on the arm [28]. Further specifications of the robot are detailed in [29].

A. Learning $g(x)$ from visual and proprioceptive data

In order to learn the sensory forward model we programmed random trajectories in the joint space that resemble to horizontal displacements of the arm. Figure 3(a) shows the data extracted: noisy joint angles and visual location of the end-effector, obtained by colour segmentation. To learn the visual forward model $s_v = g_v(x)$, each sample is defined as the input joint angles sensor values $x = (x_1, x_2, x_3)$ and the output $s_v = (i, j)$ pixel coordinates. As an example, Fig. 3(b) shows the learnt visual forward model by GP regression with 46 samples (red dots). The horizontal displacement mean (in pixels) with respect to two joints angles and the variance.

B. Extracting visuo-tactile data

We use proximity sensing information from the infrarred sensors located in every skin cell to discern when the arm is being touched. The infrarred sensor outputs a filtered signal $\in (0, 1)$. The likelihood of a cell being touched is given by the following function (Eq. 19) $a_1 e^{-b_1 \sum (s_v - s_o)^2}$, where $a_1 = 0.001$ and $b_1 = 1$. The parameters have been obtained by fitting the function to the distance-sensor output measurements. Figure 3(c) shows the raw skin proximity sensing data during the experiment (each colour represents the 117 different skin cells). From the other's hand visual trajectory and the skin proximity activation we compute the level of synchrony between the two patterns (Fig. 3(d)). Timings for tactile stimuli are obtained by setting a threshold over the proximity value: prox $> 0.7 \rightarrow$ activation. Timings for other's trajectory events are obtained through the velocity components. Detected initial and ending position of the visual touching is depicted in Fig. 3(d) (right, green circles).

IV. Results

For comparison purposes, all experiments parameters are set fixed values. $g_v(x)$ learning hyperparameters: signal variance $\sigma_n = \exp(0.02)$ and kernel length scale $l = \exp(0.1)$. The integration step is $\Delta_t = 0.05$ $(20Hz)$ and error variances are $\sigma_x \in \mathcal{R}^3 = [1, 1, 1]$, $\sigma_p \in \mathcal{R}^3 = [1, 1, 1]$, $\sigma_v \in \mathcal{Z}^2 = [5, 5]$. Finally, the learning rate of μ_x is $\lambda = 1$.

A. Robust multisensory integration

We present three different experiments to study visual and proprioceptive body estimation. The first one, described in Fig. 4 shows the proposed body estimation algorithm while deploying a similar trajectory as presented in Fig. 3(a). We analyse the error between the estimated body configuration and the ground truth joint angles for different sensor contributions. The algorithm is able to correctly estimate the joint angles but presents slow dynamics when big changes occur, due to the static nature of the generative model used. It also shows that with only visual input it is not able to estimate the elbow angle. This happens because learning trajectory was set to not provide information about the elbow. However, we can see how combining visual information and two joints sensors ($p1 + p3 + v$), reduces the estimation error. This shows the ability of the proposed method to deal with missing information.

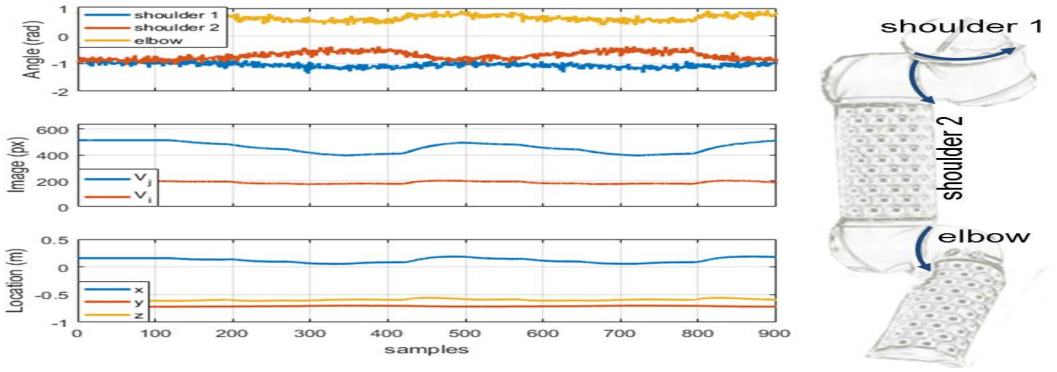
In second experiment, presented in 5(a), we test the model with non-linear proprioceptive sensors: x^2. The body configuration values plotted are in the sensor space. We have initialized the robot body belief with a wrong configuration. On the first 5 seconds the plot shows how the system converges to the "embodied" configuration and then the arm starts moving. The estimation reaction time is slightly slower than previous experiment. Furthermore, we observe an interesting effect. The joint angles vary from $-\pi$ to π, but with the function x^2 the robot cannot distinguish between positive and negative angles. Thus, when inverting the sign of one joint the robot thinks that it is in the right configuration but it is not.

The last experiment, depicted in Fig. 5(b), we study how the model deal with damaged or uncalibrated sensors. After the visual learning stage, we have added a drift error to shoulder$_1$ proprioceptive sensor. The visual prediction error should correct this anomaly. The plot shows how the system nicely reduces proprioceptive drift in shoulder$_1$. However, it induces a wrong bias on shoulder$_2$. Thus, although visual information, with the current $g_v(x)$ learning, evidences a coupling between x_1 and x_2, visual correction has appeared.

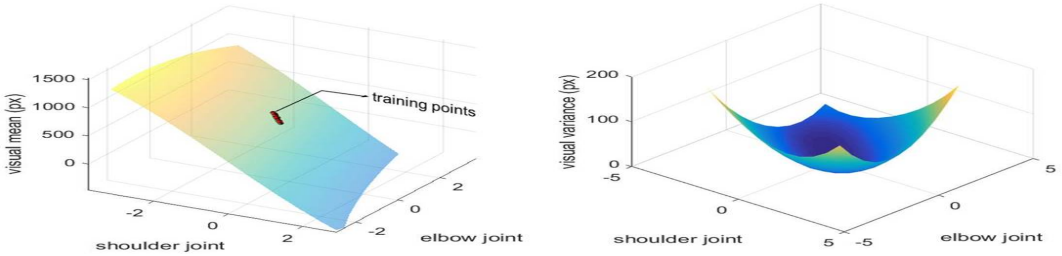
B. Adaptation with visual, proprioceptive and tactile cues

We further test the model adaptation with proprioceptive and visuo-tactile stimulation. We first analyse body estimation refinement in Fig. 6(a). Every sensor modality or feature contributes independently to improve the robot arm localization. The robot can learn more than one visual feature and incorporate them into the predictive error formulation improving body inference.

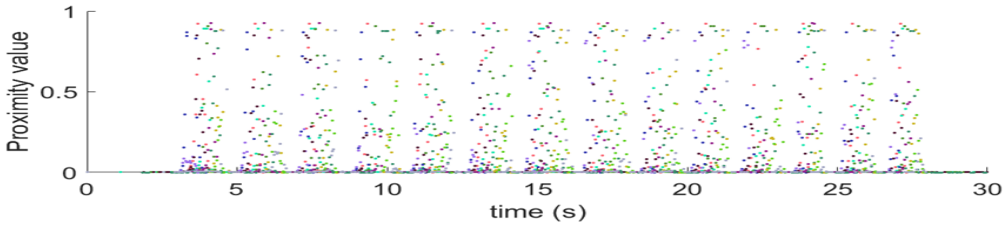
Secondly, as previously depicted in Fig. 2(b), we have introduced a test inspired on the rubber-hand illusion from psychology [30]. This experiment shows the potential of the proposed method to adapt its body inference to incoherent



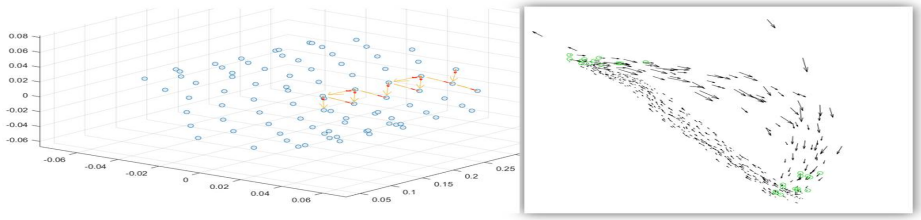
(a) Data recorded example (joints angle + noise, end-effector visual, end-effector cartesian) and schematic picture of the 3-DOF.



(b) Learnt $g_v(x)$ for visual horizontal displacement



(c) Skin proximity data



(d) Tactile (left) and visual (right) event trajectories

Fig. 3. Collected data. (a) Joints, visual and ground truth information of the end-effector. (b) Example of the mean and the variance computed by the GP, which describes the visual horizontal displacement depending on two joints. (c) 30 seconds of raw proximity sensor information of 117 forearm skin cells. (d) Touch patterns extracted from tactile and visual sources.

new situations as a human will do. We cover the arm and we add a fake arm on the table. During the experiment we inject tactile stimuli (as we were touching the robot arm) in synchrony with the movement of the operator hand on the fake hand. This visuo-tactile perturbation should induce a drift towards the new arm of the robot end-effector due to the error on the tactile term of Eq. 20. Figure 6(b) shows the view of the robot with the final end-effector estimation and 6(c) shows the evolution of the joint angles variables. During the first 5 seconds, the system adapts to the real body configuration, and when perturbing with synchronous visuo-tactile stimulation a bias appears on the body joints. This implies a drift on the horizontal axis of the robot end-effector towards the location of the perturbation (fake arm). Tactile perturbations appear as prediction error bumps (yellow line). Fig. 6(c), top plot, also shows how smooth body configuration output $\mu_{x_{1:3}}$ is (blue line). This experiment stresses the adaptability of the body estimation proposed. The robot infers the most plausible situation given the sensory information, which in this case is to include the fake arm visual input as own.

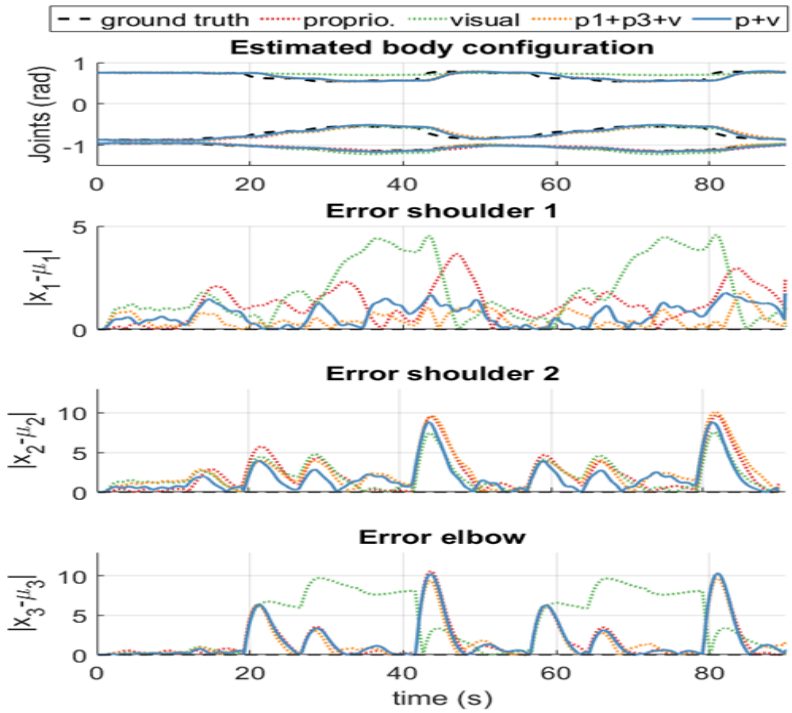
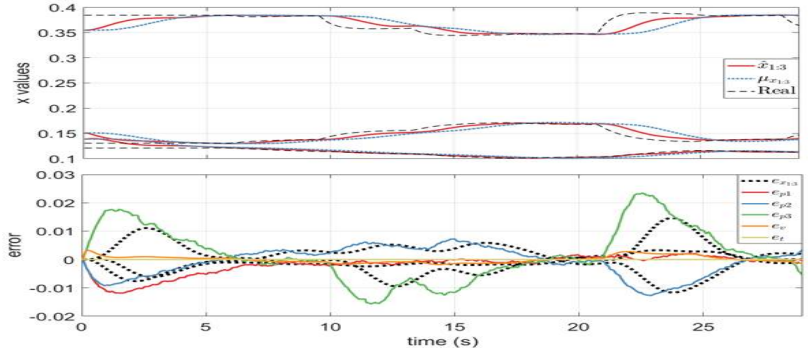
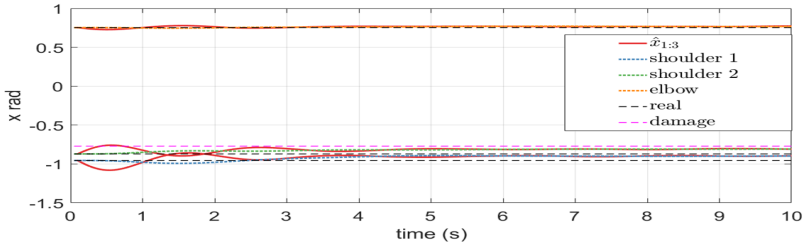


Fig. 4. Body configuration error estimation (joints angles). Proposed algorithm tested with different sensor contributions: only visual; visual + two joint sensors; and visual + proprioception.

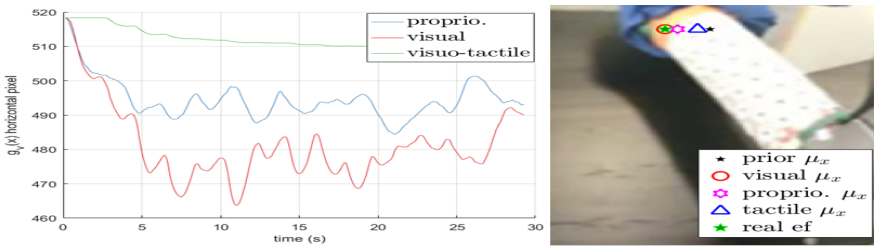


(a) Body estimation with non-linear proprioception

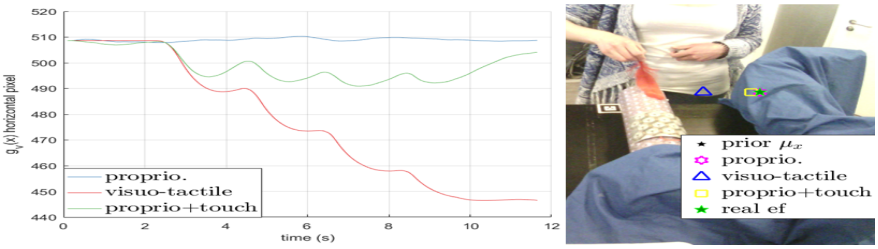


(b) Damaged proprioception compesated with visual cues

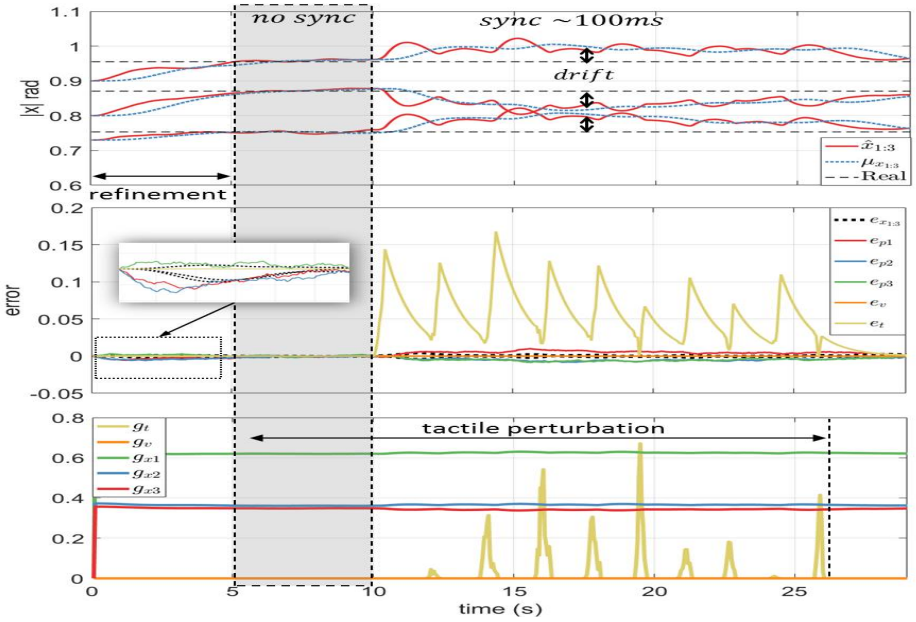
Fig. 5. Non-linear and damaged proprioception test. (a) The generative model of the proprioceptive sensing is quadratic plus Gaussian noise $x^2 + z_p$. (a.top) Body configuration estimation vs real in the *proprioceptive values space*. (a.bottom) Evolution of the predictive error for each sensor contribution. The initial prior joint angles $\mu_x = [-0.8, 0.70, 0.60]$ differs from the real one $x = [-0.9550, -0.8692, 0.7532]$. The first 5 seconds the estimation converge to the real body configuration and then the arm starts moving and body estimation changes accordingly.



(a) Body estimation refinement with biased prior



(b) Body estimation with visuo-tactile perturbation



(c) Evolution of body estimation and predictive errors

Fig. 6. Adaptation analysis. (a) Body refinement with biased prior using different modalities. (b) Visio-tactile illusion. (c) Body configuration, errors and $g(x)$ evolution of the visuo-tactile illusion. In the first 5 seconds there is no tactile stimulation then estimation is refined. Then we inject visuo-tactile stimulation while other is pretending to touch another arm. When it becomes synchronous a horizontal drift appears.

C. A note on scalability

The learning using Gaussian process regression has a computational complexity of $\mathcal{O}(n^3)$ and the prediction of the sensor forward model depends on the covariance kernel complexity $\mathcal{O}(\text{kernel})$. For N independent sensor contributions, M internal variables and N_S samples, the prediction of forward models is $\sim \mathcal{O}(N \times N_S)$. Finally, the free-energy optimization is $\mathcal{O}(N + M)$ using Euler integration method.

V. Discussion: I sense, therefore I am?

We have stressed that robot body estimation can be computed just by means of sensory information. Every sensing modality or feature, when available, contributes to the final body estimation through the prediction error and the variance of each error describes the precision of every sensor with respect to body internal variables. For instance, outside of the field of view proprioceptive and tactile cues define the arm configuration. When the arm appears in the visual field, other features are included into the inference. We have also shown that when the robot has a broken proprioceptive sensor it can

rely on visual features to complete the lack of information. Finally, we have underscored embodiment showing how the sensor function influences body estimation. Hence, we have defined adaptive body learning and estimation as providing the most plausible solution according to the current information available from the sensors. As a collateral effect, the model has been showed to be prone to visuo-tactile illusions, something that has been also evidenced in humans.

Nevertheless, we have only focused on passive perception and omitted deliberatively the generative model of the body dynamics. Moreover, where is the action? We have not considered it in the model, something core for interacting with the body. In order to obtain the full construct, which properly reduces the KL-divergence between the robot belief and the posterior probability of the body configuration given the sensors, we need to include the robot dynamics. However, this is a hard task from the learning perspective. The advantage with this approach is that we only need an approximation of the dynamics because free-energy minimization should solve the discrepancy. With the full construct we expect to improve prediction accuracy and to incorporate the action into the body estimation framework.

VI. Conclusion

We have presented an adaptive robot body learning and estimation algorithm based on predictive processing, able to integrate information from visual, proprioceptive and tactile cues. The robot independently learns the sensor forward generative functions and then it use them to refine its body estimation by a free-energy minimization scheme.

The model has been tested on a robot with a standard industrial arm to facilitate ground truth comparison. Results have shown how the model deals with missing and noisy sensory information, reducing the effect of sensor failures. The algorithm has also displayed adaptability to wrong body prior initialization and unexpected situations. In addition, we have shown how other's touch can refine body robot estimation, opening interesting questions about improved localization and mapping by means of interaction. Altogether reflects the potential of the proposed approach for complex robots, where estimating body location is a hard task and a requirement for safe interaction.

In a follow-up work, we will investigate the full dynamics construct, the role of the action on body estimation and how the variances of each sensory contribution can be dynamically tuned. Besides, we will improve intermodal (visuo-tactile) relations learning with schemes that permit temporal and spatial encoding instead of hand-crafted likelihoods.

References

[1] K. Friston, "A theory of cortical responses," *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 360, no. 1456, pp. 815–836, 2005.

[2] R. P. Rao and D. H. Ballard, "Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects," *Nature neuroscience*, vol. 2, no. 1, pp. 79–87, 1999.

[3] L. Pio-Lopez, A. Nizard, K. Friston, and G. Pezzulo, "Active inference and robot control: a case study," *Journal of The Royal Society Interface*, vol. 13, no. 122, p. 20160616, 2016.

[4] P. Lanillos, E. Dean-Leon, and G. Cheng, "Yielding self-perception in robots through sensorimotor contingencies," *IEEE Trans. on Cognitive and Developmental Systems*, no. 99, pp. 1–1, 2016.

[5] P. Lanillos, E. Dean-Leon, and G. Cheng, "Enactive self: a study of engineering perspectives to obtain the sensorimotor self through enaction," in *Developmental Learning and Epigenetic Robotics, Joint IEEE Int. Conf. on*, 2017.

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

[7] S. Vijayakumar, A. D'souza, and S. Schaal, "Incremental online learning in high dimensions," *Neural computation*, vol. 17, no. 12, pp. 2602–2634, 2005.

[8] D. Nguyen-Tuong, J. R. Peters, and M. Seeger, "Local gaussian process regression for real time online model learning," in *Advances in Neural Information Processing Systems*, 2009, pp. 1193–1200.

[9] B. Damas and J. Santos-Victor, "An online algorithm for simultaneously learning forward and inverse kinematics," in *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ Int. Conf. on*, 2012, pp. 1499–1506.

[10] C. Nabeshima, Y. Kuniyoshi, and M. Lungarella, "Adaptive body schema for robotic tool-use," *Advanced Robotics*, vol. 20, no. 10, pp. 1105–1126, 2006.

[11] E. Wieser and G. Cheng, "Progressive learning of sensory-motor maps through spatiotemporal predictors," in *Developmental Learning and Epigenetic Robotics (ICDL-Epirob), IEEE Int. Conf. on*, 2016.

[12] A. Stoytchev, "Self-detection in robots: a method based on detecting temporal contingencies," *Robotica*, vol. 29, no. 01, pp. 1–21, 2011.

[13] M. Rolf, J. J. Steil, and M. Gienger, "Goal babbling permits direct learning of inverse kinematics," *IEEE Trans. on Autonomous Mental Development*, vol. 2, no. 3, pp. 216–229, 2010.

[14] H. Mori and Y. Kuniyoshi, "A human fetus development simulation: Self-organization of behaviors through tactile sensation," in *Development and Learning (ICDL), IEEE 9th Int. Conf. on*, 2010, pp. 82–87.

[15] G. Schillaci, V. V. Hafner, and B. Lara, "Exploration behaviors, body representations, and simulation processes for the development of cognition in artificial agents," *Frontiers in Robotics and AI*, vol. 3, p. 39, 2016.

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

[17] K. J. Friston, J. Daunizeau, J. Kilner, and S. J. Kiebel, "Action and behavior: a free-energy formulation," *Biological cybernetics*, vol. 102, no. 3, pp. 227–260, 2010.

[18] A. Ahmadi and J. Tani, "Bridging the gap between probabilistic and deterministic models: a simulation study on a variational bayes predictive coding recurrent neural network model," in *Int. Conf. on Neural Information Processing*. Springer, 2017, pp. 760–769.

[19] C. Angulo and J. M. Acevedo-valle, "On dynamical systems for sensorimotor contingencies. a first approach from control engineering," in *Recent Advances in Artificial Intelligence Research and Development, Proc. Int. Conf. of the Catalan Association for Artificial Intelligence*, vol. 300. IOS Press, 2017, p. 46.

[20] Y. Nagai and M. Asada, "Predictive learning of sensorimotor information as a key for cognitive development," in *Proc. of the IROS 2015 Workshop on Sensorimotor Contingencies for Robotics*, 2015.

[21] K. Friston, "Hierarchical models in the brain," *PLoS computational biology*, vol. 4, no. 11, p. e1000211, 2008.
[22] C. L. Buckley, C. S. Kim, S. McGregor, and A. K. Seth, "The free energy principle for action and perception: A mathematical review," *arXiv preprint arXiv:1705.09156*, 2017.
[23] R. Bogacz, "A tutorial on the free-energy framework for modelling perception and learning," *Journal of mathematical psychology*, 2015.
[24] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.
[25] A. McHutchon, "Differentiating gaussian processes," 2013.
[26] M. Samad, A. J. Chung, and L. Shams, "Perception of body ownership is driven by bayesian sensory inference," *PloS one*, vol. 10, no. 2, p. e0117178, 2015.
[27] E. Dean-Leon, B. Pierce, F. Bergner, P. Mittendorfer, K. Ramirez-Amaro, W. Burger, and G. Cheng, "Tomm: Tactile omnidirectional mobile manipulator," in *Robotics and Automation (ICRA), IEEE Int. Conf. on*, 2017, pp. 2441–2447.
[28] P. Mittendorfer and G. Cheng, "Humanoid multimodal tactile-sensing modules," *IEEE Trans. on robotics*, vol. 27, no. 3, pp. 401–410, 2011.
[29] E. Dean, K. R. Amaro, F. Bergner, I. Dianov, and G. Cheng, "Integration of robotic technologies for rapidly deployable robots," *IEEE Trans. on Industrial Informatics*, 2017.
[30] M. Botvinick and J. Cohen, "Rubber hands feeltouch that eyes see," *Nature*, vol. 391, no. 6669, p. 756, 1998.