Sensorimotor learning for artificial body perception

German Diez-Valencia, Takuya Ohashi, Pablo Lanillos*, Gordon Cheng
Department of Electrical Engineering and Computer Sciences
Technical University of Munich
Munich, Germany
*p.lanillos@tum.de

Index Terms

Sensorimotor learning, Body perception, Hierarchical Bayesian estimation, Predictive coding, Deep learning.

I. Introduction

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 [1]. In other words, the spatio-temporal patterns that relate its sensors (e.g. visual, proprioceptive, tactile, etc.), its actions and its body latent variables are responsible of the distinction between its own body and the rest of the world. This paper describes some of the latest approaches for modelling artificial body self-perception: from Bayesian estimation to deep learning. Results show the potential of these free-model unsupervised or semi-supervised crossmodal/intermodal learning approaches. However, there are still challenges that should be overcome before we achieve artificial multisensory body perception.

II. Hierarchical Bayesian models

A first approach on self-perception was integrating multimodal tactile, proprioceptive and visual cues [1] by means of Hierarchical Bayesian models and signal processing, extending [2] and [3] ideas. Results showed that the robot was able to discern between inbody and outbody sources without using markers or simplified segmentation. Figure 1 shows the proto-object saliency system [4] used as visual input and the computed probability of the image regions belonging to the robot arm. Body perception was formalized as an inference problem while the robot was interacting with the world. In order to infer which parts of the scene belong to the robot we integrated visual and accelerometers information. We defined the visual receptive field as a grid where each node (i.e., the decimation of the pixel-wise image) should be decided whether it belongs to the body or not. For that purpose, we adapted Bayesian inference grids to estimate the probability of being its body along the time. The prediction step was computed by learning the pixel-wise velocity in four directions (i.e., up, down, left, right). Furthermore, this method was successfully applied to simplify the problem of discovering objects by interaction [5].

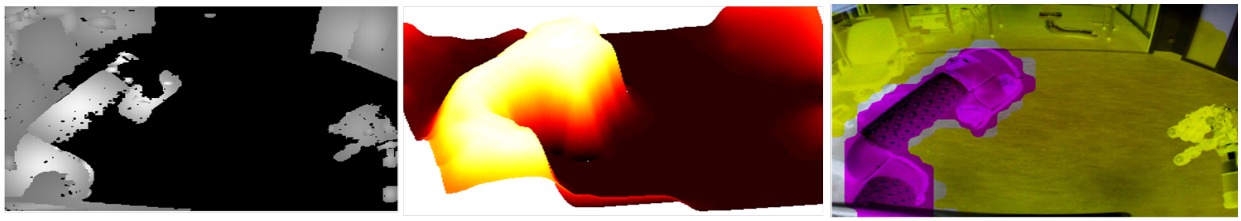
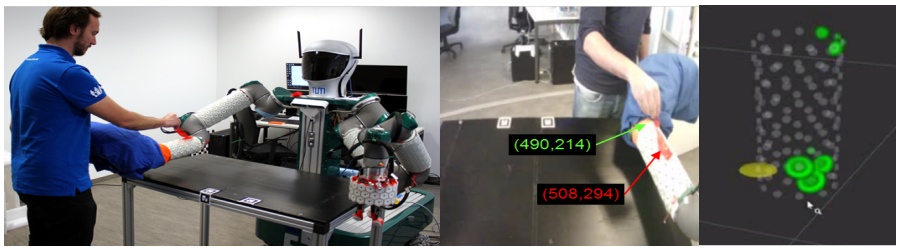


Fig. 1. Self-detection combining visual attention and Bayesian filtering [1]. (left) Saliency segmentation; (middle) inference of the body parts; and (right) inbody vs outbody sources in the visual field.

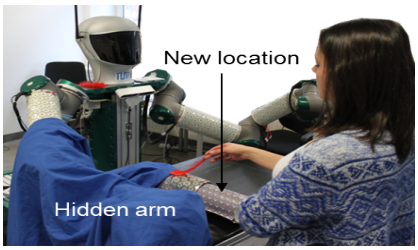
III. Predictive coding models

A biologically plausible body perception model based on predictive processing [6] was also proposed in [7]. Here, body perception was transformed into approximating the latent space distribution $q(\mu)$ that defines the body schema to the real process distribution with the sensory information (posterior) $p(x|s)$. In this approach, the forward sensory model $s = g(\mu) + z$ for each modality was learnt using Gaussian process regression. Sensory fusion was computed by means of inference approximation of the body latent variables μ minimizing the free-energy bound [6]. Furthermore, with this model, the authors were able to replicate the proprioceptive drift pattern of the rubber-hand illusion on a robot with visual, proprioceptive and tactile sensing capabilities [8].

This work was supported by SELFCEPTION project (www.selfception.eu) European Union Horizon 2020 Programme (MSCA-IF-2016) under grant agreement no. 741941. Workshop on Crossmodal Learning for Intelligent Robotics. IEEE Int. Conference on Intelligent Robots and Systems (IROS 2018)



(a) Sensory data acquisition



(b) Rubber-hand illusion test

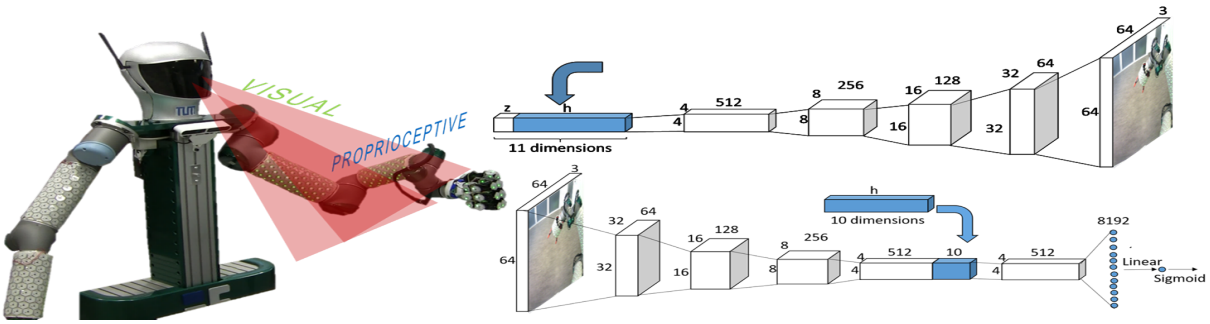
Fig. 2. Body learning and estimation through predictive coding. (a) Gathering proprioceptive (joint angles), visual robot (green) and other (red) end-effector pixel coordinates and tactile sensory data from the robot (proximity values) for different arm configurations. Green circles represent the likelihood of being touched. (b) Adaptation test where we change the visual location of the arm and we induce synchronous visuo-tactile perturbations.

IV. Generative adversarial networks (GANs)

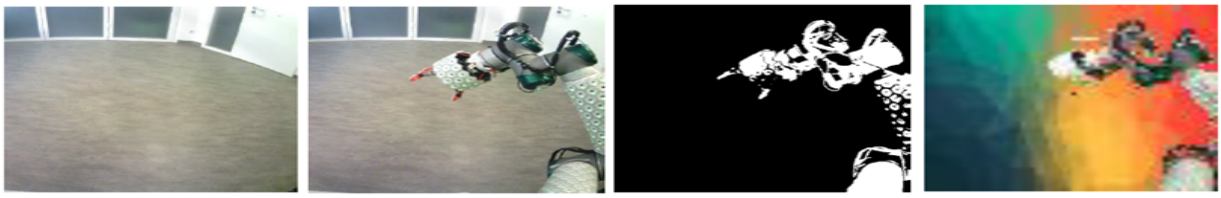
In order to generalize the features used in the previous predictive coding approach, we investigated deep neural networks architectures for learning the generative functions and the cross-modal relations. The advantage of using GANs as a model for self-perception is that the discriminator is a potential self/other distinction mechanism learnt in a unsupervised manner.

A. Visual forward model learning

We analysed how the forward function that relates the joint angles (body) and the visual input can be learnt using GANs. Visuomotor learning has been also approached with recurrent neural networks in [9]. Conversely, here we employ a Deep Convolutional Generative Adversarial Network (DC-GAN) [10], [11]. This method literally generated the arm visual shape depending on the joint angles. Figure 3(a) shows the network architecture and the robot used to extract the data.



(a) Robot and network architecture



(b) Training with synthetic generated data

(c) Generated arm visual appearance for four different joint angles configuration

Fig. 3. Forward visual-kinematic model learning using a DC-GAN. (a) Robot used and GAN architecture. (b) Example of training data generation with synthetic backgrounds. (c) Generated image from different joint arm angles configurations and unlearned background type.

The great challenge was to generalize the reconstruction of the arm for any background without using segmentation. For that purpose, several background images were synthetically generated and were overlaid by automated labelled masks (i.e., boolean mask of the arm in the visual field) by means of background subtraction (Fig. 3(b)). An example of the results of the generated arm given the a joint angle configuration is shown in Fig. 3(c). The most right generated image shows difficulties of the model to properly reconstruct the robot arm when the majority of it is outside the field of view. Anyhow, the statistical evaluation of the network, over all experiments, showed an accuracy of 84.4% when comparing the matching between the original versus the generated image mask.

B. Cross-modal learning

We further analysed self-perception from the cross-modal point of view. Instead of generating the body visual appearance from the joints angles, we extended the architecture to enable signal reconstruction from different sensor modalities using denoising autoencoders [12] but without shared representation. We used the iCub simulator [13] to generate the following visuo-tactile-proprioceptive data: the left arm joint angles, the activation of the skin sensors on the left hand and left forearm and the 2D position of a red cube in the robot left eye image plane. The detection of the red cube was performed by colour blob segmentation. The skin sensors delivered a fixed length array with value 255 if there was contact or 0 otherwise. The image size was fixed to 640×480. We generated several left arm joints configuration, where the forearm appeared in different positions in the visual field. The red cube trajectory was then computed to touch the forearm from the hand to the elbow. After a phase of data synchronization a Wasserstein type GAN [14] was trained off-line. Figure 4 shows the experimental setup of the simulation and the data reconstruction results. Convolution operators were able to extract the inherent structure of the skin data but not reliable enough to provide accurate tactile reconstruction as those operators were thought for images only.

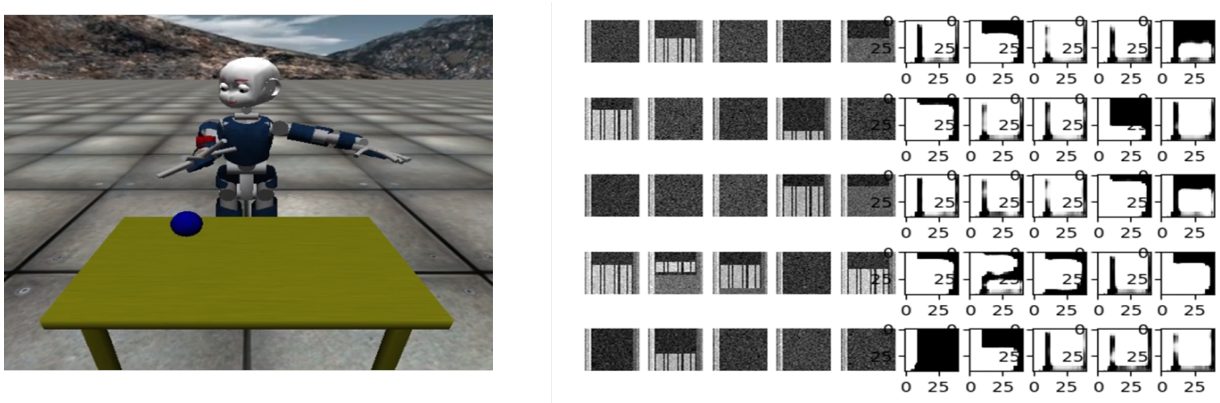


Fig. 4. Learning spatio-temporal multisensory patterns using denoising autoencoders. (left) simulator used to generate the data; (right) results: columns 1-5 original multimodal data, columns 6-10 reconstructed data.

V. Conclusion

We have presented body learning and perception as one of the most representative and challenging cross-modal learning applications. In particular, self-perception in robots has direct applications on adaptability, safety and human-robot interaction. Through examples, we identified at least three important characteristics for modelling artificial body perception: (1) body latent space estimation through noisy sensorimotor fusion; (2) cross-modal signal recovering from multimodal information; and (3) unsupervised self-generated patterns classification. Accordingly, we have shown different techniques for partially solving the problem, such as hierarchical Bayesian models for self-detection on the visual field [1], predictive processing with GP regression for body estimation [7], and deep nets for learning the forward visual-kinematics or visuo-tactile-proprioception relations. Further research will focus on developing a full cross-modal architecture able to properly tackle the nature of the different modalities and allowing sensor relevance tuning.

References

[1] 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.

[2] 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.

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

[4] 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.

[5] 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.

[6] 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.
[7] P. Lanillos and G. Cheng, "Adaptive robot body learning and estimation through predictive coding," *arXiv preprint arXiv:1805.03104*, 2018.
[8] N.-A. Hinz, P. Lanillos, H. Mueller, and G. Cheng, "Drifting perceptual patterns suggest prediction errors fusion rather than hypothesis selection: replicating the rubber-hand illusion on a robot," *arXiv preprint arXiv:1806.06809*, 2018.
[9] J. Hwang, J. Kim, A. Ahmadi, M. Choi, and J. Tani, "Predictive coding-based deep dynamic neural network for visuomotor learning," in *IEEE Int. Conf. Dev. Learn. Epigenetic Robot.(ICDL-EpiRob), Lisbon, Portugal*, 2017.
[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.
[11] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," *arXiv preprint arXiv:1605.05396*, 2016.
[12] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 689–696.
[13] V. Tikhanoff, A. Cangelosi, P. Fitzpatrick, G. Metta, L. Natale, and F. Nori, "An open-source simulator for cognitive robotics research: the prototype of the icub humanoid robot simulator," in *Proceedings of the 8th workshop on performance metrics for intelligent systems*. ACM, 2008, pp. 57–61.
[14] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International Conference on Machine Learning*, 2017, pp. 214–223.