

A deep active inference model of the rubber-hand illusion

Thomas Rood, Marcel van Gerven, and Pablo Lanillos

Department of Artificial Intelligence
 Donders Institute for Brain, Cognition and Behaviour
 Montessorilaan 3, 6525 HR Nijmegen, the Netherlands
 p.lanillos@donders.ru.nl

Abstract. Understanding how perception and action deal with sensorimotor conflicts, such as the rubber-hand illusion (RHI), is essential to understand how the body adapts to uncertain situations. Recent results in humans have shown that the RHI not only produces a change in the perceived arm location, but also causes involuntary forces. Here, we describe a deep active inference agent in a virtual environment, which we subjected to the RHI, that is able to account for these results. We show that our model, which deals with visual high-dimensional inputs, produces similar perceptual and force patterns to those found in humans.

Keywords: Active inference · Rubber-hand illusion · Free-energy optimization · Deep learning.

1 Introduction

The complex mechanisms underlying perception and action that allow seamless interaction with the environment are largely occluded from our consciousness. To interact with the environment in a meaningful way, the brain must integrate noisy sensory information from multiple modalities into a coherent world model, from which to generate and continuously update an appropriate action [13]. Especially, how the brain-body deals with sensorimotor conflicts [8,16], e.g., conflicting information from different senses, is an essential question for both cognitive science and artificial intelligence. Adaptation to unobserved events and changes in the body and the environment during interaction is a key characteristic of body intelligence that machines still fail at.

The rubber-hand illusion (RHI) [2] is a well-known experimental paradigm from cognitive science that allows the investigation of body perception under conflicting information in a controlled setup. During the experiment, human participants cannot see their own hand but rather perceive an artificial hand placed in a different location (e.g. 15 cm from their current hand). After a minute of visuo-tactile stimulation [10], the perceived location of the real hand drifts towards the location of the artificial arm and suddenly the new hand becomes part of their own.

We can find some RHI modelling attempts in the literature; see [12] for an overview until 2015. In [18], a Bayesian causal inference model was proposed to estimate the perceived hand position after stimulation. In [8] a model inspired by the free-energy principle [5] was used to synthetically test the RHI in a robot. The perceptual drift (mislocalization of the hand) was compared to that of humans observations.

Recent experiments have shown that humans also generate meaningful force patterns towards the artificial hand during the RHI [1,16], adding the action dimension to this paradigm. We hypothesise that the strong interdependence between perception and action can be accounted for by mechanisms underlying active inference [7].

In this work, we propose a *deep active inference* model of the RHI, based on [14,17,19], where an artificial agent directly operates in a 3D virtual reality (VR) environment¹. Our model 1) is able to produce similar perceptual and active patterns to human observations during the RHI and 2) provides a scalable approach for further research on body perception and active inference, as it deals with high-dimensional inputs such as visual images originated from the 3D environment.

2 Deep active inference model

We formalise body perception and action as an inference problem [11,3,7,17]. The unobserved body state is inferred from the senses (observations) while taking into account its state prior information. To this end, the agent makes use of two sensory modalities. The visual input s_v is described by a pixel matrix (image) and the proprioceptive information s_p represents the angle of every joint of the arm – See Fig. 1a.

Computation of the body state is performed by optimizing the the variational free-energy bound [7,17]. Under the mean-field and Laplace approximations and defining μ as the brain variables that encode the variational density that approximates the body state distribution and defining a as the action exerted by the agent, perception and action are driven by the following system of differential equations (see [6,4,19] for a derivation):

$$\dot{\mu} = -\partial_\mu F = -\partial_\mu e_p^T \Sigma_p^{-1} e_p - \partial_\mu e_v^T \Sigma_v^{-1} e_v - \partial_\mu e_f^T \Sigma_\mu^{-1} e_f \quad (1)$$

$$\dot{a} = -\partial_a F = -\partial_a e_p^T \Sigma_p^{-1} e_p \quad (2)$$

$$e_p = s_p - g_p(\mu) \quad (3)$$

$$e_v = s_v - g_v(\mu) \quad (4)$$

$$e_f = -f(\mu) \quad (5)$$

Note that this model is a specific instance of the full active inference model [5] tailored to the RHI experiment. We wrote the variational free-energy bound in

¹ Code will be publicly available at <https://github.com/thomasroodnl/active-inference-rhi>

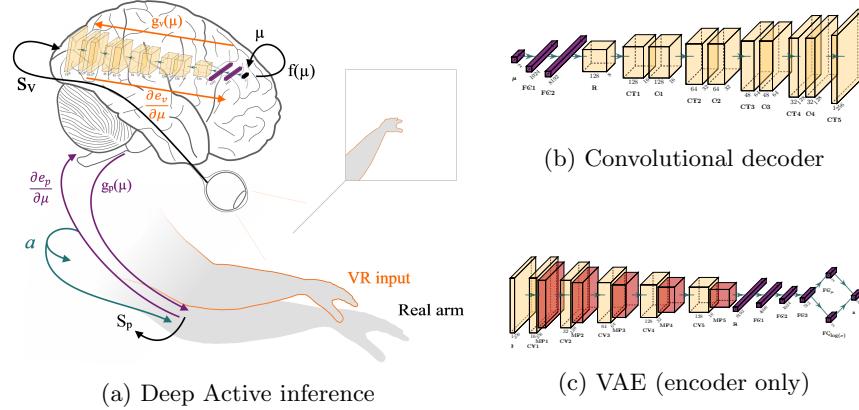


Fig. 1: Deep active inference model for the virtual rubber-hand illusion. (a) The brain variables μ that represent the body state are inferred through proprioceptive e_p and visual e_v prediction errors and their own dynamics $f(\mu)$. During the VR immersion, the agent only sees the VR arm. The ensuing action is driven by proprioceptive prediction errors. The generative visual process is approximated by means of a deep neural network that encodes the sensory input into the body state through a bottleneck. (b,c) Visual generative architectures tested.

terms of the prediction error e and for clarity, we split it into three terms that correspond to the visual, proprioceptive and dynamical component of the body state. The variances $\Sigma_v, \Sigma_p, \Sigma_\mu$ encode the reliability of the visual, proprioceptive and dynamics information, respectively, that is used to infer the body state. The dynamics of the prediction errors are governed by different generative processes. Here, $g_v(\mu)$ is the generative process of the visual information (i.e. the predictor of the visual input given the brain state variables), $g_p(\mu)$ is the proprioceptive generative process and $f(\mu)$ denotes internal state dynamics (i.e. how the brain variables evolve in time)².

Due to the static characteristics of the passive RHI experiment we can simplify the model. First, the generative dynamics model does not affect body update because the experimental setup does not allow for body movement. Second, we fully describe the body state by the joint angles. This means that the s_p and the body state match. Thus, $g(\mu) = \mu$ plus noise and the inverse mapping $\partial_\mu g_p(\mu)$ becomes an all-ones vector. Relaxing these two assumptions is out of the scope of this paper. We can finally write the differential equations with the

² Note that in Equation (5), the prediction error with respect to the internal dynamics $e_f = \mu' - f(\mu)$ was simplified to $e_f = -f(\mu)$ under the assumption that $\mu' = 0$. In other words, we assume no dynamics on the internal variables.

generative models as follows:

$$\dot{\mu} = \Sigma_p^{-1}(s_p - g_p(\mu)) + \partial_\mu g_v(\mu)^T \gamma \Sigma_v^{-1}(s_v - g_v(\mu)) \quad (6)$$

$$\dot{a} = -\Delta_t \Sigma_p^{-1}(s_p - g_p(\mu)) \quad (7)$$

where γ has been included in the visual term to modulate the level of causality regarding whether the visual information has been produced by our body in the RHI – see Sec. 2.2. Equation 7 is only valid if the action is the velocity of the joint. Thus, the sensor change given the action corresponds to the time interval between each iteration $\partial_a s = \Delta_t$.

We scale up the model to high-dimensional inputs such as images by approximating the visual generative model $g_v(\mu)$ and the partial derivative of the error with respect to the brain variables $\partial_\mu e_v$ by means of deep neural networks, inspired by [19].

2.1 Generative model learning

We learn the forward and inverse generative process of the sensory input by exploiting the representational capacity of deep neural networks. Although in this work we only address the visual input, this method can be extended to any other modality. To learn the visual forward model $g_v(\mu)$ we compare two different deep learning architectures, that is, a convolutional decoder (Fig. 1b) and a variational autoencoder (VAE, Fig. 1c).

The convolutional decoder was designed in similar fashion to the architecture used in [19]. After training the relation between the visual input and the body state, the visual prediction can be computed through the forward pass of the network and its inverse $\partial g(\mu)/\partial\mu$ by means of the backward pass. The VAE was designed using the same decoding structure as the convolutional decoder to allow a fair performance comparison. This means that these models mainly differed in the way they were trained. In the VAE approach we train using the full architecture and we just use the decoder to compute the predictions in the model.

2.2 Modelling visuo-tactile stimulation synchrony

To synthetically replicate the RHI we need to model both synchronous and asynchronous visuo-tactile stimulation conditions. We define the timepoints at which a visual stimulation event and the corresponding tactile stimulation take place, denoted t_v and t_t respectively. Inspired by the Bayesian causal model [18], we distinguish between two causal explanations of the observed data. That is, $C = c_1$ signifies that the observed (virtual) hand produced both the visual and the tactile events whereas $C = c_2$ signifies that the observed hand produced the visual event and our real hand produced the tactile event (visual and tactile input come from two different sources). The causal impact of the visual information on the body state is represented by

$$\gamma = p(c_1 | t_v, t_t) = \frac{p(t_v, t_t | c_1)p(c_1)}{p(t_v, t_t | c_1)p(c_1) + p(t_v, t_t | c_2)p(c_2)} \quad (8)$$

where $p(t_v, t_t \mid c_1)$ is defined as a zero-mean Gaussian distribution over the difference between the timepoints ($p(t_v - t_t \mid c_1)$) and $p(t_v, t_t \mid c_2)$ is defined as a uniform distribution since under c_2 , no relation between t_v and t_t is assumed. This yields the update rule

$$\gamma_{t+1} = \begin{cases} \frac{p(t_v, t_t \mid c_1) \gamma_t}{p(t_v, t_t \mid \gamma_t) \cdot \gamma_t + p(t_v, t_t \mid c_2) (1 - \gamma_t)} & \text{if visuo-tactile event} \\ \gamma_t \cdot \exp\left(-\frac{(t - \max(t_v, t_t))^2}{\Delta_t^2} \cdot r_{decay}\right), & \text{otherwise} \end{cases} \quad (9)$$

Note that γ is updated only in case of visuo-tactile events. Otherwise, an exponential decay is applied.

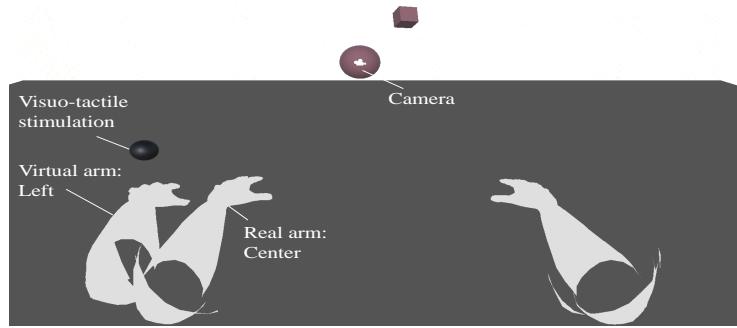


Fig. 2: Virtual environment and experimental setup modelled in the Unity engine.

3 Experimental setup

We modelled the RHI in a virtual environment created in Unity, as depicted in Fig. 2. This environment was build to closely match the experimental setup used in the human study described in [16]. This experiment exposed human participants to a virtual arm located to the left and right of their real arm, and applied visuo-tactile stimulation by showing a virtual ball touching the hand and applying a corresponding vibration to the hand. Here, the agent's control consisted of two degrees of freedom: shoulder adduction/abduction and elbow flexion/extension. The environment provided proprioceptive information on the shoulder and elbow joint angles to the agent. Visual sensory input to the model originated from a camera located between the left and the right eye position, producing 256×256 pixel grayscale images. Finally, the ML-Agents toolkit was used to interface between the Unity environment and the agent in Python [9]. The agent arm was placed in a forward resting position such that the hand was located 30 cm to the left of the body midline (center position). Three virtual arm location conditions were evaluated: Left, Center and Right. The Center condition matched the information given by proprioceptive input. Visuo-tactile

stimulation was applied by generating a visual event at a regular interval of two seconds, followed by a tactile event after a random delay sampled in the range [0, 0.1) for synchronous stimulation and in the range [0, 1) for asynchronous stimulation. The initial γ value was set to 0.01 and we ran $N = 5$ trials each for 30 s (1500 iterations).

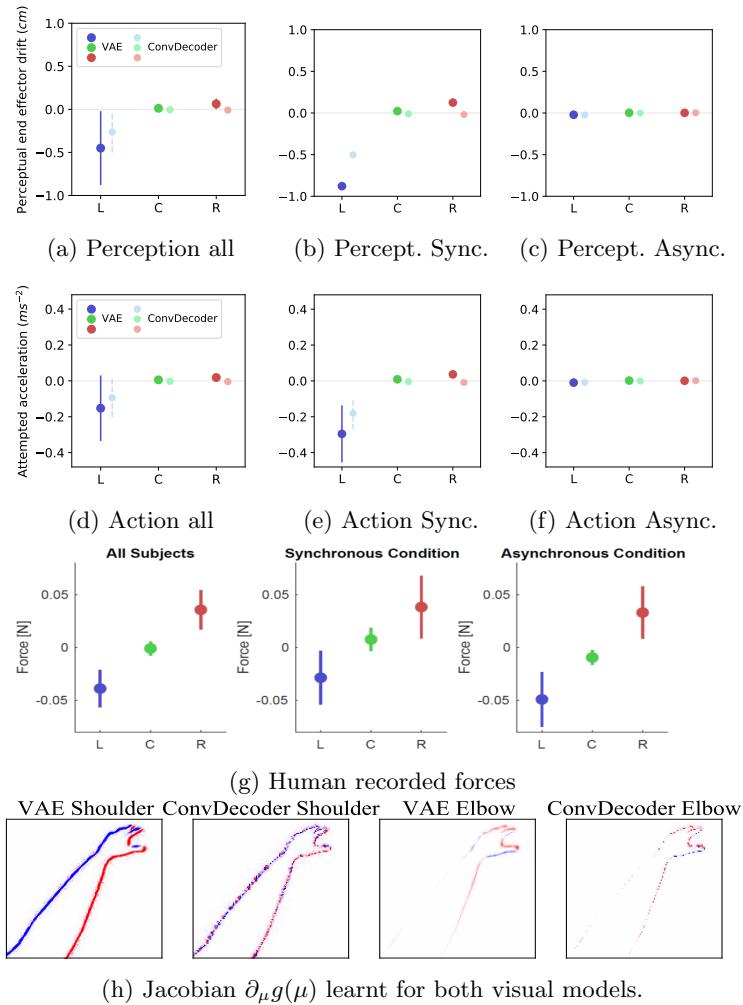


Fig. 3: Model results. (a, b, c) Mean perceptual end-effector drift (in cm). (d,e,f) Mean horizontal end-effector acceleration. (g) Mean forces exerted by human participants in a virtual rubber-hand experiment (from [16]). (h) Visual representation of the Jacobian learnt for the visual models.

4 Results

We observed similar patterns in the drift of the perceived end-effector location (Fig. 3a) and the end-effector action (Fig. 3). These agree with the behavioural data obtained in human experiments (Fig. 3g). For the left and right condition, we observed forces in the direction of the virtual hand during synchronous stimulation (Fig. 3e). However, non-meaningful forces were produced using the convolutional decoder for the right condition. For the center condition, both models produced near-zero average forces. Lastly, asynchronous stimulation produced, with both models, attenuated forces (Fig. 3f). The learnt visual representation differed between the VAE and the Convolutional decoder approaches (Fig. 3h). The VAE obtained smoother and more bounded visual Jacobian values, likely due to its probabilistic latent space.

5 Conclusion

In this work, we described a deep active inference model to study body perception and action during sensorimotor conflicts, such as the RHI. The model, operating as an artificial agent in a virtual environment, was able to produce similar perceptual and active patterns to those found in humans. Further research will address how this model can be employed to investigate the construction of the sensorimotor self [15].

References

1. Asai, T.: Illusory body-ownership entails automatic compensative movement: for the unified representation between body and action. *Experimental brain research* **233**(3), 777–785 (2015)
2. Botvinick, M., Cohen, J.: Rubber hands ‘feel’ touch that eyes see. *Nature* **391**(6669), 756–756 (Feb 1998). <https://doi.org/10.1038/35784>, <https://doi.org/10.1038/35784>
3. Botvinick, M., Toussaint, M.: Planning as inference. *Trends in cognitive sciences* **16**(10), 485–488 (2012)
4. Buckley, C.L., Kim, C.S., McGregor, S., Seth, A.K.: The free energy principle for action and perception: A mathematical review. *Journal of Mathematical Psychology* **81**, 55–79 (2017)
5. Friston, K.: The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience* **11**(2), 127–138 (Feb 2010). <https://doi.org/10.1038/nrn2787>, <https://doi.org/10.1038/nrn2787>
6. Friston, K., Mattout, J., Trujillo-Barreto, N., Ashburner, J., Penny, W.: Variational free energy and the laplace approximation. *Neuroimage* **34**(1), 220–234 (2007)
7. Friston, K.J., Daunizeau, J., Kilner, J., Kiebel, S.J.: Action and behavior: a free-energy formulation. *Biological Cybernetics* **102**(3), 227–260 (Mar 2010). <https://doi.org/10.1007/s00422-010-0364-z>, <https://doi.org/10.1007/s00422-010-0364-z>

8. Hinz, N.A., Lanillos, P., Mueller, H., Cheng, G.: 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. Juliani, A., Berges, V.P., Vckay, E., Gao, Y., Henry, H., Mattar, M., Lange, D.: Unity: A general platform for intelligent agents (2018)
10. Kalckert, A., Ehrsson, H.H.: The onset time of the ownership sensation in the moving rubber hand illusion. *Frontiers in Psychology* **8**, 344 (2017). <https://doi.org/10.3389/fpsyg.2017.00344>, <https://www.frontiersin.org/article/10.3389/fpsyg.2017.00344>
11. Kappen, H.J., Gómez, V., Opper, M.: Optimal control as a graphical model inference problem. *Machine learning* **87**(2), 159–182 (2012)
12. Kilteni, K., Maselli, A., Kording, K.P., Slater, M.: Over my fake body: body ownership illusions for studying the multisensory basis of own-body perception. *Frontiers in human neuroscience* **9**, 141 (2015)
13. Kording, K.P., Wolpert, D.M.: Bayesian integration in sensorimotor learning. *Nature* **427**(6971), 244–247 (Jan 2004). <https://doi.org/10.1038/nature02169>, <https://doi.org/10.1038/nature02169>
14. Lanillos, P., Cheng, G.: Adaptive robot body learning and estimation through predictive coding. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 4083–4090. IEEE (2018)
15. Lanillos, P., Dean-Leon, E., Cheng, G.: 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)
16. Lanillos, P., Franklin, S., Franklin, D.W.: The predictive brain in action: Involuntary actions reduce body prediction errors. *bioRxiv* (2020). <https://doi.org/10.1101/2020.07.08.191304>, <https://www.biorxiv.org/content/early/2020/07/08/2020.07.08.191304>
17. Oliver, G., Lanillos, P., Cheng, G.: Active inference body perception and action for humanoid robots. arXiv preprint arXiv:1906.03022 (2019)
18. Samad, M., Chung, A.J., Shams, L.: Perception of body ownership is driven by bayesian sensory inference. *PloS one* **10**(2), e0117178–e0117178 (Feb 2015). <https://doi.org/10.1371/journal.pone.0117178>, <https://pubmed.ncbi.nlm.nih.gov/25658822>
19. Sancaktar, C., van Gerven, M., Lanillos, P.: End-to-end pixel-based deep active inference for body perception and action. arXiv preprint arXiv:2001.05847 (2020)