A Deep Convolutional Neural Network Model for Sense of Agency and Object Permanence in Robots

Claus Lang
Bernstein Center for Computational Neuroscience
Berlin, Germany
claus.lang@bccn-berlin.de

Guido Schillaci
Adaptive Systems Group
Humboldt-Universität zu Berlin
Berlin, Germany
guido.schillaci@informatik.hu-berlin.de

Verena V. Hafner
Adaptive Systems Group
Humboldt-Universität zu Berlin
Berlin, Germany
hafner@informatik.hu-berlin.de

***Abstract*—This work investigates the role of predictive models in the implementation of basic cognitive skills in robots, such as the capability to distinguish between self-generated actions and those generated by other individuals and the capability to maintain an enhanced internal visual representation of the world, where objects covered by the robot's own body in the original image may be visible in the enhanced one.**

A developmental approach is adopted for this purpose. In particular, a humanoid robot is learning, through a self-exploration behaviour, the sensory consequences (in the visual domain) of self-generated movements. The generated sensorimotor experience is used as training data for a deep convolutional neural network that maps proprioceptive and motor data (e.g. initial arm joint positions and applied motor commands) onto the visual consequences of these actions.

This forward model is then used in two experiments. First, for generating visual predictions of self-generated movements, which are compared to actual visual perceptions and then used to compute a prediction error. This error is shown to be higher when there is an external subject performing actions, compared to situations where the robot is observing only itself. This supports the idea that prediction errors may serve as a cue for distinguishing between self and other, a fundamental prerequisite for the sense of agency.

Secondly, we show how predictions can be used to attenuate self-generated movements, and thus create enhanced visual perceptions, where the sight of objects - originally occluded by the robot body - is still maintained. This may represent an important tool both for cognitive development in robots and for the understanding of the sense of object permanence in humans.

I. Introduction

Human perception is partially guided through a predictive process [1], [2]. The way we perceive the world seems to be strongly shaped by what our brain *expects* to perceive. These expectations would be made up from past experience of our bodily interaction with the world. Growing evidence suggests that similar processes shape also the way we perceive *ourselves* [3]–[7]. How do we recognise our own body and our own movements? Indeed, this is a very intriguing skill that we, as humans, possess and that allows us to naturally interact with other people and with our surroundings.

Researchers usually identify two main phenomena characterising self-recognition in humans [8]: a sense of body ownership and a sense of agency. Sense of body ownership is intended as the subjective experience where we feel that we are observing *our own body*[1]. Sense of agency is intended, instead, as the subjective experience where we feel in *control* of our own actions.

One of the most accepted proposals explaining the functioning of these phenomena relies on the existence of a self-monitoring mechanism in our brain [9]. Such a monitoring system would constantly anticipate the sensory consequences of our own actions. The discrepancy between predicted and observed sensory outcomes of bodily actions would thus serve as a cue for the sense of agency and body ownership. The original proposal [9] has been further developed and also debated. An interesting alternative proposal explaining sense of agency supports also the existence of a self-monitoring mechanism in our brain, but shapes it around a predictive coding framework [4].

Recently, interest in these phenomena has been raising also in the developmental robotics community [10]. Here, scientists are investigating whether and how self-awareness - and, in general, motor and higher cognitive skills - can be implemented into artificial agents adopting a developmental approach. Such an approach would consist in implementing into the artificial agent those fundamental behavioural and computational components that would allow it to go through an autonomous onthogenetic process, where the experienced embodied interaction with the external environment is used to form up knowledge about the body and its capabilities.

Although we are still far from having self-aware robots, many interesting studies have been conducted in the last years, which provide insights in the understanding and in the implementation of such a skill in artificial systems. For instance, predictive models - trained with sensorimotor data generated by the robot itself - have been shown to allow the implementation of basic cognitive skills (for a review, please refer to [11] and [12]).

In previous works, we investigated such models in humanoid robots, implementing predictive capabilities in different sensory modalities, e.g. proprioceptive, visual [13] and auditory [14]. In [13], for instance, we presented an implementation of a self-monitoring mechanism that allowed a simulated humanoid

[1] *Observing*, here, is meant more generally as *perceiving*, e.g. visually or through our tactile sense.

robot to anticipate the sensory consequences of self-generated movements. We demonstrated that prediction errors in the visual modality can be used as a cue for distinguishing between self-generated actions and actions generated by other agents. Moreover, we investigated whether the same self-monitoring mechanism can be behind the development of a sense of object permanence. Sense of object permanence is the feeling that objects continue to exist even if they cannot be seen, heard, or otherwise observed [15]. Several studies suggest that object constancy perception relies on predictive processes implemented by our brains [16], [17]. More importantly, this phenomenon is considered crucial in human cognitive development. In [13], we showed that a humanoid robot can maintain a mental representation of an occluded stationary object, through the same predictive and sensory attenuation processes that are implementing the self-monitoring mechanism described above.

However, in [13], the hypotheses were only confirmed on data from a simulated robot with two degrees of freedom. Furthermore, the applied k-NN algorithm operated on extracted hard-coded features instead of the higher-dimensional original images. Here, we extend our previous work by confirming the hypotheses on real-world visual and real-robot sensorimotor data obtained from of a robot with four degrees of freedom. For that, we replace the limited k-NN forward model with a more sophisticated one not requiring predefined feature extraction - implemented as a deep convolutional neural network capable of image generation.

The rest of the paper is structured as follows. In section II, we describe the methodology, the robotic platform and the models adopted in this work. Section III illustrates the experimental setup. In section IV, we present the results of our study, which are finally discussed in section V.

II. Methodology

The aim of this work is to enable the robot to learn to predict sensory consequences of its own motor actions. For that, a mapping between current sensory state plus intended motor actions and resulting sensory state has to be learned. This notion resembles that of a forward model, depicted in figure 1, an internal model that can predict the sensory consequences of an action [18]. In this work though, other than in the usual case, the modalities of the current sensory state (t) and the predicted sensory state (t+1) differ.

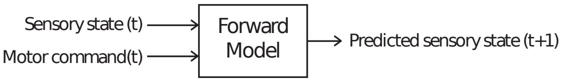

Fig. 1. Schematic representation of a forward model

Our model takes as inputs the current robot joint configuration, i.e. the current *sensory* state, denoted by $S(t)$, and the intended motor command, denoted by $M(t)$. It outputs the predicted image of the robot's camera, i.e. the future *visual* state, denoted by $Img^*(t+1)$. We trained the model with sensorimotor and visual data generated executing a self-exploration behaviour (see section III). Once the mapping is learned with the forward model, it can be used for implementing self-monitoring mechanisms. Sensory states, in this case visual ones, can be anticipated when feeding the model with the current sensorimotor state and the intended motor command. We use this capability in two experiments. First, the predicted visual state $Img^*(t+1)$ is compared to the actual visual input $Img(t+1)$, resulting in a prediction error, that is then expected to serve as a cue for the unexpectedness of an event, which in turn enables the robot to distinguish between own actions and actions of external agents. This capability, as discussed in the introduction of this paper, is thought to be fundamental for the development of a *sense of agency*. The second experiment aims at supporting speculating about the development of a sense of *object permanence*. The training images are preprocessed so that they, and consequently also the predictions of the model, represent a segmentation of the visual field into the two categories *robot body part* and *background*. In the experiment, this segmentation is then used to replace the parts in the image input where robot body parts occlude the robot's sight with previously seen image data. This generates an enhanced internal visual representation of the world, where objects hidden by the robot's own body in the original image may be visible in the enhanced one.

The approach in this work differs from our previous work [13] in that here, we use a real robot instead of a simulation environment, hence deal with real (noisy) robot movements and images from its camera. Instead of applying a 20x20 grid on the images, they are slightly downsized to 128x128 pixels. Furthermore, the model learns to predict the actual visual outcome instead of the number of pixels that change per grid area. In order to achieve this, a convolutional neural network was used as a model instead of the simpler nearest-neighbor approach. For details, see the following subsections.

A. Robotic Platform

We used a SoftBank Robotics NAO humanoid robot. It has 25 degrees of freedom, 10 of which are arm joints. We used 4 (instead of 2 in [13]) of the 5 joints of the left arm, namely shoulder roll, shoulder pitch, elbow roll, and elbow yaw. We reckoned the 5th arm joint, wrist yaw, would not have significant visual impact. We recorded the visual input from the top (forehead) camera of the robot with an image size of 320x240 pixels. In order to control the robot movements and to retrieve its camera images, we used the python API of the NAOqi SDK v2.1.4 [2].

B. Model Architecture

In order to learn the mapping of the forward model

$$(S(t), M(t)) \quad \longrightarrow \quad Img(t+1)$$

we use a deep convolutional neural network (CNN). In many cases, these models are used for image classification, for example in [19], [20] or semantic segmentation [21], where the image data is (part of) the model input. Other use cases,

[2] http://doc.aldebaran.com/2-1/dev/python/

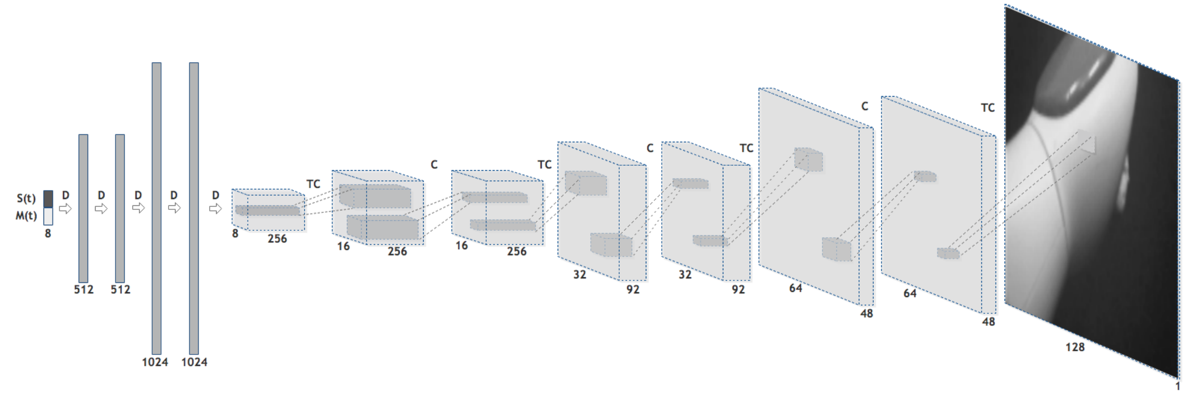

Fig. 2. Schematic representation of the convolutional neural network model architecture. S(t): sensory state at time t. M(t): Motor command sent at time t. D: Dense, i.e. fully connected, neural network layer. C: Convolutional neural network layer (kernel size: 3, stride: 1, 'same' padding). TC: Transposed Convolutional neural network layer (kernel size: 4, stride: 2, 'same' padding). Every layer except the last (output) one is followed by a ReLU activation unit (not shown).

conceptually related to this work, are image generation [22]. Even more closely related are works that aim at learning an implicit low-dimensional representation of image data, which can then also be used to generate new images [23]. Most of these approaches, however, operate in an unsupervised setting, i.e. only the image data itself is available for training, since it is often hard, expensive or infeasible to acquire large amounts of labelled image data. In our supervised setting on the other hand, the robot's self-exploration behaviour described in section III generates image data labelled with joint configuration data at no extra cost. Therefore, we build a CNN inspired by and very similar to the one in [24]. The model architecture is depicted in figure 2. The 8-dimensional input (4 dimensions each for $S(t)$ and $M(t)$) is followed by 5 fully connected layers comprising 512, 512, 1024, 1024, and finally 16,384 nodes. The final fully connected layer is reshaped to 8x8x256. Subsequently, we upsample the vector further by 3 transpose convolution layers (kernel size: 4, stride: 2, 'same' padding) each followed by a standard convolution layer (kernel size: 3, stride: 1, 'same' padding), step by step decreasing the number of channels from 256 to 48 while increasing the image dimension from 8x8 to 64x64. Finally, a last transpose convolution layer (kernel size: 4, stride: 2, 'same' padding) is added that outputs the resulting 128x128 one channel (i.e. grayscale) image representing $Img(t+1)$. Every layer except the last (output) one is followed by a ReLU activation unit.

III. Experiment

As described in section II, we implemented a self-exploration behaviour into the robot to generate the sensorimotor data needed for training the model. For simplicity, random motor babbling has been used as an exploration strategy, where a number of random commands are sent to the four joints. The commands were drawn uniformly from a restricted joint space in order to increase the relative time during training at which the robot's arm is actually visible. To further improve visibility of the robot arm, the NAO's head was turned to the left by 0.3 radians (head yaw).

Two different models were trained for the two different experiments. For both models, random motor babbling was implemented in front of a monochrome blue background, while 30 frames per second were sampled from the NAO's front camera. With each frame, the current joint configuration was recorded as well. A training sample then consisted of the joint configuration associated with the previous frame (4-dimensional) and the difference between current and previous joint configuration (4-dimensional) as inputs, plus the current recorded image as output.

For the sense of agency model, 500 motor commands were sent, resulting in 23,945 training samples. The recorded image frames were greyscaled and resized to 128x128 pixels. No further preprocessing was applied.

For the object permanence model, 200 motor commands were sent, resulting in 9,446 training samples. The recorded image frames were greyscaled and resized to 128x128 pixels. In order to segment the images into background and robot arm, further preprocessing was applied. Gaussian blur was applied to each image. The absolute difference of this result and a pre-recorded image showing only the background was taken and thresholded, resulting in strictly black and white segmented images, i.e. each pixel was either black (representing background) or white (representing the robot arm).

Each model was trained with shuffled training samples and a batch size of 128 for 100 epochs using the Adam optimizer [25].

A. *Sense of Agency*

For the sense of agency experiment, 20 trials were recorded. In each trial, one random motor command, drawn from the same uniform distribution as during training, was executed under two conditions. In the first condition, at most the robot arm movement was present in the visual field of the robot. In the second condition, additional external movement was present in the scene, for example the movement of a human hand. In both cases, the experiment was executed with a slightly tilted robot head and in front of the same blue screen as in the training session, and the resulting images were recorded from the robot's forehead camera again at 30 frames per second. Note however, that the data recorded for the experiment was not part of the data used to train the predictive model.

Each recorded frame was then greyscaled and resized to 128x128 pixels to match the format of the model output. The result was then compared to the prediction of the trained model given the input angles associated with the recorded frame, yielding a prediction error (simply the absolute difference between image and prediction) for each pixel. Taking the mean error over all pixels gives a prediction error for each frame. The two conditions are then compared in terms of this sample of prediction errors, where the sample size equals the number of frames in the trial.

Figure 3 shows the four steps. First, the originally recorded image. Second, the greyscaled images. Third, the prediction of the model, and finally the error map, showing the absolute difference between the greyscaled image and the prediction, thresholded, i.e. only pixels with a difference greater than a thresholding value are shown. The expectation is that the prediction error distributions under the two conditions differ significantly, with the prediction error being higher on average under the second condition. This would support the claim that the prediction error may be used as a cue for sense of agency, where a higher prediction error signifies a higher amount of "surprise" and therefore a higher probability of the presence of external agent activity.

B. Object Permanence

For the object permanence experiment, 20 trials were recorded in front of a blue screen, with a round object, for example a red ball, present in the robot's visual field at different locations. In each trial, one random motor command, drawn from the same uniform distribution as in the training part, was executed. The commands were selected such that the robot arm was visible to the robot and that it would occlude the robot's sight of the round object in part of the image sequence. A subsequence was then chosen around this part in order to ensure relevance of the trial for the task of maintaining object permanence.

For each frame, the joint configuration associated with it was fed as input to the forward model to generate a prediction of where in the input image the robot arm would occlude the robot's sight. The prediction (128x128 pixels) is then resized to 320x240 pixels in order to match the shape of the images recorded by the camera. Recall that the forward model used for the object permanence experiment was trained with black and white segmentation images where black pixels represented background, while white pixels represented a robot body part, but the predictions are float-valued. Therefore, the predictions were normalized to the usual range of greyscale images (0 - 255) and then thresholded in order to get a true mask image that assigns to each pixel a true or false value representing whether the model predicts that this pixel would be occluded by the NAO's arm or not.

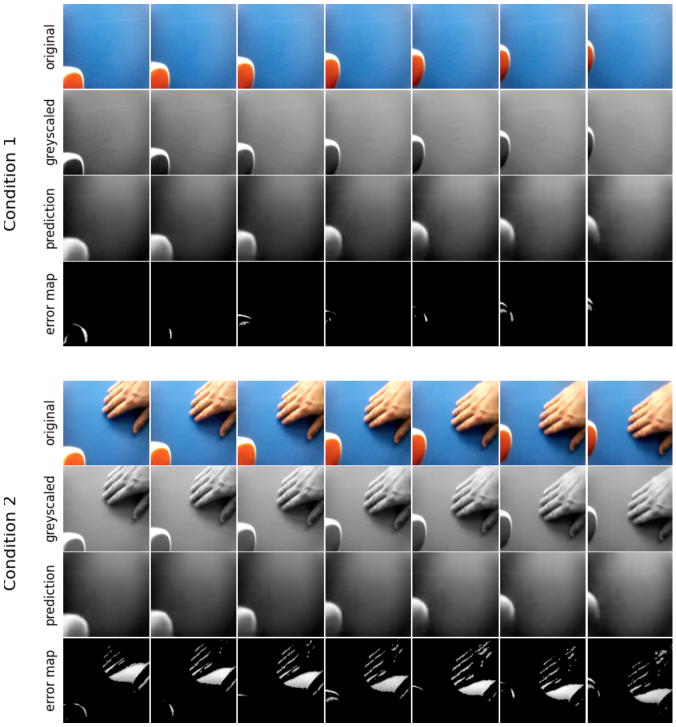

Fig. 3. Example sequence showing one partial trial of the sense of agency experiment under two conditions. First row: original image resized to 128x128 pixels. Second row: Greyscaled original image. Third row: Prediction of the forward model. Fourth row: Error map showing thresholded absolute difference between prediction and greyscaled image.

Using this mask, two approaches to obtain an enhanced image serving as an enhanced internal representation of the visual input for the robot were tested. Following the first approach, the pixels predicted to be occluded by the robot hand were simply replaced by the pixels from a pre-recorded static image showing the experimental setup without any robot body parts, i.e. the blue background with the additional round object. Following the more complex second approach, the pixels predicted to be occluded by the robot hand were replaced by the pixels from a dynamically maintained background image that was updated in each time step at the locations where we could be sure, according to the prediction, that the robot's view was not occluded. The second approach is conceptually closer to what we suggest is happening in reality and may prove useful in future experiments, where the background might change over time within one trial, for example when the object to be recognized is moving.

In order to quantify the effectiveness of the approaches, the Hough circle detection algorithm provided by the opencv-python library (v3.4.0.12) was applied to both the original and the enhanced image in every time step to determine whether a round object is visible or not. For an effective technique, the expectation is that the circle detection rate is significantly higher for the enhanced images.

Figure 4 shows the four steps. First, the original image. Second, the prediction by the forward model. Third, the normalized and thresholded mask. In the fourth row, the enhanced image obtained when using the first approach and finally the enhanced image obtained when using the second approach. The reason for resizing the prediction to 320x240 pixels rather than resizing the recorded image to 128x128 pixels is that the circle detection algorithm works much better when the original aspect ratio is preserved.

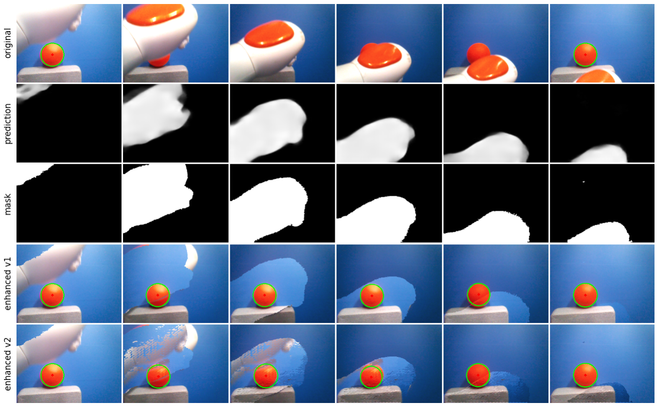

Fig. 4. Example sequence showing one partial trial of the object permanence experiment (showing every second frame). First row: original image. Second row: Prediction of the forward model. Third row: Normalized and thresholded prediction (mask). Fourth row: Enhanced image (first approach). Fifth row: Enhanced image (second approach). A green circle indicates that the red ball was detected.

TABLE I
SENSE OF AGENCY RESULTS

Trial	p-value	Condition 1		Condition 2	
		MEAN	STD	MEAN	STD
1	4.19e-01	19.5	6.58	20.85	7.76
2	2.83e-03	18.22	6.74	21.87	7.9
3	1.89e-12	11.55	2.15	31.7	11.99
4	1.47e-16	11.09	0.75	22.34	8.09
5	4.54e-18	13.56	2.28	25.84	7.46
6	1.47e-06	16.58	2.95	24.32	7.62
7	3.32e-06	17.45	6.39	28.35	11.99
8	4.24e-13	11.53	1.85	20.89	8.71
9	1.48e-10	16.21	6.61	24.64	3.0
10	1.14e-25	13.56	7.85	31.61	9.05
11	2.57e-14	17.19	6.37	27.41	7.32
12	5.99e-11	20.75	5.26	29.78	5.83
13	1.16e-15	10.51	0.4	26.59	13.47
14	6.65e-23	13.83	5.03	30.28	6.81
15	4.92e-11	11.36	1.39	28.42	10.74
16	4.66e-20	10.83	0.69	33.97	13.14
17	4.75e-03	14.21	5.4	25.49	20.01
18	2.89e-01	27.08	12.49	30.9	11.68
19	4.86e-53	12.53	1.02	31.69	3.32
20	3.35e-12	18.32	6.92	29.88	3.86

TABLE II
OBJECT PERMANENCE RESULTS

algorithm	v1		v2	
	Original	Enhanced	Original	Enhanced
average detection rate	28.03%	80.32%	28.03%	75.62%
p-value	6.11e-12		8.75e-10	

IV. RESULTS

A. Sense of Agency

Table I shows the results of the 20 sense of agency trials. For each condition, the mean and the standard deviation of the frame prediction error are shown. Frame prediction error refers to the sample vector of prediction errors, where each vector entry corresponds to one frame of the sequence, and its value is the average over all pixel prediction errors in that frame. In all 20 trials, the mean frame prediction error is greater under the second condition where external agent activity was present in the robot's visual field. In 18 out of 20 trials, the frame prediction error distribution differs significantly, with a p-value less than 0.05, supporting the claim that it can serve as a cue for a sense of agency. We suspect that the reason why the p-value is larger than 0.05 in two of the trials (trial 1 and trial 18) is that in these trials, the robot arm was very close to the robot's camera, thereby obstructing the robot's view of the external activity. Indeed, the external objects in both these trials were visible in less than 50% of the frames. Selecting the subsequences where external activity was present and running the same sense of agency experiment on the subsequences results in p-values of $0.024 < 0.05$ and $5.08e\text{-}4 < 0.05$ for trial 1 and trial 18, respectively.

B. Object Permanence

Table II shows the aggregated results of the object permanence experiment with the first and second version of the algorithm. 20 trials were run as described in section III. In every trial and for both approaches, the ball detection rate was higher for the enhanced image than for the original image. On average, the ball was detected in 28.03% of the original frames, 80.32% of the frames enhanced using the first approach, and 75.62% of the frames enhanced using the second approach. This shows that the proposed mechanism works well for developing a sense of object permanence.

We suspect there are two reasons, why the results are not perfect, i.e. 100% ball detection rate in every trial for the enhanced images. First, the circle detection algorithm does not work perfectly on these real image data. Its parameters could be fine-tuned in order to further improve the result. Second, the prediction of the NAO's arm position does not always match the actual arm position perfectly. See the first frame in figure 4 as an example. This is especially relevant when the second pixel replacement strategy is used. Future work could address this issue by increasing the size and variance of the training data set and by improving the segmentation algorithm in order to obtain less noisy training data.

V. CONCLUSIONS

We studied the development of fundamental cognitive skills in robots, such as the capability to distinguish between actions generated by the robot itself and those generated by a human subject, and the capability to generate and to maintain enhanced perceptual information, where objects occluded by the robot's own body may still be visible. For this, a deep convolutional neural network was trained as a forward model to be able to predict visual consequences of robot actions from the current sensorimotor state and intended movement of the robot. We showed that these predictions can be employed in two ways. Their absolute difference to the actual visual input, i.e. the prediction error, serves as a cue for a sense of agency for the robot. Furthermore, the predictions can be used to generate an enhanced internal visual representation of the world, in which objects originally obstructed by the robot's body parts are still visible, which enables the robot to develop a sense of object permanence.

These results confirm those of our previous work, presented in [13]: predictive models trained on self-generated sensorimotor data can represent promising tools for the implementation of basic cognitive capabilities in artificial systems. As described in the previous sections, we extended the work presented in [13] as follows: 1) we adopted a more sophisticated forward model - implemented as a deep convolutional neural network; 2) we used more complex real-world and real-robot sensorimotor data for training and testing the model. The deep convolutional neural network allowed us to work on high-dimensional data (e.g. raw images), instead of hard-coded features extracted from them (as in [13]). Moreover, generating predictions on the high-dimensional image space can allow easier visualisation of the outcomes of the model.

To further improve these results in terms of stability and detection rate, different model architectures and improved training may be considered as well as superior algorithms for object detection. There are many potential extensions to this work that shall be considered in the future. More efficient and intelligent exploration strategies than random babbling [12] could be used. In both training and experimental sessions, more complex visual scenes would increase the difficulty of the problem. This could be due to different backgrounds or the robot moving around in its environment. For the object permanence experiment, the object to be detected might be moving instead of staying at the same position. Another challenge would be to increase the number of degrees of freedom, for example by unlocking the NAO's head joints or involving its right arm, too. In order to deal with these extensions, more complex models might be needed. It could be especially useful to have the visual modality as an additional input to the forward model. Finally, we expect the same approach to work with other robots, too. Given proprioceptive data as well as the corresponding visual data catching the robots' body parts, the same CNN forward model can be applied, potentially with minor changes in the architecture accounting for different input and output dimensionalities.

Acknowledgment

This work has partially received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 773875 (EU-H2020 ROMI, Robotics for Microfarms).

References

[1] P. C. Fletcher and C. D. Frith, "Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia," *Nature Reviews Neuroscience*, vol. 10, no. 1, p. 48, 2009.

[2] A. Clark, *Embodied prediction*. Open MIND. Frankfurt am Main: MIND Group, 2015.

[3] S. Gallagher, "Philosophical conceptions of the self: implications for cognitive science," *Trends in Cognitive Neuroscience*, vol. 4, no. 1, pp. 14–21, January 2000.

[4] K. Friston, "Prediction, perception and agency," *International Journal of Psychophysiology*, vol. 83, no. 2, pp. 248–252, 2012.

[5] F. Picard and A. Craig, "Ecstatic epileptic seizures: a potential window on the neural basis for human self-awareness," *Epilepsy & Behavior*, vol. 16, no. 3, pp. 539–546, 2009.

[6] D. D. Ridder, J. Verplaetse, and S. Vanneste, "The predictive brain and the "free will" illusion," *Frontiers in psychology*, vol. 4, p. 131, 2013.

[7] S. S. Shergill, G. Samson, P. M. Bays, C. D. Frith, and D. M. Wolpert, "Evidence for sensory prediction deficits in schizophrenia," *American Journal of Psychiatry*, vol. 162, no. 12, pp. 2384–2386, 2005.

[8] E. Van Den Bos and M. Jeannerod, "Sense of body and sense of action both contribute to self-recognition," *Cognition*, vol. 85, no. 2, pp. 177–187, 2002.

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

[10] A. Cangelosi and M. Schlesinger, *Developmental robotics: From babies to robots*. MIT Press, 2015.

[11] G. Schillaci, V. V. Hafner, and B. Lara, "Re-enacting sensorimotor experience for cognition," *Frontiers in Robotics and AI*, vol. 3, no. 77, 2016.

[12] ——, "Exploration behaviors, body representations, and simulation processes for the development of cognition in artificial agents," *Frontiers in Robotics and AI*, vol. 3, no. 39, 2016.

[13] S. Bechtle, G. Schillaci, and V. V. Hafner, "On the sense of agency and of object permanence in robots," in *ICDL-EpiRob, 2016*. IEEE, 2016, pp. 166–171.

[14] G. Schillaci, C. N. Ritter, V. V. Hafner, and B. Lara, "Body representations for robot ego-noise modelling and prediction. towards the development of a sense of agency in artificial agents," in *International Conference on the Simulation and Synthesis of Living Systems (ALife XV)*, 2016, pp. 390–397.

[15] T. Ahsen, M. O'Brien, and E. Lanzilla, "Object permanence in human development and robotics - a survey," 2017.

[16] E. C. Leek, C. J. Atherton, and G. Thierry, "Computational mechanisms of object constancy for visual recognition revealed by event-related potentials," *Vision Research*, vol. 47, no. 5, pp. 706 – 713, 2007.

[17] H. E. Schendan and C. E. Stern, "Where vision meets memory: prefrontal–posterior networks for visual object constancy during categorization and recognition," *Cerebral Cortex*, vol. 18, no. 7, pp. 1695–1711, 2007.

[18] D. M. Wolpert, Z. Ghahramani, and J. R. Flanagan, "Perspective and problems in motor learning," *Trends in Cognitive Sciences*, vol. 5, no. 11, pp. 487–494, 2001.

[19] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back, "Face recognition: a convolutional neural-network approach," *IEEE Transactions on Neural Networks*, vol. 8, no. 1, pp. 98–113, Jan 1997.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *NIPS 25*, 2012, pp. 1097–1105.

[21] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640–651, April 2017.

[22] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *ICLR*, 2016.

[23] X. Chen, X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," in *NIPS 29*, 2016, pp. 2172–2180.

[24] A. Dosovitskiy, J. T. Springenberg, M. Tatarchenko, and T. Brox, "Learning to generate chairs, tables and cars with convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 692–705, 2017.

[25] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 12 2014.