Enabling the Sense of Self in a Dual-Arm Robot

Ali AlQallaf[1] and Gerardo Aragon-Camarasa[1]

***Abstract*—While humans are aware of their body and capabilities, robots are not. To address this, we propose a first step towards a basic, minimal self-awareness in a robot. That is, we propose an experimental methodology to evaluate whether the robot can differentiate itself from the environment, and to test whether *artificial self-awareness increases a robot's self-certainty in an unseen environment*. For this, we implemented a simple neural network architecture that enables a dual-arm robot to differentiate its limbs from an environment using visual and proprioception sensory inputs. The proposed experimental approach allows us to evaluate whether the robot can differentiate itself from the environment. Our results indicate that a robot can distinguish itself with an accuracy of 88.7% on average in different environmental settings and under confounding input signals.**

I. INTRODUCTION

When we become self-aware, we can recognise ourselves in any environment. This is possible because we can distinguish our body as a separate entity from the world; allowing us to adapt to different situations and scenarios. Robots, however, lack this capability because they are limited to fixed configurations, engineered to work in constrained environments. Researchers have theorised [1], [2], [3], [4] that an adaptable robot can increase its productivity, and that a self-aware robot can increase the task efficiency over different settings and environments. Hafner et al. [5] has stated that the pathway to reach a similar level of human performance in robotics, a robot requires to have a minimal self and, ultimately, knowledge of self. Current research in self-awareness for robotics [2], [6] has focused on enabling robots and agents to acquire self-awareness by interacting with the environment and task using end-to-end approaches. Self-awareness is, however, learned as a byproduct of the robot interacting with the environment in order to increase its autonomy.

Rochat [7] has argued that self-awareness in humans is an incremental process which humans start acquiring it by learning their body and capabilities first, then adapt their self-agency within the environment to fulfil high-level tasks. Rochat [7] has proposed five levels of self-awareness, each representing a competence that humans use to learn and adapt to its body and then to environments. In this paper, we propose that a robot starts by learning how to construct a sense of *self*, before interacting and dealing with the environment and objects, as shown in Fig. 1.

Our approach contrasts to previous and current approaches, where the self is built following a top-down approach via the interaction with the environment [4], [8], [9], [10]. Hence, we investigate the first basic level of self-awareness which serves as the building block for enabling a robot to become an adaptable and flexible autonomous machine. For this, the robot needs to correlate its visual and internal sensing modalities to initiate the sense of self. We frame this basic level of self-awareness as a binary classification task in which we let the robot to answer whether it can differentiate itself as an entity in an environment with a degree of certainty (i.e. certainty is the accuracy of the classification prediction).

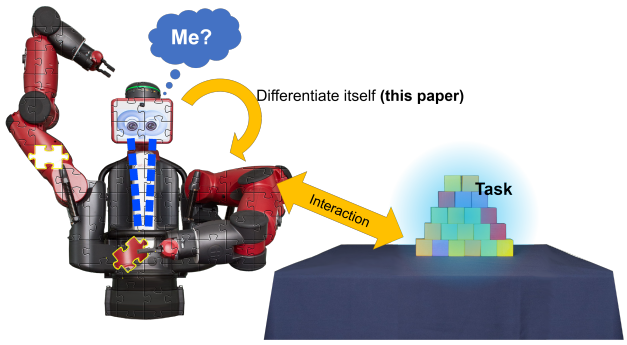

Fig. 1. A robot differentiates, recognises and situates itself first with its body, and then interacts with the environment.

Rochat's self-awareness proposition consists of a continuous development of human's self; in this paper, we frame this development as discretised levels in which each level represents a degree of self-awareness competence. For level 1 - differentiation, we therefore designed a simple multi modal neural network and a binary classification task to allow us to understand how a sense of self can be elicited within a neural network, and ultimately how we can endow a robot with the basic, minimal concept of self [5]. That is, our hypothesis in this paper is that *Level 1 for artificial self-awareness in the robot increases its self-certainty in an unseen environment*. We evaluate our approach to artificial self-awareness in a dual-arm robot in four different experimental scenarios and carry out an ablation study to investigate whether the robot can differentiate itself with a certain degree of certainty while presented with confounding signals. Our contributions are therefore threefold:

1) We propose an experimental methodology that allows us to test whether a neural network and a robot can learn (supervised) the sense of self through differentiation.
2) We compiled a dataset comprising synchronised RGB images (egocentric view) and proprioception sensor readings of a dual-arm robot which is used to test our hypothesis. Also, this dataset can be utilised for other multimodal integration studies.
3) A real robot demonstration showing the robot being able

[1] Computer Vision and Autonomous group, School of Computing Science, University of Glasgow. Email: a.alqallaf.1@research.gla.ac.uk, gerardo.aragoncamarasa@glasgow.ac.uk

to differentiate itself.

Our code and dataset for this paper are available at https://github.com/cvas-ug/towards-sa.

II. RELATED WORK

Rochat [7] has classified self-awareness into five levels, starting from sensing self as a separate entity in the world (Level 1) to self-consciousness (Level 5). Later, Rochat [11] proposed that *self-unity* (Level 0) is the primary phase of newborns which comprises the initial experience of sensory during the first hours of life, and concluded that self-unity could endow machines to learn about their body within an environment. The ordering of the five levels of self-awareness is based on their relative complexity and are further divided into implicit (from zero to two) and explicit (from three to five) levels [7], [12]. That is, Legrain et al. [12] have formulated that the implicit self-awareness levels are related to correlating the internal states with the body based on the experience of the self within the environment. The explicit self-awareness levels are those that link the environment to how the environment influences the person. In this paper, we focus on the first self-awareness level (Level 1) and, for completeness, we summarise the implicit levels of self-awareness according to [7]:

- Level Zero – "Self-unity". An individual is born with basic multi-sensory and motor control capabilities which they use to learn about itself.
- Level One – "Differentiation". The individual gets a sense that there is something unique in the experience between what is out there and the felt movements to initiate the sense of self (the focus of this paper but in robots).
- Level Two – "Situation". An individual situates within its body by experiencing the relationship between seen movements and body stimulation over time.

In robotics, Torras [3] and Chatila et al. [4] have stated that there is a need for robots to be capable of handling different environments while showing high adaptability to any environment. However, Agostini et al. [13] have argued that robots cannot accommodate all human environments, and hard-coding all possible situations is a challenging task. To mitigate this, researchers [6], [14] have proposed to learn an awareness model inspired by the free-energy principle [15] in robotics which states that the interactions with the environment are aimed at reducing the internal entropy (i.e. maximising the robot's self-certainty) of an agent. For example, [6], [14] has shown that a robot or its environment might change, and the capability of the robot to adapt to different environments is predicated on the assumption that a robot learns continuously using an active inference model. They thus enabled a robot to adjust its control to the task at hand by minimising the distance between the robot's hand and the target object [14] or where the robot's hand is to its internal belief [6]. However, the authors constrained the robot to have reduced visual perception capabilities in order to simplify the inference task, relying on an observed action within an uncluttered, simple operating environment.

Kwiatkowski et al. [2] have shown that a robot can model itself without prior knowledge of its structure, and constructs a self-model that can adapt to mechanical changes that occur to the robot. Their work has demonstrated that self-modelling is the conduit to adaptable and resilient robotic systems. However, the proposed self-model architecture learns about the robot's internal mechanical structure, and it is not able to make a distinction of itself as an entity in the environment without being explicitly defined. The basic robot's existence as an entity reflects the first level of self-awareness, and Kwiatkowski's self-model is not aware of the distinction between itself and the environment. In this paper, we, therefore, propose that a robot learns how to distinguish itself from the environment before acting. For this, we investigate and develop the first level of self-awareness [7] and demonstrate that a robot can get a sense of itself by simplifying the learning task to distinguishing itself in different contexts.

Hafner et al. [5] have reviewed biological studies and robotics studies that are related to self. The authors have concluded that the self-exploration of behaviours, body representations, and sensory-motor simulations and predictive processes are the three components that could represent self-agency in a robotic system. They have also suggested that self-agency can be measured as the prediction error. The latter is similar to what Amos et al. [16] have demonstrated where they based self-awareness on predictive control models to allow a robot to create a link between itself and the environment. Similarly, Haber et al. [17] has developed an intrinsically motivated agent by using world-model predictions via a supervised learning strategy to model agent awareness in order to generate different behaviours in complex environments. Similar to this work, Lanillos and Cheng [18] have used a hierarchical Bayesian computational model to define the self in a robot and argued that the understanding of sensory mapping changes is core to self-perception. Also, Gold and Scassellati [19] have observed that motor actions and visual motion using probabilistic reasoning can allow a robot to self-recognise in front of a mirror and test for self-recognition of their robot's self. In this paper, we considered more environments go beyond simpler backgrounds as we include clutter within the environment in order to introduce distractors while acquiring the self. We also use a deep neural network to fuse visual and proprioception inputs to acquire a sense of self. We must note that their approaches require the robot induce motion to achieve a prediction of 81.9% for self-distinction [18].

III. MATERIALS & METHODS

A. Design and Rationale

Our approach to artificial self-awareness focuses on building an initial sense of self in the robot by enabling it to differentiate itself (i.e. Level One in Rochat's self-awareness levels, Section II) from the environment using proprioception and vision. For this, we design a Deep Neural Network architecture (Fig. 2) to support and understand the *sense of self* in the robot. The levels of implicit self-awareness (Section II) inspire

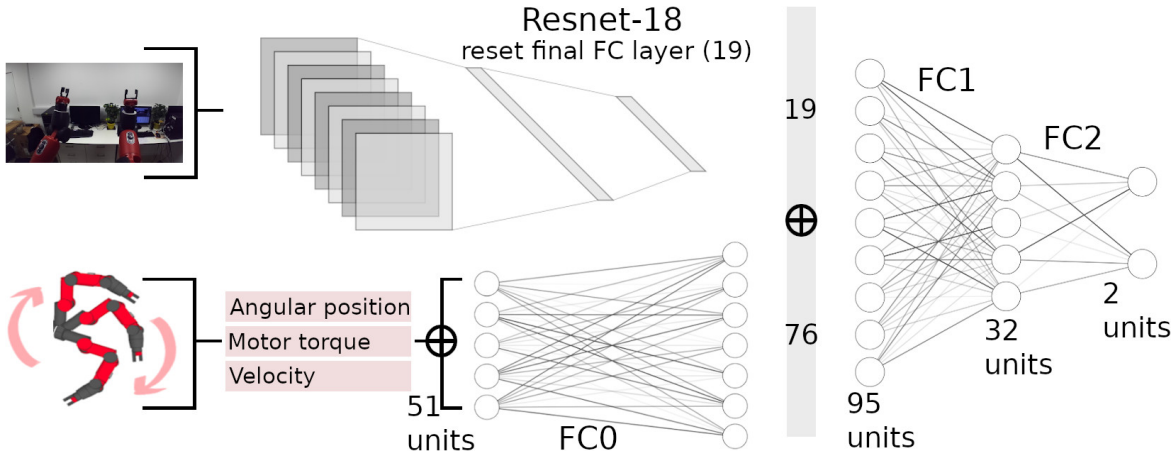

Fig. 2. The Level 1 architecture combines vision and proprioception inputs of the robot sensors to predict *self* or *environment*. As shown above, the model process vision and proprioception through two subnetworks (Resnet18 and Linear layer, respectively), then we concatenate the output features from both subnetworks and then passed to 3 fully connected layers to carry out a prediction.

our architecture design, and we, therefore, propose that these implicit levels can be mapped for robots as follows:

- Level 0 – "Self-Unity": We propose that this level corresponds to the robot's physical, mechanical and sensory capabilities. That is, the required sensing and motor devices and software that allow a robot to deal with the world, e.g. robot's kinematics, dynamics, sensor definition and configuration, motion planning, etc. These capabilities are interfaced via software APIs and software drivers, e.g. the Robot Operating System (ROS)[20].
- Level 1 – "Differentiation": This level is the initial self, and we propose that this level is about learning how to differentiate itself by seeing its arms and gripers in association with its proprioception without temporal connection between observations. The assumption at this level is that the robot has a high-level description of its limbs via forward and inverse kinematics, and can move its arms via motion planning. A binary classification therefore enables a robot to have a high-level prediction of its sense of self. The objective is then to confirm if the arms and grippers belong to the robot.

In Level 1, the output is the initiation of self and consists of sensing the distinction between robot and environment. In Level One for humans (Section II), [7] stated that the individual starts to think that there is a unique experience on what the person sees with respect to its proprioception. Hence, we anticipate that vision and body movement are crucial components to initiate the self in a robot. Similarly, the human brain is a prediction machine that makes inference based on mapping different sensory inputs [6], [15]; therefore, in our approach, the prediction of *self* depends mainly on two elements: the presence of the limbs in the robot's field of view and the sense of its movement.

The rationale behind our approach is to define a neural network architecture that provides a way to learn the first level of self-awareness and to understand the internal mechanisms of level one - Differentiation. The predicted output of the neural network is, therefore, a supervised binary classification task that predicts the sense of self of the robot.

B. Implementation

To achieve Level 1, the robot uses its visual sense to discriminate its limbs together with proprioception. For this, we used the robot's vision and proprioception capabilities as the sensory inputs and combined proprioception with static visual representations to get an initial snapshot of self. Vision comprises RGB images captured using a stereo ZED camera from Stereolabs configured to output images at 720p resolution. Captured images contain a representation of the robot's arms or environment. Proprioception consists of the robot's joint states; being velocity, angular position, and motor torque.

Our architecture for Level 1 of self-awareness is shown in Fig. 2) and consists of a Resnet18 network [21] to process the visual state of the robot. Resnet18 is a state-of-the-art architecture used widely for object detection and classification. Similarly, for proprioception, we used a single, fully connected network layer - $FC0$ to process the internal state of the robot. The output from Resnet18 is a tensor size of 19 that is concatenated with the output of $FC0$ proprioception tensor of size 76. The concatenated tensor is of size 95 and is passed to a fully connected layer - $FC1$ followed by a *Relu* activation function. The output of *Relu* is a tensor of 32 that inputs into the fully connected layer, $FC2$, that predicts *self* or *environment*.

We implemented a ROS node to capture and store synchronised visual and proprioceptive sensor information. During data capturing, we recorded Baxter moving its hands using random gripper poses for both arms and within its working volume. A total of 30k images and proprioception states were captured over four different environmental settings, as shown in Fig. 3. Each scene represents a unique group that ranges from simple (front towel and front glass) to cluttered (*front computers* and *in lab*) environments. Each experimental

Fig. 3. Sample images from captured scenes, ref. Table I

TABLE I
EXPERIMENTAL GROUPS AND UNSEEN TEST GROUP DATASETS

	Experimental Groups Sets	Unseen Test Group
Group-1	{ In lab, Front glass, Front towel }	Front computers
Group-2	{ Front computers, In lab, Front glass }	Front towel
Group-3	{ Front computers, Front glass, Front towel }	In lab
Group-4	{ Front computers, In lab, Front towel }	Front glass

group includes two classification labels, namely *self* and *environment*.

An experimental group is a combination of three scenes while leaving one out for testing purposes, as shown in Table I. The objective is to have broader and diverse data groups for training our proposed architecture (Fig. 2). We split 80/20 proportions each experimental group for training and validation, respectively [22]. Accordingly, the training and validation sets represent about 20k and 5k images and proprioception states, respectively. We used PyTorch to implement our level 1 of artificial self-awareness architecture with PyTorch's default cross-entropy loss. We used a pre-trained version of Resnet18 and fine-tuned it with our dataset. Training consisted of 24 epochs each with a 64 batch and a learning rate of 0.001.

IV. RESULTS & DISCUSSION

The robot's capabilities frame our experimental design. That is, the robot cannot move to a different place in the room. Similarly, the robot's vision sensor is fixed on top of the robot's head and cannot actively move its head. The robot can move its arms freely within its predefined working volume, and there are no obstacles included in each of the experimental groups in Table I. According to Rochat [7], newborns wave their hands randomly in order to try to identify the objects in front of them. We, therefore, commanded the robot to wave its limbs without a predefined task in the environment to enable the robot to learn to perceive and differentiate itself from the environment.

To further test our hypothesis that *Level 1 for artificial self-awareness in the robot increases its self-certainty in an unseen environment*, we carry out an ablation study represented by four confounding cases to understand the effectiveness of the combination of the proprioception and vision within our proposed model. Therefore, this experiment framework consists of four confounding experimental cases (Table III) that compare unseen experimental groups against confounding perceptual signals. The objective is *to confirm that the robot can differentiate itself with a degree of certainty while presented with confounding sensor signals*. Thus, *Case-1* comprises the unseen test group where images and proprioception corresponds to the *self* class; while *Case-2*, images and proprioception correspond to the *environment* class. *Case-3* comprises confounding samples where the robot's arms are in the visual field of the robot, but the robot's proprioception corresponds to the *environment* class. While *Case-4* is composed of environment images but the robot's proprioception comes from the *self* class. The two sensory inputs are important for the model to produce the decision of self or environment. Based on the defined confounding cases, the model will most presumably output an environment as the classification decision if any sensory input fails.

To understand Level 1 and process it in different environments, we adopted a leave-one-out cross-validation strategy to test each trained experimental group (Table I). By having an unseen experimental group, we are able to verify the validity of our hypothesis. Accordingly, confusion matrices for each unseen test group in Table I are shown in Table II. The classification accuracy for each unseen test group is: Group-1 is 88.1%; Group-2, 90%, Group-3, 82.1%, and Group-4, 94.7%. We can, therefore, state that our architecture enables the robot to differentiate itself from the environment with an average certainty of 88.7%.

We have used FlashTorch [23] to gain insights on how our Level 1 approach perceives images by means of a saliency map, as shown in Fig. 4. A saliency map is an image showing which pixel regions the neural network focused on in order to predict the underlying output. Accordingly, FlashTorch allows us to gain insights on which regions in the image contributed the most to predict *self* or *environment*. For this, vision and proprioception are used to predict the class, but we only evaluate the vision component of our architecture with FlashTorch.

As observed in Fig. 4, our Level 1 architecture is biased towards bright colours and distinctive features that remained constant in the robot's visual field. We must note that clutter is semi-consistent in the scene during data capturing, i.e. clutter was randomly placed. For example, in the *Front Computers* test group (ref. Table I), 10.3% of the *environment* class is classified as *self* (Table II, unseen test groups row). The reason is that the network interpreted the yellow pot in Fig. 4-A as part of the robot (i.e. *self* class). Close inspection of Fig. 4-B and -C (*front towel group*) shows that when the robot's hands are within the field of view, environment features such as the edges of the towel do not contribute to predicting the correct

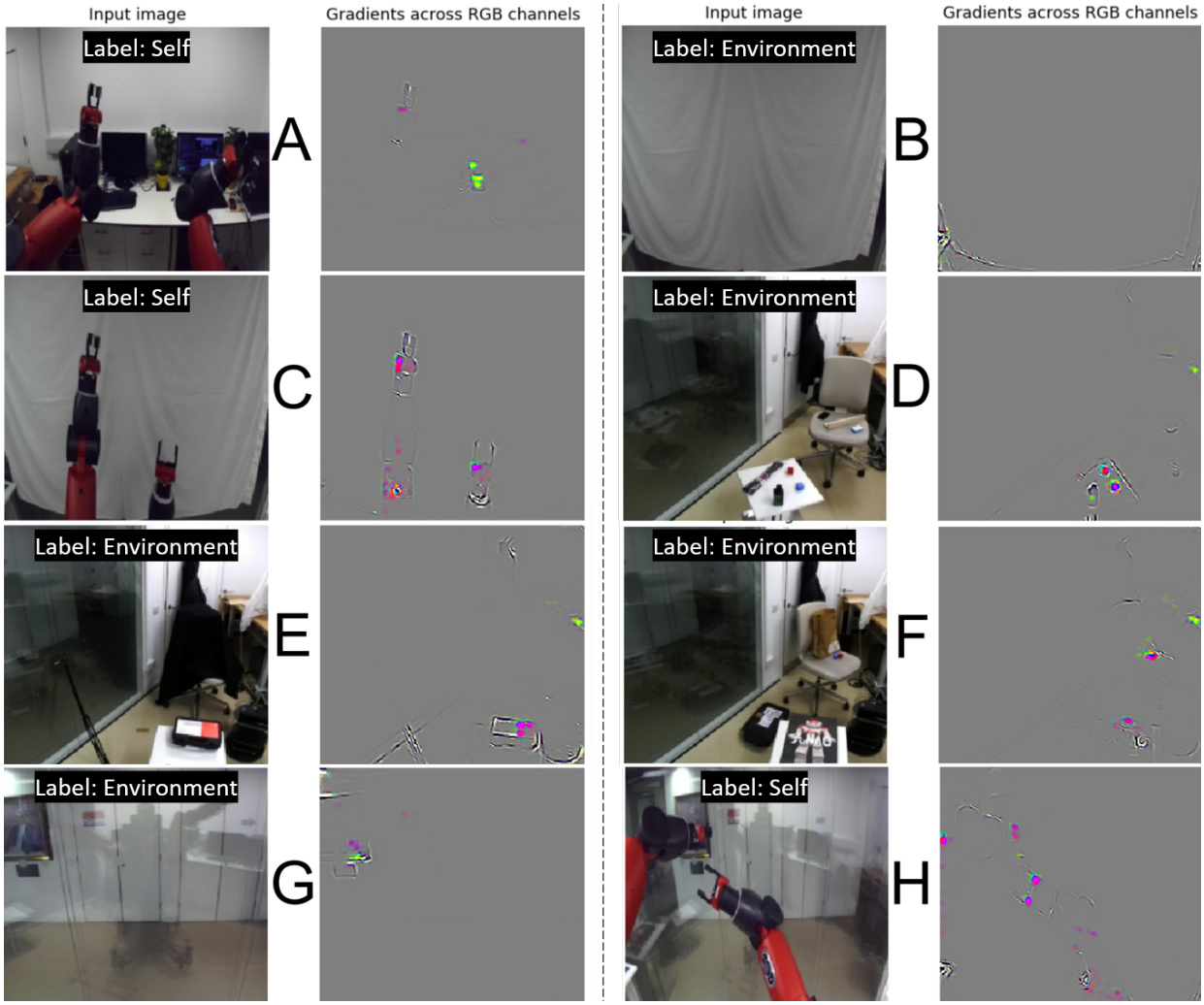

Fig. 4. This images are representing the saliency maps of different environment groups as described in Table I, were A corresponds to Group-1; B and C, to Group-2; D to F, to Group-3; and, G and H, to Group-4. For each group, the right image shows the predicted label, and the left images shows the regions the model focused on.

class, observing 90% accuracy for predicting the correct class. Figures 4-D, -E, and -F (*in lab* unseen test group) reveal the sensitivity of the network towards similar colours to the robot used in our experiments (i.e. red) and bright colours. The latter represents 17.9% of incorrect classifications. The *front glass* unseen test group (Fig. 4-G and -H) achieves a 94.7% of classification accuracy; however, our Level 1 architecture is biased towards bright regions in the images (i.e. picture frame in the background and table corner in the left-bottom).

The unseen test Group-1 (*front computers*) and Group-3 (*in lab*) have noticeable classification errors of 11.9% and 17.9%, respectively. To investigate these errors, we split the unseen test set into four confounding cases, as described in Table III. These results are shown in Table II, and reveal that Case-4, is the most misclassified case in both Group-1 and Group-3. The reason for these misclassifications is that Resnet18 is biased towards bright colours as discussed above. If we compare with Group-2 and Group-4 where the robot is facing uncluttered environments, Case-4 yields higher classification scores since visual false positives are kept at a minimum. That is, we can observe in Fig. 4-C and -B that when the robot's hands become more predominant on uncluttered backgrounds, our Level 1 architecture predicts the correct classification regardless of confounding signals coming from proprioception. We also noticed in Table II that proprioception signals have a high contribution to predicting the correct class. For instance, in case-3 where images contain the robot's arms, but proprioception corresponds to the *environment* class, our architecture can predict the correct class with high accuracy for all groups.

To further understand whether our Level 1 architecture learns to differentiate the robot from the environment, we computed the Mutual Information [24] for each group's train dataset (Table I). Our objective is to measure and compare if four Level 1 trained architectures have a degree of similar knowledge that it is invariant to the training set. Mutual

TABLE II
TEST GROUPS CLASSIFICATION ACCURACIES FOR EACH UNSEEN GROUP AND CONFOUNDING CASES.

		Group 1 Front Computers		Group 2 Front towel		Group 3 In lab		Group 4 Front Glass	
		Predicted Self	Predicted Env.	Predicted Self	Predicted Env.	Predicted Self	Predicted Env.	Predicted Self	Predicted Env.
Unseen test groups	True Self	23.4%	1.6%	21.5%	3.5%	23.5%	1.5%	23.8%	1.2%
	True Env.	10.3%	64.7%	6.5%	68.5%	16.4%	58.6%	4.1%	70.9%
Case 1 Class: Self	True Self	93.7%	6.3%	86.1%	13.9%	94.0%	6.0%	95.2%	4.8%
	True Env.	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Case 2 Class: Env.	True Self	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	True Env.	0.0%	100%	0.0%	100%	0.0%	100%	0.0%	100%
Case 3 Class: Env.	True Self	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	True Env.	0.0%	100%	2.4%	97.6%	0.2%	99.8%	2.2%	97.8%
Case 4 Class: Env.	True Self	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	True Env.	41.1%	58.9%	23.6%	76.4%	65.6%	34.4%	14.3%	85.7%

TABLE III
CONFOUNDING EXPERIMENTAL CASES

	Class	Description
Case-1	Self	Vision and proprioception correspond to the robot's arms being in the field of view
Case-2	Environment	Vision and proprioception correspond to the robot's arms not being in the field of view
Case-3	Environment	The robot's arms are in the field of view but the proprioception matches the environment class
Case-4	Environment	Proprioception corresponds to the self class but the robot's arms are not in the field of view

information allows us to compare multimodal sources and measure how well two sources are matched by mutual dependence between two variables. That is, different sources of information means more distributed points in the joint histogram and, consequently, low mutual information metric.

The spread in the joint histogram is associated with uncertainty, and in Fig. 5, joint histograms show minor variability in the correlation between the group's models weights. The latter shows that there are no significant differences between the trained models despite the differences in the training datasets (Table I), and the misclassification in the confusion matrices results (Table II) are based on the environment noise as other objects within the environment distract the network attention. Since mutual information is computed at the last layer of our architecture, proprioception is taken into account during the classification. Therefore, this demonstrates that our Level 1 network architecture captures a degree of self-awareness and, consequently, certainty. We can, therefore, conclude that our experimental hypothesis holds for the experiments presented in this paper because mutual information between architectures do not vary significantly, thus representing a basic, minimal self within the neural network and robot. In the companion video to this paper[1], we demonstrate our approach to Level 1 artificial self-awareness. That is, we use a Baxter robot to predict the sense of self while observing what is in front of

[1]https://youtu.be/woZUa2QWJxw

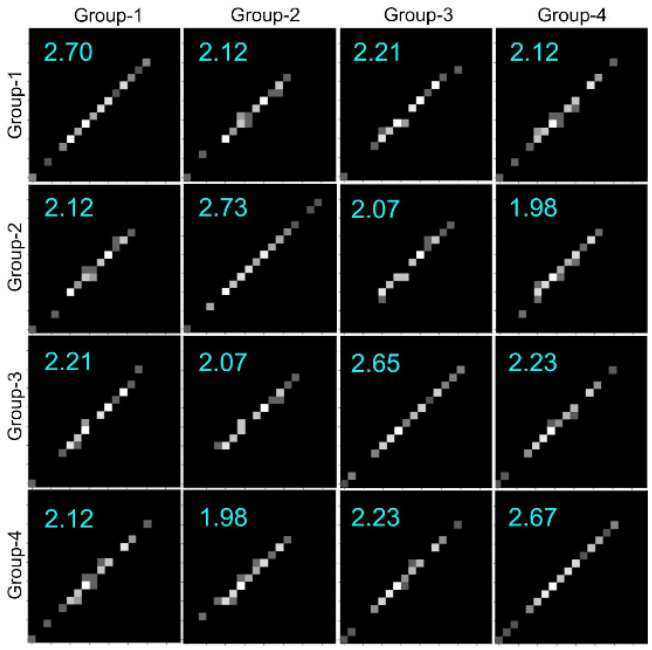

Fig. 5. Mutual information and joint 2D histograms of the trained weights for four Level 1 architectures. The mutual information is noted at the top left corner on each joint histogram plot.

it, i.e. self or environment.

V. CONCLUSIONS & FUTURE WORK

In this paper, we presented an approach to Level 1 of artificial self-awareness in a dual-arm robot. Our approach is inspired by the first level of self-awareness defined by Rochat [7]. By using vision and proprioception, we have demonstrated that a robot can differentiate itself from the environment with an average classification accuracy of 88.7% using unseen test samples and across four different scenes' groups presented in Table II.

An initial self is defined in the robot, but the robot cannot locate its limbs within the environment and put them into context for a task. Thus, future work comprises developing Level 2 (Situation; Section II) of artificial self-awareness. The idea is to employ temporal sequences of the robot's arms, and model visual and proprioception experiences in terms of a recurrent network architecture, which we believe is the next step to let a robot to be able to identify itself with higher self-certainty in an environment.

ACKNOWLEDGMENT

Ali AlQallaf thanks the Kuwait Institute for Scientific Research. We also thank the support of NVIDIA Corporation for the donation of the Titan Xp GPU used in this research.

REFERENCES

[1] J. P. Vasconez, G. A. Kantor, and F. A. A. Cheein, "Human–robot interaction in agriculture: A survey and current challenges," *Biosystems engineering*, vol. 179, pp. 35–48, 2019.

[2] R. Kwiatkowski and H. Lipson, "Task-agnostic self-modeling machines," *Science Robotics*, vol. 4, no. 26, 2019.

[3] C. Torras, "From the turing test to science fiction: The challenges of social robotics," in *Proceedings of the 16th International Conference of the Catalan Association of Artificial Intelligence*, 2013, pp. 5–7.

[4] R. Chatila, E. Renaudo, M. Andries, R.-O. Chavez-Garcia, P. Luce-Vayrac, R. Gottstein, R. Alami, A. Clodic, S. Devin, B. Girard, and M. Khamassi, "Toward self-aware robots," *Frontiers in Robotics and AI*, vol. 5, p. 88, 2018. [Online]. Available: https://www.frontiersin.org/article/10.3389/frobt.2018.00088

[5] V. V. Hafner, P. Loviken, A. Pico Villalpando, and G. Schillaci, "Prerequisites for an artificial self," *Frontiers in Neurorobotics*, vol. 14, p. 5, 2020. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnbot.2020.00005

[6] C. Sancaktar, M. A. van Gerven, and P. Lanillos, "End-to-end pixel-based deep active inference for body perception and action," in *2020 Joint IEEE 10th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*. IEEE, 2020, pp. 1–8.

[7] P. Rochat, "Five levels of self-awareness as they unfold early in life," *Consciousness and cognition*, vol. 12, no. 4, pp. 717–731, 2003.

[8] J. Tani, "An interpretation of the self from the dynamical systems perspective: a constructivist approach," *Journal of Consciousness Studies*, vol. 5, no. 5-6, pp. 516–542, 1998. [Online]. Available: https://www.ingentaconnect.com/content/imp/jcs/1998/00000005/f0020005/880

[9] Y. Nagai, Y. Kawai, and M. Asada, "Emergence of mirror neuron system: Immature vision leads to self-other correspondence," in *2011 IEEE International Conference on Development and Learning (ICDL)*, vol. 2, Aug 2011, pp. 1–6.

[10] P. Lanillos, J. Pages, and G. Cheng, "Robot self/other distinction: active inference meets neural networks learning in a mirror," 04 2020.

[11] P. Rochat, "Self-unity as ground zero of learning and development," *Frontiers in Psychology*, vol. 10, p. 414, 2019. [Online]. Available: https://www.frontiersin.org/article/10.3389/fpsyg.2019.00414

[12] L. Legrain, A. Cleeremans, and A. Destrebecqz, "Distinguishing three levels in explicit self-awareness," *Consciousness and Cognition*, vol. 20, no. 3, pp. 578–585, 2011.

[13] A. G. Agostini, C. Torras, and F. Wörgötter, "Integrating task planning and interactive learning for robots to work in human environments," in *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.

[14] P. Lanillos and G. Cheng, "Active inference with function learning for robot body perception," in *International Workshop on Continual Unsupervised Sensorimotor Learning, IEEE Developmental Learning and Epigenetic Robotics (ICDL-Epirob)*, 2018.

[15] K. Friston, "The free-energy principle: a unified brain theory?" *Nature reviews neuroscience*, vol. 11, no. 2, pp. 127–138, 2010.

[16] B. Amos, L. Dinh, S. Cabi, T. Rothörl, S. G. Colmenarejo, A. Muldal, T. Erez, Y. Tassa, N. de Freitas, and M. Denil, "Learning awareness models," in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018.

[17] N. Haber, D. Mrowca, S. Wang, L. Fei-Fei, and D. L. K. Yamins, "Learning to play with intrinsically-motivated, self-aware agents," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018, p. 8398–8409.

[18] P. Lanillos, E. Dean-Leon, and G. Cheng, "Yielding self-perception in robots through sensorimotor contingencies," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 9, no. 2, pp. 100–112, 2017.

[19] K. Gold and B. Scassellati, "A bayesian robot that distinguishes" self" from" other"," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 29, no. 29, 2007.

[20] "ROS.org | Powering the world's robots." [Online]. Available: https://www.ros.org/

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2016.

[22] I. Guyon, "A scaling law for the validation-set training-set size ratio," *AT&T Bell Laboratories*, pp. 1–11, 1997.

[23] M. Ogura and vainaijr, "Misaogura/flashtorch: 0.1.1," Sep. 2019. [Online]. Available: https://doi.org/10.5281/zenodo.3461737

[24] H. Fang, V. Wang, and M. Yamaguchi, "Dissecting deep learning networks—visualizing mutual information," *Entropy*, vol. 20, no. 11, p. 823, 2018.