Robot self-recognition via facial expression sensorimotor learning

Zhegong Shangguan[1], Mengyuan Ding[2], Chuang Yu[3] , Chaona Chen[4] and Adriana Tapus[1]

***Abstract*— To develop robots that can show cognitive functions, we must learn from the knowledge of human cognition. Existing biological and psychological evidence suggests that self-face perception and sensorimotor learning mechanisms play a crucial role in self-recognition. However, one of the most important self-identity cues – facial information – has not been extensively studied in the robot self-recognition task. Current research on robot self-recognition primarily relies on the recognition of high-precision targets and tracking of manipulator motions, where the self-perception of facial information is not well studied. In this work, we propose a novel approach to achieve self-recognition via self-perception of facial expressions. Specifically, we developed a Conditional Generative Adversarial Network (CGAN) model using the knowledge on human cognitive and sensorimotor functions. It allows the robot to be aware of self-face (i.e., off-line model). Passing the observed visual variations in a mirror and comparing them to self-perceptive information, the robot can recognize the self through an online Bayesian learning regression. The results of our first experiment show that the robot can recognize itself in a mirror. The results from the second experiment show that our algorithm could be tricked by a similar robot with the same facial expressions, which is similar to the rubber hand illusion (RHI).**

I. INTRODUCTION

Sensorimotor learning refers to a type of cognitive learning pattern in humans where an individual learns the correlation between their actions, the effects of their actions, and the sensory signals generated as a result [1]. This type of learning is often observed in infants as they begin to walk, with frequent touching of their own bodies creating interactions between multiple sensory subsystems [2]. Through this goal-directed process, infants gradually acquire the ability of self-recognition by learning the neural organization of multi-modal stimuli from physical activities. The Rubber Hand Illusion (RHI) is a fascinating phenomenon in human cognition, which occurs when visual illusions mislead individuals to perceive a rubber hand as part of their own body [3]. This phenomenon highlights the importance of sensorimotor learning in the formation of self-recognition among healthy adults. Self-recognition in primates is understood as the ability to become the object of one's own attention [4]. In order to replicate human consciousness and behavior, humanoid robots also need to possess the ability of self-recognition through sensorimotor learning.

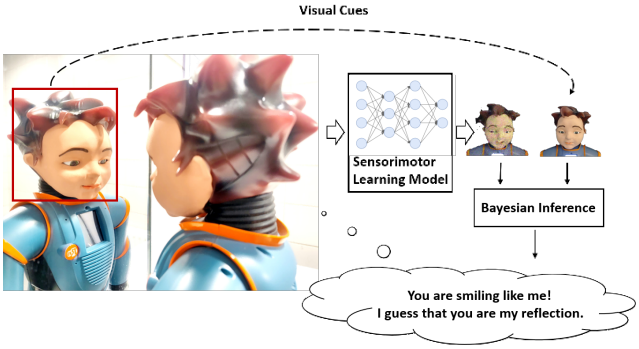


Fig. 1. Robot self-recognition. With the help of our developed algorithm, the robot can understand that the facial expressions variations in the mirror are the effect of its motor actions.

Apes and human beings possess the remarkable ability of self-recognition, as demonstrated by their ability to identify themselves and locate marks on their own bodies when presented with a mirror [4], [5]. This ability requires high-level cognitive processes such as self-awareness and self-consciousness. To assess animals' self-recognition abilities, researchers often use the Mirror Self-Recognition (MSR) test [4]. In robotic self-recognition tasks, researchers typically use body actions, such as robotic manipulators, and visual cues to enable humanoid robots to determine whether the motion changes appearing in a mirror are the result of their movements in the world [6], [7]. However, recent studies have shown that self-face recognition is a unique representation of oneself, possessing processing advantages over other faces. Recognition of one's own face is significantly faster and more accurate in various tasks than recognition of other faces [8], [9]. In fact, some patients are unable to recognize their faces when morphed with a famous face if their right hemispheres are anesthetized [10]. Self-face recognition is so distinctive that the ability to discriminate between one's own face and another is a subliminal process [11].

In this paper, we focus on the self-face recognition ability, which has been proven to be crucial for the sense of identity and for constructing and maintaining self-awareness [12]–[14]. Inspired by sensorimotor theory and cognitive developmental robotics [15], we try to make the robot understand self-face perception and recognize its face by passing an inference process from visual cues in its field of view. We

[1] Zhegong Shangguan and Adriana Tapus are with Autonomous Systems and Robotics Lab/U2IS, ENSTA Paris, Institut Polytechnique de Paris, 828 Boulevard des Maréchaux, Palaiseau 91120, France {zhegong.shangguan, adriana.tapus}@ensta-paris.fr

[2] Mengyuan Ding is with the National Engineering Laboratory for Visual Information Applications, college of Artificial Intelligence, Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an 710049, China. dmy1295@outlook.com

[3] Chuang Yu is with Cognitive Robotics Lab, Department of Computer Science, University of Manchester, Kilburn Building, Oxford Rd, Manchester M13 9PL, UK chuang.yu@manchester.ac.uk

[4] Chaona Chen is with the School of Psychology and Neuroscience, University of Glasgow, Glasgow, UK. Chaona.Chen@glasgow.ac.uk

present a novel algorithm that allows the robot to recognize its face with self-reflected images by using a Bayesian Inference process, which is addressed to answer the question: *"Are those changes in the mirror because I am changing my facial expressions?"* (see Figure 1). This algorithm allows the robot to recognize its face without feature extraction. Specifically, our approach focuses on the development of a self-face perception model from sensorimotor learning. Then, in a self-recognition scenario, the robot will change the facial expressions several times. Finally, by continuing to compute the motion changes process and update its prediction model, the robot can recognize itself.

The main contributions of this paper are as follows:

- Our approach is the first attempt to make robot self-recognition by using facial expressions.
- We developed a Conditional Generative Adversarial Network (CGAN) model to achieve robot self-perception. With the help of this off-line model, the robot can recognize self through an on-line Bayesian learning regression.
- The experimental results show that our method makes the robot capable of self-recognition (i.e., the robot can recognize itself in the mirror). Similar to rubber hand illusion (RHI), the robot can also be tricked by a similar robot with the same facial expression.

The rest of the paper is structured as follows: Section II introduces the related works about current self-recognition and self-modeling research in robotics. Section III describes the framework of our algorithm and its substructure of Sensorimotor Learning, Visual Cues Accumulation, and Bayesian Inference; Section IV shows the learning details and experimental design. The results are summarized in Section V. Our discussion is in VI. And finally, Section VIII concludes the paper.

II. RELATED WORK

Similar to the human ability to identify themselves, self-perception enables robots to understand the relationship between inner states and outside motions [16]. In [17], sensorimotor contingencies were considered as the key to body awareness. The authors used sensory consequences observed to infer where is the robot's body. Similarly, we use face cues consequences observed by the visual sensor to learn whether it is the robot's face. In [6], authors used free energy minimization [18] to infer whether the body configuration of the robot in the mirror is the same as the robot. Their approach relies on the actions of robot arms.

Furthermore, the authors in [7], [19] used Bayesian inference to collect self-evidence. However, in their methods, still objects sometimes were recognized as moving ones due to unstable segmentation. In [20], a self-morphology (e.g., space occupancy queries and robot states) model was developed to achieve better motion planning and control tasks. Different from their full-body reconstruction model, our facial expression self-perception model (see section III) reconstructs the robot's face by using sensorimotor learning.

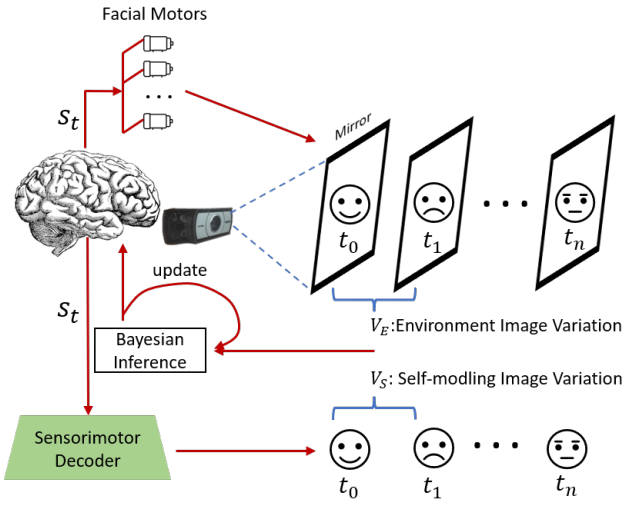


Fig. 2. Self-recognition framework

Our approach includes facial expression self-perception learning (i.e., off-line model) and Bayesian inference to accumulate evidence (i.e., online learning) and can achieve self-recognition tasks (i.e., cognitive ability). In addition, our approach can be used in social scenarios by endowing social robots with cognitive capabilities [21].

III. SYSTEM MODELLING

Our system is modeled as follows. We, first, observe the variations of images in the robot's view and the variations of images are environment visual cues. Secondly, we use our self-modeling method to generate images of robot faces. The variations of self-modeling images are self-reflection cues. Self-modeling images are generated by the sensorimotor learning model. Thirdly, the robot keeps changing its facial expression and tries to infer a prediction model from self-reflection cues distribution to environment visual cues and tries to decrease the estimation error. Finally, after several iterations, the robot updates its inference model and gives the confidence of whether an image is its face. In other words, *"The visual variations are caused by my face expression variations, so that's me."* The whole framework is shown in Figure 2. Each module is detailed in the next subsections.

A. Sensorimotor Learning

Inspired by sensorimotor theory, our idea is to make the robot learn "How do I look like?" when executing actions. By continuously capturing the robot face images pixels y_{robot} (i.e., In our experiments, it is a $640 \times 480 \times 3$ array) combined with motor signals s_{motor}, we train a deep generative model to map the robot's face from motor signals. In order to add uncertainty to our model, the inputs include motor signals and Gaussian noise $z \sim \mathcal{N}(\mu, \sigma^2)$. In this way, a Conditional Generative Adversarial Network (CGAN) is considered to provide this distribution transfer: $y_{robot} \sim P(y \mid z, s)$. CGAN is not only able to learn log-likelihood estimation but also able to control the generation from input [22]. In training, a

generator G and a discriminator D are learned together [23]. Equation (1) shows the training process of CGAN (see also Figure 3).

$$\min_G \max_D V(D,G) = \mathbb{E}_{y\sim p_{\text{data}}(y)}[\log D(y \mid s)] + \mathbb{E}_{z\sim p_z(z)}[1-\log D(G(z \mid s))] \quad (1)$$

After training, the Generator G is the sensorimotor decoder module (see Figure 2). The robot can get self-modeling morphology $y'_t = G(z_t, s_t)$ at time t.

B. Visual Cues Accumulation

Generated self-modeling y'_t is a static robot self-perception based on prior knowledge from the sensorimotor learning data before the self-recognition task. A contingent effect always exists in the environment, such as lighting, background objects, or other factors. In [6], they use contingency learning and optical flow to classify the contingency effects and robot effects. In our approach, we use the Environment Visual Variation $V_{e,t_n} = y_{t_n} - y_{t_{n-1}}$ as outside cues and the self-perception Visual Variation $V_{s,t_n} = y'_{t_n} - y'_{t_{n-1}}$ as inside cues. The Visual Variation V_{t_n} (i.e., V_{e,t_n} or V_{s,t_n}) is calculated with the help of the Euclidean Distance d_E [24]. The Equations are shown in (2) and (3).

$$V_{e,t_n} = d_E(y_{t_n}, y_{t_{n-1}}) = \|y_{t_n} - y_{t_{n-1}}\| \quad (2)$$

$$V_{s,t_n} = d_E(y'_{t_n}, y'_{t_{n-1}}) = \|y'_{t_n} - y'_{t_{n-1}}\| \quad (3)$$

In this way, we can compute the action effect from time t_{n-1} to time t_n, and the noise of the background is filtered. With the Visual Cues Accumulation process, our model can achieve a robust self-recognition system.

C. Bayesian Inference

There is no doubt that the decoder can never generate self-perception images with 100% accuracy. If the robot is partly broken, the generated images will be unreliable because of the changing in the robot morphology or the environment. In the other world, there is an *error* between estimated cues V_{e,t_n} and the observed cues V_{s,t_n} as shown in (4).

$$error_{t_n} = V_{e,t_n} - f(V_{s,t_n}) \quad (4)$$

In Friston's theory [25], the perception of the body and the action of the motor are driven by *surprise* minimization.

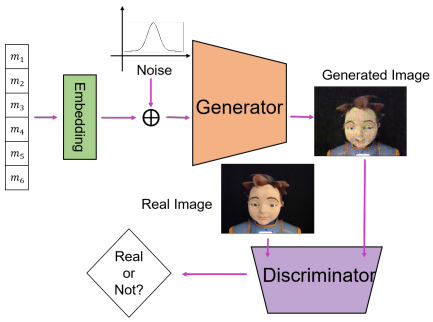


Fig. 3. CGAN pipeline

It means that we try to update our actions to make the reality the same as what we have predicted. The *error* between reality and estimation is the direction that guides our optimization. Free Energy Minimization (FEM) is used to keep updating this process in Friston's theory [25]. Inspired by this theory, *error* is the *surprise* parameter in our algorithm. We assume that if the facial expression changing in the visual sensor is the same process with its self-reflected face image, it exists a correlation as indicated in Equation (5).

$$V_{e,t_n} = f(V_{s,t_n}) \mid_{error_{t_n}\to 0} \quad (5)$$

Bayesian Inference is used to optimize the function and to update the function parameters online. The goal is to decrease the error gradually and estimate real-world changing V_{e,t_n}. If function $f(V_{s,t_n})$ successfully estimates the V_{e,t_n} with high confidence p_{self}, it means that the system has learned the effects of the robot motions and the robot is no longer surprised about the visual cues, and therefore, the self-recognition is established. We use linear regression to fit these correlations and V_{e,t_n} is assumed to be Gaussian distributed by $V_{s,t_n}w$ as indicated in (6).

$$f(V_{s,t_n}) \to p(V_e \mid V_s, w, \alpha) = \mathcal{N}(V_e \mid V_s w, \alpha) \quad (6)$$

The α is a random variable and will be updated by observing the *error*. The w is the coefficient. Based on [26], the prior of w is a spherical Gaussian as in (7) and we get (8). The α and λ given in equations are from gamma distributions.

$$p(w \mid \lambda) = \mathcal{N}(w \mid 0, \lambda^{-1}\mathbf{I}_p) \quad (7)$$

$$\ln p(w \mid V_e) = -\frac{\lambda}{2}\sum_{n=1}^{m}\frac{error_n^2}{2} - \frac{\alpha}{2}w^T w + const \quad (8)$$

By using Bayesian Theory as in [27], the update process is as in 9.

$$p(w \mid V_{e,t_n}) \propto p(V_{s,t_0}, V_{s,t_1}, ..., V_{s,t_{n-1}} \mid w) p(w \mid V_{e,t_n}) \quad (9)$$

D. Algorithm

The Algorithm 1 describes the self-recognition process. Firstly, we need a Sensorimotor Decoder G, which has been trained already to generate real-time facial expression images. G is used to calculate p_{Self} and V_{s,t_n}. Then, the online learning module estimates V_{e,t_n} from a sequence of actions. Finally, if the effects of actions can learn from a sequence of the robot actions and the $error_n$ converges to zero, the robot can conclude: *It is me!*.

IV. EXPERIMENTS

Because the execution time of the facial motor in our robot system is very long (about 1 second), real-time and visual information cannot be compared. So in our experiment, the visual information will be compared to the static expression after the motor completes the action. Moreover, our robot does not have a camera in its eyes, so the robot's head does not turn from side to side during the experiment.

Algorithm 1 Self-recognition algorithm

Input: Sensorimotor Decoder: G, Visual Sensor: y, Time t

$s_{t_0}, y_{t_0} \leftarrow$ Initial Robot Face Motor State, Visual Sensor
$V_{s,t_0}, V_{e,t_0} \leftarrow$ Initial Self and Environment Variation
$z \sim \mathcal{N}(0,1) \leftarrow$ uncertainty noise
$\alpha = \lambda = 1e-6 \leftarrow$Initial Bayesian Inference
while recognized(p_{self}) **do**
  $y'_{t_{n-1}} = G(s_{t_{n-1}}, z_{n-1})$
  $y'_{t_n} = G(s_{t_n}, z_n)$
  $V_{s,t_n} = Euclidean(y'_{t_n}, y'_{t_{n-1}})$
  $V_{e,t_n} = Euclidean(y_{t_n}, y_{t_{n-1}})$
  $p_{self} = p(V_{e,t_n} \mid f(V_{s,t_n}) \sim \mathcal{N}(V_{s,t_n} w, \lambda))$
  $error_n = V_{e,t_n} - f(V_{s,t_n})$
  $error_{accumulation} = \sum_{n=1}^{m} absolute(error_n)/n$
  $w_n \propto \arg\min -\frac{\lambda}{2}\sum_{n=1}^{m} \frac{error_n^2}{2} - \frac{\alpha}{2} w^T w$
  update$(w_n, \alpha_n, \lambda_n)$
  if $error_{accumulation}$ converges and $error_n \leftarrow 0$ **then**
    It is me ! My confidence is of p_{self}
  else
    It is not me !
  end if
end while

Output: Estimated V_e, Self-recognition probability p_{self}

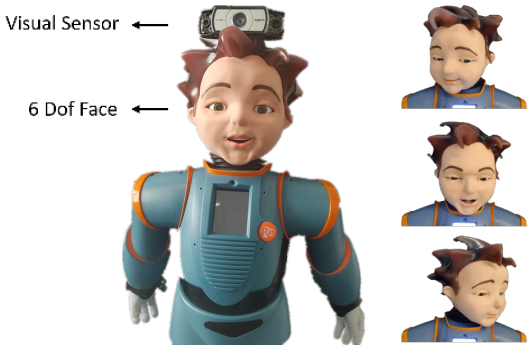
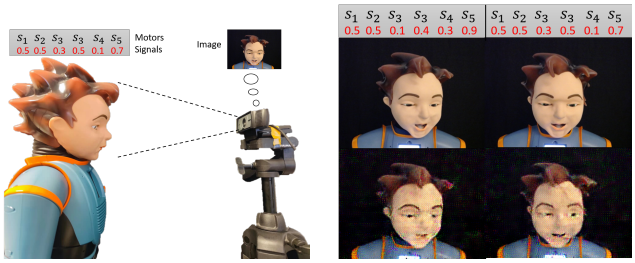


Fig. 4. Zeno robot. There are 6 motors embedded in Zeno's face, The motor control of the entire face structure is symmetrical, with the left and right sides showing the same facial movement.

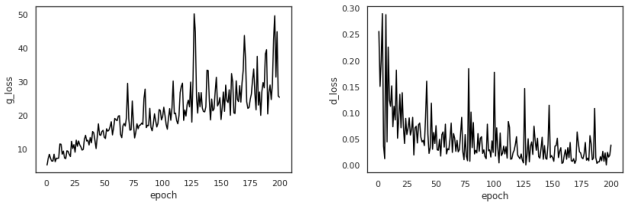


(a) Sensorimotor Learning Process (b) Model Result

Fig. 5. Data collection and model results

model. We use a multi-layer Convolutional Neural Network (CNN) as discriminator D and Deconvolutional Neural Network (DeCNN) as generator G. The number of layers in the generator and discriminator is 4. We resized the images to $320 \times 240 \times 3$. Then, we normalized the tensor images with a mean ($[0.5, 0.5, 0.5]$) and standard deviation ($[0.5, 0.5, 0.5]$). The batch size is 64. The Adam optimizer was used with an initial learning rate of 2×10^{-4}. LeakyReLU and 2D Batch normalization is used to avoid over-fitting. Training loss is shown in Figure 6. After 100 epochs of learning steps, the robot gradually acquired the self-face image generation ability. By using this module, we can estimate the expressions of the robot from motors signals (see Figure 5 (b)).

In order to evaluate our work, we have designed two types of experiments: a self-recognition experiment described in Section IV-B, and a visual illusion experiment presented in Section IV-C.



(a) Generator Loss (b) Discriminator Loss

Fig. 6. Training Loss of learning process

A. Sensorimotor Learning

Zeno robot is used in our experiment (see Figure 4). Zeno's face is driven by 6 PMW motors, which can make abundant facial expressions in 6 Dof (i.e., brow, eyelid, gaze, eye-turning, corners of the mouth, jaw, head-turning). Inspired by human sensorimotor theory, we have the robot continuously drive its motor to present different expressions. At the same time, the robot keeps watching its expression change in the mirror in order to learn the co-relationship. Specifically, in order to learn a sensorimotor module, the robot shows more than 200,000 different expressions in front of a camera and records the face images combined with the motors' signal data (see Figure 5 (a)). The resolution of each image is $640 \times 480 \times 3$.

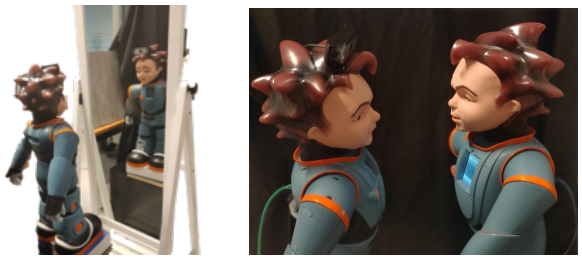
After data collection, we use Pytorch to train the CGAN

B. Self-recognition experiment

The robot is positioned in front of a mirror (see Figure 7 (a)), and the visual sensor is fixed on the robot's head to observe the information in its field of view (see Figure 4). The robot presents several different facial expressions for recognition. Examples of expressions are shown in Figure 8. The robot keeps observing and updating the fitting algorithm. Afterward, it indicates how confident it is about "It is me in the mirror." In our hypothesis, the robot should give a low probability in the initial stage. After several iterations of learning (limited to 20 steps), the robot will give a high probability from visual cues.

C. Visual illusion experiment

The robot is put in front of another Zeno robot (see Figure 7 (b)) (both robots are identical). Several expressions are



(a) Zeno in front of a mirror (b) Two Zeno face to face

Fig. 7. Experiment diagram. The camera was fixed on the head to detect optical information for the mirror before the experiment. The first experiment was in the left figure, Zeno was placed in front of a mirror and showed different facial expressions to recognize himself. Then the algorithm was trying to recognize itself in the mirror. The second experiment was in the right figure, Zeno was placed in front of the same Zeno robot and shows the same facial expression. Then the algorithm also tried to recognize its face.

given to both robots as in the above experiment. The robot will see a similar robot that shows the same expressions in its field of view. In this experiment, we aim to test whether Zeno robot will give a visual illusion result that it is itself. As presented in Section I, human beings have the rubber hand illusion (RHI) phenomenon because of sensorimotor learning. In this experiment, we will show that our algorithm makes the robot have a similar behavior.

V. RESULTS

In both experiments, 15 random expressions are given, and the robot recognizes itself successfully. In our experiment, no arm or feature recognition methods are used. No prior information about the environment is learned in advance. We used offline sensorimotor learning to generate real-time face appearance, and on-line inference to learn the difference between the real world and estimation. The face of the robot was detected and extracted to compare with each other by the python program. Finally, the robot learns the real-world expressions variations within 15 iterations with more than 98% confidence. Also, the robot is tricked when there is a robot that looks like it and shows the same facial expressions, which means that the visual illusion also appears in our robots.

A. Self-recognition in front of a mirror

As shown in Figure 8, the robot keeps moving different parts of its face (e.g., eyelids, mouth, jaw) to show different expressions $y_{t_0}, y_{t_1}, ..., y_{t_n}$. The robot keeps observing the variation $V_{s,t_0}, V_{s,t_0}, ..., V_{s,t_{n-1}}$. In Figure 10, it can be seen that at the 5_{th} iteration (i.e., 4 iterations evidence to estimate the 5_{th} iteration), the real variations V_{e,t_5} (blue line in Figure 8 (a)) drops into prediction confidence interval of $f(V_{s,t_5})$ (orange line and pink area in Figure 8 (a)). The accumulation error $error_{accumulation}$ has not converged (red line in Figure 8 (a)). The algorithm gives 49.76% confidence in self-recognition. However, in the 14_{th} iteration, after collecting 13 pieces of evidence, the algorithm gives 99.97% confidence. The $error_{accumulation}$ tends to be stable and the error tends to be 0 (green line in Figure 8 (b)), which means that the robot has recognized itself.

B. Self-recognition in front of a robot

Similar to Section V-A, at the 5_{th} iteration, the real variation V_{e,t_5} drops out of the prediction confidence interval. The algorithm gives 0% confidence. At the 12_{th} iteration, the variation between $f(V_{s,t_{12}})$ and real $V_{e,t_{12}}$ is small and $error_{accumulation}$ tends to converge, which means that the robot mistakenly thinks it is standing in front of itself with 98.34% confidence.

VI. DISCUSSION

For human beings, facial information is the most important piece of evidence for self-identity [10]. Compared to robot arm actions recognition, the use of facial information is more in line with cognitive science research [11] and fits with human intuition. However, no previous research uses facial expression context to make robots recognize themselves. In Human-Human Interaction (HHI), the face provides an interface of an underlying emotional state [28]. Hence, if we want to equip the robot with self-recognition abilities, self-face understanding and self-perception are indispensable. We posit that cognitive abilities should combine off-line knowledge learning and online environment fitting. In this way, the cognitive robot will acquire a robust intelligent system to overcome general tasks. This path is the same with Free Energy Minimization (FEM) theory [18] to use a generative model (off-line knowledge model) to learn how to achieve the goal (online optimization).

In our experiments, the robot can achieve self-recognition in front of a mirror. Furthermore, the robot is tricked by another robot's face, through an inference process. However, the full version of Mirror Self-Recognition (MSR) is still a complex cognitive task that requires the intelligent agent to not only recognize the self in the mirror but also touch a mark on the body by observing the mirror. This requires the self-perception of the whole body and the self-reflection of minimal self [1]. Our algorithm has provided the path to the self-perception model and methodology to pass the Mirror Self-Recognition (MSR).

VII. CONCLUSION

In this work, we propose a novel algorithm to achieve self-recognition ability in a humanoid robot. This framework is the first attempt to make self-recognition by using robot facial expressions. Our self-perception module is inspired by sensorimotor theory to learn self-face generation through the CGAN method. To estimate the real environment, we use a Bayesian inference regression to predict the facial expressions variations from self-perception expressions variations. This is a combination of online inference and offline learning algorithm. With the help of our experiments, we demonstrated that our algorithm successfully makes the Zeno robot recognize itself within 15 expressions. Nevertheless, our last experiment shows that our algorithm would be tricked if the recognizing object is showing the same facial expressions

Fig. 8. Face in the mirror

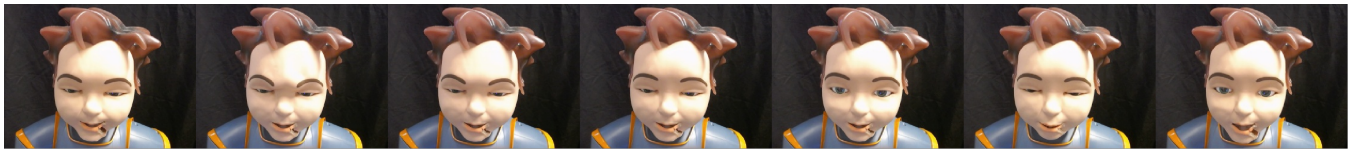
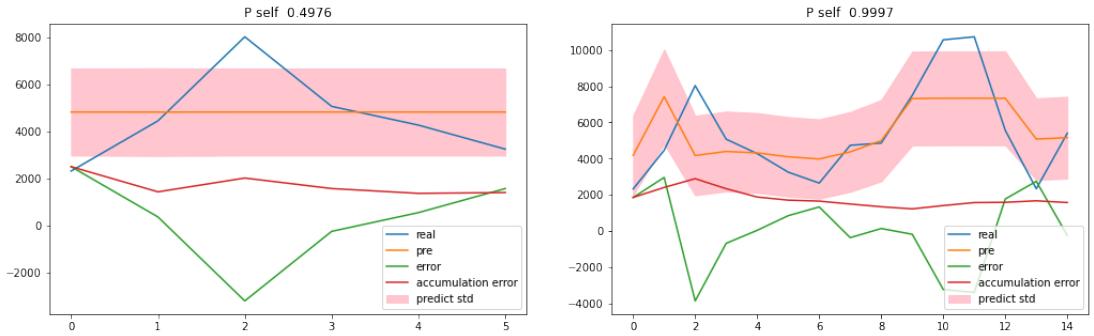
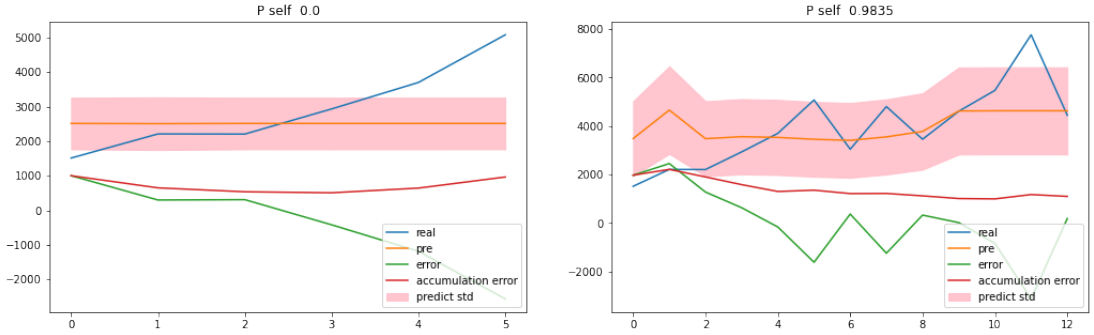


Fig. 9. Another face of Zeno in the field of view



(a) On-line learning in 5 iterations (b) On-line learning in 14 iterations

Fig. 10. Self-recognition process. After 4 iterations, the robot gives a 49.76% probability to think it is itself. After 13 iterations, 99.97% probability is given by the robot.



(a) On-line learning in 5 iterations (b) On-line learning in 12 iterations

Fig. 11. Visual illusion experiment results. After 4 iterations, the robot gives 0% probability to think it is itself. After 11 iterations, 98.35% probability is given by robot.

with a similar face. This visual illusion phenomenon is usually led by the human cognitive system. The limitation of this research is that the application of our model is limited to the face. Fully body self-perception is needed for real Mirror Self-Recognition (MSR).

VIII. LIMITATION AND FUTURE WORK

In our study, our generative model could generate real-time facial expressions, but the execution of the motor lagged behind the visual generation, which was caused by the experimental platform. This is part of one of the main challenges of humanoid robots: "How to make them more flexible and agile?" Another limitation of our system is that the background used was constant (black background). We plan to examine in further research how changes in the background can interfere with the robot's recognition, and how self-recognition of faces can be applied to a broader cognitive robot framework.

ACKNOWLEDGMENT

This work was supported by ENSTA Paris, Institut Polytechnique de Paris, France and the CSC PhD Scholarship.

REFERENCES

[1] Y. K. Georgie, G. Schillaci, and V. V. Hafner, "An interdisciplinary overview of developmental indices and behavioral measures of the minimal self," in *2019 Joint IEEE 9th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*. IEEE, 2019, pp. 129–136.

[2] B. L. Thomas, J. M. Karl, and I. Q. Whishaw, "Independent development of the reach and the grasp in spontaneous self-touching by human infants in the first 6 months," *Frontiers in psychology*, vol. 5, p. 1526, 2015.

[3] M. Tsakiris and P. Haggard, "The rubber hand illusion revisited: visuotactile integration and self-attribution." *Journal of experimental psychology: Human perception and performance*, vol. 31, no. 1, p. 80, 2005.

[4] J. R. Anderson and G. G. Gallup Jr, "Which primates recognize themselves in mirrors?" *PLoS Biology*, vol. 9, no. 3, p. e1001024, 2011.

[5] ——, "Mirror self-recognition: a review and critique of attempts to promote and engineer self-recognition in primates," *Primates*, vol. 56, no. 4, pp. 317–326, 2015.

[6] P. Lanillos, G. Cheng, *et al.*, "Robot self/other distinction: active inference meets neural networks learning in a mirror," *arXiv preprint arXiv:2004.05473*, 2020.

[7] 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.

[8] L.-Y. Wang, M. Zhang, and J. Sui, "Self-face advantage benefits from a visual self-reference frame." *Acta Psychologica Sinica*, 2011.

[9] Z. R and Z. A., "When i am old: The self-face recognition advantage disappears for old self-faces," *Frontiers in psychology*, vol. 10, p. 1644, 2019.

[10] J. P. Keenan, A. Nelson, M. O'connor, and A. Pascual-Leone, "Self-recognition and the right hemisphere," *Nature*, vol. 409, no. 6818, pp. 305–305, 2001.

[11] C. Ota and T. Nakano, "Self-face activates the dopamine reward pathway without awareness," *Cerebral Cortex*, vol. 31, no. 10, pp. 4420–4426, 2021.

[12] M. Martini, I. Bufalari, M. A. Stazi, and S. M. Aglioti, "Is that me or my twin? lack of self-face recognition advantage in identical twins," *PLoS One*, vol. 10, no. 4, p. e0120900, 2015.

[13] M. Hoffmann, S. Wang, V. Outrata, E. Alzueta, and P. Lanillos, "Robot in the mirror: Toward an embodied computational model of mirror self-recognition," *KI-Künstliche Intelligenz*, vol. 35, pp. 37–51, 2021.

[14] P. Michel, K. Gold, and B. Scassellati, "Motion-based robotic self-recognition," in *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)(IEEE Cat. No. 04CH37566)*, vol. 3. IEEE, 2004, pp. 2763–2768.

[15] M. Asada, K. Hosoda, Y. Kuniyoshi, H. Ishiguro, T. Inui, Y. Yoshikawa, M. Ogino, and C. Yoshida, "Cognitive developmental robotics: A survey," *IEEE transactions on autonomous mental development*, vol. 1, no. 1, pp. 12–34, 2009.

[16] P. O. Haikonen, "Reflections of consciousness: The mirror test." in *AAAI Fall Symposium: AI and Consciousness*, 2007, pp. 67–71.

[17] 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, 2016.

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

[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] B. Chen, R. Kwiatkowski, C. Vondrick, and H. Lipson, "Fully body visual self-modeling of robot morphologies," *Science Robotics*, vol. 7, no. 68, p. eabn1944, 2022.

[21] C. Breazeal, *Designing sociable robots*. MIT press, 2004.

[22] K. Khalvati, S. A. Park, S. Mirbagheri, R. Philippe, M. Sestito, J.-C. Dreher, and R. P. Rao, "Modeling other minds: Bayesian inference explains human choices in group decision-making," *Science advances*, vol. 5, no. 11, p. eaax8783, 2019.

[23] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[24] L. Wang, Y. Zhang, and J. Feng, "On the euclidean distance of images," *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, no. 8, pp. 1334–1339, 2005.

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

[26] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*. Springer, 2006, vol. 4, no. 4.

[27] D. J. MacKay, "Bayesian interpolation," *Neural computation*, vol. 4, no. 3, pp. 415–447, 1992.

[28] R. Buck, "Social and emotional functions in facial expression and communication: The readout hypothesis," *Biological psychology*, vol. 38, no. 2-3, pp. 95–115, 1994.