Autonomous Development Of Social Referencing Skills

S. Boucenna[1], P. Gaussier[1,2], L. Hafemeister[1], K. Bard[3]

[1]ETIS, CNRS UMR 8051, ENSEA, Univ Cergy-Pontoise, [2]IUF, [3] Psychology University of Portsmouth
{boucenna,gaussier,hafemeister}@ensea.fr, kim.bard@port.ac.uk

Abstract. In this work, we are interested in understanding how emotional interactions with a social partner can bootstrap increasingly complex behaviors such as social referencing. Our idea is that social referencing as well as facial expression recognition can emerge from a simple sensori-motor system involving emotional stimuli. Without knowing that the other is an agent, the robot is able to learn some complex tasks if the human partner has some "empathy" or at least "resonate" with the robot head (low level emotional resonance). Hence we advocate the idea that social referencing can be bootstrapped from a simple sensori-motor system not dedicated to social interactions.

1 Introduction

How can a robot or a human learn more and more complex tasks? This question is becoming central in robotics and psychology. In this work, we are interesting in understanding how emotional interactions with a social partner can bootstrap increasingly complex behaviors. This study is important both for robotics application and understanding development. In particular, we propose that social referencing, gathering information through emotional interaction, fulfills this goal. Social referencing, a developmental process incorporating the ability to recognize, understand, respond to and alter behavior in response to the emotional expressions of a social partner, allows an infant to seek information from another individual and use that information to guide his behavior toward an object or event [14].

Gathering information through emotional interaction seems to be a fast and efficient way to trigger learning. This is especially evident in early stages of human cognitive development, but also evident in other primates [19]. Social referencing ability might provide the infant, or a robot, with valuable information concerning the environment and the outcome of its behavior, and is particularly useful since there is no need for verbal interactions. In social referencing, a good (or bad) object or event is identified or signaled with an emotional message, not with a verbal label. The emotional values can be provided by a variety of modalities of emotional expressions, such as facial expressions, voice, gestures, etc. We choose to use facial expressions since they are an excellent way to communicate important information in ambiguous situations but also because they can be learned autonoumously very quickly [4]. Our idea is that social referencing as well as facial expression recognition can emerge from a simple sensori-motor system. All the work is based on the idea of the perception ambiguity: the inability at first to differentiate our own body from the body of others if they are correlated with our own actions. This perception ambiguity associated to a homeostatic system is sufficient to trigger first facial expression recognition and next to learn to associate an emotional value to an arbitrary object. Without knowing that the other is an agent, the robot is able to learn some complex tasks. Hence we advocate the idea that social referencing can be bootstrapped from a simple sensori-motor system not dedicated to social interactions.

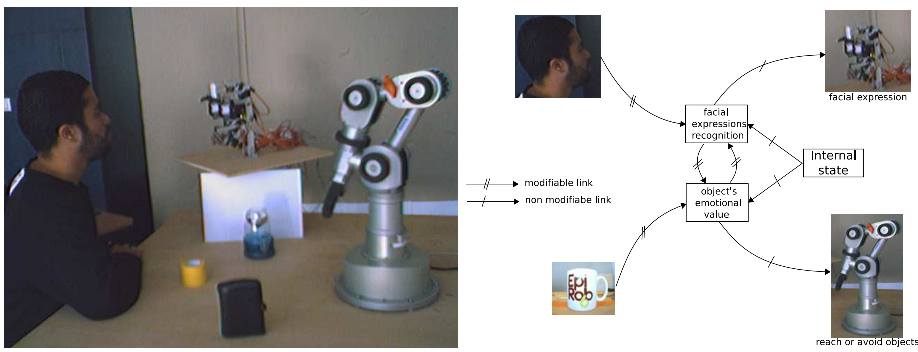

Fig. 1. Experimental set-up for social referencing. The robot relies upon the use of its expressive head which is also able to recognize facial expressions. the robotic arm will reach the positive objects and avert the negative objects after emotional interactions with a human partner.

2 Overview

Our social referencing experiment (fig. 1,2) has the following set-up: a robotic head having one camera is able to recognize facial expressions and another camera is turned toward a workspace where a Katana arm is able to reach an object. As a consequence, the robot (head plus arm) can interact with the environment (human partner) and can manipulate objects. In the developed architecture, the robot learns to handle positive objects and to avoid negative objects as a direct consequence of emotional interactions with the social partner. The robotic head learns to recognize emotional facial expressions (sadness, joy, anger, suprise and neutral face) autonomously [4]. The internal emotional state of the robot triggers one specific expression and the human mimicks the robot face to face. The robot can learn to associate its internal emotional state with the human's facial expression. The robot associates what it is doing with what it is seeing. After few minutes of real time learning (typically less than 3 minutes), the robot is able to recognize the human facial expressions as well as to mimick them. In

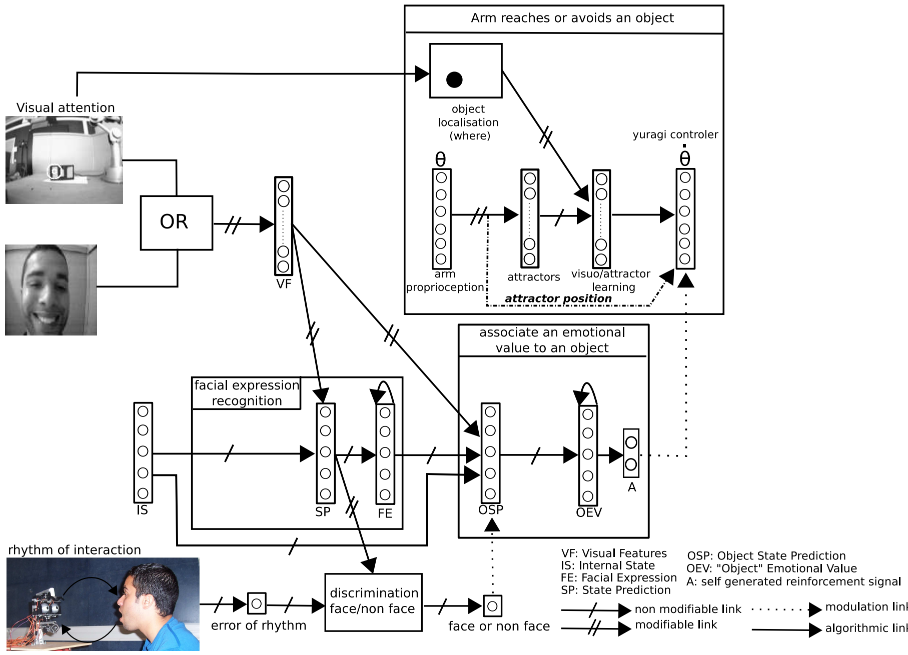

Fig. 2. Social referencing model. Social referencing emerging from the sensori-motor interactions between facial expression recognition, objects emotional value and visuo-motor learning. A simple sensori-motor architecture is able to learn and recognize the facial expressions, and then to discriminate between facial/non facial stimuli. Using a simple chain of conditioning, the robot learns the emotional value of an object as a result of the interactions with the human (face discrimination). The robot focuses on an object using a visual attention processus (Gabor filters, color). After a visuo-motor learning, the robot arm reaches or avoids some objects in the workspace thanks to the self generated reinforcement signal A(emotional value coming from the facial expression recognition). A is built as the result of the facial expression recognition (with A_1 neuron corresponding to happy facial expression, the A_2 neuron corresponding to angry facial expression)

parallel, the eye-arm system can learn throw visuo-motor asociation to reach several positions in the workspace [1]. A reinforcing signal is used to reach or avoid a position in the workspace [9]. The signal can be an emotional signal (e.g joy facial expression is a positive signal and an angry facial expression is a negative signal).

The tested scenario is the following: The robot is in a neutral emotional state, a human displays a joy facial expression in the presence of an object, consequently the robot moves to a joy state and associates a positive value to the object. On the contrary if the human displays an anger facial expression, the value associated to this object is negative. The robot arm can handle or avoid the objects according to their associated emotional value. In other words, the emotional value associated to the object is the reinforcing signal that the arm uses so as to move.

In the following paper, we will see a developmental approach of the social referencing namely: the development of facial expressions recognition, the association of emotional value to an object and finally an aspect of motor control according to emotional stimuli.

3 Online learning of facial expression recognition

At this stage of the development, the robot must be able to recognize and understand the caregiver facial expressions. We summarize here an architecture that we developped for online learning of facial expression recognition. A simple sensory-motor architecture is able to express several emotions and to recognize online the facial expression of a caregiver if this latter naturally tends to imitate the system or to resonate with it. In particular, we showed that autonomous learning of face/non face discrimination is more complex than the facial expression recognition [4].

Using the cognitive system algebra [11], we showed that a simple sensory-motor architecture based on a classical conditioning paradigm [20, 2] can learn to recognize facial expressions online. Furthermore, the dynamics of the human-robot interaction bring important but non explicit signals, such as the interaction rhythm that helps the system to perform the face/non face discrimination. The interaction rhythm is used to allow first a robust learning of the facial expression without face tracking and next to perform the learning of the face/non face discrimination. Psychologists underline the importance of the synchrony during the interaction between the mother and the baby [7]. If a rhythmic interaction between baby and mother involves positive feelings and smiles (positive reward), a social interaction interuption involves negative feelings (negative reward). In our case (following [1]), the rhythm is used as a reward signal. It provides an interesting reinforcement signal to learn to recognize an interacting partner(face/non face).

We adopt the following experimental protocol: the facial expressions of the robotic head have been calibrated by FACS experts [8]. In the first phase of interaction, the robot produces a random facial expression during 2s (among the following: sadness, happiness, anger, surprise), then returns to a neutral face during 2s to avoid human misinterpretations of the robot facial expression (same procedure as in psychological experiments). The human subject is explicitely asked to mimic the robot head (even without any instruction, psychologist have shown that the human subject resonates with the facial expressions of the robot head [17]). This first phase lasts between 2 and 3 minutes depending on the subject "patience". Then, in the second phase, the random emotional states

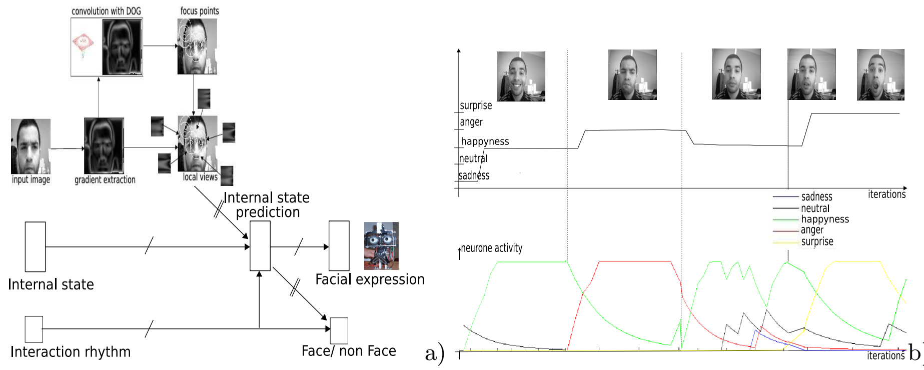

Fig. 3. a)The global architecture is able to recognize and imitate a facial expression and to perform a face/non face discrimination.A visual processing allows to extract sequentially the local views. The *internal state prediction* learns the associations between the local views and the internal state. b)Temporal activity of the neurons associated to the triggering of the different facial expressions when the robot imitates the human (after learning).

generator is stopped. After the N.N (Neural Network) has learned, the robot mimics the human partner facial expressions. This architecture (see fig. 3) allows the robot to recognize the subjects visual features and to learn if these features are correlated with the robot own facial expressions. Moreover, another sub network learns to predict the interaction rhythm allowing the robot to detect if an interacting agent (a human) faces the robot head. In this case, the facial expression recognition is a bootstrap to discriminate face from non face images. At the end of this development stage, the robot head is able to recognize and understand the emotional facial expressions. They can now be seen as a way to communicate.

4 Associating an emotional value to an object

After the human partner has imitated during 2 to 3 minutes the robot head, the robot is able to recognize and display the human facial expressions. As soon as this learning is performed, the human can interact with the robotic head to associate an emotional value to an object (positive or negative). The emotional expression is a way to communicate, that will help the robot to interact with objects according to the human will.

The N.N processes (see fig. 2) in the same way signals from the robot's internal state and infomation correlated with this internal state. An internal state can trigger a robot facial expression and a human facial expression can trigger also the robot facial expression. In case of conflict, the weights from the internal state to control the facial expression are higher than those coming from the facial expression recognition. That allows to prefer the display of the internal state rather than facial expression recognition (this is an apriori to avoid the use of much more complex structures that could be useful to allow a volontary control of the facial expression). In the absence of the internal state, the recognized facial expression induces an internal state which is associated with the object (a simple conditionning chain: fig. 2). Classical conditioning is used to perform the association between the emotional value that the human transmits and some areas of the image. The attentional process used in this model is very simple (see [12, 6] for more information), the robot focuses on colored patches and textures (fig. 4). When focusing on an object, the robot extracts some focus points and associates the recognition of the local view surrounding each focus point with the emotional value of the robot. The focus points are the result of a DOG (Difference of gaussian) filter convolved with the gradient of the input image. This process allows the system to focus more on corners or end of lines in the image. Its main advantages over the SIFT [15] method are its computational speed and the few number of needed focus points. One after another, the most active focus points are used to compute local views (a log polar[1] transform centered on the focus point and its radius is 20 pixels). Each local view is learned

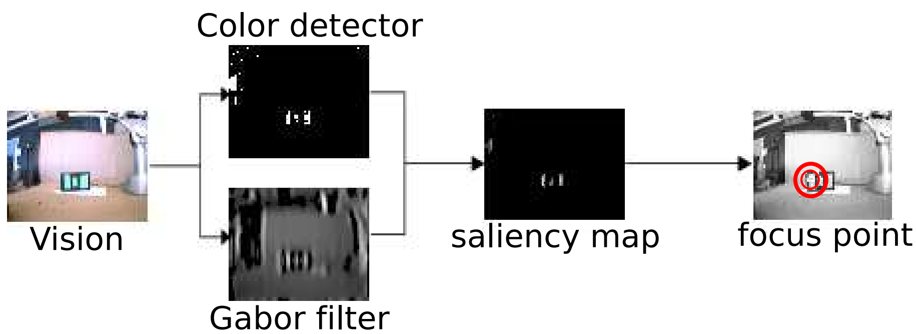

Fig. 4. Visual attention. The system focuses on some relevant features of the image. A saliency map is performed in order to focus an interesting area in the image. Visual primitives are calculated independently (gabor filters, color detector), a fusion of these primitives is performed in order to find the area that the robot must analyze.

by a VF_j (Visual Features) neuron:

$$VF_j = net_j.H_\theta(net_j) \qquad \theta = max(\gamma, \overline{net} + \sigma_{net}) \tag{1}$$

$$net_j = 1 - \frac{1}{N}\sum_{i=1}^{N}|W_{ij} - I_i| \tag{2}$$

[1] The local polar transform increases the robustness of the extracted local views to small rotations and scale variations

VF_j is the activity of neuron j in the group VF. $H_\theta(x)$ is the Heaviside function[2]. $\gamma = 0.95$ is the vigilance (if the prototype recognition is below γ then a new neuron is recruited). $\overline{net}$ is the average of the output, σ_{net} is the standard deviation, I is the input image (N size of I) and W are the weights between I and VF. The learning rule for the local view categorization allows both one shot learning and long term averaging. The modifications of the weights W are computed as follow:

$$\Delta W_{ij} = \delta_j{}^k (a_j(t) I_i + \epsilon(I_i - W_{ij})(1 - VF_j)) \tag{3}$$

with $k = ArgMax(VF_j)$, $a_j(t) = 1$ only when a new neuron is recruited otherwise $a_j(t) = 0$, $\delta_j{}^k$ is the Kronecker symbol[3] and $\varepsilon = 0.001$ is a positive constant inferior to 1. When a new neuron is recruited, the weights are modified to match the input (term $a_j(t)I_i$). The other part of the learning rule $\varepsilon(I_i - W_{ij})(1 - VF_j)$ is used to average the already learned prototypes. The more the input will be close to the weights, the less the weights are modified. Conversely the less the inputs will be close to the weights, the more they are averaged. If ε is chosen too small then it will have a small impact. Conversely, if ε is too big, the previously learned prototypes can be forgotten. With this learning rule, the neurons in the VF group learn to average the prototypes of objects.

The object state prediction (OSP) group associates the activity of VF with the recognized facial expression (FE) by the robot (simple conditioning mechanism using the Least Mean Square rule [22]):

$$OSP_j = \sum_i w_{ij}.VF_i \qquad \Delta w_{ij} = \epsilon_1.VF_i.(FE_j - OSP_j) \tag{4}$$

OSP corresponds to the object emotional value for one focus point and w_{ij} is the synaptic weights between VF and OSP. OEV (object emotinal value) corresponds to a short term memory. OEV is used to recursively sum and filter on a short period ($N < 1$), the emotional value OSP associated with each explored local view. OEV corresponds to the emotional value to object (accumulation of all focus points), the OEV_i highest activity triggers the i^{th} ($0 < i \leq 5$) emotional value (WTA mechanism). After learning, the associations between VF the view recognition and OSP the emotional state are strong enough to bypass the low level reflex activity coming from the FE. Each focus points has an emotional value (OSP) and OEV is the accumulation of all focus points which corresponds to the object emotional value:

$$OEV_i = OSP_i + N.OEV_i \tag{5}$$

[2] Heaviside function:

$$H_\theta(x) = \begin{cases} 1 \text{ if } x > \theta \\ 0 \text{ otherwise} \end{cases}$$

[3] Kronecker function:

$$\delta_j{}^k(x) = \begin{cases} x \text{ if } j = k \\ 0 \text{ otherwise} \end{cases}$$

At this stage of development, the robot is able to use the emotional facial expression of the human partner in order to assign an emotional value to an object. As a result of the interaction with the partner, the robot recognizes and understands the human's expression in the aim of disambiguating some situations (a new object in the workspace).

5 Visuo-motor learning and Yuragi Controler

At this stage of the development, the robot must be able to modulate his behavior as the result of the emotional interaction. After visuo-motor learning (learning between the extremity of the arm and the proprioception), several positions in the workspace are reached by the robot arm [1]. One visual position corresponds to one or several motor configurations (e.g attractors). These attractors pull the arm in an attraction basin (the position target). This control is performed with a dynamical system to smooth the trajectory [9]. This dynamical system also uses a reinforcing signal in the aim of attaching a lot of or little importance to some attractors, for instance a reward can be given if the arm follows the right direction, otherwise a punition. The reinforcing signal can be emotional (joy facial expression as a positive signal and angry facial expression as negative signal). Following[9] attractor selection model can be represented by Langevin equation as:

$$\tau_x \dot{x} = f(x) * A + \epsilon \tag{6}$$

where x and f(x) are the state (arm proprioception)and the dynamics of the attractor selection model, $\tau_x = 0.1$ is time constant and ϵ represents noise. A is the reinforcing signal which indicates the fitness of the state x to the environment and controls the behavior of the attractor selection model. That is to say, $f(x) * A$ becomes dominant when the activity is large, and the state transition approaches deterministic behavior (converge towards the goal). On the other hand, the noise ϵ becomes dominant when the activity is small and the state transition becomes more probabilistic.

$$f(x) = \sum_{i=1}^{n_a} N_i \frac{(X_i - x)}{||X_i - x||} \tag{7}$$

$$N_i = \frac{g_i(x)}{\sum_{j=1}^{n_a} g_j(x)} \tag{8}$$

$$g_i(x) = exp\{-\beta ||X_i - x||^2\} \tag{9}$$

With n_a the number of selected attractors, X_i (i=1, ... , n_a) a vector reprensenting the center of the i-th attractor and the function N_i a normalized Gaussian. The behavior of this system is such that the arm approaches to the nearest attractor.

Figure 5 shows the important steps of the social referencing model. Figure 5a shows the object's emotional value associated with the facial expressions of the

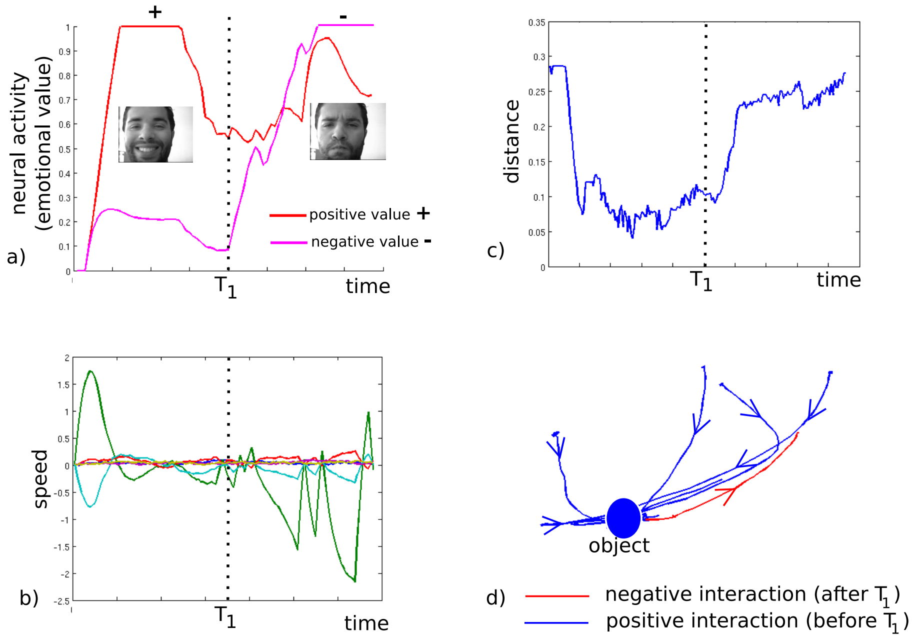

Fig. 5. These curves show: a) the emotional value transmits to the object thanks to the interaction with the human's partner (before T_1 human transmits a positive value after T_1 the human transmits a negative value) b)the speeds of each arm's motor (6 degrees of freedom) c) the distance to the object d) the robotic arm trajectories from different starting points: the arm is able to reach the object associated with the happy facial expression and avoid the object when it is associated with the angry facial expression.

human partner. Before T_1, the partner displays a happy facial expression in presence of the object,the human associates a positive emotional value to this object (A_1 is activated). We can see (fig. 5b,5c) more the distance between the gripper and the object decreases more the speed of the arm's motors decreases in order to tend to 0 when the object is reached. After T_1, the human partner displays an angry facial expression (transmits a negative value), the object value is modified (negative emotional value, A_2 is activated). We can see that the emotional value is now negative although, due to noise, the positive emotional value is high. This shows the learning robustness to the noise. Now, the arm avoids the object as if the object appears to be "dangerous" to the robot.

At this development stage, the robot can reach an object if the self generated reinforcing signal A is positive (the emotional value is positive) and avoid an object if A is negative (the emotional value is negative). The human emotional expression is able to communicate an emotional value to an object (for instance a dangereous object or a interested object) and moreover can modulate the robot behavior.

6 Conclusion

This work suggests the robot/partner system is an autopoietic social system [16] in which the emotional signal and empathy are important elements of the network to maintain the interaction and to allow the learning of more and more complex skills for instance the social referencing. The emotional facial expression is an excellent way to communicate in some ambiguous situations. The relationship between the robot and the partner is improved because an emotional communication can exist. It allows the robot to learn and manipulate an object. This work also emphasizes that the recognition of the other is built through interaction.

Social cognition, including social referencing, may have a stronger emotional foundation and less of a need for complex cognition, than previously thought (e.g. [3]). New neuropsychological studies of the mirror system in emotions[13], the neural basis of intersubjectivity (e.g. [10]) and the current study highlight the important role played by emotion in the developmental emergence of social referencing.

To our knowledge, this is the first system that autonomously learns a coupling between emotion (facial expression recognition) and sensory-motor skills. We developped a real self-supervised developmental sequence contrary to others autors [5, 21]. Here, we don't solve the question of joint attention which is an social referencing skill. Joint attention may also be reached using a learning protocol similar to Nagai[18] (developmental model for the joint attention). We think this approach can provide new interesting insights about how humans can develop social referencing capabilities from sensorimotors dynamics. In contrast to current developmental theory that social referencing is a complex cognitive process of triadic relations, the current work suggests 1) the primacy of emotion in learning, 2) the simple classical conditionning mechanisms by which anothers emotional signal assumes identity with internal emotional states, and 3) a simple system of pairing internal emotional state with object-directed behavior. To improve the functioning of the system, there may be a need to modulate the internal emotional state as a function of intensity of emotional expressions, and to modulate the behavior to the object in accordance, e.g an intense angry expression might involve withdrawing, an intense happy expression might involve picking up more quickly. On going work suggest it might be possible.

Acknowledgments The authors thank J. Nadel, M. Simon and R. Soussignan for their help to calibrate the robot facial expressions and P. Canet for the design of the robot head. Many thanks also to L. Canamero for the interesting discussions on emotion modelling. This study was supported by the European project "FEELIX Growing" IST-045169 and also the French Region Ile de France (Digiteo project). P. Gaussier thanks also the Institut Unisersitaire de France for its support.

References

1. P. Andry, P. Gaussier, S. Moga, J.P. Banquet, and J. Nadel. Learning and communication in imitation: An autonomous robot perspective. *IEEE transactions on Systems, Man and Cybernetics, Part A*, 31(5):431–444, 2001.
2. C. Balkenius and J. Moren. Emotional learning: a computational model of the amygdala. *Cybernetics and Systems*, 6(32):611–636, 2000.

3. K.A. Bard, D.A. Leavens, D. Custance, M. Vancatova, H. Keller, O. Benga, and C. Sousa. Emotion cognition: Comparative perspectives on the social cognition of emotion. *Cognitie, Creier,Comportament (Cognition, Brain, Behavior), Special Issue: "Typical and atypical development"*, 8:351–362, 2005.
4. S. Boucenna, P. Gaussier, and P. Andry. What should be taught first: the emotional expression or the face? *epirob*, 2008.
5. Cynthia Breazeal, Daphna Buchsbaum, Jesse Gray, David Gatenby, and Bruce Blumberg. Learning from and about others: Towards using imitation to bootstrap the social understanding of others by robots. *Artif. Life*, 11(1-2):31–62, 2005.
6. Sylvain Chevallier and Philippe Tarroux. Covert attention with a spiking neural network. In *International conference on computer vision systems*, volume 5008 of *Lecture notes in computer science*, pages 56–65. Springer, 2008.
7. E. Devouche and M. Gratier. Microanalyse du rythme dans les échanges vocaux et gestuels entre la mère et son bébé de 10 semaines. *Devenir*, 13:55–82, 2001.
8. P. Ekman and W.V. Friesen. Facial action coding system: A technique for the measurement of facial movement. *Consulting Psychologists Press, Palo Alto, California*, 1978.
9. Ippei Fukuyori, Yutaka Nakamura, Yoshio Matsumoto, and Hiroshi Ishiguro. Flexible control mechanism for multi-dof robotic arm based on biological fluctuation. *From Animals to Animats 10*, 5040:22–31, 2008.
10. V. Gallese. The roots of empathy: The shared manifold hypothesis and neural basis of intersubjectivity. *Psychopathology*, 36:171–180, 2003.
11. P. Gaussier, K. Prepin, and J. Nadel. Toward a cognitive system algebra: Application to facial expression learning and imitation. *In Embodied Artificial Intelligence, F. Iida, R. Pfeiter, L. Steels and Y. Kuniyoshi (Eds.) published by LNCS/LNAI series of Springer*, pages 243–258, 2004.
12. L. Itti and C. Koch. Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2(3):194–203, Mar 2001.
13. C. Keyser J. Bastiaansen, M. Thioux. Evidence for mirror systems in emotions. *Phil. Trans. R. Soc. B*, 364:2391–2404, 2009.
14. M.D. Klinnert, J.J. Campos, J.F. Sorce, R.N. Emde, and M. Svejda. The development of the social referencing in infancy. *Emotion in early development*, 2:57–86, 1983.
15. D.G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2:91–110, 2004.
16. H.R. Mataruna and F.J. Varela. *Autopoiesis and Cognition: the realization of the living.* Reidel, Dordrecht, 1980.
17. J. Nadel, M. Simon, P. Canet, R. Soussignan, P. Blanchard, L. Canamero, and P. Gaussier. Human responses to an expressive robot. *In Epirob 06*, 2006.
18. Yukie Nagai, Koh Hosoda, Akio Morita, and Minoru Asada. A constructive model for the development of joint attention. *Connect. Sci.*, 15(4):211–229, 2003.
19. C.L. Russell, K.A. Bard, and L.B. Adamson. Social referencing by young chimpanzees (pan troglodytes). *journal of comparative psychology*, 111(2):185–193, 1997.
20. N.A. Schmajuk. A neural network approach to hippocampal function in classical conditioning. *Behavioral Neuroscience*, 105(1):82–110, 1991.
21. Andrea Lockerd Thomaz, Matt Berlin, and Cynthia Breazeal. An embodied computational model of social referencing. *In IEEE International Workshop on Human Robot Interaction (RO-MAN)*, 2005.
22. Bernard Widrow and Marcian E. Hoff. Adaptive switching circuits. In *IRE WESCON*, pages 96–104, New York, 1960. Convention Record.