Development of first social referencing skills: Emotional interaction as a way to regulate robot behavior

Sofiane Boucenna, Philippe Gaussier,Laurence Hafemeister

***Abstract*—In this work, we study 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 sensory-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 sensory-motor system not dedicated to social interactions.**

***Index Terms*—Human-Robot interaction, emotion, social referencing, sensory-motor architecture**

I. 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 applications and development understanding. In particular, we propose that social referencing, gathering information through emotional interaction, fulfills this goal. Social referencing is a developmental process incorporating the ability to recognize, understand, respond to and alter behavior in response to the emotional expressions of a social partner. It allows an infant to seek information from another individual and to use that information to guide his/her behavior toward an object or event [43]. Gathering information through emotional interactions seems to be a fast and efficient way to trigger learning. This is especially evident in early stages of human cognitive development, but also in other primates [65]. Social referencing ability might provide the infant (or a robot) valuable information concerning the environment and the outcome of its behavior. In social referencing, a good (or bad) object or event is identified or signaled with an emotional message. There is no need for verbal interactions. 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 their recognition can be learned autonomously very quickly [12]. Our idea is that social referencing as well as facial expression recognition can emerge from a simple sensory-motor system. All our work is based on the idea of the perception ambiguity. In this case, the inability at first to differentiate our own body from the body of other if the actions of the other 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 first the existence of others, our robot is able to learn to catch or avoid object not related to any direct reward. Hence, we advocate the idea that social referencing can be bootstrapped from a simple sensory-motor system not dedicated to social interactions.

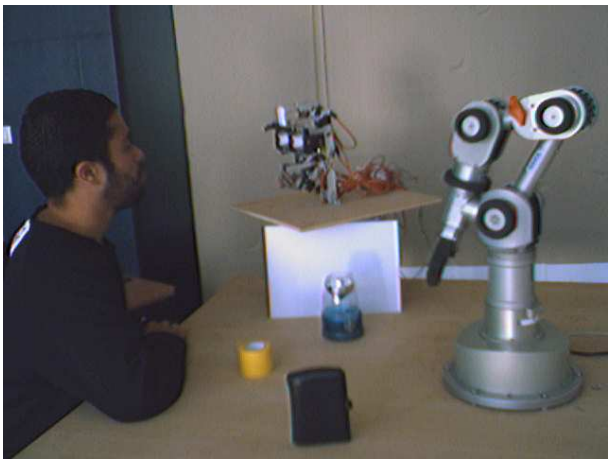

Fig. 1. Experimental set-up for social referencing. The robot relies upon the use of its expressive head which is also able to recognize facial expressions. the robotic arm will reach the positive objects and avert the negative objects after emotional interactions with a human partner.

In the next section, we will show a developmental approach of social referencing, where all the robot abilities such as the development of facial expressions recognition (section V), the association of emotional value to an object (section VII) and finally the control of the arm according to emotional stimuli (section VIII), are learned through interactions with its environment. Moreover, each ability can be learned autonomously and online, and, the social referencing may emerge once all these cognitive abilities have been learned. An important point is that the sensory-motor architecture can resolve these differents tasks based on a cascade of conditioning networks (section IV).

II. Related Work

Many researchers emphasize that the emotions involve "physiological arousal, expressive behaviors, and conscious experience" [54] or, are important for survival [20], [22], [45]. However, there are clearly no agreements on the underlying mechanisms. For instance, James and Lange [38], [44] consider emotions as direct consequences of physiological modifications in reaction to the interactions with the environment. Cannon-Bard [8], [17] supports that emotion is the result of a brain processing (centralist theory: physiological changes are the results of the triggering in the brain of a given emotional state). There is a wide spectrum of models, mostly dedicated to address only one aspect of emotions. For instance, if we focus on emotion expression then the opposition will be between discrete models of emotions (Facial Action Coding System [26]) versus dimensional/continuous models of emotions that suppose any emotion may by expresses as a point in a low dimensional space [66]. Classical models of emotions consider either the communicational aspect of emotions (for instance the emotions conveyed by the facial expressions) or the second order control necessary for survival purpose when the autonomy of the system is an issue. [32] show the interdependence of communication and meta-control aspects of emotion. They propose the idea that emotions must be understood as a dynamical system linking two controllers: one devoted to social interactions (i.e. communication aspects) and another one devoted to the interactions within the physical world (i.e. metacontrol of a more classical controller).

Starting from the neurobiological substrate of the visceral brain [60] (with the regulation loop connecting the thalamus, the hypothalamus, the hippocampus and the cingular cortex), we would like to understand how basic emotions [62] can emerge and become complex cognitive processes involving planning and inhibition of action [20]. From this literature [2], [3], [15], [21], [34], [59], [58], we know that a lot of structures are involved even for the "basic" emotions. Yet, physical and social interactions are certainly not governed by independent controllers and must share some common substructures. Moreover, we want to investigate how emotions can bootstrap complex tasks such as the social referencing [43], [65] and how an agent (robot) can develop this cognitive task.

The development of social referencing skills implies the recognition of emotional signals (providing a value to a stimulus), the recognition of stimuli/objects and the ability to perform some simple actions on these objects. Here, we will suppose the existence of a very simple reflex pathway allowing the simulation of pain and pleasure from an adhoc tactile sensor (e.g. conductive[1] objects). This signal allows the association of objects with positive or negative values and next their grasping or the avoidance according to a sensory-motor controller (see section X). One very difficult part is related to the facial expression recognition and to a lesser extends to object recognition which is generally performed with specific algorithms.

[1]measure of the object conductivity: $R = 1K\Omega$ for positive objects, $R = 0K\Omega$ for negative objects and $R > 10K\Omega$ for neutral objects (usual objects) with no hidden resistor.

In the field of image processing, solutions for the facial expressions recognition usually use algorithms to frame the image around the face [69] before performing the expression recognition. When these techniques involve some learning or optimization, the problem of autonomous learning is not addressed. Some methods are based on Principal Components Analysis (PCA) and use a batch approach for learning (offline learning). For example, the LLE (Locally Linear Embedding) [48] and [74] perform a dimension reduction on the input vectors. Neuronal methods have also been developed for facial expression recognition. In Franco and Treves [27], the network uses a multi-layer network with a classical supervised learning rule (again offline learning from a well-labeled database). The designer must determine the number of neurons that are associated with different expressions according to their complexity. Other methods are based on face models that attempt to match the face (see, for example, the appearance model [5]). Yu [73] uses a support vector machine (SVM) to categorize the facial expressions. Wiskott [71] uses Gabor wavelets to code the facial features, such as with 'jets'. These features are inserted into a labeled graph in which the nodes are 'jets' and the links are the distances between the features in the image space (i.e., the distance between both eyes); the recognition is performed through graph matching. Other sophisticated models compute head-pose invariant facial expression recognition from a set of characteristic facial points [64]. However, all of these techniques use offline learning and need to access the entire learning database. They attempt to introduce a substantial amount of a priori analysis to improve the performances of the system. Moreover, the databases are usually cleaned before use: the faces are framed (or only the face is presented in the image), and human experts label the facial expressions. Hence, the problem of online and autonomous learning is usually not a relevant issue.

With respect to interactive robots, our focus on the online development of interactive behaviors induces specific constraints that are usually forgotten. Breazeal [14] designed Kismet, a robot head that can recognize human facial expressions. Because of an interaction game between the human and the robot, kismet learns to mimic the human's facial expressions. In this study, there is a strong a priori belief about what is a human face. Important focus points, such as the eyes, the eye-brows, the nose, and the mouth, are pre-specified and thus expected. These strong expectations lead to a lack of autonomy because the robot must have specific knowledges (what is a human face) to learn the facial expressions. Breazeal [14] manages a large number of different sensory inputs and motor outputs, showing that the diversity of sensory signals and action capabilities can strongly improve the recognition performances and the acceptability of the robot as a partner. Other studies using robot heads, such as Einstein's robot [72], explore the process of self-guided learning of realistic facial expression production by a robotic head (31 degrees of freedom). Facial motor parameters were learned using feedback from real-time facial expression recognition from video. These studies are complementary to our approach because they show that learning to produce facial expressions can be accomplished

by using the same approach as the approach that we use for expression recognition.

More and more robotics studies are interested in using emotions to regulate the robot behavior [14], [68], [39]. [46] and [47] shows a social robot with empathic capabilities that acts as a chess companion for children. The robot is endowed with empathic capabilities to improve the relationship between children and the robot. In this model, the robot needs to model the childs affective states and adapt its affective and social behavior in response to the affective states of the child. However, these models have a number of a priori and they don't allow acquiring generic models. In these experiments, the robots have many degree of freedom but their adaptive behaviors are minimum. From our point of view, these studies don't focus on the development of cognitive capabilities; they are not interested about how the robot can develop skills autonomously.

Contrary to all these studies, in the following sections, we will underline the robot's ability to develop autonomous and online.

III. Material and Method

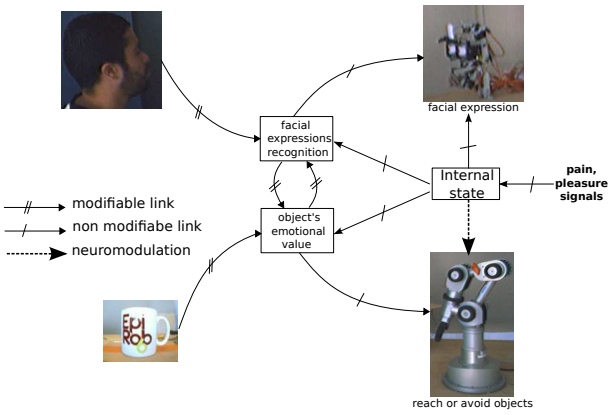

Fig. 2. Simplified model for social referencing. This model highlights the bidirectional interactions. The emergence of the social referencing capability is possible only through the interaction with a human partner.

Our social referencing experiment (Fig. 1) uses the following set-up: a robotic head able to recognize and imitate facial expressions, and, a Katana arm able to interact with different objects. One camera is turned toward the workspace where the Katana arm can reach objects. In this experiment, we used 2 cameras to simplify and to avoid the problem of alternating attention. As a consequence, the robot (head, arm, and camera) can interact with the human partner and can manipulate the objects. In this case, the robot can interact with the social environment as well as the physical environment. 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. This study shows that the emotional interaction allows changing the robot emotional state in order to regulate a robot's behavior (communication of an emotional state). We will attempt to highlight a developmental trajectory where the robot learns skills such as the facial expressions recognition, the face detection and the control of arm (visuo-motor learning). The autonomous learning of these abilities allows the emergence of the social referencing.

For each skill, the visual processing is the same. The visual attention on potentially interesting regions (or object/face) is controlled by a reflex mechanism that allows the robot to focus its gaze. The focus points are the result of a local competition performed on the convolution between a DOG (difference of Gaussians) filter and the norm of the gradient of the input image (we use the same architecture for place and object recognition [31], [50]). This process allows the system to focus more on the corners and ends of the lines in the image. The main advantages of this process over the SIFT (Scale Invariant Feature Transform) [49] method are its computational speed and a smaller number of extracted focus points. For each focus point in the image, a local view centered on the focus point is extracted: either a log-polar transform or Gabor filters are applied (Fig. 3) to obtain an input image or a vector more robust to the rotations and distance variations.

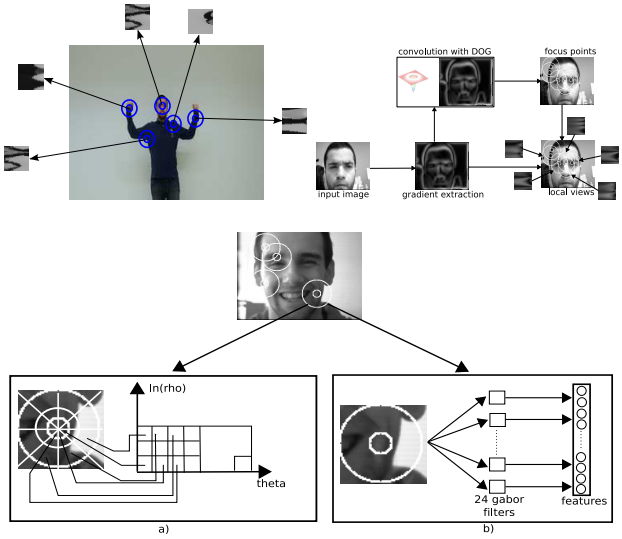

Fig. 3. The robot visual system uses a sequential exploration of the image. A gradient extraction is performed on the input image (256x192 pixels). A convolution with a Difference of Gaussian (DOG) provides the focus points. At last, the local views are extracted around each focus point. The visual features are: a) the local log-polar transform increasing the robustness of the extracted local views to small rotations and scale variations (its radius is 20 pixels). b) Gabor filters are applied to obtain a more robust signature (the Gabor filters are 60x60); the features extracted for each convolution with a Gabor filter are the mean and the standard deviation.

The robotic head learns to recognize emotional facial expressions autonomously [13]. The facial expressions learning can be learned through an imitation between the robot and the human partner. First, the robot internal emotional state triggers one specific expression and the human mimics the robot face[2]. The robot can learn to associate its

[2]in natural condition, [55] showed that the human resonates to the robot facial expression. Here, the instruction was to mimic the robot head facial expressions

internal emotional state with the human's facial expression. The robot associates what it is doing with what it is seeing. After 2 minutes of real time learning, the robot is able to recognize the human facial expressions as well as to mimic them.

After the learning of these capabilities, the eye-arm system can learn visuo-motor associations to reach several positions in the workspace [4], [23] and appendix. A dynamic system is used to smooth the trajectory [28]. This dynamic system uses a reinforcing signal in order to reach or avoid a position in the workspace. The signal can be either related to the reflex pathway (object conductivity associated to positive or negative signals) or learned through the association to an emotional signal; for example, a joy facial expression will be associated to a positive signal and an angry facial expression to a negative signal.

The tested scenario is the following: The robot is in neutral emotional state, 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 a negative facial expression (anger), the value associated to this object becomes 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 becomes the reinforcing signal that the arm uses so as to move. In this scenario, we attempt to emphasise the emotional dimension. The emotion is a way to communicate with the robot. The recognition of the emotional state regulates the robot internal state and adapts the robot's behavior to the environment.

IV. PerAc architecture: as a building block

In this section, we summarize the properties of the generic sensory-motor architecture (PerAc architecture) used as a building block in the following section (Fig. 4). PerAc learns sensory-motor conditionings [30] in order to form a perception as a dynamical sensory-motor attractor. The low-level pathway consists in reflex behaviors. The conditioning pathway allows anticipating reflex behaviors through the learning. This learning performs associations between the recognition of sensory information (high-level) and the reflex behavior (low-level).

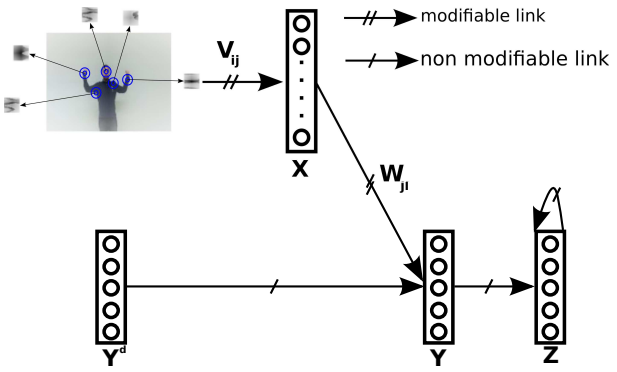

Fig. 4. Sensory-motor architecture based on Neural Networks.

For each focus point in the image, a local view centered on the focus point is extracted (Fig. 3). The extracted local view around each focus point is learned and recognized by a group of neurons X (visual features) using a k-means variant that allows online learning and real-time functions [42] called SAW (Self Adaptive Winner takes all):

$$X_j = net_j \cdot H_{max(\gamma, \overline{net}+\sigma_{net})}(net_j) \quad (1)$$

$$net_j = 1 - \frac{1}{N}\sum_{i=1}^{N}|V_{ij} - I_i| \quad (2)$$

X_j is the activity of neuron j in the group X. $H_\theta(x)$ is the Heaviside function [3]. Here, γ is a vigilance parameter (the threshold of recognition). When the prototype recognition is below γ, then a new neuron is recruited (incremental learning). $\overline{net}$ is the average of the output, and σ_{net} is the standard deviation. This model allows the recruitment to adapt to the dynamics of the input and to reduce the importance of the choice of the vigilance γ. Hence, the vigilance γ can be set to a low value to maintain only a minimum recruitment rate. The learning rule allows both one-shot learning and long-term averaging. The modification of the weights is computed as follows:

$$\Delta V_{ij} = {\delta_j}^k(a_j(t)I_i + \epsilon(I_i - V_{ij})(1 - X_j)) \quad (3)$$

with $k = ArgMax(a_j)$, $a_j(t) = 1$ only when a new neuron is recruited; otherwise, $a_j(t) = 0$. Here, ${\delta_j}^k$ is the Kronecker symbol [4], and ε is the adaptation rate for performing long-term averaging of the stored prototypes. When a new neuron is recruited, the weights are modified to match the input (the term $a_j(t).I_i$). The other part of the learning rule, $\varepsilon(I_i - V_{ij}).(1 - X_j)$, averages the already learned prototypes (if the neuron was previously recruited). The more the inputs are close to the weights, the less the weights are modified. Conversely, the less the inputs are close to the weights, the more they are averaged. The quality of the results depends on the ε value. If ε is chosen to be too small, then it will have only a small impact. Conversely, if ε is too large, then the previously learned prototypes can be unlearned. Because of this learning rule, the neurons in the X group learn to average the prototypes of the objects. One neuron can be recruited to store a new pattern when the none of the neurons is sufficiently activated. The initial number of neurons has to be large enough to avoid recruitment failure (lack of neurons to be recruited).

In our network, the Y group associates the activity of the visual features X with the proprioception Y^d of the robot (a simple conditioning mechanism using the Least Mean Square (LMS) rule [70]):

[3]Heaviside function:

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

[4]Kronecker function:

$${\delta_j}^k = \begin{cases} 1 & \text{if } j = k \\ 0 & \text{otherwise} \end{cases}$$

During the learning phase:

$$Y_l = \sum_j W_{jl} \cdot X_j \qquad \Delta W_{jl} = \epsilon_1 \cdot X_j \cdot (Y_l^d - Y_l) \quad (4)$$

After the learning phase:

$$Y_l = \sum_j W_{jl} \cdot X_j + Y_l^d \quad (5)$$

Y corresponds to the sensory-motor association and W_{jl} is the synaptic weights between X and Y. Y predicts Y^d, based on the input X. Hence, Y^d is the target output. Y^d is a vector with real components (continuous values $>= 0$). Y is also a vector but a Winner Takes All procedure is used to transform the analog values into binary values according to the WTA law.

Z corresponds to a short term memory (accumulation of all focus points). Z is used to sum and to filter the Y activities on a short period ($T < 1$). The Z_i highest activity triggers the i^{th} motor action (WTA mechanism). After learning, the associations between X the view recognition and Y are strong enough to bypass the low level reflex activity coming from the Y^d. Each focus point is associated with a motor action (Y) and Z is accumulation over all the focus points:

$$Z_i(t + dt) = T \cdot Y_i(t) + (1 - T) \cdot Z_i(t) \quad (6)$$

We will show that the robot can develop cognitive abilities thanks to this architecture. A cascade of this architecture allows the learning of increasingly complex tasks such as the social referencing. This sensory-motor architecture will allow building some complex behaviors such as the facial expressions recognition (section V), the association of an emotional value to an object (section VII) or the control of the robotic arm (section X).

V. Online learning of facial expression recognition

Here, the robot must learn to recognize (and to understand) the caregiver's facial expressions. We investigate how a robot can develop the recognition of facial expressions such as the baby could perform it. In our case, we limit our work to the recognition of basic facial expressions. The tests are limited to 4 prototypical facial expressions: happiness, sadness, hunger and surprise [36], [26], [25], [61], plus a neutral face (Fig. 5 for the experimental setup). In other studies, we have shown that using an imitation procedure with first prototypical facial expression can be generalized to more analog states (such as more or less happy and more or less smiling) and next that secondary emotional state can be recognized [11].

For sake of simplicity, we focus on the online learning of prototypical facial expression without having a teaching signal that associates a facial expression with a given abstract label (e.g., 'sadness', 'happiness'). In a first series of robotic experiments, we showed that a simple sensory-motor architecture based on a classical conditioning paradigm could learn online to recognize facial expressions if and only if we assume that the robot first produces some facial expressions according to his/her internal emotional state and that the parents next imitate the facial expression of their robot, which

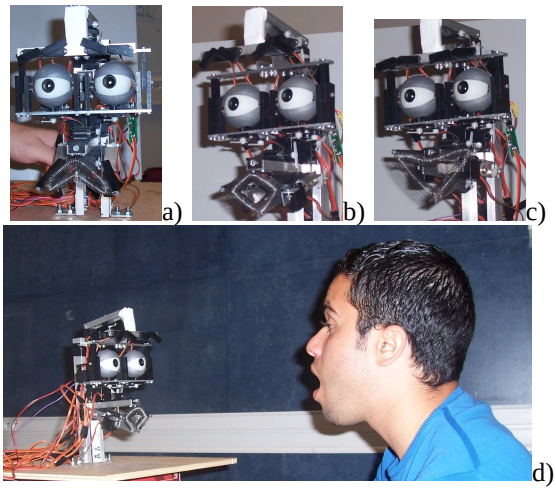

Fig. 5. Examples of robot facial expressions: a) sadness, b) surprise, c) happiness. d) Example of a typical human / robot interaction game (here, the human imitates the robot).

helps the robot to associate these expressions with his/her internal state [13]. In the present study, the robot will be considered as a baby and the human partner will be considered as a parent (the father or mother). At first, the robot knows almost nothing about the environment. Through the interaction with a human, the robot will learn to recognize different facial expressions.

Each of the four facial expressions has been controlled by FACS experts [25]. The validity of this choice could be discussed (especially for the surprise and/or for the choice of the expression names) [41]. However, for our purpose, we need only a small set of facial expressions that are easily recognized and that induce a resonance from the human partner (allowing learning while the human partner is mimicking the robot head). The sensory-motor architecture (Fig. 6 and section IV) learns

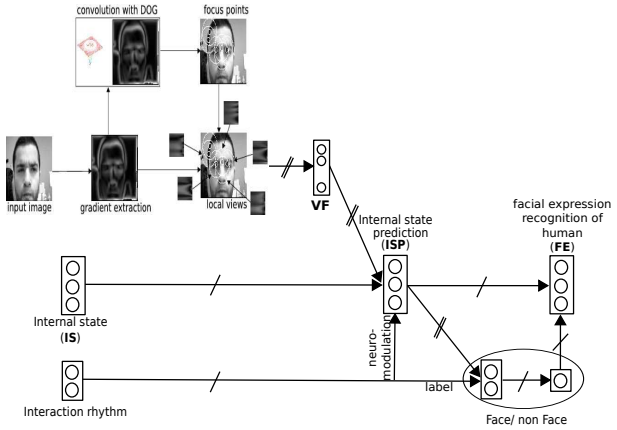

Fig. 6. The global architecture to recognize facial expressions, to imitate and to recognize face from non-face stimuli. Visual processing allows the extraction of sequential local views. The VF group (local view recognition) learns the local views (each group of neurons. A tensorial product is performed between ISP (emotional state: internal state prediction) and a neuromodulation signal, to select the neuron that must learn. The face/non face discrimination is learned by a neural network.

the association between the internal state of the robot and the

visual features. Here, X corresponds to visual features learned and recognized when the visual system explore a face (face features). Y^d corresponds to internal state of the robot. Y corresponds to the facial expression associated (internal state prediction) to one focus point and Z corresponds to the facial expression recognized after the sequential exploration of the image (integration of the answers).

Moreover, the following experimental protocol was adopted: In the first phase of the interaction, the robot produces a random facial expression (sadness, happy, anger, or surprised) plus the neutral face for 2s; then, the robot returns to a neutral face for 2s to avoid human misinterpretation of the robot facial expression (the same procedure is used in psychological experiments). The human participant is asked to mimic the robot head. After this first phase, which lasts between 2 and 3 minutes according to the participant's "patience", the generator of the random emotional states is stopped. If the N.N. (neural network) has learned correctly, then the robot is able to mimic the facial expression of the human partner.

Fig. 7 shows that the interaction with the robot head for a period of 2 minutes can be sufficient for the robot to learn the facial expressions, and, then to imitate the human partner. This incremental learning gains in robustness when the number of human partners increases (expression of sadness can be quite different among people with the lack of action of some action units). These results show the robot capability to recognize the facial expressions of participants who interacted with the robot during the learning phase. Note that this result is sufficient to accomplish the social referencing task because the robot interacts only with known participants (learned during the learning phase).

Moreover, Fig. 8 shows the measure of generalization capabilities which is approximately 38% when 10 subjects interacted with the robotic head during the learning phase, and the success rate is approximately 50% when the robotic head learned with 20 subjects. In our experiment, 20 persons imitated the robot, and then, we asked to a new person to perform facial expressions (the success rate is 65% for joy, 73% anger, 47% for surprised, 4% for sadness and 56% for neutral face).

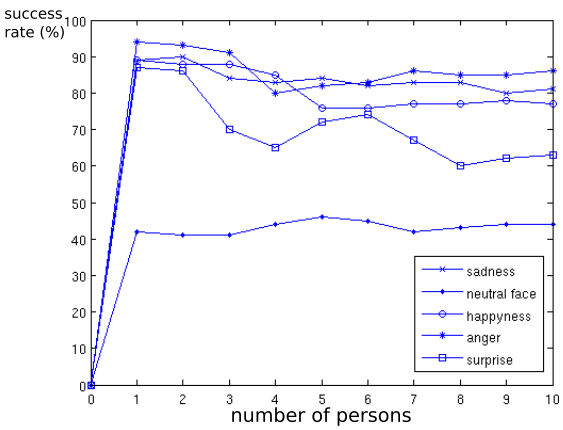

Fig. 7. The success rate for each facial expression. These results are obtained during the natural interaction with the robot head. A total of 10 persons interacted with the robot head. During the learning phase, these humans imitate the robot, and then the robot imitates them. To perform the statistical analyses, each image was annotated with the response of the robot head. The annotated images were analyzed, and the correct correspondence was checked.

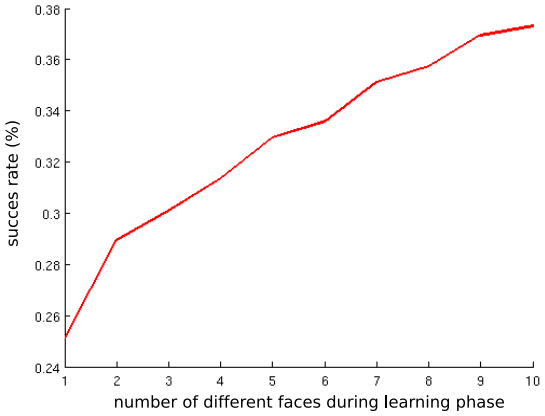

Fig. 8. Measure of generalization capabilities averaged over our 5 categories (4 facial expressions plus the neutral face). This result shows the success rate (y axes) of the facial expression recognition as a function of the number of faces (x axes) that the system learned during the learning phase. The success rate was measured on a database built from images of 10 other participants (1600 images that were never learned). The generalization improves after interaction with increasing numbers of people.

VI. Face from non-face discrimination can emerge through the emotional interaction

Recognizing a face from a non-face can be accomplished autonomously if we accept that learning to recognize a face can occur after learning to recognize a facial expression, and not the opposite, as is classically considered. To perform autonomous learning, we introduced the capability of predicting the rhythm of the interaction [4] to avoid learning when there is no human participant in front of the robot or when the human is not paying attention to the robot (for example, when the human partner is leaving or talking with someone else).

When a participant displays a facial expression, he/she performs whole face or body motions. If the participant imitates the robot, then his/her movement peaks have a frequency that depends on the frequency of changes in the robot facial expressions (in our case, this frequency is constant because the robot facial expression changes after 4s). The interaction rhythm can be predicted by using a prediction of the timing between 2 visual peaks (a stable frequency of interaction of the human partner). A measure of the prediction error can easily be built from the difference in activity between the predicted signal and the non-specific signal itself. In our study, the non-specific signal is the movement produced by the human. The non-specific signal is related to the presence or absence of the human partner. If the error is important, then there is a novelty (the participant is not in the rhythm). Otherwise, the prediction error is small, which involves a good interaction between the participant and the robot. Many studies in psychology underline the importance of synchrony during the interaction between a mother and a baby. For example,

babies are extremely sensitive to the interaction rhythm with their mother [53], [52], [24]. An interruption of the social interaction involves negative feelings (e.g., agitation, tears). However, a rhythmic interaction between a baby and his/her mother involves positive feelings and smiles. These studies show the importance of the interaction rhythm. In our case, the rhythm is used as a neuromodulation signal or a label (see [4] for the application of the same principle to the learning of an arbitrary set of sensory-motor rules and the details of the N.N.):

- a rhythmic interaction is equivalent to a positive neuromodulation: the robot head and the participant produce a coherent action at each instant.
- conversely, an interruption of the interaction is interpreted as a negative neuromodulation.

We consider this second network for the face/non face discrimination that functions in parallel with facial expression recognition. This network learns to predict the rhythm of the interaction, allowing detection if an interacting agent (a human) faces the robot head. The interaction rhythm provides the reinforcement signal to learn to recognize an interacting partner, which is a human, and, more specifically, to learn to recognize his/her face at a short interaction distance (the robot sees the human face and not the other parts of his/her body).

The results linked to this online learning of the face are highly positive. When the face detection is learned and tested using the same participant, the system success rate with that participant tends toward 100%. However, when the face detection is learned with a single participant and is tested on 4 other participants, the system success rate ranges between 29% (for people with beards) and 90% for more ”similar” participants. It is important to consider that the learning was performed during a period of only 2 minutes (in real-time: frame rate 10 Hz) with a single participant. This scenario shows the generalization capabilities of our visual system when focusing the robot's attention on particular visual features. Now, when face detection is learned on 4 participants and the tests are performed on 21 different participants, the system success rate tends toward 95% for face detection (see Fig. 9). The performances improve after the interactions with an increasing number of people.

At this development stage, the robot head is able to recognize and understand the emotional facial expressions and to discriminate the face from a non-face. In the following section, we will show how the robot can assign an emotional signal to an arbitrary object.

VII. Associating an emotional value to an object

This section shows how the facial expressions recognition and the face detection are integrated in order to associate the emotional value to an object using always the same PerAc building block. When the human partner interacts with the robot, the robot uses the human's expressiveness to regulate its behavior.

In our scenario, the robot must consider the output of its facial recognition system only when the human was interacting

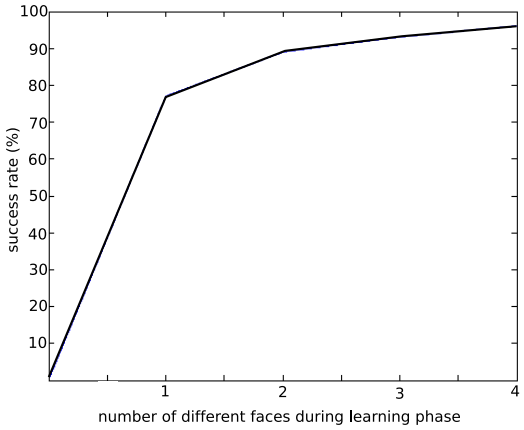

Fig. 9. Face/Non-face recognition and generalization. This result shows that the success rate of face recognition is a function of the number of faces that the system learned on during the learning phase. The results are obtained with 21 people (3360 images). After interacting with only 4 people, the system generalizes to 21 people.

with the robot. Because the robot head performs facial expressions with a known rhythm, it is easy for the N.N. to attempt to predict the visual signal according to its own rhythm. When predictions match the robot action rhythm, this means that one human is interacting with the robot. This solution avoids propagating the emotional recognition when the human doesn't interact with the robot.

As soon as the recognition of human facial expressions has been learned, the human partner can interact with the robotic head to associate an emotional value to an object (positive or negative). The N.N. processes (see Fig. 10 and 16) in the same

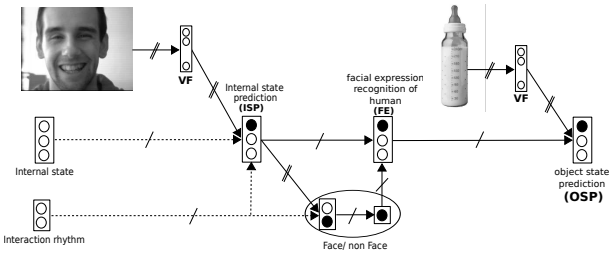

Fig. 10. This sensory-motor architecture shows how the facial expressions recognition and the face detection are integrated in order to associate the emotional value to an object.

way signals from the robot's internal state and information 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 (Refer to (5)). Note that in real life condition the reflex associations should rarely be activated since they are only related to low level signals (internal levels, tactile signals). During the learning of the facial expressions recognition, we bypass natural interactions by a fast and random activation of the different states to obtain enough feedback from the human partners. In case of conflict, between the internal state (IS) and the facial expression recognition (FE), the reflex links connecting IS to

the control of FE (through ISP) are higher than the learned links coming from the recognition of visual features (VF) to ISP. The internal state remains dominant. This means that if the robot touches an object inducing some pain (because of a "tactile" hardwired feedback), the pain signal will win on any previous positive association regarding this object (through social referencing for instance). Recognized visual stimuli (VF) will either be conditioned to internal state prediction (ISP) or to the object state prediction (OSP) (the ISP being "priority" on OSP because of the reflex link from ISP to OSP). In recent works, we have generalized this association capability adding a feedback loop from OSP to ISP to build second order conditioning and to allow the robot learning complex chains of conditioning [1] but this is out of the scope of the present paper. In our experiment, the internal state is absent (the internal state neurons have all null values), the recognized facial expression induces an internal state which is associated with the object (a simple conditioning chain: Fig. 16). Classical conditioning is used to perform the associ-

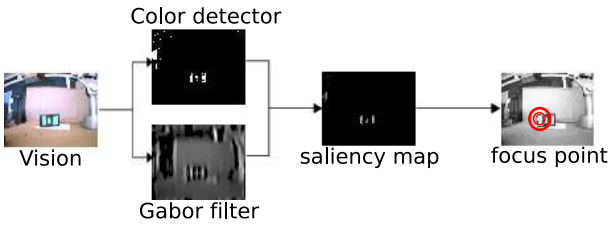

Fig. 11. Visual attention. The system focuses on some relevant features of the image. A saliency map is performed in order to focus an interesting area in the image. Visual primitives are calculated independently (Gabor filters, color detector), a fusion of these primitives is performed in order to find the area that the robot must analyze.

ation between the emotional value transmitted by the human and some local views of the image. An attentional process is also introduced to avoid that the robot spends too much time looking its own arm (see [35], [18] for more information). The robot focuses on colored patches and textures (Fig. 11 and 12). We use a very simple spatial competition between different maps (colors, textures). 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 recognized by the robot. Starting again from our generic architecture (see section IV and the Fig. 4), X corresponds to visual features (object features), Y^d corresponds to the recognized facial expression. Y corresponds to the object emotional value for one focus point and Z corresponds to the global object emotional value after the sequential exploration of one image. According to our sensory-motor architecture, the proprioceptive signal is considered as a training signal for the Y layer. This training signal corresponds to the internal state prediction (facial expression recognized by the robot). Consequently, the sensory-motor architecture associates the internal state prediction with the visual perception (object). To associate an emotional value to an object, the training signal is provided by the facial expression of the human.

The tests were performed with 10 objects (5 positive objects and 5 negative objects). We put on the workspace the different

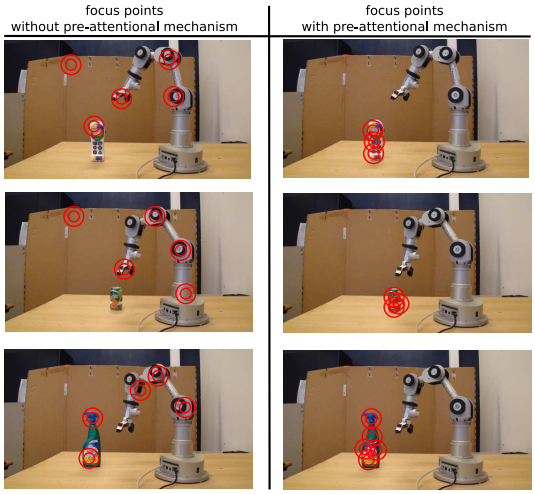

Fig. 12. Visual processing with or without pre-attentional mechanism.

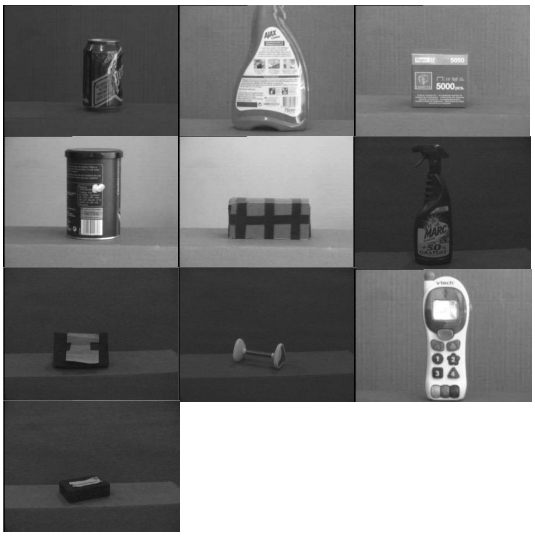

Fig. 13. The 10 objects use during the social referencing experiment.

objects one after another (Fig. 13). Each object is put few seconds in the robot workspace (Fig. 12) and each object is learned as the result of the emotional interaction with the robotic head. During the learning phase, the objects position is fixed (the object doesn't move) and the human partner sends an emotional signal to the robot: a positive signal when the object can give pleasure and a negative signal when the object is dangerous. The recognition of emotional value is 87% for the negative objects and 98% for the positive objects. The success rate difference between the positive and negative objects is only related to the variability of the objects complexity. The success rate shows the robustness of the model despite some variations such as the distance and the object position in the image (the objects are put at different locations).

Hence, the robot is now 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 new situations.

VIII. Emotional interaction regulates the robot's behavior

At this stage of the development, the robot has some capabilities such as the facial expressions recognition, to associate an emotional value to an object and to control his multi-DoF robotic arm (see section X). In this section, we show how the robot can integrate all these capacities to regulate its behavior. In our experimental set-up, the emotional interaction with the human partner can bias the object approach. The objects and the human facial expressions can provide a reinforcing signal allowing the robot's adaptation. Here, A is an emotional reinforcement signal according to the robot emotional state. In others words, the robotic arm can reach or avoid an object according to the parameter A:

$$A = \begin{cases} 1 & \text{if the emotional value is positive} \\ -1 & \text{if the emotional value is negative} \end{cases} \quad (7)$$

In this experiment, one object is put in the robot workspace. If the object is associated to pleasure and/or a smile from the human then the robot reaches the object. On the contrary, if the object is dangerous and/or is associated to a negative expression from human partner, then, the robot avoids the object. Fig. 14 shows the important dynamics induced by the social

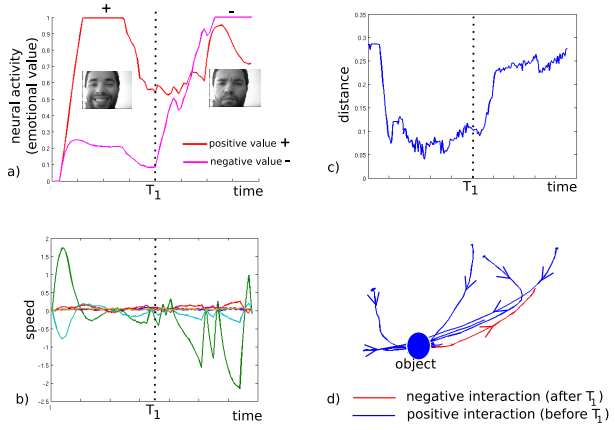

Fig. 14. These curves show: a) the emotional value transmits to the object thanks to the interaction with the human's partner (before T_1 human transmits a positive value after T_1 the human transmits a negative value) b)the speeds of each arm's motor (6 degrees of freedom) c) the distance to the object 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.

referencing architecture. Fig. 14a shows the object's emotional value associated with the facial expressions of the 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. We can see (Fig. 14b,14c) the more the distance between the gripper and the object decreases the 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 face (transmitting a negative value), the object value is modified (negative emotional value). We can see that the emotional value is now negative although, due to noise, the positive emotional value is still 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.

To provide more quantitative results on the robotic arm capability to catch the objects having a positive emotional value. We performed the following experiment: the objects are put at different positions to show that the robot can catch and recognize the objects in the whole workspace. Fig. 15 shows that the robot is able to reach the positive object (92% success rate) and to catch it (with successful prehension: 82% success rate). The robot fails only when the object can't be reached (environment area not surrounded by attractors). In the case the object is negative, the robot avoids the different objects all time (100% success rate). These results highlight the robot's capability to adapt its behavior according to the emotional signal (emotional value associates to object).

	reach	catch
object 1	90%	90%
object 2	90%	80%
object 3	100%	80%
object 4	90%	80%
object 5	90%	80%
average	92%	82%

Fig. 15. Success rate when the robot attempts to catch objects in the environment. The 5 positive objects are put one after the others in the workspace. Each object is put at 10 different positions in the workspace allowing obtaining quantitative results for the prehension of positive objects.

At this level, 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 dangerous object or an interested object) and moreover can modulate the robot behavior.

IX. Conclusion

In our study, the social referencing is seen as a cascade of sensory-motor architecture (Fig. 16). We showed that the robot can learn different behaviors (or tasks) through the interaction with the environment (Fig. 17): facial expressions recognition (B1), face/non face discrimination (B2), the association of an emotional value to an object (B4) or the control of the robotic arm (B3). Each ability can be learned automously and online therefore the social referencing may emerge once all these cognitive abilities will be learned. However, some abilities such as the association of an emotion to an object are inefficient while the facial expressions recognition has been not learned. Consequently, some skills must be learned first so that others can be learned correctly. In our experiment, two different cameras are used: one looking in the direction of the human and another one looking in the direction of the object. Therefore, it was possible to learn the visuo-motor control of the arm in parallel with the learning of the facial expression recognition (Fig. 17a). In practice, the facial expression recognition and the face/non face recognition

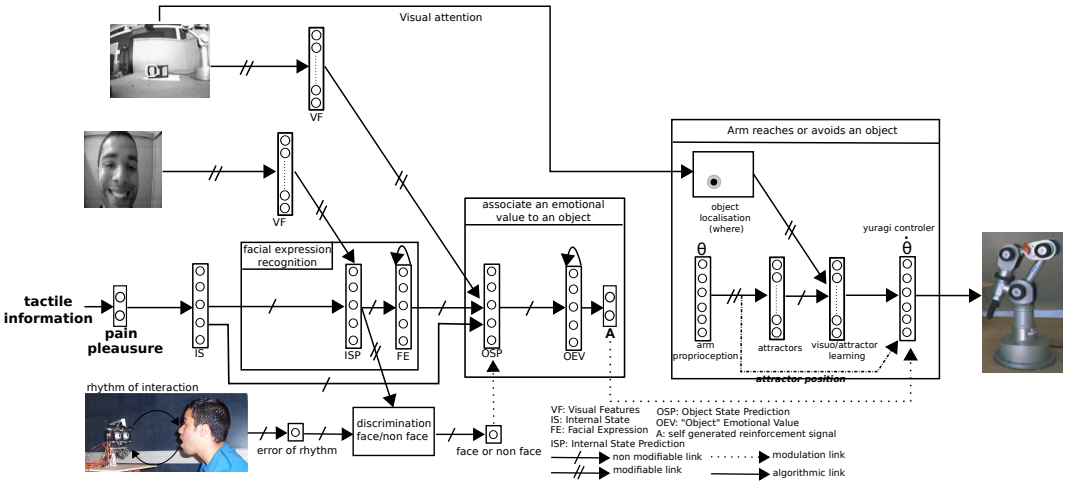

Fig. 16. Global architecture for the social referencing model. Social referencing emerges from the sensory-motor interactions between facial expression recognition, objects emotional value and visuo-motor learning for the arm control. A simple sensory-motor architecture is able to learn and recognize the facial expressions, and then to discriminate between face/non face stimuli (face detection). 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 processes (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)

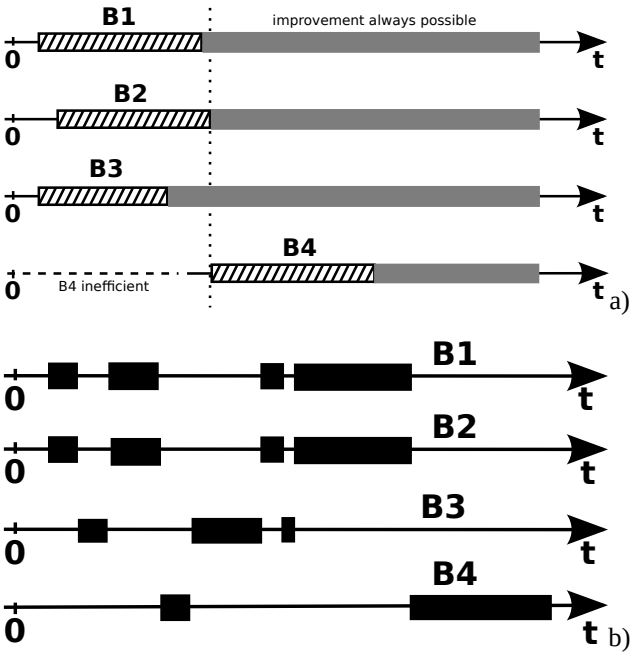

Fig. 17. The different behaviors learned by the robot. a) with 2 cameras (our experimental set-up), the robot learns the different behaviors (or tasks) through the interaction with the environment. B1: the learning of facial expressions. B2: learning of face/non face discrimination. B3: learning of visuo-motor coordination. B4: Emotional interaction regulates the robot's behavior. The gray shows that the learning can always be improved. Learning of these cognitive skills. b) with 1 camera, the development could be closer to the baby development with the need to alternate between the social and physical interactions.

were performed first. Next, the arm control was learned and finally the social referencing was learned. Starting to the baby development, the learning should be continuous and alternate (Fig. 17b). If the two cameras are replaced by a single camera or if the cameras have to look in the same global direction (as for human gaze) then there is a need to add a mechanism to alternate the attention between two directions. A simple oscillatory mechanism could be sufficient to control the visual attention. However, for the learning of the arm control, it would be better to perform this task until some progress has been made in this learning. Hence it is clear that some complex self-evaluation need to be added [67], [57], [6]. In [40], we propose a possible solution but it has not been tested for our problem.

To our knowledge, our architecture is the first one that learns a coupling between emotion (facial expression recognition) and sensory-motor skills. We developed a real self-supervised developmental sequence contrary to others authors [14], [68]. Yet, we don't solve the question of joint attention which is an important issue. Joint attention may also be reached using a learning protocol similar to Nagai [56] (developmental model for the joint attention).

We think our sensory-motor approach can provide new interesting insights about how humans can develop social capabilities from sensorimotor dynamics. For example, studies [9] show that humans use the theory of mind (to assign mental states to the self and to others [63]) for complex social interactions. For example, the false-belief task is became the test for crediting a child with a theory of mind [9]. One consequence of this definition is an emphasis upon representational mental states and knowledge rather than upon emotions, intentions, perceptions. In contrast to current developmental theory which considers the social interactions as a complex cognitive process [9], our works suggest 1) the primacy of emotion in learning, 2) the effectiveness of using a simple conditioning for the learning of facial expressions through an imitation game with the human partner 3) the efficiency of a simple system of pairing internal emotional

state with object-directed behavior. New neuropsychological studies related to the mirror system in emotions [37], the neural basis of intersubjectivity (e.g. [29]) as well as our study highlight the important role played by emotion in the emergence of social referencing. Social cognition, including social referencing, may have stronger emotional foundation and less need for complex cognition than previously thought (e.g. [7]). Our works show that the robot can develop social interactions without a theory of mind, and, we argue that the theory of mind can emerge from social interactions. Therefore, the theory of mind should be considered as a developmental processes [19].

To improve the functioning of our architecture, 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, and an intense happy expression might involve picking up more quickly. Ongoing work suggests it might be possible by using a population coding at the different stages of the architecture.

The facial expressions are an excellent way to bootstrap complex sensory-motor learning. The relationship between the robot and the partner is dramatically changed thanks to an emotional communication. It allows the robot to learning and manipulating an object. The dynamical interactions between the robot and the human participant allows to simplify learning, for example, the robot can learn autonomously and online the facial expressions if the human partner mimics the robot (resonate to the robot facial expression). Consequently, we show that the dynamics of interaction and simple rules (PerAc architecture) are sufficient to have an autonomous robot. This work suggests the robot/partner system is an autopoietic social system [51] 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.

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 and K. Bard for the interesting discussions on emotion modeling. This study was supported by the European project "FEELIX Growing" IST-045169, French ANR INTERACT and also the French Region Ile de France (Digiteo project). P. Gaussier thanks also the Institute Universitaire de France for its support.

REFERENCES

[1] D. Vidal A. Karaouzene, P. Gaussier. A robot to study the development of artwork appreciation through social interactions. *ICDL-EPIROB*, page to appear, 2013.

[2] R. Adolphs, D. Tranel, and A.R. Damasio. The human amygdala in social judgment. *Nature*, 393(6684):470–474, 1998.

[3] R. Adolphs, D. Tranel, H. Damasio, and A.R. Damasio. Fear and the human amygdala. *The Journal of neuroscience*, 15(9):5879–5891, 1995.

[4] 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.

[5] M. Dang B. Abboud, F. Davoine. Facial expression recognition and synthesis based on an appearance model. *Signal Processing: Image Communication*, 19:723–740, 2004.

[6] A. Baranès and P.Y. Oudeyer. R-iac: Robust intrinsically motivated exploration and active learning. *Autonomous Mental Development, IEEE Transactions on*, 1(3):155–169, 2009.

[7] 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.

[8] P. Bard. A diencephalic mechanism for the expression of rage with special reference to the central nervous system. *American journal of psychology*, 84:490–513, 1928.

[9] S. Baron-Cohen, A.M. Leslie, and U. Frith. Does the autistic child have a theory of mind? *Cognition*, 21(1):37–46, 1985.

[10] A. Billard, S. Calinon, R. Dillmann, and S. Schaal. Survey: Robot programming by demonstration. In *Handbook of Robotics*, volume chapter 59. MIT Press, 2008.

[11] S. Boucenna. *De la reconnaissance des expressions faciales à une perception visuelle partagée: une architecture sensori-motrice pour amorcer un référencement social d'objets, de lieux ou de comportements*. PhD thesis, Université de Cergy Pontoise, 2011.

[12] S. Boucenna, P. Gaussier, and P. Andry. What should be taught first: the emotional expression or the face? *epirob*, 2008.

[13] S. Boucenna, P. Gaussier, P. Andry, and L. Hafemeister. Imitation as a communication tool for online facial expression learning and recognition. In *IROS*, pages 5323–5328, 2010.

[14] C. Breazeal, D. Buchsbaum, J. Gray, D. Gatenby, and B. Blumberg. Learning from and about others: Towards using imitation to bootstrap the social understanding of others by robots. *Artificial Life*, 11(1-2):31–62, 2005.

[15] J. Burgdorf and J. Panksepp. The neurobiology of positive emotions. *Neuroscience & Biobehavioral Reviews*, 30(2):173–187, 2006.

[16] S. Calinon, F. Guenter, and A. Billard. On learning, representing and generalizing a task in a humanoid robot. *IEEE transactions on systems, man and cybernetics, Part B. Special issue on robot learning by observation, demonstration and imitation*, 37(2):286–298, 2007.

[17] W.B. Cannon. Bodily changes in pain, hunger, fear and rage. *Southern Medical Journal*, 22(9):870, 1929.

[18] S. Chevallier and P. 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.

[19] C. Colonnesi, C. Rieffe, W. Koops, and P. Perucchini. Precursors of a theory of mind: A longitudinal study. *British Journal of Developmental Psychology*, 26(4):561–577, 2008.

[20] A. Damasio. *Descartes' error: Emotion, reason, and the human brain*. Penguin Books, 2005.

[21] A.R. Damasio. Emotion in the perspective of an integrated nervous system. *Brain Research Reviews*, 1998.

[22] C Darwin. The expression of emotion in man and animals. *Chicago:University of Chicago Press (Originally Published in 1872)*, 1965.

[23] Antoine de Rengervé, Sofiane Boucenna, Pierre Andry, and Philippe Gaussier. Emergent Imitative Behavior on a Robotic Arm Based on Visuo-Motor Associative Memories. In *IEEE/RSJ International Conference on Intelligent Robots and systems (IROS'10)*, pages 1754–1759, Taipei, Taiwan, October 2010.

[24] 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.

[25] 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.

[26] P. Ekman, W.V. Friesen, and P. Ellsworth. Emotion in the human face: Guide-lines for research and an integration of findings. *New York: Pergamon Press*, 1972.

[27] L. Franco and A. Treves. A neural network facial expression recognition system using unsupervised local processing. *2nd international symposium on image and signal processing and analysis. Cognitive neuroscience*, 2:628–632, 2001.

[28] I. Fukuyori, Y. Nakamura, Y. Matsumoto, and H. Ishiguro. Flexible control mechanism for multi-dof robotic arm based on biological fluctuation. *From Animals to Animats 10*, 5040:22–31, 2008.

[29] V. Gallese. The roots of empathy: The shared manifold hypothesis and neural basis of intersubjectivity. *Psychopathology*, 36:171–180, 2003.

[30] P. Gaussier and S. Zrehen. Perac: A neural architecture to control artificial animals. *Robotics and Autonomous Systems*, 16(2-4):291–320, 1995.

[31] C. Giovannangeli, Ph. Gaussier, and J.-P. Banquet. Robustness of visual place cells in dynamic indoor and outdoor environment. *International Journal of Advanced Robotic Systems*, 3(2):115–124, jun 2006.

[32] C. Hasson, P. Gaussier, and S. Boucenna. Emotions as a dynamical system: the interplay between the meta-control and communication function of emotions. *Paladyn*, 2(3):111–125, 2011.

[33] A. Ijspeert, J. Nakanishi, and S. Schaal. learning attractor landscapes for learning motor primitives. In *advances in neural information processing systems 15*, pages 1547–1554. cambridge, ma: mit press, 2003.

[34] S. Ikemoto and J. Panksepp. The role of nucleus accumbens dopamine in motivated behavior: a unifying interpretation with special reference to reward-seeking. *Brain Research Reviews*, 31(1):6–41, 1999.

[35] L. Itti and C. Koch. Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2(3):194–203, Mar 2001.

[36] C. Izard. The face of emotion. *Appleton Century Crofts*, 1971.

[37] C. Keyser J. Bastiaansen, M. Thioux. Evidence for mirror systems in emotions. *Phil. Trans. R. Soc. B*, 364:2391–2404, 2009.

[38] W. James. What is an emotion. *Mind*, 9:188–205, 1884.

[39] H. Jasso, J. Triesch, and Gedeon Deak. A reinforcement learning model of social referencing. *Development and Learning. ICDL*, pages 286–291, 2008.

[40] Cuperlier N. Gaussier P. Jauffret A., Grand C. and Tarroux P. How can a robot evaluate its own behaviour? a generic model for self-assesment. *International Joint Conference on Neural Networks (IJCNN)*, 2013.

[41] N. L. Stein J.M. Jenkis, K. Oatley. *Human Emotions*, chapter The communicative theory of emotions. Blackwell edition, 1998.

[42] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24:881–892, 2002.

[43] 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.

[44] C.G. Lange. The mechanism of the emotions. *The Classical Psychologists. Boston: Houghton Mifflin*, 1912, 1885.

[45] J.E. LeDoux. Brain mechanisms of emotion and emotional learning. *Current opinion in neurobiology*, 2(2):191–197, 1992.

[46] I. Leite, G. Castellano, A. Pereira, C. Martinho, and A. Paiva. Long-term interactions with empathic robots: Evaluating perceived support in children. In Shuzhi Sam Ge, Oussama Khatib, John-John Cabibihan, Reid Simmons, and Mary-Anne Williams, editors, *ICSR*, volume 7621 of *Lecture Notes in Computer Science*, pages 298–307. Springer, 2012.

[47] I. Leite, G. Castellano, A. Pereira, C. Martinho, and A. Paiva. Modelling empathic behaviour in a robotic game companion for children: an ethnographic study in real-world settings. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, HRI '12, pages 367–374, New York, NY, USA, 2012. ACM.

[48] D. Liang, J. Yang, Z. Zheng, and Y. Chang. A facial expression recognition system based on supervised locally linear embedding. *Pattern recognition Letter.*, 26:2374–2389, 2005.

[49] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2:91–110, 2004.

[50] M. Maillard, O. Gapenne, L. Hafemeister, and Ph. Gaussier. Perception as a dynamical sensori-motor attraction basin. In Capcarrere et al., editor, *Advances in Artificial Life (8th European Conference, ECAL)*, volume LNAI 3630 of *Lecture Note in ArtificialIntelligence*, pages 37–46. Springer, sep 2005.

[51] H.R. Mataruna and F.J. Varela. *Autopoiesis and Cognition: the realization of the living*. Reidel, Dordrecht, 1980.

[52] D. Muir and J. Nadel. Infant social perception. In A. Slater, editor, *Perceptual development*, pages 247–285. Hove: Psychology Press, 1998.

[53] L. Murray and C. Trevarthen. Emotional regulation of interaction between two-month-olds and their mother's. In *Social perception in infants*, pages 177–197. N.J. Norwood: Ablex, 1985.

[54] David G Myers. Theories of emotion. *Psychology: Seventh Edition, New York, NY: Worth Publishers*, 2004.

[55] 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.

[56] Y. Nagai, K. Hosoda, A. Morita, and M. Asada. A constructive model for the development of joint attention. *Connect. Sci.*, 15(4):211–229, 2003.

[57] P. Y. Oudeyer, F. Kaplan, and V. Hafner. Intrinsic motivation systems for autonomous mental development. *IEEE Trans. Evolutionary Computation*, 11(2):265–286, 2007.

[58] J. Panksepp. Feeling the pain of social loss. *Science*, 302(5643):237–239, 2003.

[59] J. Panksepp et al. A critical role for "affective neuroscience" in resolving what is basic about basic emotions. *Psychological review*, 99(3):554–560, 1992.

[60] J.W. Papez. A proposed mechanism of emotion. *Archives of Neurology and Psychiatry*, 1937.

[61] R. Plutchick. A general psychoevolutionary theory of emotion. *Emotion: Theory, research and experience*, pages 3–33, 1980.

[62] R. Plutchik. A general psychoevolutionary theory of emotion. *Emotion: Theory, research, and experience*, 1(3):3–33, 1980.

[63] D. Premack and G. Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526, 1978.

[64] O. Rudovic, M. Pantic, and I. Patras. Coupled gaussian processes for pose-invariant facial expression recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (in press)*, October 2012.

[65] 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.

[66] J. Russell and L. Feldman-Barrett. Core affect, prototypical emotional episodes, and others things called emotion: Dissecting the elephant. *Journal of Personality and Social Psychology*, 76:805–819, 1999.

[67] J. Schmidhuber. Curious model-building control systems. In *Neural Networks, 1991. 1991 IEEE International Joint Conference on*, pages 1458–1463. IEEE, 1991.

[68] A.L. Thomaz, M. Berlin, and C. Breazeal. An embodied computational model of social referencing. *In IEEE International Workshop on Human Robot Interaction (RO-MAN)*, 2005.

[69] P. Viola and M. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57:137–154, 2004.

[70] B. Widrow and M. E. Hoff. Adaptive switching circuits. In *IRE WESCON*, pages 96–104, New York, 1960. Convention Record.

[71] L. Wiskott. Phantom faces for face analysis. *Pattern Recognition*, 30:586–191, 1991.

[72] T. Wu, N.J. Butko, P. Ruvulo, M.S. Bartlett, and J.R. Movellan. Learning to make facial expressions. *International Conference on Development and Learning*, 0:1–6, 2009.

[73] J. Yu and B. Bhanu. Evolutionnary feature synthesis for facial expression recognition. *Pattern Recognition Letters*, 2006.

[74] X. Zhao and S. Zhang. Facial expression recognition using local binary patterns and discriminant kernel locally linear embedding. *EURASIP J. Adv. Sig. Proc.*, 2012:20, 2012.

X. Appendix: Visuo-motor learning and control of a multi-DOF robotic arm

A. Introduction

In this section, we show how the robot can learn to control its arm through a visuo-motor learning (babbling phase).

Robots have the ability to learn a task as a result of a demonstration provided by a human teacher [10]. This demonstration can be done through passive manipulation. Dynamic Motion Primitives [33] is a framework for learning to perform demonstrated tasks. In [16], the learning system uses a statistical model based on Gaussian Mixture that is adapted to fit the data from training demonstrations using passive manipulation. These systems have proved to be efficient for performing different tasks. Even if the learning by demonstration is very interesting, this technique is difficult to use in our experiment, because we want to develop an autonomous control of a multi-DoF robotic arm without human physical interaction during the learning.

In this paper, the autonomous control of a multi-DoF robotic arm requires the learning of a visuo-motor map [4]. During the "babbling phase", the robot arm produces random movements,

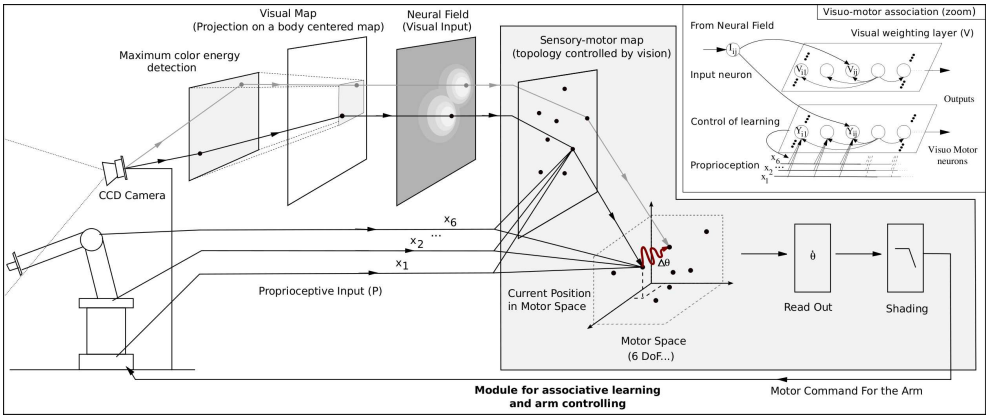

Fig. 18. Model of the arm controller (see [23]). The sensorimotor map can learn to associate visual stimulus and proprioceptive information of the arm. A competition between visuo-motor neurons enable to associate current proprioception with the most activated visual input neuron. Thus, neurons on this layer can activate one or several attractors (constructed from visuo-motor neurons) in the motor space. If the current position of the arm is different of the generated attractor, a non-null motor command is read out by Fukuyori's adapted dynamical equations and given to the robotic arm. In the social referencing experiment, this model is used to catch or to avoid objects according to the emotional interaction.

and, the visuo-motor controller can learn the correspondence between the attractors in the joint space and the visual position of the arm end-effector. After this learning, the robot arm can reach several positions in the workspace. One visual position corresponds to one or several motor configurations (e.g. attractors). These attractors pull the arm in an attraction basin (the target position). Recently, [28] has proposed a solution (Yuragi/fluctuation method) for arm control in the motor space. This model is based on the use of a limited number of attractors allowing the arm to converge reliably toward one of these motor configurations. The robot modulates the strength of the nearest attractors in the joint space allowing creating a virtual attractor in the joint space. Yuragi equation allows with a fitness signal to control the end-effector displacement so as to minimize the fitness function. The exploration allows avoiding possible local minima by creating new states when necessary and by playing with the visual associations.

Taking inspiration from this model, our working hypothesis is that proprioceptive configurations associated with the visual positions of the arm end effector can be used as attractors to achieve the visuo-motor control. The dynamical equations of the Yuragi controller allow smoothening the trajectory. The interest of this controller is the capability to control the exploration/exploitation dilemma according to a reinforcement signal. If the fitness signal increases, the strength of the attractors is increased and the noise is decreased (exploitation) and vice versa if the fitness signal decreases and the random exploration increases (see [23] for more details). In our case, the "happy face" activation is associated to a positive fitness signal for the "Yuragi" controller while the "anger face" is related to a negative fitness signal.

B. Yuragi Controler

Following Langevin equation (Refer to (8)) used to describe Brownian movements, [28] proposed that using random configurations expressed in the joint space of the arm (x is the current proprioception) combined with a noise parameter is enough to move a robotic arm toward any position by controlling the speed $\dot{x}$.

$$\tau_x \dot{x} = f(x) * A + \epsilon \tag{8}$$

$$f(x) = \sum_{i=1}^{n_a} N_i \frac{(X_i - x)}{||X_i - x||} \tag{9}$$

$$N_i = \frac{g_i(x)}{\sum_{j=1}^{n_a} g_j(x)} \tag{10}$$

$$g_i(x) = exp\{-\beta ||X_i - x||^2\} \tag{11}$$

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.

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. Modulating A of the command $\dot{x}$ or the noise level ϵ enables to switch from converging toward one of the selected motor configurations (Fig. 19.a) to exploring randomly the working space by "jumping" from an attraction basin to another (Fig. 19.b) and vice versa.

C. Results

After verifying the convergence of the arm to a learned position, we tested the convergence to a visual position spotted between four learned attractors. If the arm/target distance is lower than 3 pixels then the movement is stopped, the target is considered as reached. Fig. 20 shows the results when the robot has to reach a not learned position. The "virtual"

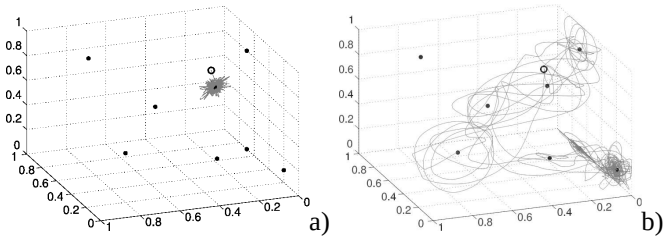

Fig. 19. Simulation of 3 DoF arm proprioception using (Refer to (8)). **a**: The trajectory converges to the nearest attractors. Simulation parameters are following: number of iterations=1000; beta (Gaussian parameter)=20; noise level ϵ_{max}=0; Number of attractors = 8; shading parameter=0.01; A=1. **b**: When the ratio of A to noise level decreases, noise has a stronger effect on the speed command and allows an exploration of the motor space with jumps from an attractor to another. Simulation parameters are following: number of iteration=about 5000 ; beta (Gaussian parameter)=20 ; noise level ϵ_{max}=1 ; A=1; shading parameter=0.01.

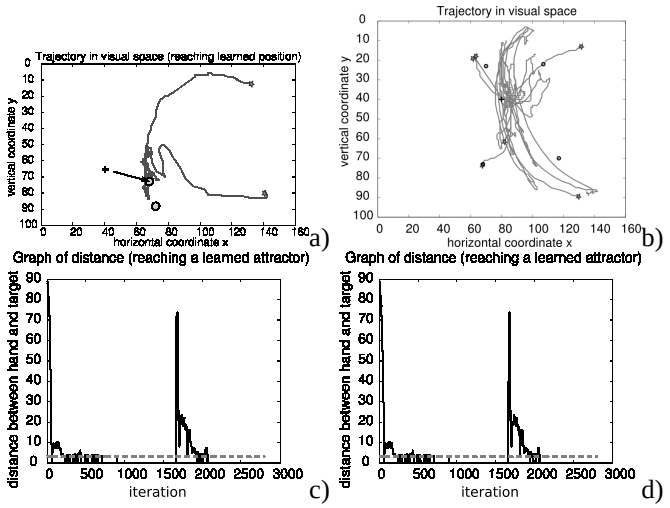

Fig. 20. Trajectory of the robot arm end effector in visual space. Experiments are made of several trials. For each trial the arm is initialized at a different position. The black circles correspond to the learned attractors and the black cross is the visual target to be reached. The stars are the starting positions for each trial. a) Reaching a learned attractor, 2 attractors activated. b) Reaching a not previously learned position, 4 attractors activated. We also record the distance between the arm end effector and the target in the visual space (number of pixels). c) Reaching a learned position, 2 trials. d) Reaching a not previously learned position, 6 trials. The light gray line shows the threshold under which the target is reached.

attractors are built as a linear combination of the real attractors. The results show the robot's capability to reach target in the robotic arm workspace. Fig. 20a) and Fig. 20b) show the trajectories of the robot arm towards a target. The arm is able to reach the target whatever the starting positions. The robotic arm succeeds in reaching a visual stimulus at arbitrary places. These results show that the cooperation of the sensory-motor map and the Yuragi method (Refer to (8)) offers an interesting basis for the control of a robotic arm with a self-learning of the associations between visual and motor spaces. This architecture (Fig. 18) has some interesting properties:

- the learning of a few attractors is sufficient to reach any position; the robotic arm reaches the target with high precision. The accuracy can be improved by recruiting a new attractor close to the target.
- the architecture can merge attractors in order to make a "virtual attractor". For example, the Fig. 20b) shows 4 attractors activated to reach a not previously learned position.
- the trajectories of the robotic arm are curvilinear which involve smooth movements of the robotic arm (Fig. 20c and Fig. 20d).

At this point, the robot can reach a neutral object in its workspace as the result of the cooperation of sensorimotor map and the Yuragi method.

Sofiane Boucenna is postdoctorant at Pierre et Marie Curie University in the Institut des Systémes Intelligents et de Robotique lab (ISIR). He obtained its PhD at the Cergy Pontoise University in France in 2011, where he worked with the Neurocybernetic team of the Image and Signal processing Lab (ETIS). His research interests are focused on the modelling of cognitive mechanisms and the development of interaction capabilities such as imitation, emotion and social referencing. Currently, he attempts to assess the effect of the type of partners (adults, typically developing children and children with autism spectrum disorder) on robot learning.

Philippe Gaussier is Professor at the Cergy Pontoise University in France and leads the neurocybernetic team of the Image and Signal processing Lab. He is also member of the Institut Universitaire de France. Currently his research interests are focused on the modelling of cognitive mechanisms and brain structures such as the hippocampus and its relations with cortical structures like parietal, temporal and prefrontal areas, the dynamics of visual perception,the development of interaction capabilities (imitation, emotions...). Current robotic applications include autonomous and online learning for motivated visual navigation, object manipulation, emotion, social referencing.

Laurence Hafemeister is associate professor at the Cergy Pontoise University in France and works in the neurocybernetic team of the Image and Signal processing Lab. She received a Ph.D. degree in computer science from the Universty of Paris XI (1994). Currently, his research interests are focused on the visual attention, the perception and the development of interaction capabilities.