Noname manuscript No.
(will be inserted by the editor)

A robot learns the facial expressions recognition and face/non-face discrimination through an imitation game

Sofiane Boucenna · Philippe Gaussier · Pierre Andry · Laurence Hafemeister

Received: date / Accepted: date

Abstract In this paper, we show that a robotic system can learn online to recognize facial expressions without having a teaching signal associating a facial expression with a given abstract label (e.g., 'sadness', 'happiness'). Moreover, we show that recognizing a face from a non-face can be accomplished autonomously if we imagine that learning to recognize a face occurs after learning to recognize a facial expression, and not the opposite, as it is classically considered. In these experiments, the robot is considered as a baby because we want to understand how the baby can develop some abilities autonomously. We model, test and analyze cognitive abilities through robotic experiments. Our starting point was a mathematical model showing that, if the baby uses a sensory motor architecture for the recognition of a facial expression, then the parents must imitate the baby's facial expression to allow the online learning. Here, a first series of robotic experiments shows that a simple neural network (N.N.) model can control a robot head and can learn online to recognize the facial expressions of the human partner if he/she imitates the robot's prototypical facial expressions (the system is not using a model of the face nor a framing system). A second architecture using the rhythm of the interaction first allows a robust learning of the facial expressions without face tracking and next performs the learning involved in face recognition. Our more striking conclusion is that, for infants, learning to recognize a face could be more complex than recognizing a facial expression. Consequently, we emphasize the importance of the emotional resonance as a mechanism to ensure the dynamical coupling between individuals, allowing the learning of increasingly complex tasks.

ETIS, CNRS UMR 8051, ENSEA, Cergy-Pontoise University
{sofiane.boucenna}@gmail.com

Keywords Human-Robot Interaction · Emotional Interaction · Imitation · Sensory-motor Architecture

1 Introduction

Human/Robot interactions are a complex issue that challenges both engineering and cognitive sciences. The recognition of human emotional states can provide important information either as a direct stimulus (stop or continue a given behavior) or as a way to bias human-machine interactions (avoiding specific topics, attempting to understand why the human partner is in a negative mood or, in contrast, detecting the conditions inducing a positive mood in the human partner so that they can be triggered at the right time). Most of the proposed architectures use strategies that are specific to the stimulus to be analyzed. For example, model-based or offline learning techniques are used first to locate a face in an image and next to analyze the face to find emotional signatures. The same scenario is conducted for other learning tasks, for example, those that involve gestures, gaze direction, and vocalizations.

We believe robotic behaviors implying rich interactions must be investigated from a developmental perspective to avoid the symbol grounding problem[1] [22] and to provide better adaptation capabilities in unforeseen situations. Our study focuses on the cross fertilization between developmental psychology and developmental robotics. A new understanding of how human cognitive functions develop is sought using a synthetic approach and robotic devices [3, 34].

[1] The symbol grounding problem is related to the problem of how symbols (words) get their meanings (without a human expert providing the knowledge)

Our starting point was motivated by the question of how a "naive" system can learn to respond correctly to another person's expressions during a natural interaction. "Natural" here means that the interaction should be as unconstrained as possible, without an explicit reward or ad-hoc detection mechanism or a formatted teaching technique. The baby-mother interaction provides a good framework for addressing this question. A newborn has a set of expressions that are linked with his/her own internal state, for example crying and displaying a sad face when he/she needs food or expressing happiness/pleasure after being fed. Yet, the link with the expressions of others must be built. How does the link between his/her own emotions and the expression of others emerge from non-verbal interactions? The problem here is to understand how babies learn to recognize facial expressions without having an explicit teaching signal that allows the association, for example, between the vision of "happy face" and their own internal emotional state of happiness [17]. The development of social referencing allows to propose a persuasive scenario [5]. In this context, we suppose the existence of a very simple reflex pathway allowing the simulation of pain and pleasure from an ad hoc tactile sensor (e.g. conductive[2] objects). This signal allows the association of objects with positive or negative values, but also, to express the facial expression corresponding to this value. Consequently, if the robot produces a facial expression according to a reflex pathway and if a caregiver is present and imitates the robot's face, then, the robot can link this reflex pathway (internal emotional state) with the vision of a "happy face". Psychological experiments [41] have shown that humans "reproduce" involuntarily the facial expression of our robot face. This low-level resonance with the facial expression of another person could be a bootstrap for robot learning ("sympathy" and perhaps not "empathy" for the robot head) similar to what has been proposed for humans [48,8]. To fasten the learning, in our protocol, we ask the caregiver (non-professional actors) to imitate the robot's face. A minimal robotic set-up is used to learn the task (Fig. 4), i.e., the robotics head does not attempt to represent perfectly a human face (no skin, a minimal number of degrees of freedom), was important, first, to be sure to avoid problems linked to the uncanny valley [36] and, next, to test what are the truly important features for the recognition of a given facial expression. Our expressive head was developed both as a tool for man-machine interfaces and for psychological studies [40] with infants and adults.

In the framework of interactive learning based on imitation games, we will attempt to show that online incremental learning is easy and could be used to bootstrap more and more complex learning, such as learning the gaze direction or providing a social referencing for objects and places [6,23]. Our starting point is that, during natural interactions, the parents provide unintended social feedback to the baby by allowing a fast online self-supervised learning of more and more complex social signals. In the next section, we will summarize first our formal model for the online learning of facial expression recognition. We show that a simple sensory-motor architecture based on a classical conditioning paradigm learns online to recognize facial expressions if and only if the robot produces facial expressions and that the caregiver imitates the facial expression of the robot. Next, the implementation of this theoretical model without a face detection module will be presented. The constraints from the online learning will be discussed. Using a reinforcement signal built from the interaction rhythm prediction, we will show how our robot can use the recognition of facial expressions to learn to discriminate what a face is.

In conclusion, we will discuss the possibility of using the presented mechanism as a general bootstrapping strategy for the autonomous learning of complex tasks in a social context (see [8]). The problem of the need to build an ad hoc number of recognition systems to establish long-term man-machine interactions will then be replaced by the life-long learning of more and more complex, interactions avoiding the symbol grounding problem [22].

2 Related work

Many solutions have been proposed for the problem of face localization and facial expression recognition. The emphasis in these studies has been on the quality and performance of the algorithms. These two topics have always been linked. The general process of facial expression recognition follows several steps. The first step concerns face registration (face detection and landmarks localization). Many authors, such as Viola&Jones [50], Littlewort [32] or Rowley [45], have developed models for robust face detection. They use some a priori analysis of the face (e.g., eyes, mouth, and nose) and/or involved offline and supervised learning. The second concerns the image coding and classification (facial expressions recognition). Solutions for the recognition of facial expressions usually use these algorithms to frame the image around the face before performing the expression recognition. An overview of facial expression recognition can be found in [55]. Some methods are

[2] 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.

based on Principal Components Analysis (PCA) and use a batch approach. For example, the LLE (Locally Linear Embedding) [31] performs a dimension reduction on the input vectors. Neuronal methods have also been developed for facial expression recognition. In Franco and Treves [12], the network uses a multi-layer network with a classical supervised learning rule (again an 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 [1]). Yu [54] uses a support vector machine (SVM) to categorize the facial expressions. Others studies as [47] propose to combine different types of features to recognize the facial expressions and action units, and, the recognition is performed through a multi-kernel SVM. Wiskott [52] 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. Moreover, many studies focus on continous emotions [21,42] (assigning an intensity level of recognition) and the recognition of action units [47].

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 performance 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 [7] 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 certain specific knowledge (what is a human face) to learn the facial expressions. [7] 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. An other study [30] shows how people's social response toward a humanoid robot can change when we vary the number of the active degrees of freedom in the robot's head and face area. The study of Lee [30] is interesting because it questions about the design and the acceptability of the robot. Other studies using robot heads, such as Einstein's robot [53], 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.

3 Theoritical model for Online learning of facial expression recognition: an interactive model

The theoretical model considers a single system that is composed of the two agents interacting in a neutral environment (Fig. 1).

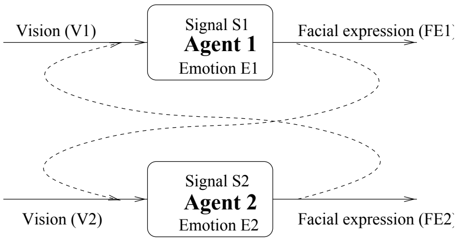

Fig. 1 The bidirectional dynamical system studied. Both of the agents face each other. Agent 1 will be considered to be a newborn and agent 2 will be considered to be an adult mimicking the newborn facial expressions. Both of the agents are driven by internal signals that can induce the feeling of specific emotions.

One agent is assumed to be an adult with perfect emotion recognition and reproduction capabilities. The second agent is considered to be a newborn without any previous learning of the social role of emotions. Formally, the 'baby' agent is described as a conditioning system (Fig. 2).

We assume that both agents receive a visual signal V_i with $i \in \{1, 2\}$. V_i can be learned and recognized in the VF_i group, VF_i being the result of, for example, unsupervised pattern matching strategy, such as a winner takes all (WTA)[46] or an adaptative resonance theory (ART) network[20] or a Kohonen map[28] or a real time kmeans algorithm. Hence, the vision of a face displaying a particular expression should trigger the activation

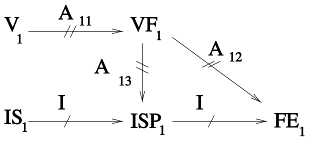

Fig. 2 Schematic representation of an agent that can display and recognize facial expressions. Arrows with one stroke represent "one-to-one" reflex connections. Arrows with labels and 2 parallel strokes represent "one-to-all" modifiable connections. V: visual stimulus, VF: local view recognition (visual features), IS: internal state, ISP: "emotional" state (internal state prediction), FE: facial expression (motor command). The synaptic weights: A corresponds to links one to all modifiable (pathway learned) and I corresponds to links one to one non-modifiable (reflex pathway)

of the corresponding node in VF_i:

$$VF_i = c(A_{i1}.V_i) \quad (1)$$

where c is a competitive mechanism. A_{i1} represents the weights of the neurons in the recognition group of agent i, allowing for direct pattern matching. When a new local view is learned (Visual Feature: VF), $VF_i = (0, 0, ..., 1, 0, 0, ..., 0)$, with a single 1 for the new winner. IS_i is the internal state, which represents the physiological inputs that are related to emotions, including fear and anger. The recognition of a specific internal state will be called an emotional state ISP_i (internal state prediction). We assume also that ISP_i depends on the visual recognition VF_i (visual features) of the visual signal V_i. Last, the agents can produce a facial expression FE_i. If one agent acts as an adult, then it must have the ability to recognize the facial expression of someone else's face (sympathy or empathy[3]). At least, one connection between the visual features and the group representing its emotional state must exist. To display an emotional state, we must also assume that there is a connection between the internal signals and the triggering of the facial expression. For sake of homogeneity, we will assume that the internal signals activate, through unconditional links I (reflex pathway), the emotion recognition group, which activates, through an other reflex pathway I (unconditional connections), the display of a facial expression (which is equivalent to a direct activation of FE_i by IS_i - see [13] for a formal analysis of this type of property). Hence, the sum of both flows of information is the following:

$$ISP_i = c\,(I.IS_i + A_{13}.VF_i) \quad (2)$$

Last, we can also assume that the teaching agent can display a facial expression without "feeling" it (just by a mimicking behavior obtained from the recognition of the other facial expressions). The motor output of the teacher's facial expression then depends on both the facial expression recognition and the will to express a specific internal state (here, to simplify a discrete emotional state):

$$FE_i = c\,(I.ISP_i + A_{12}.VF_i) \quad (3)$$

If we suppose both agents use the conditioning architecture presented in Fig. 2, then we have shown in [16] that in the symetric interaction proposed in Fig. 1. The network becomes equivalent to Fig. 3a only if the second agent acts as a mirror.

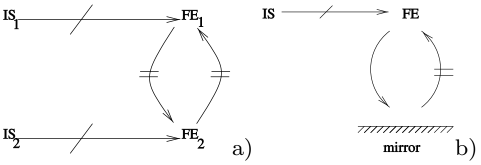

Fig. 3 a) Final simplification of the network, representing the interaction between the 2 identical emotional agents. b) The minimal architecture that allows the agent to learn the "internal state" - "facial expression" associations.

Fig. 3 shows also that if both agents display their internal emotional states IS_1 and IS_2, then the learning is impossible if both agents have independent emotional states (there is no correlation between IS_1 and IS_2). In this case, the learning cannot stabilize. If we assume that there is no way to control the internal state of the baby agent, then the only solution is that the second agent mimics or resonates [41] to the facial expressions of the baby agent, thus allowing for an explicit correlation (the parent is no more than a mirror - see fig. 3). If this condition is verified, then the system can learn; agent 1 (baby) can learn to associate the visual recognition of the 'parent' facial expressions with its own internal state (ISP_1). The agent can learn how to connect the felt but unseen movements of self with the observed but unfelt movements of the other. A surprising conclusion is that the teacher has to imitate the learner agent.

4 Set-up experimental

4.1 Material

To test this theorical model, we propose in this paper to build a robotic head based on the network presented in fig. 2. Our robot head has 2 eyes which use classical

[3] To be precise, this scenario can involve only low-level resonance and perhaps sympathy because, according to Decety & Meyer, empathy is "a sense of similarity in the feelings experienced by the self and the other, without confusion between the two individuals" [8].

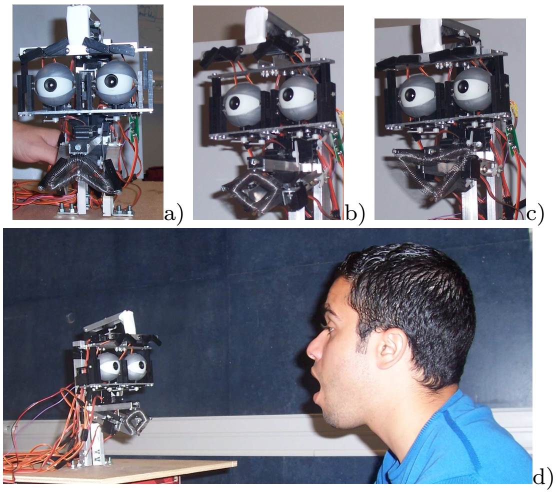

Fig. 4 Examples of robot facial expressions: a) sadness, b) surprise, c) happiness. d) Example of a typical human / robot interaction game (here, the human imitates the robot).

PAL cameras[4] (a single camera is used here). Two eyebrows and one mouth were added to provide minimal facial expression capabilities. A total of 13 servomotors are used to control the motion of the different mobile parts of the face. A mini SSC3 servomotor controller card allows the generation of PWM (pulse width modulation) signals to move and maintain the servo motors in a given position. 4 motors control the eyebrows (bending), 1 motor controls the forehead (to move up and down) and 5 motors control the mouth (opening and bending). Finally, 3 motors control the orientation of the 2 cameras that are located in the robot "eyes" : 1 motor controls the vertical plane (pan movement) and 2 motors control the horizontal plane (1 servo for each camera and an independent tilt movement). The low-level software that controls the robot head can reproduce prototypical facial expressions. All servo motors move in parallel to attend the angular position given by the controller. This process induces a dynamical and homogeneous movement in which all of the parts of the face change to form a given expression (which is programmed as a list of the final positions for each joint). Because of the servomotor dynamics, the robot head can produce a high number of facial expressions. Depending on the distance in the joint space between two specific facial expressions, one change in the facial expression is achieved in approximately 200 to 400 ms.

Using a real device instead of a virtual face causes several difficulties but induces a visible "pleasure" linked to the physical "presence" of the robot for the human partner. The robot head is also very interesting because the control of the gaze direction (pan/tilt camera) can be used both as an active vision system and as a communication tool. In other works, the head has been placed on a mobile platform, to be used for visual navigation [18] and object manipulation [6] and, of course, to display the robot's internal state [23].

4.2 Protocol

In the following, the robot will be considered to be a baby and the human partner will be considered to be a parent (the father or mother). At first, the robot knows almost nothing about the environment. The robot learns through the interaction with the human partner. The two fundamental assumptions in this study are: the existence of low level resonance (the humans mimic the robot head) and the presence of a reflex pathway connecting the internal emotional state to the facial expression.

To test this model, a neural network architecture was developed, and the following experimental protocol was adopted: First, we ask human partners to sit in front of the robot head (the distance between the two partners is around 1 meter) and we ask human to imitate the robotic head. In the first phase of the interaction, the robot produces random facial expressions (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 subject mimics the robot head. This first phase lasts between 2 and 3 min according to the subject's "patience". At the end, the generator of the random emotional states is stopped. If the N.N. has learned correctly, then the robot must mimic the facial expression of the human partner.

The tests are limited to 4 prototypical facial expressions: happiness, sadness, anger and surprise [25, 11, 10, 44], plus a neutral face (Fig. 4 for the experimental setup). Each of the four facial expressions has been controlled by FACS experts [10]. The validity of this choice could be discussed (especially for the surprise and/or for the choice of the expression names) [26]. However, for our purpose, we need only a small set of facial expressions that are easily recognized and that induce a low level resonance from the human partner.

5 Facial expression recognition and imitation

Our initial approach followed classical algorithms: first, face localization was performed, which used, for ex-

[4] The standard PAL camera provides a 720x580 color image used only for grey levels.

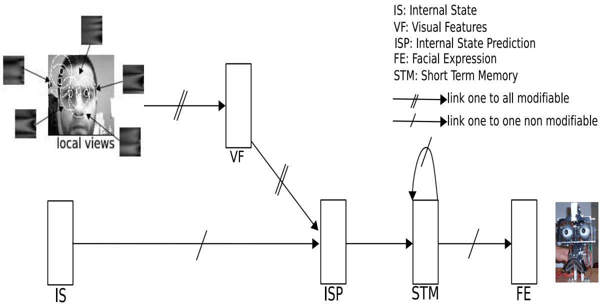

Fig. 5 The architecture for facial expression recognition and imitation. The visual processing allows for the sequential extraction of the local views. The *internal state prediction* (*ISP* group) learns the association between the local views and the internal state (*IS* group).

ample, [24] or [50]; then, face framing was performed, and third, facial expression recognition of the normalized image was performed. In this case, the quality of the results was highly dependent on the accuracy of the face framing (the generalization capability of the N.N. can be affected). Moreover, if we use this approach, the robot head cannot be truly autonomous because we were unable to find a way to perform first (or alone) face/non-face discrimination. An online learning of face/non-face recognition cannot directly use our mimicking framework because the human (and our robotic head) does not have an internal signal to trigger a specific face/non-face reaction of the human partner, while it is natural for the human to mimic the facial expressions.

Our solution consists of skipping the framing step to directly use all of the most activated focus points in the image, and, we show the least mean square equation used in our conditioning architecture is sufficient to discriminate contingent stimuli from background information (non contingent).

The system performs a sequential exploration of the focus points in the image (see appendix 9.1) and attempts to condition them to a given facial expression. The network shown in fig. 5 is directly derived from the theoretical model presented in the previous section. The extracted local view around each focus point is learned and categorized by a group of neurons VF (visual features) using a k-means variant that allows online learning and real-time functions [27] called SAW (Self Adaptive Winner takes all):

$$VF_j = net_j.H_{max(\gamma,\overline{net}+\sigma_{net})}(net_j) \qquad (4)$$

$$net_j = 1 - \frac{1}{N}\sum_{i=1}^{N}|W_{ij} - I_i| \qquad (5)$$

VF_j is the activity of neuron j in the group VF. $H_\theta(x)$ is the Heaviside function [5]. 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 method allows the recruitment to adapt to the dynamics of the input and to reduce the importance of the choice of γ. Hence, 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. It allows to learn quickly and have prototypes averaged, this mechanism is essential for an online learning. The modification of the weights is computed as follows:

$$\varDelta W_{ij} = {\delta_j}^k \cdot (a_j(t)I_i + \epsilon(I_i - W_{ij})(1 - VF_j)) \qquad (6)$$

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 [6], 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, $\varepsilon(I_i - W_{ij})(1 - VF_j)$, 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. 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 VF group learn to average the prototypes of the facial features (for example, a mean curved lip for a happy face).

Of course, there is no constraint on the selection of the local views (no framing mechanism). This scenario means that many distractors can be present (there are local views in the background or inexpressive parts of the head). This scenario also means that any of these distractors can be learned on VF. The figure/ground discrimination can be learned because the local views in the background are not statistically correlated with a given facial expression. In a face-to-face interaction, the

[5] Heaviside function:

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

[6] Kronecker function:

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

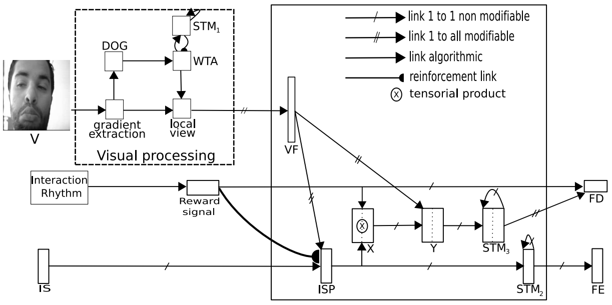

Fig. 6 The global architecture to recognize facial expressions, to imitate and to recognize face from non-face stimuli. Visual processing allows the extraction of sequential local views. The VF group (local view recognition) learns the local views (each group of neurons, IS, ISP, STM_2 and FE, contains 5 neurons that correspond to the 4 facial expressions plus the neutral face). A tensorial product is performed between ISP (emotional state: internal state prediction) and a reward signal, to select the neuron that must learn. Y learns the correlation between a local view and a facial expression on a specific neuron that is activated depending on whether a reward linked to the interaction has been obtained or not. The FD group (Face detection) learns the correlation between the tensorial product and the reward signal (its activity corresponds to the recognition of a face).

distractors are present for all of the facial expressions, and their correlation with an emotional state will tend to zero. Only the local views on the face (Fig. 6) correlated with a given robot expression will be reinforced. The use of the Widrow and Hoff rule (derived from a least mean square (LMS) optimization) will learn correctly if, during the period that is allowed for the exploration of one image, enough focus points can be found on the face. In our network, The Internal State Prediction ISP (or "emotional" state) associates the activity of the visual features VF with the current internal state IS of the robot (a simple conditioning mechanism using the Least Mean Square (LMS) rule [51]):

$$\Delta w_{ij} = \epsilon.VF_i.(IS_j - ISP_j) \tag{7}$$

STM_2 is Short Term Memory used to sum and filter over a short period (N iterations), and the emotional states $ISP_i(t)$ associated with each explored local view are as follows:

$$STM_{2,i}(t+1) = \frac{1}{N} \cdot ISP_i(t+1) + \frac{N-1}{N} \cdot STM_{2,i}(t) \tag{8}$$

Here, i is the index of the neurons; for example, ISP_i corresponds to the i^{th} "emotional" state ($0 < i \leq 5$).

Arbitrarily, a limited amount of time is fixed for the visual exploration of one image, to obtain a global frequency of approximately 10 Hz (100 ms per image). The system can analyze up to 10 local views on each image. This is a small number of views; however, because the system usually succeeds in taking 3 to 4 relevant points on the face (e.g., on the mouth, eyebrows), it is sufficient in most cases and it allows real-time interactions to be maintained.

The control of the robot facial expression is performed via the FE group. The highest activity of the FE_i triggers the i^{th} facial expression because of a WTA. To increase the robustness, FE also uses a short-term memory to give more importance to the present than to the past. This scenario ensures high performances when the correct action is selected more than 50% of the time during the chosen time window. This feature is important for online interactions, and it allows for a reduction of the constraints on the intrinsic recognition quality if the frame rate is sufficiently high.

6 Experimental results during the interaction

This section presents the experimental results obtained during the interaction with the robot. Before going into details, we present the 3 different setups and databases recorded to analyze the different aspects of the systems and its learning capabilities. They correspond to three different experimental conditions where the procedures and the number of participants vary: The first database is composed of the images of 10 persons interacting with the robot head (1600 images) during a "natural" interaction. During the learning phase, the partners imitate the robot, and then the robot imitates them. Each image is annotated with the robot facial expression during the online learning[7], enabling the statistical analyses to be performed offline. The behavioral performances show the robot capability to recognize the facial expressions of participants who interacted with the robot during the learning phase. This database will be used in the section 6.2. The second and third setup allow focusing on the robot generalization capabilities to imitate new persons who were not observed during the learning phase. The databases are composed of 20 persons who interacted with the robot head (3200 images). Two tests have been performed to study: in G1 the capability of the robot to generalize when the robot learns with 20 human partners (the statistics were obtained with the twenty-fold cross validation methodology); and in G2 the success rate of the facial expression recognition as a function of the number of faces that the system learned during the learning phase (this database will be used in the section 6.3).

[7] Since the human partner is supposed to be imitating the robot head, the label of the robot facial expression should be the correct label of the image. We will see later it is not always the case because of the human reaction time and because of some misrecognition from the human partners.

6.1 Constraint caused by the online learning

The online learning can involve specific problems regarding real-time learning because the human reaction time to the robot facial expressions is not immediate (Fig. 7). First, [49] has shown the minimal duration for a human to differentiate an image with an animal from another image is about 150 ms. The minimal duration to recognize a facial expression for a human seems quite similar to these 150 ms. In our case, we consider the minimal period T of an interaction loop is the sum of t_1, the delay for the robot to perform a facial expression, plus t_2, the delay for the human to recognize the facial expression plus t_3, the delay for the human subject to mimic the recognized expression ($T = t_1 + t_2 + t_3$). When the robot is only an automata producing arbitrary facial expressions, we measured a minimal period T of approximately 800 ms for expert subjects (a person knowing the robot) and 1.6 s for novice subjects (a person who never interacted with the robot) to produce a facial expression matching the robot head expression. This time lag can greatly disturb the learning process: the first images available for a given expression are always associated with the human's previous facial expression. To avoid the unlearning of the previous expression using an LMS algorithm, the presentation time of each expression must be long enough (in our case 3 seconds) to be sure there are more correct matching pairs of robot/human expressions than incorrect one related to the transition between one expression and the following.

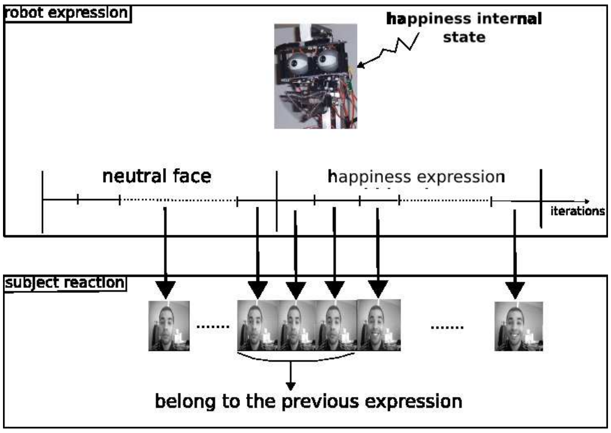

Fig. 7 Phase shift between the human facial expression and the robot facial expression during an imitation game (the human imitating the robot).

6.2 Behavioral performances

After learning, the associations between the visual features VF and the internal state prediction ISP are strong enough to bypass the low-level reflex activity that comes from the internal state IS (see section 3). In this case, the facial expression FE will result from the temporal integration of the emotional state associated with the different visual features analyzed by the system (features will have an emotional value if they are correlated with the robot facial expression, basically the expressive features of the human head).

First, the low part of Fig. 8 shows the neural activity, and facial expressions recognition during a natural interaction. The upper part of Fig. 8 shows that the robot recognizes and imitates correctly the first two facial expressions of the human partner (happiness and anger expressions) but makes an error when the human displays a neutral face (the robot recognizes the happiness expression). Second, we study the labels associated with the different local views after learning. Fig. 9 shows the focus points associated with each facial expression, i.e., these focus points vote for the recognition of a given facial expression.

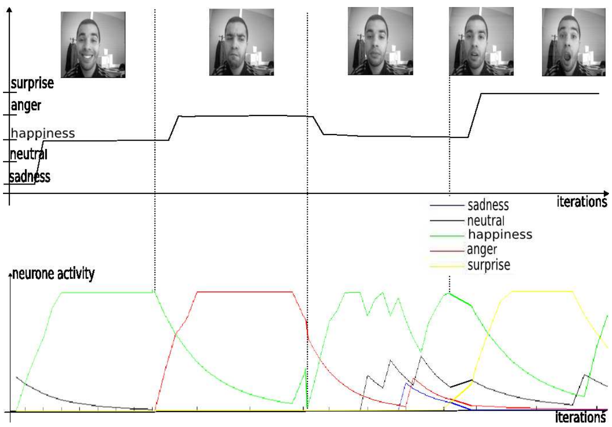

Fig. 8 Temporal activity of the neurons associated with the triggering of the different facial expressions when the robot imitates the human (after learning).

The focus points are usually taken in the areas of the face where the features are useful to recognize facial expressions (mouth, eyebrows etc). Moreover, some facial expressions are characterized by a specific set of focus points corresponding to local areas on the face relevant for the recognition of a specific expression. For example, the local views corresponding to the corner of the mouth are used to characterize the "happiness" while the local views around the eyebrows can characterize the "anger" and "sadness" or the local views taken at the center of the mouth characterize the ”surprised” expression.

The different results show our low level visual processing involving a gradient extraction and a focus on the maxima of the convolution with a DOG filter are sufficient in most part of the cases to allow the focus on the useful areas of the face in order to discriminate each expression[8]

Fig. 10 shows that the interaction with the robot head for a period of 2 min can be sufficient for the robot to learn the facial expressions and next to imitate the human partner. The Y axis (success rate) corresponds to the rate of correct facial expression recognition by the robot. Fig. 10 shows the robot capability to recognize the facial expressions of participants who interacted with the robot during the learning phase. In this experiment (Fig. 10), the success rate of the facial expressions recognition is computed according to the number of participants that the robot meets during the learning phase, i.e. the persons already learned during the learning phase. For example, if four persons have interacted with the robot during the learning phase, then, the validation is limited to these four persons. The results on 10 persons interacting each during 2 min with the robot show our incremental learning is robust, although the variability of the expressiveness of the different subjects (for example, sadness was expressed in different ways from one person to another).

6.3 Generalization results

In this subsection, we want to show that the robot is able to generalize on new persons. 20 persons interacted with the robot head during the learning phase. During this period, the 20 caregivers imitated the robot and the robot analyzed the images, learned the task and recorded all of the images. The database was created to perform offline processing and analysis. Each image was annotated with the response of the robot during the online learning. The results show that when learning is over, the robot can imitate new persons who were not observed during the learning phase. The statistics were obtained with the twenty-fold cross validation methodology by using the database created online.

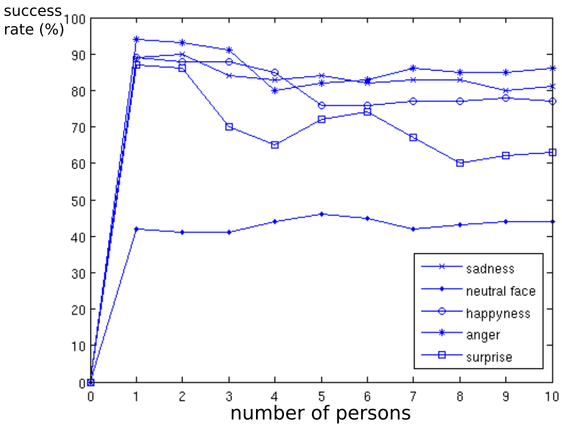

Fig. 10 The success rate for each facial expression (sadness, neutral face, happiness, anger, and surprise) when using a log-polar transform for the local view recognition. These results are obtained during the natural interaction with the robot head. A total of 10 persons interacted with the robot head (32 images by facial expression per person: 32x5x10 images). During the learning phase (only a 2-minute period), 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 by other human volunteers, and the correct correspondence was checked.

The Table 1 shows the generalization test of the system obtained during a ”natural” interaction with the robot. These tables show the confusion rate for each facial expression using three different visual features extracted around each focus point (log-rho-theta image, signature of Gabor filters or the fusion between both features). The success rates with the log-polar local views are 62% for happiness, 52% for anger, and only 27% for sadness. Using the Gabor filters, the success rate is 70% for happiness, 80% for anger and 4% for sadness. These results show the good generalization capability of Gabor filters that have to be combined to the log-polar views to avoid overgeneralization. The Table 1c) underlines that the average recognition rate is better with the fusion between both visual features except for the sadness expression. A possible explanation for the sadness result is that people have difficulty displaying sadness without a real social context. Each partner imitating the robot displays sadness in a different way (see Fig. 11). To investigate this hypothesis, we asked 7 non-expert humans to annotate our database (Table 2). The results show the confusion rate for each facial expression. We observe that the non-expert humans also have difficulty in recognizing some of the facial expressions, especially sadness.

The detailed analysis of the Table 1 corresponding to confusion rate for each facial expression, shows that

[8] Problems can occur with pale faces and dark airs since they can induce a lot of focus points with a high value around the face reducing the probability to select focus points really on the face. One solution is to increase the number of selected focus points to be sure to take focus points on the face but then the exploration time increases and the frame rate is reduced. In future works, we plan to use this simple solution as a bootstrap mechanism for a spatial attentional mechanism to focus on the face in order next to come back to a fast image analysis in the selected area.

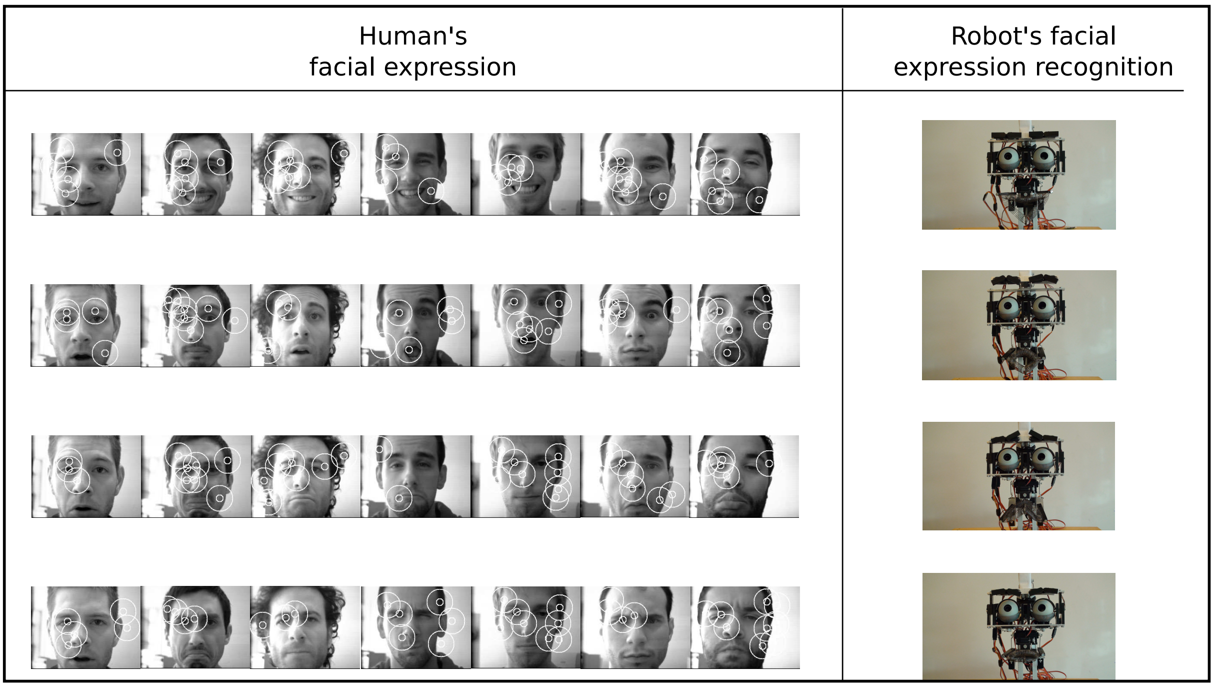

Fig. 9 Study of the labels associated with the different local views after learning. The displayed focus points correspond to the facial expressions that were associated with happiness, surprise, sadness and anger.

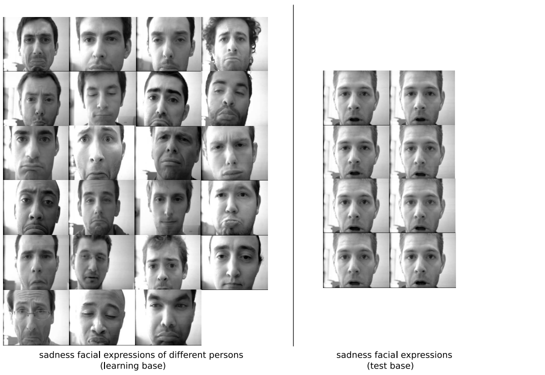
sadness facial expressions of different persons
(learning base)
sadness facial expressions
(test base)

Fig. 11 Several examples of the sadness expression show that sadness is expressed differently from one person to another.

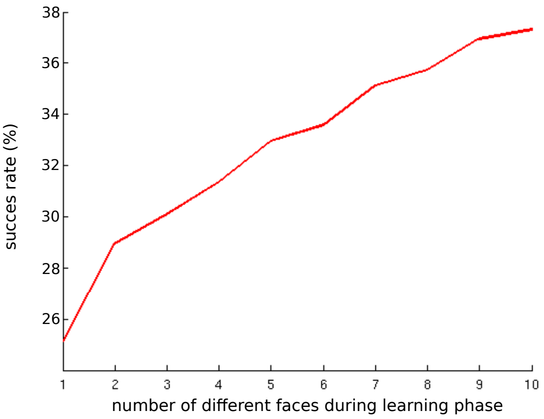

Fig. 12 Measure of generalization capabilities averaged over our 4 facial expressions plus the neutral face. This figure shows the success rate (y axes) of the facial expression recognition as a function of the number of faces (x axes) that the system learned during the learning phase. The measure was performed on a database of 10 unlearned subjects (1600 images that were never learned). The generalization performance improves after interacting with an increasing number of people.

the robot confuses the sadness with the anger. We observed the images of sadness and anger and we underlined that the eyebrows are furrowed for these two facial expressions. Table 1 show also that the robot confuses the surprised with the neutral face. Analyzing the image database, we note that these two expressions are very similar for some human partners. In these two cases, it is difficult for the robot to discriminate robustly these expressions.

Moreover, the success rate of the facial expressions recognition is a function of the number of faces that the robot learns during the learning phase. To test further the generalization capabilities of the N.N. during an incremental learning, the N.N. performances has been measured on 10 other people (10 unlearned subjects). The success rate improves after interacting with an increasing number of people (Fig. 12). The success rate is approximately 38% when 10 subjects interacted with

		robot's facial expressions				
		sadness	neutral face	happiness	anger	surprised
human's facial expressions	sadness	**27%**	18%	23%	24%	8%
	neutral face	27%	**26%**	23%	19%	5%
	happiness	18%	9%	**62%**	9%	2%
	anger	23%	11%	10%	**52%**	4%
	surprised	26%	12%	13%	14%	**35%**

a)

		robot's facial expressions				
		sadness	neutral face	happiness	anger	surprised
human's facial expressions	sadness	4%	18%	20%	**38%**	21%
	neutral face	7%	16%	27%	**36%**	15%
	happiness	0.08%	9%	**70%**	18%	4%
	anger	0%	0.08%	14%	**80%**	2%
	surprised	6%	14%	23%	13%	**44%**

b)

		robot's facial expressions				
		sadness	neutral face	happiness	anger	surprised
human's facial expressions	sadness	4%	32%	3%	**43%**	17%
	neutral face	7%	**56%**	6%	26%	5%
	happiness	4%	17%	**65%**	14%	0%
	anger	3%	11%	10%	**73%**	3%
	surprised	5%	33%	6%	9%	**47%**

c)

Table 1 Generalization test of the system obtained during a "natural" interaction with the robot: After 20 persons interacted with the robot head (during the learning phase), the robot had to imitate new persons who were not observed during the learning phase. These tables show the success rate and the confusion rate for each facial expression (sadness, happiness, anger, surprise and neutral face). a) shows the confusion rate when using the log-rho-theta features, and b)shows the confusion rate when using the Gabor filter features, and c) shows the confusion rate when using the fusion between both features. These statistics were obtained with the twenty-fold cross validation methodology.

		facial expressions annotated by no-expert humans				
		sadness	neutral face	happiness	anger	surprised
human's facial expressions	sadness	**45%** ± 2.2	10% ±1.1	0% ± 0	19% ±0.55	26% ±1.6
	neutral face	4% ± 0	**89%** ±0.36	3% ±0.13	1% ± 0	3% ±0.13
	happiness	0% ± 0	6% ±0.10	**90%** ±0.43	1% ± 0	3% ±0.19
	anger	9% ±0.24	15% ±3.86	10% ± 0	**46%** ±2.46	20% ±1.43
	surprised	2% ±0.12	12% ±0.78	5% ± 0	6% ±0.19	**75%** ±0.64

Table 2 Performance of 7 non-expert subjects when attempting to annotate our database of other human subjects imitating the robot head (forced response choice). This table shows the confusion rate and the standard deviation for each facial expression. The non-expert subjects have difficulty in recognizing the facial expression produced by human partners.

the robotic head during the learning phase (Fig. 12), and the table 1 shows that the success rate is approximately 50% when the robotic head learned with 20 subjects. Nevertheless, the results show that the N.N. can generalize to people who were not present during the learning phase.

In this section, we have shown that the robot head can learn and recognize autonomously facial expressions if, during the learning phase, the robot head performs facial expressions and if the human partners mimic them. The different results show that our architecture can recognize, with some success, the facial expressions without any exterior supervision or any face detection (or even a precise model of the face). The different results also show that the integration of both log-polar transform and Gabor filters as visual features provides more robust results. The model only learns to associate the recognition of local views with the robot's own facial expression through the human mimicking behavior.

However, the robot could learn non-pertinent associations if there is not a human partner. As a result, the learning is still not completely autonomous. Learning should not be turned on and off by the researches according to the presence or absence of the human partner. In the next section, we will focus on solving this issue by adding a mechanism for modulating the learning and allowing a face/non-face discrimination.

7 Face/non-face recognition from facial expression recognition

To perform autonomous learning, we introduced the capability of predicting the rhythm of the interaction [2] to avoid learning when there is no human subject in front of the robot or when the human is not paying attention of the robot (for example, when the human partner is leaving or talking with someone else). 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 [39,38,9]. A social interaction rupture 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 will be used as a reward signal (see [2] for the application of the same principle to the learning of an arbitrary set of sensory-motor rules):

- A rhythmic interaction is equivalent to a positive reward: The robot head and the subject produce a coherent action at each instant.
- Conversely, an interaction rupture is interpreted as a negative reward.

When a subject displays a facial expression, he/she performs whole face or body motions. If the subject imitates the robot, then his/her movement peaks have a frequency that depends on the frequency of changes in the robot face (in our case, this frequency is constant because the robot facial expression changes every 4s). The interaction rhythm can be predicted either by using a prediction of the timing between 2 visual peaks (a stable frequency of interaction of the human partner) or using the prediction of the delay between the triggering of the robot facial expression and the motions perceived by its CCD cameras (a reaction of the human partner to the robot expression). A measure of the prediction error can easily be built from the difference in activity between the predicted signal and the non-specific signals. These non-specific signals (motions) are related to the presence or absence of a human partner. If the error is important, then there is a novelty (the subject is not in the rhythm). Otherwise, the prediction error is small, which involves a good interaction between the subject and the robot (Fig. 13b)).

We consider a second network working in parallel with the 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. Hence, this network can also provide an internal supervision signal for the face/non-face discrimination (without a need for an external supervision). The production and recognition of facial expressions can be a bootstrap for the online learning of face/non-face discrimination. The details of the neural network used for the rhythm prediction were presented in [2,4]. The Neural Network uses three groups of neurons, with each group having a different functionality. A Derivation Group (DG) receives the input signal. A Temporal Group (TG) is composed of a battery of neurons (15 neurons) with different temporal activities and is triggered from the DG. Last, the Prediction Group (PG) learns the conditioning between DG (the present) and TG (the past) information. This model (Fig. 13 a) is grounded in the following rule: a PG neuron can learn and can also predict the delay between two events from DG. The equations are provided in the appendix 9.3 and in [15,2] for more details.

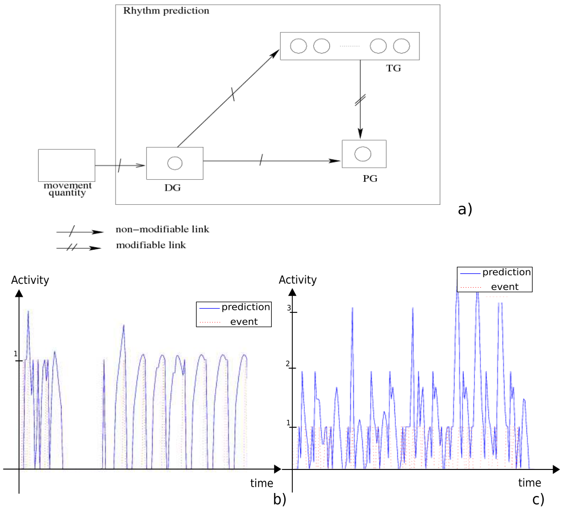

Fig. 13 a) The model for the prediction of the interaction rhythm between the subject and the robot. b) The activity of neuron PG for the rhythm prediction and the activity of neuron DG for the event, when the robot performs the facial expressions and the human imitates the robot head. c) The activity of neuron PG for the rhythm prediction and the activity of neuron DG for the event, when the robot performs facial expressions and the human does not imitate the robot head.

The interaction rhythm provides an interesting 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 reward group is composed of 2 binary neurons. The first one being activated when an interaction is predicted while the second is activated when there is no interaction. A tensorial product between the internal state prediction and the reward group (X group of neurons) allows building an unconditional stimulus for detecting both the correlations with the human facial expression and a predicted rhythmic activity in Y group (Fig. 14). A simple conditioning mechanism using the LMS rule is used to associate the activity of the neurons in the recognition of the local views VF (visual features) with the current X activity; the group Y learns this conditioning (if $X_{i,j}$ is activated, then $Y_{i,j}$ must learn). After learning, the associations between VF activity and Y activity are sufficiently strong to bypass the low-level activity that comes from $X_{i,j}$. Next, a STM_3 (5*2 neurons) is used to accumulate the focus points of an image:

$$STM_{3,(i,j)}(t+1) = \frac{1}{N} Y_{(i,j)}(t+1) + STM_{3,(i,j)}(t) \quad (9)$$

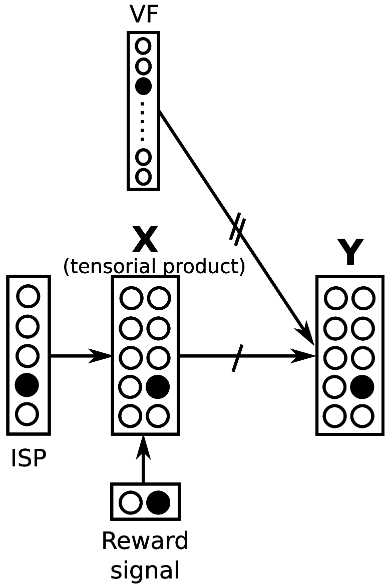

Fig. 14 A tensorial product is performed between the reward signal (2 neurons) and ISP (5 neurons). This tensorial product corresponds to the X group of neurons (a matrix of 10 neurons: 5 lines and 2 columns). The group Y learns the conditioning between VF and X (a simple conditioning mechanism using the LMS rule).

where (i, j) is the index of the neurons ($0 < i \leq 5$ and $0 < j \leq 2$), and N is the number of focus points.

After learning, the STM_3 matrix tends to activate the first column of neurons when there is a face more than the second column, which is more activated when there is no face. Next, a simple group of 2 neurons (FD Face Detection) using classical conditioning [51] is used to predict the reward (rhythmic or non rhythmic activity) from the visual recognition of the local views. If only humans provide synchronization activities with the robot then, in our setup, the reward prediction is equivalent to detecting a face from other stimuli. The group of neurons FD (2 neurons) can recognize the face after learning.

Here, the question is to evaluate the capability of the robot to discriminate a face from a non-face. To validate our model, the following experimental protocol is used: during the learning phase, the database is composed of: first, 4 persons who interacted correctly with the robot (640 images). Second, the robot produces facial expressions but nobody is in front of the robot (there is no human subject in front of the robot). In this latter case, the robot can see objects or the wall (without a human partner) or a human paying no attention to the robot (600 images are collected). During this period, the robot learns to discriminate a face from a non-face. When learning is finished, the robot had to discriminate a face from a non-face who was not observed during the learning phase. A database has been created to perform a complete offline analysis.

In our experiment, the success rate corresponds to the performance of the face/non-face discrimination. The preliminary results linked to this online learning of the face are highly positive. When the face detection is learned and tested using the same subject, the system success rate with that subject tends toward 100%. However, when the face detection is learned with a single subject and is tested over 4 other subjects, the system success rate ranges between 29% (for people with beards) and 90% for more ”similar” subjects. It is important to consider that the learning was performed during a period of only 2 minutes (in real-time) with a single subject. This scenario shows the generalization capabilities of our visual system when focusing the robot's attention on particular visual features and the good generalization properties of the DOG filters. Fig. 15 shows the success rate depending on a function of the number of participants that the robot learns during the learning phase. When, the face discrimination is learned on 4 subjects and the tests are performed on 20 other different subjects, the system success rate tends toward 95% for face detection and 99% for non-face. This figure shows that the performances improve when increasing numbers of interacting partners.

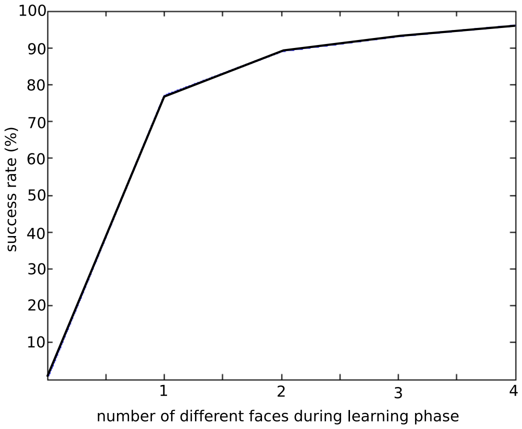

Fig. 15 Face/Non-face generalization. This result shows that the success rate of face recognition is a function of the number of faces that the system learned on during the learning phase. The results are obtained with 20 people (3200 images). After interacting with only 4 people, the system generalizes correctly to 20 people.

8 Discussion & Conclusions

As we have seen in the state of the art (section 2), the usual techniques for facial expression recognition use offline learning and need to access the entire learning database. The aim of these models is to improve the performance of the system. The problem of online and autonomous learning is usually not tackled. For long-term interactions, these methods cannot adapt to facial expressions that were not present in the learning database: for example, the sadness expression in fig. 11. The expression of sadness shown in fig. 11 was different from the way that previous people interacting with the robot had expressed sadness. Hence, even if the main reason for designing our N.N. architecture was at first to try developing a realistic model of a baby learning, we obtain a N.N. architecture that could be used when the databases cannot be made statistically complete. Moreover, we obtained a counter-intuitive result: the development of a fine capability to discriminate faces (and more generally partners or agents) from other objects might not precede the capability of recognizing the facial expressions. At the opposite, our model suggests the pre-eminence of emotional resonance and the recognition of emotional features to enable subsequent face/non-face discriminations.

In our model, facial expressions can be learned without an ad hoc algorithm to locate and frame the face. The only constraint is that the human face will be near enough the robot head to guaranty that on each image enough focus points are taken on the human face. This procedure could be improved by using visual movement detection to focus more on the moving part of the face but it fits well with the short interaction between baby and parents that could sustain such a learning during baby development. Our theoretical model has allowed us to show that to learn online to recognize facial expressions, the learner must produce facial expressions first and be mimicked by his/her caregiver (which is the opposite of using imitation as a way of learning; here, the imitator is the teacher). With the real robot, once the learning period is complete (after 2 minutes), the robot can interact in real time with the human partner (at up to 10 images/second with quad-core Intel CPU). The learning can be fully autonomous thanks to the imitation game between the robot and the caregiver providing both a self supervision signal and a face/non face signal through rhythmic prediction capabilities. Without the caregiver's emotional resonance, the learning would not be able to work. The emotional resonance is the keystone of the model. The face discrimination is learned autonomously during the emotional interaction. It appears to be a byproduct of the emotional interaction and not the first step of an emotional system, as is the case usually in an artificial vision system for facial expression recognition. The emotional interaction is a bootstrap for learning to recognize a human face. Without this bootstrap, we could not find a way to autonomously learn face/non-face discrimination.

An important idea used in this paper was to introduce the prediction of the interaction rhythm as a way to build an intrinsic reinforcement signal that allows modulating the learning rate to avoid learning when there is no partner. Actually, this system is a detector of the interaction based on the temporal dimension of the interaction. The system predicting the timing (here the rhythm) of the interaction is very useful to detect the engagement of agents who interact in front of the robot. In our case as the interacting agent is a Human, it was easy to derive a neural network for face/non-face discrimination. If the human partner is near the robot head, then face/non-face discrimination can be learned. For longer human-robot distances, one can imagine that the present approach could be generalized to human/thing discrimination. Finally, the model described here is more than a simple of face/non face discriminator. It is a generic model capable to detect a caregiver interacting with the robot or more generally a partner or an agent similar in some sense to the robot (providing perhaps the premise for agentivity detection).

In our architecture, the correct learning is a consequence of the interaction with the human partner. Hence, the quality or specificity of the interaction is crucial for the performance level. First, online learning can encounter problems because the human reaction time can be long. This delay perturbs the learning phase because, when the robot changes its facial expression, the first acquired images correspond to the previous facial expression of the human partner, and the robot will attempt to associate them with its new facial expression. To obtain statistically correct results, the display of the expression must be long enough to counter the transitory state (more images from the well-paired situation than from the transitory state). This phase difference between what the robot and the human partner do, is a recurrent problem in all of the interaction experiments that can be solved thanks to the detection of the statistical contingency (in our case using a LMS). Another difficulty arises from the diversity of the human expressions. In particular, some humans seem less expressive and, therefore, the robot has difficulty when categorizing their facial expressions; however, when we ask other human subjects to label the images obtained by the robot, their performances are the same as our system. The compared performances between the robot and the humans show that humans also have great difficulty in recognizing out of context facial expressions performed by non expert subject as opposed to actors for instance (Table. 1 and Table. 2). We underline the difficulty for both humans and our robots to recognize a facial expression such as sadness. This difficulty is certainly linked to a lack of context. Because the imitation game is intrinsically meaningless, there is no signal allowing us to distinguish between sadness or anger or surprise, for example. However, our results show that the robot head recognizes better the happiness and anger facial expressions than the others. These two facial expressions are essential for the development of the robot. For example, these two expressions are sufficient to bootstrap complex capabilities such as the social referencing [6]. In future studies, using facial expressions in meaningful interactions (involving objects, goals) will certainly provide contextual information that will help to solve these problems. Moreover, in this paper, the main limitation (and a priori) was that the robot can learn and recognize only 5 basic facial expressions. An ongoing work focuses on how the robot can produce an infinity of facial expressions by using the same experimental paradigm. In this case, the robot head can perform expressions mixed such as smile and frown using more elementary motor primitive such as opening more or less the mouth (a basic expression being an analog mixture of motor primitives).

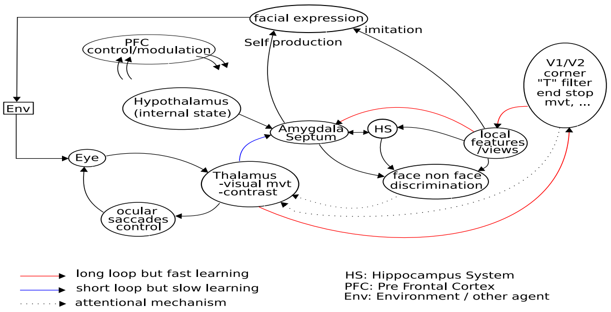

Fig. 16 the thalamo-cortico-amygdala pathway and the thalamo-amygdala pathway in the mammalian brain.

Our results also show that working with the gradient image and the extraction of local views centered on focus points obtained from the convolution of the gradient image with a DOG filter constitute a sufficiently low level visual processing for facial expression recognition, even if these treatments do not allow us to always find the corner of the lips, eyes and eyebrows, which are necessary for model-based methods that are classically used to recognize facial expressions [7,52,1]. The global performances of our architecture could certainly be improved using face-related masks, such as Haar filters, but it was not the purpose of the present work. The focalization strategy corresponds to a sequential and time-consuming analysis of the image. It could be observed as a simple implementation of the thalamo-cortico-amygdala pathway in the mammalian brain [29]. In previous studies [14], we tested simpler and faster architectures using the whole image. The use of primitives based on Gabor filters allows us recognizing a facial expression after parallel processing. However, the learning is more time consuming because the extraction of the pertinent components in the Gabor vector are not easy due to the presence of more background information (as compared to the use of several local views). Such a processing scenario could correspond to the short thalamo-amygdala pathway [43,29], which is implied in rapid emotional reactions. Our works allow to propose a model composed of two parallel networks: (1) as low neural network corresponding to the thalamo-cortico-amygdala network [14]; and (2) a fast neural network corresponding to the thalamo-amygdala network (see Fig. 16). In future studies, we will attempt to verify the idea that the slow neural network could be learned first and used as a way to control the learning of the fast neural network, allowing at the end both a quick recognition of the facial expressions and their precise labeling.

In conclusion, this paper has shown that an emotional interaction can be used as a way of structuring learning (the emotional interaction is a bootstrap for the face/non-face discrimination, which cannot be learned easily in an autonomous way, as opposed to the facial expressions). Ongoing studies indicate that this approach can be generalized to the learning of more complex tasks that involve object manipulation. An emotional interaction can bootstrap the social referencing of objects and then allow the robot to avoid negative objects and to grasp positive objects (the whole learning process is based on a chain of simple emotional conditioning learned autonomously without any predefined signal). Hence, we can suggest that the baby/parents system is an autopoietic social system [37] in which the emotional signal and the adult emotional resonance are important elements for maintaining the interaction and for allowing the learning of more and more complex skills. Future studies using our robotic head will attempt to test this hypothesis in more dynamical situations involving human/robot and robot/robot interactions.

Downloads

Video:
http://www.etis.ensea.fr/˜neurocyber/Videos/emotion_head/ expressionFaciale.mp4
http://www.etis.ensea.fr/˜neurocyber/Videos/emotion_head/ social_referencing.wmv
Database:
For interested researchers, The database is available on demand

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 modeling. This study is part of the European project "FEELIX Growing" IST-045169 and also the French Region Ile de France (DIGITEO project) and ANR INTERACT. P. Gaussier thanks also the Institut Unisersitaire de France for its support.

References

1. B. Abboud, F. Davoine, and M. Dang. Facial expression recognition and synthesis based on an appearance model. *Signal Processing: Image Communication*, 19:723–740, 2004.
2. 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.
3. M. Asada, K. Hosoda, Y. Kuniyoshi, H. Ishiguro, T. Inui, Y. Yoshikawa, M. Ogino, and C. Yoshida. Cognitive developmental robotics: A survey. *Autonomous Mental Development, IEEE Transactions on*, 1(1):12–34, 2009.
4. J.P. Banquet, P. Gaussier, J.C. Dreher, C. Joulain, A. Revel, and W. Günther. Space-time, order, and hierarchy in fronto-hippocampal system: A neural basis of personality. In Gerald Matthews, editor, *Cognitive Science Perspectives on Personality and Emotion*, volume 124, pages 123–189, Amsterdam, 1997. North Holland.
5. S. Boucenna, P. Gaussier, and L. Hafemeister. Development of first social referencing skills: Emotional interaction as a way to regulate robot behavior. *IEEE Transactions on Autonomous Mental Development*, pages 1–14, 2013.
6. S. Boucenna, P. Gaussier, L. Hafemeister, and K. Bard. Autonomous development of social referencing skills. In *From Animals to Animats 11*, volume 6226 of *Lecture Notes in Computer Science*, pages 628–638. 2010.
7. 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.
8. J. Decety. Dissecting the neural mechanisms mediating empathy. *Emotion review*, 3:92–108, 2011.
9. 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.
10. 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.
11. 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.
12. 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.
13. P. Gaussier. Toward a cognitive system algebra: A perception/action perspective. *In European Workshop on Learning Robots (EWRL).*, pages 88–100, 2001.
14. P. Gaussier, S. Boucenna, and J. Nadel. Emotional interactions as a way to structure learning. *epirob*, pages 193–194, 2007.
15. P. Gaussier, S. Moga, M. Quoy, and J.P. Banquet. From perception-action loops to imitation processes: a bottom-up approach of learning by imitation. *Applied Artificial Intelligence*, 12(7-8):701–727, Oct-Dec 1998.
16. 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.
17. G. Gergely and J. Watson. Early socio-emotional development: contingency perception and the social-biofeedback model. *In P. Rochat, (Ed.), Early Social Cognition: Understanding Others in the First Months of Life*, pages 101–136, 1999.
18. C. Giovannangeli and P. Gaussier. Interactive teaching for vision-based mobile robot: a sensory-motor approach. *IEEE Transactions on Man, Systems and Cybernetics, Part A: Systems and humans. In Press*, to appear 2009.
19. C. Giovannangeli, P. 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.
20. S. Grossberg. Competitive learning: From interactive activation to adaptive resonance. *Cognitive Science*, 11:23–63, 1987.
21. H. Gunes, B. Schuller, M. Pantic, and R. Cowie. Emotion representation, analysis and synthesis in continuous space: A survey. In *Automatic Face Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 827–834, 2011.
22. S. Harnad. The symbol grounding problem. *Physica D*, 42:335–346, 1990.
23. C. Hasson, S. Boucenna, P. Gaussier, and L. Hafemeister. Using emotional interactions for visual navigation task learning. *International conference on Kansei engineering and emotion researchKEER2010*, pages 1578–1587, 2010.
24. R.L Hsu, M. Abdel-Mottaleb, and A.K. Jain. Face detection in color images. *Pattern Analysis and Machine Intelligence, IEEE Transactions*, 24:696–706, 2002.
25. C. Izard. The face of emotion. *Appleton Century Crofts*, 1971.
26. J.M. Jenkis, K. Oatley, and N.L. Stein. *Human Emotions*, chapter The communicative theory of emotions. Blackwell edition, 1998.
27. 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.
28. T. Kohonen. Self-organization and associative memory. *Heidelberg: Springer-Verlag, Berlin, 3rd edition.*, 1989.
29. J.E. LeDoux. *The Emotional Brain.* Simon & Schuster, New York, 1996.
30. J.K. Lee and C. Breazeal. Human social response toward humanoid robot's head and facial features. In *CHI '10 Extended Abstracts on Human Factors in Computing*

Systems, CHI EA '10, pages 4237–4242, New York, NY, USA, 2010. ACM.

31. 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.
32. G. Littlewort, M.S. Bartlett, I. Fasel, T. Kanda, H. Ishiguro, and J.R. Movellan. Towards social robots: Automatic evaluation of human-robot interaction by face detection and expression classification. volume 16, pages 1563–1570. MIT Press, 2004.
33. D.G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2:91–110, 2004.
34. M. Lungarella, G. Metta, R. Pfeifer, and G. Sandini. Developmental robotics: a survey. *Connection Science*, 15(4):151–190, 2003.
35. 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.
36. M. Masahiro. The uncanny valey. *Energy*, 7:33–35, 1970.
37. H.R. Mataruna and F.J. Varela. *Autopoiesis and Cognition: the realization of the living.* Reidel, Dordrecht, 1980.
38. D. Muir and J. Nadel. Infant social perception. In A. Slater, editor, *Perceptual development*, pages 247–285. Hove: Psychology Press, 1998.
39. 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.
40. J. Nadel, M. Simon, P. Canet, P. Soussignan, P. Blancard, L. Canamero, and P. Gaussier. Human responses to an expressive robot. In *Epirob 06*, 2006.
41. 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.
42. M.A. Nicolaou, H. Gunes, and M. Pantic. Output-associative rvm regression for dimensional and continuous emotion prediction. *Image Vision Comput.*, 30(3):186–196, March 2012.
43. J.W. Papez. A proposed mechanism of emotion. *Archives of Neurology and Psychiatry*, 1937.
44. R. Plutchick. A general psychoevolutionary theory of emotion. *Emotion: Theory, research and experience*, pages 3–33, 1980.
45. H.A. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20:23–38, 1998.
46. D. E. Rumelhart and D. Zipser. Feature discovery by competitive learning. *Cognitive Science*, 9:75–112, 1985.
47. T. Sénéchal, V. Rapp, H. Salam, R. Seguier, K. Bailly, and L. Prevost. Facial action recognition combining heterogeneous features via multi-kernel learning. *IEEE Transactions on Systems, Man, and Cybernetics–Part B*, 42(4):993–1005, 2012.
48. B. Thirioux, M.R. Mercier, G. Jorland, A. Berthoz, and O. Blanke. Mental imagery of self-location during spontaneous and active self other interactions: An electrical neuroimaging study. *The Journal of Neuroscience*, 30(21):7202–7214, 2010.
49. S. Thorpe, D. Fize, and C. Marlot. Speed of processing in the human visual system. *Nature*, 381:520–522, 1996.
50. P. Viola and M. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57:137–154, 2004.
51. B. Widrow and M. Hoff. Adaptive switching circuits. In *IRE WESCON*, pages 96–104, New York, 1960. Convention Record.
52. L. Wiskott. Phantom faces for face analysis. *Pattern Recognition*, 30:586–191, 1991.
53. 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.
54. J. Yu and B. Bhanu. Evolutionnary feature synthesis for facial expression recognition. *Pattern Recognition Letters*, 2006.
55. Z. Zeng, M. Pantic, G.I. Roisman, and T.S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(1):39–58, 2009.

9 Appendix

9.1 Visual processing

The visual attention is controlled by a reflex mechanism that allows the robot to focus its gaze on potentially interesting regions. 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 [19,35]). This process allows the system to focus more on the corners and ends of the lines in the image (e.g., eyebrows, corners of the lips). The main advantages of this process over the SIFT (Scale Invariant Feature Transform) [33] method are its computational speed and a smaller number of extracted focus points. The intensity of the focus points is directly linked to their level of interest. Through a recurrent inhibition of the already selected points, a sequence of exploration is defined (a scan path). A short-term memory allows maintaining the inhibition of the already explored points. This memory is reset after each new image acquisition[9].

To reduce the computational load (and to simplify the recognition), the image is sub-sampled from 720x580 to 256x192 pixels. 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. 18) to obtain an input image or a vector that is more robust to the rotations and distance variations. In our case, the log-polar transform has a radius of 20 pixels and projects to a 32x32 input image. The Gabor filtering is used to extract the mean and the standard deviation for the convolution of the 24 Gabor filters (see appendix). Hence, the input vector obtained from the Gabor filter has only 48 components. It is a smaller vector than the result of the log-polar, which transform induces an a priori better generalization but also has a lower discrimination capability.

9.2 Gabor filters

A Gabor filter has the following equation:

$$G(x,y) = \frac{1}{2\pi\sigma} \exp(-\frac{x^2+y^2}{2\sigma^2}) \cos(2\pi f(x\cos(\theta) + y\sin(\theta))) \tag{10}$$

$$f = \frac{1}{\sigma\gamma} \tag{11}$$

[9] The weak point of this technique is that the DOG size must fit the resolution of the objects to analyze. With a standard PAL camera, this constraint is clearly not a problem because there is not much choice if the size of the face must be large enough to allow a correct recognition (for an HD camera, a multi-scale approach would be necessary). For more general applications, a multi-scale decomposition could be performed (for example using 3 scales).

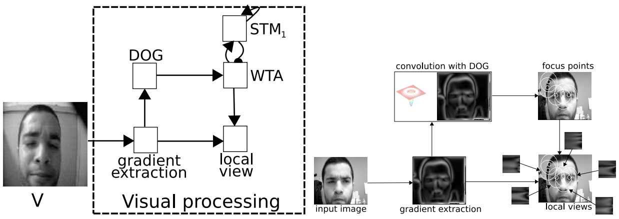

Fig. 17 Visual processing: This visual system is based on a sequential exploration of the image focus points. A gradient extraction is performed on the input image (256x192 pixels). A convolution with a Difference Of Gaussian (DOG) provides the focus points. Last, the local views are extracted around each focus point.

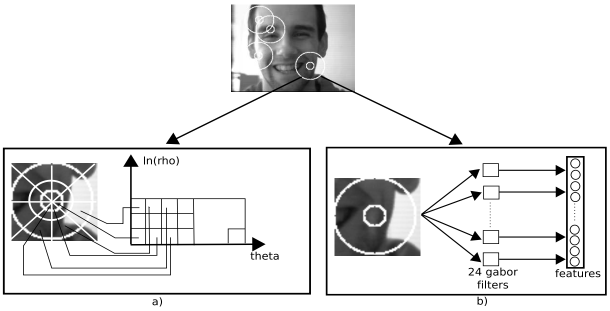

Fig. 18 Visual features: a) The local log-polar transform increases the robustness of the extracted local views to small rotations and scale variations (a log-polar transform centered on the focus point is performed to obtain an image that is more robust to small rotations and distance variations. Its radius is 20 pixels). b) Gabor filters are applied to obtain a signature that is more robust than a log-polar transform (the Gabor filters are 60x60); the features extracted for each convolution with a Gabor filter are the mean and the standard deviation.

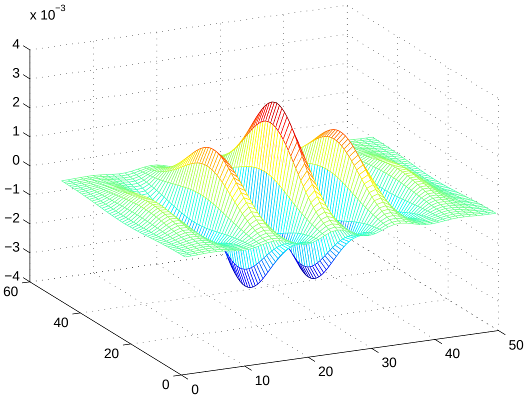

Fig. 19 A Gabor filter with these parameters $\sigma = 8$, $\gamma = 3$, and $\theta = \pi/3$.

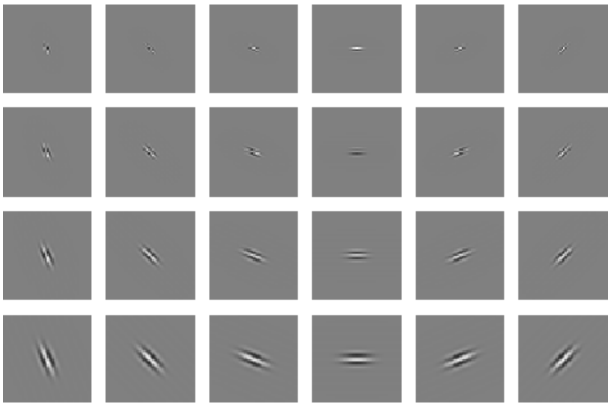

Fig. 20 Gabor filters for different frequencies and orientations.

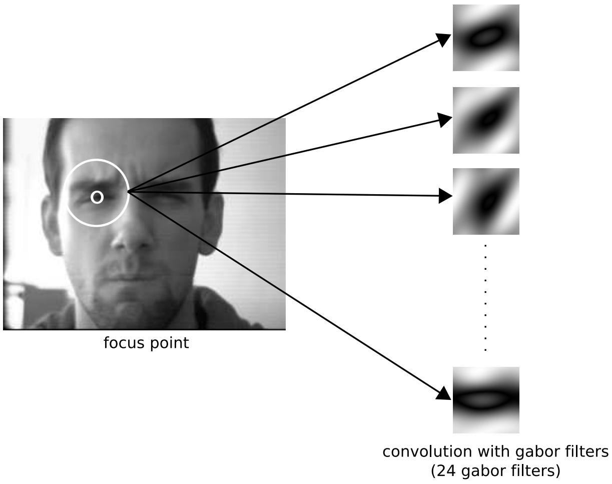

Fig. 21 Result of the convolution with 24 Gabor filters (different frequencies and orientations)

9.3 Model for the rhythm prediction

The Neural Network uses three groups of neurons. The Derivation Group (DG) receives the input signal, and the Temporal Group (TG) is a battery of neurons (15 neurons) with different temporal activities. The Prediction Group (PG) learns the conditioning between DG (the present) and TG (the past) information. In this model (Fig. 13, a PG neuron learns and also predicts the delay between two events from DG. A DG activation sets to zero the TG neurons because of the links between DG and TG. The neuronal activity (DG) involves instantaneous modifications in the weights between PG and TG. After each reset (set to zero) by DG, the TG neurons have the following activity:

$$Act_l^{TG}(t) = \frac{1}{m} \cdot \exp - \frac{(t-m)^2}{2 \cdot \sigma} \tag{12}$$

l corresponds to the cell subscript, m is a time constant, σ is the standard deviation, and t is the time. The TG activity presents an activity trace DG (of the past). PG receives the PG and TG links, and the TG information corresponds to the elapsed time since the last event, whereas the DG information corresponds to the instant of the appearance of a new event. PG sums the TG and DG inputs with the following equation:

$$Pot^{PG} = \sum_l W_{pg}^{tg(l)} \cdot Act_l^{TG} + W_{pg}^{dg} \cdot Act^{DG} \tag{13}$$

Act_l^{TG} is the l cells activity of TG, $W_{pg}^{tg(l)}$ are the weight values between TG and PG, Act^{DG} is the DG neuron activity and W_{pg} is the weight value between DG and PG. The PG activation is triggered by the maximal value detection of its potential (the maximal value is equal to the potential's derivative when it is equal to zero):

$$Act^{PG} = f_{PG}\left(Pot^{PG}\right) \tag{14}$$

$$f_{PG}\left(x\left(t\right)\right) = \begin{cases} 1 \quad if \ \frac{dx(t)}{dt} < 0 \ and \ \frac{dx(t-1)}{dt} > 0 \\ 0 \ else \end{cases} \tag{15}$$

Finally, only the W_{PG}^{TG} are learned, and we perform a one-shot learning process. The learning rule is the following:

$$W_{PG}^{TG(l)} = \begin{cases} \frac{Act_l^{TG}}{\sum_l \left(Act_l^{TG}\right)^2} \ if \ Act^{TG} \neq 0 \\ inchange \qquad else \end{cases} \tag{16}$$

Thereby, the PG is activated if and only if the TG cells' activity sum is equal to that of learning.

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 received the M.S. degree in electronics from Aix-Marseille University, Marseille, France, in 1989, and the Ph.D. degree in computer science from the University of Paris XI, Orsay, France, for his work on the modeling and simulation of a visual system inspired by mammals vision. From 1992 to 1994, he conducted research in neural network applications and in control of autonomous mobile robots at the Swiss Federal Institute of Technology, Zurich, Switzerland. He is currently a Professor at Cergy-Pontoise University, Cergy-Pontoise, France, and leads the Neurocybernetic Team of the Image and Signal processing Lab (ETIS). His research interests are focused on the modeling of the cognitive mechanisms involved in visual perception, motivated navigation, and action selection, and on the study of the dynamical interactions between individuals with a particular research in the fields of imitation and emotions.

Pierre Andry received the M.Sc. degree in artificial intelligence from Paris VI University, Pierre et Marie Curie, France, in 1999, and the Ph.D. degree in computer science from the Cergy-Pontoise University, Cergy-Pontoise, France, in 2002, investigating the role of imitation in learning and communication skills of autonomous robots. He is currently a Researcher at the ETIS Lab, University of Cergy Pontoise, Cergy Pontoise, France. His research interests include epigenetic robotics, i.e., the way autonomous robots could develop in order to adapt to their physical and social environment. This concern includes issues such as imitation, emotions, preverbal interactions processes, turn-taking, and emergent behaviors.

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.