HAL
open science

# Readability of the gaze and expressions of a robot museum visitor : impact of the low level sensory-motor control

Aliaa Moualla, Ali Karaouzene, Sofiane Boucenna, Denis Vidal, Philippe Gaussier

## To cite this version:

Aliaa Moualla, Ali Karaouzene, Sofiane Boucenna, Denis Vidal, Philippe Gaussier. Readability of the gaze and expressions of a robot museum visitor : impact of the low level sensory-motor control. 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2017), Aug 2017, Lisbonne, Portugal. 10.1109/ROMAN.2017.8172381 . hal-01706159

## HAL Id: hal-01706159
## https://hal.science/hal-01706159

Submitted on 10 Feb 2018

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Readability of the gaze and expressions of a robot museum visitor: impact of the low level sensory-motor control

Aliaa Moualla[1], Ali Karaouzene [1], Sofiane Boucenna [1], Denis Vidal[2] and Philippe Gaussier[1]

***Abstract*— In this paper we propose a neural network allowing a mobile robot to learn artwork appreciation. The learning is based on the social referencing approach. The robot acquires its knowledge (artificial taste) from the interaction with humans. We present and analyze specifically the visual system, its impact on the robot behavior, and at the end, we analyze the readability of our robot behavior according to visitors comments. We show that the low level spatial competition between the values associated to areas of interest in the image are important for the coherence of the robot's object evaluation and the readability of its behavior.

***Index Terms*— Artificial intelligence, neural networks, Human robot interactions, computer vision.**

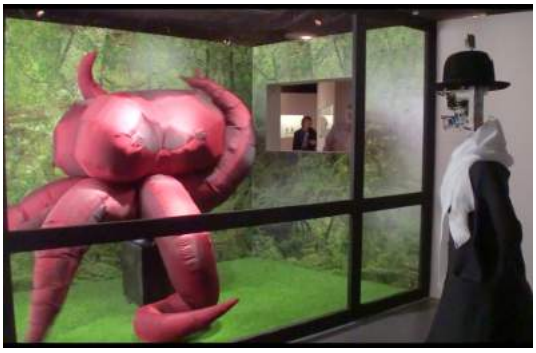

Fig. 1. The robot Berenson at the right looking at an artwork in the Quai Branly Museum.

## I. Introduction

This work belongs to an interdisciplinary project between robotics and anthropology. Our global goal is to find a simple neural model to study the emergence of "Artificial Aesthetic" in robots. Our robot Berenson exhibits complex behavior based on a simple sensory-motor architecture (PerAc)[13]. Using this architecture, the robot learns social referencing skills [16]. It develops in the museum of Quai Branly in Paris a new kind of art appreciation (artificial aesthetics taste) through social interactions. Here, we present the model of Berenson's visual system and its impact on low level control of the robot actions. Two types of visual information are processed in parallel : the what and where information [12] (the recognition of some local views and their position).

At the social interaction level, it is primordial to understand the intentions of people who are involved in the interaction with us. However, could visitors understand the intentions of our robot or at least its artistic preferences without any explicit language ? How do people expect the robot to communicate ? Can they explain and predict the behavior of our robot when the robot behavior is only controlled thanks to a sensory-motor architecture ?.

As a first step for our robot to participate in social interactions, we expect that its behavior should be understandable to the museum visitors. Analyzing the movement of our robot and capturing its expressivity in front of artworks should give the visitors an indication of its preferences. Often, when a robot is used in museums, it is used either as an artwork as in Tinguely's work or as a guide for visitors [1], [2], [3]. In our case, our robot Berenson is a new kind of visitor. It has to develop its own artwork preferences thanks to the social interaction. And even if there are still very few studies in robotics with long-term interactions in a real life context, we chose the challenge of putting a humanoid at the Quai Branly Museum, an environment where no one expects a robot to walk alone. In this paper, we begin with a presentation of the experiment in the museum. Next, we present the architecture of the visual system, the sensory-motor architecture allowing to associate an emotional values with an observed object. We study the performance of the visual system with and without a spatial competition mechanism. Then we present the model controlling the robot's navigation. At the end, we analyze the results of the visual system performance and some tests that were done in the museum to evaluate the readability of the robot's behavior (its movements and facial expressions).

## II. Material and methods

The setup is composed of a robot and a distant workstation. Berenson is composed of a robulab 10 from Robosoft, associated with an embedded computer, and an expressive head. Its weight is almost 20 kg and height 1.80 meters. To avoid obstacles, Berenson is equipped with some proximity sensors, 15 infrared sensors and 9 ultrasonic rangers and a laser. The sensors are placed all around its frame. One magnetic compass is used to navigate. Berenson uses the camera in its right eye to perform the artwork recognition task. The expressive head has 9 degrees of freedom (DoF), 4 for eyebrows, 3 for the mouth, 1 for the front tilt and 1 for the eyes tilt. The embedded computer manages the sensors (including camera) and the actuators. It computes low-level algorithms (artwork recognition). A WiFi connexion can be used to debug the robot behavior.

As explained above, Berenson learns with museum visitors

This research was supported by the Labex PATRIMA and The Quai Branly museum

[1] ENSEA, University of Cergy-Pontoise, UMR CNRS 8051 ETIS, FRANCE.

[2] IRD, URMIS-Paris Diderot, 4, Rue Enghien, 75010 Paris, FRANCE.

to relate an artwork with an emotional value (positive or negative) and in a second step it will move towards those artworks and will express the associated value. Museum mediators asked visitors (those who agree to interact with Berenson and teach it their preferences) to select an artwork they found more interesting or impressive than the others (positive judgment) or at the opposite end one less interesting than the others (negative judgment). A joystick is used to drive the robot in front of the artworks during the teaching phase. It controls the direction of the robot's attention. The goal is to center the desired object (selected by the mediators or the visitors) in the robot field of view. After that, using a two-button mouse, the mediator assigns the visitor's appreciation to the observed object. Then, the robot associates the recognized artwork with the given emotional value. More natural interactions were done using facial expression recognition in previous works [7].

Berenson visits randomly the museum avoiding the obstacles (objects or visitors). When an interesting object (positively learned) enters its field of view, Berenson heads to this object and changes its direction to center the positive object in its field of view, expressing the associated facial expression (joy). If Berenson perceives a negative object it expresses a negative facial expression and heads in the direction of any other object (positive object or neutral object if no positive object can be found in its field of view).

The robot's navigation is controlled by a dynamical neural field [15]. The system for object tracking is inspired by [10]. Our neural network associates the where information (local view positions in the image) with the what information (visual features) and the system takes into account the what and where information during the recognition of the object and the choice of the object to follow. The field of view is discretized in a set of neurons (population coding). The most active neuron on the field gives the direction of the motor angular command.

The first experiment took place at the Quai Branly Museum, in Paris in the Insulinde (Insular South-est Asia) area art display for 4 hours a day. A second one took place in the same museum in PERSONA exposition from 26 january to 13 November 2016.

### A. The What channel estimation

This section describes how Berenson's visual system functions and how Berenson learns to associate an artwork with a positive/negative or neutral emotional value. Berenson uses a bio-inspired visual system. The visual system input is a subsampled grayscale image (320x240 pixels). A gradient image is computed from this image. Then a convolution between the gradient image and an off-center filter is used to extract focus points through a local competition between all the salient areas. Around each focus point, the system extracts a local view of a radius of 60 pixels (5 local views per image are extracted during the learning mode and 15 local views in test mode). Since there is no simple way to segment one object from the others (one artwork may be confused with other proximal objects). During the learning phase the field of view is reduced to the central area of the image to avoid a noisy learning.

Local views are transformed into log/polar coordinates to allow robustness to scale variation and small perspective changes. Then the SAW (selective adaptive winner) algorithm (a real-time K-means algorithm) learns and categorizes compressed local view pattern on a set of neurons [5] (Fig. 2).

$$VF_j = net_j.H_\gamma(net_j) \tag{1}$$

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

$VF_j$ is the activity of neuron $j$ in the group $VF$ (visual feature). $net_j$ is the complement to 1 of the sum of the distances between the input feature and the nearest similar feature learned. $N$ is the local view size, $I_i$ is the input visual feature, and the learned features are coded on the neuron's weight $W_{ij}$. $H_\gamma(x)$ is the Heaviside function [1]. $\gamma$ is a vigilance parameter (the threshold of recognition).

Incoming local views are compared with learned patterns. If the maximum activity is below a given threshold, the observed local view is learned as a new pattern and associated to a recruited neuron, which means if the recognition activity $net_j < \gamma$ one new neuron is recruited (learning of a new local view). Otherwise the SAW algorithm adapts the link between the winner neuron and the input pattern as in the K-means algorithm.

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

In eq.3, when a new neuron is recruited $a_j = 1$, otherwise, $a_j = 0$. The threshold $\gamma$ was set to 0.98 in learning mode and 0.78 in test mode. The vigilance $\gamma$ is set at a high value when the robot is instructed to learn something to ensure strong learning. Low vigilance in test mode allows generalization to unlearned objects. The robot can attribute valences to objects it has never seen before. This ability to generalize on unlearned objects is very interesting to generalize learned aesthetic preferences to new objects. Nevertheless, it is necessary to have a compromise between generalization and discrimination. We added a system of normalization and competition between categories to filter out stronger generalizations (see Section. II-C). The results in Section. III will show the performance enhancement due to this competition mechanism. The emotional value association is supervised. Associations of a local view $VF$ and an emotional value $EV$ (indicated by the visitor) is made thanks to a Pavlovian conditioning (Fig. 2), based on a least mean square (LMS) algorithm [14]. LMS minimizes the error between desired output (one to one unconditional links) and the actual output (one to all conditional links). $OEV$ is

1. Heaviside function :

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

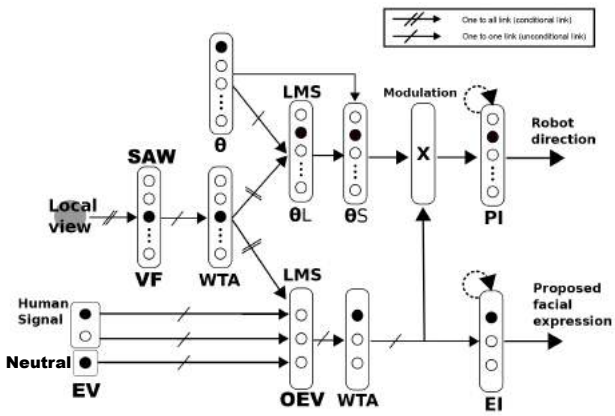


Fig. 2. Artwork appreciation architecture. Architecture inputs are local views (from a camera) and human signal (from a two-button mouse). The outputs are a direction and a proposed facial expression redirected respectively to the robot control (neural field).

the learned object emotional value. After learning, the local views are associated to a positive or negative value according to the visitor's instructions.

$$\Delta W_{ij}^{VF-OEV} = \epsilon_1 . VF_i . (EV_j - OEV_j) \tag{4}$$

The groups $VF$ learns and categorises the visual features (the SAW group in our model). $\epsilon_1$ is the learning rate. The system learns to associate an emotional state with an object or a scene.

We added a constant input to associate neutral objects to a non-emotional value (neutral state). Association with the neutral state is done 10000 times slower than others. When facing an undesired object, like a wall or a window, the neutral association can be sped up manually to reduce the teaching duration. The slow learning rate allows Berenson to forget very old learned artworks, and to avoid paying attention to undesired objects in order to express a neutral facial expression when facing those objects. After the LMS, (Fig. 2) a WTA activates only the neuron with the highest activity, the first neuron represents a positively learned local view ; the second one is for negatively learned views and the last for neutral ones. This represents Berenson's internal feeling state $EI$.

### B. The Where channel estimation

The way Berenson navigates in the museum is influenced by its previous interactions with visitors. It will move in a preferred way by going to the objects it appreciates (positive objects) according to its learning. An object is considered as a set of local views. Estimating the object position (location) in the camera field of vision remains to estimate the relative positions of its components according to a given referential. The $Where$ information associated to a positive value is projected on a population of neurons coding for the robot orientation. More precisely, the $\theta_L$ group associates the predicted $Where$ information with each local view $VF$ in order to compute the artwork shift in the image and then the robot angular command. Next equations shows the weight modification for each iteration (the $t$ time variable is not represented). The $\theta_L$ group corresponds to the learned position.

$$\theta_L = [W].VF \tag{5}$$

$$\Delta W_{ij}^{VF-\theta_L} = \epsilon_1 VF_i(\theta_j - \theta_{Lj}) \tag{6}$$

The groups $\theta$, $\theta_L$, $\theta_S$, use population coding for angle computation (Fig. 2). The first neuron codes the image's left border position and the last neuron in the field codes the image's right border. The number of neurons in $\theta$, $\theta_L$, $\theta_S$, depends on the desired population coding quantization. Here, arbitrarily, 60 neurons are used, one neuron per camera degree. Our camera has a field of view of 60°. The neurons in the $\theta$ group corresponds to the focus point's position along the $x$ coordinates in the input image (the same model is used for the $y$ coordinates).

$$\theta_t(x) = \sum_{m=0}^{M} \theta'_t(m) . \frac{1}{2\pi\sigma_1^2} e^{\frac{-(x_t-m)^2}{2\sigma_1^2}} : \theta_t'(m) = \delta_{d_t}(m) \tag{7}$$

$$\Rightarrow \theta_t(x) = \frac{1}{2\pi\sigma_1^2} e^{\frac{-(x_t-d_t)^2}{2\sigma_1^2}} \tag{8}$$

$\theta_t(x)$ is the resultant vector of the convolution of the local view position with a gaussian kernel. $M$ is the vector size. During the learning phase the $\theta_L$ group associates each local view explored with the central position in the image thanks to the LMS rule eq.6 (the central position should match with the global position of the object if the object was well centered by the subject). The learning between neurons associated to the view recognition and the angular position of the object center is done via one to all conditional links. They work like a memory storing the local views position during the learning phase. Thus, in use mode, when a local view is recognized the neuron coding for its learned position is activated. the $\theta_S$ group corresponds to the shifted position. $\theta_S$ computes the distance between the learned and the current position as formalized in eq.9. When the local view is at the learning position then $\theta_S = 0$. If the local view is translated by a $\Delta d$ distance then $\theta = \theta_L + \Delta d$ the neuron $\Delta d$ is activated in $\theta_S$.

$$\theta_S(Circ(x - \arg\max(\theta_L)) = \theta(x) \tag{9}$$

$$Circ(x) = \begin{cases} x, x > 0 \\ x + N, x < 0 \end{cases} \tag{10}$$

The $\arg\max(\theta_L)$ give a position at which $\theta$ is maximized. $\theta_S$ is the vector $\theta$ circularly shifted. Now, if we assume that the object to learn is well centered in the camera field, the referential becomes the object center (in the $x$ abscissa). The local views belonging to this object predict in $\theta_S$ their distance to the object center. The system can estimate the object pose by integrating the activity of the neurons in $\theta_S$.

The *Position Integration* $PI$ group integrates the local views distances to the referential with a Gaussian kernel

| | Q1 | Q2 | Q3 | Q4 | Q5 |
|---|---|---|---|---|---|
| PV | 0.3301 | 0.722 | 0.8902 | 0.0076 | 0.5947 |
| | P1/P2 | P1/P2 | P1/P2 | P1/P2 | P1/P2 |
| Mean | 3.7/3.4 | 3.6/3.5 | 3.4/3.5 | 3.7/2.4 | 3.2/3.0 |
| SD | 1.0/1.1 | 0.9/1.1 | 1.0/1.3 | 1.1/1.6 | 1.3/1.3 |
| SEM | 0.2/0.2 | 0.2/0.2 | 0.2/0.3 | 0.2/0.3 | 0.3/0.3 |
| N | 17/20 | 16/20 | 17/19 | 17/20 | 17/20 |

TABLE I

SURVEY STATISTICAL ANALYSIS,PV :P-VALUE

summation eq.11. In the learning phase when the object is in the center of the field of view the sum of the activity creates a peak at the image center. When the object is translated, the peak is also translated in the image referential. The Gaussian kernel summation provides a robustness to small rotation and perspective changes. The robustness is driven by the Gaussian standard deviation (sd). The larger the sd, the more robust the system is, the less discriminant it becomes.

$$PI(x) = \sum_{t=1}^{\tau} \frac{1}{2\pi\sigma_1^2} e^{\frac{-(x_t - ds_t)^2}{2\sigma_1^2}} \tag{11}$$

$$ds_t = (Circ(d_t - \arg\max(\theta_L)).$$

An example is shown in Fig.3, Fig.4. If an object is composed by slightly the same $VF$ as the learned object but put in different position, the system creates small peaks scattered in the $PI$ field. The activity in $PI$ group may represent the confidence level the system has in the estimated pose.

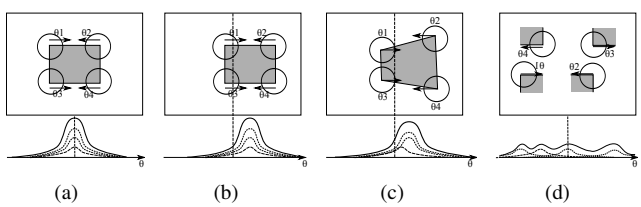


Fig. 3. Schematic example of pose estimation applied to a square. 3(a) is the learned object at the image center. Below is the activity in the $PI$ group. It creates a peak at the object location. 3(b) shows the same square translated and the translation result in $PI$ group. 3(c) is the same square with deformation, and its estimated pose below. In 3(d) contains same local view as 3(a) scattered in the image.

$\sigma$ is the kernel sd. All kernels have the same sd. $ds_t$ is the distance of each local view to the the referential. When $ds_t - ds_{t-n} < 2\sigma$, the local views $VF_t$ and $VF_{t-n}$ predict the same pose. Thus, the estimation uncertainty could be driven by the kernel sd. Fixing the $\sigma$ depends on the wanted application, the environment and the object sizes. At this level, we already could make a decision about the object recognition confidence. A high neuron activity level in $PI$ represents good object recognition and a low activity represents bad recognition. The object pose estimation system is parallel to the object recognition.

### C. Adding the normalization and the competition mechanism to the model

In this section, we propose some methods to take into account the What and Where information during the recognition of the object and the choice of the object to follow. In

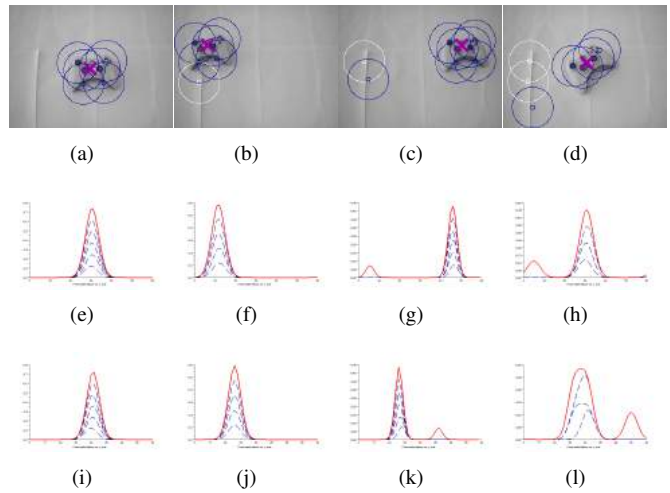


Fig. 4. Schematic example of pose estimation applied to a an XBox pad. 4(a) is the learned phase. Below is the activity in the $PI$ group. All the patterns vote for the object center. The pink cross shows the predicted object location. 4(b) and 4(c) show the same object translated to the upper left and upper right. The resulting predictions are depicted in 4(f) and 4(g). 4(d) is the same object rotated, and its estimated pose below. The system is invariant under translation changes and robust to small rotation, scale and perspective changes.

the learning phase, the robot associates the local views of an artwork with the emotional value attributed by the visitor.

In test phase, high activity in the PI group represents a high probability of recognizing the object while low activity represents a low probability of recognizing the object. The robot assigns an emotional value to some artworks or visitors faces that present some similarities with the learned objects.

Like mentioned above estimating the position of an object in the image reference is equivalent to estimating the position of its local views with respect to the center of the object and in order to know the valence associated with an object, the responses of the local views are summed. In the beginning we used a simple product between the What and Where information. This solution was problematic because with the generalization, a lot of neurons associated with the recognition of local views looking more or less like the learned views were activated as well as the direction associated to their learned position in the image. Fig. 6 shows how we can use the emotional value (EI) to affects the choice of the winner object when Berenson is facing some distractors. In this figure, the pentagon on the left is associated with a negative emotional value. Five focal points and their associated local views are extracted on this object. The triangle on the right, with only three local views extracted, is associated with a positive value. The result of the position estimation is shown below the objects. A higher bubble is associated with the pentagon because it is the sum of five bubbles of activity while the triangle is only associated with 3 local views. A direct competition between those two bubbles induces the selection of the pentagon as the winner object (the object to follow). If we consider that the robot should select objects that please him (positive valence objects), the robot reaction is incorrect in this case.

Then we will add a modulation with the emotional value

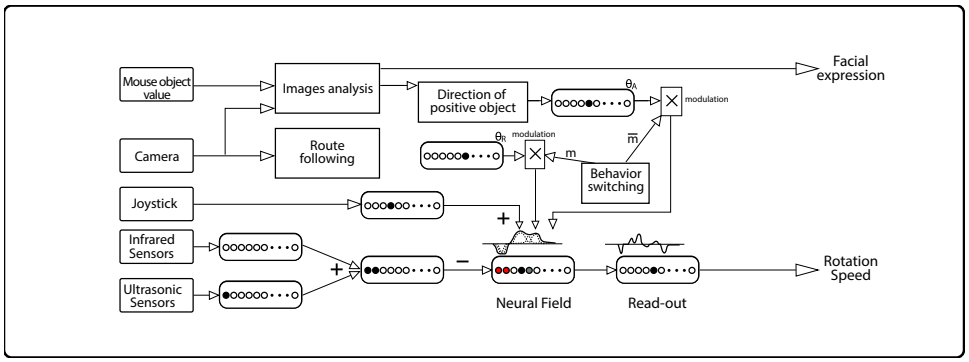


Fig. 5. Functional diagram of the control architecture. The direction of positive object and image analysis are part of the artwork appreciation behavior. The main behaviors (route following and artwork appreciation) determine the directions $\theta_A$ and $\theta_R$ using camera input. behavior switching alternatively inhibits one of the main behaviors. The neural field merges activities from the main behaviors, manual control and sensors. Rotation speed is computed as the derivative of the neural field activation as read-out mechanisms.

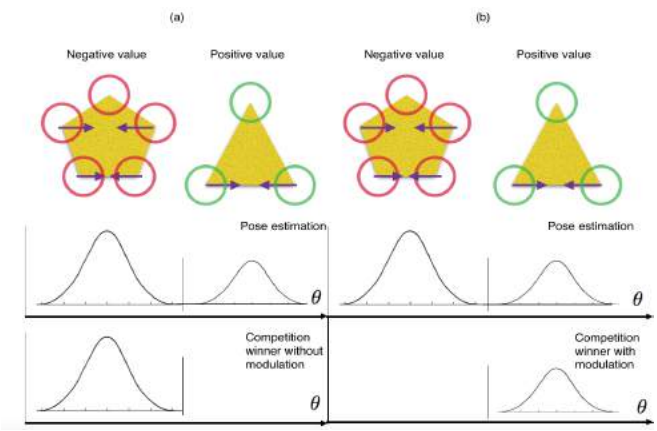


Fig. 6. Example of the effect of modulation by valence

see eq.12 and a normalization see eq.14. (We use the annotation of $A(i,j)$ to represent a matrix, $A_i(j)$ to the vector that represent the $i_{th}$ row of the matrix $A(i,j)$, $A(i)$ for vectors). On one hand, the valence modulation of local views facilitates the reconstruction of the object and assists decision making.

$$\beta(e,x) = V(e)^T.\theta_s(x) \tag{12}$$

$$PVI_e(x) = \sum_{t=1}^{\tau} \beta_{te}(x) \tag{13}$$

$\beta(e,x)$ is a matrix representing the estimated positions $\theta s(x)$ weighted by the emotional valence $V(e)$.

On the other hand, in heavily textured environments, the number of local views varies from one object to another. To give the same order of magnitude to all predictions, the position field has to be normalized. For this, the position estimated by the integration of the estimates of all the local views is divided by the number of views in this same position. $NPVI(e,x)$ is the result matrix of the prediction after normalization. $NPVI_e(x) = [a1, a2, ..., a_i]$ is a row of this matrix that correspond to specific emotionol value $e$. Like mentioned above, this field results from the division of $PVI$ by the estimated positions $\theta s(x)$ without multiplying them by the emotional value.

$$a_i = \frac{PVI_e(x_i)}{\sum_{g=1}^{G} \theta_{sg}(x_i) + A} \tag{14}$$

$PVI_e(x)$ is the integration field that allows estimation of the position of the object in the image for a specific emotional value $e$. It is obtained by the sum of $\beta(e,x)$ for each emotional category. $G$ is the number of local view. $A$ is a normalization constant which makes possible to avoid attributing too great importance to the very small number of thumbnails. Then we will apply the competition mechanism see eq.15.

$$C_e(x) = NPVI_e(x) - NPVI_\alpha(x) \; where: \tag{15}$$
$$NPVI_\alpha(x) = max[NPVI(\alpha,x)] \; and \; e \neq \alpha$$

$$C(e,x) = \sum_{e=1}^{E} \delta_e(x)^T.C_e(x) \tag{16}$$

$$\delta_e(x) = [a_{e1}, a_{e2}, ....a_{eE}], a_{ex} = 1 \; when \; e = x$$

$C_e(x)$ is the competition between activities for each emotional value. $NPVI_\alpha(x)$ is the winner row where is the maximum activity, $\alpha \in E$ correspond to the winner emotional value for each competition $C_e(x)$. The $C(e,x)$ represent the matrix resulting from all competition. and $e \in E$ is the emotional value. Thus, the object with the highest emotional value is selected independently of the extracted local views.

### D. The movement of Berenson in the museum

The system for object tracking is inspired from [10]. The robot must also avoid obstacles but turn softly and give the impression that it moves from one artwork to another. In the visual system 15 frames are processed each second (which correspond to the 15 focus points extracted) and the motor command is controlled at a frequency of 10 Hz (limitation of the Robulab platform). Berenson's linear speed is bounded by a constant value (0.3 m/s) corresponding to the speed of a visitor walking through the museum and is reduced depending on the activity of its sensors. Moreover,

we must merge several behaviors within certain priorities, such as artwork appreciation, manual control and obstacles avoidances are in competition to determine robot direction. A dynamical neural field [15] allows merging of the motor commands coming from all behaviors, see Fig. 5. This gives soft transitions between behaviors and a smooth movements. Moreover, the hysteresis of the field allows robustness to intermitent detections. The robot speed is the value of the spatial derivative of the field at the position provided by the robot proprioception. This allows the robot to avoid unstable behavior when the levels of activities change. For instance, if a new bubble (new goal) appears on the field, the robot will remain in the direction of the first actived bubble until and only if this bubble disappears because of the lateral inhibition or because of the visual disappearance of the first target. Linear speed is calculated as follows :

$$s_j = V_{max} * (1 - max_{x \in [-\theta,\theta]}(A(x) * C(x))) \qquad (17)$$

where $x$ is the $x_{th}$ neuron of the field, $x \in [-\theta, \theta]$. In this experiment $\theta$ is equal to 180 degrees. Sensors values are added in the neural field where the index depends on the angle between the sensor and the front of the robot. $C(x)$ is a Gaussian kernel, centered at 0, which has to decrease the influence of out of center sensors. $A(x)$ is the activity of the $x_{th}$ neuron and is normalized in [0 ;1]. Rotation speed is computed as the derivative of the neural field activation (read-out mechanisms). We use the equation proposed by Amari [15] to compute the neural field activation :

$$\tau.\frac{u(x,T)}{dT} = -u(x,T) + I(x,T) + h + \int_{z \in V_x} w(z).g(u(x-z,T))dz \qquad (18)$$

where $I(x,T)$ are the inputs to the system provided by main behavior (route following and artwork appreciation), manual control and sensors activities. $\tau$ is the relaxation rate of the system. $w(z)$ is the interaction kernel in the neural field activation. These interactions are modeled by a Gaussian function. $g(u(x-z,T))$ is the activity of neuron $x$ according to the potential $u(x,T)$. We use a standard ramp function.

## III. Results

### A. Enhancement of coherent objects (normalization/competition)

After we added the mechanisms of normalization and competition, we wanted to test their efficiency on the object and valence recognition performance. We compare the new model (NM) with normalization and competition with the old model (OM) in which the competition principle is missing, to study how the addition of the concept of spatial competition improves the object recognition performance.

For this test, 13 objects are used, 10 images per object were learned and 100 images per object were chosen for the test Fig. 7. The test measures the success rate of object recognition but also the success rate of object position recognition (in pixels). The two histograms shows the performance difference between the system without OM and with the competition NM by comparing how many objects were well recognized and what is the position estimation error (between 1 pixels and 320 pixels, the width of our image) Fig. 8. From 1300 test images, the object is well recognised in 870 images, we accept a merge of error of 30 pixels (10% error) to consider an object well located, for the model with competition NM objects are well located in 696 images, corresponding to $80\%$ of total images where the object is recognized, for the old model OM without competition, the objects are well located in 604 images, corresponding to $69.4\%$ of total images where the objects is recognized. The results show a higher success rate of object recognition for the model with competition, and a lower error of recognition of the position of the object. Table. II shows the mean of position estimating error is about 25 pixels with competition, and 34.5 pixels without competition. This shows the improvement of object recognition when the competition is added to the N.N. model. The standard deviation (sd) of position estimating error is 37.4 pixels with competition and 46.4 pixels without competition. This means we must be careful with the ranking of the error in position since the sd is hight as compared to the mean values. Yet the fact that the sd is clearly lower when using the competition show the important improvement we observe in practice (the high variance of the numerical results is related to the objects which are badly recognized and not only to the presence of distractors).

Fig. 7. TOP : Learning mode, local views associated with the emotional value, circles represents valence. BOTTOM : Test mode, Berenson associates the recognized object to the given emotional value.

| | with competition | without competition |
|---|---|---|
| mean of distances | 25 | 29.83 |
| Standard deviation | 37.4 | 46.4 |

TABLE II

Recognition performance with and without competition mechanism

### B. Results of the survey

For the second experiment in the museum and after giving time to the visitors to observe and interact with the robot, we invited the visitors to answer a survey. The participants rated the overall quality of the robot interaction. The majority of visitors (according to the answers on our survey) felt curiosity and surprise when they saw the robot during their museum visit. Here we should mention that the majority of the visitors (70% of the museum visitors) indicate it was the

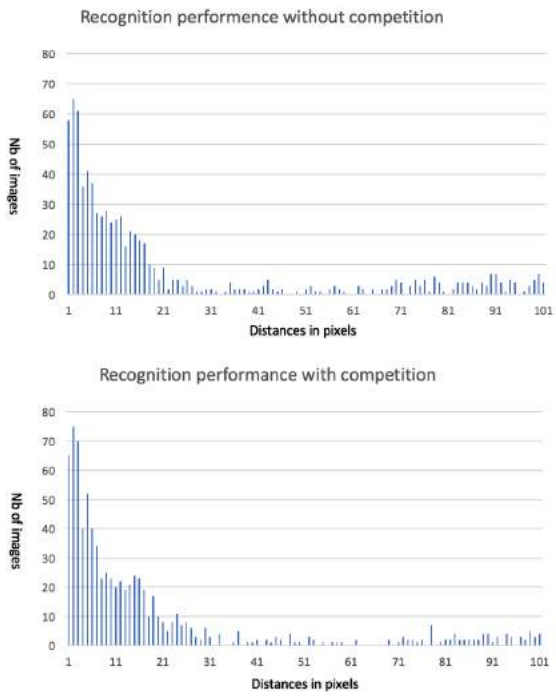


Fig. 8. TOP : The histogram represents the number of images (in which the object was well recognized) according to the distance between the correct position and the (answer) position) without competition. BOTTOM : The histogram represents the number of images (in which the object was well recognized) according to the distance between the correct position and the (answer) position) with competition.

| Answers | AF | | CF | |
|---|---|---|---|---|
| | IP | UP | IP | UP |
| **Never** | 1 | 2 | 0,05 | 0,1 |
| **Rarely** | 1 | 0 | 0,05 | 0 |
| **Sometimes** | 3 | 10 | 0,17 | 0,5 |
| **Often** | 8 | 4 | 0,47 | 0,2 |
| **Always** | 4 | 4 | 0,23 | 0,2 |

TABLE III

IP : INFORMED POPULATION, UP : UNINFORMED POPULATION, AF : ABSOLUTE FREQUENCY, CF : CUMULATIVE FREQUENCY

first time they met a robot. We tried to ask some questions about the consistency in the robot behavior and if the visitors could understand what Berenson was doing but the questions were problematic for the subject : what is the consistency for a robot behavior (for a naive subject) ? Yet looking and speaking about the interaction with Berenson was easy for the majority of the visitors. We ended with the following set of questions where answers is a 5 value scale (Never, Rarely, Sometimes, Often, Always) : Q1 : Is Berenson guided by what it sees ? Q2 : Did Berenson's behavior seem coherent ? Q3 : Did the interaction with Berenson seem easy to you ? Q4 : Did you understand what Berenson was doing ? Q5 : In your opinion, does Berenson react on its own ? (Table.II represent answers to Q1).

This experiment took place at the Quai Branly Museum, in the Insulinde (Insular South-est Asia) area art display, comparing two visitor populations $P_1 = 17$ and $P_2 = 20$. The $P_1$ population was informed about the robot and how it

| | Q1 | Q2 | Q3 | Q4 | Q5 |
|---|---|---|---|---|---|
| PV | 0.3301 | 0.722 | 0.8902 | 0.0076 | 0.5947 |
| | P1/P2 | P1/P2 | P1/P2 | P1/P2 | P1/P2 |
| Mean | 3.7/3.4 | 3.6/3.5 | 3.4/3.5 | 3.7/2.4 | 3.2/3.0 |
| SD | 1.0/1.1 | 0.9/1.1 | 1.0/1.3 | 1.1/1.6 | 1.3/1.3 |
| SEM | 0.2/0.2 | 0.2/0.2 | 0.2/0.3 | 0.2/0.3 | 0.3/0.3 |
| N | 17/20 | 16/20 | 17/19 | 17/20 | 17/20 |

TABLE IV

SURVEY STATISTICAL ANALYSIS,PV :P-VALUE

works before letting them observe and respond to our survey. By contrast the population $P_2$ was completly naive (uninformed). We gave them our survey without any explanation, and therefore the members of this population relied solely on their observation to answer our questions, which allows us to judge how readable Berenson's behavior was. The two-tailed P value from the Table.III for the Q1, Q2, Q3, Q5 show by conventional criteria, that this difference between the two populations is considered to be not statistically significant. This means that without our explanations the naives (uninformed) visitors manage more or less to read the behavior of our robot. But when the question become more specific (Did you understand what Berenson was doing ?) The naive visitors hesitated a lot and were not able to answer. The P-value shows this results P=0.0076 for the question number 4 showing that the difference between informed and not informed population is considered to be very significant statistically.

### C. Readability of Berenson behavior

The third experiment took place at the same museum in the PERSONA exposition one year later. We wanted to know at what point the movement of our robot is informative. What are the advantages and disadvantages of its way to navigate in the museum ? Could visitors understand the intentions of our robot or at least its artistic preference ? This time we wanted to test the readability of Berenson's behavior. The facial expressions give already strong indications. Also, the choice of objects towards which to head should enrich the behavior and make it more readable. A map with the objects learned by the robot was provided to the visitors. They were asked to note on this map one or several of these objects if they noticed that Berenson was interested in some objects and to add information if Berenson likes or dislikes these objects. We had 28 objects in our working area that were learned and associated to an emotional value (positive or negative). The majority of the objects (18 objects) have been quoted very few times with a high rate of error on the emotional reaction of Berenson. At the opposite, some objects (11 objects) were cited many times and with a high success rate (70% of success) for the recognition of Berenson appreciation. This shows that on certain object visitors succeed to read the behavior of our robot correctly see Table.IV. There is a clear link between the incorrect evaluation of the subjects and the over-generalization of Berenson on the same objects. The limitation is related to its recognition performance for objects that can be easily confused with other objects. When the point of view is changed from the one used for learning the object can be confused with another object. Learning

| N.of | objects | success | observation | success/observed |
|---|---|---|---|---|
| OM $< 3$ | 18 | 12 | 21 | 0,57 |
| OM $\geqslant 3$ | 11 | 46 | 60 | 0.76 |

TABLE V

FREQUENCY OF SPECIAL CHARACTERS, OM :OBJECT MENTIONED

should be improved or the robot should indicate the risk of confusion on some way (in the network when a local view is ambiguous several neurons associated to the recognition of the looking like learned views have almost the same level of activation).

## IV. CONCLUSION AND DISCUSSION

We showed that the system can generalize on the data that it does not learn, which is a very interesting property. It was necessary to have the compromise between generalization and discrimination. With the vigilance parameter that allows us to filter in low-level visual stimuli and to filter out strong generalizations, we added the mechanism of normalization and competition between categories. Berenson's way of navigation in museum is influenced by the object winner, so by the spacial computation mechanism added to the model. Once Berenson learns the association between an artwork and a positive/negative emotional value, it will express the facial expression according to what it sees in its visual fields, and will move in a preferred way by going to the objects it appreciates according to its learning. It will control its direction to reach positive artworks. This behavior seem to be clear and readable generally. From the last experiment we can summarize that the behavior of the robot is more clear for some learned object than others. We deduce that a possible similarity between the objects learned with different emotional value reduces the recognition rate of each one in the test phase. The more singular objects (that share fewer visual features with the others) retain a high level of recognition. Then the robot goes more often to these objects, the visitors notice it more, and they understand more easily the robot's behavior. We can also find another cause to explain this result, but in any case, this hypothesis remains to be verified. The spot was really hard because the object was learned very few times and then the robot was asked to generalize enormously. We can conclude therefore that the way of the robot navigation based on the neural field is informative. The dynamic neural field and the normalization allow to choose the right object and keep it during the approach, and the robot expressions associated with the correct value shows well the robot preference. The motricity (direction of choice of the object) and the expressivity (expression in front of the objects) of the robot is almost readable, and the visitors come to understand the robot's preferences many times. But the robot doesn't stop in front of its preferred object for some duration because of obstacle avoidance mechanism. This could also affect the readability of Berenson's behavior negatively. For that, we will add a mechanism that will stop the robot in front of artworks (the time to habituate to the visual scene) and enhance its behavior readability.

## REFERENCES


[1] Burgard, W., Cremers, A. B., Fox, D., Hahnel, D., Lakemeyer, G., Schulz, D., ... & Thrun, S. (1998, July). The interactive museum tour-guide robot. In Aaai/iaai (pp. 11-18).

[2] Faber, F., Bennewitz, M., Eppner, C., Gorog, A., Gonsior, C., Joho, D., ... & Behnke, S. (2009, September). The humanoid museum tour guide Robotinho. In Robot and Human Interactive Communication, 2009. RO-MAN 2009. The 18th IEEE International Symposium on (pp. 891-896). IEEE.

[3] Chella, A., & Macaluso, I. (2006, October). Sensations and perceptions in Cicerobot, a museum guide robot. In Brain Inspired Cognitive Systems Conference (BICS, 2006).

[4] Ekman, P., & Friesen, W. Facial action coding system : a technique for the measurement of facial movement. 1978. Consulting Psychologists, San Francisco.

[5] Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2002). An efficient k-means clustering algorithm : Analysis and implementation. IEEE transactions on pattern analysis and machine intelligence, 24(7), 881-892.

[6] Widrow, B., & Hoff, M. E. (1960, August). Adaptive switching circuits. In IRE WESCON convention record (Vol. 4, No. 1, pp. 96-104).

[7] Boucenna, S., Gaussier, P., Andry, P., & Hafemeister, L. (2010, October). Imitation as a communication tool for online facial expression learning and recognition. In Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on (pp. 5323-5328). IEEE.

[8] A. de Rengerve, S. Boucenna, P. Andry, and P. Gaussier, ?Emergent Imitative behavior on a Robotic Arm Based on Visuo-Motor Associative Memories, ? in IEEE/RSJ International Conference on Intelligent Robots and systems (IROS10), Taipei, Taiwan, October 2010, pp. 175-1759.

[9] Quoy, M., Moga, S., & Gaussier, P. (2003). Dynamical neural networks for planning and low-level robot control. IEEE Transactions on Systems, Man, and Cybernetics-Part A : Systems and Humans, 33(4), 523-532.

[10] Lepretre, S., Gaussier, P., & Cocquerez, J. P. (2000). From navigation to active object recognition.

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

[12] Ungerleider, L. G., & Haxby, J. V. (1994). What and where in the human brain. Current opinion in neurobiology, 4(2), 157-165.

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

[14] Widrow, B., & Hoff, M. E. (1960, August). Adaptive switching circuits. In IRE WESCON convention record (Vol. 4, No. 1, pp. 96-104).

[15] Amari, S. I. (1977). Dynamics of pattern formation in lateral-inhibition type neural fields. Biological cybernetics, 27(2), 77-87.

[16] Boucenna, S., Gaussier, P., & Hafemeister, L. (2014). Development of first social referencing skills : Emotional interaction as a way to regulate robot behavior. IEEE Transactions on Autonomous Mental Development, 6(1), 42-55.