Contingency Perception and Agency Measure in Visuo-Motor Spiking Neural Networks

Alexandre Pitti, Hiroki Mori, Shingo Kouzuma, and Yasuo Kuniyoshi, *Member, IEEE*

***Abstract*—Agency is the sense that I am the cause or author of a movement. Babies develop early this feeling by perceiving the contingency between afferent (sensor) and efferent (motor) information. A comparator model is hypothesized to be associated with many brain regions to monitor and simulate the concordance between self-produced actions and their consequences. In this paper, we propose that the biological mechanism of spike timing-dependent plasticity, that synchronizes the neural dynamics almost everywhere in the central nervous system, constitutes the perfect algorithm to detect contingency in sensorimotor networks. The coherence or the dissonance in the sensorimotor information flow imparts then the agency level. In a head-neck-eyes robot, we replicate three developmental experiments illustrating how particular perceptual experiences can modulate the overall level of agency inside the system; i.e., 1) by adding a delay between proprioceptive and visual feedback information, 2) by facing a mirror, and 3) a person. We show that the system learns to discriminate animated objects (self-image and other persons) from other type of stimuli. This suggests a basic stage representing the self in relation to others from low-level sensorimotor processes. We discuss then the relevance of our findings with neurobiological evidences and development psychological observations for developmental robots.**

***Index Terms*—Contingency detection, self-agency, sensorimotor integration, spiking neural networks.**

I. Introduction

THE arising of self-agency and body-ownership constitutes undoubtedly one of the most important stage of infants development. The sense of agency corresponds to the prereflective experience or sense that I am the cause or author of a movement (e.g., an experience that I am in control of my action [1]–[3]). Babies, early in their first months, acquire rapidly this sense of self distinct from others that allows them to develop later on higher cognitive skills such as social interactions and imitation. For many developmental roboticists, this issue is considered to be central for replicating infants developmental stages in robots and to permit them to apprehend themselves as distinct entities (e.g., for interacting with people); see [4]–[7]. To this aim, computational models should be confronted to biological and psychological data. Many evidences suggest that ownership and agency are perceptual experiences that are likely to be generated by 1) low-level sensorimotor processes and 2) contingency perception.

For instance, Gibson suggested that this stage corresponds to the period when infants are caught themselves in the act and "co-perceive" the world in the same time as the perceiver and actor in their environment [8]—emphasizing perceptuo-motor integration. During this period, babies construct a mental representation of their body dynamics in the central nervous system—a body image, whereby they expand their sensorimotor capabilities of interaction within the environment (i.e., acquisition of the first affordances [9]).

Other proponents, on the other hand, insist on babies' ability to sense contingency between different modalities (the timing between events) as an important paradigm to bring forth self-awareness (see [10]–[14]): experiencing sensorimotor contingency makes them to perceive the agency of their own body, whereas lags in sensorimotor information flow generate conflicts that disrupt this feeling. Watson hypothesized that this skill is owed by contingency detectors in the central nervous system (acquired or ad hoc) dedicated to learning and to reasoning about temporal events during exploration of their own body and during their early interactions with the environment. This corresponds to a first stage of intermodal calibration where the babies explore *systematically* their body (e.g., self-touch, predominance to look at their hand) to define the stable and robust cross-modal causal relations.

In line with Watson, Rochat envisions it as the primordial bootstrap to the acquisition of the sense of self-awareness (see [14] and [15]). Babies calibrate their proprioceptive information to define a mental representation of their own body distinct from other individuals by detecting the causal effects between the self-produced contingent activity of their own actions and the induced sensory information (e.g., spatial position, somatosensory or visual or sonorous information). From this stage, they soon distinguish the *even* of integrated proprioceptive information from the *odds* of nonstrict contingent information. This permits them to *extract* themselves from the external environment—for instance, to recognize themselves in front of a mirror or from a live video projection [16]. The disrupting of sensorimotor synchrony as little as a few hundreds of milliseconds delay, in contrast, alters or even destroys the feeling of self-perception (see [15] and [17]).

Recent findings in cognitive neuroscience comfort these results and reveal the importance of the temporoparietal cortex, premotor cortex, insula, and primary somatosensory cortex for agency and ownership (see [1] and [2]); abnormal sensorimotor and/or multisensory processing due to damages in these areas

Manuscript received April 19, 2005; revised January 11, 2007. First published April 21, 2009; current version published May 29, 2009. This work was supported by the JST Asada ERATO Synergistic project.

The authors are with the Department of Mechano-Informatics, Graduate School of Information Science and Technology, University of Tokyo, Tokyo, Japan (e-mail: alex@jeap.org).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TAMD.2009.2021506

1943-0604/$25.00 © 2009 IEEE

reveal modulation in performance and experience of agency and ownership. Mainly two neurocomputational theories of agency are advanced [18]: the comparator model, which emphasizes the functional role of the motor system (for comparing reafference), and simulation theory, which assumes shared representations of self and others. Taken together, these considerations suggest to us that understanding the neural mechanisms underlying timing integration in sensorimotor networks for a biological system exposed to multimodal sensory information is a primordial step toward the understanding of self-agency. Modeling self-agency in robots may permit then to understand how the further developmental stages, more complex, might arise in infants.

From a biological viewpoint, the mechanism responsible for neural communication and information propagation in the central nervous system is the one of spike timing-dependent synaptic plasticity (STDP); see [19]–[22]. It regulates the spike timing delays between the neural pairs and sustains the phasic (temporal) information processing for both encoding and retrieving tasks such that neurons that fire contingently wire together. Within a network, the many neural pairs constituted can assemble and aggregate themselves into consistent long-range spatiotemporal clusters [23], [24]. In a previous paper, we suggested that these clusters can encode actions sequences and represent action primitives in cross-modal networks [25]. In this paper, we further develop our study and propose that this mechanism of STDP can constitute a biologically plausible model for detecting contingency between multimodal events and permits, for an agent, to "experience" its own agency during motion in line with developmental and neurocomputational theories; these two works pursue some preliminary works in which we first addressed the agency problem in the context of dynamical systems exploiting the mechanism of phase synchronization [26]. For babies, it means to learn the temporal rules existing between afferent and efferent information (sensorimotor integration), which can produce then representations of their own body and of their own actions [27], [28].

In experiments with a head-neck-eyes vision system, we investigate how such primary skill can emerge within sensorimotor networks. We show that the embodied system self-drives its dynamics through the continuous interplay with the environment. Without any other bias except the one to move, the system generates sensorimotor information, focusing and shifting its attention to salient objects. Over time, it explores its parameters' space in a self-organized manner and learns the simple sensorimotor causal relations associated to its own motion (e.g., turning the head in direction of salient objects). It follows the emergence of sensorimotor links (contingency detectors), which anticipate and estimate within milliseconds order in advance the next motor status and visual response. These short *scripts* synchronize dynamically the sensorimotor maps and maintain the overall integrity of their dynamics. The level of this global coherency provides then a quantitative measure of the system's agency. Beside the normal case of live enaction, we show that other types of perceptual experiences can modulate differently its level such as delaying the visual and proprioceptive information, scrutinizing its own reflection in front of a mirror, or observing someone else, each corresponding to different sensorimotor coordinations.

This paper is organized as follows. In the first part, we present the head-neck-eyes robotic device used in our experiments, the visual processing we are relying on, and the neural architecture that represents the system's self-produced visuo-motor information. We conduct four experiments replicated from developmental psychology to reveal the system's agency (see [11], [15], and [17]). In the first experiment, we explain how the neural system acquires simple sensorimotor primitives by interacting with the environment. Over time, the system learns to anticipate the saccades and the foveations with the specific temporal relationship associated between the vision and the motor events. We define then a measure of one agent's agency based on information retrieval theory [29] to calculate the distance between the predicted sensorimotor state and the realized ones. Then, we explain how different perceptual experiences can affect dynamically the system's behavior and its associated agency level. For instance, delaying the visual information disrupts the sensorimotor coordination flow, whereas fixating its own reflection in front of a mirror or facing another person produces a strong global entrainment in the neural dynamics, which modulates the overall agency level. Based on our results, we discuss the possible relevance of our findings for rising agency and self-perception limited to the here and now in infants and its link to the further development stages involving self–other distinction and social interactions.

II. Framework

We present the characteristics of our head-like device, its visual system, and its embedded neural system. The network processes the robot's intensity-based visual information and motor information into spike trains and controls back its orientation.

A. Description of the Robot

Our robot is a camera-based device aimed to replicate the basic kinematics of the human head-neck-eyes system (see Fig. 1). It is composed of two cameras that pan (move left and right) with relative angles $\{\theta_{\text{left}}, \theta_{\text{right}}\}$ on their own pan axis and mounted on a common platform that can also rotate in the vertical axis with absolute angle θ_{neck}. Three servo-motors control the respective kinematics of the parameter set $\{\theta_{\text{neck}}, \theta_{\text{left}}, \theta_{\text{right}}\}$ limited between $[-\pi/6; +\pi/6]$. The angles $\{\theta_{\text{left}}, \theta_{\text{right}}\}$ are joint variables. We set their values with the formula $\theta_{\text{eyes}} = (\theta_{\text{left}} + \theta_{\text{right}})/2$ such that they fixate the same focal point. As a result, the overall system complexity can be reduced to two dimensions $\{\theta_{\text{neck}}, \theta_{\text{eyes}}\}$: this system, although simple, is sufficient to study agency arising in agents and sensory-motor integration.

B. Visual System

Infants are not passively exposed to multimodal sensory information. By moving their body, they actively structure their sensory input and generate statistical regularities (e.g., it has been found that neonates rhythmically agitate their limbs to acquire the general movements). In turn, these regularities enable appropriate perceptual experiences that can produce certain developmental changes in learning (see, [30]–[32]). For instance, the morphological and visual processing done in the retina permits better representing this information to ease skills

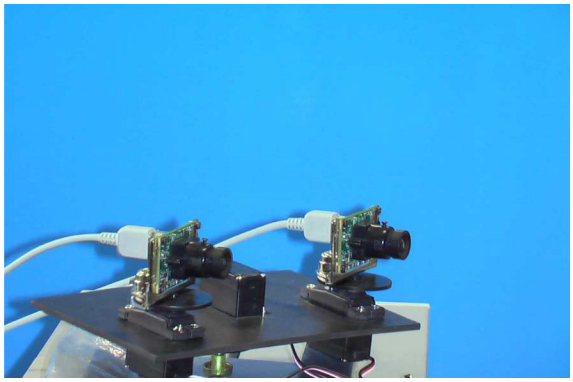

Fig. 1. Photo of our head-neck-eyes device (three degrees of freedom). Our purpose is to induce sensorimotor coordination and contingency perception of salient objects in the scene during motion. Its architecture, though simple, will permit us to study the system's visuo-motor (mis)matching and its agency.

like foveating and attention to others (see [33] and [34]). The particular log-polar topology of the retina involves eye movements to center the objects of interest in the fovea, where receptors are denser (and the information more detailed), rather than in the periphery. It produces a first "morphological computation" of the visual attention by decaying exponentially information from the fovea (see [35] and [36]). Moreover, the successive transformations done in the ganglions layers realize a complex filtering on the images into a set of spikes, which permits discriminating the salient information from a scene; the so-called saliency maps (see [37] and [38]). Four principles guide the Itti–Koch model.

1) Visual attention is based on multiple features.
2) The saliency of a region is affected by the surrounding context (conspicuity).
3) The saliency of locations is represented by a saliency map.
4) The winner-take-all and inhibition of return are suitable mechanisms to allow attentional shift [39].

In our experiments, we limit the modeling of visual attention by computing the saliency map from the image's intensity, the color opponency filters, and the relative motion gradient features; we neglect the orientation, winner-take-all, and inhibition of return from the original Itti–Koch model.

The different stages of the visual processes are done as follows. The incoming RGB visual information is first filtered into three feature maps to discriminate the color-based spatiotemporal patterns. Their respective center-surround maps are then combined into a unique saliency map transformed next into log-polar coordinates in order to emulate the foveal vision. We first extract the intensity I and the two color opponency filters from the image by applying the classic formula $I = 0.3 * R + 0.59 * G + 0.11 * B$ and by separating then the color information R^+G^- and B^+Y^- ("yellow") according to $(R-G)/I$ and $(B-Y)/I$.

Next, we extract the relative motion gradient. Motion detection is a very salient feature. When the head moves, it generates lot of information aligned in the direction of displacement. This global flow becomes the referent gradient so that an object moving in the opposite direction becomes then the most salient object in the scene: its relative saliency is inversely proportional to the global motion orientation. To compute its saliency, we first calculate the motion gradient of the derivatives $\mathtt{Mx}$ and $\mathtt{My}$ from two consecutive intensity-based images $I(x, y, t-\Delta t)$ and $I(x, y, t)$

$$\begin{aligned}\mathtt{Mx}(x,y,t) =& I(x,y,t-\Delta t)\cdot I(x+\Delta x,y,t)\\ &- I(x,y,t)\cdot I(x+\Delta x,y,t-\Delta t)\\ \mathtt{My}(x,y,t) =& I(x,y,t-\Delta t)\cdot I(x,y+\Delta y,t)\\ &- I(x,y,t)\cdot I(x,y+\Delta y,t-\Delta t).\end{aligned} \tag{1}$$

The gradient orientation φ is then assigned from the formula

$$\varphi = \begin{cases} 0, & \mathtt{Mx}=0, \mathtt{My}=0 \\ \arctan\left(\dfrac{\mathtt{My}}{\mathtt{Mx}}\right), & \text{else.}\end{cases} \tag{2}$$

Then, from the orientation histogram built, the global orientation angle φ_{global} is calculated from the histogram maximum. A possible measure of one object saliency is to compute the angle difference between φ located at (x, y), $\varphi(x, y)$, and the referent motion orientation φ_{global} normalized between $[0:1] : M(x, y, t) = (1 + \cos(\varphi(x, y) + \varphi_{\text{global}}))/2$.

The three feature maps processed, we transform them into their respective conspicuity maps, which highlight the parts of the scene that strongly differ from their surroundings depending on that feature. Practically, this is achieved by using a center-surround mechanism (a difference of Gaussians filters). It is implemented as the difference between fine and coarse scales (implementation details in [37], [39], and [40]). The resulting conspicuity maps are combined into one saliency map weighted equally. The map with Cartesian coordinates (x, y) is finally transformed into the (ξ, ρ) log-polar coordinates

$$\begin{bmatrix}\xi\\ \rho\end{bmatrix} = \begin{bmatrix}\log\sqrt{x^2+y^2}\\ \arctan\dfrac{y}{x}\end{bmatrix}. \tag{3}$$

From there, the pixel output of the saliency map Z located at $\{\xi, \rho\}$ is normalized between [0,1] to provide the excitatory current distribution $I_i(t)$ associated to its ith neuron

$$I_i(t) = 20.0 * Z(\xi, \rho, t). \tag{4}$$

Fig. 2 summarizes the successive image processing performed from pixels to spikes.

C. Neuron Model

We define the neurons dynamics with the neuron model proposed by Izhikevich [41]

$$\begin{aligned} v' &= 0.04v^2 + 5v + 140 - u + I\\ u' &= a(bv - u)\end{aligned} \tag{5}$$

where I is the external input, with the auxiliary after-spike resetting

$$\text{if } v \geq +30 \text{ mV, then } \begin{cases} v \leftarrow c\\ u \leftarrow u + d\end{cases} \tag{6}$$

with v representing the membrane potential of the neuron, u a membrane recovery variable (see [41] and [42]), and v' and u' their temporal derivate. The variables set $\{a, b, c, d\}$ defines the neuron attributes—whether it is excitatory $[(a; b) = (0.02; 0.2)$

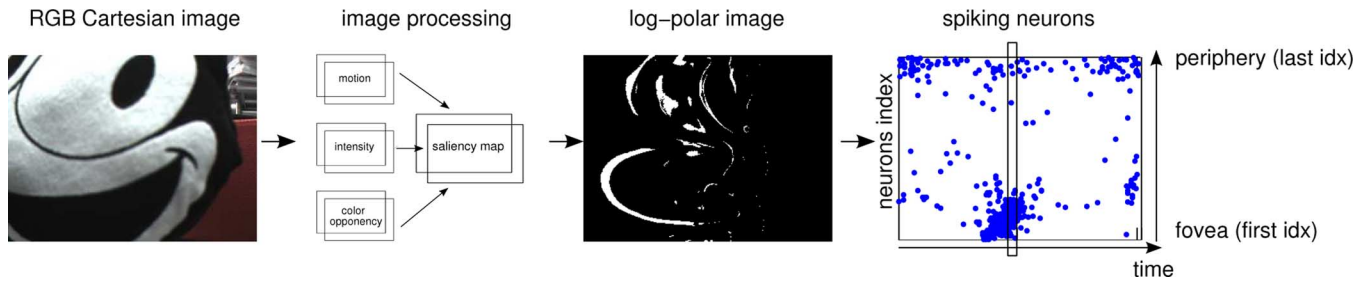

Fig. 2. From pixels to spikes. Intensity, color, and relative motion detection features are extracted from the Cartesian RGB images and filtered by a center-surround mechanism. The three resulting feature maps are combined then into a saliency map in log-polar coordinates so that the first indexed neurons of the vision map receive first the pixels information from the fovea and the last-ranked neurons those from the periphery.

and $(c; d) = (-65; 8)$] or inhibitory [$(a; b) = (0.02; 0.25)$ and $(c; d) = (-65; 2)$].

D. Neural Network Architecture

Four specific maps compose the entire neural network. They process separately the incoming information from the two cameras (visual input) and the motors (the neck and eyes' orientations). The motor maps deliver then the motor commands. The schematic in Fig. 3 describes the coupling between the neural system and the robot head. The two vision maps receive, respectively, the left and right camera visual information (640×480 pixels) subsampled to 80×60 resolution (4800 neurons) and filtered as described in Section II-B. The pixel's value of the retina layer Z located at $\{\xi, \rho\}$ provides the excitatory current distribution $I_i(t)$ to its associated neuron with index i such that $i = \rho * 80 + \xi$ [see (4)], transforming the two-dimensional map into a one-dimensional vector. It is noteworthy that the first-ranked neurons receive first the activity from the fovea, whereas the last-ranked neurons receive those from the periphery. The first neurons are therefore topologically biased to receive more information than the last ones. The two motor maps are discretized into the interval $[-(\pi/6); +(\pi/6)]$ in 256 bins, such that the ith neuron is associated to the angle a with the formula $i = 256 \cdot a(3/\pi) + 128$. The motor command is retrieved back from a winner-take-all-like mechanism, where the location of the spikes is denser (i_{dense}) with the inverse formula $\texttt{motor\_command} = (\pi/3)(i_{\text{dense}} - 128)/256$.

All the neurons within the four maps are excitatory. We add a hidden layer composed exclusively of inhibitory neurons to stabilize the overall activity. This layer is composed of 1000 neurons. For all the neurons, whether excitatory or inhibitory, we initialize the neurons with 100 synaptic links randomly selected so that the one-third correspond to local connections with spatially proximate neurons inside their maps and the other two-thirds correspond to long-distance links to other maps. The local connections between neural pairs $\{i, j\}$ follow a normal distribution centered to the neuron with index i and variance $\sigma = 50$ so that the neuron with index j is comprised between $[-150 + i, +150 + i]$. The distant neural pairs $\{i, j\}$, on the other hand, follow a uniform distribution within the entire network. The neural pairs within the same map are responsible for the intramap neural activity, and the neural pairs belonging to different maps are responsible for the intermap neural activity. We justify our choice to have an initial ratio within the network between specialized processing (intramap activity) and intermodal integration (intermap activity). Before learning, we initialize all the synaptic links with same weight $c = 5$ so that the network has no particular organization. We expect that during enaction the embodied network integrates the visuo-motor neural pairs and builds up a multimodal representation of its own actions.

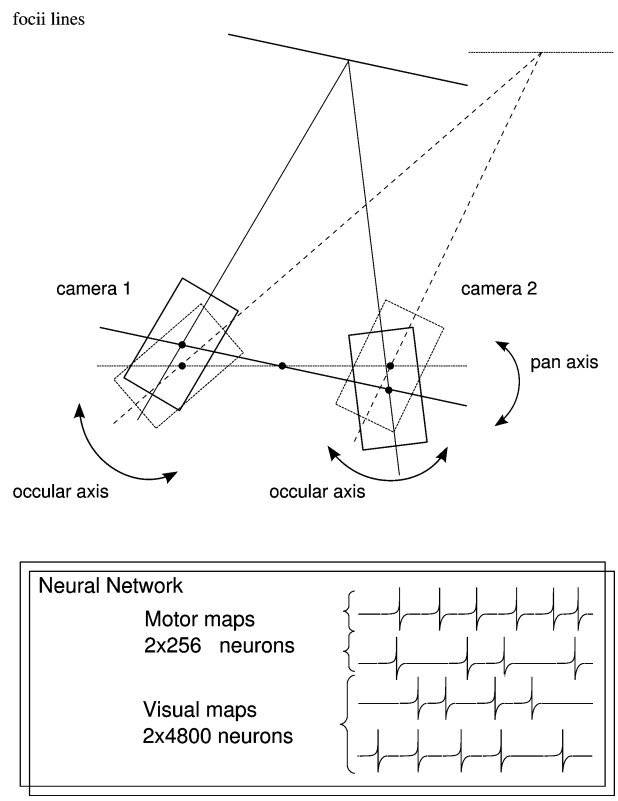

Fig. 3. Schematic of the head-neck-eyes vision system coupled to its neural system. The device is constituted of two cameras controlled jointly, converging to the focal point equidistant to the neck central axis. Two motors control each camera's relative angle, and a third one commands the orientation of the neck-like ensemble. Four maps of spiking neurons compose the neural network receiving the respective visual and motors signals and commanding back the motors.

E. Learning Mechanism

STDP is the bidirectional adaptation mechanism that dynamically regulates the long-term potentiation (LTP) and long-term depression (LTD) in synaptic plasticity, readjusting the synaptic weights to the precise timing interval between the initiating

and the targeting neurons (see [19]–[21]). They are significant mechanisms for both activity-dependent development of neural circuitry and adult memory storage. The time delay $\Delta t = t_{\text{post}} - t_{\text{pre}}$ between the presynaptic neuron spiking t_{pre} and the postsynaptic neuron firing t_{post} corresponds to the interval range of activation of their synaptic plasticity and weight adaptation Δc; see Fig. 4

$$c_{\text{pre,post}} = c_{\text{pre,post}} + \Delta c \tag{7}$$

$$\Delta c(\Delta t) = \begin{cases} A_+ \exp\left(\dfrac{\Delta t}{\tau_+}\right), & \text{if } \Delta t < 0 \\ -A_- \exp\left(\dfrac{-\Delta t}{\tau_-}\right), & \text{if } \Delta t \geq 0. \end{cases} \tag{8}$$

The synaptic weights decay exponentially depending on the time delay Δt between the pre- and postsynaptic neurons in the interval range $[\tau_-, \tau_+]$. Each time a postsynaptic neuron fires, its synaptic weights $c_{\text{pre,post}}$ are decreased by A_- (LTD), and each time a synapse receives an action potential, its synaptic weight $c_{\text{pre,post}}$ is incremented by an amount A_+ (LTP); we set $-A_- = A_+ = 1$ and $\tau_- = \tau_+ = 20$ ms in all our experiments. It follows that the contingent neurons strengthen their links, whereas the incongruent neurons weaken their ones (Hebb's law). Within the network, the neural pairs $\{i, j\}$ can be viewed as small conditional *scripts,* which can detect/encode the contingency at the local level

if neuron_i fires at time t_i

then neuron_j fires at time $t_j = t_i + \Delta t$.

This mechanism, although simple at the neurons' scale, can generate very complex dynamics as the neural pairs can aggregate themselves into long-range spatiotemporal clusters [e.g., Fig. 4(b)]; see [42] and [43]. In sensorimotor networks, we propose that these assembled spatiotemporal patterns constitute a repertoire of commands or action primitives as Wolpert conceives them (see [25] and [43]). They represent internal models, which are the building blocks used to construct intricate motor behaviors with an enormous range. In our experiments, the (mis)match between the online visuo-motor information with one of these rules will correspond to the level of sensorimotor coordination and, thus, to the perception of agency associated with it (explanations thereinafter).

F. Agency Index

According to Tsakiris, the coherent experience of the body depends on the integration of efferent information (motor) with afferent information (proprioception) in action contexts [2]. The sense of agency is generated by or at least linked to the motor commands sent to the muscles and the accompanying efference copy that is internally processed within the predictive models of the motor system [27]. A possible quantitative measure of agency is then to compute the accuracy between the afferent visual and proprioceptive information and the efferent predictions using the F-measure in information retrieval theory [29]; another measure of the neural dynamics coherency could have been the synchrony index used in [24]. In our framework, this situation occurs when a presynaptic neuron #1 and an input stimulus #2 activate both at the same time their common neuron

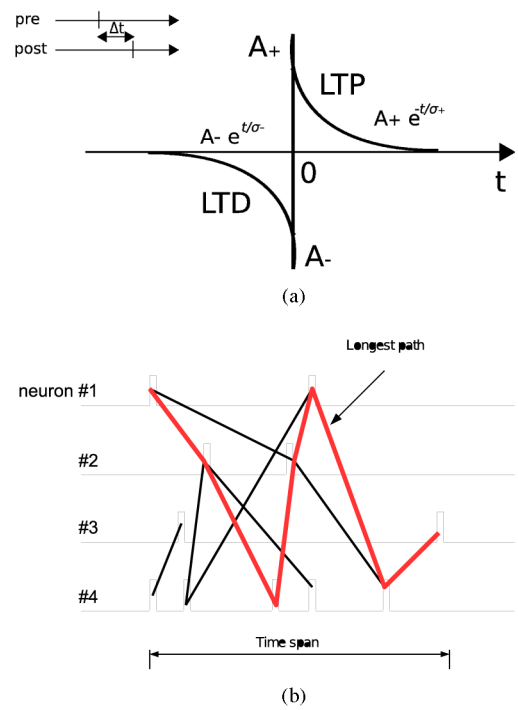

Fig. 4. Mechanism of STDP with $-A_+ = A_- = +1$ and $\tau_- = \tau_+ = 20$ ms. (a) Each time a post-synaptic neuron fires, its synaptic weights are decreased by A_-, and each time a synapse receives an action potential, its synaptic weight is incremented by an amount A_+. (b) Based on this mechanism, different neural pairs can assemble themselves into asynchronous neuronal groups (polychronized groups; see [23]).

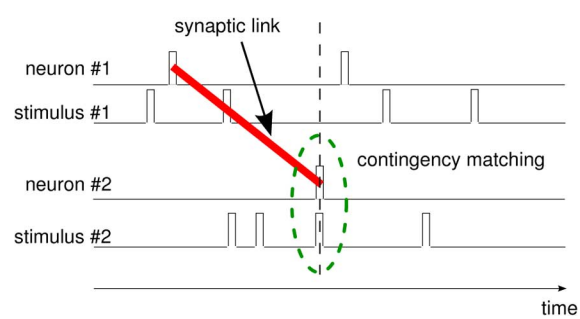

Fig. 5. Contingency detection. The sense of agency arises from the matching between afferent (proprioception) and efferent information (motor prediction). In our model, it corresponds to the synchronization between the presynaptic neuron #1 and the incoming input stimulus #2 that activate in the same time neuron #2. To fire contingently with the input, the presynaptic neuron #1 must trigger in advance.

#2 (i.e., contingency matching); see Fig. 5. The agency index at current time t is therefore computed as the number of correctly predicted stimuli divided by the number of all returned predictions done at time t, either correct or false (the "precision"), and the number of correctly predicted stimuli divided by the number of predictions that should have been returned (the "recall")

$$\text{precision} = \frac{\text{nb\_correctly\_predicted}}{\text{nb\_predicted}}$$

$$\text{recall} = \frac{\text{nb\_correctly\_predicted}}{\text{size\_current\_input}}$$

$$\text{agency\_idx} = \frac{2.\text{precision.recall}}{(\text{precision} + \text{recall})}$$

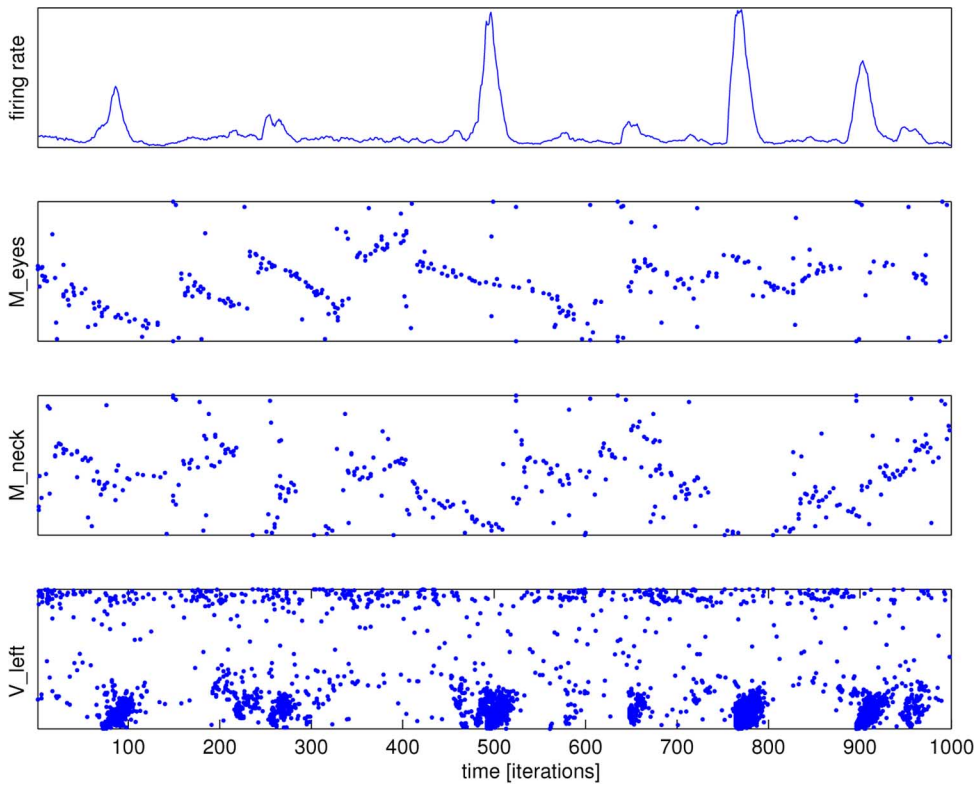

Fig. 6. Vision and motor neural activity during enaction. The embodied system generates a spontaneous activity through the interaction with the environment dynamics switching between foveation and attentional shift (vision map), saccades, and long spike trains (in the motor maps).

where the size_current_input corresponds to the size of the current input pixels values binarized (size_current_input $\in [0; 2 \times 4800]$). In this form, the agency index is normalized such that the lowest values (e.g., agency_idx $= 0$) correspond to complete mismatch between sensory inputs and motor predictions (no agency), whereas the highest ones correspond to perfect contingency prediction (maximum level of agency).

III. Experiments

We propose to study how agency arises in visuo-motor networks for different perceptual experiences. During enaction, sensorimotor maps integrate their dynamics and learn simple perceptuo-motor patterns between the current motors state (eyes and neck orientation) and the visual scene. We show that the pattern (mis)matching between the neural activity and the incoming information flow modulates its agency level. Moreover, other perceptual experiences can affect the system's agency level in various fashions, e.g., when a visual feedback delay is imposed or when it scrutinizes its own reflection in front of a mirror (perfect synchrony and maximum relative motion saliency) or someone else. We reproduce these three experiments and study the possible relations entangling agency, self-recognition, and self–other distinction.

A. Sensorimotor Primitives Learning

Before enaction, the network's architecture has no specific organization. The neural network is initialized with local and global couplings, and the node connectivity has a uniform distribution (equal synaptic weights); see Section II-E. The connected neurons inside the same map sustain the intra-map information processing, whereas the connected neurons belonging to different maps support the intermap information transmission. We expect the sensorimotor maps to self-organize under the action of STDP and of its embodiment (i.e., the log-polar representation combined with the saliency maps). We place the head-neck-eyes device in the normal daily-life conditions—enclosed and enlightened environment with salient objects and quite dynamic persons (the laboratory room and its researchers)—and start our experiment in this situation without prior knowledge or supervised learning.

During enaction, the incoming signals starting with the initial motors and visual inputs proceed to the visuo-motor maps. The first spikes ignite then their respective associated postsynaptic neurons, whether located inside or outside the originating map. These neurons, at their turn, trigger others, affecting either the vision maps or the motor maps; thus the motor commands. At this point, the new state of the motor commands changes the neck and the cameras' orientation, and a new cycle can begin with this new configuration. Gradually, a spontaneous activity bootstraps in the whole network in closed loop with the environment. The device dynamically switches from one orientation to another, punctuated by short-range attentional fixation (see Fig. 6). The long spike trains in the motor maps correspond to fixation to certain locations (same motor angles), whereas the bifurcations correspond to jumps to new locations (new motor angles). Saccades to new locations in the motor maps happen almost in phase (at same timing). If salient information is found in the fovea where the neurons are concentrated, a burst of spikes

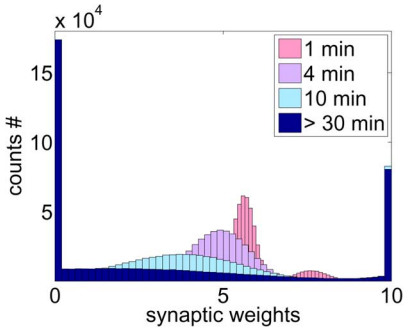

Fig. 7. Snapshots for different periods of time of the network synaptic weight distribution. The neural network achieves its self-organization by strengthening the most robust neural groups (weights' value > 9) and deleting the inaccurate ones (weights' values < 1).

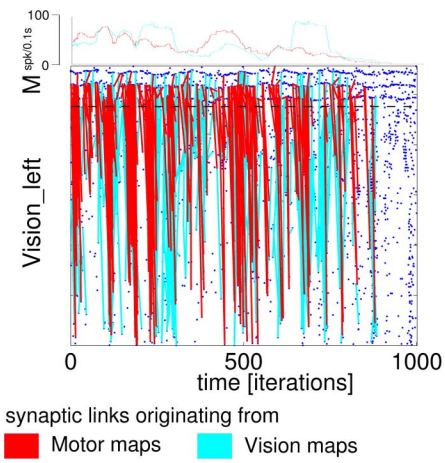

Fig. 8. Superposed synaptic links on the neural dynamics. The most robust anatomical synaptic links, extracted from the connection matrix $(c > 9)$, are sur-imposed between the spikes of presynaptic and postsynaptic neurons [same as Fig. 4(b)]. The links having their presynaptic neurons inside the vision map (left vision map) are plotted in red. Those with presynaptic neurons from the motor maps (neck and eye maps) are plotted in cyan; above, their respective rate (number of links per 100 ms).

then activates the low-ranked index neurons. In the other case, when the salient information is located in the cameras periphery, it is this time the few highly ranked neurons that trigger.

During this random-walk stage, the embodied system rapidly learns to differentiate the visuo-motor patterns associated to salient objects in the fovea (corresponding to high firing rate) and the other stimuli (i.e., salient objects not centered from the fovea having a low firing rate). Via STDP, the causal relationships at the neural level between the visions and the motor maps are reinforced, whereas the uncorrelated neurons weakened their links. The snapshots at different period of the synaptic weights histogram in Fig. 7 describe this evolution. After 10 min, most of the neural links are strengthened $(c > 9)$ or suppressed $(c < 1)$. In that situation, the system does not freeze to one pattern but still continues to switch between foveation and attentional shift, even though the synaptic weight values do not evolve anymore. The network has reached its stable configuration.

We analyze now the visuo-motor information flow exchanged within and between the maps to understand the mechanisms underlying functional integration inside the network. In Fig. 8, we

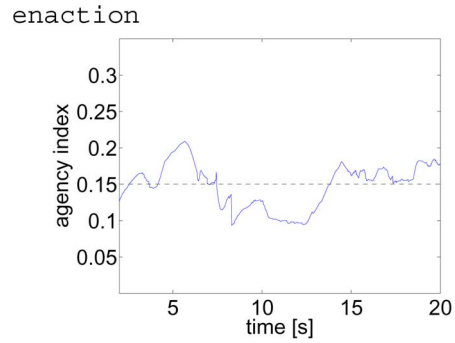

Fig. 9. Agency index during enaction. The agency index between is calculated as the distance between the visuo-motor input patterns and their prediction by the neural maps (see Section II-F). The system's action produces the change of its own degree of agency. Its averaged value represents the reference level for coherent sensorimotor information.

superpose the synaptic links between neurons, extracted from the connection matrix from the most robust links $(c > 9)$, on the neural dynamics. We color those with presynaptic neurons within the vision maps in cyan and the others in the motor maps in red to illustrate how information propagates within and between the vision and motor maps. The graph above plots their respective rate and evolution. As we explained in Section II-E and -F, these links constitute conditional *scripts* forming contingency predictors as the anticipation by one neuron (the presynaptic one) of not-yet-realized stimuli arriving at the right locations (the postsynaptic neurons) and the right time. The correctness of the predictions, by comparison with the inputs, ensures then the sensorimotor integrity; i.e., the system's agency.

Fig. 9 displays the agency index computed from the synaptic links and the input patterns with the formula in Section II-F. This graph reveals that the agency level is not static but dynamic and can rapidly switch within seconds. Its values, on average above 0.15, correspond to the system's agency reference level. It will be our standard level for comparison with in the next experiences. It indicates a certain confidence level of the predictions on the sensorimotor inputs when the device moves: its actions afford its agency, which is in line with developmental findings. Since the "signature" of live enaction corresponds to a certain agency index, it follows that other kind of perceptual experiences with the environment can modulate its level (e.g., visual or somatic illusions with mirrors, feedback delays). We propose to reproduce some of these in the following parts.

B. Visual Feedback Delays

Agency relies on the contingency anticipation and validation of incoming signals during enaction by forward models. To this respect, timing is a critical factor to sustain the systems' integrity. To expose its incidence on our system, we artificially add a delay between proprioceptive and visual information. A complete cycle takes approximately 15 ms for processing data and controlling the robot. From the referent current cycle T and a desired delay of X cycles, we provide therefore the visual input of the cycle $T - X$ to the vision maps.

For this experiment, we delay the visual feedback of 300 ms (20 cycles) and plot in Fig. 10 the corresponding vision map

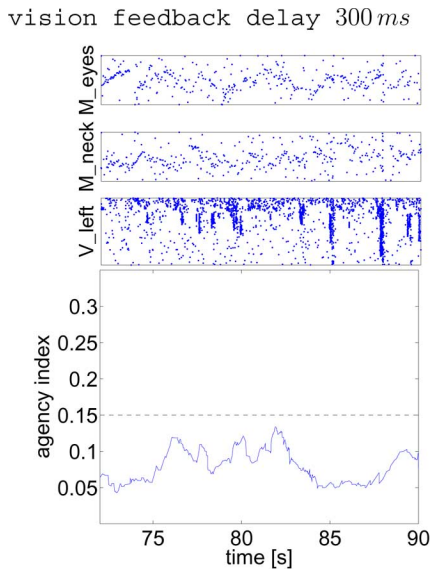

Fig. 10. Agency index delay for feedback delay of 300 ms between proprioceptive and visual information. The top shows the neural dynamics in the left vision map. Despite the delay added, the vision and motor maps present very similar neural dynamics as during enaction. The system's behavior is however erratic, and its agency index is low (<0.1), revealing the discrepancy between the vision and the motor maps.

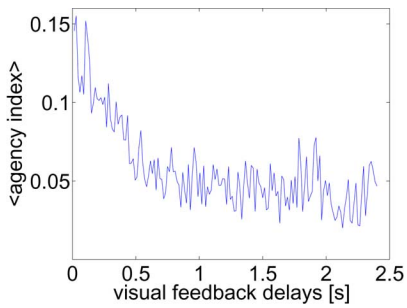

Fig. 11. Agency index for different interval range delays between proprioceptive and visual feedbacks. Each value correspond to the temporal delay of visual feedback. Delays decrease the agency index and a drastic fall occurs around 500 ms, indicating sensorimotor mismatching (functional visuo-motor discrepancy).

and agency index. In this situation, the system behaves differently from live enaction: motion is more erratic, but the neural dynamics appear however preserved in the vision and motor maps (see the top of Fig. 10). The agency index confirms this and reflects the sensorimotor mismatch and the network dysfunctioning. Its level falls below 0.1, revealing discrepancy between the system dynamics and the inputs. A phase transition has occurred. Timing, or the incongruence between the vision and motor STDP neural pairs, compels therefore the network's enaction.

We reproduce in Fig. 11 the averaged agency level relative to temporal delays in the interval between [0, 2500 ms]. The graph presents a distinct separation between live enaction (delay $= 0$) and the other cases when a feedback delay is added (delay > 0), which is in line with developmental psychology experiments [17]. The agency index varies above 0.1 for the enactive case, while the values for the asynchronous cases decrease in two phases. The first phase is for delays below 500 ms with a linear decreasing, and the other at 0.05 for delays above 500 ms. In infants, this phase transition is estimated around 300 ms delays (visual processing in the visual cortex takes about 70 ms and integration is hypothesized to take about 200 ms in the parietal cortex) and discrepancy is linearly proportional to the delay interval (see [17]).

To conclude on these two first experiments, one can say that the system's agency is restricted to the here and now; it does not hold multiple representations of itself by combining present, past, and future events. This stage might correspond to the very initial period of infant development—the first out of five levels in Rochat's graduation of self-awareness [15]—limited to the "present self" or the "ecological self," which entails some basic perceptual differentiation by sensing contingency in sensorimotor flow. Three-year-olds fail to appreciate the temporal distinction of delayed self-image and to recognize themselves even for short delays as small as 1 or 2 s (see [14], [17], [44], and [45]). Moreover, the temporal discrepancy between actions and visual feedback affects the performance of self-recognition (for example, in front of a mirror) and of social interaction (eye contact). We study these two cases in the following experiments.

C. The Mirror Test

The perceptual experience of self-perception in front of a mirror (or from self-touch) is interesting since it involves the perfect synchronization between afferent information (proprioception and vision) and efferent information (motor), which might hugely contrast with other type of stimuli (e.g., observing a visual scene or grasping an object). In our visual processing, the observation of a scene by our system produces in normal conditions a very low saliency since all the motion vectors are aligned in the same direction of the global motion (saliency is distribution in all the image). At reverse, the observation of its own reflection, in front of a mirror and depending also on the distance to it (less than 1 m in our case), will produce a highly contrasting saliency map due to the high ratio between the relative motion (from its own reflection) and the global motion (from the background motion). To increase this effect and ease contingency detection, we put a salient mask on the device; we found that it was difficult to achieve good results without it. The salient information in the image is self-centered at the robot head and cameras levels such that the most salient information is by product located at the fovea where most of the neurons are [see Fig. 12(b)].

We describe now the system's behavior when we place a mirror in front of it. The plots in Fig. 13 display the neural dynamics of the neck motor map and of the left camera visual map, plus the firing rate. The dashed line at $t = 98$ s corresponds to the right moment when we position the mirror. Before this period, the system is under the normal conditions of live enaction. But rapidly after, the robot starts to fixate and to saccade in front of its own reflection. The neural dynamics in the vision map reveal that an entrainment effect has occurred with a strong synchrony. As expected, the motion information is very salient and perfectly contingent to its own action (i.e., proprioceptive information), which makes the mirror perceptual experience very unique. We plot in Fig. 14 its corresponding agency index. Compared with the normal situation in Fig. 9, the agency index jumps

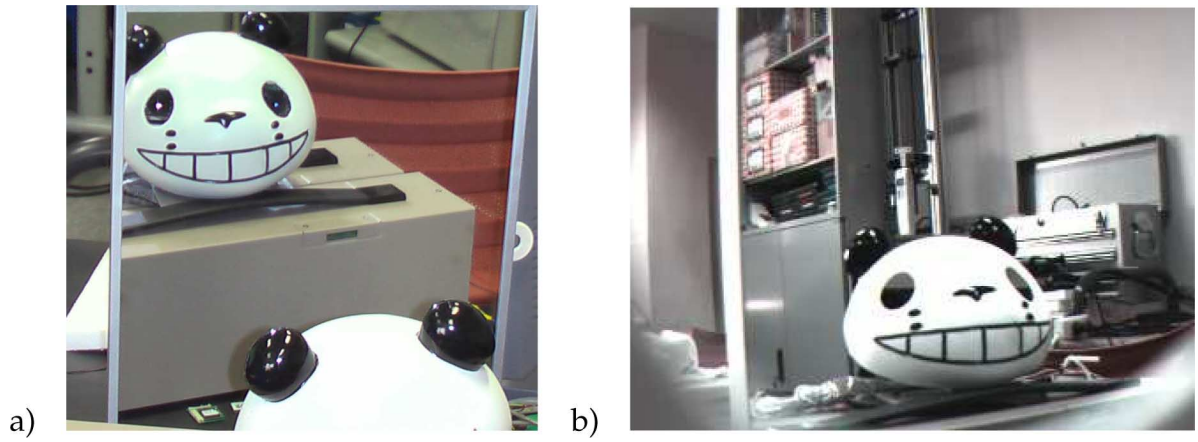

Fig. 12. Mirror experience. When our head-like robot scrutinizes its own reflection in front of the mirror [camera view in b)], most of the salient information gets centered in the middle of the scene, which is not the case in normal conditions of observation of a scene. This perceptual experience is different. To increase this effect, we covered the device with a salient mask.

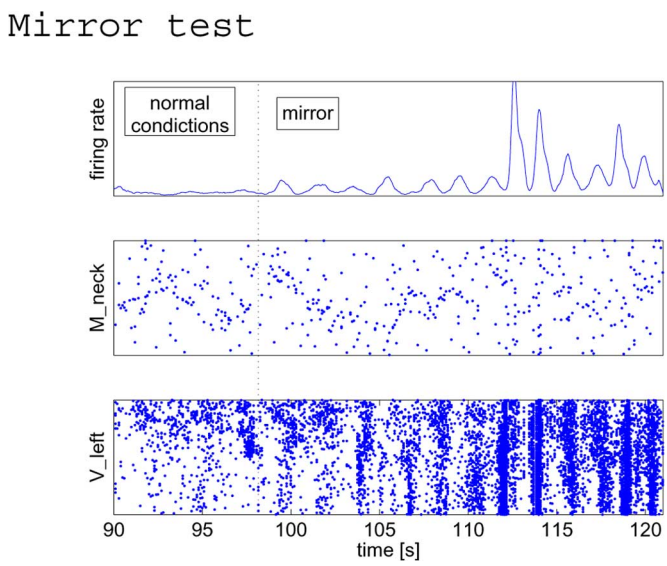

Fig. 13. Neural dynamics of the vision map in front of the mirror. When the robot starts to observe its own reflection, a rapid phase transition emerges in the neural space corresponding to a global entrainment. The self-produced motion centered in the image trigger the neural dynamics in a situation not possible for the normal case.

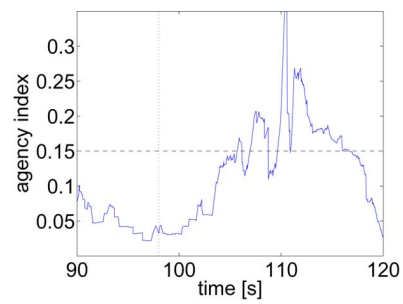

Fig. 14. Agency index in front of a mirror (see Fig. 13). Soon after presenting the mirror in front of the robot, the agency index rise too a very high peak above the normal situation of live enaction.

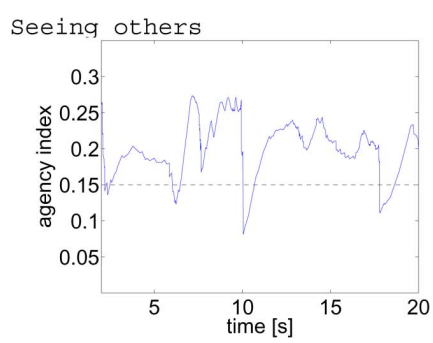

Fig. 15. Agency index in front of another person. Facing a person produces values similar to the mirror experience but higher values compared with the observation of a visual scene. The visuo-motor maps separate self and other stimuli with those of the static scenes. Rather than a measure of dissimilarities, the agency index indicates the degree of coherence in sensorimotor interactions—the shared level of agency between the embodied system and the person that occurs during social interaction.

to a very high value (> 0.2), revealing the strong matching between the inputs and the motor actions. The two maps mutually influence from each other.

D. Seeing Other Persons

In comparison with the preceding experiment, we study now how the system's agency index will evolve when a person faces it; see Fig. 15. Compared to the later case, a perceptual experience with another person produces also high agency indexes, and both are a little higher than for the case in Fig. 9 when it observes a visual scene. This result means that one person's motion or its own reflection induces more saliency in terms of sensorimotor conflicts and coordinations rather than for objects and visual scenes. The embodied system, in a way, modulates and combines its own agency with those of other people, sharing the same circuits. Hence, rather than strict self–other distinction, the agency index is a measure of self–other similarities and dissimilarities: the first condition for developing the self in relation to others (see social resonance [13], [17], [46], [47]).

Fig. 16(a) summarizes the agency indexes for the three studied cases but this time by separating the quantities relative to afferent to efferent information ($S \rightarrow M$, red circles) and to efferent to afferent information ($M \rightarrow S$, blue crosses). Their amount and ratio vary depending on the type of perceptual

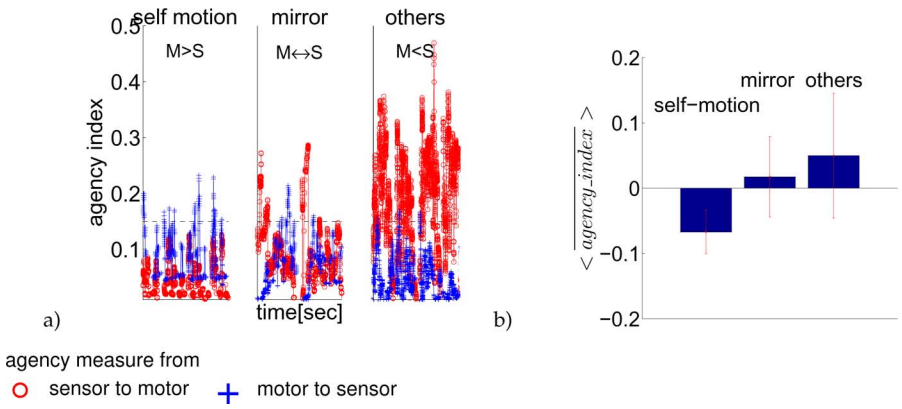

Fig. 16. (a) Summary of the agency indexes of the three cases studied, separating the relative quantity from afferent to efferent information ($S \rightarrow M$, red circles) and from efferent to afferent information ($M \rightarrow S$, blue crosses). (b) Mean scores of agency in the three situations studied after being averaged and centered (in blue). The red vertical lines indicate their variance.

experiences. Fig. 16(b) resumes the relative mean scores of the three agency indexes after being averaged and zero-centered.

IV. Discussion

In this paper, we propose a plausible biological mechanism underlying the emergence of agency based on sensorimotor timing integration in shared circuits, a description that we hypothesize is consistent with simulation theory viewpoint [46], [48] but also with the comparator model [2], [18], [28]. The regulatory mechanism of STDP observed in many brain regions updates and governs the synaptic conduction delays between the neural pairs. Nevertheless, without the constraining from the environment, this mechanism can only generate a random network with anything common to sensorimotor rules. It is the action of embodiment under live enaction (synchronization) that "shapes" the sensorimotors dynamics and aligns them to each other [31]. By doing so, the embodied system preserves its global integrity. Over time, the most congruent sensorimotor neural pairs are reinforced, whereas the incongruent ones are weakened and inhibited. These robust neural pairs represent thus the contingency detectors at a very short time scale (20 ms order) that anticipate the ongoing sensorimotor activity and predict the embodied system's next state. We believe that this model is in line with Wolpert *et al.*'s idea of internal forward models (see [28], [49], and [50]), which make predictions of sensory feedback based on the motor commands to the actuator (i.e., efferent copies). The coordination or the dissonance between the actual state and the predicted ones becomes then the system's "signature" of its own agency. This integration is suspected to occur in the parietal cortex with a separated functioning for synchrony and asynchrony detection [17] but sharing the same circuits for representing self and other [47]. It follows that different perceptual experiences can produce other kinds of agency patterns, as in well-known developmental experiments (see [1], [3], and [51]); i.e., delays between vision and proprioceptive information or the mirror test.

Our model restricts agency to the here and now (i.e., limited to live visual feedback). Said in the reverse, it entails perceptual distinctiveness between contingent and noncontingent information only, which is suggested to represent one of the first (i.e., its most basic) levels of self-awareness (see [14] and [47]). Children 18–24 months old can discriminate the self-produced motion in front of mirrors as those of other persons, whereas in the case of inanimate objects, it corresponds to completely different type of stimuli and sensorimotor patterns. In this case, the system's relative motion is perfectly aligned with the global visual feedback (low saliency), whereas the relative motion with animated objects or persons, or from their self-reflection, produces a different saliency. To this respect, we identified that the distance to the mirror or to the person is an important parameter (i.e., vicinity), as well as the saliency of the robot's face in order to increase the matching. In future research, we plan to expand this framework to better understand how to bridge the low-level sensorimotor interactions described here to possible higher ones including social interactions and imitation.

Acknowledgment

The authors thank K. Kinjo, and H. Sugiyama for valuable discussions.

References

[1] M. Jeannerod, "Being oneself," *J. Physiol. Paris*, vol. 101, no. 4–6, pp. 161–168, 2007.

[2] M. Tsakiris, S. Schutz-Bosbach, and S. Gallagher, "Subjectivity and the body on agency and body-ownership: Phenomenological and neurocognitive reflections," *Conscious. Cogn.*, vol. 16, no. 3, pp. 645–660, 2007.

[3] L. Schwabe and O. Blanke, "Cognitive neuroscience of ownership and agency," *Conscious. Cogn.*, vol. 16, no. 3, pp. 661–666, 2007.

[4] M. Kuperstein, "Infant neural controller for adaptive sensory-motor coordination," *Neural Netw.*, vol. 4, pp. 131–145, 1991.

[5] J. Tani, "An interpretation of the self from a dynamical system perspective: A constructivist approach," *J. Conscious. Studies*, vol. 5, pp. 516–542, 1998.

[6] B. Scassellati, K. Gold, and P. Michel, "Motion-based robotic self-recognition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, 2004, pp. 2763–2768.

[7] Y. Yoshikawa, Y. Tsuji, K. Hosoda, and M. Asada, "Is it my body? Body extraction from uninterpreted sensory data based on invariance of multiple sensory attributes," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, 2004.

[8] J. Gibson, *The Ecological Approach to Visual Perception*. Boston, MA: Houghton Mifflin, 1979.

[9] P. Zukow-Goldring, "Assisted imitation: Affordances, effectivities, and the mirror system in early language development," in *Action to Language Via the Mirror Neuron System*, M. A. Arbib, Ed. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[10] J. Watson, "Detection of self: The perfect algorithm," in *Self-Awareness in Animals and Humans: Developmental Perspectives*, S. Parker, R. Mitchell, and M. Boccia, Eds. Cambridge, U.K.: Cambridge Univ. Press, 1994.
[11] J. R. Movellan and J. S. Watson, "The development of gaze following as a Bayesian systems identification problem," in *Proc. Int. Conf. Develop. Learn.*, 2002, pp. 34–40.
[12] G. Prince and J. Hollich, "Taking synchrony seriously: A perceptual-level model of infant synchrony detection," *Epigen. Robot.*, vol. 7, pp. 89–96, 2005.
[13] J. Nadel, K. Prepin, and M. Okanda, "Experiencing contingency and agency: First step towards self-understanding in making a mind?," *Interact. Studies*, vol. 6, no. 3, pp. 447–462, 2005.
[14] P. Rochat, "Five levels of self-awareness as they unfold early in life," *Conscious. Cogn.*, vol. 12, pp. 717–731, 2003.
[15] P. Rochat, "Self-perception and action in infancy," *Exp. Brain Res.*, vol. 123, pp. 102–109, 1998.
[16] P. Rochat and T. Striano, "Who's in the mirror? Self-other discrimination in specular images by four and nine-month-old infants," *Child Develop.*, vol. 73, no. 1, pp. 35–46, 2002.
[17] K. Hiraki, "Detecting contingency: A key to understanding development of self and social cognition," *Jpn. Psychol. Res.*, vol. 48, no. 3, pp. 204–212, 2006.
[18] N. David, A. Newen, and K. Vogeley, "The sense of agency and its underlying cognitive and neural mechanisms," *Conscious. Cogn.*, vol. 17, pp. 523–534, 2008.
[19] G. Bi and M. Poo, "Activity-induced synaptic modifications in hippocampal culture, dependence of spike timing, synaptic strength and cell type," *J. Neurosci.*, vol. 18, pp. 10464–10472, 1998.
[20] L. Abbott and S. Nelson, "Synaptic plasticity: Taming the beast," *Nature Neurosci.*, vol. 3, pp. 1178–1182, 2000.
[21] S. Song, K. Miller, and L. Abbott, "Competitive hebbian learning and through spike-timing-dependent synaptic plasticity," *Nature Neurosci.*, vol. 3, pp. 919–926, 2000.
[22] P. Larimer and B. Strowbridge, "Timing is everything," *Nature*, vol. 448, pp. 652–654, 2007.
[23] E. Izhikevich, "Polychronization: Computation with spikes," *Neural Comput.*, vol. 18, pp. 245–282, 2006.
[24] N. Masuda and K. Aihara, "Global and local synchrony of coupled neurons in small-world networks," *Biol. Cybern.*, vol. 90, pp. 302–309, 2004.
[25] A. Pitti, H. Alirezaei, and Y. Kuniyoshi, "Cross-modal and scale-free action representations through enaction," *Neural Netw.*, to be published.
[26] A. Pitti, M. Lungarella, and Y. Kuniyoshi, "Synchronization: Adaptive mechanism linking internal and external dynamics," in *Proc. 5th Int. Workshop Epigen. Robot.*, 2006, pp. 127–134.
[27] D. Wolpert, "Computational approaches to motor control," *Trends Cogn. Sci.*, vol. 1, no. 6, pp. 209–216, 1997.
[28] S. Blakemore, D. Wolpert, and C. Frith, "Abnormalities in the awareness of action," *Trends Cogn. Sci.*, vol. 6, no. 6, pp. 237–242, 2002.
[29] C. van Rijsbergen, *Information Retrieval*. London, U.K.: Butterworth, 1979.
[30] M. Lungarella and O. Sporns, "Information self-structuring: Key principle for learning and development," in *Proc. 4th Int. Conf. Develop. Learn.*, 2005, pp. 25–30.
[31] R. Pfeifer and J. Bongard, *How the Body Shapes the Way We Think, A New View of Intelligence*. Cambridge, MA: Bradford, 2006.
[32] E. Thelen and L. Smith, *A Dynamic Systems Approach to the Development of Cognition and Action*. Cambridge, MA: MIT Press, 1995.
[33] A. Bernardino, J. Santos-Victor, and G. Sandini, "Model-based attention fixation using log-polar images," in *Visual Attention Mechanisms*. New York: Plenum, 2002.
[34] G. Metta, A. Gasteratos, and G. Sandini, "Learning to track colored objects with log-polar vision," *Mechatronics*, vol. 14, pp. 989–1006, 2004.
[35] R. Pfeifer, F. Iida, and J. Bongard, "New robotics: Design principles for intelligent systems," *Artif. Life*, 2004.
[36] M. Lungarella and O. Sporns, "Mapping information flow in sensorimotor networks," *Plos Comput. Biol.*, vol. 2, no. 10, pp. 1301–1312, 2006.
[37] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, no. 11, pp. 1254–1259, 1998.
[38] A. Wohrer, P. Kornprobst, and T. Viéville, "From light to spikes: A large-scale retina simulator," in *Proc. Int. Conf. Neural Netw.*, 2006, pp. 4562–4570.
[39] V. Singh, S. Maji, and A. Mukerjee, "Confidence based updation of motion conspicuity in dynamic scenes," in *Proc. IEEE 3rd Comput. Robot Vision (CRV'06)*, 2006.
[40] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural Netw.*, vol. 19, pp. 1395–1407, 2006.
[41] E. M. Izhikevich, "Simple model of spiking neurons," *IEEE Trans. Neural Netw.*, vol. 14, pp. 1569–1572, 2003.
[42] E. Izhikevich, A. Gally, and G. Edelman, "Spike-timing dynamics of neuronal groups," *Cerebral Cortex*, vol. 14, pp. 933–944, 2004.
[43] D. Wolpert, Z. Ghahramani, and J. Randall, "Perspectives and problems in motor learning," *Trends Cogn. Sci.*, vol. 5, no. 11, pp. 487–494, 2001.
[44] D. Povinelli, K. Landau, and H. Perilloux, "Self-recognition in young children using delayed versus live feedback: Evidence of a developmental asynchrony," *Child Develop.*, vol. 67, pp. 1540–1554, 1996.
[45] D. Povinelli and B. Simon, "Young children's understanding of briefly versus extremely delayed images of the self: Emergence of the autobiographical stance," *Develop. Psychol.*, vol. 34, pp. 188–194, 1998.
[46] J. Decety and J. Sommerville, "Shared representations between self and other: A social cognitive neuroscience view," *Trends Cogn. Sci.*, vol. 7, no. 12, pp. 527–533, 2003.
[47] J. Decety and T. Chaminade, "When the self represents the other: A new cognitive neuroscience view on psychological identification," *Conscious. Cogn.*, vol. 12, pp. 577–596, 2003.
[48] C. Keysers and V. Gazzola, "Towards a unifying neural theory of social cognition," *Progr. Brain Res.*, vol. 156, pp. 379–401, 2006.
[49] S. Blakemore, D. Wolpert, and C. Frith, "Central cancellation of self-produced tickle sensation," *Nature Neurosci.*, vol. 1, pp. 635–640, 1998.
[50] S. Blakemore, C. Frith, and D. Wolpert, "Spatio-temporal prediction modulates the perception of self-produced stimuli," *J. Cogn. Neurosci.*, vol. 11, pp. 555–559, 1999.
[51] P. Rochat, *The Infant's World*. Cambridge, MA: Harvard Press, 2001.

Alexandre Pitti received the M.Eng. degree and DEA from Engineer School, ESME Sudria, Paris, France, and the University Pierre et Marie Curie, Paris, France, respectively, in 2002, and the Ph.D. degree from the University of Tokyo, Tokyo, Japan, in 2007.

Currently, he is a Researcher at the ERATO Asada Synergistic Intelligence Project, JST. His research interests include complex systems, emergence and development of embodied cognition, and human motion coordination.

Hiroki Mori received the M.Eng. degree in engineering from Toyohashi University of Technology, Japan, in 2004. Currently, he is working towards the Ph.D. degree at the University of Tokyo, Tokyo, Japan.

His research interests include computational neuroscience, developmental neuroscience, embodied cognitive science, and epigenetic robotics.

Shingo Kouzuma received the M.Eng. degree in information science and technology from the University of Tokyo, Tokyo, Japan, in 2009.

Currently, he is working at Sony. His research interests include whole body motion control of humanoid robots, infants general movement, tactile sensors and hardware design, and integration.

Yasuo Kuniyoshi (M'03) received the M.Eng. and Ph.D. degrees from the University of Tokyo, Tokyo, Japan, in 1988 and 1991, respectively.

From 1991 to 2000, he was a Research Scientist and then a Senior Research Scientist at Electrotechnical Laboratory, AIST, MITI, Japan. From 1996 to 1997, he was a Visiting Scholar at Massachusetts Institute of Technology (MIT) AI Lab. Since 2001, he has been an Associate Professor at the University of Tokyo. Currently, he is a Professor at the Department of Mechano-Informatics, School of Information Science and Technology, University of Tokyo. He is the author of over 200 technical publications, editorials, and books. His research interests include emergence and development of embodied cognition, human action understanding systems, and humanoid robots.

Dr. Kuniyoshi has received an Outstanding Paper Award from the International Joint Conference on Artificial Intelligence, Best Paper Award from the Robotics Society of Japan, Sato Memorial Award for Intelligent Robotics Research, and other awards. He is a member of the Robotics Society of Japan, Japan Society for Artificial Intelligence, Japanese Society of Baby Science, and other societies.