Neural Networks 22 (2009) 144–154

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

2009 Special Issue

Cross-modal and scale-free action representations through enaction

Alex Pitti [a,*], Hassan Alirezaei [b], Yasuo Kuniyoshi [a,b]

[a] ERATO Synergistic Intelligence Project, JST, The University of Tokyo, 113-8656 Tokyo, Japan

[b] ISI Laboratory, Department of Mechano-Informatics, The University of Tokyo, 113-8656 Tokyo, Japan

ARTICLE INFO

Keywords:
Mirror neurons
Action understanding
STDP
Polychronization

ABSTRACT

Embodied action representation and action understanding are the first steps to understand what it means to communicate. We present a biologically plausible mechanism to the representation and the recognition of actions in a neural network with spiking neurons based on the learning mechanism of spike-timing-dependent plasticity (STDP). We show how grasping is represented through the multi-modal integration between the vision and tactile maps across multiple temporal scales. The network evolves into a small-world organization with scale-free dynamics promoting efficient inter-modal binding of the neural assemblies with accurate timing. Finally, it acquires the qualitative properties of the mirror neuron system to trigger an observed action performed by someone else.

© 2009 Elsevier Ltd. All rights reserved.

1. Introduction

Before articulating the first words, the development of social cognition starts with non-verbal communication and the understanding of actions performed by others. Perception of movements, gestures and actions of someone else can help us understand (or guess) about his intentions, his desires, and his emotions.

These capacities of non-verbal communication are argued to be formed from the existence of pragmatic representations, generally implicit arose from the intertwining between perception and action within the brain (Hiraki, 2006; Rizzolatti, Fadiga, Fogassi, & Gallese, 1996), they constitute the body schema that activate automatically the motoric representations in the prefrontal and frontal area. It follows that, observing someone else acting, recognizing it, and understanding it may result then from a direct pairing between the visually observed action and our own motoric representation of it. Differently said, the observer mentally "simulates" the action from his own experience of it (Gallese, 2005), leading then to a "resonance entrainment" in his motor system (Rizzolatti & Craighero, 2004; Rizzolatti et al., 1996; Rizzolatti, Fogassi, & Gallese, 2001). This phenomenon, termed mirror neurons – located in the *F*5 area in the pre-motor cortex (Gallese, Fadiga, Fogassi, & Rizzolatti, 1996) – describes the neurons' response to action-related visual stimuli, such as graspable object or action of other individuals.

Of particular importance, mirror neurons show temporal congruence between visual and motor neurons (Oztop, Kawato, & Arbib, 2006): mirror neurons fire with *accurate* timing to both observed and to hidden end-state actions. Visual representations of an observed action are therefore temporally linked to our own motor representations of the same action, a product of associative learning in line with the generalist theories of imitation (Brass & Heyes, 2005; Heyes, 2001). According to them, what facilitates imitation is due to the general organization of motor control rather than a special purpose mechanism dedicated to imitation. Mirror neurons are thus not innate systems, but rather acquired from learned perceptual-motor links. Other evidences from developmental psychology tend to confirm that timing between sensory and motor representation is crucial for babies in order to acquire the significance of one action. For instance, infants identify soon the timing correlations and the sequential order of events; e.g., synchrony and contingency (Prince & Hollich, 2005). Moreover, in interceptive actions such as reaching and grasping, synchrony detection between different sensory and/or motor channels is particularly important for detecting the right timing for contact or that of preparatory actions (Corbetta, Thelen, & Johnson, 2000; Prince & Hollich, 2005). More complex cognitive abilities – e.g., imitation, self-agency and social interaction – may be developed from these newly acquired affordances (Heyes, 2004; Meltzoff & Moore, 1977; Nadel, Prepin, & Okanda, 2005; Rochat, 2003; Zukow-Goldring, 2005). Taken together, these considerations suggest that exploiting the mechanism(s) regulating timing at the neural level can reveal some of the principle(s) behind action representation, cognitive development and social interaction.

At the neural level, the regulation mechanism responsible for the timing delays between the spikes is the one of spike-timing-dependent plasticity termed STDP (cf. Bi and Poo (1998) and also Abbott and Nelson (2000)). Temporal structure of complex

* Corresponding author.
E-mail address: alex@jeap.org (A. Pitti).

0893-6080/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2009.01.007

actions, for instance, are decomposed with millisecond order precision into ordered sequences of neural rules in canonical motor neurons and in the mirror neuron system (Changeux & DeBevoise, 2004; Lestou, Pollick, & Kourtzi, 2008; Rizzolatti et al., 1996). Precisely, information processing in large networks of spiking neurons is performed both in the temporal domain (i.e., time delay between the spikes) and in the spatial domain (i.e., spatial location of the neurons). It is therefore the coherency of the local dynamics among the neural pairs that will (or will not) produce a coherency at the network scale—we mean a functional integration among the different parallel processes in the maps into a dynamical representation of the body in action.

Our main objective is to understand how such global integration in the neural dynamics is produced during physical interactions. How functional connectivity in the network permits the representation of one action from the differentiated processes done in the sensor and the motor maps having a structured multi-modal activity of the neural code. In this paper, we demonstrate how actions are represented at the neural level as accurate spatio-temporal clusters sparsely encoded over distant neural maps ruled by the learning mechanism of STDP. We set up an experiment of grasping, in which the temporal sequence of the action is acquired (or "represented") through the neural interaction between vision and tactile modalities. This functional integration – termed "vertical association" by Brass and Heyes (2005) – between the sensory and motor maps produces the emergent structure of reentrant or mirroring maps, a result of their entanglement due to embodiment. Interestingly, reentry achieves the cross-modal linkage between the tactile and vision maps making the neural system earn the capabilities of associative memory. For instance, inter-modal activation (capacity to trigger one modality from another) and anticipation (capacity to anticipate the next state of the other modality) combining the feature of a coupled forward and inverse model, predicting the sensory consequences of a motor command and transforming a desired sensory state into a motor command that can achieve it (Oztop et al., 2006). Since the network produces inter-modal associations, information may be retrieved back from the activation of another modality. It follows that the observation of one action (i.e., visual information available only) will induce the simulation of the missing modality (i.e., haptic perception). The qualitative property observed in the mirror neuron system.

In the first section, we present the framework employed to design our neural network. Thereinafter, we study how the network acquires appropriate perception–action matching from repeated experiences of seeing and touching permitting to reproduce the qualitative properties of canonical and mirror neurons: firing to executed actions and to observed actions. We then discuss the relevance of our findings to cross-modal binding and to functional integration in the brain. We advance that the neural organization of the mirror neuron system is mediated by the regulatory mechanism of STDP for action representation and action understanding using the same pathways.

2. Framework

In comparison with classical feed-forward neural networks, information processing in recurrent networks of spiking neurons is not based on the statistical modeling of the available data but rather on the parallel processing of the neurons combined in a self-organized fashion (i.e., assembling the relative spatio-temporal coordinations). We define, in this part, the network architecture, the neuron model used in our experiments and the reinforcement mechanism of spike-timing-dependent plasticity (STDP) that regulates the dynamics of the neurons from each other. We then detail the design of the retina model (biologically inspired) used for the visual processing in our experiments to the transforming of intensity-based images into spike trains.

2.1. Spiking neuron model

The neurons are defined with the formal model (temporal derivatives) proposed by Izhikevich [cf. Izhikevich (2003)]:

$$\begin{aligned} v' &= 0.04v^2 + 5v + 140 - u + I \\ u' &= a(bv - u) \end{aligned} \tag{1}$$

with v representing the membrane potential of the neuron in mV and u a membrane recovery variable – v' and u' their respective temporal derivatives. The neurons are externally triggered by the signal I and their dynamics are resetted after any spiking

$$\text{if} \quad v \geq +30\ \text{mV}, \quad \text{then} \begin{cases} v \leftarrow c \\ u \leftarrow u + d. \end{cases} \tag{2}$$

The variables set $\{a, b, c, d\}$ defines the neuron attributes whether excitatory $(a; b) = (0.02; 0.2)$ and $(c; d) = (-65; 8)$, or inhibitory; $(a; b) = (0.02; 0.25)$ and $(c; d) = (-65; 2)$. For further details, see Izhikevich (2003) and Izhikevich, Gally, and Edelman (2004).

2.2. Recurrent neural network architecture

In our experiments, the networks are composed of large ensembles of neural units. The neurons are connected to each other with arbitrarily short- and long-range synaptic connections (up to one hundred synaptic links for each neurons) and with variable time delays between the neurons (arbitrarily defined up to 20 ms). By doing so, information is sparsely coded in the recurrent networks which facilitates the recall of memories from partial cues and allow for denser and more reliable storage (Aoki & Aoyagi, 2007). Without external constraints from the environment, no particular organizational structure is visible at the system level which gives rise to a spontaneous-like activity in its dynamics.

We explain in the following part the details on the mechanism of spike-timing-dependent plasticity (STDP) on which the networks rely on.

2.3. Reinforcement mechanism of spike-timing-dependent plasticity

STDP is the bidirectional adaptation mechanism which 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 (Abbott & Nelson, 2000; Bi & Poo, 1998; Song, Miller, & Abbott, 2000). 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 pre-synaptic neuron spiking t_{pre} and the post-synaptic neuron firing t_{post} corresponds to the interval range of activation of their synaptic plasticity and weight adaptation Δc.

$$c_{\text{pre,post}} = c_{\text{pre,post}} + \Delta c \tag{3}$$

$$\Delta c(\Delta t) = \begin{cases} A_+ \exp(\Delta t/\tau_+) & \text{if } \Delta t < 0 \\ -A_- \exp(-\Delta t/\tau_-) & \text{if } \Delta t \geq 0. \end{cases} \tag{4}$$

The synaptic weights decay exponentially depending on the time delay Δt between the pre- and post-synaptic neurons in the interval range $[\tau_-, \tau_+]$ [see Fig. 1(a)]. Each time a post-synaptic 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). In all our experiments, we set $-A_- = A_+ = 1$ and $\tau_- = \tau_+ = 20$ ms. Over time, pairs of neurons are consolidated and can form long-range clusters of parallel processes [see Fig. 1(b) and Fig. 2], the idea behind polychronization coined by Izhikevich (2006) and Izhikevich et al. (2004) that we present hereinafter.

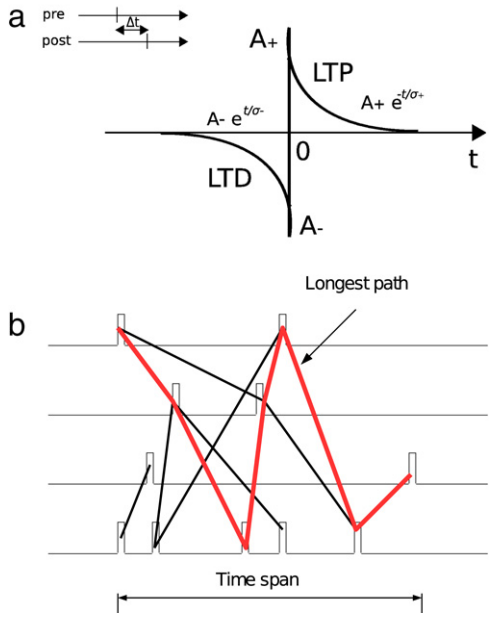

Fig. 1. Mechanism of STDP with $-A_+ = A_- = +1$ and $\tau_- = \tau_+ = 20$ ms. (a) Each time a post-synaptic neuron fires, its synaptic weight is decreased by A_-, and each time a synapse receives an action potential, its synaptic weight is incremented by an amount A_+. (b) Neuronal groups are formed from the dynamical linkage between the neural pairs (hierarchical representations).

Spike-timing in Neuronal Groups. STDP coordinates the dynamics between only neural pairs. Far from being a disadvantage, its action is interesting since it permits to produce a flexible system organization based on many very small *scripts*. Rich information, for instance, can be represented in the network spatially and temporally at the lowest level by neural pairs built into hierarchies of assembled complex patterns. Following this idea, STDP has for some respect similar attributes with the Bayesian rule. The timing-dependent synaptic activation of the neuron $neuron_{\text{post}}$ by the activating neuron $neuron_{\text{pre}}$ can be devised as a conditional rule between the two units in the form of a script, for instance: `if` $neuron_{\text{pre}}$ `fires at time` t_{pre} `, then` $neuron_{\text{post}}$ `fires at time` $t_{\text{post}} = t_{\text{pre}} + \Delta t$.

Such pragmatic timing rules between two neurons represent the smallest "quanta" of information possible to encode. They form, inside the network, a repertoire of primitives that can be used for example to model the motoric system in order to constitute a "grammar" of action primitives (Rizzolatti & Arbib, 1998). It follows that more complex rules – or abstract representations of actions and behaviors – can be constructed from the dynamical assembling of these basic pairs into long-range spatio-temporal clusters of very short conditional codes see Fig. 2.

Moreover, if the network is sufficiently large, neurons may exhibit non-trivial connection assembling apparent to a spontaneous activity or to self-organization. This spontaneous activity may enable then the system to process information beyond its availability as exposed in Fig. 2: the activation at precise timing of particular neurons before t_1 generates the reconstruction process of the *whole* spatio-temporal cluster till t_5. The sequence is retrieved from partial information and one may not see the complete sequence if the first neurons do not fire. To some extent, this retrieval of spatio-temporal patterns can be seen as a "trajectory attractor". Once an event is re-activated, it follows the ongoing synchronization of other units firing—the idea behind chaos itinerancy (Kuniyoshi, Yorozu, Inaba, & Inoue, 2003; Tsuda, 1991; Tsuda, Fujii, Tadokoro, Yasuoka, & Yamaguti, 2004). To pursue our analogy with Bayesian statistics, we can interpret the long-range spatio-temporal clusters as *enfolded* causal chains of scripts (as for Markovian tree) e.g.,

`if` X `fires at` t_1, `then` Y, Z `fire at resp.` t_2 and t_3, and `if` Y, Z `fire at resp.` t_2 and t_3 `then etc . . .`

In this fashion, the neuronal groups form hierarchies of different level descriptions set up from their basic neural bricks in a bottom–up fashion, in line with recent biological data supporting that the motoric system is organized into hierarchical representations (Lestou et al., 2008). Although some computational frameworks have been proposed to model hierarchical representations for action representation (Demiris & Simmons, 2006; Wolpert, Doya, & Kawato, 2003; Wolpert, Ghahramani, & Flanagan, 2001), they do not emphasize the importance of timing as the neuroscience dynamical systems viewpoints do (Edelman, 1987; Kelso, 1995; Rabinovich, Varona, Selverston, & Abarbanel, 2006; Tsuda, 1991), which we think important for its functioning. Besides it, polychronization of neural pairs might establish a "vertical association" between parallel neural processes to represent actions and to re-enact them.

2.4. Retina model

The vision map is coarsely inspired by the serial processing done in the retina transforming a video sequence into spike trains (Wohrer, Kornprobst, & Viéville, 2006). The successive transformations done in the ganglion layers realize a complex filtering on the images into a set of spikes (see Fig. 3). The mechanism discriminates both the spatial and the temporal information from a scene, apparent to a spatio-temporal convolution filtering: the combination of a contrast enhancement on the images (a spatial edge refining) sensitive to "temporal edges" (a processing apparent to optical flow making the neurons trigger to moving objects only). The filtering is modeled with the two-pass Gaussian blurring in the spatial domain in the two layers V_A and V_B (detecting the smooth areas in the image) convolved temporally (detecting temporal edges, high-pass temporal filter):

$$\begin{aligned} V_{A_{\mu_A,\sigma_A}}(x,y,t) &= \frac{1}{2\pi\sigma_A^2}\, e^{-f(x,y,\mu_A)/2\sigma_A^2}. \\ V_{B_{\mu_B,\sigma_B}}(x,y,t) &= \frac{V_{A_{\mu_A,\sigma_A}}}{2\pi\sigma_B^2}\, e^{-f(x,y,\mu_B)/2\sigma_B^2}. \end{aligned} \tag{5}$$

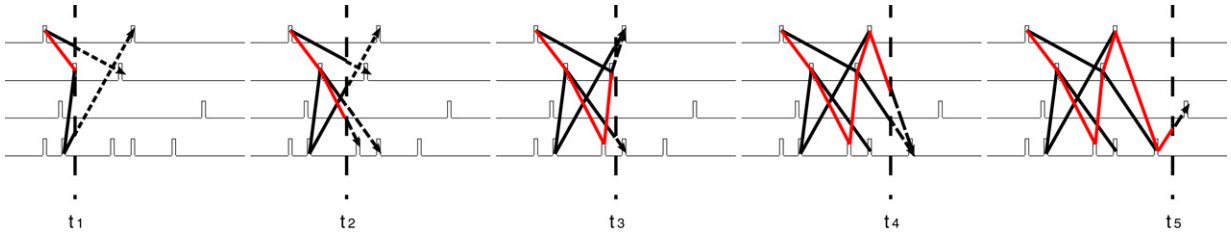

Fig. 2. Self-organization and hierarchical representations. High level neuronal groups are formed from the assembling of lowest level neural pairs self-arranged in a bottom–up fashion. Accurate information about the timing and the location of spikes permit to have a flexible system organization to represent complex sensori-motor structures into long-range spatio-temporal clusters. One of the ideas behind polychronization of Izhikevich (2006).

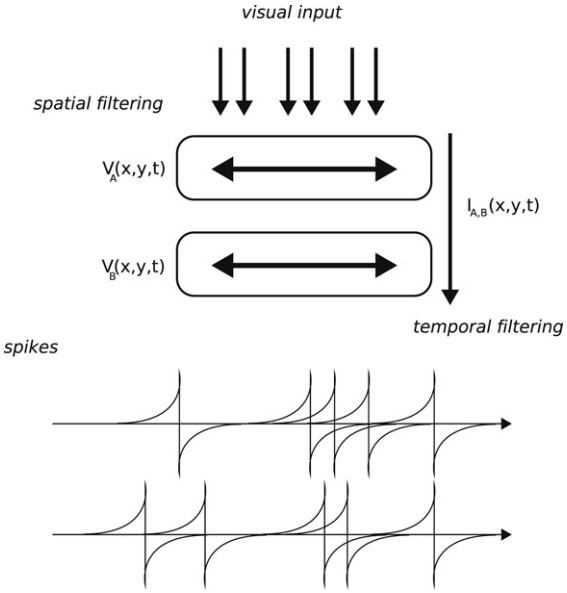

Fig. 3. Retina model. Rough architecture of the information treatment done in the retina transforming a video sequence into spikes trains. The two layers V_A and V_B process a successive spatial filtering of the visual inputs which are then convolved together producing a spatio-temporal filtering of the visual inputs into spike trains.

with $f(x, y, z) = (x-z)^2+(y-z)^2$, μ and σ respectively the center and the variance of the convolution with values $\{\mu_A, \sigma_A\} = \{0, 1.0\}$ and $\{\mu_B, \sigma_B\} = \{0, 0.5\}$. The pixel output of the retina layer located at $\{x, y\}$ provides the excitatory current distribution, $I_i(t)$ to its associated neuron i Eq. (1) with:

$$I_i(t) = V_{B_{\sigma_B,\tau_B}}(x, y, t) - V_{A_{\sigma_A,\tau_A}}(x, y, t+1). \quad (6)$$

3. Experiments of eye–hand coordination and grasping

We reproduce the experimental series conducted by Rizzolatti et al. (1996) illustrating the qualitative aspects of mirror neurons and of canonical neurons: inter-modal binding, action representation and action understanding with temporal constraint. These neurons combine visuo-motor properties to represent one action sequence and to fire at precise timing. In our experiments, we investigate the conditions for such situation to arise in a network of spiking neurons that would lead from the temporal linkage between the visuo-tactile maps to actions representation. We count, to this end, on the regulating roles of STDP and of the body (embodiment) to coordinate the neuron dynamics to the timing integration among the maps.

In the first part, we conduct some repeated experiences of visually perceived acts (i.e., seeing and touching one object) to be mapped in the neural system in the form of linked visuo-tactile representations (encoding both vision and tactile information). Over time, we expect the network to acquire the direct matching from behaviors to neural dynamics. As the representation of physical actions is fetched into the network as multi-modal patterns, it would be possible then to access one modality from the activation of the other. In the second part, we consider how the network integration will permit to access one missing modality from the activation of another one for instance to the understanding of actions performed by others, when no tactile information is received.

3.1. Description of the experiment

The experiment consists of repeated executions of the action sequence associated to grasping (i.e., reaching the cup – time to contact – grasp – tearing) till convergence of the network dynamics to a stable organization. A schematic of the experiment is presented in Fig. 5. During physical interactions, the vision and tactile maps receive their respective information; a time-line of the action sequence "seen" from the sensor maps is presented in Fig. 4. The vision map receives the pre-processed signals from the retina, the tactile map receives the associated force gradient at the object surface. Since we consider the timing information particularly important (e.g., time-to-contact), we assume that this information can either come from the fingertips or from the object surface. We chose the latter solution for practical reasons.

The visual map is composed of 5400 neurons receiving the output signal from the pixel associated with; 90×60 camera resolution which corresponds to 5400 pixels. Their value, binarized after being filtered by the retina layers, are then normalized to [0; 20] and fed to their relative neurons input I in Eq. (1). The correspondence equation between the pixel coordinates $\{i, j\}$ to the neural index, *neuron_ID*, is: $neuron\_ID = i \times 90 + j$.

Information at the tactile sensor surface is sampled with 0.5 mm^2 resolution into a data grid of 1000 samples (details in A.1). Each sample is associated to one neuron in the tactile map and 1000 neurons are composing the tactile map. Below 1 N force pressure, the tactile neurons receive no input value from their corresponding sample, $I = 0$, whereas above 1 N force pressure, each sample triggers their corresponding neuron with $I = 20$ (see Eq. (1)). The two maps are composed of eighty percent of the neurons present in the whole network, all excitatory (6400 units). The other twenty percent (2000 units) are inhibitory neurons added to stabilize the global system. Finally, each neuron, either excitatory or inhibitory, is initially connected to one hundred others arbitrarily selected within the global network with equal synaptic weight ($c_{init} = 5$). Under this condition, the ensemble forms a sparse network with no functional connections before learning. Since the inhibitory neuron activity do not correspond to any representational patterns, we will not display them in the following sections.

3.2. Learning eye–hand visuo-tactile coordination

Experiencing grasping. Perceiving the inter-modal and temporal correlations is an important factor for learning the significance of actions, i.e., to decompose the sequential order of goal-directed movements and to distinguish between them the means and ends (Falck-Ytter, Gredeback, & von Hofsten, 2006). Implicit temporal relations between neural dynamics permit to recognize and to detect if one particular event in the sequence order has occurred or not. This role is held in our experiments by STDP to the detection of coincidental events at the neural level.

For instance, when experiencing grasping, the neurons trigger to the synchronous spatio-temporal events of the scene extracted from the sensors. The firing patterns permit to deduce the signals' "hidden causal structure" that discriminates the whole sequence order into action primitives from its preparation to its execution.

The processing done in the visual map and in the tactile sensors permits easily to distinguish the three phases of grasping and to discern its temporal structure (see Fig. 4):

phase 1 *reaching*, the period before contact with the object,
phase 2 *grasping*, the time to contact and touching,
phase 3 *tearing*, the period when the person handles the object.

The events in the two maps can serve then to construct time-based conditional rules through STDP. The saliency map produced from the retina (Section 2.4), for instance, plays a major role to filter all the static objects from the scene: the output of the retina gets the vision neurons to be sensitive to the moving objects only. Hence, before the grasping at $t = t_0$, the device is completely filtered from the scene and only the hand motion (i.e., its spatial contour) is actively retrieved at $t = t_1$. This interval corresponds

Fig. 4. Temporal structure of grasping. Grasping sequence seen from the retina and of the skin sensors. Before grasping ($t < t_1$), the retina detects only the temporal changes about the hand motion in the direction of the cup: the spatial information about the cup is filtered. When grasping the object ($t = t_1$), joint detection of hand motion contingent to the cup motion and the tactile activity corresponding to a coordination in the neural dynamics (synchronization among the maps). Temporal rules about the sequential order of the event are then associated to a neural representation into the network.

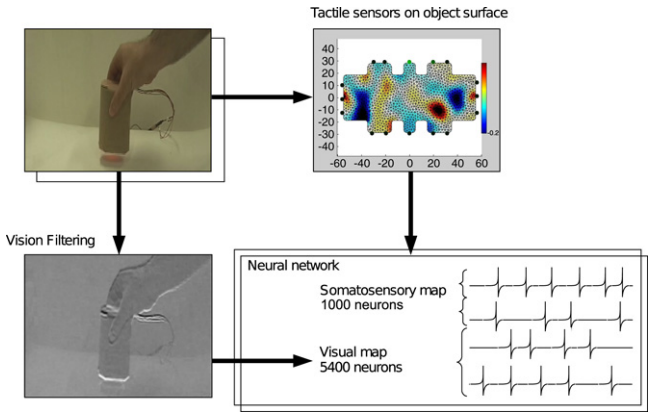

Fig. 5. Schematic of the experiment. The experiencing of co-occurrent visuo-tactile perception during grasping (in the upper-left corner) by the network (bottom-right corner) is done by receiving the incoming information from the camera and from the pressure sensitive device.

to the first phase of the action sequence. The following one stands for the period of time-to-contact at $t = t_2$ and of grasping, when the hand induces some involuntary small perturbances and position changes in the object. These disturbances, accurately detected in the two maps, constitute a unique event distinguished in both maps as the inter-modal grouping of the "hand-device" representation. We analyze in the next section how the network structures its dynamics to categorize the information coming from the sensory inputs.

Network structuring. When experiencing grasping, the network structures its dynamics following the STDP rule: the neuron firing within $\Delta t = 40$ ms latency are wiring together. The temporal coherency between the neurons is the principal factor for their linkage; the location of the neurons is taken into account through the synaptic conduction latency between two spikes. Following this, the neurons can have therefore either short- and long-range connections which can then support segregation within the maps and integration between them. Over time, they may form coherent pairs and clusters associated to the particular experience of grasping with other neurons belonging to the same map or to others. It is this aspect of the network, its functional integration, that we want to analyze in Fig. 6: Fig. 6(a) displays the evolution of the synaptic weights distribution during the learning stage and Fig. 6(b) reproduces the distribution of the paired neurons belonging to the same map or to different maps. This second measure tries to capture quantitatively the network's structural organization evolution. On the one hand, the network level of intra-modal specialization, I_{intra}, corresponds to the information processed between the neurons of the same maps (i.e., the number of synaptic links). On the other hand, the network level of inter-modal integration, I_{inter}, corresponds to the information exchanged between the neurons belonging to different maps (i.e., the number of synaptic links). The third graph in Fig. 6(c) plots the connection matrix between the pre- and post-synaptic neurons belonging to the two maps.

An interpretation of the graphs can be given as follows. The distribution of the neurons' synaptic weights, sharply centered around the unique value 6.0 in Fig. 6(a) in red, indicates a