From Self-Observation to Imitation:

Visuomotor Association on a Robotic Hand

THIERRY CHAMINADE*†§, ERHAN OZTOP†‡, GORDON CHENG†‡ and MITSUO KAWATO†‡

‡ JST-ICORP Computational Brain Project, 2-2-2 Keihanna Science City. Soraku-gun, Kyoto, 619-0288, Japan

† ATR Computational Neuroscience Laboratory, 2-2-2 Keihanna Science City. Soraku-gun, Kyoto, 619-0288, Japan

§ Present address: Institut de Neurosciences Cognitives de la Méditerranée, 31, Chemin Joseph Aiguier, 13402 Marseille Cedex 20, France

*Corresponding author. tchamina@gmail.com

Abstract

Being at the crux of human cognition and behavior, imitation has become the target of investigations ranging from experimental psychology and neurophysiology to computational sciences and robotics. It is often assumed that the imitation is innate, but it has more recently been argued, both theoretically and experimentally, that basic forms of imitation could emerge as a result of self-observation. Here, we tested this proposal on a realistic experimental platform, comprising an associative network linking a 16 degrees of freedom robotic hand and a simple visual system. We report that this minimal visuomotor association is sufficient to bootstrap basic imitation. Our results indicate that crucial features of human imitation, such as generalization to new actions, may emerge from a connectionist associative network. Therefore, we suggest that a behavior as complex as imitation could be, at the neuronal level, founded on basic mechanisms of associative learning, a notion supported by a recent proposal on the developmental origin of mirror neurons. Our approach can be applied to the development of realistic cognitive architectures for humanoid robots as well as to shed new light on the cognitive processes at play in early human cognitive development.

Keywords: imitation, associative learning

1 Introduction

In the course of human development, imitation entails two key abilities: social interaction and learning of motor skills [29]. Facing an imitator triggers a positive emotional response from infants around one year of age, who later engage in reciprocal imitation [24]. Motor learning can be achieved with a "look-at-me and do-like-me" procedure more efficiently than through "trial-and-error", even though these procedures are not mutually exclusive and are probably used in alternation. Cultural learning, another essential aspect of human cognition, also uses imitation to spread codes shared by a group within and between generations, a process Tomasello called the ratchet effect [39]. The fact that autism, characterized by abnormalities of social behaviours, has been associated with impairment in infants' imitation capacities highlights the putative fundamental role of this behaviour in normal social development.

Imitation covers a set of behaviours sharing a common factor, the transformation of an observed action into an executed action, widely differing in terms of what type of action and what part of the action is imitated, or whether the imitator has access to the internal representation of the goal of the model. It covers a continuum of behaviours ranging from simple, automatic and involuntary action contagion to intentional imitation and emulation [7]. Jacobs and Jeannerod recently emphasized that "[imitation] is a folk psychology concept whose boundaries are presently too ill-defined for scientific purposes" [18]. It is difficult to realize the number of complex mechanisms involved in imitation, from body correspondence to extraction of task-relevant features [4]. Because of this complexity, the understanding of imitation benefited from the multi-disciplinary approach inherent to cognitive neuroscience

[37], built on a variety of scientific fields such as experimental psychology and neuropsychology, neurophysiology or computational sciences. Our aim here is not to investigate or review the whole scope of imitative behaviours. Instead, we used the opportunity offered by robotics and computational sciences to test a specific hypothesis. We did not take an engineering stance but a cognitive science perspective in order to test the hypothesis that automatic and non intentional imitation of a simple action, or action contagion, can emerge from the intrinsic properties of a neural associative network fed by spontaneous actions and visual feed-back of these actions available during neonates' motor babbling.

Most developmental theories emphasize that social interactions, in particular understanding of other individual's intentions, could be first achieved through imitation, yet the discussion on the origin of low-level imitative abilities is often neglected, referring instead to the possibility of its innateness. The question of the origin of imitation has indeed been highly controversial following the seminal paper reporting neonatal imitation [25]. The finding that 'neonates between 12 and 21 days of age can imitate both facial and manual gestures' is often cited as evidence of innate abstract representational systems of actions. Yet despite its elegance, this finding has been criticized on two separate fronts, its results and their interpretation [2]. The poor reproducibility of the neonatal imitation has shed doubts on their validity so that altogether, it has been claimed that only tongue protrusion has been repeatedly shown to be imitated by neonates [2].

Other lines of evidence suggest that newborn infants come into the world with innately specified, though crude, visual representation of faces [38]. Using 2-dimensional stimuli, it was found that neonates preferably track a schematic face-like pattern than other patterns consisting of the same facial features in different, not face-like, arrangements. This led to the proposal that neonates were provided with innate modules such as a face-detecting device consisting of a perceptual system sensitive to specific arrangements of 2-dimensional shapes [reviewed in 38]. Recent finding of neonates imitation of oral gestures in chimpanzees [28] suggests a release mechanism ability that is likely to have evolved in relation to feeding behaviours and to be restricted to oral gestures. In an evolutionary psychology perspective, these mechanisms could have been positively selected for the advantages they provide to neonates not only in feeding but also in initiating social interactions. It is noteworthy that innate recognition and imitation of facial and oral gestures in humans facilitate early social interaction with caregivers and are important for social cognitive development.

Keeping in mind that imitation is not a unitary behaviour but covers a range of different behaviours, this attractive scenario on the innate origin of oral and facial gestures imitation cannot be generalized to other, in particular visible, body parts. Interestingly, another visuo-motor ability has been clearly demonstrated in human neonates less than a month old despite the poor resolution of their visual system: they can control their actions purposefully in order to bring their hand back in the field of view, even when it is being pulled by an external force [41]. This shows that babies perceive their hands as objects of particular interest, and can make use of spontaneous arm waving to build an embodied frame of reference for their actions [41]. The central aspect of our hypothesis is derived from this result. We propose that in the case of hand movements the temporal synchrony between the motor command and the corresponding visual (and somatosensory) feed-back is sufficient to acquire visuomotor associations which can sustain early forms of imitation. The observation of another agent would automatically retrieve the visuo-motor association which execution results in a sensory input corresponding, in a loose sense, to the observed one. Such a visuo-motor association would thus uphold action contagion, an early form of involuntary imitation. This developmental path could be particularly important when compelling evidence about innate mechanisms are absent, as in the case of hand and finger gestures.

Recent neurophysiological findings on the cerebral bases of perception, imitation and understanding of actions provide us with a rich set of results [9,13,17]. Of particular interest are mirror neurons, activated both when monkeys perform a goal-directed action and when they see the same action performed by an experimenter. These neurons were found in the reciprocally connected ventral premotor region F5 and inferior parietal region PF [36], and it has been proposed that a similar mirror system exists in humans. These neurons are the best example of a motor resonance system, in which not only brain activity related to both observation and execution of action, but also behavioural markers of this activity, have a reciprocal influence on each other. In the most classical behavioural example, observing an action hinders the execution of a different one [22,32]. Key regions for the human imitation are the left inferior parietal lobule [9], possible homolog to the monkey's PF, and the ventral premotor cortex [17], putative homolog to F5 [34]. Both visual and motor properties have been reported for these two reciprocally connected regions. One study showed that parietal and premotor cortices involved in producing a specific action are recruited when understanding the precise goal of another individual performing a similar action [8].

A classical view related to synaptic plasticity and learning in the brain is Hebbian learning (‘when an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased.’) In its original form this learning is not very useful from a computational point of view. However slight relaxations to the original statement allow one to construct so-called associative memories [30]. The crucial property of an associative memory is its ability to retrieve a stored pattern based on a partial representation of it. So these architectures are also called content addressable memories. Assuming that similar mechanisms are at work in the cerebral cortex, we can think of how early automatic forms of imitation ability may emerge. When the system (a learning robot or an infant) generates motor commands, the representations of this command and the sensed effects of the command can be associated through Hebbian-like learning. These associations between motor command and their sensory effects are reminiscent of internal models, defined as central representations of movement which mimic the input-output relationships of a controlled system, the human body, for movement generation. The model that predicts the sensory consequences of a motor command is referred to as a forward model [26,43]. Conversely, the model that outputs the required motor command to reach a desired sensory state is called an inverse model. A computational structure storing input-output relations of a control system and capable of retrieving the input-output pair given either the input or the output, as in an auto-associative memory, is effectively a combined inverse and forward model of the whole system (e.g. the MOSAIC control architecture [44], that employs explicit multiple paired inverse and forward models and has been proposed to sustain imitation [42]).

Thus, in the computational framework our hypothesis can be stated as follows: motor babbling could induce Hebbian-like acquisition of sensory-motor associations able to sustain early imitative abilities. The synchrony between motor commands and sensory feed-back during motor babbling could allow Hebbian learning of the associations between the two types of events, i.e. motor and sensory, in infants. The activation of such multimodal representations when actions from another person are perceived could result in automatic and involuntary production of the motor output, a behaviour that would be considered as action contagion. This Hebbian acquisition of sensory-motor associations and its relation with mirror neurons and imitation have been proposed elsewhere [15,21]. The proposal that associative learning could provide the developmental link between automatic imitation and the representation of goal-directed and meaningful actions by mirror neurons is particularly interesting. Other studies have used Hebbian-like associative learning during self-observation to bootstrap early forms of imitation in embodied agents [23].

This bootstrapping of imitation could be particularly important for certain body parts, most notably the hands in the visual domain and the vocal tract in the auditory domain, whose circuits appear to overlap on the human cortex [3,35]. To assess the extent of imitation features that can be bootstrapped we combined a robotic human-like hand with a minimal visual retina via biologically-inspired associative network to reproduce an infant’s early visuo-motor experiences. Results indicate that this simple system depicts contagion of finger postures demonstrated by an external agent, and, more importantly, that it can generalize to unseen hand postures, raising the possibility that this mechanism is at play during early infancy.

2 Methods

2.1 Associative memory and hardware

As detailed in the Introduction, accumulating evidence suggests that visual and motor processing overlap in most aspects of the sensory-motor transformation, which at the neuronal level can be realized through a Hebbian-like learning. The goal of the present experiment is to explore to what extent the features of imitation, such as the ability to produce unseen postures through generalization, can be bootstrapped via self-observation and Hebbian-like association. In the next section, we describe the visuomotor representations that are stored in the higher order Hopfield (HHOP) network, the associative memory we used for this study. The details of HHOP network is given in the following sections.

2.1.1 Visuomotor representation: the HHOP encoding

[insert Figure 1 about here]

Figure 1 shows the flow of visual and motor information converging onto HHOP network forming a visuomotor representation over the array of units making up the HHOP network. The visual input initiates from a video source. We have circumvented the requirement for extensive computer vision algorithms (e.g. skin colour tracking, body segmentation etc.) by providing a background-free visual input in an approximately fixed angle. The image processing we applied emulates a very simple retina: the colour video is converted into greyscale and smoothed. A bounding rectangle is determined based on the smoothed image, and the contents are scaled to a fixed size set by the HHOP network (in this study: width 32, height 12). Thresholding and removal of isolated points complete the preprocessing of the input video resulting in a reduced image, referred to as the retina from now on.

The retina is directly connected to the visual units of the HHOP network, which receive binary pixels (+1 or -1) in a one-to-one manner, where +1 indicates the existence of a bit at the corresponding location on the retina. The motor code that is used to drive the Gifu hand is represented by five bits for each finger, also directly connected to the HHOP network. The units that receive motor input are referred to as motor units. The representation in the motor units is redundant as in the current setting, fingers can only be in two states (up or down) which could be coded by one bit per finger, but will allow additional flexibility in future developments, for example encoding of joints angles. The motor units of HHOP receive these representations of motor patterns whose values are +1 or –1. The visuomotor representation produced by the combination of the retina and motor inputs form an input pattern (orange square in Figure 1) which can be either stored or used to retrieve a stored pattern by the HHOP network. We will often use ‘pattern’ to indicate a visuomotor memory trace, which has been stored or supplied as a key for the retrieval of a visuomotor memory trace. In most sophisticated architectures, one could revert to a feature space framework and extract a set of powerful features (e.g. higher order moments) with desirable invariance properties and use those as visual input to the HHOP network. The result would be a robust imitation engine, thanks to the sophisticated feature encoding. However in this study our concern is not so much to provide robust imitation but rather to present a simple but yet biologically realistic connectionist framework that may be thought of as a model for the origin of action contagion. For this purpose we have avoided complex computer vision and pattern recognition techniques.

2.1.2 HHOP: Higher-order Hopfield network

The standard Hopfield network [16] is a classical associative memory, which is composed of units that are fully connected with symmetrical weights. The computation of the connection strength between units in a Hopfield network follows a Hebbian-like update rule, and is thus considered as a biologically plausible network. Although Hopfield network is suitable for small problems, in its standard form, the performance of the network degrades when the patterns to be stored are too closely correlated (e.g. overlapping patterns). Preliminary experiments showed that Hopfield network was not suitable for storing visuomotor patterns since the patterns are highly correlated. Motivated with the fact that the use of higher-order units increases the computational power of neural networks [14], we implemented an extension of the Hopfield network using higher order units (named Higher-order Hopfield network or HHOP [31]). Here we present the basic equations that are sufficient to implement the HHOP network on a computer. The HHOP network representation is bipolar (i.e. -1, +1). Each unit in a HHOP network receives input from all possible products of the other units as illustrated in Figure 2. Output of a unit (S_i) is given by

$$S_i = \operatorname{sgn}\left(\sum_{jk} w_{ijk} S_j S_k \right) \tag{1}$$

Where sgn(.) is defined as

$$\operatorname{sgn}(x) = \begin{cases} -1, & x < 0 \\ +1 & x \geq 0 \end{cases}$$

Denoting the p^{th} pattern to be stored with ξ^p, and representing the k^{th} bit of pattern p with ξ_k^p, the connection weight of the product $S_i S_j$ to unit k is calculated using

$$w_{ijk} = \frac{1}{N} \sum_p \xi_i^p \xi_j^p \xi_k^p \tag{2}$$

The running of the network is asynchronous. After initial loading (assignment of S_i's), the network is run by choosing a random unit and applying the update rule (1) until convergence is reached. In practice it is possible to stop updating at a fixed iteration since usually, several passes for each unit suffices to reach equilibrium with HHOP. The network was iterated four times in all simulations reported here. If the initial loaded pattern is close to one of the stored patterns then the network state converges to that pattern, called the attractor pattern. The main advantage gained by using higher-order units is the increased ability to deal with correlated patterns compared with the standard Hopfield model. In addition less iterations are required to reach stable equilibrium states [20].

[insert Figure 2 about here]

2.1.3 Experimental Platform

We used "The Gifu Hand III" (Dainichi Co. Ltd., Japan; referred to as "Gifu Hand"[1] in this report) as the test-bed for action contagion using HHOP associative memory. The Gifu Hand consists of a thumb and four fingers (Figure 1, left). The thumb possesses four joints with four degrees of freedom (DOF) while the fingers possess four joints with 3-DOF. One hand contains a total number of 20 joints, encompassing 16-DOF. These DOF closely approximate those of a human hand.

A network control framework was developed for maximal flexibility and work load distribution. It contains three computers as shown in Figure 3. The Video Capture computer (C) is connected to a video camera and a video capture board. The task of C is to capture frames and transfer them to the target machine H. The colour video is sent at 30 frames/second with a resolution of 320x240. The High Level Coordinator

[1] This article includes a word that is a proprietary term. Its inclusion does not imply it has acquired for legal purposes a non-proprietary or general significance.

(H) has three main tasks. The first one is to preprocess the incoming video as described in the previous section (see also Figure 1). The second task is to run the HHOP network based on the video processing result. Finally the last task is to send commands to the Low Level Hand Control Server (S) that is directly connected to the Gifu Hand. Computer S in return implements a PD servo driving the Gifu Hand to the desired postures. In this sense the High Level Coordinator (H) is the main imitation system while Capture and Controller servers (C and S) serve as the input and output channels, similar to the visual and motor pathways in humans.

[insert Figure 3 about here]

2.2 Experimental procedure

2.2.1 Network simulations

A first series of simulations were run off-line on synthetic data to ensure correct function of the associative network and assess its properties. The aim of these simulations was to confirm that HHOP can reliably be used in real-time as an associative memory bridging video input and controller software that actuates the Gifu Hand. For these simulations the artificial input patterns consisted of all possible hand postures with 4 fingers (all but thumb) up or down, and the expected retinal images for the posture coded by the motor patterns. The motor bits of the input patterns were set to ones if the corresponding finger was up and to zero otherwise (zeros are treated as -1 when loaded to the HHOP). To obtain the retinal image, the 32 columns of the retina were divided into four equal parts. Each part was used to represent a finger, and its six central columns were filled with one if the motor code of the corresponding finger was one and zero otherwise. The two-pixel gap between fingers was always kept null. Two lines at the bottom were filled with ones in all postures to represent the palm of the hand. Thus a total of 2^4=16 synthetic visuo-motor patterns were generated for testing. Two properties of the HHOP network were tested: its robustness with regard to noise and its ability to generalize across patterns.

Noise Robustness

To assess noise robustness, random noise was uniformly added to theinput patterns by flipping the value of n percent of the bits, with n varying from 0% to 100% in 10% increments. Note that the noise contamination is applied to the motor and visual units without distinction. Before testing, HHOP was trained with the full data set of 16 patterns. Then, for each noise level, the recall capability of the HHOP network was tested by loading each of the 16 patterns and contaminating them with noise at the appropriate noise level. The network was then iterated four times (each unit received 4 updates in a random order) after which the overlap with the original ('clean') pattern was recorded. This process was repeated 100 times for each pattern. The average over repetitions and patterns gives the noise robustness curve of HHOP, given in Figure 5A.

Generalization across patterns

To assess the ability to generalize across postures, a random training set was generated which could have n=1 to n=15 training patterns. The testing was conducted on the 16 possible hand postures, 16-n of which were not in the training set. For each level of n, 500 randomized training and testing session were run. If the network can 'infer' the motor patterns corresponding to patterns which are not in the training set, then m, the number of correctly recalled motor patterns, must be greater than n (m>n). Notice that after a pattern not belonging to the training set is loaded into the network, the motor part is randomized, and the network is iterated. A correct functioning would recover the randomly initialized motor bits such that they reflect the posture represented by the visual bit patterns. In other words, this generalization tests the ability of the network to reconstruct the motor information on the basis of the visual information of a posture the network has never experienced before. Figure 5B shows the generalization ability of the HHOP network averaged over the 500 repetitions.

2.2.2 Robotic validation

To assess whether this network depicts action contagion in a real-life situation, we used real hands, either the Gifu hand or other hands (Human hands, Wood hand) to provide visual input to the system in the testing phase. Since our hypothesis is that a network trained by self-observation displays action contagion when observing another individual, we performed training with the robotic hand: the robot watched its own hand postures while associating the motor commands with the perceived hand postures through the simple retina. The testing is carried on with other hands as well as the robot hand itself. Two properties of the HHOP network were tested: its ability to generalize between different agents and to generalize to new postures.

Generalization between agents

Generalization between agents entails the ability of a network trained by self-observation to respond to other agents, i.e. other hands (Human hands, Wood hand) used to form the retinal bit pattern (preprocessed real-time video). Upon presentation of a posture, the motor part of the input was set randomly as in the off-line simulations. Then, the network was iterated, yielding a motor code that was send to the robot for visual inspection of the action contagion behaviour. In addition, the motor code was recorded for estimating an average performance value for the generalization. In the real experiments with the robot, we used two reduced sets of hand postures as human hands could not perform all of the sixteen postures that were used in the off-line simulations. Both sets contain four postures as illustrated by Figure 4.

[insert Figure 4 about here]

The first set consists of 4 postures: all fingers flexed, all extended, index finger extended, and little finger extended. The second set consists of two or three-fingers postures. We created four different versions of each set by producing the respective postures with the respective agent (Gifu hand, wood hand, each of the two human hands) four times independently. The training, emulating the self-observation learning, is carried on by letting the Gifu hand generate all postures from one set, and associating the motor code and the retinal feedback of its posture perceived in one HHOP associative network. The ability of this network to generalize to different agents is tested by recording its responses when presented with another set of retinal images of hand postures. The other set could be from the same (Gifu) or another (Wood, Human) hand, and the four versions of each set were used for testing in order to take into account the existence of noise during posture presentation. For each posture, the number of fingers in the correct position on the robotic hand output was counted after iterating the network four times. For example, if the output of the robotic hand has two fingers in the correct position and two in the incorrect position, the ratio of correct response is ½, which corresponds to chance level for finger configurations. Finally, for each set and agent, four networks are formed from each of the four versions of the set, and each are tested across the four versions of set, so that

4 (postures) x 4 (trained network) x 4 (tested sets), i.e. 64, observations are converted into the percentage of correct reproduction given in Table 1.

Generalization between postures

A follow-up was conducted to specifically test generalization to new finger postures. Offline simulation indicates that networks trained with a small number of gestures could not generalize to the whole set of remaining patterns. The Gifu hand was used for both training and testing, allowing the use of all sixteen postures available with four fingers. Five simple postures were used for training (all fingers up, each of the four finger up individually), and the trained HHOP networks were tested against a set containing all possible configurations except the ones used for training. The fully closed hand was not used in this experiment. We considered that the posture was correctly reproduced when all fingers were flexed or extended correctly as this shows that the motor code that would yield the observed posture was correctly inferred. Therefore the chance level for correct imitation was 6.25% in this experiment.

3 Results

3.1 Network simulations

Noise Robustness

Hopfield networks construct associative memories by creating attractor dynamics around the stored patterns. Given proper conditions, when loaded with a pattern near a stored one, the network will settle to an attractor that will coincide with the stored pattern. The ability to return a stored pattern given a noisy version of it is thus intrinsic to Hopfield networks including our higher-order Hopfield network.

The effect of noise on the recall performance of HHOP is illustrated in Figure 5A. Statistical analysis shows that the effect of noise on recall performances is highly significant. For levels of noise between 0 and 30%, recall performance is perfect and still superior to 90% when the noise amounts to 40%. In other words, when 40 out of 100 pixels of canonical hand postures used for training are randomly flipped, the HHOP still recognizes more than 9 out of 10 postures presented. Recall performance is reduced at 50% noise (32%) and 60% noise (16%), but bounce back for increased levels of noise. When the input is at 100% noise, the input pattern is the inverted version of the original. The increase in recall performance can be intuitively explained when we look closer at the workings of HHOP. Assume we have flipped some number of bits in the input pattern so that the ratio of flipped bits to all bits is p ($0=< p<=1$). The effective input to a unit is all the double products of the input bit patterns (see Eq. 1), which are -1 or +1, therefore $p^2 + (1-p)^2$ of the input channels to each unit remains the same as if no noise was applied. Thus for the first iteration, the ratio of number of disturbed input channels to a unit has a maximum at p=0.5, where we may expect to see the worst recall performance. Although, this simple argument does not consider the subsequent iterations, it gives a reasonable approximation to the noise level where the recall performance is minimum ($p \approx 0.6$; see Figure 5A).

Generalization across patterns

In contrast to their ability to return a stored pattern to the presentation of a noisy pattern, Hopfield networks are not easy to craft to generate 'plausible' new memories out of the stored patterns, an essential feature for generalization. It is known that spurious or 'ghost' memories will be created when a given set of patterns are stored, but the possibility that these spurious memories would coincide with what we understand as generalization is not clear. Although, it is possible to understand the ghost memories of the standard Hopfield network as linear combinations of the stored patterns, it is harder when higher order or more complex memories are concerned as in HHOP network. Therefore, we tested the ability to generalize in an off-line simulation before using the network in a real-life environment. Results show that HHOP network is capable of correctly inferring motor codes other than those used for training (Figure 5B). On average, when 8 to 12 postures are used for training, the network can correctly reproduce 3.5 postures that were not part of the training set, and it extrapolates to the 16 postures when more than 12 postures are used for training. Between 5 and 7 postures the network shows limited ability to generalize, reproducing 1 to 3 new postures.

[insert Figure 5 about here]

3.2 Robotic validation

Generalization between agents

The two sets of pattern inputs from the Gifu hand shown on Figure 4 were repeated four times to investigate generalization over agents. HHOP network were trained with one of this set, and tested with similar sets, one identical and three different versions recorded using the Gifu hand or one of the other hands available, two humans hands and a wood hand. Results, given in the first lines of Table 1, reveal an average percentage of 90.7% of correct finger imitation for the first set and of 97.8% for the second set.

[insert Table 1 about here]

It appears that the networks trained by self observation demonstrate perfect reproduction when tested with the Gifu hand, despite being tested with the exact same set as well as three other versions of the same set. The system also depicts high level of generalization to other agents. Though the number of patterns used in the two datasets was similar, generalization to other agents is improved in the second dataset. A likely explanation is the similarity of the retinal image of the closed and open hand patterns of the first set due to scaling.

Generalization between gestures

Because of the small number of postures available in the previous simulation, it is not fit to test the generalization ability of the system. Indeed, the offline simulation demonstrated positive generalization results when 5 gestures or more are used. Yet technical and practical limitations forbid the use of extensive tests similar to those used in the offline simulation. To test generalization more specifically, a similar experiment was thus conducted using five simple postures to train the network and all remaining postures to test it with the Gifu hand. The assumption here was that complex postures could be described as combinations of the simple ones. The mean ability to generalize to the postures presented in the new data set is about 30%, significantly higher than the chance level of 6.25% for whole hand configuration. There was a huge variability depending on the postures tested. In particular, we found that the network was able to imitate perfectly (100%) the two-finger posture involving the index and the middle fingers even though it had never produced two-fingers postures during training. This indicates that the system is able to reproduce postures it had never experienced before. On the other hand there were some postures it was never able to reproduce.

4 Discussion

The aim of the simulations reported here was to test whether the ability to imitate could emerge from Hebbian-like learning of sensori-motor associations resulting from self-observation. Perception of the visual consequences of our hands actions would suffice to acquire internal models of these actions, which could then be triggered by the observation of other individuals' hand. This automatic and involuntary action in response to the perception of another individual's action is referred to as action contagion [6], as in contagious yawning. This Hebbian learning of sensory-motor associations and its relation with imitation have been proposed elsewhere [15,21], but not tested empirically on a robotic hand.

We investigated this hypothesis using a robotic hand and a simple associative network. First we will describe the properties of the system in relation to the biological inspiration, and argue that associative network was more suitable for the current simulation than another possible approach. Then we will discuss the results in relation to the hypothesis, and finally describe possible extensions of the system.

4.1 Properties of the computational network and robotic implementation

In addition to testing a hypothesis derived from cognitive sciences, one long-term goal of this research is to implement biologically-inspired robotic systems. It was thus not possible to limit this investigation to the off-line simulation aspects, and the system had to be implemented in a biologically inspired robotic system. The hand was chosen for two reasons: in a biological perspective, it is one of the effectors clearly visible to infants from birth and it is possible to relate our results to developmental psychology studies [41]; in a roboticist perspective, it offers a large number of degrees of freedom in a limited space with limited security or control issues. Nevertheless, the same principles could be applied to most parts of the human body, the most notable exceptions being the head and face.

The associative memory we employed is a connectionist architecture relying on Hebbian-like learning mechanisms with units resembling neurons, and is thus a credible biological simulation. The experiment setup was voluntarily kept minimal, as a way to limit the hypotheses required to describe the homologies between our artificial and biological systems. For instance, no claims are made here on the biochemical mechanisms that may underlie association in the cortex. Two fundamental properties were necessary for the associative network (Higher-Order Hopfield net, see Methods) to be used in this experiment. First, in order to be usable in real life environments when implemented on a robot it needs to be resistant to noise. The ability to return a stored pattern given a noisy version of it is intrinsic to Hopfield networks, and results from the off line simulations indicate that when 40% or less of the pixels of the input pattern used for training are randomly flipped, the HHOP still retrieves on more than 9 out of 10 postures presented (see Figure 5A). This robustness to noise is fit for correctly reproducing gestures. It was also found that the network is capable of generalizing, on average to more than three new gestures when 8 or more are used for training (Figure 5B). This series of off-line tests of the associative network ensures that it is robust to noise and supports generalization. It can thus be used to test our hypothesis that some imitative abilities can emerge from self-observation.

4.2 Emergence of imitative abilities in the associative network

We investigated this system's ability to imitate not in an engineering point of view, but from a cognitive science perspective in order to acquire knowledge on the possibility for simple imitation abilities -comparable to those described by Piaget in its early developmental stages- to be bootstrapped by experience given the simple (innate) capacities of the system (the newborn). As explained in the previous parts, efforts were made for the network to be biologically realistic so that as few *a priori* hypotheses as possible are needed.

The main result from the robotic implementation is that this associative network trained by self observation of hand postures is capable of action contagion, depicting two features of imitation: reproducing actions regardless of the actor, and more importantly exhibiting one-shot imitation, that is without training the motor code corresponding to a new posture presented can be inferred, and hence executed. In the first experiment, we tested the network response when it was tested with either visual input from itself or from another hand (Table 1). As expected, the network was 100% correct when it was tested with its own visual input, and largely above chance (superior to 80%) when tested with another hand. This was irrespective of which set of finger postures and which hand was used, though there may be individual differences that are beyond the scope of the present report. In accordance with the theoretical framework which inspired this experiment, our interpretation for this ability is that observation of actions from the self can be used to associate (near) synchronous visual and motor aspects of an action by application of Hebbian type of learning rule. The stored visuomotor patterns can be seen as internal models of actions. Hands used in the experiment differ in relative size and shape of fingers, but their general aspect is similar to the Gifu hand. After preprocessing, the content of the retina can be regarded as a noisy hand input similar to a stored pattern. In other words, the differences among the hands are treated like noise by the associative network. Because of robustness to noise, the observation of another individual's action can retrieve the visual aspect of the corresponding stored pattern. The ability of the network to return a stored pattern to the presentation of a noisy pattern leads to the retrieval of the original motor code stored along with the visual code produced during self-observation. If this motor code is used to drive the hand, we obtain the action contagion reported in Table 1. This automatic and involuntary behaviour is defined as action contagion in psychology, and the present result suggests self observation could bootstrap this initial step in the development of imitative abilities.

The experiment testing the capacity of the associative network to generalize to unseen postures in a real-life setting used only the robotic hand, with the intention to stress the capability to generalize by checking all possible postures which could not be performed satisfactorily with a human hand. The results showed that the overall ability to generalize (~30%) is significantly higher than chance (6.25%) but highly variable. Unexpectedly, some new postures are always reproduced while others never are. This indicates that the system is able to reproduce postures it had never experienced before, but didn't generalize to the whole data set with the simple and limited number of postures provided for training. This result demonstrates that the ability to imitate postures that are not in its existing repertoire of actions emerged from the associative memory network without any tuning. In accordance with our hypotheses, action resonance could eventually be used to learn new gestures after some simple visuomotor primitives, in this case individual fingers' patterns, have been used to form an associative memory. It would be interesting to compare the present results with the development of generalization capability, or more generally of action contagion, in infants.

4.3 Possible extensions of the imitation system

Acquisition of visuomotor representation of the body

The neonatal ability to recognize faces or to move the hands so that they remain in the visual field, even when the visual feedback is given through a camera and screen display [41], strongly suggests innate features detectors for faces and maybe hands too. In addition, these feature detectors could correspond to brain areas for which specific activity for faces [Fusiform Face Area or FFA; see 19 for example] or for body parts [Extrastriate Body Area or EBA; 10] has been reported. In addition the FFA was activated during face perception in 2-month-old infants using Positron Emission Tomography (PET [40]), and a recent study reported different electrophysiological response to normal and scrambled bodies in 3-month-old infants, suggesting that very young infants already recognize faces and bodies [12].

The visual processing of information and in particular the difficult computational problem of the rotation of point of view when watching hands from self and from others, which constitutes one of the main difficulties for modelling imitation in an associationist framework, was largely simplified by the segmentation of the visual input. On the basis of very young infants' abilities, we postulated that their visual system could segment the visual field so as to isolate a region of interest on the basis of its intrinsic features, in our case a hand. Alternatively, some motor control schemes, such as proposed for reaching [33], could be extended for viewpoint independent hand posture control. Thus the issue of the origin of mechanisms used to segment visual input and isolate visible body parts is still opened. It is theoretically possible to use a similar visuo-motor associative network approach using motor babbling and first-person perspective visual feedback to investigate whether it can learn to recognize its own body parts [23]. As for imitation, self-observation could bootstrap recognition of visible body parts.

From posture to complex action

Considering the attraction of neonates to dynamic stimuli [11], we believe the extension of the current network to dynamic actions is crucial to further explore the current hypothesis in a biologically valid framework. Dynamic features may also facilitate the acquisition of a body schema through visuomotor associative learning, an hypothesis that could be investigated with our model. Dynamic actions require the addition of time as a variable to the network as the current system does not consider the timing of events, and uses static sensory stimuli. More widely explored, in particular in robotics, is the use of motion, the time variation of visual stimuli, as input for imitation [1,23]. A more general system would need to learn spatio-temporal relations over the sensory stimuli, for example by utilizing spatio-temporal associative memories [5,27]. However the use of these networks for complex and highly correlated set of movement patterns is a challenging task that constitutes a logical extension of the current approach.

Here we assumed that the network would be in the learning phase first, and then in the testing phase. In a biological setting learning and testing ensue in an interleaved manner. Additional iterations, similar to reciprocal imitation games during development, could improve the efficiency of the system. We can speculate on the recursive learning possibility provided by the generalization ability. For example, learning visuomotor patterns A and B by self observation provides the ability to imitate a novel action C. Then by self-observation associative learning of C, the repertoire of visuomotor patterns can be expanded to store A, B and C. This expanded network could then allow more new actions D, E to be imitated and learnt. It would be interesting to expand the one-iteration of generalization explored in this article by using such a recursive learning mechanism, which would illustrate a “ratchet effect” at the level of individuals.

5 Conclusion

This report emphasizes that robotic systems are suited for testing hypotheses about human visuomotor development by showing that a biologically realistic neural network coupled with a simple visual system and a motor apparatus is capable of producing imitative skills. The current study is the first step towards an imitative agent that ‘lives’ in social learning loop, where the recursive nature of imitation can be expressed by the agent, better emulating the learning cycle of a growing infant in a socially active environment. The works building on this study then could help answer questions about imitation including the self-observation versus innate interpretations of infants’ imitative abilities.

Figures legends

Figure 1. Flow of visual and motor information, which form the input pattern when combined together. The whole retina and all fingers are associated to form the input pattern of the HHOP network.

Figure 2. Illustration of the Higher order Hopfield Network (HHOP) with three units i, j and k. The weight w_{ijk} represent the higher-order effect of units j and k on unit i.

Figure 3. The network framework built around Gifu Hand comprises three computers, to capture video (C), to control the Gifu hand (S), and to run the simulation (H) respectively. Arrows indicate the flow of information illustrated in Figure 2.

Figure 4. First (top) and second (bottom) sets of 4 finger postures used in the real-life experiment. For each set, the top row shows the image of the posture shown by the robotic hand which is recorded by the camera, and the bottom row one example of the visual part of the retina after preprocessing of the video signal.

Figure 5. A Percentage of bits correctly recalled as a function of the noise added. Dotted lines correspond to individual postures and straight line with squares to the average across the 16 postures. B Percentage of generalization (number of new postures imitated divided by the total number of postures not experienced by the network during training) as a function of the number of postures used for training. Other details are available in the methods section of the main text.

Table

Table 1. Percentage of finger configuration correctly reproduced by the HHOP networks during testing. The first column indicate which set of postures were used to train the HHOP networks (simple or complex sets of Gifu hand's patterns), the second, which set was used for testing the networks (Simple or Complex sets of hand patterns), the following columns whose hand was presented when testing the network (Gifu hand, Wooden hand, or one of the experimenters', EO and TC, hand). The last column gives the average correct recollection rate.

Set of postures	Gifu hand	Wooden hand	EO	TC	Average
First	100	82	82	98	90.7
Second	100	99	95	94	97.8

6 References

[1] P. Andry, P. Gaussier, S. Moga, J.-P. Banquet and J. Nadel, Learning and Communication in Imitation: An Autonomous Robot Perspective, IEEE Transaction on Systems, Man and Cybernetics, Part A: Systems and Humans 31 (2001) 431-444.

[2] M. Anisfeld, Only tongue protrusion modeling is matched by neonates, Developmental Review 16 (1996) 149-161.

[3] M.A. Arbib, From monkey-like action recognition to human language: an evolutionary framework for neurolinguistics, Behav Brain Sci 28 (2005) 105-124; discussion 125-167.

[4] A. Billard, Y. Epars, S. Calinon, S. Schaal and G. Cheng, Discovering optimal imitation strategies, Robotics and Autonomous Systems 47 (2004) 69-77.

[5] A. Billard and G. Hayes, DRAMA, a Connectionist Architecture for Control and Learning in Autonomous Robots, Adaptive Behavior 7 (1999) 35-63.

[6] S.J. Blakemore and C. Frith, The role of motor contagion in the prediction of action, Neuropsychologia 43 (2005) 260-267.

[7] R.W. Byrne and A.E. Russon, Learning by imitation: A hierarchical approach, Behavioral and Brain Sciences 21 (1998) 667-721.

[8] T. Chaminade, D. Meary, J.P. Orliaguet and J. Decety, Is perceptual anticipation a motor simulation? A PET study, Neuroreport 12 (2001) 3669-3674.

[9] J. Decety, T. Chaminade, J. Grezes and A.N. Meltzoff, A PET Exploration of the Neural Mechanisms Involved in Reciprocal Imitation, Neuroimage 15 (2002) 265-272.

[10] P.E. Downing, Y. Jiang, M. Shuman and N. Kanwisher, A cortical area selective for visual processing of the human body, Science 293 (2001) 2470-2473.

[11] R.L. Freedland and J.L. Dannemiller, Detection of stimulus motion in 5-month-old infants, J Exp Psychol Hum Percept Perform 13 (1987) 566-576.

[12] T. Gliga and G. Dehaene-Lambertz, Structural encoding of body and face in human infants and adults., J Cogn Neurosci 17 (2005) 1328-1340.

[13] K.Y. Haaland, D.L. Harrington and R.T. Knight, Neural representations of skilled movement, Brain 123 (2000) 2306-2313.

[14] U. Halsband, J. Schmitt, M. Weyers, F. Binkofski, G. Grutzner and H.J. Freund, Recognition and imitation of pantomimed motor acts after unilateral parietal and premotor lesions: a perspective on apraxia, Neuropsychologia 39 (2001) 200-216.

[15] C. Heyes, Causes and consequences of imitation, Trends Cogn Sci 5 (2001) 253-261.

[16] J.J. Hopfield, Neural networks and physical systems with emergent collective computational abilities, Proc Natl Acad Sci U S A 79 (1982) 2554-2558.

[17] M. Iacoboni, R.P. Woods, M. Brass, H. Bekkering, J.C. Mazziotta and G. Rizzolatti, Cortical mechanisms of human imitation, Science 286 (1999) 2526-2528.

[18] P. Jacob and M. Jeannerod, The motor theory of social cognition: a critique, Trends in Cognitive Sciences 9 (2005) 21-25.

[19] N. Kanwisher, J. McDermott and M.M. Chun, The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception, J. Neurosci. 17 (1997) 4302-4311.

[20] J.D. Keeler, Comparison Between Kanerva's SDM and Hopfield-type Neural Networks, Cognitive Science 12 (1988) 299-329.

[21] C. Keysers and D.I. Perrett, Demystifying social cognition: a Hebbian perspective, Trends Cogn Sci 8 (2004) 501-507.

[22] J.M. Kilner, Y. Paulignan and S.J. Blakemore, An interference effect of observed biological movement on action, Current Biology 13 (2003) 522-525.

[23] Y. Kuniyoshi, Y. Yorozu, M. Inaba and H. Inoue, From Visuo-Motor Self Learning to Early Imitation - A Neural Achitecture for Humanoid Learning. International Conference on Robotics & Automation, IEEE, Taipei, Taiwan, 2003.

[24] A.N. Meltzoff and J. Decety, What imitation tells us about social cognition: a rapprochement between developmental psychology and cognitive neuroscience, Philos Trans R Soc Lond B Biol Sci 358 (2003) 491-500.

[25] A.N. Meltzoff and M.K. Moore, Imitation of facial and manual gestures by human neonates, Science 198 (1977) 74-78.

[26] R.C. Miall and D.M. Wolpert, Forward Models for Physiological Motor Control, Neural Networks 9 (1996) 1265-1279.

[27] M. Morita and A. Suemitsu, Computational modeling of pair-association memory in inferior temporal cortex, Cognitive Brain Research 13 (2002) 169-178.

[28] M. Myowa-Yamakoshi, M. Tomonaga, M. Tanaka and T. Matsuzawa, Imitation in neonatal chimpanzees (Pan troglodytes), Dev Sci 7 (2004) 437-442.

[29] J. Nadel, A. Revel, P. Andry and P. Gaussier, Toward communication: First imitations in infants, low-functioning children with autism and robots, Interaction studies 5 (2004) 45-74.

[30] M. Ogino, H. Toichi, Y. Yoshikawa and M. Asada, Imitation faculty based on a simple visuo-motor mapping towards interaction rule learning with a human partner. Proceedings in The Social Mechanisms of Robot Programming by Demonstration, Workshop in IEEE International Conference on Robotics and Automation, 2005.

[31] E. Oztop, A New Content Addresable Memory Model Utilizing High Order Neurons. Computer Engineering, Master Thesis, Middle East Technical University, Ankara, 1996.

[32] E. Oztop, D. Franklin, T. Chaminade and C. Gordon, Human-humanoid interaction: is a humanoid robot perceived as a human, Internation Journal of Humanoid Robotics 2 (2005) 537-559.

[33] E. Oztop, D. Wolpert and M. Kawato, Mental state inference using visual control parameters, Brain Res Cogn Brain Res 22 (2005) 129-151.

[34] M. Petrides, G. Cadoret and S. Mackey, Orofacial somatomotor responses in the macaque monkey homologue of Broca's area, Nature 435 (2005) 1235-1238.

[35] G. Rizzolatti and M.A. Arbib, Language within our grasp, Trends in Neurosciences 21 (1998) 188-194.

[36] G. Rizzolatti, L. Fadiga, V. Gallese and L. Fogassi, Premotor cortex and the recognition of motor actions, Cognitive Brain Research 3 (1996) 131-141.

[37] S. Schaal, Is imitation learning the route to humanoid robots?, Trends in Cognitive Sciences 3 (1999) 233-242.

[38] A. Slater and R. Kirby, Innate and learned perceptual abilities in the newborn infant, Experimental Brain Research 123 (1998) 90-94.

[39] M. Tomasello, The Cultural Origins of Human Cognition, Harvard University Press, 2001.

[40] N. Tzourio-Mazoyer, S. De Schonen, F. Crivello, B. Reutter, Y. Aujard and B. Mazoyer, Neural correlates of woman face processing by 2-month-old infants, Neuroimage 15 (2002) 454-461.

[41] A.L. van der Meer, F.R. van der Weel and D.N. Lee, The functional significance of arm movements in neonates, Science 267 (1995) 693-695.

[42] D.M. Wolpert, K. Doya and M. Kawato, A unifying computational framework for motor control and social interaction, Philosophical Transaction of the Royal Society of London B Biological Sciences 358 (2003) 593-602.

[43] D.M. Wolpert, Z. Ghahramani and J.R. Flanagan, Perspectives and problems in motor learning, Trends Cogn Sci 5 (2001) 487-494.

[44] D.M. Wolpert and M. Kawato, Multiple paired forward and inverse models for motor control, Neural Networks 11 (1998) 1317-1329.