

Towards a Shared Action Code for Human-Robot Interaction

Roberto Prevete, Giovanni Tessitore, Ezio Catanzariti, Guglielmo Tamburrini
Dipartimento di Scienze Fisiche, Università di Napoli Federico II, Italy
{prevete, tessitore, ezio, tamburrini}@na.infn.it

Abstract

Recent experiments show that the observation of robots controlling arm and hand movements in order to reach and manipulate objects activates the human mirror system. This suggests the possibility of developing a basic action code which is shared between humans and robots. What are the central features of this code, and its functional roles in the complex processes of action generation and understanding? A computational framework for addressing these issues is presented and critically compared to extant computational models of mirror neurons.

1. Background and Motivations

Motor cognition comprises both the processes involved in planning and executing our own actions and the processes involved in understanding and predicting actions carried out by other agents [12]. Crucial functionalities of motor cognition in both human and non-human primates appear to be supported by the neural mechanisms of cortical areas that jointly form so-called mirror systems [9]. Recent experimental findings provide evidence that the observation of robots controlling arm and hand movements in order to reach and manipulate objects (object-directed actions) activates the human mirror system [1]. Additional evidence for robots activating human mirror systems, notably in non-object-directed action conditions, is provided in [4].

These findings on mirror system response to robot actions open up a new research perspective in human-robot interaction (HRI): by endowing robots with functionalities of the human mirror system, one may now aim at establishing a basic action code which is shared between humans and robots. The operation of this shared action code in HRI environments may facilitate epistemic and intentional modeling of others by both human and robotic agents involved in HRI scenarios, thereby leading to improved mutual action legibility, and the development of trust relationships in human-robot multi-agent systems.

In order to evaluate in a more detailed manner the potential contribution of computational modeling of mirror systems to HRI inquiry, we turn to examine now the following questions: What are the central features of a mirror system action code? And what are its functional roles in the complex processes of action generation and understanding?

Salient features of this action code are identified in section 2, chiefly by reference to inquiries into primate mirror systems. In section 3, perceptual (typically visual) inputs that enable one to classify an object-oriented action are critically discussed in their relationship to biological data about mirror neurons behavior. Perceptual input data for mirror systems are examined in more detail in section 4, where a computational model is presented for a view-independent visual classification of hand-state features. In section 5, a mirror system model is outlined, which crucially includes an expected perception module relying on a shared action code. This model is critically compared to extant computational models of mirror systems, and its potential for providing a causal account for a wider variety of biological data is highlighted.

2. Primate Action Codes

The development of an action code which is shared and used by humans and robots presupposes that (1) there is such an action, (2) one identifies its central mechanisms, and (3) these mechanisms are sufficiently flexible to be exported from human-human to human-robot interaction contexts.

Let us briefly summarize the evidence for (1), which is provided by biological inquiries into mirror systems. The more extensive neuroscientific data and models concern macaque pre-motor cortical areas that are strongly involved in controlling arm, hand, and mouth movements in object-directed action [5, 9, 10]. These studies provide evidence for a common neural substrate representing both observed and executed actions. In particular, a population of neurons was discovered in macaque pre-motor area F5 which are active (high spike rate) when the monkey executes an object-

directed action *or* observes another individual (either monkey or human experimenter) executing an object-directed action. In view of their characteristic activation properties, these neurons were called mirror neurons. Mirror properties have been identified in neural cells from other cortical regions too. On this account, the cortical circuits including F5 and these additional cortical areas have been collectively called mirror system [9, 10].

Additional evidence for a neural substrate coding for action properties is supported by finer-grained motor properties of F5 neurons, that is, by distinctive forms of neural activation during the execution of object-directed actions. F5 neurons usually respond with high spike rates during the execution of object-directed actions classified in terms of their overall goal (such as grasping, tearing, manipulating, holding) [9,10]. In particular, most of these neurons respond to just one type of object-directed action.

Additional specificities of mirror neuron activation have been detected. Consider reach-to-grasp actions: the set of F5 neurons that are involved in grasping a sphere using the *whole hand* and all fingers in opposition differs from the set of F5 neurons involved in a *precision-grip* grasping of a cylinder, as the latter requires the opposition of thumb and forefinger only. Moreover, many F5 neurons discharge during movements that have a similar goal (for example, grasping some object), but are performed with different effectors (for example, grasping an object with one's own right hand, left hand, or even mouth). Finally, it turns out that many F5 neurons correlate with different action phases: some neurons exhibit a high spike rate during hand-opening phases preceding closure phases, some neurons discharge mainly during hand closure, and some other neurons discharge throughout action execution. In accordance with these findings, one may cluster area F5 into different subsets of neurons, each cluster being associated to different aspects of the action temporal segmentation [10,15].

Taken together, these findings suggest that F5 codes for some sort of *vocabulary of movements* [5,15]. But what is this vocabulary of movement and action types used for? We advance an answer to this question in terms of the following hypothesis, which appears to be consistent with findings and functional hypotheses about mirror neurons: the mirror system organizes movement and action tokens into sequences which represent object-oriented actions. Accordingly, *sequences* of movement and action tokens are processed by the mirror system, each sequence being a code for an individual action. In a sequence, the initial token may stand for, say, a general *type of action* (hold,

grasp, tear, manipulate), another token for how the effector must adapt to the grasped object, that is, for distinctive features of the *action mode* (precision grip, whole hand grip, etc), another token yet for effector configuration at action beginning, and so on.

Is this action code, hypothesized on the basis of motor properties of F5 neurons, a *shared* action code? An affirmative answer to this question is suggested by F5 mirror neuron behavior during action observation. Mirror neurons have been classified as *strictly congruent* and *broadly congruent* [9], according to whether their perceptual (typically visual) and motor properties do correspond or do not correspond to each other in terms of both type and mode of action. The above findings, mostly based on single-neuron recording experiments, suggest the presence of a shared action code in monkeys. Evidence for a similar mirror system and shared action code in humans flows from a wide variety of neurophysiological and brain-imaging experiments [9]. One should be careful to note, however, that this evidence is less direct, insofar as single-neuron recordings have not been performed in humans, and one significantly draws on analogies with the monkey model.

3. Visual data and action understanding

A central functional role hypothesized for the mirror system concerns *action understanding* [9]. The mirror system is supposed to play a crucial role in the process of understanding actions performed by others. More specifically, the mirror system is involved into a mechanism which transforms perceptual (typically visual) information (observed action by another individual) into activity of premotor cortices (mirror neuron activity) which is akin to the activity generated in executing congruent or coincident actions. Accordingly, action understanding involves a transformation from visual information to a motor coding whose effects are known to the observing agent. These activations of mirror neurons are usually interpreted as codes for action *goals*, which require no strict correspondence with visual features associated to the observed action phases. Thus, visual features are subservient to the understanding of an observed action in terms of its goal, but their processing role can be filled in by non-visual perceptual cues as well. For example, if action understanding is possible on the basis of sounds (produced, say, in tearing a sheet of paper), then mirror neurons signal the action even in the absence of visual stimuli [9,10].

This interpretation of mirror neurons as *codes for the goal of an action* requires careful critical

examination in view of the behavioral properties of broadly congruent mirror neurons, which are triggered by observed actions which do not correspond, in terms of both type and mode, to the actions they code motorically for. Some broadly congruent mirror neurons respond to observed and executed actions widely differing from each other. For instance, some of these neurons discharge for goalwise different actions, when the monkey observes another monkey placing an object onto a table *or* performing a reach-to-grasp action. These goal-incongruent behaviors have been accounted for in the framework of a prevailing interpretation of mirror neurons as *codes for action goals* by claiming that these mirror neurons represent actions which are logically or conceptually linked to each other [10]. According to an alternative interpretation [8], the behavior of these broadly congruent mirror neurons is accounted for as the coding of actions which share similar effector/target configurations. Consider the putting down and the grasping of objects. In this alternative view, mirror neurons fire in both conditions because the pattern corresponding to the final part of object-putting-down actions is similar to the pattern corresponding to the final part of object-grasping action.

These various interpretations illustrate extant gaps in our current understanding of what mirror neurons code for and the causal mechanisms of mirror systems. This incomplete understanding is reflected into different accounts of mirror behavior provided by computational models, which we turn now to consider. To begin with, we explore aspects of the perceptual pre-processing needed by the mirror system.

4. Visual processing for mirror activation

In various computational models of mirror neurons, the characteristic matching of activity occurring during both action execution and observation is achieved by means of the *same effector/target description* [3, 5, 6, 8]. For example, in [3], the effector/target description is provided by means of a feature vector, called *hand state*, which assumes the same values during both action execution and observation. The hand state is input to a classification module which outputs mirror-like behavior: the hand state (the effector/target description) brings about the same behavior of the classification module during the observation of both one's own action and the same action performed by another agent. Note that in this model mirror behavior is caused by the *observation* of some actions performed by oneself and by others, rather than by execution and observation of some given actions.

The ability to compute the same effector/target description is based on the possibility of measuring view-invariant visual features related to the hand, to the object, or to hand-object pairs. Consider, for example, the specific grasp action of *grasping food by a precision grip*; consider, moreover, just one feature of the hand-state vector, the grip-size (that is, the aperture between index finger and thumb). The measured value of the grip-size must be “almost” the same during both observation of the self-executed action and observation of the same action performed by another agent, insofar as the same effector/target description must be extracted in either case. Thus, a mechanism is needed to extract the same grip-size value from different visual inputs (relative to the agent's viewpoint when performing the action and when observing the same action performed by another agent).

The need for a view-invariant feature extraction system is further motivated by the developmental interpretation of mirror neurons [5] and by neurophysiological data regarding the temporal area STS, which is connected to F5 via parietal area PF. In particular, in [6] it is shown that STS contains mirror-like cells selectively responding to a wide variety of body movements, including object-directed actions. These neurons respond during the observation of both self-executed actions and actions performed by others. Unlike mirror neurons, STS neurons do not respond when a closed-eyed monkey executes an object-directed action, that is, STS neurons lack motor properties. It has been argued that STS provides a “pictorial description” or high-level visual features of the ongoing action to be processed by the mirror system. These findings suggest that STS is likely to be involved in the computation of a view-invariant effector/object description of an ongoing action. It is not clear, however, which features are included in this description and how they are computed.

An initial simplification of the hand control problem is achieved in [16] by introducing the concept of “virtual finger” as a mechanism enabling one to reduce the number of degrees of freedom, and thereby the complexity of the hand control problem. Various neurophysiological data are consistent with a simplified hand control mechanism: the response of several hand muscles is evoked during the stimulation of one site in the primary motor cortex, and a single neuron in primary motor area F1 generally discharges in connection with multiple finger movements.

These various findings jointly support the idea that a simplified control strategy is reflected, at the output stage, in the reduction of the shapes that the hand can assume and, thus, in the number of features that are

needed to describe hand movements. This seems to be confirmed by the work of Santello [11]. Santello carried out a principal component analysis of hand features showing that the first principal component is sufficient to explain most hand feature variability, while the first two principal components explain most of whole hand feature variability.

These observations suggest that relatively few features suffice to monitor the evolution of hand motion during reach-to-grasp action. In [2] grasping actions are supposed to be coded in terms of changes in grip aperture, i.e., the separation between the thumb and the index finger. It is shown that during a reach-to-grasp action, grip aperture increases at the initial stage until a maximum value that exceeds the object dimension is reached, and then gradually decreases until the object’s actual size is matched. Moreover, the time at which the thumb-finger opening is largest (maximum grip size) is linearly correlated to the object size and always occurs within 60-70% of the grasp action duration.

Accordingly, grip aperture appears to be a good candidate to describe overall hand movement, since it concerns the fingers (thumb and index finger) that are primarily involved in grasp actions, while the position of other fingers can be computed indirectly because of their correlation to the grip movement, resulting from the above mentioned synergic hand mechanism.

The computational model for extracting grip-size that we now turn to describe is view-independent., and inspired to the functional modeling of temporal areas IT and STS. However, an analysis of its biological plausibility goes clearly beyond the scope of this paper. In our approach, we assume that grip-aperture can be measured as the superposition of a small number of basic hand-shapes corresponding to predefined grip-apertures. For example, Fig. 1 shows three basic hand-shapes: fully open grip-size (BS_1), middle grip-size (BS_2), and fully closed grip-size (BS_3).

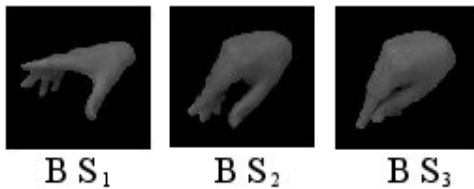


Figure 1. Basic hand shapes for a given viewpoint. From left to right: fully open grip-size, middle grip-size and fully closed grip-size.

The basic idea underlying this model is to create view-independent units (GS units) that are selective to basic hand-shapes (BS units), and to integrate the output of these units to compute grip-aperture. This is

done by interpreting the output of the GS units, given a novel hand-shape, as a similarity measure to each basic hand-shape, and integrating this information to estimate the grip-aperture in a view-independent manner. View-independence of GS units is obtained by a “pooling operation” over a set of view-dependent units (GV units). One should be careful to note that the basic hand-shapes do not depend on the specific action.

The overall approach can be schematized as follows. There is a set of computing units organized in hierarchical way into three ordered layers. The first layer is composed of view-dependent units, called GV, which are selective to a specific hand-shape, and generalize across transformation of the preferred stimulus to changes in scale and position (but not to change in viewpoints). The second layer is composed of viewpoint independent units, called GS, selective to a basic hand shape. A GS unit generalizes across transformation of the preferred stimulus to changes in scale, position, and view-points. The third layer is composed of just one unit (called GA unit). The GA unit’s input is the similarity measure, computed by the GS units, of the current shape with respect to all basic hand-shapes. On the basis of these inputs the GS units compute the grip-aperture.

An implemented version of this model is described in [7], and a series of tests are reported, which were performed in order to verify the model’s effectiveness in computing both the actual grip-aperture and the view-independent property. It turns out that grip-aperture values as measured by this system satisfy the following properties [2]: the maximum grip aperture values bear a linear relation to the target dimension (linear regression, determination index ≈ 0.90), and the values of the maximum grip size occur, on the average, at roughly 80% of action duration. Moreover, grip aperture as measured by this system assumes an essentially viewpoint independent value (see Figure 2).

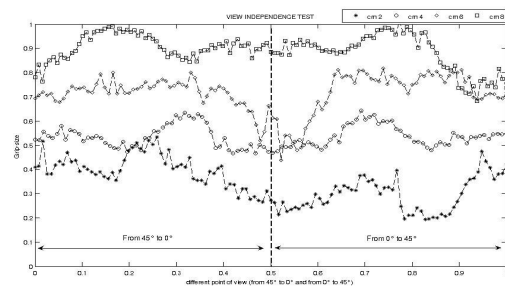


Figure 2. Viewpoint Independence: Grip-size values measured for four different actual grip sizes versus different viewpoints

5. MS computational modeling

It was mentioned above that several computational models inspired by mirror system findings have been introduced. Some of these, see [5] for a review, directly link mirror activity to motor activity, i.e. mirror activity causes motor activity. While accounting for the “mirror” property of the mirror system (same response on both action execution and observation), these models fail to account for other significant data, notably concerning the behavior of broadly congruent mirror neurons [9], and the fact that mirror neuron inhibition causes a slowdown of motor activity only, without utterly preventing action execution [14].

One of the more interesting and detailed mirror system models to date (the Oztop-Arbib model, OA from now on) was presented in [3]. OA is based on three chief assumptions: (1) an interpretation of canonical neurons as computing a hand-program aimed at controlling hand movements on the basis of object features processed in area AIP and oriented to performing an action; (2) the *hand state* hypothesis, according to which the observed action is coded by a vector of observer-independent features (hand-state); under this hypothesis, the moving hand can either be another agent's hand or else the observer's own hand; (3) a functional interpretation of mirror neurons as a recognition module classifying actions on the basis of a sequence of hand-states.

In OA, mirror neurons are basically assigned the role of classifying arm+hand movements with respect to a sequence of hand-states. In particular, mirror neurons are collectively modeled as a classification module which has to *observe* the whole action from beginning to end in order to give the correct responses, in contrast with the biological findings that mirror neurons can present a high spike rate at different phases of an observed action [10]. Moreover, in the OA model mirror neuron inhibition results into the suppression of classification module activity, while the hand program is unaffected. This behavior contrasts with the fact that mirror neurons inhibition causes motor slowdown only [14].

Overall, the OA functional interpretation of mirror neurons as a classification system appears to require some amendments in order to account for a wider variety of known experimental findings. A step in this direction was undertaken in [8], by modeling the mirror system as a system crucially including an *expected perception* mechanism [13]. In a sensory-motor control loop involving an expected perception mechanism, pre-planned action execution is monitored by means of computationally inexpensive comparisons between

actual and expected perceptions. Sensory data are more extensively processed for the purpose of action re-planning only, when a mismatch occurs between actual and expected perception.

The proposed model is based on the supposition that an object-directed action can be segmented and represented in terms of a sequence of *base units*. These units enable one to generate expected sensory features of motor patterns. For example, one may represent a grasping action by a precision grip as a sequence composed of two base units, representing “grip aperture equal to a value exceeding object dimensions” and “grip aperture equal to a value matching object dimensions” respectively. Each base unit is used by a classification mechanism, which verifies whether the actual perception generated by the motor pattern associated to the base unit corresponds to the expected perception associated to the base unit.

In this model, mirror neurons do not code for action goals only. Rather, their behavior is interpreted in terms of action base units and their sequences. Notice that in this model a temporal relationship holds between mirror activity and action unfolding. This is consistent with biological data showing that there is a strict temporal coupling between F5 neuronal activity and the temporal unfolding of an object-direct action.

A related functional interpretation of mirror system involvement in forward sensory prediction is provided in [5]. There one assigns mirror neurons a twofold role: 1) while observing object-directed action, mirror activity enables mind reading inferences; 2) while executing object-directed action, mirror activity enables one to eliminate sensory processing delays in motor control processes. One should be careful to note that canonical neurons are supposed to process movement plans during both action execution and observation according to this interpretation. This is controversial, however, insofar as available data support canonical activation during action execution only. In contrast with this, in the conceptual model we are advancing here one has that 1) an action is represented by mirror neurons in terms of a sequence of perceivable effects (expected perceptions) selected by canonical neurons on the basis of object sight only, and 2) the activity of canonical neurons is not required while observing object-directed actions.

Overall, this functional interpretive framework promises to overcome limitations of other (computational) models. In particular, this framework enables one to sketch out a causal account for a wider variety of biological data, notably including the following: (i) mirror neurons are known to be active in both action execution and observation, (ii) inhibiting

mirror neurons causes motor slowdown, without utterly preventing action execution, (iii) mirror neurons may manifest behaviors that are only broadly congruent, (iv) there are clear temporal correlations between mirror activity and the unfolding of an action.

Within the broad framework outlined in this final section, there are many crucial design and implementation problems that have to be addressed in order to move ahead towards a shared action code for HRI based on mirror system functionalities. In particular, a better understanding is needed of experiments showing that robots engaged in object-directed actions activate the human mirror system [1, 4]. There, central interpretive issues concern the role, if any, of a tight match between robotic arm kinematics and primate arm kinematics, and the need, if any, to avoid exact repetitions of robotic movements. One can apparently dispense with a tight match insofar as complex actions are concerned, which were found to determine mirror system activation. However, it is suggested in [1] that a tighter match between human and robotic kinematics may be required in the context of less complex actions, when contextual information is not sufficiently rich to be diagnostic of action goal independently of kinematic information. These “simpler” actions, as emphasized in [12, 17], are just actions whose interpretation seems to rely primarily on the mirror system, more complex action understanding crucially involving cortical pre-frontal cortical processing. These findings suggest some specific functional roles of a shared action code exclusively based on mirror system functionalities, thereby pointing to both potentialities and limitations of what can be achieved in the way of mutual action legibility by deploying such shared action codes in the context of HRI.

Acknowledgments. The authors wish to thank Edoardo Datteri and Matteo Santoro for stimulating discussions on mirror systems and biological anticipation mechanisms.

6. References

- [1] V. Gazzola, G. Rizzolatti, B. Wicker, and C. Keysers, "The anthropomorphic brain: The mirror neuron system responds to human and robotic actions," *NeuroImage*, vol. 35, 4, pp. 1674-1684, 2007.
- [2] M. Jeannerod, "The timing of natural prehension movements," *Journal of Motor Behavior*, vol. 16(3), 3, pp. 235-54, 1984.
- [3] E. Oztop and M. A. Arbib, "Schema design and implementation of the grasp-related mirror neuron system," *Biological Cybernetics*, vol. 87, pp. 116-140, 2002.
- [4] L. M. Oberman, J. P. McCleery, V. S. Ramachandran, and J. A. Pineda, "EEG evidence for mirror neuron activity during the observation of human and robot actions: Toward an analysis of the human qualities of interactive robots," *Neurocomput.* vol. 70, 13-15, pp. 2194-2203, 2007.
- [5] E. Oztop, D. M. Kawato, and M. Arbib, "Mirror neurons and imitation: a computationally guided review", *Neural Networks*, vol. 19, pp. 254-271, 2006.
- [6] C. Keysers and D. I. Perrett, "Demystifying social cognition: a Hebbian perspective," *Trends Cogn Sci*, vol. 8, iss. 11, pp. 501-507, 2004.
- [7] R. Prevede, M. Santoro, E. Catanzariti, and G. Tessitore, "A Neural Network Model for a View Independent Extraction of Reach-to-Grasp Action Features," *Advances in Brain, Vision and Artificial Intelligence*, Mele F. et al. (Eds.) LNCS 4729 Springer, pp. 124-133, 2007.
- [8] R. Prevede, M. Santoro, and F. Mariotti, "A biologically inspired visio-motor control model based on a deflationary interpretation of mirror neurons," Proceedings of COGSCI05 (XXVII Annual Meeting of the Cognitive Science Society), Stresa, Italy, July 21-23, 2005.
- [9] G. Rizzolatti and L. Craighero, "The mirror-neuron system." *Annual Rev Neurosci*, vol. 27, pp. 169-192, 2004.
- [10] G. Rizzolatti and C. Sinigaglia, *Mirrors in the brain: How our minds share actions, emotions, and experience*, Oxford University Press, Oxford, 2007.
- [11] M. Santello, M. Flanders, and J. F. Soechting, "Patterns of hand motion during grasping and the influence of sensory guidance," *Journal of Neuroscience*, vol. 22, 4, pp. 1426-1235, 2002.
- [12] J. A. Sommerville and J. Decety, "Weaving the fabric of social interaction: Articulating developmental psychology and cognitive neuroscience in the domain of motor cognition," *Psychonomic Bulletin & Review*, vol. 13, 2, pp. 179-200, 2006.
- [13] E. Datteri, G. Teti, C. Laschi, G. Tamburrini, P. Dario, E. Guglielmelli, "Expected Perception: an anticipation-based perception-action scheme in robots", *Proc. IROS 2003*, 2003.
- [14] L. Fogassi, V. Gallese, G. Buccino, L. Craighero, L. Fadiga, and G. Rizzolatti, "Cortical mechanism for the visual guidance of hand grasping movements in the monkey. A reversible inactivation study," *Brain*, vol. 124, pp. 571-586, 2001.
- [15] L. Fadiga, L. Fogassi, V. Gallese, and G. Rizzolatti, "Visuomotor neurons: ambiguity of the discharge or 'motor' perception?" *International Journal of Psychophysiology*, vol. 35, pp. 165-177(13), 2000.
- [16] M. A. Arbib, T. Iberall, and D. M. Lyons, "Coordinated control programs for movements of the hand," in A. W. Goodwin and T. Darian-Smith eds., *Hand Function*, Springer Verlag, Berlin, 1985, pp. 111-129.
- [17] T. Zalla, P. Pradat-Diehl, A. Sirigu, "Perception of action boundaries in patients with frontal lobe damage", *Neuropsychologia* vol. 41, 1619-1627.