Cogn Comput (2011) 3:419–435
DOI 10.1007/s12559-010-9082-z

Embodied Gesture Processing: Motor-Based Integration of Perception and Action in Social Artificial Agents

Amir Sadeghipour · Stefan Kopp

Received: 3 May 2010 / Accepted: 14 October 2010 / Published online: 11 November 2010
© The Author(s) 2010. This article is published with open access at Springerlink.com

Abstract A close coupling of perception and action processes is assumed to play an important role in basic capabilities of social interaction, such as guiding attention and observation of others' behavior, coordinating the form and functions of behavior, or grounding the understanding of others' behavior in one's own experiences. In the attempt to endow artificial embodied agents with similar abilities, we present a probabilistic model for the integration of perception and generation of hand-arm gestures via a hierarchy of shared motor representations, allowing for combined bottom-up and top-down processing. Results from human-agent interactions are reported demonstrating the model's performance in learning, observation, imitation, and generation of gestures.

Keywords Computational model · Interactive artificial agents · Nonverbal communication · Gestures · Perception-action links

Introduction

In social interactions, one is continuously confronted with an intricate complexity of verbal and nonverbal behavior, including hand-arm gestures, body movements or facial expressions. All of these behaviors can be indicative of the other's referential, communicative, or social intentions [1]. In this paper, we focus on hand-arm gestures. Interlocutors in social interaction incessantly and concurrently produce and perceive a variety of gestures. The generation of a hand-arm gesture, coarsely, consists of two steps. First, finding the proper gesture for an intention that is to be realized under current context constraints. Second, performing the gesture using one's motor repertoire. Similarly, the recipient perceives and analyzes the other's movement both at motor and at intention levels. Cumulating evidence suggests that these two processes are not separate, but that recognizing and understanding a gesture is grounded in the perceiver's own motor repertoire [2, 3]. In other words, a hand movement is understood, at least partially, by evoking the motor system of the observer. This is evidenced by so-called *motor resonances* showing that the motor and action (premotor) systems become activated during both performance and observation of bodily behavior [4–6]. One hypothesis is that these neural resonances reflect the involvement of the motor system in deriving predictions and evaluating hypotheses about the incoming observations. This integration of perception and action enables imitating or mimicking the observed behavior, either overtly or covertly, and thus forms an embodied basis for understanding other embodied agents [7], and for communication and intersubjectivity of intentional agents more generally (cf. *simulation theory* [8]). Hence, perception-action links (and resulting resonances) are assumed to be effective at various levels of a hierarchical perceptual-motor system, from kinematic features to motor commands to goals and intentions [9], whereas these levels interact bi-directionally; bottom-up and top-down [10]. Further, a close perception-action integration can be assumed to support two important ingredients of social interaction: First, fast and often subconscious inter-personal coordinations (e.g., alignment, mimicry, interactional synchrony) that lead to rapport [11]

A. Sadeghipour (✉) · S. Kopp
Sociable Agents Group, Cognitive Interaction Technology (CITEC), Bielefeld University, P.O. Box 100 131, 33501 Bielefeld, Germany
e-mail: asadeghi@techfak.uni-bielefeld.de

S. Kopp
e-mail: skopp@techfak.uni-bielefeld.de

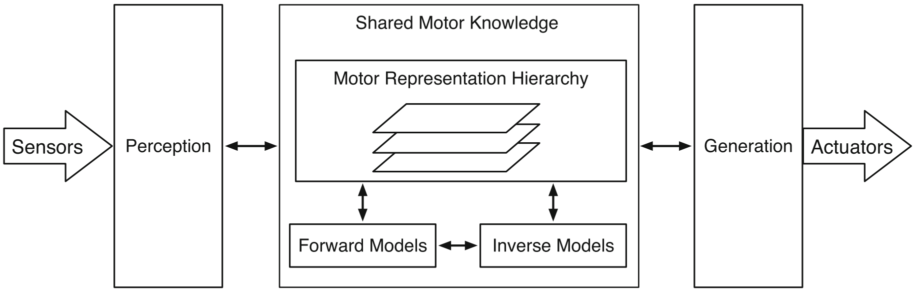


Fig. 1 Overall model for cognitive processes of embodied perception and generation, integrated in a shared motor knowledge

and social resonance [12] between interactants. Second, social learning of behavior by means of imitation, which helps to acquire and interactively establish behavior through connected perceiving, processing, and reproducing of their pertinent features. All of these aforementioned effects may also apply—at least to a certain extent—to the interaction between humans and embodied agents, be it physical robots or virtual characters (see [12] for a detailed discussion). For example, brain imaging studies [13, 14] showed that artificial agents with sufficiently natural appearance and movements can evoke motor resonances in human observers.

Against this background, we aim for interactive embodied systems ultimately able to engage in social interactions, in a human-like manner, based on cognitively plausible mechanisms. A central ingredient is a computational model for integrated perception and generation of hand-arm gestures. This model has to fulfill a number of requirements: (1) perceiving and generating behavior in a fast, robust, and incremental manner, (2) concurrent and mutually interacting perception and generation, (3) concurrent processing at different levels of motor abstraction, from movement trajectories to intentions; (4) incremental construction of hierarchical knowledge structures through learning from observation and imitation.

In this paper, we present a cognitive computational model that has been devised and developed to meet the above-mentioned requirements for the domain of hand-arm gestures. Focusing on the motor aspect of gestures, it should also serve as a basis for future modeling of higher cognitive levels of social intentions. In the section "Shared Motor Knowledge Model", we introduce the Shared Motor Knowledge Model that serves as a basis for integrating perception and action, both of which operate upon these knowledge structures by means of forward/inverse models. In "A Probabilistic Model of Motor Resonances" we present a probabilistic approach to simulate fast, incremental and concurrent resonances and their exploitation of these structures in both perceiving and generating behavior. Section "Perception-Action Integration" details how the integration of perception and action is achieved in this model and how this helps to model and cope with characteristics of nonverbal human social interaction. Results of applying this model to real-world data (marker-free gesture tracking) from a human-agent interaction scenario are reported in "Results". In the final section we discuss our work in comparison to other related work.

Shared Motor Knowledge Model

In previous work [15], we have presented a cognitive model for hierarchical representations of motor knowledge for hand-arm gestures, and we proposed how these structures can be utilized for probabilistic "embodied" behavior perception. Here, we present an extended version of this model that serves as a unified basis for both perception and generation of hand-arm movements (wrist position trajectories, to be specific) as they occur in natural gesturing by human users in interaction with a humanoid virtual agent. Overall, the model consists of three main modules (see Fig. 1): shared motor knowledge, perception and generation. This model allows for parallel gesture generation and perception processes grounded in shared motor knowledge. Further, the hierarchical model enables bottom-up processing (mainly for perceptual tasks) interacting bidirectionally with top-down processing (for action production as well as attention and perception guidance). In the remainder of this section, we describe each module separately.

Shared Motor Knowledge

The central shared motor knowledge module (see Fig. 2 for a detailed view) consists of a hierarchical representation of hand-arm gestural movements, and a pair of forward and inverse model submodules, interacting with the movement representation.

Shared Representation of Motor Knowledge

The representation of motor knowledge for hand-arm gestures is a hierarchical structure comprising different levels of abstraction, starting from the form of single gesture performances in terms of movement trajectories and leading into less contextualized motor levels, toward meaning. As illustrated in Fig. 2, the representation

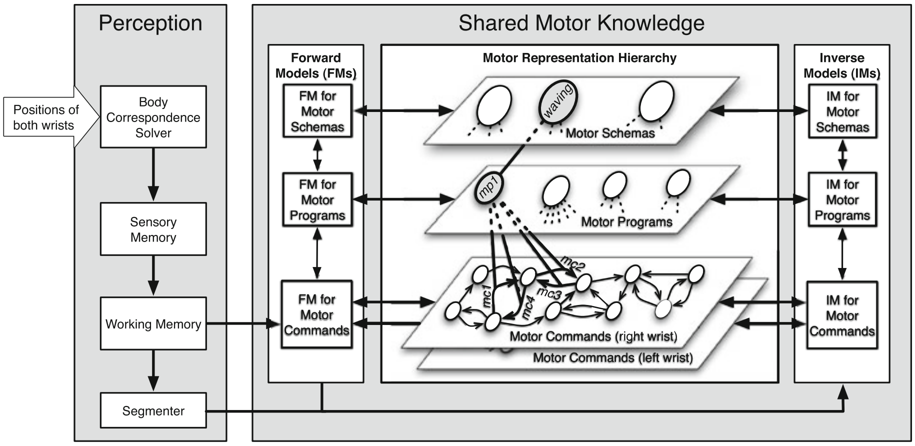


Fig. 2 Modules of "perception" and "shared motor knowledge" in detail. Shared motor knowledge comprises three different levels of abstraction representing motor knowledge as a hierarchical graph. The representation of a sample waving gesture is highlighted in *bold lines*

hierarchy consists of three levels: motor commands, motor programs and motor schemas.

At the lowest level, a directed graph is used to store motor knowledge about gestural movements for each hand. Nodes represent spatial positions of the wrists and edges represent trajectories of movement segments. Edges are referred to here as *motor commands* (MC). These are extracted from the trajectories of observed movements, regarding their velocity profiles. The resulting motor commands encode one of the three basic forms of movement segments: straight, curved or s-shaped trajectories. These motor commands only parametrize the significant spatiotemporal features of movement trajectories (e.g. extent, shape, timing) as needed by our motor control engine ACE [16]. In this way, a movement corresponds to a sequence of motor commands, i.e., a path through the graph. Neurobiological studies showed that motor systems use a similar principle of decomposing complex movements into simpler elements, called motor primitives, and performs them in parallel or sequence [17, 18].

The next level consists of nodes associated with *motor programs* (MP), each of which represents a whole performance of a gesture. That is, a motor program is associated with a sequence of motor commands for each hand. For example, the representation of a waving gesture could consist of four motor commands: raising the right hand (*mc1*), moving it to the right (*mc2*), moving it to the left (*mc3*) and retracting it back to the rest position (*mc4*). This performance of a waving gesture with the right hand can be represented by a motor program (*mp1*), which sequentially connects *mc1*, *mc2* and *mc3* repeatedly to represent the swinging movement, and finally *mc4*. A different gesture, like drawing a circle at the same height as waving, can also encompass some of these motor commands in its motor program (e.g., *mc1* and *mc4* for raising and lowering the right hand). That way, the agent has a compact representation of various gesture performances stored in its motor program repertoire.

However, in general, gestures are neither limited to a specific performance nor exhibit only invariant spatiotemporal features. The variant features are the performance parameters that, when varied, do not change the meaning and intention of the gesture but the way of performing it. Consequently, understanding a gesture might *not only* involve a direct matching, but also needs inference of intended function. For example, seeing a demonstrator waving should be recognized as waving, the symbolic meaning of which is independent of the absolute spatial position of the hands, the swinging frequency, the performing hand or to some degree the velocity of the movement. Although different persons have different styles of waving and different ways of modulating its meaning (e.g. cheerful vs. hesitant), those performances can still be recognized by an observer as instances of waving. And, when reciprocating, the observer likewise performs it in an individual manner. Thus, the motor representation must be able to cluster numerous variants of a gestural movement into one "schema". Therefore, we define *motor schema* (MS) as a generalized representation that groups different possible performances (motor programs) together. Such a generalization process is an important capability and can foster the understanding and imitation of behavior in two ways. First, it forwards the problem of inferring the intention behind a gesture from a specific performance to a more abstract, yet less complex level, namely schema interpretation. Second, an agent can retain its own personal form of performing a gesture while being able to relate other performances of the same gesture to the same schema.

Forward and Inverse Models

The motor system employs two internal models for prediction and motor control. These internal models have been hypothesized to exist in the human cerebellum [19]. A *forward model* implements causal and temporal predictions of a movement, providing likely next movement states, given the current state and possible efferent control signals. In other words, forward models are able to predict the continuation of movements during both perception and generation processes, using sensory and/or proprioceptive feedback. In contrast, an *inverse model* provides motor commands that are likely to achieve a desired movement state. That means learning a movement skill is largely equivalent to acquiring corresponding inverse and forward models. On this basis, our motor knowledge model is endowed with a pair of generic inverse and forward models, which can operate on the hierarchical motor representation at all levels. Many computational models assume a multitude of pairings of such internal models, containing the necessary motor knowledge for prediction and control of individual movements [20]. In contrast, our model is geared to the flexibility and generativity of gestural communication. Motor knowledge is thus integrated in the shared motor representation as an expandable graph and the forward and inverse submodules are seen as generic processors that perform the corresponding tasks on arbitrary elements of the graph representation.

Perception Process

The perception module receives visual stimuli about movements of relevant body parts of a demonstrator (positions of the wrists in this case). First, these are preprocessed such that they can be directly operated upon for recognizing familiar gestures or learning new ones. The preprocessing pipeline, illustrated in Fig. 2, consists of four submodules: (1) the observations are first transformed (rotated and scaled) by a *body correspondence solver* from external coordinate system to egocentric space of the virtual agent, which stays face-to-face to the human interlocutor, (2) the *sensory memory* is an ultra short-term memory that receives the transformed positions and buffers them in chronological order; (3) the *working memory* holds a continuous trajectory for each hand through agent-centric space and, (4) the *segmenter* submodule decomposes the received trajectory into movement segments called *guiding strokes* [16], based on their spatial and kinematic features (velocity drops, changes of movement direction). A guiding stroke represents a simple and short movement segment in 3D space and describes the movement path as well as kinematics along this segment in terms of a few parameters (see [21] for similar preprocessing steps). Since the focus of this paper is on intransitive (i.e., not object-directed) movements, all parameters attributed to the segments refer only to their morphological features and are not defined relative to an object. Such parametrized segments are the atomic movement components that form the motor commands in the shared motor knowledge representation.

While perceiving hand movements, the perception module employs the shared motor knowledge as follows: Candidates of known gestures that might correspond to the movement currently being observed are passed on to the forward models which derive predictions of the corresponding likely continuation of the movement. This principle is applied at all levels of motor knowledge in parallel. At the lowest level, predictions are evaluated against the positions of the wrist received at each time step from the working memory; at higher levels (programs, schemas), predictions are derived and evaluated against the corresponding structures at lower levels. The results of these evaluations are fed back into the shared motor knowledge as "motor resonances" in the graph. Those resonances and the evaluation processes that spawn them are modeled probabilistically and described in "A Probabilistic Model of Motor Resonances" in detail. Here we note that strong motor resonance of a motor component (motor command, program or schema) indicates a successful prediction of the observed movement by the corresponding forward model at that level. Strong motor resonance corresponds to high confidence of the virtual agent in recognizing the corresponding movement, grounded in its own motor experience. In contrast, if none of the motor components at one level resonated sufficiently, i.e., no sufficiently similar motor representation exists at that level, an unfamiliar movement segment or gesture performance is likely being observed. In this case, the analyses switch from the forward model to the inverse model at the corresponding level.

As aforementioned, inverse models capture motor specifications to achieve a desired state. For this purpose, inverse models receive movement segments from the segmenter and augment the motor representation graph with new components, which represent the new movement at the proper level of abstraction (for more details about inverse models at the motor command level see [22]). In this way, the virtual agent learns new gestures by extending its motor repertoire. Inverse models could insert new segments as motor commands, create a new motor program or associate an old one with a new sequence of motor commands, and create a new motor schema or associate a new motor program with a known schema. However, a new association between a motor program (performance of a gesture) and a schema (core meaning of a gesture) cannot be determined only with the help of the spatiotemporal features available in working memory. Concerning those features, totally different gestures can signal the same social intention, and therefore, they have to be clustered together as one schema. However, as intention recognition is beyond the scope of the present study, we utilize explicit feedback to cluster different performances of a gesture schema: the human interlocutor labels each new gesture with a name, e.g. "waving", which is then used by the inverse model at motor schema level to associate the current performance of waving with the respective schema.

Generation Process

Generation of a gesture is a top-down process that is invoked by the prior decision to express an intention through a gesture. In our framework, the performance of a hand-arm movement by the virtual agent is built on a motor control engine described in previous work [16]. Here, we focus on the processes in shared motor knowledge which result in movement commands to be performed by this engine. The first step is to select a proper motor schema to be generated. This decision has to be made by higher cognitive levels, concerning referential, communicative and social intentions, which are beyond the scope of this paper. Therefore, the proper motor schema is currently directly given in each social situation to the virtual agent. At the next step, the agent has to select a motor program, i.e., a possible performance of the selected gesture schema. This choice is modeled to depend on two criteria: observation frequency, and previously perceived or self-generated gestures. The latter refers to the mutual effects of perception and action handled in "Perception-Action Integration". The former refers to the agent's tendency to perform a gesture in the way it has observed (and recognized) it most often. The corresponding motor program is referred to as the *prototype* of that motor schema. Hence, the prototype gesture of each schema emerges from the way the majority of human interlocutors with whom the agent has interacted have performed that gesture. After choosing a motor program to perform, the next generation step is simply to follow the unambiguous association between the selected motor program and the motor commands for both wrists.

A Probabilistic Model of Motor Resonances

As described earlier, motor resonances result from comparing predictions with observations. This basic mechanism is employed at all three levels (using different kinds of forward models to derive the predictions) and the resulting resonances indicate the agent's confidence in a correspondence between what it sees someone doing and what it knows from own experience.

Motor resonances are modeled probabilistically and are computed for each motor candidate at each level during observation. The general approach is to apply Bayesian inference and is the same for all levels. Given the evidence $\mathbf{e}$ (e.g. visual stimuli of moving hands), we define this confidence in recognizing a certain motor candidate (referred to as h for hypothesis) as the mean over time of its conditional probabilities until the current time T: $P_T(h|\mathbf{e}) := \frac{1}{T}\sum_{t=t_1}^{T} P(h|\mathbf{e}_t)$. At each time step, we employ Bayes' law and compute the probability $P(h|\mathbf{e}_t)$ as a normalized product of the likelihood and the prior probability of the same motor candidate: $aposteriori = \alpha \cdot likelihood \cdot apriori$. To compute the *apriori* term, we apply the prior feedback approach [23] to accumulate probabilities up to each time step. That means *apriori* at each time step is the *aposteriori* from the previous time step. This affords incremental processing. In other words, the more positive evidence we have, the higher the recognition confidence will be. The *likelihood* term (i.e., $P(\mathbf{e}|h)$) is modeled specially for the motor candidates at each level (see "Bottom-Up Perception").

Furthermore, the probabilities of motor candidates at the three levels influence each other mutually. A Bayesian network models how the levels of the hierarchy of motor representation interact (see Fig. 3). This approach allows motor resonances to percolate bottom-up and top-down across adjacent levels, to find (possibly a variant of) a known gesture quickly, effectively and robustly. In the following section, we first focus on the bottom-up perception process, from motor commands, to motor programs, to motor schemas. After that we consider the top-down guidance of the perception process.

Bottom-Up Perception

Level 1: Resonating Motor Commands

At this level, the spatial position of a wrist at each time step t is our evidence and motor commands in the corresponding graph are the hypotheses. That means, the agent updates its beliefs in observing each motor command as a candidate. In order to have a fast and cost-efficient algorithm, we need not consider all motor commands, but only the subset of the most probable ones, referred to as the set of active motor command hypotheses H_c. The criterion to add a motor command to this set is as follows: As soon as the first evidence, here the observation $\mathbf{o}_{t_1}$, is perceived, its

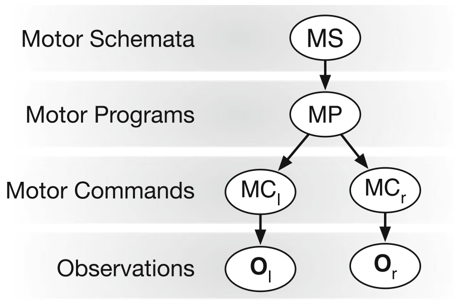


Fig. 3 Bayesian network for the relations between different levels of the motor hierarchy

probability to represent a node in the graph of motor commands is computed with the aid of Gaussian densities centered at the position of each node in three-dimensional space. Comparing it with a predefined threshold yields the most likely candidate nodes for the starting point of a gesture. All outgoing motor commands from these nodes are added to H_c. At the next time steps, the probability of each of these hypotheses is computed from the new observation (Eq. 1). If the probability of a hypothesis is smaller than a predefined threshold, it will be omitted from H_c, and if a new node becomes likely to be the start of a new hypothesis, the corresponding hypothesis will be inserted to H_c. Furthermore, the active hypotheses can change dynamically and be split into new hypotheses by a branching node, whereas one of those hypotheses indicates the belief in stopping the movement at the end of the passed motor command, and the other hypotheses refer to the beliefs in possible continuation of the movement observation.

Employing Bayes' law, the probability of a motor command hypothesis c (this equals the resonance) is updated at each time step on the basis of perceived evidence up to the current time step, T: ($\mathbf{o} = \{\mathbf{o}_{t_1}, \mathbf{o}_{t_2}, \ldots, \mathbf{o}_T\}$).

$$P_T(c \in H_c) = P_T(c|\mathbf{o}) := \frac{1}{T}\sum_{t=t_1}^{T} P(c|\mathbf{o}_t)$$
$$= \frac{1}{T}\sum_{t=t_1}^{T} \alpha P_{T-1}(c)P(\mathbf{o}_t|c) \tag{1}$$

The Bayesian normalizing constant is referred to as α. The term $P_{T\text{-}1}(c)$ is the prior probability of the hypothesis c and indicates the previous knowledge about the motor command c, which is equal to the posterior probability of c at the previous time step, $T - 1$. The likelihood term $P(\mathbf{o}_t|c)$ refers to the probability of passing the coordinate $\mathbf{o}_t = \{x_t, y_t, z_t\}$ with motor command c and, now, represents a probabilistic prediction of the forward model. In other words, it represents the probability of where the hand would be if the agent performed this motor command c. We model this as a four-dimensional Gaussian probability density function of $\{x, y, z, t\}$ (PDF, in short), formed for each possible motor command, i.e., each possible continuing movement segment of the wrist in space (see Fig. 4a). This likelihood term reaches its maximum value if the observed performance exactly matches the agent's own motor execution.

Level 2: Resonating Motor Programs

Resonance of a motor program depends on the probabilities of its components (i.e., motor commands of each wrist) and thus, indirectly, on the wrist observations. Similar to motor commands, we compute the posterior probability of the hypothesis p considering evidence from both hands ($\mathbf{o}_l$ and $\mathbf{o}_r$) until the current time step, T. The hypothesis with the highest posterior probability indicates the agent's belief in observing that specific program being executed.

$$P_T(p \in H_p) = P_T(p|H_c, \mathbf{o}_l, \mathbf{o}_r) := \frac{1}{T}\sum_{t=t_1}^{T} P(p|H_c, \mathbf{o}_l, \mathbf{o}_r)$$
$$= \frac{1}{T}\sum_{t=t_1}^{T} \alpha P_{T-1}(p) \prod_{i\in\{r,l\}} \sum_{c\in H_{c,i}} P(\mathbf{o}_{i,t}|c)P_t(c|p) \tag{2}$$

The set of motor programs considered as active hypotheses H_p contains all programs with at least one active motor command hypothesis in H_c. At each time step, the computed probability for each active motor program hypothesis corresponds to the confidence of the agent in recognizing that motor program for which, in contrast to

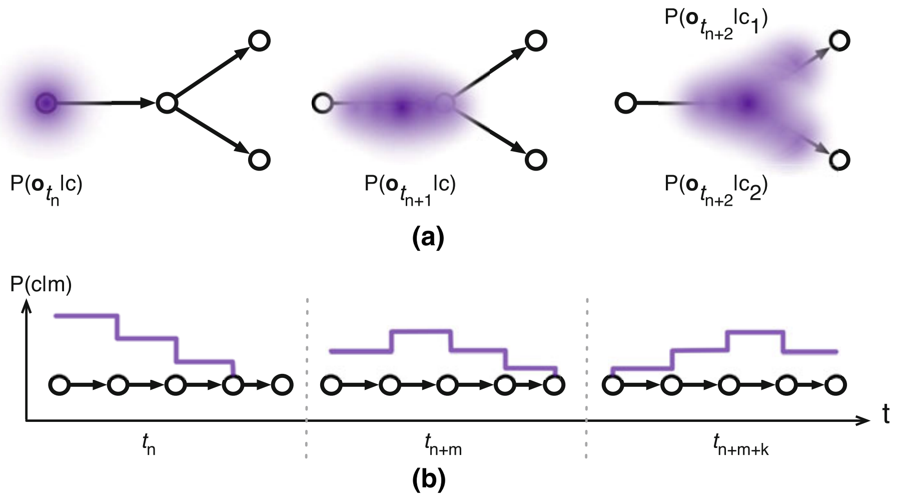

$P(\mathbf{o}_{t_{n+2}}|c_1)$
$P(\mathbf{o}_{t_n}|c)$
$P(\mathbf{o}_{t_{n+1}}|c)$
$P(\mathbf{o}_{t_{n+2}}|c_2)$
(a)
P(c|m)
t
t_n
t_{n+m}
t_{n+m+k}
(b)

Fig. 4 **a** Likelihood of motor command hypotheses at these successive time steps, modeled as four-dimensional Gaussian density functions that change in accord with the motor command. **b** Likelihood of a motor program hypothesis, modeled as one-dimensional discrete Gaussian density function, stretched over associated motor commands. The density function moves over time along the sequentially connected motor commands the commands, the morphological properties of the whole gesture performance are considered. Motor programs with too small probabilities will be removed from H_p.

Same as motor commands (Eq. 1), the prior probability is equal to the posterior probability from the previous time step and α indicates the Bayesian normalization constant.

Most of the gestures which are performed with only one hand should be recognized regardless of the hand they are performed with. Since a motor program comprises a sequence of motor commands for both hands, it can specify how their probabilities affect the probability of the associated motor program. In Eq. 2 both wrists are assumed to have their own task during performance (therefore, an AND relation: $\prod_{i\in\{r,l\}}$). For example, in referring to a round shape with a symmetrical gesture, each hand draws half of a circle. Alternatively, the hands may also be combined using an OR relation ($\sum_{i\,\in\,\{r,l\}}$). For instance, a waving gesture as a motor program can be performed by the right "or" left hand. The choice whether AND or OR is used depends on the particular gesture and is determined by the corresponding schema.

The likelihood term $P_t(c|p)$ indicates the probability of performing the command c at time t, if the demonstrator were to perform the program p. This probability is time-dependent and is modeled using a PDF as a function of t and the motor commands c. The mean of the Gaussian moves through the commands of a motor program, according to the velocity of each motor command (Fig. 4b). Thus, this term along with $P(\mathbf{o}_t|c)$ yields the highest likelihood when observing a gesture performance with exactly the same movement trajectory and velocity as the one represented in the own motor repertoire.

Note, however, that these probabilities are incrementally computed and adjusted from the evidence at hand, also during perception where only parts of the gesture have been observed. That is, the agent does not need to specify the start and end point of gestures, but can recognize gestures that were started at a later point of a trajectory, e.g., in the case of several successive gestures without intermediate rest position, or when observing a gesture partially.

As mentioned in "Shared Motor Knowledge", in the case of observing an unfamiliar gesture, which cannot be predicted confidently, the performance will be learned through the inverse models. However, the decision between observing a new gesture performance or a familiar one cannot be made only on the basis of the posterior probabilities of all hypotheses, i.e., all familiar gestures, whose sum at each time step is 1. Observing an unfamiliar gesture will assign the highest probability to the most similar motor program, despite of high deviations in performance. Therefore, the confidence in recognizing a familiar gesture depends on the likelihood average over the whole performance which should be above a predefined rejection threshold. If this is not the case, the model switches to acquire the new performance

Level 3: Resonating Motor Schemas

Motor schemas group different motor program hypotheses into a single one. The probability of each motor schema hypothesis thus depends on the probabilities of the associated motor program hypotheses, and indirectly on the related motor commands and evidence about each wrist. Figure 3 illustrates these causal influences between the graph nodes in a Bayesian network.

In detail, the probability of each schema hypothesis s is computed as:

$$P_T(s\in H_s)=P_T(s|H_p,H_c,\mathbf{o}_l,\mathbf{o}_r):=\frac{1}{T}\sum_{t=t_1}^{T}P(s|H_p,H_c,\mathbf{o}_l,\mathbf{o}_r)$$
$$=\frac{1}{T}\sum_{t=t_1}^{T}\alpha P_{T-1}(s)\sum_{p\in H_p}P(p|s)$$
$$\prod_{i\in\{r,l\}}\sum_{c\in H_{c,i}}P(\mathbf{o}_{i,t}|c)P_t(c|p) \quad (3)$$

The likelihood $P(p|s)$ is taken to be uniformly distributed among the associated motor programs p with the active motor schema hypothesis s, and 0 otherwise. Since a schema can be performed by any of the associated performances, there is an OR relation among the connected motor programs ($\sum_{p\in H_p}$) and the probability of a schema s is the sum of the probabilities of its possible performances.

Top-Down Guidance

The described probabilistic model simulates the bottom-up emergence of resonances in the hierarchical representation, where probabilistic motor resonances at each level induce resonances at higher levels. The other way around, higher levels can also affect the perception process at lower levels by way of *priming*. For instance, having recognized a motor schema unequivocally should yield expectations to perceive the remaining part of the motor program over the next time steps. Similarly, recognizing a motor program should increase the expectation of the associated subsequent motor commands. This top-down information processing and attention guidance may also directly be extended into higher social cognitive levels. For instance, expecting the closing of a dialog could prime an agent to observe a waving schema.

To achieve this capability, we extend our model by basically running two update processes for each time step, one bottom-up and one top-down. The first one acts as described above and calculates posterior probabilities of all motor structures given bottom-up evidence using a Bayesian update. The second one also performs a Bayesian update but now on the probability of each motor unit conditioned on the higher-level hypotheses. For example, consider the probability of each hypothesis at the motor command level: First, the posterior probability $P(c|\mathbf{o})$ is determined bottom-up as given in Eq. 1. Then, this posterior probability is used as prior probability $P(c)$ for a second, top-down update which determines the posterior probability $P(c|p) = \alpha\ P(p|c)P(c)$, where the likelihood term $P(p|c)$ is the current probability of the motor program p. The resulting probability $P(c|p)$ represents the posterior probability for the motor command c and is, in turn, taken as prior probability in the first update process at the next time step. That way, bottom-up and top-down processing are connected and contribute both to the emergence of motor resonances. Likewise, a resonating motor schema affects the expectation of its comprised motor programs by applying the top-down update Bayes' law $P(p|s) = \alpha$ $P(s|p)P(p)$. Overall, we do not only percolate probabilities of active hypotheses upward, but also adjust the prior probabilities of current or future hypotheses top-down in a context-dependent way. Section "Results" presents results obtained with this approach applied to real gesture data.

Perception-Action Integration

As discussed in "Introduction", humans employ their motor expertise for both perception and generation [24]. Similarly, in our model the basic idea is to allow an embodied virtual agent to create and augment its motor knowledge by observing others' gestural movements, and then to use that knowledge for both perceiving others and generating movements. This sharing of motor knowledge directly enables an interaction of both processes in ways that are observable in humans: on the one hand, behavioral tendencies of humans are influenced by their perception and resulting motor resonances [25, 26]; on the other hand, self-generation of behavior guides attention and increase sensitivity for subsequent perception of similar movements (called *perceptual resonance* [27]). Both resonance phenomena are assumed to play important roles for the contingent processing and coordination of social signals.

In the previous sections, we have presented a hierarchical model of motor knowledge and we have shown how bottom-up and top-down processes probabilistically operate upon these structures, for perception and generation of social behavior. To model how perception and action influence each other, we define a notion of *motor neural activation*: neural activation of the motor system is evoked during both generation and perception processes, it is assumed to persist in shared knowledge structure, and to cease slowly over time such that subsequent processing is affected. Each motor component (motor command, program or schema) is assigned a value between 0 and 1 indicating its *relative* activation. At each time step, this value is either updated by a generation or perception process or, if not, will decay following a sigmoidal decrease function toward 0. For a motor component m we have:

$$activation(m,t) = \begin{cases} 1, & m \text{ is being performed at } t \\ P_t(m), & m \text{ is being observed with probability } P \text{ at } t \\ decrease(m), & \text{otherwise} \end{cases} \tag{4}$$

The time needed for an activation to cease is set to depend on the abstraction level: motor command activations live shortest, while motor schema activations last longest. Note that these neural activations arise from perception and generation processes and influence them in turn. Basically, the activations serve as a "bridge" between perception and generation processes.

The perception process results in probabilistic motor resonances at each level, updated at each time step, with the prior probability of motor candidates set to its previous posterior probability (see "A Probabilistic Model of Motor Resonances"). At the first time step t_1 the prior probabilities of the active hypotheses at each level are set to their current activation values normalized by the sum of the activations of all other active hypotheses at the same level. In this way, a highly active motor candidate attains a relatively higher prior probability which corresponds to stronger priming. At the level of motor commands, the hypotheses can be split into several *child* hypotheses at each node. The prior probability of each child hypothesis is set to a fraction of its *parent* posterior probability that is proportional to its relative activation with respect to other child hypotheses.

During movement generation, the neural activations are currently only taken into account at the level of motor programs, since the decision to choose a motor schema is made by higher levels and motor commands have unambiguous relations to the associated motor programs. Selecting a motor program for a given schema, i.e., selecting a specific performance of a gesture, is done probabilistically according to two criteria: observation frequency (as explained in "Generation Process") and the given activations of the candidate motor programs. We again employ Bayes' law and select the motor program p with the highest posterior probability $P(p|s)$. The likelihood term $P(s|p)$ is thereby set to the observation frequency value of p, relative to other performances of the schema s. The prior term $P(p)$, similar to the perception case, is

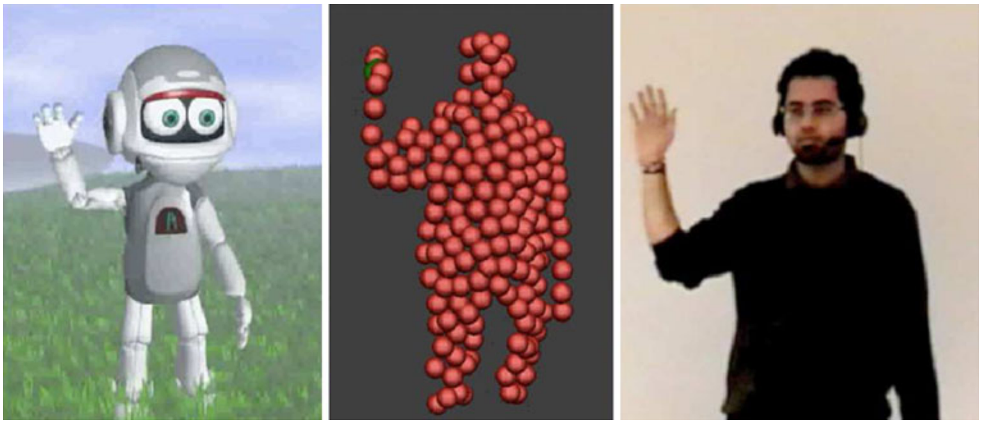


Fig. 5 Setup: the virtual agent Vince (*left*) and the perceived body posture augmented with wrist positions (*middle*), while a human user performs a waving gesture (*right*)

assigned to the current neural motor activation of p, normalized by the sum of activations among all other candidate performances. In this way, we consider a combination of two criteria: (1) how strong the agent is accustomed to its own individual prototype performance of a schema, and (2) how high the current neural activations of the corresponding motor components are.

As a result, this use of neural motor activations realizes perception-action integration by way of probabilistically biased decision-making. The next section shows simulation results of this.

Results

The presented model of embodied social signal processing has been implemented and evaluated against real-world hand-arm gesture data. In our setup with a 3D time-of-flight camera (a SwissRanger$^{\mathrm{TM}}$ SR4000[1]) and the marker-free tracking software iisu,[2] our humanoid virtual embodied agent *Vince* observes the wrist trajectories of a human demonstrator freely performing gestures (see Fig. 5). We demonstrate the capabilities of the presented model in an interaction scenario between Vince and the human interlocutor.

The interaction scenario consists of a game between the human and the agent. To this end, Vince has been equipped with a dialog manager component that manages the interaction and controls corresponding verbal behavior. The overall course of the interaction is as follows:

(1) Vince greets the human interlocutor and explains the game

(2) *Human's turn*: The human interlocutor performs a gesture while Vince observes

(3) If Vince recognizes the gesture as familiar:

(3.1) *Recognition*: Vince says the name of the recognized gesture schema

(3.2) Vince imitates the gesture by performing the gesture schema prototype and asks for confirmation

(3.3) If the interlocutor rejects the imitation: go to (4.2) *learning*

(3.4) Otherwise, if the interlocutor confirms the guess: go to (5) *Vince's turn*

(4) Otherwise, if Vince detects an unfamiliar gesture:

(4.1) Vince states the performed gesture was unknown to him

(4.2) *Learning*: Vince asks for the label of the gesture

(4.3) Human interlocutor gives a label for the gesture

(4.4) Vince acquires the observed gesture and labels the schema accordingly

(4.5) Vince re-produces the newly learned gesture

(5) *Vince's turn*: Vince randomly performs a gesture from his own repertoire and asks for its name

(6) The interlocutor guesses the label of the performed gesture

(7) If the label matches, Vince confirms the guess; if not, Vince corrects the interlocutor

(8) If the interlocutor does not end the game: go to (2) *Human's turn*

(9) Vince says goodbye.

This scenario imposes a number of challenges of processing and using social signals, which are prevalent also in natural human face-to-face interaction. The following sections report how our model accomplishes them: "Detecting and Learning New Gestures" discusses how the

[1] http://www.mesa-imaging.ch.

[2] http://www.softkinetic.net.

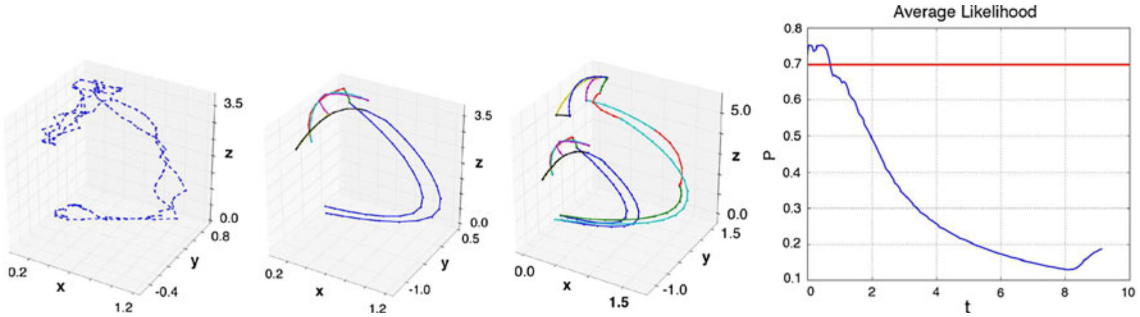


Fig. 6 *From left to right* the observed trajectory of the *wave*1 gesture; the created guiding strokes for *wave*1; guiding strokes after learning a second waving performance, *wave*2; the likelihood average of *wave1* while observing gesture *wave*2 (the *horizontal line* indicates the recognition threshold of 0.7 for the average likelihood)

virtual agent detects and acquires new gestures from observation. Section "Recognizing Familiar Gestures" describes how Vince recognizes familiar gestures in a fast, incremental, and robust manner. Furthermore, top-down and bottom-up resonances in this process are illustrated and discussed. Finally, "Perception-Action Integration" shows how the perception-action integration improves the interaction through behavior coordination and attention guidance on the part of the agent.

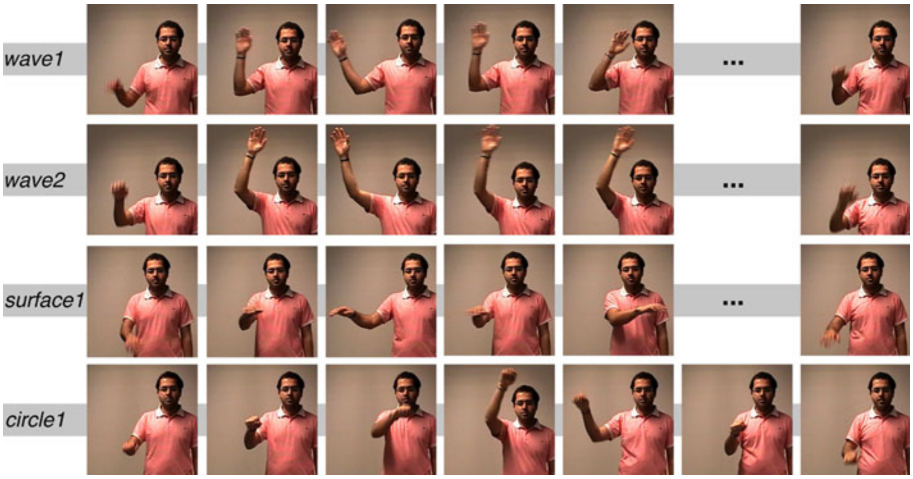
Detecting and Learning New Gestures

At the beginning of the interaction, Vince has an empty motor knowledge and is ready to observe new gestures. The following analysis refers to steps (2–4) of our scenario. As first gesture, we presented Vince a waving gesture, performed with the right hand at about the height of the head (*wave1* in Fig. 7; see Fig. 6 (left) for the trajectory). Since the motor knowledge does not possess any candidate at either level, the model switches immediately to inverse models. Preprocessing and motor command inverse models yield movement segments and guiding strokes (shown in Fig. 6) which are added as motor commands at the lowest level of motor knowledge. At the next level, this motor command sequence is acquired as a new motor program, referred to as *wave*1. Afterward, Vince asks for a label and recognizes the word "waving" which is then assigned to a newly created motor schema *waving*, associated with the motor program *wave1*. The resulting structures is shown in Fig. 8a.

In the next round, we presented Vince another waving gesture *wave*2, performed with a more outstretched right arm (see Fig. 7). Since there is only one motor hypothesis at each level, all hypotheses attain a probability of 1. However, the likelihood average, shown in Fig. 6 (*right*), is too low to push Vince's confidence in observing the gesture *wave1* above the threshold (see "Bottom-Up Perception" for further details about this issue). Vince hence detects another new gesture (step 4) in the interaction) and creates new motor commands (shown in Fig. 6) as well as a new motor program *wave2*. However, since the user labeled the new performance "waving", the motor program is associated with the same motor schema *waving* (Fig. 8a). Figure 11 shows the learned gesture performed by Vince.

In two further rounds the interlocutor performed two other gestures (see Fig. 7): shaping a flat surface by

Fig. 7 Snapshots from a human interlocutor performing four different gestures, labeled as the corresponding motor programs: *wave*1, *wave*2, *surface*1 and *circle*1



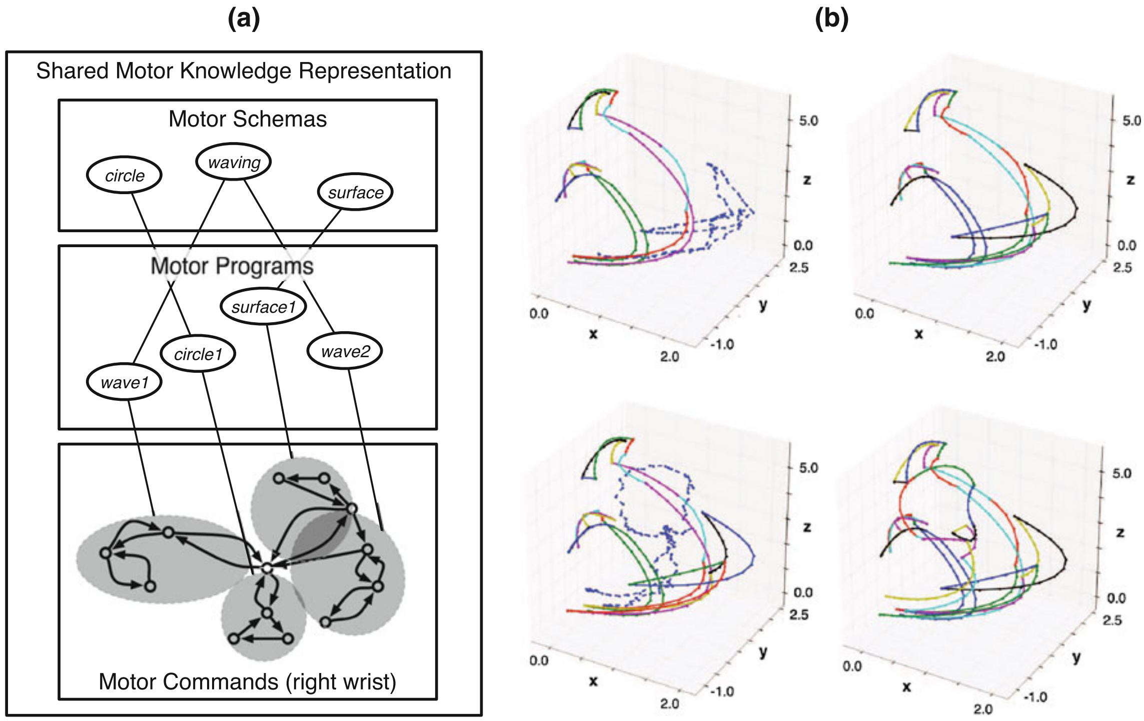


Fig. 8 **a** The acquired hierarchical motor knowledge after observing four gestures within the scenario; **b** observed trajectories (*dashed*) overlaid on the space of guiding strokes (*solid*). *Top-left*: trajectory for the *surface*1 gesture and waving guiding strokes; *top-right* learned *surface*1 motor commands; *bottom-left*: trajectory of gesture *circle*1; *bottom-right*: guiding stroke space containing learned motor commands for all four gestures

moving the right hand horizontally in front of himself, and a circle gesture. In the former case, at the end of the performance, the motor program hypothesis of *wave*2 attains stronger motor resonance because of a greater similarity to the starting movement of the demonstrated gesture. However, the average likelihood value for that hypothesis is 29% of the maximum likelihood which is clearly lower than the rejection threshold, empirically set to 70%. Therefore, the surface gesture is determined as unfamiliar; it is learned and inserted as a new motor program (*surface*1) into the motor knowledge and a new schema (*surface*) is created. Likewise, in the case of the circle gesture, Vince's motor knowledge is augmented with further motor commands, a motor program and a motor schema. Figure 8a shows the shared motor knowledge of Vince after these four rounds of interaction.

Recognizing Familiar Gestures

After the previous interaction, we meet Vince as an embodied agent with some motor expertise on his own. Thus, we can investigate the case of recognizing familiar gestures based on resonances in the own repertoire (steps (2–3) in the our scenario). We analyze this for two cases: First, we turn off top-down motor guidance and focus on the bottom-up perception process. Afterward, we compare this with combined both bottom-up and top-down processing applied to exactly the same recognition scenario.

Simulation of Bottom-Up Perception

We presented Vince a waving gesture similar to the first performance *wave*1. Figure 9 (top) shows how motor resonances in Vince's motor system (viz. confidences in hypotheses) evolve during the course of perceiving this gesture. The forward model at motor program level creates one hypothesis for each of the four known gesture performances. Overall, there are three known gesture schemas: waving, circle and surface. Since the demonstrated gesture starts similarly to the known circle gesture, Vince at first "thinks" that the interlocutor is going to draw a circle. However, after 3.5 s, the resonance of motor program *wave*1 is stronger than that of *circle*1, and from about second 4.0 on Vince recognizes the *wave*1 performance reliably. Note also that right from the beginning of the performance Vince is quite sure he is observing a waving schema. The reason is that he has experienced twice as many waving gestures as other gestures in his short life. This effect is wanted and emphasizes a developmental perspective of our approach to learning social signal processing. This "assimilation bias" will wash out as Vince sees more and more different performances (in fact, when

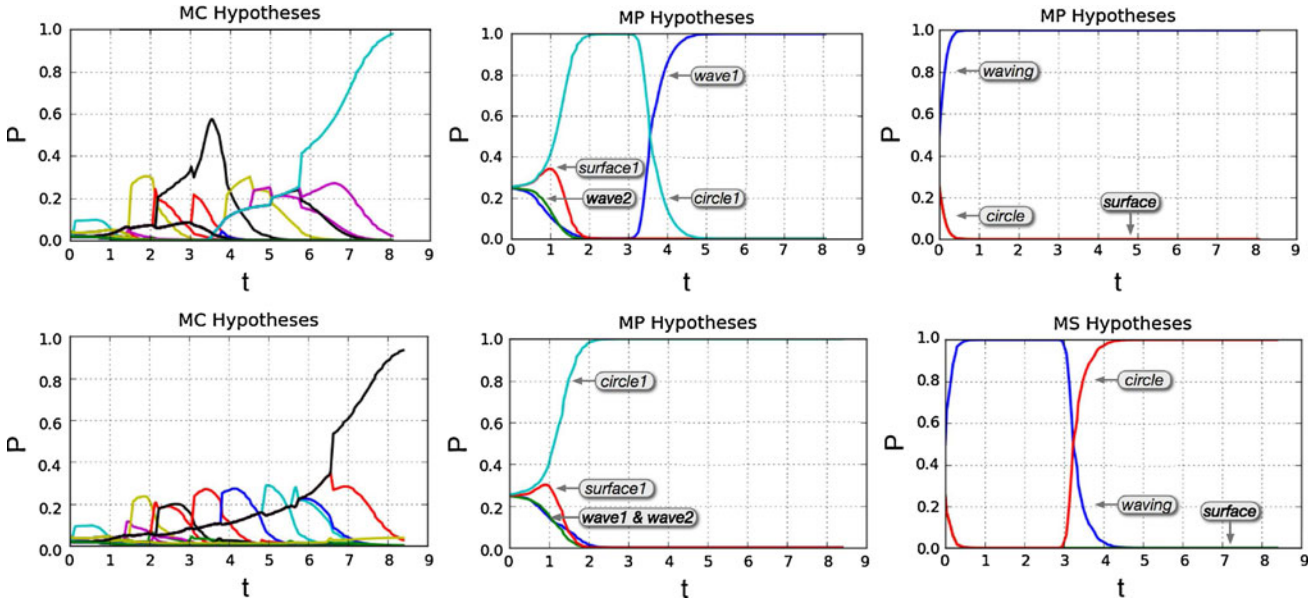


Fig. 9 Bottom-up motor resonances at all three levels (*MC* motor commands, *MP* motor programs, *MS* motor schemas) while observing *wave*1 gesture (*top*) and while observing gesture *circle*1 (*bottom*)

presenting this gesture to Vince after he has learned not two but only one waving gesture, the motor schemas confidences evolve identically to those at the motor program level). When we now present Vince another performance of *circle*1 (Fig. 9, bottom), the agent again recognizes this motor program already 1.5 s after the onset of the demonstration. The bias toward waving gestures, as evidenced by the early resonances of the *wave* schema, diminishes after about 3 s and the *circle* schema prevails.

Simulation of Integrated Bottom-Up and Top-Down Motor Guidances

Now, consider the resulting motor resonances when the agent, employing the same motor knowledge and being demonstrated exactly the same waving gesture, uses both bottom-up and top-down processing. Figure 10 compares the emerging motor resonances, with and without top-down guidance. Since hypotheses are now confirmed or rejected by both higher and lower levels, the motor resonances are more stable and respond faster, such that Vince determines likely hypotheses earlier. In this case, the motor program *wave*1 is also activated by the more probable *waving* schema and is recognized about 2.5 s earlier than in the sole bottom-up processing case. Likewise, associated motor commands are also more expected to be observed.

Note that both bottom-up and top-down processes update the probabilistic motor resonances "at each time step". That means, the gradient of belief variation depends on the frequency with which these update methods are applied: the more frequently the hypotheses are updated, the steeper their belief variations are. In the results presented here, we have applied both processes with frequencies of 10 Hz which have to be set empirically as parameters according to two criteria: (1) how fast the beliefs should be updated given the frame rate of the

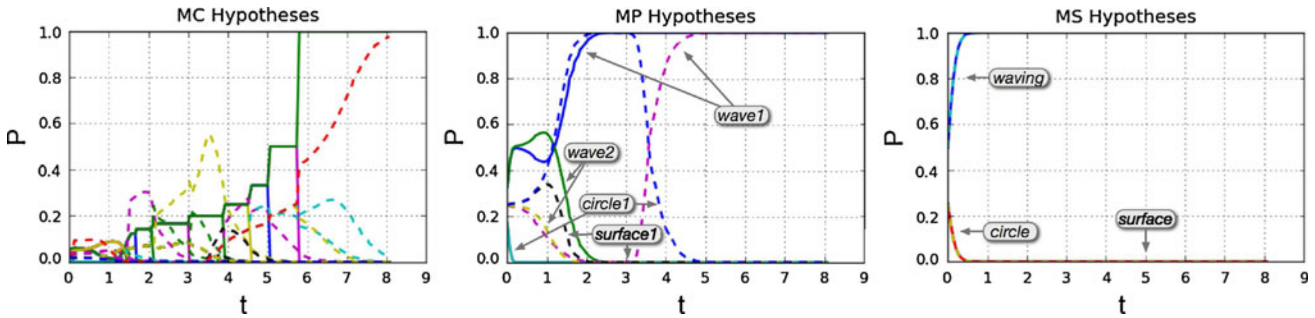


Fig. 10 Comparison of motor resonances at all three levels while observing *wave*1 with only bottom-up percolation (*dashed lines*) or top-down guidance integrated (*solid lines*)

tracking system and, (2) how strong the relative effects of top-down and bottom-up processes should be. This means, for example, performing bottom-up process more frequently than top-down process simulates a virtual agent who relies primarily on his sensory input when recognizing a movement, rather than his higher-level beliefs and speculations.

Perception-Action Integration

In our interaction scenario, there are situations which demonstrate, and benefit from, the interaction between perception and generation of gestural behavior. Next, we analyze an example of how perception affects the following generation process. Then we show how generating a gesture primes the agent's attention in the following perception process.

Generation after Observation

After conducting the same training session as above in which Vince has learned three schemas (see Fig. 8), we presented the agent three further performances of *wave*1 during further rounds of the game (i.e. step 3). Therefore, the corresponding motor program *wave*1 has been encountered four times overall and the alternative waving performance *wave*2 has been seen once. If Vince now decided to wave (i.e., it selects the schema *waving* at scenario step 5), he would perform *wave*1 as the prototypical performance. Now, in the next round of the game at step (2), we performed a gesture similar to *wave*2. This performance is recognized by Vince as a familiar gesture and increases the observation frequency of *wave*2, which however is still half of the frequency of *wave*1. That is, the generation likelihood $P(s|p)$ is 0.5 (cf. "Perception-Action Integration"). Now, it is Vince's turn (step 5) to generate a gesture and he decides to wave. The previous perception of the *wave*2 gesture has evoked an according motor activation, which remains active for a while (here, set to 2 s for motor commands, 4 s for motor programs and 6 s for motor schemas). Since Vince intends to wave before the motor activation of *wave*2 decreases to a value lower than its likelihood (equal to 0.5, in this case), the agent chooses the same waving performance as the one it has recently observed, i.e. *wave*2, and not its individually preferred way of waving, namely *wave*1 (see Fig. 11).

Observation After Generation

The effect of the generation process on perception is seen at step (2) when the human user performs a gesture after Vince has generated one himself (5–8). We simulate two recognition cases: First, when Vince did perform *wave1* previously and, second, when Vince did not generate any gesture previously (recognition case described in "Recognizing Familiar Gestures"). Figure 12 shows the resulting motor resonances at the motor program level during observation of *wave*1 in both cases. When Vince has

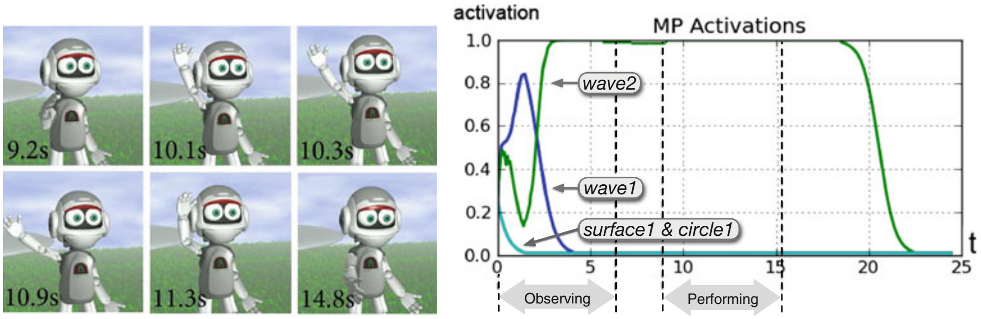


Fig. 11 *Left* Vince performing the *wave*2 gesture, after observing the same gesture performed by the human user; *right* motor activations while observing and performing waving gestures; high activations of the *wave*2 motor program prime the performance of the same gesture, albeit *wave*1 is generally preferred

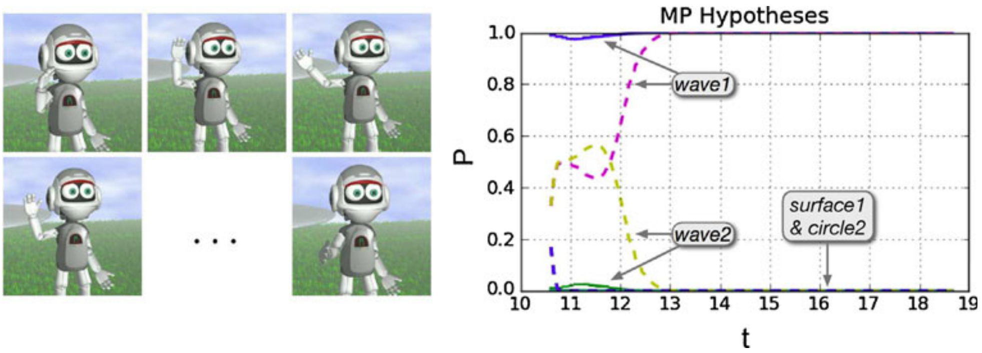


Fig. 12 *Left* Vince while performing *wave*1 gesture; *right* motor resonances at motor program level during observation of *wave*1 with (*solid lines*) or without (*dashed lines*) previous self-performance of the same gesture

previously performed *wave*1 himself, the corresponding motor components are activated and his sensitivity is primed toward the same behavior. In result, Vince immediately recognizes the demonstrated *wave*1 gesture right after the interlocutor has started to move, and the following observations confirm this expectation. He thus responds about 2 s earlier to the human's demonstration (step 3).

Related Work

Coupling of perception and generation for transitive and intransitive actions is a hot topic in computational modeling, from both the engineering view of robotics [28–31] and from the (social) cognitive science perspective [21, 32–37]. In almost all of these studies, the main focus is on imitation as a learning mechanism which links perception to action in an artificial agent. The applied methods for this aim to fulfill a continuum of requirements, from (neuro) biologically inspired ones, to more technical and task-oriented approaches.

Hidden Markov Models (HMMs), which are commonly used for automatic speech recognition [38], are the most popular modeling tool used for analyzing movements in technical robotics [28, 30, 31, 39]. Although HMMs have become established in movement recognition and even generation [30], they are bounded to some methodological restrictions, which arise especially in social interaction between humans and artificial agents. In order to apply HMMs as movement classifiers, the number of HMM states needs to be found empirically [28, 31] or by applying additional criteria to "available" data [40, 41]. Hence, such a model cannot guarantee its flexibility in classifying new unpredicted movements. Furthermore, HMMs as movement recognizers need to be fed with the whole movement sequence to compute the corresponding generation probability (usually by applying the *Viterbi algorithm* [38]). This is due to the necessary preprocessing step which maps observed data to clusters, as inputs of the corresponding HMM states (e.g., Calinon and Billard [41] have applied principle component analysis (PCA); Aleotti and Caselli [28] have used distance-based geometric clustering). Therefore, although those HMMs are also applied as online and incremental learning methods [30, 42], and they simulate human recognition ability better than many classical off-line and batch-learning approaches employed by robotic studies [36, 41], they are still not fast enough for human social interaction. In many cases, humans need to recognize a communicative movement already during its performance to (re)act fast and in a social manner. Our probabilistic method updates the agent's belief in observing familiar movements at each time step during observation. Furthermore, in recognition mode there is no restriction on the length of the given observation sequence. Hence, the model is robust against duration variability of movements to any extent.

Besides other probabilistic methods (such as Dynamic Bayesian Networks [29], or Gaussian Mixture Models [42]), connectionist algorithms are neurobiologically inspired approaches, which are mostly applied to model mirror neurons as links between visual perception of movement and motor commands [21, 43, 44]. The disadvantage of such connectionist algorithms, on the one hand, is the fact that they need a high number of training data to converge and be applied as classifiers (which make them suitable for developmental modeling). On the other hand, their parameters and subprocesses are not analyzable with respect to the given problem. In contrast, applying symbolic probabilistic algorithms allows further interpretation of single terms and components of the method concerning modeled cognitive processes. For example, in the present model, the terms likelihood, prior and posterior probabilities also indicate their role in the corresponding cognitive processes, similar to their mathematical denotations.

Modular architectures of forward and inverse models were initially proposed for motor control as the MOSAIC architecture [19, 45]. In the following work [32, 33] this architecture has shown its capability in action recognition by applying forward models as predictors, which can be employed to assess their corresponding movement hypotheses against observation. The MOSAIC architecture represents movements through related predictors and controllers in a modular system. However, in our model, we have applied these internal models as generic processors performing their tasks on a central representation of motor knowledge. In this context, the forward and inverse models are functionally similar to the concept of *simulators* as in the *perceptual symbol system theory* [46]. These simulators perform forward and inverse simulations by applying the shared hierarchical motor knowledge. Haruno et al. [47] have extended MOSAIC to a hierarchical architecture (HMOSAIC), which can perform bidirectional information processing (top-down and bottom-up) between different levels of motor knowledge. The hierarchical representation is similar to our shared motor knowledge and consists of three different levels of abstraction: kinematics movements, sequences of actions, and goals and intentions of actions. Hierarchical representation of actions has been applied by many other studies, for instance, in order to solve correspondence problem of transitive actions [48], or to use motor primitives as building blocks for more complex actions [30, 36]. Alternatively, in recent work [49] Krüger et al. have applied parametric HMMs (PHMMs) for recognizing and synthesizing transitive movements. In this way, different HMMs representing different movements that vary due to different parametrizations can be collapsed into a single HMM. However, except for the HMOSAIC architecture in [47], none of the aforementioned approaches so far have considered the top-down aspect of recognition as a cognitive process, in form of information propagation coming from more abstract motor representations. Concerning the probability propagation in terms of bidirectional interaction between levels of motor knowledge, the approach employed in [47] is close to our work: the input of higher-level modules is the bottom-up posterior probability and the output is a set of top-down probabilities, acting to prioritize lower-level module selection.

Conclusion

In this paper, we have proposed an approach for realizing social artificial agents, based on principles of embodied cognition. We assume two key components for this: First, "horizontal integration" between perception and action such that own motor knowledge is utilized for a better recognition and understanding of others' behavior. This can be achieved by prediction-evaluation schemes, which are likely to reside on various levels of the sensorimotor hierarchy. Second, "vertical integration" which refers to a combination of bottom-up and top-down flow of information in this hierarchy. Bottom-up resonance amounts to attributing an intention to a gesture performed by the interaction partner, whereas top-down guidance informs and guides this process with hypotheses derived from prior and context-dependent knowledge.

We have presented, for the case of natural hand-arm gestures, a model that accounts for both kinds of integration. The main components of our model are (1) a hierarchy of shared motor knowledge from kinematic features of movement segments (modeled through motor commands) to complete movements (motor programs) to more general prototype representations (motor schemas); (2) a probabilistic approach to create resonances in these structures when applied for predicting and evaluating movement hypotheses against incoming observations; (3) inverse models to build up and augment these motor knowledge structures from observing and imitating novel gestures; (4) generation processes that exploit these structures for behavior production. With these components, an interaction of perception and generation could easily be modeled by imposing dynamic activation upon the shared structures, and by devising ways to evoke and respond to such activations in behavior perception and production.

While a growing interest in cognitively and (neuro)biologically inspired modeling could be observed over the last years, tenets of embodied cognition like perception-action links are mainly adopted in technical approaches to imitation learning. We, however, argue that the integration of perception and action, both horizontally and vertically, is still not sufficiently considered in computational modeling of social behavior. Imitation has been treated mostly as a one-way interaction, from a human demonstrator to an artificial agent as imitator. In this context, further bidirectional social phenomena, resulting from perception-action links, are ignored. For instance, the direct effect of perception on action, and action on perception (which lead to phenomena like priming, alignment and mimicry) are highly relevant topics in social cognitive science and should be considered as requirements for modeling social interaction. In the present work, we have tried to fulfill these requirements to some extent, while further refinements are required to position the model more clearly with respect to such cognitive phenomena. However, we have shown that the proposed computational model has the potential to capture and simulate such cognitive requirements. We have evaluated our model with real-world data (noisy gesture trajectories obtained with marker-free, camera-based body tracking). The results we have obtained are promising. Future work will have to scale this model up in terms of the number of gestures, further significant gesture features like hand shape, as well as higher levels of social and referential meaning. Nevertheless, we are confident that an integrated model as presented here is an important step, not only in investigating cognitively plausible computational models of social interaction, but also in building interactive artificial agents that can engage with their users in reciprocal interactions in a more adaptive, human-like and sociable way.

Acknowledgments This research is supported by the Deutsche Forschungsgemeinschaft (DFG) in the Center of Excellence "Cognitive Interaction Technology".

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

References

1. Tomasello M. Origins of human communication. Cambridge: Mit Press; 2008.
2. Montgomery KJ, Isenberg N, Haxby JV. Communicative hand gestures and object-directed hand movements activated the mirror neuron system. Soc Cogn Affect Neurosci. 2007; 2(2):114–22.
3. Kröger B, Kopp S, Lowit A. A model for production, perception, and acquisition of actions in face-to-face communication. Cogn Process. 2010; 11(3):187–205.
4. Brass M, Bekkering H, Prinz W. Movement observation affects movement execution in a simple response task. Acta Psychologica 2001; 106(1–2):3–22.
5. Fadiga GPL, Fogassi L, Rizzolatti G. Motor facilitation during action observation: a magnetic stimulation study. J Neurophysiol. 1995; 73(6):2608–11.

6. Buccino G, Binkofski F, Fink GR, Fadiga L, Fogassi L, Gallese V, Seitz RJ, Zilles K, Rizzolatti G, and Freund H-J. Action observation activates premotor and parietal areas in a somatotopic manner: an fMRI study. Eur J Neurosci. 2001; 13(2): 400–04.
7. Wilson M, Knoblich G. The case for motor involvement in perceiving conspecifics. Psychol Bull. 2005; 131(3): 460–73.
8. Gallese V, Goldman A. Mirror neurons and the simulation theory of mind-reading. Trends Cogn Sci. 1998; 2(12):493–501.
9. Hamilton A, Grafton S. The motor hierarchy: from kinematics to goals and intentions. In Attention and performance RY, KM, and HP (eds) Oxford University Press; 2007.
10. Zacks JM. Using movement and intentions to understand simple events. Cogn Sci. 2004; 28(6):979–1008.
11. Lakin JL and Chartrand TL. Using nonconscious behavioral mimicry to create affiliation and rapport. Psychol Sci. 2003; 14(4):334–39.
12. Kopp S. Social resonance and embodied coordination in face-to-face conversation with artificial interlocutors. Speech Commun Special Issue Speech Face-to-Face Commun. 2010; 52(6):587–97.
13. Schilbach L, Wohlschlaeger AM, Kraemer NC, Newen A, Shah NJ, Fink GR, Vogeley K. Being with virtual others: neural correlates of social interaction. Neuropsychologia 2006; 44(5): 718–30.
14. Oztop E, Franklin DW, Chaminade T, Cheng G. Human-humanoid interaction: is a humanoid robot perceived as a human?. Humanoid Robot. 2005; 2(4):537–59.
15. Sadeghipour A, Kopp S. A probabilistic model of motor resonance for embodied gesture perception. In: Ruttkay Z, Kipp M, Nijholt A, Vilhjálmsson H, editors. Intelligent Virtual Agents, vol. 5773 of lecture notes in computer science. Berlin: Springer; 2009. pp. 90–103.
16. Kopp S, Wachsmuth I. Synthesizing multimodal utterances for conversational agents. Comput Animat Virtual Worlds. 2004; 15(1):39–52.
17. Flash T, Hochner B. Motor primitives in vertebrates and invertebrates. Current Opin Neurobiol. 2005; 15(6): 660–66.
18. Mussa-Ivaldi F, Solla S. Neural primitives for motion control. IEEE J Ocean Eng. 2004; 29(3):640–50.
19. Wolpert DM, Miall RC, Kawato M. Internal models in the cerebellum. Trends Cogn Sci. 1998; 2(9):338–47.
20. Wolpert DM and Kawato M. Multiple paired forward and inverse models for motor control. Neural Netw. 1998; 11(7–8):1317–29.
21. Billard A, Schaal S. Robust learning of arm trajectories through human demonstration. In: Proceedings 2001 IEEE/RSJ international conference on intelligent robots and systems 2001; 2:734–39.
22. Sadeghipour A, Yaghoubzadeh R, Rüter A, Kopp S. Social motorics—towards an embodied basis of social human-robot interaction. Human CenterRobot Syst. 2009; 6:193–203.
23. Robert CP. Prior feedback: a Bayesian approach to maximum likelihood estimation. Comput Statistic 1993; 8:279–94.
24. Mukamel R, Ekstrom AD, Kaplan J, Iacoboni M, Fried I. Single-neuron responses in humans during execution and observation of actions. Current Biol. 2010; 20(8):750–56.
25. Dijksterhuis A, Bargh J. The perception-behavior expressway: Automatic effects of social perception on social behavior. Adv Exp Soc Psychol. 2001; 33:1–40.
26. Cook SW and Tanenhaus MK. Embodied communication: speakers' gestures affect listeners' actions. Cognition 2009; 113(1): 98–104.
27. Schutz-Bosbach S, Prinz W. Perceptual resonance: action-induced modulation of perception. J Trends Cogn Sci. 2007; 11(8):349–55.
28. Aleotti J, Caselli S. Robust trajectory learning and approximation for robot programming by demonstration. Robot Auton Syst. 2006; 54(5):409–13.
29. Rett J, Dias J. Gesture recognition using a marionette model and dynamic Bayesian networks DBNs. In: Campilho A, Kamel M, editors. Image analysis and recognition, vol. 4142 of lecture notes in computer science. Berlin: Springer; 2006, pp. 69–80.
30. Kulic D, Takano W, Nakamura Y. Incremental learning, clustering and hierarchy formation of whole body motion patterns using adaptive Hidden Markov chains. Int J Robot Res. 2008; 27(7):761–84.
31. Shon A, Storz J, Rao R. Towards a real-time Bayesian imitation system for a humanoid robot. IEEE Int Conf Robot Autom. 2007:2847–52.
32. Demiris J, Hayes GR. Imitation as a dual-route process featuring predictive and learning components: a biologically plausible computational model. In: Imitation in animals and artifacts. Cambridge: MIT Press; 2002. pp. 327–61.
33. Demiris Y, Johnson M. Distributed, predictive perception of actions: a biologically inspired robotics architecture for imitation and learning. Conn Sci. 2003; 15(4):231–43.
34. Mataric MJ. Sensory-motor primitives as a basis for imitation: Linking perception to action and biology to robotics. In: Dautenhahn K, Nehaniv CL, editors. Imitation in animals and artifacts. Cambridge: MIT Press; 2002. pp. 391–422.
35. Schaal S, Ijspeert A, Billard A. Computational approaches to motor learning by imitation. Philos Trans R Soc Lond. 2003; 358(1431):537–47.
36. Amit R, Mataric M. Learning movement sequences from demonstration. in ICDL '02: Proceedings of the 2nd international conference on development and learning. Cambridge: MIT Press; 2002. pp. 203–08.
37. Oztop E, Kawato M, Arbib M. Mirror neurons and imitation: a computationally guided review. Neural Netw 2006; 19(3): 254–71.
38. Rabiner LR. A tutorial on hidden Markov models and selected applications in speech recognition. In: Proceedings of the IEEE vol. 77. San Francisco: Morgan Kaufmann Publishers Inc.; 1989. pp. 267–96.
39. Calinon S, D'halluin F, Sauser E, Caldwell D, Billard A. Learning and reproduction of gestures by imitation. Robot Autom Mag IEEE 2010;17(2):44–54.
40. Billard AG, Calinon S, Guenter F. Discriminative and adaptive imitation in uni-manual and bi-manual tasks. Robot Auton Syst. 2006; 54(5):370–84. (The Social Mechanisms of Robot Programming from Demonstration).
41. Calinon S, Billard A; Learning of gestures by imitation in a humanoid robot. In: Dautenhahn K, Nehaniv CL, editors. Imitation and social learning in robots, humans and animals: behavioural, social and communicative dimensions. Cambridge: Cambridge University Press; 2007. pp. 153–77
42. Calinon S, Billard A. Incremental learning of gestures by imitation in a humanoid robot. In: HRI '07: Proceedings of the ACM/IEEE international conference on Human-robot interaction. New York: ACM; 2007. pp. 255–62.
43. Oztop E, Arbib MA. Schema design and implementation of the grasp-related mirror neuron system. Biol Cybern. 2002; 87(2): 116–40.
44. Tani J, Ito M, Sugita Y. Self-organization of distributedly represented multiple behavior schemata in a mirror system: reviews of robot experiments using RNNPB. Neural Netw. 2004; 17(8–9):1273–89.
45. Haruno M, Wolpert DM, Kawato M. MOSAIC model for sensorimotor learning and control. Neural Comput. 2001; 13(10): 2201–20.
46. Barsalou LW. Perceptual symbol systems. Behav Brain Sci. 1999; 22(04):577–660.
47. Haruno M, Wolpert DM, Kawato M. Hierarchical MOSAIC for movement generation. International congress series vol. 1250,

2003. pp. 575–590. Cognition and emotion in the brain. Selected topics of the international symposium on limbic and association cortical systems.

48. Johnson M, Demiris Y. Abstraction in recognition to solve the correspondence problem for robot imitation. In: TAROS; 2004. pp. 63–70.

49. Krüger V, Herzog D, Baby S, Ude A, and Kragic D. Learning actions from observations. Robot Autom Mag IEEE 2010; 17(2):30–43.