Learning through Observation and Imitation: An Overview of the ConSCIS Architecture

Antonio Chella and **Haris Dindo** and **Salvatore Gaglio**
Department of Informatics Engineering, University of Palermo
Viale delle Scienze, 90128 Palermo, Italy
chella@unipa.it, dindo@csai.unipa.it, gaglio@unipa.it

Abstract

Imitation in robotics is seen as a powerful means to reduce the complexity of robot programming. It allows users to instruct robots by simply showing them how to execute a given task. Through imitation robots can learn from their environment and adapt to it just as human newborns do. In order to be useful as human companions, robots must act for a purpose by achieving goals and fullfiling human expectations. But, what is the goal behind the surface of the demonstrated behavior? How to extract, encode and reuse eventual regularities observed? These questions are indispensable for the development of cognitive agents capable of being human companions in everyday life. In this paper we present ConSCIS, a framework for robot teaching through observation and imitation inspired by recent findings in cognitive sciences, biology and neuroscience. In ConSCIS we regard imitation as the process of manipulating high-level symbols in order to achieve goals and intentions hidden in the observation of task. The architecture has been tested both in simulation and on an anthropomorphic robot platform.

Introduction

State of the art robots show only limited capabilities of perception, reasoning and action in new and unstructured environments. A new generation of robotic agents, able to perceive and act in new and unstructured environments, and to learn from it, is needed. They should be able to pay attention to the relevant entities in their environment, to image, predict and to effectively plan their actions, and to acquire new skills, behaviors and knowledge through social learning and interaction with other agents (both humans or robots).

In the field of robot teaching two major approaches have been traditionally adopted: *programming* (explicitly tell the robot each detail of what to do trough the use of a programming language) and *learning* (let the robot figure out itself what to do given a description of an objective). Both approaches have proved to be too weak to provide an effective paradigm of robot teaching. Programming does not scale well to unforeseen situation, and requires a considerable effort. Learning, on the other hand, is an ill-posed problem due to the dimensionality of the search space the robot is faced with. A novel approach, which tries to overcome these limitations, is emerging in the past years. It is based on the idea that the robot could learn new behaviors by observing and imitating the behavior of others: *imitation learning*.

Imitation, as a powerful mechanism of social learning, has received a great deal of interest from researchers in the fields of behavioral and cognitive sciences, neuroscience, artificial intelligence and robotics. It is believed that imitation represents the manifestation of a higher-cognition and the most complex form of animal learning. In addition, the ability to learn by watching others (and in particular the ability to imitate) is thought to be a precursor to the development of appropriate social behavior, and ultimately the means to reason about the thoughts, intents and desires of others. From an engineering perspective, a mechanism that allow to imitate the actions of others would provide a simple means for a robot to acquire new skills and tasks without any additional programming. Imitation eases the task of robot programming and facilitates the transmission of complex skills. The same ideas may be employed in the more general context of interaction between software agents and humans.

What could an efficient implementation of imitative mechanisms provide to robots?

- **Robustness**: by observing other agent's actions, the robot can figure out new behaviors that are likely to be useful, so it can continuously learn new skills, or adapt if the environment changes;
- **Learning**: a robot can learn from other skilled agents; the learning process is significantly speeded-up compared to trial-and-error approaches;

In this paper we will restrict ourselves to a particular imitative process, namely the *goal-level imitation*, where focus is put on effects of actions on objects, without taking into account their kinematic or dynamic properties. We believe that such capabilities should involve the generation of a high-level declarative description of the world as it is being perceived by the robot. This process requires both *bottom-up* data driven processes that associate data coming in from robot's sensors to symbolic knowledge representation structures, and *top-down* processes in which high-level symbolic information is employed to drive robot's actions. From a robotic point of view, the introduction of abstraction and conceptualization into the imitation paradigm seems a

promising approach towards more sophisticated imitative architectures.

In the following section we will try to briefly give a comprehensive overview of different facets of imitative behavior from the point of view of psychology, cognitive sciences, neuroscience and robotics, aiming to provide a unified view of the same problem.

Related work

Psychologists have been studied imitation since the beginning of the last century. However,despite a century of research and a great deal of interest, the processes underlying imitation remain largely unknown. However, it is an essential mechanism in studying the acquisition of novel skills in agents, being they biological or artificial.

The earliest studies on imitation are that of Thorndike (Thorndike 1898) and Piaget (Piaget 1962) who claimed that "imitating" or "mimicking" is not an expression of higher intelligence. However, research recently performed by Meltzoff and Moore (Meltzoff and Moore 1977) showed that 2- to 3-week-olds imitated tongue protrusion, mouth opening, lip protrusion, as well as simple finger movements, without having never neither seen their own faces nor been exposed to viewing faces of other humans for any significant amount of time. Thus, the ability to map a perceived facial gesture to their own gestures was concluded to be innate and contradicted Piaget's ontogenetic account of imitation.
Unfortunately, imitation does not have a unique meaning, since different similar behaviors may be characterized as imitation. Several taxonomies are possible which have a significant impact on how imitation is developed in artificial agents. Byrne and Russon provide an overview of imitative behaviors observed in animals and humans. Several "imitative" behaviors are analyzed and divided into low- and high-level imitation. Low-level imitation is seen as a process of copying an observed behavior, where primitive motion patterns which compose that behavior are already in the ones repertoire. On the other hand, the high-level imitation does actually lead to the acquisition of a novel behavior through imitation (Byrne and Russon 2000).

As cognitive beings, humans go beyond the surface of the observed behavior. Persons have *beliefs*, *desires*, and *intentions*. Research on the *theory of mind* investigates the development of this framework. In humans, this ability is accomplished in part by viewing the other as being "*like me*" (Goldman 2001). This may help us to predict and explain others emotions, behaviors and other mental states, and to formulate appropriate responses based on this understanding. For instance, it enables us to infer the intent or goal enacted by another's behavior: an important skill for enabling richly cooperative behavior (Breazeal 2002). Recently, it has been postulated that imitation provides a foundation for developing the theory of mind (Meltzoff 1999).
Advances in neuroscience postulate there may be a common neural substrate for coding action, understanding goals/intentions, and processing theory of mind What are the brain regions for involved in imitation? Among other neuron types, area F5 contains a class of neurons that discharge both when the monkey *performs* a particular hand or

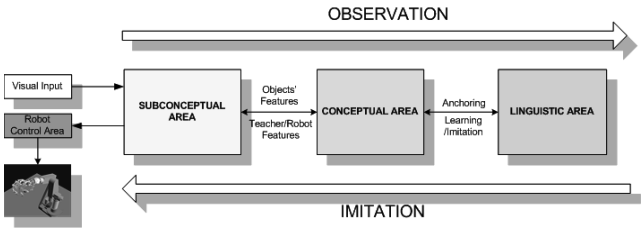


Figure 1: Overview of the ConSCIS architecture

mouth actions and when the monkey *observes* another individual (being its conspecific or a human) performing a similar action. These neurons have been called *mirror neurons* after this very peculiar property. These neurons are triggered only by the observation of an agent and an object, and do not respond to the sight of the hand miming an action, or to the observation of an object alone. It has been hypothesized that mirror neurons represent the neural substrate for action recognition performed by others (Gallese et al. 1996; Rizzolatti, Fogassi, and Gallese 2001). Neurons with similar characteristics were found also in humans through *TMS* (Transcranial Magnetic Stimulation) and *fMRI* (functional Magnetic Resonance Imaging) studies in the so-called Brocca's area, a region considered to be devoted to speech production. These data are being investigated as an interesting evolutionary scenario: linking the origin of language with the comprehension of hand actions (Rizzolatti and Arbib 1998). Some studies links mirror neurons with experiential understanding of the emotions of others and empathy (Gallese, Keysers, and Rizzolatti 2004).

In artificial agents, the implementation of imitation mechanisms has been recognized to be an important milestone toward a new generation of intelligent machines (Schaal 1999; Arbib 2003; Breazeal et al. 2005) Probably the first work regarding imitation in robotics is reported by Kuniyoshi in (Kuniyoshi, Inaba, and Inoue 1989). In this approach, the robot learns how to perform a new task by watching a human perform the same task. In other work with highly articulated humanoid robots, learning by demonstration has been explored as a way to achieve efficient learning of dexterous motor skills (Schaal, Ijspeert, and Billard 2003). The state-action space for such robots is prohibitively large to search for a solution in reasonable time. To address this issue, the robot observes the human's performance, using both object and human movement information to estimate a *control policy* for the desired task. Another way to accelerate learning is to encode the state-action space using a more compact representation. Researchers have used biologically-inspired representations of movement, such as *movement primitives*, to encode movements in terms of goal-directed behaviors rather than discrete joint angles (Matarić, Zordan, and Mason 1998). More recent works deal with the so called goal-level imitation and intention reading. In (Jansen and Belpaeme 2006), the authors propose a computational model for learning the goal, or intent, of a demonstration using a mode which draws inspiration from psychological models of imitation. Agents then imitate the goals of other agents

behavior rather than their exact actions. Another important issue in imitation is that of determining a measure of the similarity across demonstrator and imitator motions A closely related problem in imitation is the so called *correspondence problem*, which deals with mapping action sequences of the demonstrator and the imitator agent. This problem becomes particularly obvious when the two agents do not share the same embodiment and affordances (Alissandrakis, Nehaniv, and Dautenhahn 2002).

ConSCIS framework for imitation learning

This work will give an overview of**ConSCIS** (*Conceptual Space based Cognitive Imitation System*), a framework for imitation learning.with the focus on how and what to imitate questions. The how to imitate question involves the problem of inverse kinematics and robot control: how to reach a certain state given your body configuration and constraints from the robot's actuators and from the environment. The what to imitate question is much more intricate: how can an agent know what it should imitate? Humans solve this problem by inferring the intention of the demonstrator and imitating the intention only, which is the approach followed in the present work

In the following, we will give a brief overview of the ConSCIS architecture. For a detailed description, please refer to (Chella, Dindo, and Infantino 2007). The architecture is broadly organized into three computational areas as shown in Fig. 1, while Fig. 2 depicts an exploded view of the architecture and the main connections between various units. We have followed the well-established framework of three-layered cognitive robotic architectures described in (Chella, Frixione, and Gaglio 2001). What is each layer responsible of?

Subconceptual Area

This area is is concerned with low-level sensory data processing. It contains mainly artificial vision algorithms for robot perception. This area is phylogenetic in the sense that it is defined beforehand by the system designer and does not involve any learning. This area is the contact point between the architecture and the environment and it poses attention toward certain classes of stimuli. However, perception is not just a matter of the proximal stimulus but also a matter of the information extracted and inferred from the stimulus. This information is stored in the next area.

Conceptual Area

This are is based on the *"conceptual space"* model proposed by Gärdenfors (Gärdenfors 2000). Information is organized into geometrical conceptual structures independent of the symbolical descriptions. This area maintains a memory of the experience held by the system about its environment: property of the objects it encounters (shape, color, relative displacement, etc) and particular actions being performed by the demonstrator. The conceptual area is fundamental in relating symbolical representations to low-level data, problem known as symbol grounding.
In the current work we use two conceptual spaces:

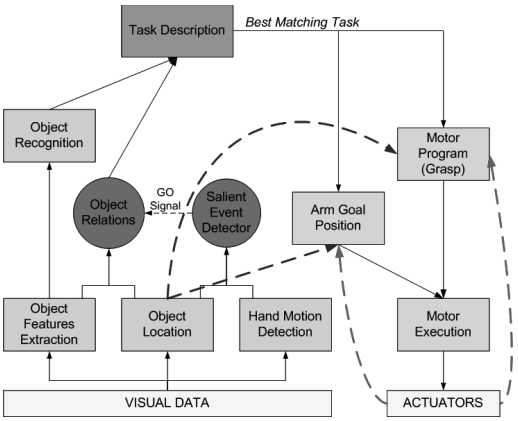


Figure 2: A detailed view of the ConSCIS architecture with main processing units and their mutual interconnections

- Perceptual space, PS, encodes common perceptual properties of the objects being observed by an agent. Our space of perceptions is composed of two different domains: *shape* and *color*. Each domain is intimately connected to the sensory capabilities of the robot.
- Situation Space, SS, encodes spatial relations between objects in a scene, as well as absolute position/orientation of detected objects. It depicts what is called *effect metrics* in the context of imitation (Alissandrakis, Nehaniv, and Dautenhahn 2002). The conceptual spaces proposed capture the *what to imitate* property in imitation based robotic architectures

Linguistic Area

Linguistic area is based on the symbolical formalisms and the elaborations are those of the logical calculus. Elements of the linguistic area are symbolic terms grounded to sensory data by mapping them on appropriate representations in different layers of the conceptual area. Symbolical labels of important concepts (e.g. shape and color of objects, their relative and absolute positions, actions performed by the user, task names) are initially obtained through human-robot interaction, and then stored in the conceptual spaces as the internal model of the robot's perceptions. Learning of concepts is gradually performed on these internal representations. The system can manipulate symbols which emerge from the interaction between the robot and its environment, and which have passed a series of levels (sensory, subconceptual and conceptual) using linguistic-logical-syntactical means which characterize serial cognitive processes.

Robot Control Area

This are is concerned with the control issues of the physical robotic system. It is composed of two main modules: *navigation*, in which robot trajectories are computed, and *grasping*, in which a neuro-genetic algorithm computes the most appropriate configuration of robotic fingers in order to firmly grasp a given object (Chella et al. 2007). This area

abstracts the rest of the architecture from a particular robotic platform. Its goal is to translate a high-level description of a task (e.g. *pick up circle-shaped object at {10cm, 20cm, 0.14rad}*) into low-level control commands to be sent to the robot's actuators.

Imitation in ConSCIS

How can an agent know "*what*" it should imitate? Humans solve this problem by *inferring the intention* of the demonstrator and imitating the intention only, namely the goal of the observed sequence of actions. This is the approach adopted in the present work: imitation will be seen as the process of manipulating high-level symbols, grounded to real-world data through a set of layers, in order to achieve goals and intentions hidden in the observation of task. This allows the robot to reuse and generalize the acquired knowledge in novel and unseen circumstances.

The process of goal extraction is achieved through a set of *imitation games*, which involve repetitive interactions between the demonstrator and imitator. Each imitation game is composed of a phase of *observation*, followed by a phase of *imitation*. During the observation phase, the robot observes a human while performing tasks in the world. Tasks we have considered regard meaningful displacements of objects in front of the robot. Through subconceptual space modules the system extracts information about important features of the world: *shape* and *colour* of objects in the scene, their *relative spatial arrangement* during several stages of the task demonstration, and *sequencing of elementary actions* performed on objects.

Some features may be known to the robot through *a priori* knowledge (which actions it can detect and perform), while some others are learned during its *lifetime*. Former may be labels of shapes, colours, or spatial relations between object, which are assigned through human-robot interaction. These high-level labels are grounded to the robot's perception of the world trough the intermediate conceptual area layer, which provides a powerful substrate for semantic interpretation of symbols. For each observed task the linguistic area stores a symbolical description of objects' properties, together with an annotation of actions performed on them.

The ability of the our architecture to recognize and represent different objects, their relative spatial arrangement, as well as elementary actions performed on objects, is crucial to the capability to imitate. Its goal is to achieve corresponding effects of arbitrary tasks seen across different observation sessions, since our focus is on final effects of actions on objects.

During the imitation phase, the robot is faced a novel scene and we expect it to autonomously decide what to do. This choice could not be a mere replication of the observed sequence of actions, since this would lead to an erroneous outcome of the imitation game in different world configurations. The robot chooses which task to imitate depending on the strength of the perceived similarities between objects (and their relations) in the current scene, and all previously observed scenes. In other words, the robot must *abstract* shape, colour and relative displacement properties of objects in a task, and reuse this knowledge in a perceptually similar scene: it must *infer* the intention of the user.

Categorization of objects

During task observation, the scene may be populated with several objects which can be directly or indirectly involved in the task. On the other hand, some objects may not carry any useful information. In order to deal with this, we introduce a categorization of objects based on their relevance in a task.

Objects seen in the observation mode are divided into three distinct *relevance categories*:

- *Relevant objects*: objects directly involved into the execution of an action, i.e. objects grasped and transferred, and objects relative to which actions are executed;
- *Situation objects*: objects having a spatial relation with a relevant object; they are relative only to the *initial* displacement of objects, before any action occur;
- *Context objects*: other objects neither directly nor indirectly involved in the task.

Recognizing which objects are carrying information about an observed task is a crucial step towards the process of intention inference. Its reliable detection would be possible only if we were able to observe hidden mental states of the demonstrator during the task execution. We use a heuristics which assigns each object to a relevance category on the basis of actions performed in the task, and *strength* of spatial relations between objects. The latter is stored in the linguistic area as the sequence of observed actions, while the former is stored in the situation space. Computation of the strength of relations is based on the perceived distance between objects and an empirically chosen threshold. In other words, if two objects are close enough their relation is encoded and processed in order to extract the object's relevance category. Furthermore, an important role in the description of a task is given by the *initial* relative displacement of the objects (an information stored in the situation space), which may carry a useful information for deciding which action to perform in a given context.

How and what does the robot imitate?

This section depicts the whole process involved in the ConSCIS architecture. We start from the observation phase, in which a human teacher performs a task. The robot, through its perceptual capabilities, is able to segment objects involved from the background and extract their shape, color and position. In addition to extracting objects and their perceptual and spatial properties, the robot represent the sequence of actions performed on objects as well. This is done by continuously tracking the human hand and segmenting its motion into meaningful events.

While observing the execution of a task, the robot extracts and stores the following information: perceptual properties of objects, relative displacement between objects in each step of the observation, the description of the task being performed, and a list of all relevant, situation and context objects.

Recognition of important properties then allows the robot to assign a label to each object, which will be used to conceptualize its knowledge about the world. Perceptual properties of each object (i.e. its shape and color parameters) are stored in the perceptual space, while the relative displacement between objects, as well as their absolute position in the world is stored in the situation space. If it is the first time the robot sees an object, it asks the user about its properties and adjust its beliefs about concepts in the world. In other words, it performs the clustering of the conceptual spaces. These features are then represented as high-level symbols in the linguistic area.

Given a scene in the imitation phase, the sequence of actions to execute by the robot (if any) is selected on the basis of similarity between the configuration and visual properties of the current scene and all observed scenes. Hence, it may be seen as a process of *pattern matching*: the robot searches an observed scene which *best* matches with the given scene. In other words, the problem may be reduced to the process of associating objects in the workspace to the objects seen during task observation, in such a way that the execution of the *same* sequence of elementary actions (reach/grasp or transfer/release) on the current objects will achieve the corresponding effects. Each possible match is given a score which is based on assigning weights to identical shape, colour and initial relative displacement properties between objects in the observation and imitation phases. The most promising matching (i.e. the one with the highest score) is then selected for the execution. In case of several competing matches across several observations, the most recent one is preferred. A detailed description of the ConSCIS matching algorithm is provided in (Chella, Dindo, and Infantino 2008).

Actions that have been observed in the observation with the highest match are then adjusted to the current configuration of objects. This is done by traversing the top-down path in the architecture, from the linguistic area to the lowest level layer which controls the robot's actuators. The absolute positions in the world reference system are computed through the situation space, and then sent to robot control module for the execution. This module computes the path to be followed, and synthesizes a grasp for a given object.

Experiments

The system has been tested both in simulation and on a arm/hand robotic platform. Tests we have performed are concerned with the problem of teaching the robot several tasks of composing workspace objects.
As a toy example, suppose the robot observes the two tasks performed by a human user in two different sessions represented in the figure 3. Each observation is processed and a linguistic description is created for each task. During the imitation phase the robot must choose which known observation matches the best with the current scene, by associating objects from the imitation phase to the objects in the observation phase. This is done by the matching process in such a way that the execution of the observed task, grounded to the novel objects, yields to similar effects.

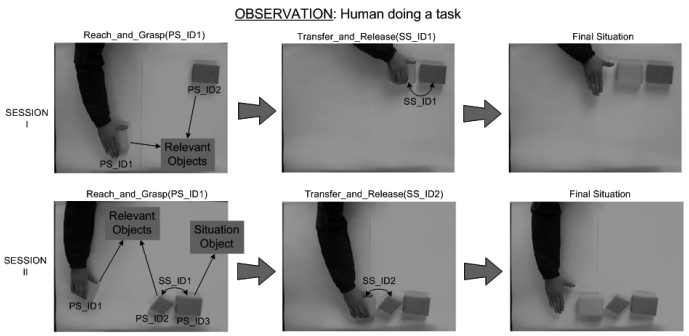


Figure 3: Observation and processing of a human performing a task

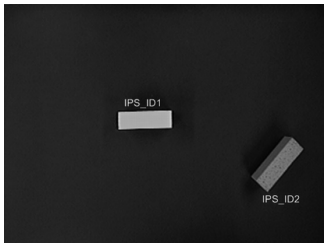


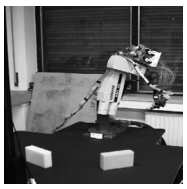
Figure 4: What should the robot imitate?

An example of the scene in which the robot should perform the imitation is shown in Fig. 4. The matching process finds the first observation more relevant to the task at hand. Thanks to the conceptual area the robot can simulate beforehand the execution of the action. Based on the , and information stored in the situation space, the robot computes the set of actions to perform, which are then sent to the robot control module in order to compute the trajectory the robot should follow. The grasping module computes the best grasp for the given object. The sequence of actions executed by the robot are shown in Fig. 5.

Conclusions

In this paper we have introduced a cognitive system for imitation learning. The architecture focuses on "how" and "what" to imitate questions, and is broadly organized into three computational areas: subconceptual, conceptual and linguistic, together with a module for robot control and object grasping. We see imitation as the process of manipulating high-level symbols in order to achieve goals and intentions hidden in the observation of task. This allows to reuse and generalise the stored knowledge in novel and unseen circumstances.

The architecture integrates several aspects of imitation, such as *perception* and *extraction of the relevant features* of the demonstrated behavior, *learning* of novel behaviors, *knowledge representation*, *action generation* and execution of *motor commands*. This differs from previous approaches where focus is put on a particular act of an imitative behavior. Another novelty is represented by the introduction of conceptual and linguistic processing areas which allow to



(a) Initial situation

(b) Reaching and grasping

(c) Transferring

(d) Final situation

Figure 5: Successful execution of the imitation task on the robotic system

represent task data through high-level symbols, grounded to real-world data trough an intermediate conceptual layer.

Current work is oriented toward integrating aspects of perspective taking and mental state inference and prediction in the ConSCIS framework. If robots can learn and infer the intentions of other agents, they can use this knowledge to predict and anticipate other agents' behaviors in a cooperative setting. This approach is intimately related to the "theory of mind", and involves several hard problems dealing with mental states of others. The ultimate goal (or a dream?) is to develop a robotic system which can autonomously learn behaviors through observation and imitation, acquire a language through its own experience, and engage in complex interaction with humans.

References

Alissandrakis, A.; Nehaniv, C.; and Dautenhahn, K. 2002. Imitation with ALICE: learning to imitate corresponding actions across dissimilar embodiments. *IEEE Transactions on Systems, Man and Cybernetics, Part A* 32(4):482–496.

Arbib, M. 2003. *The handbook of brain theory and neural networks*. MIT Press Cambridge, MA.

Breazeal, C.; Buchsbaum, D.; Gray, J.; Gatenby, D.; and Blumberg, B. 2005. Learning From and About Others: Towards Using Imitation to Bootstrap the Social Understanding of Others by Robots. *Artificial Life* 11(1-2):31–62.

Breazeal, C. 2002. *Designing Sociable Robots*. Bradford Book.

Byrne, R., and Russon, A. 2000. Learning by imitation: A hierarchical approach. *Behavioral and Brain Sciences* 21(05):667–684.

Chella, A.; Dindo, H.; Matraxia, F.; and Pirrone, R. 2007. A neuro-genetic approach to real-time visual grasp synthesis. In *Proc. of the International Joint Conference on Neural Networks (IJCNN)*, 1619–1626.

Chella, A.; Dindo, H.; and Infantino, I. 2007. Imitation learning and anchoring through conceptual spaces. *Applied Artificial Intelligence* 21(4-5):343–359.

Chella, A.; Dindo, H.; and Infantino, I. 2008. A cognitive approach to goal-level imitation. *Journal of Interaction Studies (Social Behaviour and Communication in Biological and Artificial Systems)* 9:2:301–318.

Chella, A.; Frixione, M.; and Gaglio, S. 2001. Conceptual Spaces for Computer Vision Representations. *Artificial Intelligence Review* 16(2):137–152.

Gallese, V.; Fadiga, L.; Fogassi, L.; and Rizzolatti, G. 1996. Action recognition in the premotor cortex. *Brain* 119(2):593–609.

Gallese, V.; Keysers, C.; and Rizzolatti, G. 2004. A unifying view of the basis of social cognition. *Trends in Cognitive Sciences* 8(9):396–403.

Gärdenfors, P. 2000. *Conceptual spaces: the geometry of thought*. MIT Press-Bradford Books.

Goldman, A. 2001. Desire, intention, and the simulation theory. *Intentions and intentionality: Foundations of social cognition* 207–225.

Jansen, B., and Belpaeme, T. 2006. A computational model of intention reading in imitation. *Robotics and Autonomous Systems* 54(5):394–402.

Kuniyoshi, T.; Inaba, M.; and Inoue, H. 1989. Teaching by Showing: Generating Robot Programs by Visual Observation of Human Performance. *Proc. of the 20th International Symp. on Industrial Robots* 119–126.

Matarić, M.; Zordan, V.; and Mason, Z. 1998. Movement control methods for complex, dynamically simulated agents: Adonis dances the Macarena. *Proceedings of the second international conference on Autonomous agents* 317–324.

Meltzoff, A., and Moore, M. 1977. Imitation of facial and manual gestures by human neonates. *Science* 198(4312):74–8.

Meltzoff, A. 1999. Origins of theory of mind, cognition and communication. *Journal of Communication Disorders* 32(4):251–269.

Piaget, J. 1962. Play, Dreams and Imitation in Children.

Rizzolatti, G., and Arbib, M. 1998. Language within our grasp. *Trends in Neurosciences* 21(5):188–194.

Rizzolatti, G.; Fogassi, L.; and Gallese, V. 2001. Neurophysiological mechanisms underlying the understanding and imitation of action. *Nat Rev Neurosci* 2(9):661–70.

Schaal, S.; Ijspeert, A.; and Billard, A. 2003. Computational approaches to motor learning by imitation. *Philosophical Transactions: Biological Sciences* 358(1431):537–547.

Schaal, S. 1999. Is imitation learning the route to humanoid robots. *Trends in Cognitive Sciences* 3(6):233–242.

Thorndike, E. 1898. Animal intelligence: an experimental study of the associative processes in animals.