Citation:

*E. Bicho, W. Erlhagen, L. Louro, E. Costa e Silva, R. Silva, N. Hipolito, "**A dynamic field approach to goal inference, error detection and anticipatory action selection in human-robot collaboration**", In Dautenhah, Kerstin and Saunders, Joe (Eds), New Frontiers in Human Robots Interaction, [Advances in Interaction Studies, 2] 2011. Vi, 332 (pp. 135-164, John Benjamins Publishing Company. Hardbound:* ISBN 978 90 272 0455 4 eBook: ISBN 978 90 272 8339 9

Running head: A DNF approach to human-robot joint action

A dynamic field approach to goal inference, error detection and anticipatory action selection in human-robot collaboration

Estela Bicho[a1], Wolfram Erlhagen[b], Luís Louro[a], Eliana Costa e Silva[a], Rui Silva[a], Nzoji Hipolito[a]

[a]Department of Industrial Electronics
[b]Department of Mathematics and Applications
University of Minho, Portugal

Abstract

In this chapter we present results of our ongoing research on efficient and fluent human-robot collaboration that is heavily inspired by recent experimental findings about the neurocognitive mechanisms supporting joint action in humans. The robot control architecture implements the joint coordination of actions and goals as a dynamic process that integrates contextual cues, shared task knowledge and the predicted outcome of the user's motor behavior. The architecture is formalized as a coupled system of dynamic neural fields representing a distributed network of local but connected neural populations with specific functionalities. We validate the approach in a task in which a robot and a human user jointly construct a toy 'vehicle'. We show that the context-dependent mapping from action

[1]Email addresses: estela.bicho@dei.uminho.pt (corresponding author), wolfram.erlhagen@math.uminho.pt (Wolfram Erlhagen), llouro@dei.uminho.pt (Luís Louro), esilva@dei.uminho.pt (Eliana Costa e Silva), rsilva@dei.uminho.pt (Rui Silva) and nhipolito@dei.uminho.pt (Nzoji Hipolito).

observation onto appropriate complementary actions allows the robot to cope with dynamically changing joint action situations. More specifically, the results illustrate crucial cognitive capacities for efficient and successful human-robot collaboration such as goal inference, error detection and anticipatory action selection.

1. Introduction

As robot systems are moving as assistants into human everyday life, the question how to design robots capable of acting as sociable partners in collaborative joint activity becomes increasingly important (Breazeal, 2004; Fong, Nourbakhsh and Dautenhahn, 2003). Useful and efficient human-robot collaboration requires that both teammates coordinate and synchronize their actions and decisions in a shared task. In order to decrease the workload of the human and to increase user satisfaction, the robot should equally contribute to this coordination effort. This necessarily means that the robot should be endowed with cognitive capacities such as action understanding, action monitoring and goal inference. Humans achieve their remarkable fluent organization of joint action by anticipating the motor intentions of others (Sebanz, Bekkering and Knoblich, 2006). In our everyday social interactions we continuously monitor the actions of our partners, interpret them effortlessly in terms of their outcomes and use these predictions to select adequate complementary behaviours. Very often this happens without the need for explicit verbal communication. Imagine for the instance the joint action task of preparing a dinner table. The way that a partner grasps a certain object, e.g., a coffee cup, transmits to the observer important information about the ultimate goal of the action. Depending on the grip type, the partner may want to place the cup on the table or, alternatively, has the intention to hand it over to the co-actor. Being able to predict the goal of the whole action sequence at the time of the grasping allows the observer to timely prepare for receiving the cup, or to initiate the selection of another object for the dinner table. However, even in routine joint activity the co-actor may perform actions that are in some way incorrect or inappropriate. The partner may for instance want to hand over a second spoon for the sugar bowl or may pick a cup without having already placed the saucer on the table. Being able to evaluate the predicted outcomes of the co-actor's actions with respect to the (sub)goals of the shared task is thus a fundamental capacity for efficient and successful joint action. It allows the observer to overrule a familiar response (e.g., accepting the spoon) or to initiate an adequate corrective behaviour (e.g., quickly grasping a saucer for placing it on the table).

This chapter presents results of our ongoing research towards creating socially intelligent robots that are able to flexibly adjust their goal-directed behaviours depending on the predicted outcomes of actions of their human partners. For the experiments we used a more complex version of a joint assembly task introduced in our previous work (Bicho et al., 2009; Bicho et al., 2010) in which the human-robot team has to assemble a toy object from its components. The focus is on implementing and testing in the robot context-sensitive action monitoring and anticipatory action selection capacities. Our approach is heavily inspired by recent experimental and theoretical findings about the neurocognitive mechanisms underlying joint action in humans (Bekkering et al., 2009; Sebanz, Bekkering and Knoblich, 2006). We believe that designing cognitive control architectures on the basis of these mechanisms defines a very promising research direction to reduce the significant imbalance in social and cognitive skills between human and robot that still exists today. Ultimately, implementing a human-like joint action model in the robot will contribute to more natural human-robot collaboration since the teammates will become more predictable to each other. This in turn will increase the acceptance by humans.

An impressive body of experimental evidence from behavioural and neurophysiological studies investigating action and perception in a social context shows that when we observe others' actions, corresponding motor representations in our motor system become activated (for reviews see (Wilson and Knoblich, 2005; Rizzolatti and Craighero, 2004). These findings have been interpreted as supporting the hypothesis that the mere perception of actions automatically increases the likelihood of the performance of those actions, without the person's conscious awareness. During the last decade, the idea of an obligatory and direct perception-action link has inspired robotics work mainly in the domain of learning by imitation and social development (Billard et al., 2008; Erlhagen et al., 2006; Demiris and Johnson, 2003; Alissandrakis et al., 2002; Schaal, 1999). A major insight of this work is that the matching between action observation and action execution has to solve the correspondence problems that exist between agents with dissimilar embodiment. Moreover, the metrics of the mapping should be highly task-dependent and may range from the level of movement kinematics to the level of desired end states or action goals.

While an automatic facilitation of corresponding motor representations during action observation may support social learning, it is normally not beneficial for cooperative joint action tasks that require the facilitation of complementary motor programs. An alternative proposal for a functional role of the automatic action resonance mechanism suggests that it contributes to understanding the actions of other individuals during social interactions (Rizzolatti and Craighero, 2004; Wilson and Knoblich, 2005).

The key idea is that the observer performs an internal motor simulation to predict the consequences of perceived actions using knowledge of his or her actions and motor intentions. During joint action, the representation of the inferred goal of the co-actor together with representations of prior task knowledge may then automatically bias the observer's decision process towards selecting an adequate complementary behaviour. In line with this hypothesis, the findings of a recent behavioural study suggest that the perception-action coupling appears to be indeed to some extent under the control of task and goal representations. By comparing motor planning in an imitative and a cooperative setting van Schie et al. (2008) demonstrated the reversal of the automatic action congruency effect. In the cooperative setting, people were faster to respond to an observed action with a complementary behaviour compared to the matching behaviour.

The robot control architecture for human-robot collaboration implements such a context-sensitive, i.e. flexible, mapping between action observation and action execution. The coordination of actions and decisions among the teammates is modeled as a dynamic process that builds on the continuous integration of input from representations of the inferred goal of observed actions (obtained through motor simulation), contextual cues (e.g., location of objects in the scene) and shared task knowledge (e.g., assembly plan). The representation of the complementary action that gets the strongest support will win the dynamic competition process among all possible complementary behaviours. As a theoretical framework we have used the Dynamic Neural Field (DNF) approach to robotics (Erlhagen and Bicho, 2006). Originally introduced as a simplified mathematical model for pattern formation in neural populations (Amari, 1977; Wilson and Cowan, 1973), DNFs have been later generalized and applied to the cognitive domain (for a recent review see Schöner, 2008). The architecture of DNFs reflects the hypothesis that strong recurrent interactions in local populations of neurons form a basic mechanism of cortical information processing. These interactions support the existence of self-stabilized representations that allow the cognitive agent for instance to compensate for temporally missing sensory input, or to anticipate future environmental inputs that may inform the decision about a specific goal-directed behaviour.

The DNF-model of joint action forms a complex dynamical system consisting of a distributed network of reciprocally connected neural populations that integrate and represent in their activation patterns task-relevant information. For the experimental validation of the model in the joint construction task we assume that both agents share the knowledge of the assembly plan representing the sequential execution of subgoals. Since the construction work cannot be performed alone, each agent has to continuously monitor and evaluate the co-actor's actions in order to guarantee success. The results reported here extend our previous work to a more realistic and complex joint action context which includes situations in which the co-actor's behaviour is only partially observable (due to occluding surfaces) and the robot has to select among several possible complementary actions. The focus of the study is on the dynamic interactions within the DNF network that support the selection of an appropriate action in anticipation of the co-actor's current goal. The anticipatory action control includes situations in which the predicted effect of the observed action is inconsistent with an efficient team performance and thus requires a corrective response. The timely decision for such a response is possible since the action planning process integrates continuously in time the activity from connected populations representing a mismatch between the inferred goal and the desired action effect in a specific joint action context.

The chapter is organized as follows: Section 2 introduces the joint construction task and the robotic platform. Section 3 gives an overview about the cognitive control architecture. Section 4 presents the basic concepts of the dynamic field framework. The results of the human-robot interactions are described in section 5. The chapter ends with a discussion of concepts and results in section 6 and conclusions and outlook in section 7.

2. Joint construction task

----------------Insert Figure 1 around here -------------

To test the dynamic field architecture for human-robot collaboration we have chosen the joint construction of a toy ‘vehicle’ from components that are initially distributed on a table (see Figure 1). The toy object consists of a round platform with an axle on which two wheels have to be attached and each fixed with a nut. Subsequently, 4 different columns have to be plugged into specific holes in the platform. The placing of another round object on top of the columns finishes the task. The components were designed to limit the workload for the vision and the motor system of the robot. Thus, the task is completely symmetric in that both the human and the robot can make assembly actions. It is assumed that each teammate is responsible to assemble one side of the toy. Since the working areas of the human and the robot do not overlap, the spatial distribution of components on the table obliges the team to coordinate handing-over sequences. In addition, some assembly steps require that one actor helps the other by holding still a part in a certain position. It is further assumed that both partners know the construction plan and keep track of the subtasks which have been already completed by the team. The prior knowledge about the sequential execution of the assembly work is represented in the DNF-architecture by pre-defined connections between populations encoding subsequent assembly steps. Since the desired end state does not uniquely define the logical order of the construction, at each stage of the construction the execution of several subtasks may be simultaneously possible.

The main challenge for the team is thus to efficiently coordinate in space and time the decision about actions to be performed by each of the teammates. The task is complex enough to show the impact of goal inference, action understanding and action monitoring and evaluation on complementary action selection.

The robot (ARoS) used in the experiments has been built in our lab (Silva, Bicho and Erlhagen, 2008). It consists of a stationary torus on which a 7 DOFs AMTEC arm (Schunk GmbH) with a 3-fingers dexterous gripper (Barrett Technology Inc.) and a stereo camera head are mounted. A speech synthesizer (Microsof Speech SDK 5.1) allows the robot to communicate the result of its reasoning to the human user. For the control of the arm-hand system we applied a global planning method in posture space that allows us to integrate optimization principles derived from experiments with humans (Costa et Silva et al, submitted). The goal is to guarantee robot motion that is perceived by the human user as smooth and goal-directed.

The information about object class, position and pose is provided by the vision system. The object recognition combines color-based segmentation with template matching derived from earlier learning examples (Westphal el al., 2008). The same technique is also used for the classification of object-directed, static hand postures such as grasping and communicative gestures such pointing and demanding an object.

3. Cognitive architecture for joint action

----------------Insert Figure 2 around here -------------

Figure 2 presents a sketch of the multi-layered robot control architecture. It reflects neurocognitive mechanisms that are believed to support human joint action (Bekkering et al., 2009). Each layer contains several neural populations encoding information relevant for the joint assembly task. Every population can receive input from multiple connected populations that may be located in different layers. Rather than describing in detail the schema of the hand-coded connections for the concrete assembly task (for an example see the supplementary material in Bicho, Louro and Erlhagen, 2010) we give here an overview about the functional role of the different layers and discuss the flow of information between layers with respect to experimental findings that have inspired our work.

Ultimately, the architecture implements a context-dependent mapping between observed action and executed action (Poljac, van Schie and Bekkering, 2009; van Schie, Waterschoot and Bekkering, 2008; Erlhagen, Mukovskiy and Bicho, 2006). The fundamental idea is that the mapping takes place on the level of abstract motor primitives defined as whole object-directed motor acts like reaching, grasping, placing, attaching or plugging. These primitives encode the motor act in terms of an observable end state or goal rather than in terms of a detailed description of the movement kinematics (Rizzolatti and Craighero, 2004; Schaal, 1999). An observed hand movement that is recognized by the vision system as a particular primitive (e.g., top grip or side grip) is represented in the action observation layer (AOL). The action simulation layer (ASL) implements the idea that by automatically matching the co-actor's action onto its own sensorimotor representations without executing it, the robot may simulate the ongoing action and its consequences. The neural populations in ASL represent entire chains of action primitives that are in the motor repertoire of the robot (e.g., reaching-grasping-placing/plugging or reaching-grasping-holding out). These chains are linked to representations of specific goals or end states (e.g., attach wheel to base) which are represented by populations in the intention layer (IL). Here the action chains are pre-coded, but in our previous work we have addressed how they may autonomously develop (Erlhagen, Mukovski and Bicho, 2006; Erlhagen et al., 2007). This chained organization of motor primitives is motivated by recent findings of specific neural populations in the inferior parietal lobe of monkey which is known to be part of the matching system in monkey. Fogassi and colleagues (2005) described neurons that fire at the time of a specific motor act (e.g., grasping) in dependence of the ultimate goal of the action sequence in which the act is embedded (e.g., grasping for placing versus grasping for eating). If a chain may become activated by the mere observation of the first act of the chain, the observer is able to predict future motor behaviour and the consequences of the whole action sequence before its execution, i.e. the co-actor's motor intention. However, since a single motor act may be part of several chains, or may not be directly observable, the integration of additional contextual information is necessary to disambiguate the simulation (Erlhagen et al., 2007). For the assembly task, an important input comes from layer OML representing the memorized world knowledge about the location of the different parts in the two working areas. A second source of information that may sustain the simulation process is the shared task knowledge about what the human partner could do in a particular joint action situation (Sebanz, Knoblich and Bekkering, 2006). The subgoals of the assembly work, that are currently available for the team, are represented by populations in the common subgoal layer (CSGL). They are continuously updated in accordance with the assembly plan based on visual feedback about the state of the construction and the inferred goal of the co-actor (represented in the IL). The input from the IL to the CSGL is of particular importance for pro-active behaviour since an updating of subgoals based on anticipated action outcomes allows the robot to plan ahead of time a complementary behaviour that best serves the user's future needs. The CSGL contains two sublayers each containing populations representing all possible subgoals of the assembly task. The sequential order of task execution is encoded by the connections between populations in the two layers. Input signalling the achievement (or predicted achievement) of a certain subtask activates the respective population representation in the first layer which in turn drives automatically through the connections the populations in the second layer representing the next possible assembly steps (e.g., attaching first a wheel and subsequently fixing it with a nut). The action execution layer (AEL) contains the same goal-directed action sequences as the ASL. The different populations integrate input from the IL, OML and CSGL to select among all possible actions the most appropriate complementary behaviour.

The implemented context-sensitive mapping from observed actions on to-be executed complementary actions guarantees a fluent team performance if no errors occur (Bekkering et al., 2009). To cope in an efficient manner also with unexpected or erroneous behaviour of the co-actor, populations in the error monitoring layer (EML) are sensitive to a mismatch on the goal level (integrating input from CSGL and IL, e.g., the co-actor reaches a part that has to be attached only later) or on the level of action means to achieve a valid sub-goal (integrating input from OML and ASL, e.g., the co-actor requests a certain part versus reaching the part directly in his/her workspace). Through direct connections to the AEL, population activity in the EML may bias the robot's planning and decision process by inhibiting the representations of complementary actions normally linked to the inferred goal and exciting the representations of a corrective response. Importantly, to efficiently communicate detected errors to the human partner a corrective response may consist of a manual gesture like pointing or a verbal comment to attract the co-actor's attention (Bicho, Louro and Erlhagen, 2010).

4. Basic concepts of the dynamic field framework

The dynamics of each population in the various layers of the control architecture is governed by a neural field equation (Wilson and Cowan, 1973; Amari, 1977). Dynamic neural field (DNF) architectures implement the idea that task-relevant information is encoded by means of activation patterns of local pools of neurons (Erlhagen and Bicho, 2006). These patterns are initially triggered by input from connected populations and sources external to the network (e.g., vision system, speech input). They may become self-stabilized in the absence of any external input due to the recurrent interactions within the population. Figure 3 shows an example of a self-sustained activity pattern (dashed-dotted line) representing a grasping behaviour. Importantly, there exists an instability of the field dynamics. The self-stabilized pattern coexists with a stable homogenous activation distribution (solid line) that represents the absence of specific information about the motor primitive (resting level). Only sufficiently strong input may activate the self-sustaining forces within the population. Weaker external stimuli lead to a subthreshold, input-driven activation pattern (dashed line). This preshaping of local populations by relative weak input signals may nevertheless play an important role for the processing in the joint action circuit. It brings populations closer to the threshold for triggering the self-sustaining interactions and thus biases the decision processes linked to behaviour (Erlhagen and Schöner, 2002).

----------------Insert Figure 3 around here -------------

We employed a particular form of a DNF first analyzed by Amari (1977). In each model layer *i* the activity $u_i(x,t)$ at time *t* of a neuron at field location *x* is described by the following integro-differential equation (for a discussion of analytical results see Erlhagen and Bicho, 2006):

$$\tau_i \frac{\delta u_i(x,t)}{\delta t} = -u_i(x,t) + S_i(x,t) + \int w_i(x-x') f_i\big(u(x',t)\big)dx' + h_i \qquad (1)$$

where the constants $\tau_i > 0$ and $h_i < 0$ define the time scale and the resting level of the field dynamics, respectively. The integral term describes the intra-field interactions. It is assumed that (1) the interaction strength, $w_i(x,x')$, between any two neurons *x* and *x'* depends only on the distance between locations, and that (2) nearby cells excite each other, whereas separated pairs of cells have a mutually inhibitory influence. For the present implementation we used the following integral kernel of lateral-inhibition type:

$$w_i(\Delta x) = A_i \exp\left(-\Delta x^2 / (2\sigma_i^{\ 2})\right) - w_{inhib,i} \qquad (2)$$

where $w_{inhib,i}>0$ is a constant and $A_i>0$ and $\sigma_i>0$ describe the amplitude and the standard deviation of a Gaussian, respectively. Only sufficiently activated neurons contribute to interaction. The threshold function $f_i(u_i)$ is chosen of sigmoidal shape with slope parameter β and threshold u_0:

$$f_i(u_i) = \frac{1}{1+\exp\left(-\beta(u_i - u_o)\right)}. \tag{3}$$

Normally, summed input from several connected populations is necessary to create a self-stabilized activity pattern in a target population. Each connected population contributes a Gaussian input signal of certain strength whenever its activity level is above threshold. The total input from all connected populations in the various layers of the dynamic field architecture layer u_i can thus be mathematically described by

$$S_l(x,t) = K\sum_m \sum_j a_{mj} c_{lj}(t) \exp\left(-\left(x - x_m\right)^2 / \left(2\sigma^2\right)\right) \tag{4}$$

where c_{lj}(t) is a function that signals the existence or evolution of a self-sustained activation pattern in subpopulation j in layer u_l, and a_{mj} is the inter-field synaptic connection between subpopulation j in u_l to subpopulation m in u_i. The parameter K scales the total input to the target population relative to the threshold for triggering a self-sustained pattern. This guarantees that the inter-field coupling is weak and the field dynamics is dominated by the recurrent interactions.

The existence of a single, self-stabilized pattern of activation in a dynamic field can not only be used to implement a working memory function but is also closely linked to decision making. In layers ASL, IL, AEL and AML subpopulations encoding different action chains (ASL), goals (IL), complementary actions (AEL) and detected errors (EML), respectively, interact through lateral inhibition. These inhibitory interactions lead to the suppression of activity below resting level in competing neural pools whenever a certain subpopulation becomes activated above threshold. Figure 4 shows an example of the temporal evolution of activity in a field encoding different actions. The population for which the summed input from connected populations is highest wins the competition process. Note that at the beginning all subpopulations appear to be activated to some extent, that is, all action alternatives receive input from suprathreshold activity in connected pools. At time $t=0$ an additional input to the population encoding action A2 drives the activity beyond the critical level for a self-stabilized pattern.

----------------Insert Figure 4 around here -------------

To represent and memorize simultaneously (1) the location of several objects, and (2) multiple common subgoals, the spatial ranges of the lateral interactions in layers OML and CSGL were adapted to avoid a direct competition between different populations. The updating of the memorized information is performed by defining a proper dynamics for the inhibition parameter, h_i<0, of the population dynamics (Bicho, Mallet and Schöner, 2000). A sufficiently large global inhibition destabilizes an existing activity peak. As a consequence, the population activity decays back to the stable resting state.

5. Results

In the following we validate the dynamic field architecture by presenting snapshots of the human-robot interactions in the assembly task. The examples illustrate the impact of action observation on decision making in varying context from the perspective of the robot. The focus is on showing and explaining the goal inference, error detection and anticipatory action selection capacities. In all examples here reported the sequential order of sub-tasks for the construction of the 'toy vehicle' is the following: First, mount wheels and fix them with nuts; second, insert column 1; third, insert column 2 and column 4; fourth, insert column 3, and finally mount the top floor.

The videos of the human-robot interaction and the associated dynamics of the fields can be found at http://dei-s1.dei.uminho.pt/pessoas/estela/JASTvideosBookIS.htm.

Alternatively, the videos may be seen at http://www.youtube.com/watch?v=A0qemfXnWiE (video 1) and http://www.youtube.com/watch?v=7t5DLgH4DeQ (video 2).

5.1 Pro-active behavior and goal inference based on an anticipatory model of action observation

----------------Insert Figure 5 around here -------------

----------------Insert Figure 6 around here -------------

----------------Insert Figure 7 around here -------------

----------------Insert Figure 8 around here -------------

----------------Insert Figure 9 around here -------------

An important prerequisite for successful and fluent interaction is that both team members must be committed to the fulfillment of the joint task. For the robot this means that it should be able to initiate the assembly work and take initiative whenever the human partner is taking too long to act. Moreover, the robot should be able to show its commitment to the team by selecting an action in anticipation of the consequences of the co-actor's action. Anticipation of action effects is possible since the robot simulates the co-actor's motor intentions using its own knowledge about goal-directed action sequences in the assembly task. The two capacities are tested in the experiment illustrated in Figure 5. The robot's camera view shows that all components are distributed on the table (Fig. 6, panel B). The activity patterns in the OML represent the knowledge about the corresponding distribution of the components in the two working areas (Fig. 6, panels A and C). The robot takes the initiative to start the assembly work while the co-actor is still reading the instructions and thus does not show any object-directed action. Since the robot has no wheel within its reach, it decides to request a wheel to mount it on its side of the platform (Fig.5, snapshots S1-S2). This decision is possible because the information about the available subgoals (see the CSGL, Fig.7) and the location of parts in the two working areas (see the OML, Fig.6) creates sufficient input to the AEL to trigger a self-stabilized activation peak centered at the action 'request wheel' (see Panel B in Fig.8, time interval T1-T2).

The user then grasps a wheel. However, her intention is not to transfer it to the robot but to mount it on her side. As can be seen, at the moment of grasping the wheel (Fig.5, snapshots S2-S3) the robot is able to anticipate the partner's motor intention and immediately prepares for holding the base in order to help the user while she inserts the wheel on the axle (Fig.5, snapshots S4-S5). The capacity to infer the goal of the user at the time of grasping is possible because of the way in which the partner grasps an object conveys information about what she intends do with it. The robot has sequences of motor primitives in its motor repertoire that associate the type of grasping with specific final goals. A grasping from above is used to attach a wheel to the axle whereas using a side grip is the most comfortable and secure way to hand the wheel over to the co-actor. The observation of an above grip (represented in the AOL) together with information about the currently active subgoal(s) (attach wheel on the user's side) trigger an activation peak in ASL that represents the simulation of the corresponding 'reaching-grasping-inserting' chain (see Panel A in Fig.8, time interval T2-T3), which automatically activates the underlying goal in the intention layer (see Fig.9, T2-T3). The evolving activation pattern in the AEL (panel B, Fig.8, T2-T3) reflects the decision to stabilize the base. The robot's decision to give up its own intention to attach a wheel is a result of slight differences in the connection strengths between representations in the CSGL and the AEL. These differences favor the realization of the user's subtasks over the subtasks that are under the control of the robot.

When the user has attached the wheel the corresponding activation peak in CSGL disappears and an activation pattern representing the subsequent subgoal ('insert nut on user's side' to fix the wheel) automatically evolves (see Fig.7, T3-T4). The second subgoal 'insert wheel on robot's side' remains active.

The user again takes long to act, this time because she is interrupted by a colleague entering the room (see video 1). The robot again takes the initiative and demands a wheel (Fig.5, snapshots S6-S7) since as before the information represented in the CSGL and the OML is sufficient to trigger the corresponding action representation in the AEL (Panel B in Fig.8, T3-T4). Next, the user grasps a wheel with a 'side grip' and the robot anticipates that she is going to hand it over (see ASL in Panel A in Fig.8, T4-T5). The robot prepares to receive the wheel in order to mount it on its side of the platform (see snapshots S7-S9 in Fig.5, and the AEL activation in Fig.8, T4-T5).

5.2 Understanding partially occluded actions and anticipating the user's future needs

----------------Insert Figure 10 around here -------------

----------------Insert Figure 11 around here -------------

----------------Insert Figure 12 around here -------------

----------------Insert Figure 13 around here -------------

In the previous example we have seen that the robot could infer through motor simulation the co-actor's motor intention from the way the object is grasped. But what happens when the robot cannot directly observe the hand-object interaction? In natural environments with multiple objects and occluding surfaces this is a realistic scenario. The capacity to discern the user's motor intention and to select an appropriate complementary behavior should of course not be disrupted by missing information about the grip type used. Information about the context in which the action is executed may sustain the motor simulation process. This is illustrated in the following interaction scenario in which only the 'reaching' part of the user's action sequence can be observed. The robot sees the hand disappearing behind an occluding surface but knows that there is a wheel behind (Fig.10) since the occluder has been introduced into the scene only after the robot could memorize the position of the wheel in the workspace. Figure 11 illustrates the goal inference mechanism in this situation. The AOL (not shown) only codes the reaching behavior. The currently possible subgoals represented in CSGL are ‘insert wheel on user’s side’ and ‘insert wheel on robot’s side’ (Fig.11, Panel A). The inputs from the AOL and the CSGL to the ASL are thus compatible with two competing chain representations and thus should only pre-activate these representations. The additional input necessary for goal inference comes from the information about the location of the wheel in the user’s workspace represented in the OML (see Panel B in Fig. 11). This input triggers the evolution of a self-stabilized activation peak in the ASL representing the action sequence ‘reach wheel-grasp-insert’ (see Panel C in Fig.11). This activation in turn induces a suprathreshold pattern in the IL representing the underlying goal ‘insert wheel’ (see Panel B in Fig.12; see also snapshot S4 in Fig.10). When the activation pattern in IL rises above threshold it initiates a dynamic updating process in the second layer of the CSGL, representing the next possible subgoal(s) for the user (Panel C, Fig.12; see also snapshot S5 in Fig. 10). This representation allows the robot to select a complementary action that serves user’s future needs.

In summary, in this example ‘insert wheel on robot side’ and ‘insert wheel on user’s side’ are two currently available subgoals, the robot infers that the user is grasping a wheel with the intention to mount it. The next possible subgoal for the user is ‘insert nut’ in order to fix the wheel. All the nuts and one wheel are located in the robot’s workspace. As a consequence, three complementary actions in AEL are supported by input from connected populations and thus compete for expression in the robot’s overt behavior (see Fig.13): ‘insert wheel’, ‘give nut’ and ‘hold the base’. As can be seen when comparing the pattern of activation that evolves in the AEL, the robot decides to serve the human by grasping a nut for handing it over (see Fig.10, snapshots S5-S6).

5.3 Error detection and context-sensitive interpretation of a request gesture

----------------Insert Figure 14 around here -------------

----------------Insert Figure 15 around here -------------

----------------Insert Figure 16 around here -------------

Even in known tasks the user can easily become confused and make errors that should be corrected by the co-actor before failure becomes manifested. For example, the user may get confused with the different columns (labeled as C1 to C4) and tries to insert a certain column in the wrong whole or tries to manipulate the columns in a wrong sequential order. The example shown in Fig.14, illustrates two cases in which the robot’s error monitoring layer detects mismatches between the inferred intention of the user and the state of the construction, that is, between the intention and possible subgoals. In the two cases, the robot reports the error to the user and explains what needs to be done. This interaction scenario also allows illustrating the robot's ability to infer the goal of a request gesture.

As illustrated by snapshot S16 (in Fig.14), the robot observes the user grasping column C4 with a top grip which it interprets via action simulation as belonging to a 'grasping to insert sequence' (see Fig.15, time interval T7-T8). However, the activation pattern in the DNF of the CSGL representing present available subgoal(s) indicates a conflict. The correct subtask is to 'insert column C1 on the robot's side' (Panel A in Fig.16, time interval T7-T8). Input from this field together with input from IL (Panel B in Fig.16, time interval T7-T8) triggers a suprathreshold peak of activation in a population encoding the error in user's intention (see Panel C in Fig.16, time interval T7-T8). The error related activity is linked to a subpopulation in AEL that represents a corrective response. In this case, suprathreshold activity initiate speech output to report the mismatch and to explain what needs to be done (see snapshot S17-S18 in Fig.14, and field activity in the AEL depicted in Panel D of Fig.16). The content of the speech combines the information coded in the activation patterns that have initially triggered the error related activity, i.e. intention layer (IL) and possible active sub-goals represented in CSGL, respectively ("*You cannot insert column 4 yet. First, we need to insert column 1*"). Subsequently, the user proceeds by grasping column C1 (snapshot S19, Fig.14) which the robot interprets as belonging to a 'grasp to handover' chain (see Fig. 15, time interval T8-T9). An activation pattern starts to evolve in AEL (Panel D in Fig. 16, time interval T8-T9) that represents the robot's decision to receive column 1 for inserting it in the corresponding whole of the platform (snapshots S20-S21, Fig.14).

Next the user opens up her empty hand as it moves towards the robot (snapshot 22, Fig.14). The robot has this gesture associated with 'request object' in its motor repertoire. The observation of this unspecific gesture activates to some extent all action chains in the ASL linked to requesting the components of the toy vehicle that are in the robot's workspace (in this example nut and column C3). The nuts have already been attached (working memory about already achieved subtask is represented by self-stabilized activation patterns in the "past" layer of the CSGL, not shown). Thus the robot interprets the user's gesture as a request for column C3 (see snapshot S22 in Fig14, and field activity in ASL in Fig.15, time interval T9-T10). The bimodal activation pattern in CSGL (panel A, Fig.16) indicates that the currently available subgoals are inserting columns C2 and C4 on the user's side of the platform. Hence similar to the preceding example, input from this field together with input from IL induces a suprathreshold peak of activation in a population of the error layer representing that 'inserting column C3' is an error in intention (see snapshot S23 in Fig.14, and panel D in Fig.16, in the interval T9-T10). The robot verbally signals the error and explains what sub-goals are currently possible (snapshot S24, Fig.14). Subsequently, the user grasps and inserts column C4 while the robot assists the human user by stabilizing the base.

6. Discussion

We presented a robot control architecture for human-robot collaboration that is inspired by neurocognitive theories about how humans perceive and act in a social context. The results of the validation in a joint assembly task show that the implementation of a human-like joint action model in the robot supports a fluent and flexible task execution. The ease with which humans coordinate in routine joint activity their decisions in space and time is impressive. The capacity to quickly register the co-actor's motor intention before his or her action sequence is completed is essential for a fluent performance in human team activity (Sebanz, Knoblich and Bekkering, 2006). Being able to select an action based on predicted effects of the co-actor's behaviour is thus considered a crucial skill for robots in order to be fully accepted by a human user as a social partner in cooperative tasks (Hoffman and Breazeal , 2007). Converging lines of experimental evidence support the notion of an automatic and obligatory motor simulation process as the underlying mechanism for the intention understanding capacity (for review see Rizzolatti and Craighero, 2004).
The dynamic field architecture implements the idea that goal inference, performance monitoring and action selection occur rather effortlessly and do not require a fully developed human capacity for conscious control (for a review Ferguson and Bargh, 2004). As the representation of context, goals and shared task knowledge are interconnected, the observation of a motor act together with situational cues activates through motor simulation first the self-sustained population representations of the related goal and subsequently the representation of the most appropriate complementary action. As our examples show, this automatic process includes situations in which the human partner acts in an unexpected or inappropriate manner. This view on intention communication and joint action planning contrasts with most robot control architectures that have been tested in the past in similar collaborative tasks (e.g., Alami et al., 2005; Steil et al., 2004; Gast et al., 2009; but see Breazeal, Gray and Berlin, 2009, for a conceptually similar approach). Typically, these architectures include dedicated modules that organize the high-level task of intention coordination between the co-

actors and the intelligent monitoring of the team performance using some form of symbolic manipulation and logic. Although we do not deny that the results of our robotics experiments could be implemented in a symbolic framework we argue here that the additional planning process that would be needed to link the high-level representations to the motor level of the robot's actuators would greatly reduce the effectiveness of those representations.

In the dynamic field architecture the decision process linked to complementary actions unfolds over time under multiple influences which are themselves modelled as dynamic representations with proper time scales. This is the basis of flexible behaviour in dynamic joint action conditions. The absence or delay of information about for instance the co-actor's motor intention will automatically lead to a decision that does not take into account the co-actor (Bicho et al., in press). Conversely, the dynamic updating of the currently available subgoals for the team based on predicted effects of the co-actor's ongoing action allows for anticipatory action planning. If on the other hand the predicted effect is inconsistent with the current goals for the team the obligatory integration of evolving suprathreshold activity in the action monitoring layer may override the planning of a pre-potent complementary action.

Among the many possible types of errors that might occur during joint activity of the human-robot team (for a discussion see Spexard et al., 2008) we have focused in this chapter on the detection and communication of intention errors made by the human partner. The proposed action monitoring mechanism can be easily extended to cope with other types of unexpected events including errors made by the robot. For instance, the co-actor may show a request gesture with the intention to fulfil the valid subgoal of fixing the wheel with a nut. However, since a nut is located in her workspace, responding to the request by transferring the nut would not be an efficient behaviour of the robot. Population activity in the EML that combines the information from the OML about object location and the ASL about the goal of the request gesture should instead trigger for instance a pointing gesture to attract the co-actor's attention to the nut in her workspace. Similarly, population activity in the AML that automatically integrates the information about the goal (represented in CSGL) of a self-performed sequence like reaching-grasping-attaching a wheel and proprioceptive (and/or visual) information about an accidental loss of the wheel during transportation may allow the robot to quickly start searching for a new wheel or to ask the human for assistance. Interestingly, studies investigating the functional system for action monitoring suggest that humans use similar cognitive and neural mechanisms to detect own and observed errors in joint action (Bekkering et al., 2009).

The focus of the present experiments was on implicit communication. The robot had to interpret the co-actors actions and gestures in terms of a goal to select an adequate complementary action. Although we have argued that joint activity in a familiar task often does not require explicit communication it is undeniable that being able to communicate with the robot system by natural speech would greatly facilitate human-robot interactions in many situations (e.g., Spexard et al., 2007; Pardowitz et al., 2007; Gast et al., 2009). Our robot speaks aloud to make its goal inference and action monitoring capacities transparent for the user. The knowledge about the robot's cognitive capacities supports predictability of its behaviour which is essential to an effective collaboration. Using a simplified version of a joint assembly task in which the robot merely assists the human user by handing over pieces (Bicho, Louro and Erlhagen, 2010), we have recently made first steps towards integrating in the DNF-model of joint action the capacity to understand simple action-related speech. The basic idea is that the automatic resonance of motor structures during action observation extends to the language domain. The robot understands sentences like *Give me the wheel* or *I give you the wheel* by covertly activating semantically congruent motor representations that are linked to the specific goal or end-state (e.g. hand opens up as it moves towards the co-actor for a request gesture and a reaching-grasping-holding out sequence for a handing over procedure). This embodied view on language comprehension is supported by findings in a range of recent experimental studies (for review see Fischer and Zwaan, 2008).

7. Conclusions and outlook

The work presented in this chapter wished to contribute to the development of design principles for robots that are supposed to directly collaborate with their human partners in shared tasks. As an exquisitely social species, humans are experts in coordinating actions and decisions with other in order to achieve common goals. We believe that implementing a human-like joint action model in the robot is a promising approach because it will allow the artificial cognitive agent to meet the user's expectations about a pleasant, efficient and successful interaction with a socially intelligent partner. While theories about the neurocognitive mechanisms supporting human joint action have now reached a sufficient level of detail to guide robotics work, it is also clear that these theories contain hypothesis and assumptions for which the experimental evidence is still under debate. Implementing and testing theories and hypothesis about human joint action in an embodied agent with sensory, motor and cognitive capabilities offers in our view unique opportunities for researchers from cognitive science and neuroscience.

The adopted dynamic perspective offers in general a high degree of flexibility and robustness in joint task execution. However, the present implementation of the dynamic field architecture limits the cooperative interactions to the specific assembly task since the neural representations and their connectivity were tailored by the designer. It is thus highly desirable to endow the robot with a developmental program that would allow the artificial agent to autonomously learn and represent new task-relevant representations (Weng, 2004). Learning efficient joint action coordination in a complex task is a very demanding and to a large extent unsolved problem even when starting with a "minimal" set of pre-defined capacities and knowledge. Adopting a socially guided machine learning paradigm in which a human trainer teaches a robot through demonstration and verbal or gestural commands in much the same way as parents teach their children seems to be a promising research direction (Otero et al., 2008; Thomaz and Breazeal, 2008). First experimental results of our attempt to apply a learning dynamics for establishing inter-field connections show the feasibility of the approach. Using correlation-based learning rules (Gerstner and Kistler, 2002) with a gating that signals the success of behaviour, we have shown for instance how goal-directed mappings between action observation and action execution that support an action understanding capacity may develop during learning and practice (Erlhagen, Mukovskiy and Bicho, 2006; Erlhagen et al., 2006). Importantly, the developmental process may explain the emergence of new task-specific populations which have not been introduced to the architecture by the human designer (Erlhagen et al., 2007).

We are currently applying and testing a learning by demonstration approach to systematically address the question of how ARoS may acquire and store the knowledge about the serial order of task execution that was predefined by the designer in the experiments reported here. For the present dynamic field architecture this means to autonomously develop the connections between neural populations in the two sublayers of CSGL representing subsequent steps of the assembly plan. We exploit here the self-stabilizing properties of the field dynamics determined by the recurrent interactions within each population and fixed excitatory and inhibitory connections between neuronal populations in both layers representing the same assembly step or subgoal of the construction plan. During demonstration by a human teacher, ARoS perceives changes in the state of the construction. This visual input triggers a suprathreshold activity pattern of the neural population in the second layer representing the currently achieved subgoal. Mediated by excitatory connections this activity propagates to the population in the first layer and initiates the evolution of a self-sustained activity pattern. This pattern implements a working memory function in the first layer and suppresses through inhibitory feedback connections the population activity in the second layer. Importantly, the actively maintained representation of the achieved subgoal allows the robot to learn associations between subsequent assembly steps that are separated in time. Correlation-based learning takes place whenever a perceived change in the state of construction triggers a transient population representation of a newly achieved subgoal in the second layer. Since the serial order of task execution may not be exactly the same in different demonstrations of the task (e.g., different teachers), association between a single memorized subgoal and several possible next steps may be learned during observation.

The work on learning and development in the dynamic field architecture for joint action represents first steps towards robotics systems that will ultimately be able to autonomously built representations for assisting different human users in a large variety of tasks. We believe that combining the processing principles of neural field dynamics and different machine learning techniques in the context of the socially guided learning paradigm represents a promising research direction towards achieving this demanding goal.

Acknowledgements

The present research was conducted in the context of the fp6-IST2 EU-IP Project JAST (proj. nr. 003747) and partly financed by the FCT grants POCI/V.5/A0119/2005 and CONC-REEQ/17/2001. We would like to thank Emanuel Sousa, Flora Ferreira and Toni Machado for their help during the robotic experiments.

References

Alami, R., Clodic, A., Montreuil, V., Sisbot, E. A. & Chatila, R. (2005). Task planning for human-robot interaction. In: *Proceedings of the 2005 Joint Conference on Smart Objects and Ambient Intelligence,* ACM International Conference Proceeding Series, 121, 81-85.

Allisandrakis, A., Nehaniv, C. L, & Dauthenhahn, K. (2002). Imitation with ALICE: learning to imitate corresponding actions across dissimilar embodiments. *IEEE Transactions on Systems, Man and Cybernetics-Part A,* Syst. Humans, 32, 482-492.

Amari, S. (1977). Dynamics of pattern formation in lateral-inhibitory type neural fields. *Biological Cybernetics*, 27, 77–87.

Bekkering, H., Bruijn, E.R.A. de, Cuijpers, R.H., Newman-Norlund, R., Schie, H.T. van & Meulenbroek, R.G.J. (2009). Joint action: neurocognitive and neural mechanisms supporting human interaction. *Topics in Cognitive Science, 1*, 340-352.

Bicho, E., Mallet, P. & Schöner, G. (2000). Target representation on an autonomous vehicle with low-level sensors. *International Journal of Robotics Research*, 19, 424-447.

Bicho, E., Louro, L., Hipolito, N. & Erlhagen, W. (2009). A dynamic field approach to goal inference and error monitoring for human-robot interaction. In: Dautenhahn, K. (Ed.), Proceedings of the 2009 *International Symposium on New Frontiers in Human-Robot Interaction*. AISB 2009 Convention, Heriot-Watt University Edinburgh, 31–37.

Bicho, E., Louro, L. & Erlhagen, W. (2010). Integrating verbal and nonverbal communication in a dynamic neural field architecture for human-robot interaction. *Frontiers in Neurorobotics*, 4, May, 2010. doi: 10.3389/fnbot.2010.0005.

Bicho, E., Louro, L., Costa e Silva, E. & Erlhagen, W. (in press). Neuro-cognitive mechanisms of decision making in joint action: a Human-robot interaction study. *Human Movement Science*. doi:10.1016/j.humov.2010.08.012

Billard, A, Calinon, S., Dillmann, R. and Schaal, S. (2008). Robot Programming by demonstration. In Handbook of Robotics. B. Siciliano, B. and Kathib, O. (Eds). Springer.

Breazeal, C. (2004). Social interactions in HRI: the robot view. *IEEE Transactions on Systems, Man and Cybernetics-Part C: Applications and Reviews*, 34, 181-186.

Breazeal, C., Gray, J. & Berlin, M. (2009). An embodied cognition approach to mindreading skills for socially intelligent robots. *International Journal of Robotics Research*, 28, 656–680.

Costa e Silva, E., Bicho, E., Erlhagen, W & Meulenbroek, R. (submitted). Human-like collision-free arm movements for human-robot collaboration.

Demiris, Y., Johnson, M. (2003). Distributed, predictive perception of actions: a biologically inspired robotics architecture for imitation and learning. *Connection Science*, 15, 231-243.

Erlhagen, W. & Schöner, G. (2002). Dynamic field theory of movement preparation. *Psychological Review, 109, 545-572*.

Erlhagen, W. & Bicho, E. (2006). The dynamic neural field approach to cognitive robotics. *Journal of Neural Engineering*, 3, R36–R54.

Erlhagen, W., Mukovskiy, A. & Bicho, E. (2006). A dynamic model for action understanding and goal-directed imitation. *Brain Research*, 1083, 174–188.

Erlhagen, W., Mukovskiy, A., Bicho, E., Panin, G., Kiss, C., Knoll, A., van Schie, H. & Bekkering, H. (2006). Goal-directed imitation for robots: a bio-inspired approach to action understanding and skill learning. *Robotics and Autonomous Systems*, 54, 353–360.

Erlhagen, W., Mukovskiy, A., Chersi, F. & Bicho, E. (2007). On the development of intention understanding for joint action tasks. In: *Proceedings of the 6th IEEE Int. Conf. on Development and Learning*. Imperial College London, pp. 140–145.

Ferguson, M. J. & Bargh, J. A. (2004). How social perception can automatically influence behaviour. *Trends in Cognitive Sciences, 8, 33-39*

Fischer, M. & Zwaan, R. (2008). Embodied Language: A Review of the Role of the Motor System in Language Comprehension. *The Quarterly Journal of Experimental Psychology, 61*, 825-850.

Fogassi, L., Ferrari, P. F., Gesierich, B., Rozzi, S., Chesri, F. & Rizzolatti, G. (2005). Parietal lobe: from action organization to intention understanding. *Science*, 308, 662-667.

Fong, T., Nourbakhsh, I. & Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and Autonomous Systems*, 42, 143–166.

Gast, J., Bannat, A., Rehrl, T., Wallhoff, F., Rigoll, G., Wendt, C., Schmidt, S., Popp, M. & Farber, B. (2009). Real-time framework for multimodal human–robot interaction. In: *Proceedings of the* 2^{nd} *Conference on Human*

System Interactions 2009, HIS ’09. IEEE Computer Society, Catanid, Italy, 276–283.

Gerstner, W. & Kistler, W. M. (2002). Mathematical formulations of Hebbian learning. Biological Cybernetics, 87, 404-415.

Hoffman, G. & Breazeal, C. (2007). Cost-based anticipatory action selection for human-robot fluency. *IEEE Transactions on Robotics*, 23, 952–961.

Otero, N. Saunders, J., Dautenhahn, K. & Nehaniv, C. L. (2008). Teaching Robot Companions: The Role of Scaffolding and Event Structuring. *Connection Science*, 20, 111-134.

Pardowitz, M., Knopp, S., Dillmann, R. & Zöllner, R. D. (2007). Incremental learning of tasks from user demonstrations, past experiences and vocal comments. *IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics*, 37, 322-332.

Poljac, E., van Schie H.T. & Bekkering, H. (2009). Understanding the flexibility of action-perception coupling. *Psychological Research, 73, 578-586.*

Rizzolatti, G. & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience,* 27, 169–192.

Schaal, A. (1999). Is imitation learning the route to humanoid robots?. *Trends in Cognitive Science*, 3, 233-242.

Schöner, G. (2008). Dynamical systems approaches to cognition. In: Sun, R. (Ed.). *The Cambridge Handbook of Computational Psychology*. Cambridge University Press, pp. 101–125.

Sebanz, N., Bekkering, H. & Knoblich, G. (2006). Joint action: bodies and minds moving together. *Trends in Cognitive Sciences*, 10, 70–76.

Silva, R., Bicho, E. & Erlhagen, W. (2008). ARoS: an anthropomorphic robot for human-robot interaction and coordination studies. In: *Proceedings of the 8th Portuguese Conference on Automatic Control (Control 2008)*. 21-23 July, Vila Real, Portugal, 819–826.

Spexard, T. P., Hanheide, M., & Sagerer, G. (2007). Human-oriented interaction with an anthropomorphic robot. *IEEE Transactions on Robotics*, 23, 852-862.

Spexard, T.P., Hanheide, M., Li, S. & Wrede, B. (2008). Oops, something is wrong: Error detection and recovery for advanced human-robot interactions. In: *Proceedings of the ICRA 2008 Workshop on Social Interaction with Intelligent Indoor Robots*. Kruijff, G.J., Zender, H. Handeide, M. & Wrede, B. (Eds).

Steil, J. J., Röthling, F., Haschke, R., & Ritter, H. (2004). Situated robot learning for multi-modal instruction and imitation of grasping. *Robotics and Autonomous Systems*, 47, 129-141.

Thomaz, A. L., & Breazeal, C. (2008). Teachable robots: Understanding human teaching behaviour to built more effective robot learners. *Artificial Intelligence, 172, 716-737.*

van Schie, H. T., van Waterschoot, B. M. & Bekkering, H. (2008). Understanding action beyond imitation: Reversed compatibility effects of action observation in imitation and joint action. *Journal of Experimental Psychology: Human Perception and Performance,* 34, 1493–1500.

Weng, J. (2004). Developmental robotics: Theory and experiments. *International Journal of Humanoid Robotics, 1, 199-236.*

Westphal, G., von der Malsburg, C. & Würtz, R. P. (2008). Feature-driven emergence of model graphs for object recognition and categorization. In: Kandel, A., Bunke, H., Last, M. (Eds.). *Applied Pattern Recognition*. Springer Verlag, pp. 155–199.

Wilson, H. R., Cowan, J. D. (1973). A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue. *Kybernetik*, 13, 55–80.

Wilson, M. & Knoblich, G. (2005). The case for motor involvement in perceiving co-specifics. *Psychological Bulletin*, 131, 460–473.

List of Figures & captions

(a) (b)

Figure 1: Human-robot team and scenario for the joint construction task. The team has to collaborate in the construction of a ‘toy vehicle’ (shown in panel b) from components that are initially distributed on a table (panel a).

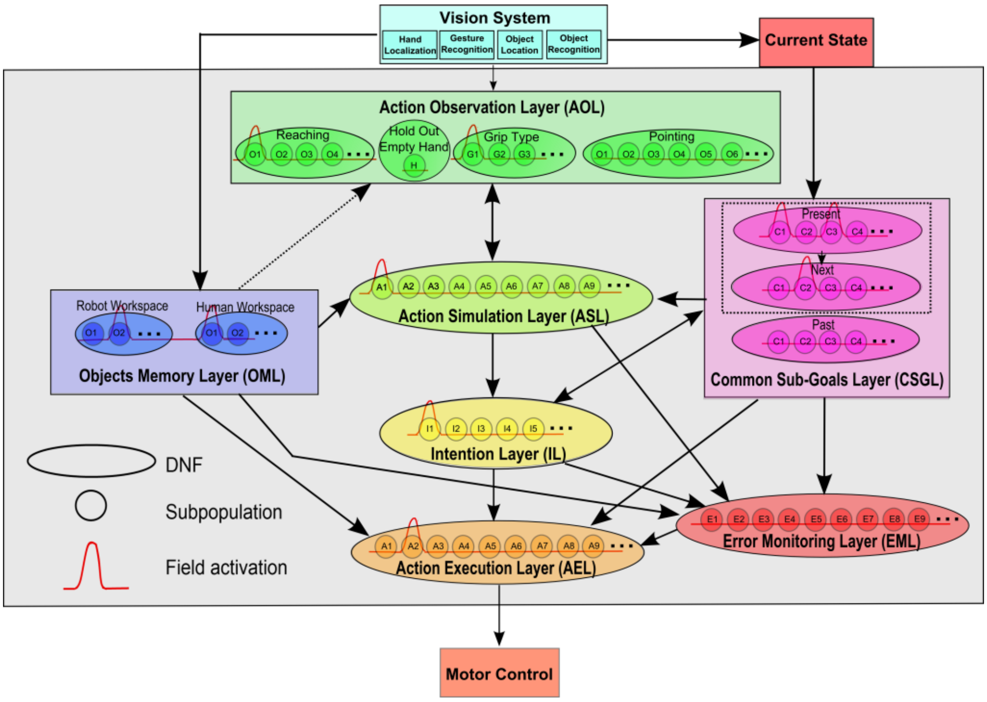

Figure 2: Schematic view of the cognitive architecture for joint action. It implements a flexible mapping from observed actions (layer AOL) onto complementary actions (layer AEL) taking into account the inferred action goal of partner (layer IL), detected errors (layer EML), contextual cues (OML) and shared task knowledge (CSGL). The goal inference capacity is based on motor simulation (layer ASL).

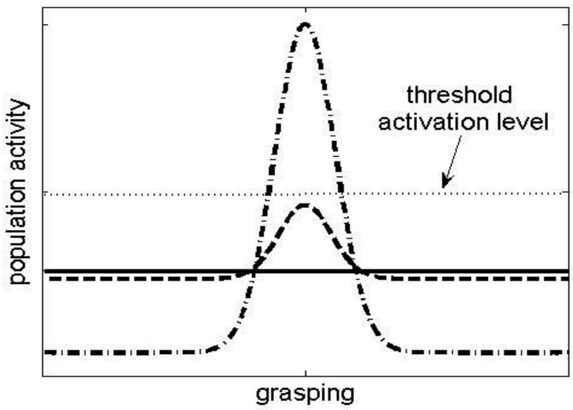

Figure 3: The activity of a neural population is shown that represents through a self-stabilized activation peak the presence of information about a grasping behaviour (dashed-dotted line), while a flat, low-level distribution

of activity (solid line) indicates that information about this motor primitive is currently not processed. In response to a weak input the population generates an activation pattern with amplitude below the threshold necessary to trigger an interaction-dominated activation peak (dashed line).

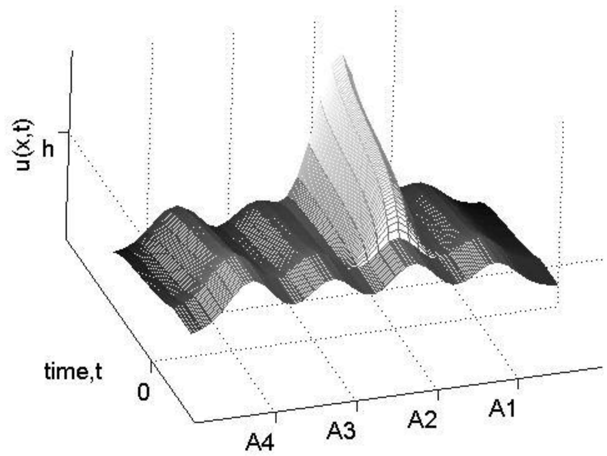

Figure 4: Decision making in a field representing different actions, A1 to A4. The decision is triggered by a sufficiently strong input at time t=0 to population A2. Note that at this time all 4 populations are activated below the critical value for a self-stabilized pattern.

Figure 5: Snapshots, S1-S9, in the time interval T1-T5 of video 1 (long interaction scenario) illustrate the robot's pro-active behavior and its goal inference capacity which is based on an anticipatory model of action observation.

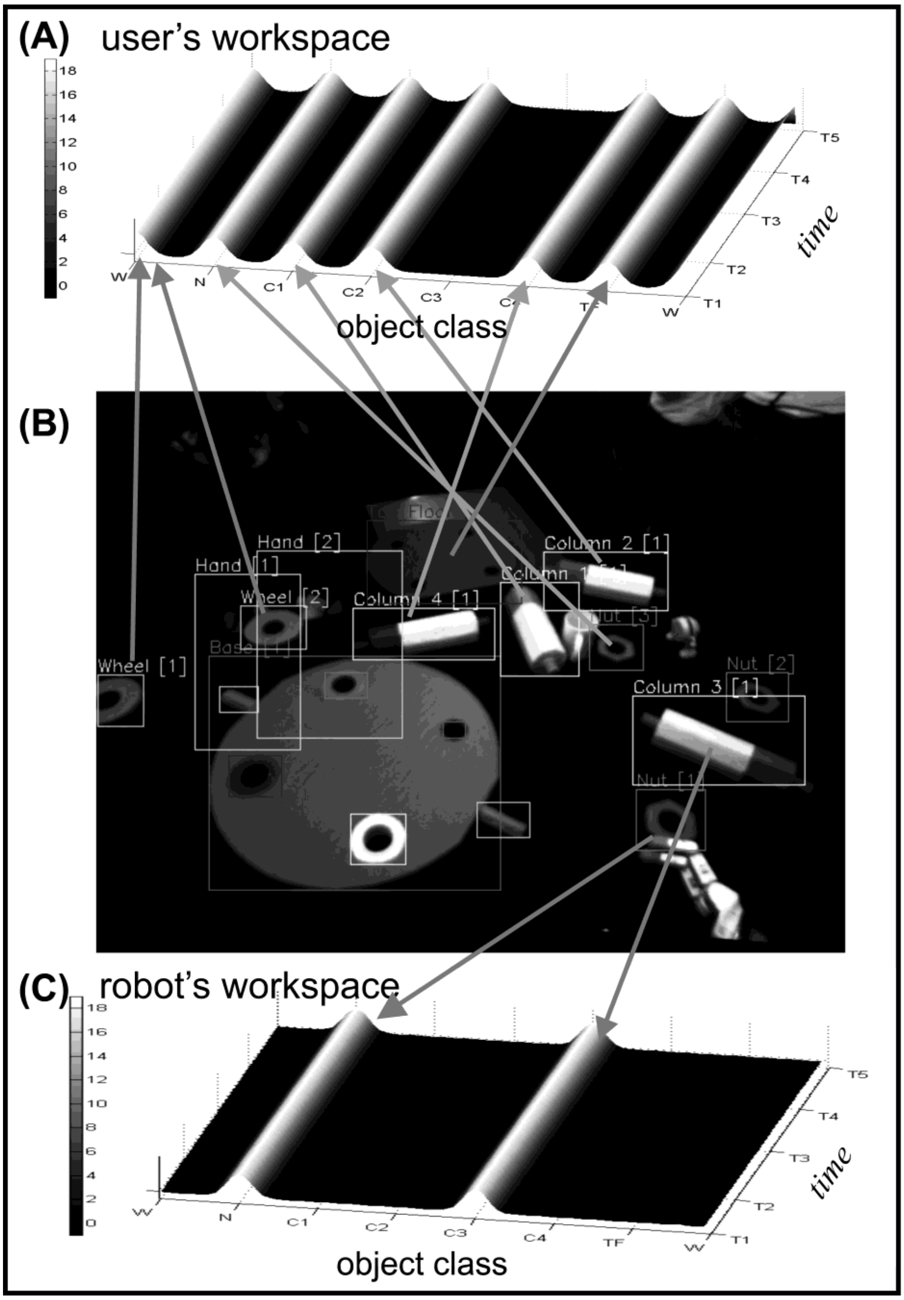

Figure 6: Snapshot of the robot's vision system (Panel B) and corresponding representations of objects in the OML. It contains two fields, one for each workspace (i.e. Panels A and C). In each field the presence of an object of a particular class is represented by a peak of activation.

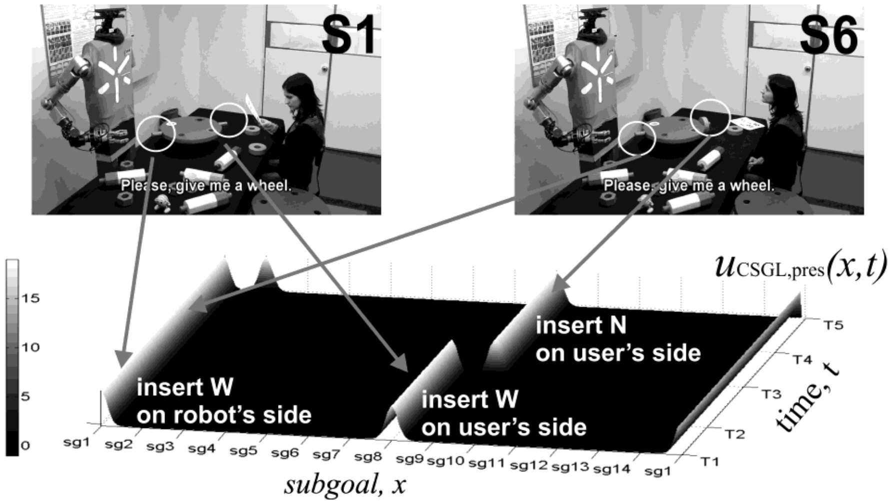

Figure 7: The temporal evolution of activity in the dynamic field encoding the currently available subgoals (in the CSGL) is shown in the interval T1-T5 (video 1). The snapshots S1 and S6 show the corresponding events of the human-robot interactions.

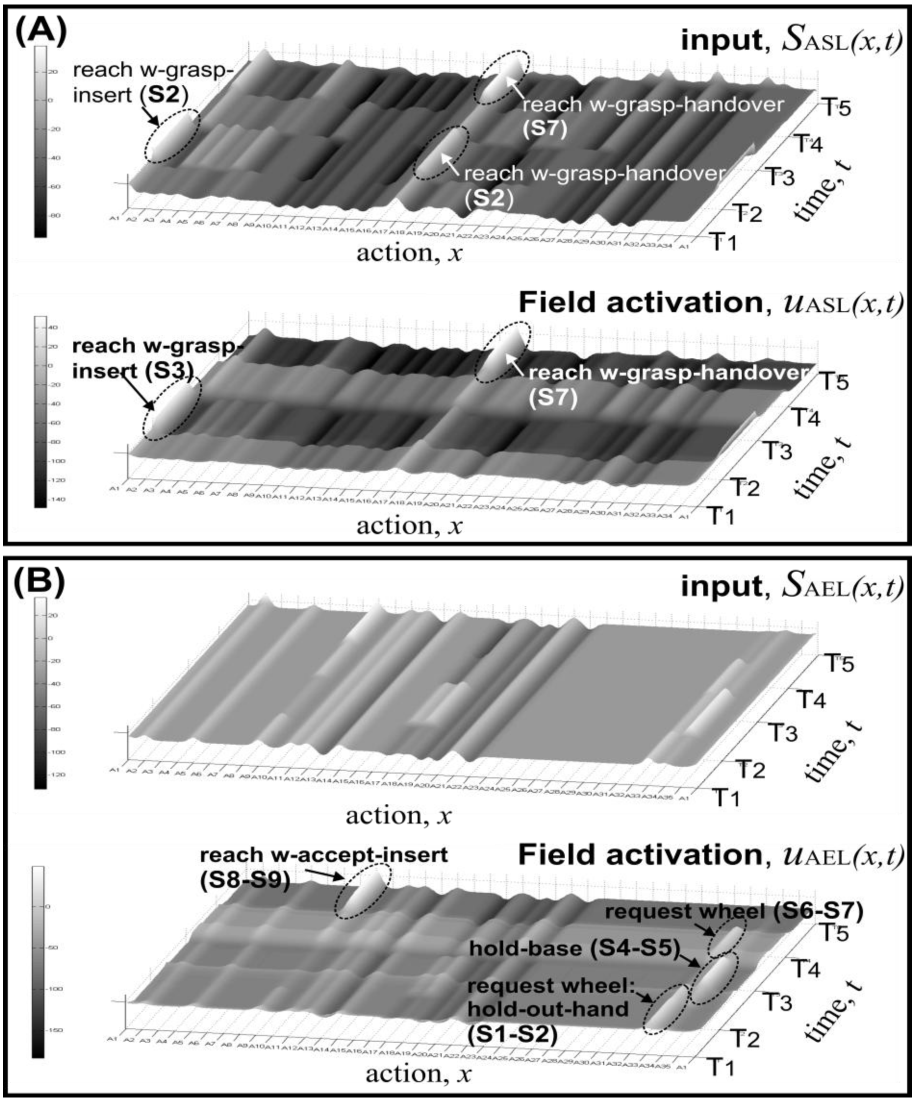

Figure 8: Proactive behavior and goal inference based on an anticipatory model of action observation (video 1, time interval T1-T5). (A) Temporal evolutions of input to ASL (top) and field activity in ASL (bottom). (B) Temporal evolutions of input to AEL (top) and field activity in AEL (bottom)

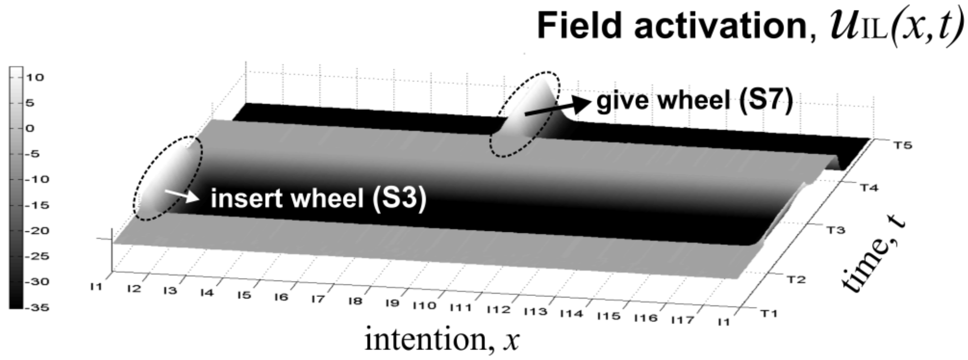

Figure 9: Temporal evolution of field activity in the intention layer (IL) during time interval T1-T5 (video 1).

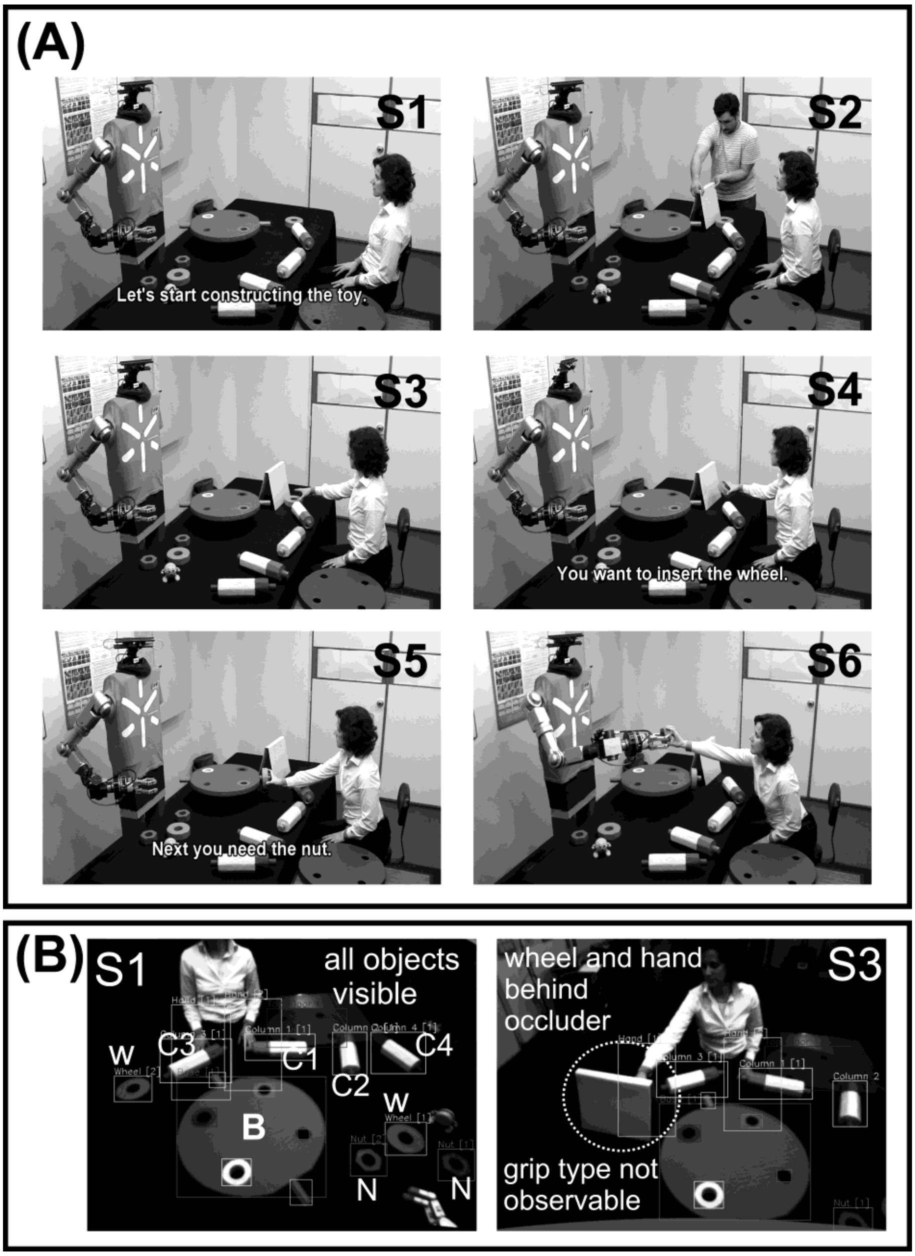

Figure 10: Video 2: (A) Video snapshots showing action understanding of partially occluded actions (S1-S4) and anticipation of the user's future needs (S4-S6). (B) Snapshots of the robot's vision system.

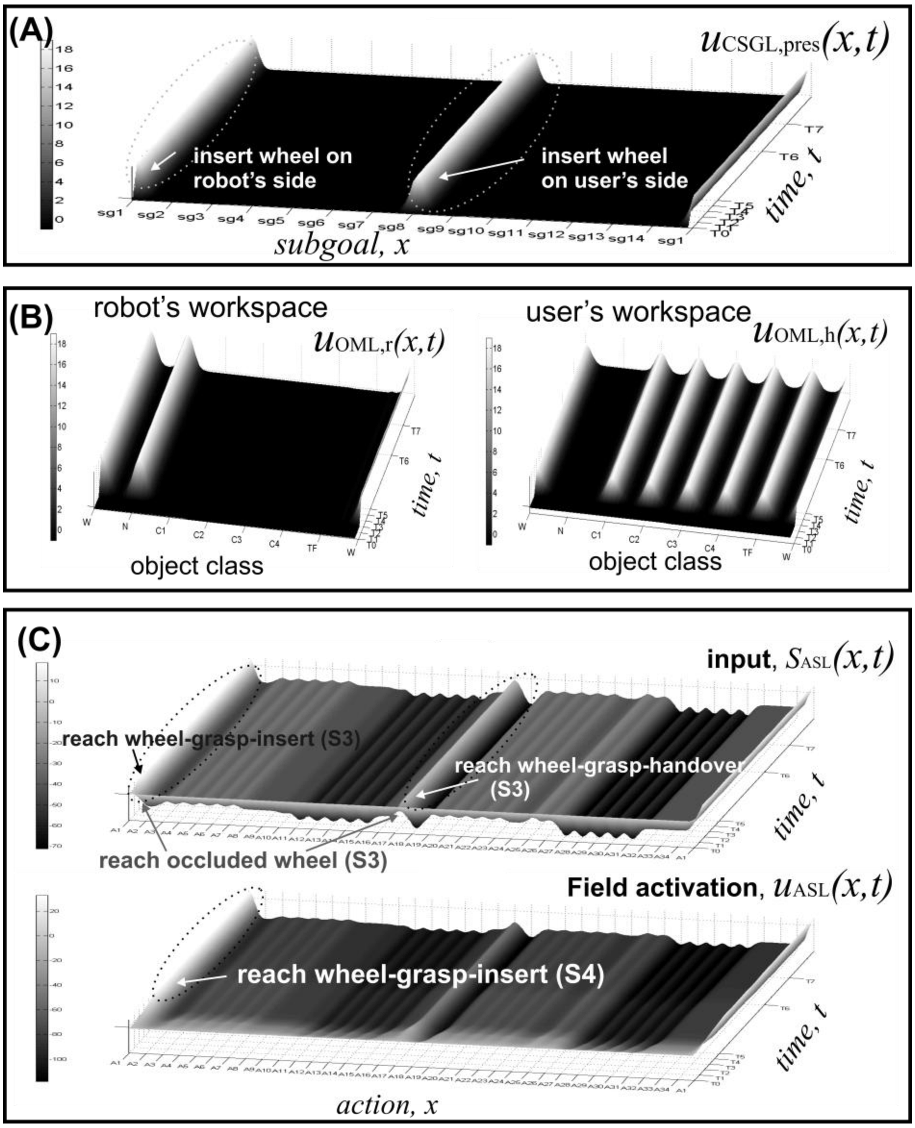

Figure 11: Video 2: (A) Temporal evolution of field activity representing present possible subgoals (in the CSGL). (B) Temporal evolutions of field activity in the OML. (C) Temporal evolutions of input to the ASL (top) and the field activity of the ASL (bottom).

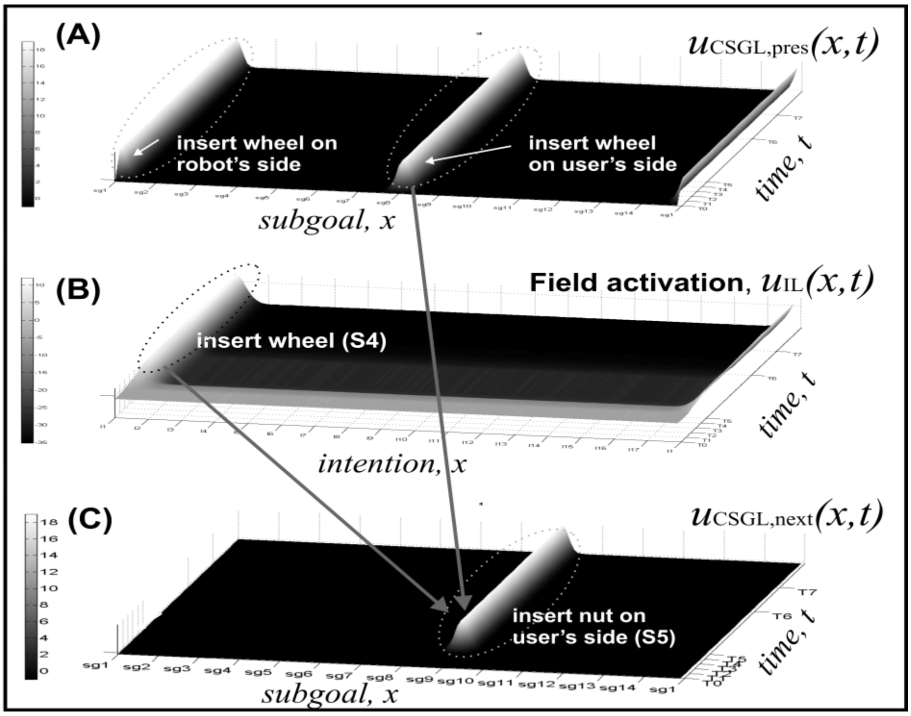

Figure 12: Updating of the dynamic field representing subsequent subgoals for the user based on a prediction of his/her current motor intention. Input from the dynamic field encoding the current subgoals (A) and input from IL (B) induces suprathreshold peak(s) of activation in the field encoding subsequent assembly steps (C).

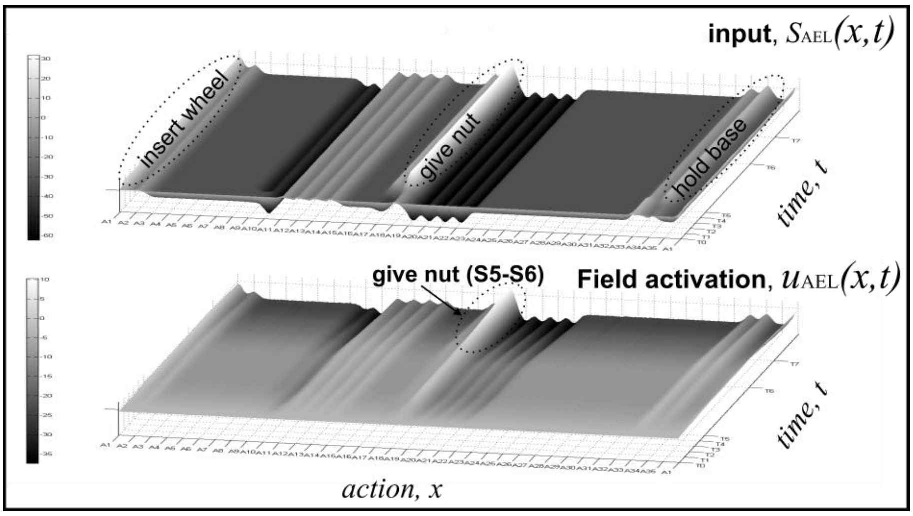

Figure 13: Video 2: Anticipatory action selection. The temporal evolution of total input to AEL (top) and field activity in AEL (bottom) are shown. The robot decides to transfer a nut to the user (´give nut` action, snapshots S5-S6, Fig.10).

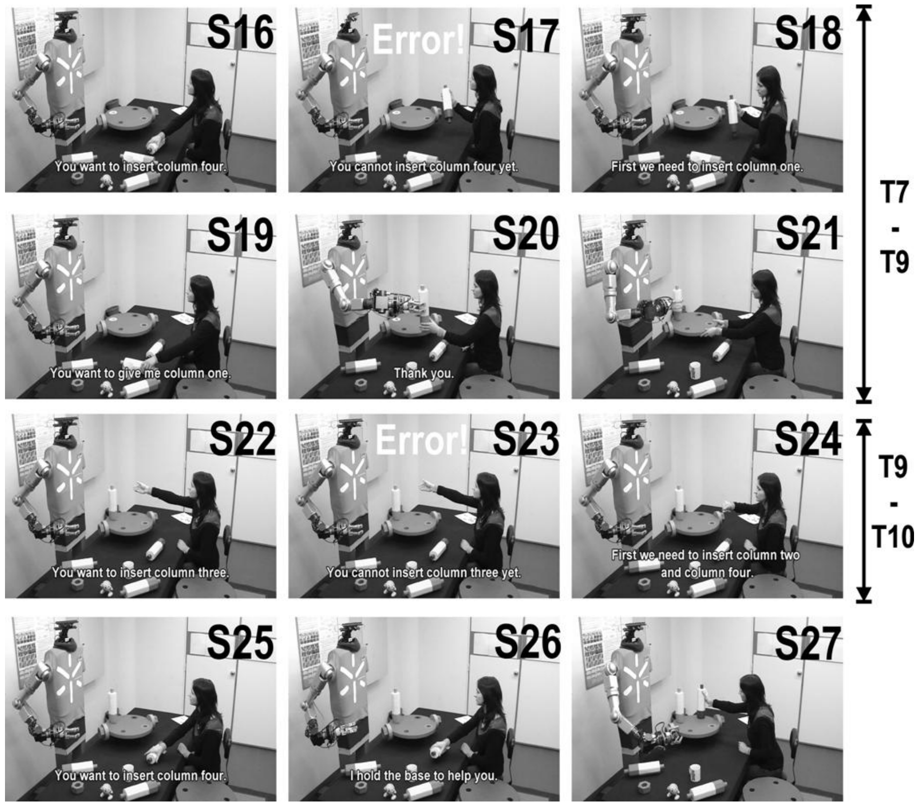

Figure 14: Snapshots of Video 1 (long interaction scenario) in the time interval T7-T11 are shown. They illustrate the robot's capacity to detect and correct the user's intention errors and interpret a request gesture in a context-sensitive manner (see the text for details).

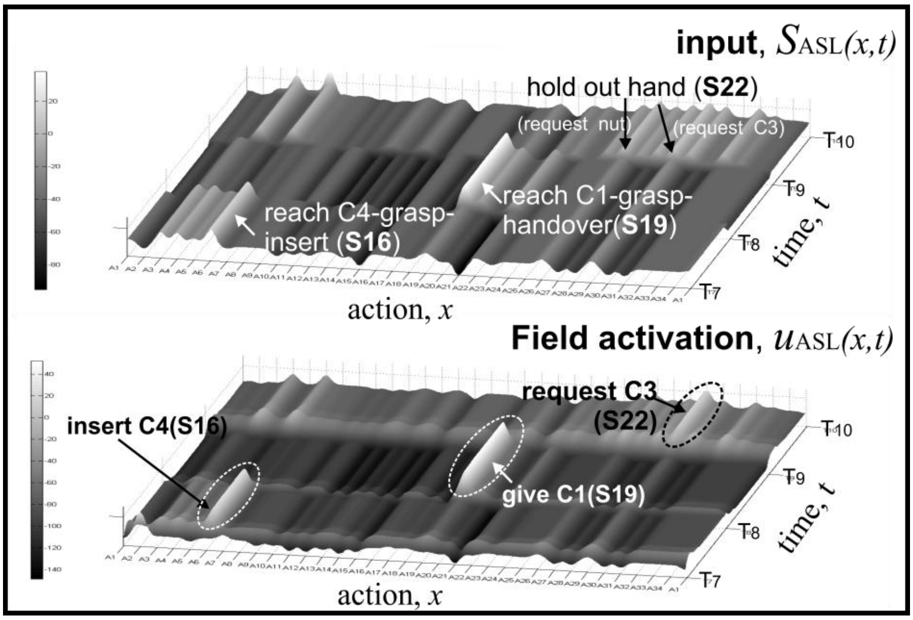

Figure 15: Error detection and context-sensitive interpretation of a request gesture (video 1, time interval T1-T10). The temporal evolution of input to the ASL (top) and the field activity in the ASL (bottom) are shown.

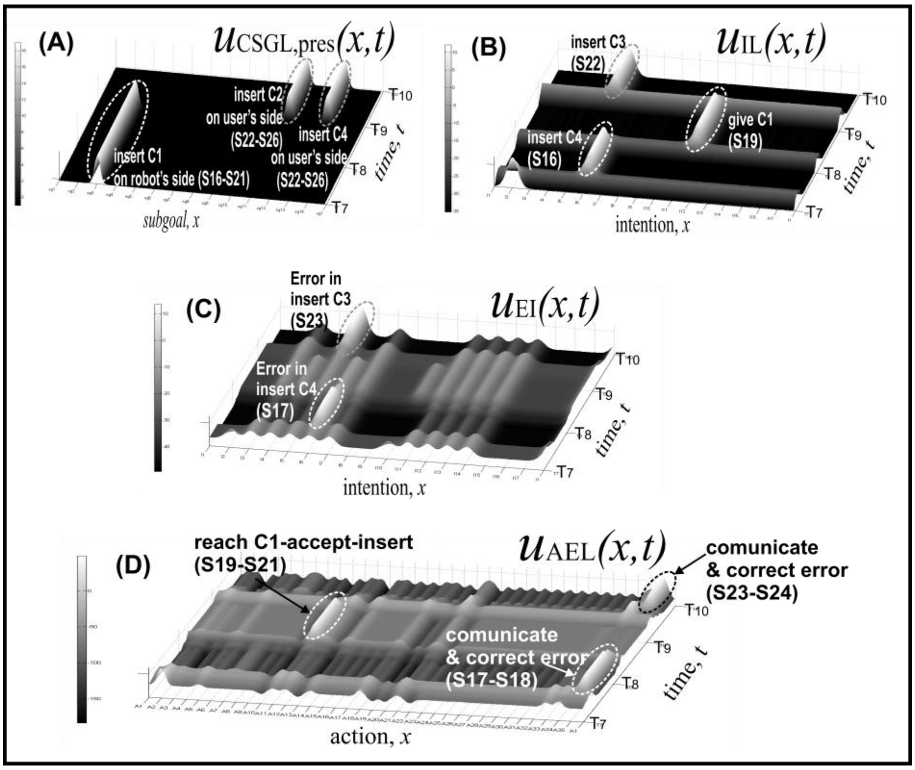

Figure 16: The temporal evolution of activity in different fields is shown for the Video 1 (long interaction scenario) in the time interval T7-T10. (A) Field of the CSGL representing currently available subgoals, (B) IL, (C) EI encoding the errors in intention, (D) AEL.