Interactive, Collaborative Robots: Challenges and Opportunities

Danica Kragic[1], **Joakim Gustafson**[2], **Hakan Karaoguz**[1], **Patric Jensfelt**[1] and **Robert Krug**[1]
[1] Robotics, Perception and Learning lab, KTH Royal Institute of Technology, Stockholm, Sweden
[2] Speech, Music and Hearing department, KTH Royal Institute of Technology, Stockholm, Sweden
dani@kth.se, jocke@speech.kth.se, hkarao@kth.se, patric@kth.se, rkrug@kth.se

Abstract

Robotic technology has transformed manufacturing industry ever since the first industrial robot was put in use in the beginning of the 60s. The challenge of developing flexible solutions where production lines can be quickly re-planned, adapted and structured for new or slightly changed products is still an important open problem. Industrial robots today are still largely preprogrammed for their tasks, not able to detect errors in their own performance or to robustly interact with a complex environment and a human worker. The challenges are even more serious when it comes to various types of service robots. Full robot autonomy, including natural interaction, learning from and with human, safe and flexible performance for challenging tasks in unstructured environments will remain out of reach for the foreseeable future. In the envisioned future factory setups, home and office environments, humans and robots will share the same workspace and perform different object manipulation tasks in a collaborative manner. We discuss some of the major challenges of developing such systems and provide examples of the current state of the art.

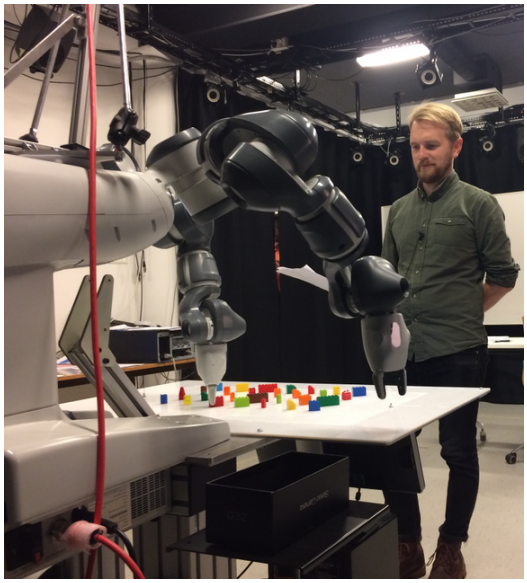


Figure 1: *Interactive Lego picking setup*: The robotic setup for our use-case is centered around a dual-arm ABB YuMi manipulator that is attached to a static table. The manipulator offers advanced force control to ensure safety during human-robot collaboration tasks. Additionally, the platform is equipped with a Kinect structured-light sensor for perceiving objects in the workspace.

1 Introduction

Industrial robots today are still largely preprogrammed for their tasks and most of the commercial robot applications have a very limited ability to interact and physically engage with humans. Full robot autonomy for challenging tasks in unstructured environments will remain out of reach for the foreseeable future. Therefore, development of various aspects of collaborative robot systems where a human can take over parts of the task that are too hard for a robot to do is of great interest. Consequently, in the envisioned future factory setups, humans and robots will share the same workspace and perform different object manipulation tasks in a collaborative manner.

Classical robot programming requires an experienced programmer and the amount of work needed is infeasible for rapidly changing tasks. Programming robots through human demonstration has been promoted as a flexible framework that reduces the complexity of programming robot tasks and allows end-users to control robots in a natural and easy way without the need for explicit programming. The traditional approach is to let the robot be a passive agent in the interaction while the human agent controls the motion of the object. For human-robot collaboration to become as efficient as human-human collaboration, a robot must be able to perform both the active and passive parts of the interaction, just as a human would. We presented an example of this for the task of human-augmented mapping in [Kruijff *et al.*, 2006]. However, when it comes to physical interaction and collaboration in a assembly or a service robot setting, the existing work is still rather limited.

For the robot to take the active part in the interaction, and to be able to plan and execute appropriate trajectories of objects and its own motion, it must have knowledge about the partnering agent, its internal state and what constraints the human imposes on the object. Unlike in the conventional

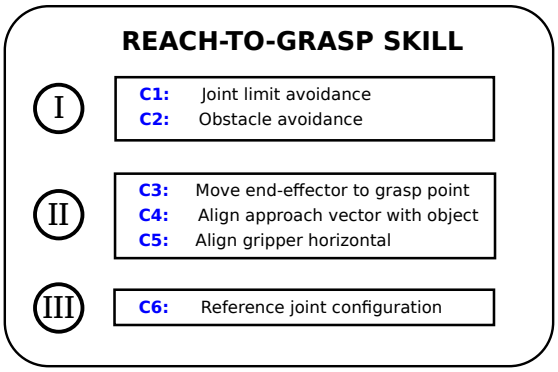


Figure 2: *Skill example:* Reach-to-grasp behavior can emerge from 6 concurrent controllers (C1 . . . C6) running on 3 different hierarchy levels (I-III).

program development process, when programming robots through demonstration the user may not be familiar with the syntax and semantics of the programming language. Thus, we need methodologies for learning and encoding tasks from multiple demonstrations and base learning on both explicit communication (natural language) as well as implicit communication (motion).

However, there are still a number of challenges to tackle: interaction modalities such as speech, gaze or gestures need to be grounded in a formalism that can be interpreted by the robots control algorithms and disambiguities need to be resolved; motion planning and control has to be performed in (or near) real-time to ensure a natural workflow and interaction. This is something that cannot be done with classical sense-plan-act architectures which form the current state-of-the art in applied robotics. Furthermore, robotic manipulation tasks require models capturing the interaction dynamics. These cannot be acquired by sparse human demonstrations alone. The most promising approach to this dilemma seems to be exploration learning. Here, the robot interacts with its environment and tries to build a corresponding model (or directly a control policy) using the gathered information. However, gathering data in robotics is time consuming and random exploration is potentially dangerous for the robot and its surroundings. Therefore, the main interest lies in sample-efficient reinforcement learning methods to find local models and control policies. How to generalize these results is an active area of research.

The purpose of this paper is to highlight these challenges as well as to suggest a framework for human-robot collaboration which utilizes the aforementioned ideas of using explicit communication, exploration learning and reactive control-based approaches to motion generation. The framework is rooted in historic ideas such as Brooks subsumption architecture [Brooks, 1986] and we demonstrate its applicability by means of an interactive pick & place use-case (see Fig. 1).

2 Motion Planning and Control

One core aspect of collaborative robotics is the necessity for reactive robot behavior generation. Classical monolithic sense-plan-act structures do not meet this demand. In contrast, it seems more promising to use reactive control-based motion generation [Kappler *et al.*, 2018]. At the core, the problem is to instantaneously generate control commands such that the resulting robot behavior satisfies the user requirements in interactive manipulation settings. One solution we are experimenting with is to base the motion planning and control architecture on a library of offline-trained/optimized controllers (policies) which can be grouped together to form skills which the user can parametrize and sequence online.

2.1 Skills

Similar in spirit to Brooks classical subsumption architecture [Brooks, 1986], we envision complex robot behavior to emerge in real time from the interplay of several concurrently running elemental controllers (policies). These controllers can potentially act in different operational spaces such as 6D task space, joint space, 3D Cartesian space or along a ray. Also, they can use different state descriptions (e. g., joint positions/velocities, interaction forces/moments, raw image data . . .). To resolve redundancy and possible conflicts between controllers they can be ordered according to a user-defined hierarchy. Here the intent is that lower-ranked controllers (e. g., responsible for motion generation) are only executed "as good as possible" such that higher-ranked ones ($e.\,g.$, for obstacle avoidance) are not affected. Figure 2 outlines an example of a skill composed of a set of hierarchically ordered controllers. Embedded optimization [Escande *et al.*, 2014] provides a possibility to invert each controllers commands from their respective operational spaces to joint space (in which the robot is controlled) in real time. Here, enforcing a hierarchy is accomplished by executing lower-ranked controllers as good as possible (in the least-square sense) in the null-space of higher ranked ones.

To accomplish complex tasks, skills need to be sequenced appropriately. To do so in a reactive manner we suggest to encapsulate skills together with appropriate success/failure conditions in Behavior Trees (BT) [Colledanchise and Ögren, 2017]. BTs are a directed tree where, at a given period, enabling signals (ticks) are sent from the root node down the tree. The main organizational units in a BT are Sequence nodes (denoted by $\rightarrow$) and Fallback nodes (denoted by ?). These are memoryless and test, each time they are executed, an ordered list of associated actions that can respond with one of the following statuses: running, success or failure. A Fallback node returns success if the first of its children succeeds or failure if all children fail (and running otherwise). In contrast, a Sequence node returns failure if the first child fails and returns success only if all children succeed. Apart from actions, Conditions are also possible. They can only return success or failure (not running). In our context, action nodes are formed by skills. An example for a pick-and-place behavior formed by skills and corresponding switching conditions is shown in Fig. 3.

2.2 Reactive Motion Generation Architecture

As shown in Fig. 4, the overall architecture rests on three legs: perception, skills and behaviors.

2.3 Behaviors

We provide a pre-defined library of skills among which the user has free choice to accomplish the task at hand. The controllers forming these skills are optimized/learned beforehand. One benefit is that existing controllers ($e.g.$, for avoidance) can be leveraged also for learning new ones and that controllers can be implemented in arbitrary operational spaces as the inversion of the corresponding kinematics/dynamics is centralized.

The idea is to use Human-Robot-Interaction (HRI) inputs in all three legs to guide the resulting behavior of the robot. To this end speech, gaze and gestures need to be mapped to numeric quantities suitable to i) parametrize the controllers underlying each skill and ii) to assemble a sequence of skills into a behavior tree. A detailed discussion on HRI in the context of collaborative robotics is given in Section 4.

3 Perception for Grasping and in-hand Manipulation

Replicating the effectiveness and flexibility of human hands in object manipulation tasks is an important open challenge. This requires a fundamental rethinking of how to exploit the multi-sensory data and the available mechanical dexterity of robot systems. In comparison to humans or primates, the dexterity of today's robotic grippers and hands is extremely limited [Feix *et al.*, 2013]. In [Bohg *et al.*, 2017], perception for manipulation approaches are divided into four categories. These are sensorless manipulation, image perception, active perception and interactive perception. Erdmann et. al. [Erdmann and Mason, 1988] investigated sensorless manipulation by means of generating motion strategies without any sensory feedback for simple manipulation tasks such as tray-tilting. Although, sensorless manipulation can work in simple scenarios, it is often insufficient for achieving complex tasks.

In visual perception approaches, static images are used to create sensory feedback for manipulation tasks. Cai et al. [Cai *et al.*, 2016] use images to understand the relation between the grasp types and object types. They analyze scenes in which a human performs grasps and a visual hand-tracking system detects and tracks the human hand. The grasped object and the attributes are inferred from the relative hand position and scale. The limitation of these approaches is that the actual object properties cannot be explicitly estimated since there is no direct interaction during perception.

In active perception approaches, the sensory feedback system is manipulated to mimic human attention and gaze. The common modality in this setting is an RGB camera. As an example, Nalpantidis et al. [Nalpantidis *et al.*, 2012] use robot motion to move the camera around the scene and segment objects reliably in cluttered environments. Compared to static camera settings, active perception approaches allow better modeling of an object with data from multiple views. However, the object properties cannot be inferred due to lack of physical interaction.

Interactive perception approaches are developed to combine physical interaction and traditional perception methods. The combination of these two allows a wider range of applications such as learning to manipulate unknown objects or object property learning. The main perception modality in these work is vision. For example, the authors in [Van Hoof *et al.*, 2014] use physical interaction and image features to segment objects in cluttered scenes. However, there are problems with these approaches. First, the outcome of the physical interaction adds more uncertainty to the task since it cannot be fully estimated. Second, the arm motion can obscure the view of the other sensors which can complicate sensor placement.

Our work has focused on the use of visual and haptic feedback for better understanding of object shapes, scene properties and in-hand manipulation. For example, object shape information is an important parameter when it comes to physical interaction with an object such as grasping and in-hand manipulation. Available object models may be erroneous especially when it comes to non-rigid objects where an object's shape may change due to frequent interaction with the object. In [Björkman *et al.*, 2013], we presented a probabilistic approach for learning object models based on visual and tactile perception through physical interaction with an object. The robot was enabled to touch objects incrementally by focusing on parts that were uncertain in terms of shape. The robot started by using only visual features to form an initial hypothesis about the object shape and then gradually added tactile measurements to refine the object model. This work was then continued in [Li *et al.*, 2014] where we focused on learning of grasp adaptation through experience and tactile sensing. We developed a grasp adaptation strategy to deal with uncertainties originating from physical properties of objects, such as the object weight and the friction at the contact points. Based on an object-level impedance controller, a grasp stability estimator was first learned in the object frame. Once a grasp was s predicted to be unstable, a grasp adaptation strategy was triggered according to the similarity between the new grasp and the training examples. Our recent work in [Li *et al.*, 2016] continued in a similar direction and addressed dexterous grasping under shape uncertainty. The uncertainty in object shape was parametrized and incorporated as a constraint into grasp planning. The proposed approach was used to plan feasible hand configurations for realizing planned contacts using different robotic hands. A compliant finger closing scheme was devised by exploiting both the object shape uncertainty and tactile sensing at fingertips.

In summary, effective and fully autonomous robot systems need the ability to interact with the physical world. Necessary behaviours range from scene understanding to in-hand object manipulation. Most of these have so far been studied in isolation and there is still a rather long way before robots can adapt their performance flexibly to dynamically changed scene, taking also into account the limitations of their own embodiment.

4 Human-Robot Interaction and Collaboration

There is an increasing interest in the area of collaborative robotics to make it possible for humans to teach robots different types of skills [Vollmer and Schillingmann, 2017]. In order for robots to learn from human demonstration, they first need to be able to learn to recognize meaningful goal-directed

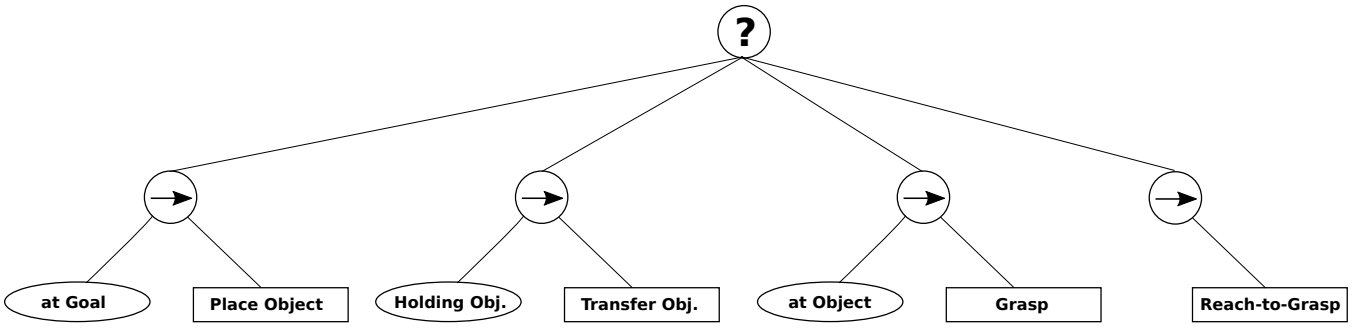


Figure 3: *Behavior example:* A pick-and-place behavior is accomplished by a reactive sequence of corresponding skills considering their respective success/failure conditions. The children of the corresponding Behavior Tree's leaves are tested from left to right.

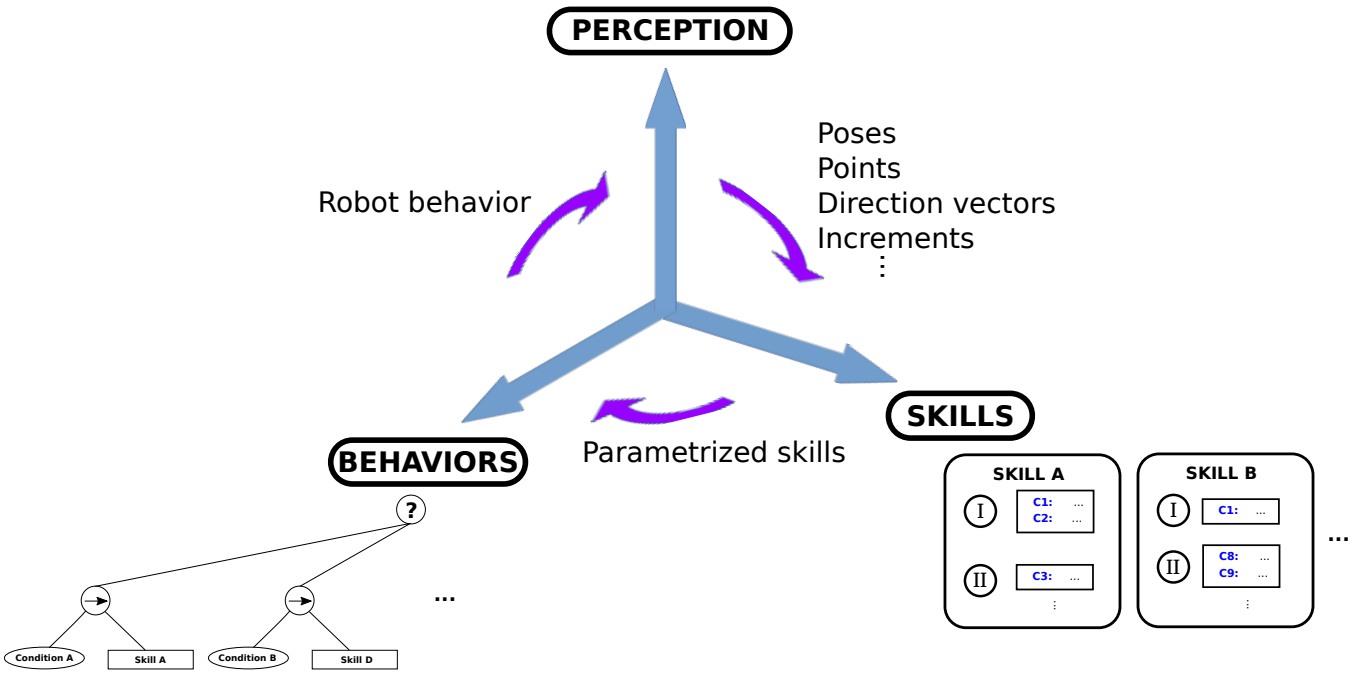


Figure 4: *Architecture layout:* The role of perception is to provide set points (*e. g.*, target alignment directions, or target points to reach) for the controllers composing the individual skills. The parametrized skills can then be composed to form complex motion plans.

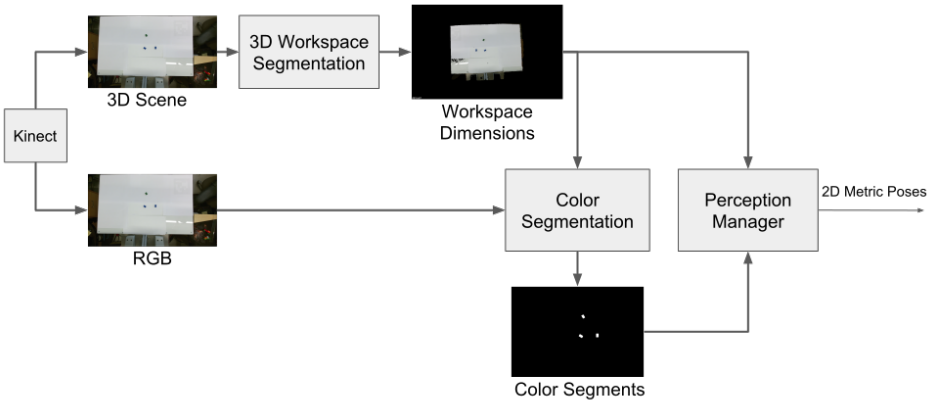
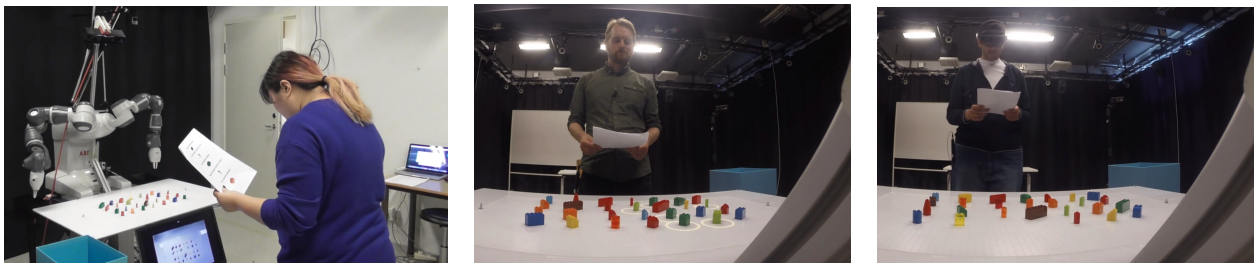


Figure 5: *Perception pipeline*: The overall perception system used in the interactive pick & place use-case.



(a) Monitor (b) Projector (c) Mixed Reality

Figure 6: *Human-Robot Interaction modalities:* (a) In Monitor mode, labels for identified objects are shown on a computer screen in front of the operator. (b) A projector is used to project labels onto the tabletop surface. (c) A Microsoft Hololens is utilized to directly visualize labels in the operators' field of view.

actions. The most efficient way to do this is to equip the robot with the ability to understand human's physical actions and their verbal and non-verbal communication [Liu and Zhang, 2017]. It then needs to learn task-driven changes of objects that the recognized human actions lead to. This will help to build a task plan that it can use to understand how different actions can be used to achieve the same goal [Pandey and Alami, 2014]. Finally, it has to be able to learn how to map the human actions onto its own body topology to achieve the same goal-directed actions. There are examples of systems that simultaneously learn both task-level goals and motor level actions from human teachers [Akgun and Thomaz, 2016]. In a dynamic environment imitation is not enough, but the robot needs to emulate what is observed to understand how to modify its actions in future slightly different situations [Vollmer *et al.*, 2014]. However, it will always be infeasible to learn just from a couple of sparse human demonstrations and it will require models of the dynamics provided by a combination of a-priori knowledge, exploration learning of the robot and large amounts of pre-collected data. This means that it is currently not realistic to build a fully autonomous robot that learns from human examples.

Nevertheless, by providing robots with the capability to understand human actions and communicative skills, we can already today build robots that collaborate with humans in assembly tasks. These collaborative robots could either assist humans that assemble, or make it possible for assembly robots to ask humans for assistance. In a scenario where the robot assists a human in assembly, it needs to know when to provide assistance. Either it simply responds to user requests for help or it automatically detects when it should assist [Baraglia *et al.*, 2016]. Humans can assist semi-autonomous assembly robots with low-level tasks such as object detection and situation-dependent adaptation of the execution of actions, or high-level tasks such as action acceptance and change of task sequence [Kyrarini *et al.*, 2018].

Our recent work has focused on studying how deep representation learning should be used for human motion prediction and classification [Bütepage *et al.*, 2017]. Generative models of 3D human motion are often restricted to a small number of activities and can therefore not generalize well to novel movements. We developed a deep learning framework for human motion capture data that learns a generic representation from a large corpus of motion capture data and generalizes well to new, unseen, motions. Using an encoding-decoding network that learns to predict future 3D poses from the most recent past, we extracted a feature representation of human motion. Most work on deep learning for sequence prediction focuses on video and speech. Since skeletal data has a different structure, we presented and evaluated different network architectures that made different assumptions about time dependencies and limb correlations. To quantify the learned features, we used the output of different layers for action classification and visualize the receptive fields of the network units. Our method outperformed the recent state of the art in skeletal motion prediction even though these use action specific training data.

Human-robot collaboration in assembly tasks requires that both parties can refer to objects in the shared space. Humans can use a combination of verbal descriptions, pointing gestures and gaze to single out one of many objects [Kennington and Schlangen, 2017]. A human that is assembling might not be able to use pointing gestures to request the next part from a third-hand helping robot. However, since verbal descriptions can be hard to disambiguate, the robot needs ways to achieve common ground of which object the user intended [Paul *et al.*, 2017]. Ambiguous referring expressions can be resolved by verbal grounding [Chai *et al.*, 2014], with pointing [Admoni *et al.*, 2016] or using gaze [Mehlmann *et al.*, 2014]. Robots can of course also use non-human ways to indicate their focus of attention, such as projecting their intentions into the shared environment [Chadalavada *et al.*, 2015] or by using augmented or mixed reality [Pereira *et al.*, 2017]. In the interactive pick & place scenario described in the next section we compare the effect of using different visualisation strategies to ground verbal descriptions of LEGO pieces.

5 Interactive Pick & Place: a Use-Case

We illustrate our approach to interactive and collaborative robotics by means of a pick & place scenario where a human operator and a robot interact with each other in a joint LEGO picking task (see Fig. 1). Here, the aim is to collect LEGO pieces from the table. The task starts with the human describing the required LEGO piece verbally. As the verbal descriptions are parsed, the robot queries the perceived LEGO pieces and labels the ones that match with the descriptions stored

in an available database. Once the required LEGO piece is identified by the robot, a pick & place behavior is executed. In order for the robot to take part in this scenario, multiple modules have to work in an integrated fashion.

The perception pipeline used in the presented use-case is illustrated in Fig. 5. The module uses the Kinect RGB-D sensor data to both segment the workspace ($i.e.$, the table top) and the objects in the workspace. The robot's workspace is segmented using depth information while the objects in the workspace are segmented using color segmentation. Subsequently, the perception manager receives the segmented workspace and segmented color blobs to filter out noisy segments and to compute 2D metric poses of the perceived objects with respect to the robots' frame of reference.

The human-robot grounding is achieved through different ways of visualizing the robot's understanding of the human's verbal descriptions of LEGO pieces, see Fig. 6. In the *monitor* modality, the robot provides feedback to the human through a monitor display. As the verbal requests of the human are parsed, the queried objects are labeled on the monitor display so that the person can continue the interaction by examining the monitor output. In the *projector* modality the queried objects are directly labeled on the workspace itself. Thus, the person can continue its interaction with the robot without changing his/her gaze and attention. In the *mixed reality* modality, the human interacts with the robot through a head-mounted display in which different layers of virtual feedback can be embedded in the person's view. Thus, as the verbal requests of the person are parsed, the interaction between robot and human is carried out by embedding virtual labels in the headset view.

In a user study, 29 subjects performed a task where a human operator and a robot took turns in asking the other to pick up one of the LEGO pieces on the table. They did three trials where they pick and placed 15 objects with the different grounding modalities. After each trial, they answered subjective questions from The Presence Inventory [Lombard *et al.*, 2009] and the Presence Questionnaire [Witmer and Singer, 1998]. In the subjective measures we found that the Mixed Reality system was most engaging, but least observable (due to the limited screen size in the head-mounted display used). Using projection onto the table was considered best overall, providing the observability with the least display interference with the task. We did not find any significant differences in completion times in the different modalities, and they led to very similar error rates. The conclusion is that all three could be used, and that the choice depends on the users, the task and the physical environment. In future studies we will investigate the benefits of indicating a robot's visual attention using pointing gestures or gaze generated by a back-project robotic head [Al Moubayed *et al.*, 2012]. When user requests are parsed and the particular object that will be grasped is identified, a pick & place behavior is triggered as described in Section 2.3. Note that, in our current implementation, we use a common state machine in place of a reactive behavior tree to transition between the individual skills forming the behavior. Skills are built from concurrently running controllers utilizing simple predefined PD-control laws. It is assumed that the workspace does not change during the robot's motion execution. Therefore, while the robot is executing the action, no feedback from the perception system is received due to Kinect's occluded field of view.

In the future, we envision to extend the HRI interaction modalities to enable verbal parametrization of movement controller setpoints such as "move here" or "a little more". These commands need to be mapped to numeric quantities suitable to be fed to the corresponding controllers. Similarly, commands such as "take the red part from here and put it over there" need to be parsed and associated with the individual skills necessary to accomplish the task. Also, recent work on reinforcement learning of manipulation policies [Levine *et al.*, 2016] has shown promise to acquire rich manipulation skills which will augment and/or replace our simple predefined control laws.

6 Discussion

Collaborative robotics has shown great promise to bring potentially complex tasks in frequently changing settings closer to automation. In that respect, especially tasks involving contact between the robot and the environment such as assembly have remained challenging. This is due to the fact that contact states are hard to detect and due to the difficulty in modeling the effects of the robots' actions. Reinforcement learning of local control policies has proven to be a promising method to obtain control policies for interaction tasks. Of particular importance here is the sample efficiency as explorative actions are costly and potentially hazardous. An open issue remains the generalization of learned policies to novel settings. We see the potential of addressing this using a-priori (partial) knowledge of the robots' model perform learning in task-invariant operational spaces. Recent policy learning approaches also achieved a tight coupling with perception [Levine *et al.*, 2016].

From the perception point of view, effective and flexible use of multisensory data in real-time will be necessary. Although sensor fusion has been demonstrated in other areas (mapping and localization), physical interaction suffers from the challenges outlined in the previous paragraph and many of the existing methodologies for sensor fusion do not meet all the challenges that physical contact, including both rigid and deformable objects, brings.

Learning and collaboration are tasks that come natural to us humans. Since the second half of the 1980's, the concept of embodied intelligence has revolutionized artificial intelligence. Instead of logical architectures and knowledge representation, the embodied view argues that intelligent behavior emerges naturally from the interplay between motor and sensory channels. Mirror neurons and social engagement together with physical constraints imposed by the environment enable and challenge the human way of learning and interacting with the environment and collaborating with each other. Although there are difficulties in conveying and generating new knowledge, the process is simplified given that we share quite similar body shape and that imitation can be used to achieve a goal. The policy or the way of achieving the goal can then be refined through training and experimentation. So, how can we achieve the same way of interaction between a

human and a machine when the differences in perception and acting are, and probably will, remain different? Maybe, in the future, we will be able to build mechanical structures that are superior to humans in perception and action, and developing collaborative setups will bring a completely new set of challenges to resolve.

Acknowledgements

This work was supported by the Swedish Foundation for Strategic Research (SSF) and Knut and Alice Wallenberg Foundation through the WASP project.

References

[Admoni *et al.*, 2016] Henny Admoni, Thomas Weng, and Brian Scassellati. Modeling communicative behaviors for object references in human-robot interaction. In *Proc. IEEE Int. Conf. Robotics and Autom.*, pages 3352–3359. IEEE, 2016.

[Akgun and Thomaz, 2016] Baris Akgun and Andrea Thomaz. Simultaneously learning actions and goals from demonstration. *Auton. Rob.*, 40(2):211–227, 2016.

[Al Moubayed *et al.*, 2012] Samer Al Moubayed, Jonas Beskow, Gabriel Skantze, and Björn Granström. Furhat: a back-projected human-like robot head for multiparty human-machine interaction. In *Cognitive behavioural systems*, pages 114–130. Springer, 2012.

[Baraglia *et al.*, 2016] Jimmy Baraglia, Maya Cakmak, Yukie Nagai, Rajesh Rao, and Minoru Asada. Initiative in robot assistance during collaborative task execution. In *Proc. ACM/IEEE Int. Conf. Human-robot Int.*, pages 67–74. IEEE Press, 2016.

[Björkman *et al.*, 2013] Mårten Björkman, Yasmin Bekiroglu, Virgile Högman, and Danica Kragic. Enhancing visual perception of shape through tactile glances. In *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, pages 3180–3186, 2013.

[Bohg *et al.*, 2017] Jeannette Bohg, Karol Hausman, Bharath Sankaran, Oliver Brock, Danica Kragic, Stefan Schaal, and Gaurav S. Sukhatme. Interactive perception: Leveraging action in perception and perception in action. *IEEE Trans. Rob.*, 33(6):1273–1291, Dec 2017.

[Brooks, 1986] Rodney Brooks. A robust layered control system for a mobile robot. *IEEE Journal on Robotics and Automation*, 2(1):14–23, 1986.

[Bütepage *et al.*, 2017] Judith Bütepage, Michael Black, Danica Kragic, and Hedvig Kjellström. Deep representation learning for human motion prediction and classification. In *Proc. IEEE Int. Conf. Comp. Vision and Pattern Recogn.* IEEE, 2017.

[Cai *et al.*, 2016] Minjie Cai, Kris M. Kitani, and Yoichi Sato. Understanding hand-object manipulation with grasp types and object attributes. In *Robotics: Science and Systems*, 2016.

[Chadalavada *et al.*, 2015] Ravi Teja Chadalavada, Henrik Andreasson, Robert Krug, and Achim J Lilienthal. That's on my mind! robot to human intention communication through on-board projection on shared floor space. In *Proc. European Conf. on Mobile Robots*, pages 1–6, 2015.

[Chai *et al.*, 2014] Joyce Y. Chai, Lanbo She, Rui Fang, Spencer Ottarson, Cody Littley, Changsong Liu, and Kenneth Hanson. Collaborative effort towards common ground in situated human-robot dialogue. In *Proc. ACM/IEEE Int. Conf. Human-robot Int.*, pages 33–40. ACM, 2014.

[Colledanchise and Ögren, 2017] Michele Colledanchise and Petter Ögren. How behavior trees modularize hybrid control systems and generalize sequential behavior compositions, the subsumption architecture, and decision trees. *IEEE Trans. Rob.*, 33(2):372–389, 2017.

[Erdmann and Mason, 1988] Michael A. Erdmann and Matthew T. Mason. An exploration of sensorless manipulation. *Robotics and Automation, IEEE Journal of*, 4(4):369–379, August 1988.

[Escande *et al.*, 2014] Adrien Escande, Nicolas Mansard, and Pierre-Brice Wieber. Hierarchical quadratic programming: Fast online humanoid-robot motion generation. *Int. J. Rob. Res.*, 33(7):1006–1028, 2014.

[Feix *et al.*, 2013] Thomas Feix, Javier Romero, Carl Henrik Ek, Heinz-Bodo Schmiedmayer, and Danica Kragic. A metric for comparing the anthropomorphic motion capability of artificial hands. *IEEE Trans. Rob.*, 29(1):82–93, 2013.

[Kappler *et al.*, 2018] Daniel Kappler, Franziska Meier, Jan Issac, Jim Mainprice, Cristina G. Cifuentes, Manuel Wthrich, Vincent Berenz, Stefan Schaal, Nathan Ratliff, and Jeannette Bohg. Real-time perception meets reactive motion generation. *IEEE Rob. Autom. Letters*, 3(3):1864–1871, 2018.

[Kennington and Schlangen, 2017] Casey Kennington and David Schlangen. A simple generative model of incremental reference resolution for situated dialogue. *Computer Speech & Language*, 41:43–67, 2017.

[Kruijff *et al.*, 2006] Geert-Jan M. Kruijff, Hendrik Zender, Patric Jensfelt, and Henrik I. Christensen. Clarification dialogues in human-augmented mapping. In *Proc. ACM/IEEE Int. Conf. Human-robot Int.*, Salt Lake City, UT, March 2006.

[Kyrarini *et al.*, 2018] Maria Kyrarini, Muhammad Abdul Haseeb, Danijela Ristić-Durrant, and Axel Gräser. Robot learning of industrial assembly task via human demonstrations. *Auton. Rob.*, pages 1–19, 2018.

[Levine *et al.*, 2016] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *J. Machine Learning Res.*, 17(1):1334–1373, 2016.

[Li *et al.*, 2014] Miao Li, Yasmin Bekiroglu, Danica Kragic, and Aude Billard. Learning of grasp adaptation through experience and tactile sensing. In *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, pages 3339–3346, 2014.

[Li *et al.*, 2016] Miao Li, Kaiyu Hang, Danica Kragic, and Aude Billard. Dexterous grasping under shape uncertainty. *Robotics and Autonomous Systems*, 75:352 – 364, 2016.

[Liu and Zhang, 2017] Rui Liu and Xiaoli Zhang. Systems of natural-language-facilitated human-robot cooperation: A review. *arXiv preprint arXiv:1701.08269*, 2017.

[Lombard *et al.*, 2009] Matthew Lombard, Theresa B. Ditton, and Lisa Weinstein. Measuring presence: the temple presence inventory. In *Proceedings of the 12th Annual International Workshop on Presence*, pages 1–15, 2009.

[Mehlmann *et al.*, 2014] Gregor Mehlmann, Markus Häring, Kathrin Janowski, Tobias Baur, Patrick Gebhard, and Elisabeth André. Exploring a model of gaze for grounding in multimodal hri. In *Proc. Int. Conf. on Multimodal Interaction*, pages 247–254. ACM, 2014.

[Nalpantidis *et al.*, 2012] Lazaros Nalpantidis, Mårten Björkman, and Danica Kragic. YES - YEt another object segmentation: Exploiting camera movement. In *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, pages 2116–2121, 2012.

[Pandey and Alami, 2014] Amit Kumar Pandey and Rachid Alami. Towards human-level semantics understanding of human-centered object manipulation tasks for hri: Reasoning about effect, ability, effort and perspective taking. *Int. J. Social Rob.*, 6(4):593–620, 2014.

[Paul *et al.*, 2017] Rohan Paul, Jacob Arkin, Nicholas Roy, and Thomas Howard. Grounding abstract spatial concepts for language interaction with robots. In *Proc. Int. Joint Conf. on Artificial Intelligence*, pages 4929–4933, 2017.

[Pereira *et al.*, 2017] André Pereira, Elizabeth J Carter, Iolanda Leite, John Mars, and Jill Fain Lehman. Augmented reality dialog interface for multimodal teleoperation. In *Proc. IEEE Int. Symp. on Rob. and Human Int. Comm.*, pages 764–771, 2017.

[Van Hoof *et al.*, 2014] Herke Van Hoof, Oliver Kroemer, and Jan Peters. Probabilistic segmentation and targeted exploration of objects in cluttered environments. *IEEE Trans. Rob.*, 30(5):1198–1209, 2014.

[Vollmer and Schillingmann, 2017] Anna-Lisa Vollmer and Lars Schillingmann. On studying human teaching behavior with robots: a review. *Review of Philosophy and Psychology*, pages 1–41, 2017.

[Vollmer *et al.*, 2014] Anna-Lisa Vollmer, Manuel Mühlig, Jochen J Steil, Karola Pitsch, Jannik Fritsch, Katharina J Rohlfing, and Britta Wrede. Robots show us how to teach them: Feedback from robots shapes tutoring behavior during action learning. *PloS one*, 9(3):e91349, 2014.

[Witmer and Singer, 1998] Bob G Witmer and Michael J Singer. Measuring presence in virtual environments: A presence questionnaire. *Presence*, 7(3):225–240, 1998.