Report from Dagstuhl Seminar 15411

Multimodal Manipulation Under Uncertainty

Edited by

Jan Peters[1], Justus Piater[2], Robert Platt[3], and Siddhartha Srinivasa[4]

1 **TU Darmstadt, DE,** mail@jan-peters.net
2 **Universität Innsbruck, AT,** justus.piater@uibk.ac.at
3 **Northeastern University – Boston, US,** rplatt@ccs.neu.edu
4 **Carnegie Mellon University, US,** siddh@cs.cmu.edu

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 15411 "Multimodal Manipulation Under Uncertainty". The seminar was organized around brief presentations designed to raise questions and initiate discussions, multiple working groups addressing specific topics, and extensive plenary debates. Section 3 reproduces abstracts of brief presentations, and Section 4 summarizes the results of the working groups.

Seminar October 4–9, 2015 – http://www.dagstuhl.de/15411
1998 ACM Subject Classification I.2.9 Robotics
Keywords and phrases Robotics, Manipulation, Uncertainty, Perception, Computer vision, Range sensing, Tactile sensing
Digital Object Identifier 10.4230/DagRep.5.10.1

1 Executive Summary

Jan Peters
Justus Piater
Robert Platt
Siddhartha Srinivasa

While robots have been used for decades to perform highly specialized tasks in engineered environments, robotic manipulation is still crude and clumsy in settings not specifically designed for robots. There is a huge gap between human and robot capabilities, including actuation, perception, and reasoning. However, recent developments such as low-cost manipulators and sensing technologies place the field in a good position to make progress on robot manipulation in unstructured environments. Various techniques are emerging for computing or inferring grasp configurations based on object identity, shape, or appearance, using simple grippers and robot hands.

Beyond grasping, a key ingredient of sophisticated manipulation is the management of *state information and its uncertainty*. One approach to handling uncertainty is to develop grasping and manipulation skills that are robust to environmental variation. Another approach is to develop methods of interacting with the environment in order to gain task-relevant information, for example, by touching, pushing, changing viewpoint, etc. Managing state information and uncertainty will require a tight combination of perception and planning. When the sensor evidence is unambiguous, the robot needs to be able to recognize that and

perform the task accurately and efficiently. When greater uncertainty is present, the robot needs to adjust its actions so that they will succeed in the worst case or it needs to gain additional information in order to improve its situation. Different sensing modalities as well as world models can often be combined to good effect due to their complementary properties.

This seminar discussed research questions and agendas in order to accelerate progress towards robust manipulation under uncertainty, including topics such as the following:

- Is there a master algorithm or are there infinitely many algorithms that solve specialized problems? Can we decompose multimodal manipulation under uncertainty into I/O boxes? If so, what would these be?
- Do we prefer rare-feedback / strong-model or frequent-feedback / weak-model approaches? Is there a sweet spot in between? Is this the way to think about underactuated hands?
- What are useful perceptual representations for manipulation? What should be the relationship between perception and action? What kind of perception is required for reactive systems, planning systems, etc.?
- How do we do deformable-object manipulation? What planning methods, what types of models are appropriate?
- How should we be benchmarking manipulation? What kind of objects; what kind of tasks should be used?
- How should humans and robots collaborate on manipulation tasks? This question includes humans collaborating with autonomous robots as well as partially-autonomous robots acting under human command.

In the area of perception, we concluded that the design of representations remains a central issue. While it would be beneficial to develop representations that encompass multiple levels of abstraction in a coherent fashion, it is also clear that specific visual tasks suggest distinct visual representations.

How useful or limiting is the engineering approach of decomposing functionality into separate modules? Although this question was heavily debated, the majority view among seminar participants was that modules are useful to keep design complexity manageable for humans, and to keep the event horizon manageable for planning systems. It seems that to build more flexible and powerful systems, modules will need to be more strongly interconnected than they typically are these days. Fundamental challenges lie in the specification of each module and of their interconnections. There is a lot of room for creative innovation in this area.

Benchmarking questions were discussed chiefly in the context of the YCB Object Set[1]. Specific benchmarks were suggested and discussed, covering perception and planning in the context of autonomous manipulation.

[1] http://www.ycbbenchmarks.com/

2 Table of Contents

3 Overview of Talks

3.1 Uncertainty during Assistive Manipulation

Brenna Argall (Northwestern University – Evanston, US)

It is a paradox that often the more severe a person's motor impairment, the more challenging it is for them to operate the very assistive machines which might enhance their quality of life. *Assistive manipulators* pose a particular challenge because of their complexity: the dimensionality of the manipulator's control space generally far exceeds the dimensionality of the control signal able to be produced by the human operator (for reasons of motor impairment, or interface limitations). By introducing robotics autonomy and intelligence, we can turn the manipulator into an autonomous robot and offload some of the control burden from the human. Under such an assistance paradigm, *multi-modality* presents itself foremost within the space of *control signals*—since there are multiple (the human, the robotics autonomy) sources controlling the robot platform. *Uncertainty* within the domain of assistive robotic manipulation presents itself in many forms. One way is in the *inference of operator intent*. For the autonomy to provide control assistance requires an idea of the operator's intended task or movement. The aforementioned mismatch in control space complicates and introduces uncertainty into this inference. A second way is in the *estimation of optimal assistance*. Exactly how much assistance is required, or desired, by the operator is a critical unknown. We further expect that the optimal amount of assistance is unique to each human operator, because of the uniqueness of their personal preferences and physical abilities. A third way is in *how to adapt the assistance paradigm*. We expect the optimal assistance solution to change over time, because people's abilities change over time. The right way to adapt the assistance paradigm—autonomously, and without engineer intervention—is unknown, and user-specific. Unknown, yet fundamental—the customization and adaptation of assistance I believe will be critical to the adoption of assistive robots within larger society.

3.2 Learning in Robotics

Leslie Pack Kaelbling (MIT – Cambridge, US)

How can we make robots that are really robust, flexible, and competent in complex relatively unstructured environments? Through a combination of high-level design of algorithms, representations, and structures, on the part of human engineers, and learning and planning, on the part of the robot.

Robots can learn *synthetic* and *analytic* knowledge. Synthetic learning gains actual information about the domain they operate in; it can be represented in terms of policies, values functions, reward models, dynamics models, or observation models; it can be short term (What is the pose of the object in front of me?) or long term (What should the gain on my motor controller be? How does rain affect the arrival time of the people I cook dinner for?); it can be more or less abstract. Analytic learning can be thought of as a kind of compilation or tuning of internal representations: it includes learning to play chess (after

you know the rules) or using a slow planner to generate training data for a policy that will be quick to execute.

As domains become more highly variable, more complex, and longer-horizon, I argue that learning structures that can be re-used is critical. Learning a predictive model of kinematics or physics or folk psychology can be re-used over and over with different objectives and can often be adapted with very few training examples. Learning a policy or value function is tied to an objective and will generally be more difficult to transfer, adapt, or re-use. Ultimately, an effective general-purpose robot will need all of these kinds of structures: from fast, specific, low-level policies to slow, abstract, general purpose knowledge and reasoning mechanisms. A critical research question is how to design an architecture that supports these kinds of learning, reasoning, and behavior.

Related questions include:

- What kinds of model representations are most useful for what kinds of problems?
- Can we formulate the objective of a learning problem to include some notion of how the learned structures will be used? (So, for example, it might be useful to learn both a very detailed and a very abstract model for predicting physical interactions of objects, and then be able to employ the most useful one in each circumstance).
- At what point to we really need to address general-purpose reasoning? How should it integrate with the basic planning and learning mechanisms that we use now in robotics?

3.3 Action Selection in Hybrid Spaces

Tomas Lozano-Pérez (MIT – Cambridge, US)

The fundamental planning problems for autonomous manipulation have high-dimensional hybrid state spaces and many actions, also with hybrid parameters. Furthermore, these planning problems, e.g. making dinner, typically have very long planning horizons and take place under substantial uncertainty both in the current state and the result of actions. How do we build effective planners for these problems? We have been exploring approaches built on the following cluster of ideas inspired by AI planning, decision-theoretic planning and motion planning:

- Factored representations of belief space
- Determinize and re-plan for probabilistic planning
- Temporal hierarchies for abstraction
- Implicit representation of pre-images for backchaining
- Relaxed planning problems for heuristic guidance
- Sample-based minimum-constraint removal motion-planning

This talk outlines how these ideas fit together into a coherent whole and discuss strengths and weaknesses.

Related questions include:

- How much of a robot planner could be robot-independent? That is, at what level (if any) could planning effectively become independent of geometry, kinematics, perception, etc.
- How do we effectively combine planning and learning to build autonomous robots?

3.4 Some Basic and/or Old Thoughts on Multimodality and Uncertainty (and New Thoughts on the Amazon Picking Challenge)

Oliver Brock (TU Berlin, DE)

Multimodality

In robotics, *multimodal* is often taken to mean the same as *multi-sensorial*. The question is: what should constitute a modality? If a modality simply is a type of sensor, the matter seems well-understood for a long time. In 1988 Hugh Durrant-Whyte identified three types of multi-sensor integration: competitive, complementary, and cooperative [Sensor Models and Multisensor Integration, IJRR 7(6):97–113, 1988]. But it seems natural to think about the goal of perception when talking about modalities. This view of modalities is what James J. Gibson described in this 1983 book *The Senses Considered as a Perceptual System*. When the senses are viewed a perceptual system, a modality corresponds to a regularity in the sensor data, irrespective of what sensor-type they originated from. This implies that perception should consist of (at least) two layers: one that extracts regularities from multi-sensor streams, and one that leverages these signals appropriately for a particular application. Extracting important regularities, i.e. regularities that are robust and useful for the task of an agent, becomes an important problem in perception. In some way, this changes the general approach: rather than thinking of an application and then identifying the information we might need for it, we should start longing for robust and generally useful regularities in multi-sensorial data. Interestingly, this is exactly what interactive perception is attempting to do.

Uncertainty

There are many approaches to address uncertainty. None of them can be used exclusively to address the uncertainty a real-world robotic system is exposed to. No matter how much effort we invest in modeling uncertainty in a POMDP, there will always be remaining uncertainty that is not reflected in our model. No matter how smart I design my mechanism to suppress uncertainty through clever engineering, there will be situations when the design is insufficient. No matter what assumptions I make about the agent, the world, and their interaction, situations will arise that we not anticipated. At the moment, the key opportunity for addressing uncertainty in real-world robotics is not to advance any one of these possibilities but instead to learn how to cleverly combine them into a robust system.

Amazon Picking Challenge

My lab won the inaugural Amazon Picking Challenge. Of course, the most interesting question is: why? What can be learned from this success? While it is difficult to learn from a single sample, we believe that there were several factors that played a major role.

1. Luck: Back home we tested all five shelf configurations (placement of objects in bins) and the one we had to solve during the competition was the one we performed best on. However, we also performed very well on all the other configurations.
2. Behavior first: Rather than improving components, we always worked on the behavior of the integrated system. We observed that improving the performance of an isolated

component does not necessarily (and maybe rarely) improve the performance of the system containing the component. This has far-reaching consequences. For example, the basic compositional entity (dare I say "module"?) was a behavior of the entire system. Traditionally, these compositional entities are vision systems, planners, controllers, etc.
3. Embodiment: We picked the robot to suit the task. This is reflected in our end-effector (vacuum cleaner with suction cup) and in the fact that we used a mobile base. These are only two choices that ended up making the solutions to other problems much simpler.
4. Prior knowledge: We thought hard about where we can exploit knowledge of the specific problem and the specific setup in our solution. This makes things simpler and therefore more robust, albeit not general, of course.

During the process of preparing for the APC, it became clear to us that "systems papers" in robotics are not very helpful for building systems. Most of them simply describe a particular system, rather than postulating principles of system building that then can be confirmed or disproven by other groups. The robotics community must start writing systems papers to grow a body of knowledge about system building, hopefully leading to some kind of system science.

3.5 On Multifingered Hands and their (Lack of) Industrial Applications

Máximo A. Roa (German Aerospace Center-DLR, DE)

Despite almost 30 years of development in multifingered hands, traditional two-finger parallel jaw grippers are still one of the most common choices for grasping objects in industrial environments. Multifingered hands were developed for solving challenging manipulation tasks, but applications are still marginal, mainly due to the mechanical complexity of the devices, the complex associated control, and their high cost. Hardware is, though, capable of a large variety of interesting behaviors, as demonstrated in teleoperated scenarios of robotic manipulation, or by amputees operating prosthetic devices. The question of the required dexterity in a robotic end effector arises, since simple devices seem to be capable of amazing behaviors, when a human is in control of the actions (planning and execution).

From a robotic perspective, handling uncertainties can be tackled at different levels: hardware, planning or control. From the control point of view, the framework of compliant control applied at object or joint level can cope with deviations of the object pose with respect to the nominal position, leading to robust grasp behaviors. Probabilistic approaches for planning have started to consider uncertainties in the loop. Also, recent hardware advances such as underactuated hands follow the principle of exploiting as far as possible the dynamics of mechanisms to simplify the control tasks, and they favor the exploitation of (instead of avoiding) the contacts with the environment to maximize grasp robustness.

Applications of grasping for manufacturing applications (especially in SMEs) requires still some effort in terms of integration, control, execution and error recovery to guarantee robust applications of robotic technology. The combination of grasp and assembly planning is an example that illustrates the introduction of highly automated workflows for productions of small batches of assembled modular structures. Closing the action-perception loop is crucial for achieving reliability in this domain, using robust error-recovery strategies.

Related questions include:

- How much complexity is required for the end effectors? (Do we really need dexterous hands, or is it enough having task-specific end effectors?)
- How to effectively combine hand and arm dexterity?
- What is the best approach to handle uncertainty in the manipulation process: through hardware, planning, control?
- Is multisensorial perception really needed, or is vision sufficient to successfully perform manipulation tasks?
- Should manipulation problems be solved as combinations of basic skills?

3.6 Haptic Perception and Other Things that Keep Me Up at Night

Veronica J. Santos (University of California – Los Angeles, US)

In the Socratic spirit of this workshop, I will present some of our truths, working assumptions, and challenges in the context of haptic perception for artificial hands. I will provide a brief summary of some (non-intuitive?) insights we have gained from experiments in which a robot hand outfitted with a deformable, multimodal tactile sensor was used to replay human-inspired haptic *exploratory procedures* to perceive edge orientation and fingertip-sized geometric features. I will then give a sneak preview of some new experiments (some more developed than others) on such grand challenge-inspired topics as bimanual manipulation, haptic search within a granular material, and manipulation of deformable materials. These experiments do not yet have "punchlines," but could be used to stimulate discussion of potential pitfalls, alternative approaches, and collaborative extensions of this work. I will end my talk with a list of things that keep me up at night and where I would love to see the field of manipulation in the next 10–20 years.

Truths and working assumptions:

- Perception is an active process.
- Learnable, consistent action-perception relationships are key.
- Solutions can be *bio-inspired* without having to be *biomimetic*.

Related questions include:

- How can we empower robots with a deeper understanding of objects and actions? How can we break away from pre-planned trajectories and teach robots to perceive when a task has been completed or that different actions must be taken to achieve task completion?
- How can we, or should we, teach robots physics? How much physics intuition is needed for a robot to reason about grasp and manipulation tasks?
- What is a practical, useful, (compact? easily searchable? modular?) representation of learned experiences[2] for robotics?
- What machine learning techniques will enable us to achieve online perception and decision-making for interactions with humans at human-like speeds? Or should we not be concerned with speed at this point?

[2] Thanks to Leslie Pack Kaelbling, Mehmet Dogar, and Kostas Bekris for this topic.

- How can we discover new solutions for artificial manipulation when current machine learning techniques are limited by our own (often subjective) hand-tuned model parameters, input features, and reward structures?
- How can we extract physical intuition from successful "black box" machine learning approaches?

3.7 Computing Motions for Robots in Healthcare Applications

Ron Alterovitz (University of North Carolina at Chapel Hill, US)

Emerging robots have the potential to improve healthcare delivery, from enabling surgical procedures that are beyond current clinical capabilities to autonomously assisting people with daily tasks in their homes. I will discuss new algorithms to enable medical and assistive robots to safely and semi-autonomously operate inside people's bodies or homes. These algorithms must compensate for uncertainty due to variability in humans and the environment, consider deformations of soft tissues, guarantee safety, and integrate human expertise into the motion planning process.

I will discuss ongoing and future research on motion planning algorithms for new tentacle-like medical instruments, including steerable needles and concentric tube robots, designed for interventional radiology, cardiothoracic surgery, and neurosurgery procedures. These new devices can maneuver around anatomical obstacles to perform procedures at clinical sites inaccessible to traditional straight instruments. To ensure patient safety, our algorithms must explicitly consider uncertainty in motion and sensing to maximize the probability of avoiding obstacles and successfully accomplishing the task. We compute motion policies by integrating sampling-based motion planners, optimal control, and parallel computation. Second, I will discuss new motion planning algorithms for autonomous robotic assistants for helping people with tasks of daily living in the home. I will present demonstration-guided motion planning, an approach in which the robot first learns time-dependent features of an assistive task from human-conducted demonstrations and then autonomously plans motions to accomplish the learned task in modified environments with never-before-seen obstacles.

Related questions include:

- Uncertainty is an inevitable implication of medical robots becoming smaller and gaining degrees of freedom and assistive robots using less precise actuators and sensors to gain compliance and decrease cost. How do we manage uncertainty in robot motion and state estimation in a manner that enables us to provide guarantees on safety?
- Medical robots operate in deformable environments, where both the surrounding soft tissues and the robot itself may deform. How do we model such settings such that we can efficiently compute high-quality motion plans that increase the autonomy and safety of medical robots?
- Physicians don't like to hand over all control to an autonomous agent, but like receiving assistance that makes their job easier. How should robots and physicians effectively share autonomy during surgery?

3.8 Data, Data, Data – How much of it do we really need in robotics?

Jeannette Bohg (MPI für Intelligente Systeme – Tübingen, DE)

Large-scale, labeled databases for classic tasks in Computer Vision or speech recognition and the ability to learn from them provide the key to the state-of-the-art performance on many benchmarks that are important in these areas. In robotics however, databases of this magnitude are rare which may be due to the effort that would be required to collect them. Traditionally, robotics is much more driven by models that are for example based on first principles. Often these provide very useful abstractions of dynamical systems to develop robust controllers. But in many cases, we have seen that the underlying assumptions of such a models do not hold in the real world. This has for example been the case for grasping. In some of our recent work, we assembled a large-scale database for learning to grasp given only partial and noisy data of the object. The data points are automatically annotated using physics simulation of the grasp. This type of data helps to synthesize stable grasps and even to predict some of the latent object properties like global object shape, category or contact locations.

However, I wonder about two related questions:

1. What is the right mixture of first-principle modelling and learning? Do we want to learn entirely from scratch to not bias the resulting model with our potentially too restrictive ideas? Or do we want to learn isolated parameters of an otherwise fixed model? In what situations is learning in one way or the other preferred?
2. Related to the above, how much data do we really need? How should it be collected: incrementally or in batch? What is the important data to collect: all the modalities or some intuitively important ones? Should it be time-series data or data collected at a particular point in time?

The relation between these two questions lies in the apparent trade-off between including expert knowledge in our models and the amount of data required to learn the remaining open parameters.

4 Working Groups

4.1 Perception

- What are useful perceptual representations for manipulation?
- What should be the relationship between perception and action?
- What kind of perception is required for reactive systems, planning systems, etc.?

4.1.1 Group 1

We began by discussing the uses of perception: disambiguation, association, and creation of a unified world model for decision-making. We discussed how perception representation must include uncertainty, and that tactile and/or proximity sensing could be used to reduce uncertainty. The value of tactile sensing seems to be (unnecessarily) dependent upon the approach to manipulation. For example, reliance on contact location information limits the usability of certain tactile sensor technologies that might otherwise be useful if a different approach were taken. There was disagreement on the richness of tactile information needed

for manipulation. We discussed whether complete 3D models of objects are necessary to grasp and manipulate the environment. Alternative approaches were suggested, such as focusing on labeling of objects, or starting with a partial model (e.g. incomplete surface models) and using tactile sensing to adjust grasp after contact is made. We discussed reactive systems as enacting direct mappings of inputs to outputs that do not project forward in time. However, planning systems can be implemented as reactive systems (e.g. reinforcement learning).

Different approaches to perception were discussed. One approach is to broadly search for patterns in sensor data. Another approach is to look specifically for relevant task-based information. We discussed the degree to which sensor data must be processed for use in the world model. Classical AI approaches use a *lazy* approach to perception in which data are stored and accessed when needed. In contrast, control theory approaches use a *greedy* approach in which filters are used to maintain online, on-demand, updated models of the world at all times. We discussed whether awareness was necessary for perception. Is it possible to have physical, but not computational awareness? For example, a jamming gripper may not be aware of its shape, and yet it still reacts and changes shape in response to stimuli.

We concluded that the explicit representation of uncertainty is essential for combining representations of the world. The network for fusing different representations of the world and queries of the environment can be task-specific. Action and perception should be considered jointly. There is no "correct" approach to perception, as the approach depends on the application.

4.1.2 Group 2

We observed that object-action representations span multiple levels of abstraction. Depending on the information available to the robot and the task at hand, the link between perception and action can be direct (e.g., motor reflex linked to a specific tactile pattern), or it can pass through abstracted representations, for instance a 3D vector representation of a world contact, surface patches with surface normals (used for hand-object interaction models, as in the work of Kopicki and Wyatt), complete object surface representations (as in grasp densities), symbolic labeling, generic hand postures or prototypical motor patterns.

The outcome of our discussion is that it would be profitable to develop representations that encompass multiple levels of abstractions, instead of using distinct representations for different levels. The representation should also allow for long term information gathering (graceful integration), and lend itself to choosing actions that disambiguate perception.

4.1.3 Group 3

The perceptual representation to use depends on the task at hand. Even during a simple grasping operation, we can create a matrix mapping each subtask to a perceptual representation: *a geometric representation* during motion planning, *an appearance-based representation* during visual servoing to align the hand with the object, *a direct representation* during the guarded move to make the first contact with the object, and so on. Some of the important and useful representations include: spatial probabilistic representations over poses, spatial probabilistic representations over occupancy grids, symbolic representations, affordances, appearance/material/inertial representations, worst/best-case geometry representations, direct representations ("world is its own model"), and predictive state representations.

4.2 General vs. Specific Solutions

4.2.1 Group A

- Is there a master algorithm or are there infinitely many algorithms that solve specialized problems?
- Can we decompose multimodal perception under uncertainty into a simple system of separate, interconnected I/O modules? If so, what would these be?

Assuming that a master algorithm composed of isolated submodules exists, one can start thinking of how such an algorithm could look like. This master algorithm would consist of separate modules for computer vision, control and others. For such a modular structure, each module is working independently of the other modules to some degree (e.g. a module can simply be replaced without any effort). However, the question came up, how modularity is defined in this setting. We define a module by the number of connections inside and outside the module (e.g. a component is modular, if some subparts have high inter-connectivity but the number of connections to other components is low). However, it still remains unclear, if such a definition is meaningful as there are different levels of modularity. Almost every system can be formulated in a modular way, even though it isn't (e.g. the human brain is known not to be strictly modular; however, on a lower level it consists of many neurons that can be seen as modules). In order to decide for the likelihood of a strictly modular approach to be successful we looked at the human brain. The brain is clearly not modular in a strict sense. Many synapses are connecting neurons from different areas within the brain. Still, one can define several areas that mainly seem to be responsible for a specific purpose. Therefore, we do not reject the idea of modularity (and the currently dominant engineering approach in robotics), but we propose the significant extension of having higher inter-connection between the modules. Each module incorporates a different prior on the input data and produces a prediction based upon it. In order to obtain much more robust functionality, these outputs could be fed back to the other modules. Such a paradigm should be able to cancel out noise significantly. In the perception domain, this relates to the fusion of different sensing modalities. For humans it is known that these affect each other. For example a human can smell, see, feel and even taste fire, where each of these separate modalities can be used to predict much more precisely if fire is present or not.

4.2.2 Group B

Can we decompose multimodal manipulation under uncertainty into I/O boxes? – Our conclusion was: yes. Large successful engineered systems typically can be represented as boxes. A key challenge is how to create them and connect them.

Creating *useful* boxes is challenging. A box should be defined by a task or a well-defined problem; it should have preconditions and postconditions (i.e., a goal) that are well defined and we should be able to develop an algorithm for it. A box needs to be thought about based on how it interacts with other boxes; if one decides on a box without thinking about other boxes, one cannot define a useful box. Smaller boxes are better, e.g., ape's primitive motions, a learned subtask, and a push grasp.

When creating boxes, a consideration is whether to take a top-down vs. bottom up approach. In the top-down approach, one starts with a general approach (e.g., a hierarchy) and then creates boxes for various operations in a hierarchy. In the bottom-up approach, one builds boxes for certain skills, then composes them for more complex tasks.

Deciding what boxes are needed is challenging. A decomposition of a task into finite boxes restricts the possible choices that a system has. By using a decomposition, the gain

is simplification, and the loss is that we restrict the system from doing certain things. For example, for performing tasks in clutter, if we only have boxes for push grasps and some simple transitions, then we eliminate the possibility of doing something entirely different that might be useful, such as throwing an object. Also, boxes may be myopic, making it difficult to do error attribution and recovery.

4.2.3 Group C

Can we decompose multimodal manipulation under uncertainty into I/O boxes? If so, what would these be?

There are results from neuroscience indicating a strong modularity in the human sensorimotor areas. E.g. the neural signals related to individual fingers are independent. I.e., these signals look the same if a person hold a cup with her own e.g. 2 fingers vs. 2 persons with a finger each. So there is some evidence from the biological side for modularity.

The next question then is, should we think in terms of boxes (and their connections) or always consider loops including feedback signals as the basic building block? The latter would seem a bit like the subsumption architecture, which worked fine for simple, reflex behaviour based agents, but fell short for more complex behaviour requiring a larger planning horizon. We do need hierarchies, with levels of abstraction to keep the planning horizon small. One problem with such is that they can be too restrictive. The designer of such a hierarchy might have been thinking about a certain set of problems, for which the particular hierarchy seemed a "natural" way to organise the various modules/boxes. But this might prevent the system from solving slightly different problems. Is there even a necessary, "natural" modularity? Different types (regarding time spans, horizons) of tasks require different representations and thus *levels*. Differences are 2D/3D, work space/task space, short/long horizon, fast/slow.

So as long as there is no one, obvious architecture/hierarchy/nodularity we will have several competing propositions. How can we make and measure methodological progress? I.e., how do we know whether a certain architecture is better than another? This is very difficult to ascertain. Also because one can not simply connect boxes in a different way to obtain a different architecture. There is an inherent connection between boxes and architectures. A decision for a specific (type of) box is already a decision about the architecture in which the box is going to live.

One pragmatic option is to take stock of what functionalities/modules/boxes are available and selecting a path connecting all the modules necessary to solve a concrete problem. What follows is an exemplary (of course incomplete) list of modules:

- Segmentation: in: RGBD; out: segments
- Recognition: RGBD, model; out: instance, pose
- Classification: in RGBD, model, priors (segments); out: classified objects
- Tracking: in: RGBD, model, prev. pose; out: next pose
- Reinforcement learning: any observation, hypothesis class; out: actions
- Supervised learning: in: D, y; out: y
- Unsupervised learning: in: D; out: f(D)
- Object state filter: in: seq. of obj detections; out: set of obj. hyp.
- Metric SLAM: in: RGBD, laser; out: occupancy grid
- Motion planning: model of world, model of robot; out: path
- Trajectory Controller: in: goal x; out: control signal
- Grasp generator R: in: model, RGBD, target object category; out: hand pose
- Grasp generator M: in: model, RGBD, grasp category; out: trajectory

- Grasp generator J: in: model, RGBD; out: hand pose
- Grasp controller: in: joint angles, torques, tactile readings; out: joint torques
- Non-prehensile single finger manipulation: in: tactile readings, joint torques; out: joint velocities

For all these modules it is critical to make all implicit assumptions which the designer made very explicit. Otherwise, connecting a set of modules will lead to a very brittle system, failing for no obvious reason as soon as any assumption is violated.

Is there a master algorithm or are there infinitely many algorithms that solve specialized problems?

So having subscribed to the idea that there are many boxes, connected within this or that architecture to form a complete system, the question is: Do we have many similar boxes (one for grasping door handles, one for grasping mugs, one for grasping cloth, ...), or is each box very general (a master type algorithm)? In the former case one would take whatever boxes available and use a good architecture to connect them smartly. In the latter case one would first aim for universally usable boxes, that always work (independent of task), and then build an architecture around these.

Not that we do not necessarily have to aim for one single algorithm, but hopefully we can formulate one framework (general unified strategy), which can then be specialized for specific tasks. Note that the SLAM community seems to have converged to one such well-understood framework. Can we hope for something similar regarding manipulation?

4.2.4 Group D

Is there a master algorithm or are there infinitely many algorithms that solve specialized problems?

There is not enough evidence to answer this question. While human studies would seem to suggest that a master algorithm can exist, the conclusion is far from irrefutable, and one can easily veer into "pop psychology" when trying to answer the question. More clear seems to be the fact that, so far, the robotics field has had success by using many algorithms, loosely following the "boxes" paradigm. Examples are too numerous to build complete lists.

Can we decompose multimodal manipulation under uncertainty into I/O boxes? If so, what would these be?

As mentioned above, this approach has served the field well so far. Still, many caveats must be mentioned. It is important not to confuse computational modularity with task-space modularity (in other words, it's one thing for the APIs of two modules to match, but do the assumptions they make about the environment also match?). Furthermore, independent design of modules that are meant to be combined later is a difficult undertaking. One must remember that it is possible to achieve a bad design when using good modules. It is also important to remember that interconnections matter. For example, it is a topic of discussion whether perception and action can be separated into two boxes. However, even if the answer is affirmative, the connection between such boxes can not be a single point in state space (i.e. perception uniquely identifies the state of the world, then passes it on to action)!

4.3 Protocols and Benchmarks

4.3.1 Perception

We began by identifying existing benchmarking and challenges in the computer vision community: ICCV Workshop on Recovering 6D Object Pose, and the Visual Object Tracking Challenge. We then discussed similar efforts in the robotics community: YCB Object Set, 3D object databases, and the Amazon Picking Challenge. The sharing of datasets was briefly discussed. While camera images are easily shared and pooled, it was less clear how tactile datasets could be shared. Should action information, such as force and motion time histories, be provided along with tactile data?

We discussed the feasibility of experiments conducted over the internet. It was recommended that the robotics community appeal to federal institutions (e.g. NIST) and/or companies (e.g. KUKA, ABB) to house, host, and maintain hardware for communal benchmarking use. It was unclear which hardware would be appropriate and whether the effects of the hardware could be eliminated. One idea was to initially have two pathways and see which (or both) gain traction with the robotics community. Pathway 1 would use a single set of universal hardware for communal use. Pathway 2 would use a single task and no constraints on hardware selection. If some groups wanted to focus on one aspect of a manipulation experiments could be designed based on assumed modules of capability. For example, a specific grasp planner would be provided and researchers would test their ability to provide accurate inputs (e.g. 6D object pose) to the grasp planner module. We discussed how performance could be evaluated. Points were deemed to be too arbitrary. The time to completion was identified one objective measurement. Statistics could be used to report accuracy and frequency of success. It was unclear when and how often to assess performance during a single task.

We discussed the possible outputs of a perception system. For instance, should the perception system output 3D coordinates, a grasp pre-shape, a delta change in movement, etc.? The exact output would likely depend on the hardware, approach, and whether the system is open- or closed-loop. We discussed how to benchmark tactile sensing-related tasks, and whether it would be premature to do so. Care should be taken to prevent the benchmarking effort from turning into a test of the sensor technology itself. A benchmarking protocol would need to focus on perceptual information and not raw sensor data. Protocols could be designed that purposely occlude or disallow computer vision. Protocols could test the ability to perform tasks that require detection of discrete events (e.g. slip), and/or the ability to performance tasks that require tracking of forces. Tactile object recognition was one suggested task.

We identified two candidate benchmarking tasks:

1. Picking up an object from a scene and dropping in a bin
   - Input: Object to be selected, Point cloud
   - Output: 6D object pose in space, Gripper pose with respect to object
   - Constraints: fixed hardware, fixed planning, fixed grasping algorithm
   - Evaluation metric: Grasp and pick success
2. Placing object
   - Input: Point cloud
   - Output: 6D object pose in space, Gripper pose with respect to object
   - Evaluation metric: Placement accuracy, success

4.3.2 Planning

Planning benchmarks can defined in terms of complexity along the following axes:
- the type of *robot*: one arm vs. multiple arms,
- the type of *interaction*: basic pick-and-place vs. dextrous manipulation,
- the types of *objects* being manipulated: rigid, articulated, deformable, or a combination,
- the type of *uncertainty* (in action and perception): none, corrected/bounded by special moves/controllers, explicit planning in belief space.

Orthogonal to a categorization of planning benchmarks along axes of complexity, we can also define a taxonomy of robot planning capabilities. Planning algorithms will typically focus only on a subset of such capabilities. Examples of such capabilities include:
- collision-free point to point planning,
- task planning, assembly, and object rearrangement,
- grasping,
- bimanual manipulation: coordinated motion (closed kinematic chains), concurrent motion,
- manipulating deformable objects and articulated objects,
- planning in dynamic (i.e., time-varying) scenes,
- planning dynamic motions (e.g., throwing, waiter-like motion for carrying a tray),
- handover of objects (to a human),
- object search (necessarily includes sensing),
- planning for tasks that require tool use.

Considerations for a useful benchmark include that it should be predictive of how well a planning problem would work on other, similar problems. The benchmark should be well-defined so that scores can be compared in a meaningful way across contributors. On the other hand, it should be made difficult for a hacky, special-purpose planner to do well. One way to do this is introduce some randomization in the benchmark's initial conditions and report average score across a number of trials. Simulation can also be a useful to evaluate parameter / pose sensitivity of planner for a given benchmark.

We arrived at the following three benchmark concepts:
1. **Page turning.** This would test the ability to manipulate deformable objects. Performance can be measured in terms of number of pages turned within some time.
2. **Object search on a shelf.** Here we can define several variants. For example, the initial object arrangements could be well-defined, but not be given to the robot. This means that the robot needs to have some exploration strategy. In other words, it would also test perception. In simulation, perfect perception can be faked to measure just the effectiveness of the planner. Alternatively, the exact location of each object can be specified and the task would "simply" be to remove one of the objects, which would entail having to move other objects out of the way. By placing the objects in a cluttered environment (e.g., by placing objects within a small cubby), the problem can be made very hard.
3. **Assembly of Duplo bricks.** This would benchmark task planning, pick-and-place planning, compliant motion, and force control in mating the Duplo pieces. The problem can be made even harder by requiring bimanual in-hand assembly (i.e., using one hand to pick up pieces while the other hand holds the partial assembly). The problem can be made slightly easier by specifying the task plan. The initial placement of the Duplo pieces can be random (to avoid hardcoding the moves) or explicitly defined. The desired assemblies can be divided into easy/medium/hard categories. The initial arrangement blocks can be made easy (all blocks right side up) or hard (blocks sideways or upside down).

5 Open Problems

The group identified several open problems:

1. **Benchmarks for robotic manipulation:** A key challenge is in developing benchmark objects and protocols that a) are of interest to all communities (planning, perception, control, learning), b) generalizable across robot platforms, and c) accessible so that every group can try them out.
2. **Developing building blocks:** Manipulation requires the tight integration of several components. A key challenge is in developing an open-source architecture that enables several groups to contribute specific modules they are expert in, and harness modules written by other groups.
3. **Perception for manipulation:** The perception community has several excellent benchmarks for computer vision. However, manipulation has specific demands: dealing with clutter, and outputting the 6D pose of objects, among others. Many manipulation systems suffer from poor perception and an open challenge is in harnessing the perception community to address the specific demands of manipulation.

6 Panel Discussions

Our seminar had several interesting discussions (detailed above) in smaller groups. We found this to be more useful because a) it allowed everyone to have a greater opportunity to speak (as our total group size was large), and b) it allowed much more focussed discussions and deep dives on a particular topic.

Participants

- Ron Alterovitz
University of North Carolina at Chapel Hill, US
- Brenna D. Argall
Northwestern University – Evanston, US
- Yasemin Bekiroglu
KTH Royal Institute of Technology – Stockholm, SE
- Kostas Bekris
Rutgers Univ. – Piscataway, US
- Dmitry Berenson
Worcester Polytechnic Inst., US
- Bastian Bischoff
Robert Bosch GmbH – Stuttgart, DE
- Jeannette Bohg
MPI für Intelligente Systeme – Tübingen, DE
- Oliver Brock
TU Berlin, DE
- Matei Ciocarlie
Columbia University, US
- Fan Dai
ABB Corporate Research, DE
- Renaud Detry
University of Liège, BE
- Mehmet R. Dogar
University of Leeds, GB
- Aaron M. Dollar
Yale University, US
- Roderic A. Grupen
University of Massachusetts – Amherst, US
- Simon Hangl
Universität Innsbruck, AT
- David Hsu
National Univ. of Singapore, SG
- Leslie Pack Kaelbling
MIT – Cambridge, US
- Marek S. Kopicki
University of Birmingham, GB
- Dirk Kraft
University of Southern Denmark – Odense, DK
- Norbert Krüger
University of Southern Denmark – Odense, DK
- Ville Kyrki
Aalto University, FI
- Ales Leonardis
University of Birmingham, GB
- Shuai Li
Rensselaer Polytechnic, US
- Maxim Likhachev
Carnegie Mellon University, US
- Tomás Lozano-Pérez
MIT – Cambridge, US
- Matthew T. Mason
Carnegie Mellon University – Pittsburgh, US
- Mark Moll
Rice University – Houston, US
- Duy Nguyen-Tuong
Robert Bosch GmbH – Schwieberdingen, DE
- Erhan Öztop
Özyegin Univ. – Istanbul, TR
- Jan Peters
TU Darmstadt, DE
- Justus Piater
Universität Innsbruck, AT
- Robert Platt
Northeastern University – Boston, US
- Maximo A. Roa
German Aerospace Center-DLR, DE
- Veronica Santos
University of California – Los Angeles, US
- Siddhartha Srinivasa
Carnegie Mellon University, US
- Ales Ude
Jozef Stefan Institute – Ljubljana, SI
- Francisco Valero Cuevas
USC – Los Angeles, US
- Jeremy L. Wyatt
University of Birmingham, GB
- Michael Zillich
TU Wien, AT