Interactive Perception: Leveraging Action in Perception and Perception in Action

Jeannette Bohg, *Member, IEEE*, Karol Hausman, *Student Member, IEEE*,
Bharath Sankaran, *Student Member, IEEE*, Oliver Brock, *Senior Member, IEEE*, Danica Kragic, *Fellow, IEEE*,
Stefan Schaal, *Fellow, IEEE*, and Gaurav S. Sukhatme, *Fellow, IEEE*

***Abstract*—Recent approaches in robot perception follow the insight that perception is facilitated by interaction with the environment. These approaches are subsumed under the term Interactive Perception (IP). This view of perception provides the following benefits. First, interaction with the environment creates a rich sensory signal that would otherwise not be present. Second, knowledge of the regularity in the combined space of sensory data and action parameters facilitates the prediction and interpretation of the sensory signal. In this survey, we postulate this as a principle for robot perception and collect evidence in its support by analyzing and categorizing existing work in this area. We also provide an overview of the most important applications of IP. We close this survey by discussing remaining open questions. With this survey, we hope to help define the field of Interactive Perception and to provide a valuable resource for future research.**

***Index Terms*—Autonomous systems, cognitive robotics, robot learning, robot vision systems.**

I. Introduction

There is compelling evidence that perception in humans and animals is an active and exploratory process [1]–[3]. Even the most basic categories of biological vision seem to be based on active visual exploration, rather than on the analysis of static image content. For example, Nöe [3] argues that the visual category *circle* or *round* cannot be based on the direct perception of a circle, as 1) we rarely look at round objects from directly above and 2) the projection of a circle onto our retina is not a circle at all. Instead, we perceive circles by the way their projection changes in response to eye movements.

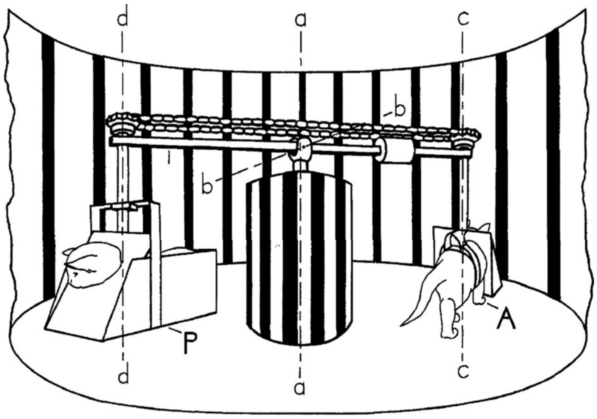


Fig. 1. Mechanical system where movement of Kitten A is replicated onto Kitten P. Both kittens receive the same visual stimuli. Kitten A controls the motion, i.e., it is *active*. Kitten P is moved by Kitten A, i.e., it is *passive*. Only the active kittens developed meaningful visually guided behavior that was tested in separate tasks. Figure adapted from [4].

Held and Hein [4] analyzed the development of visually guided behavior in kittens. They found that this development critically depends on the opportunity to learn the relationship between self-produced movement and concurrent visual feedback. The authors conducted an experiment with kittens that were only exposed to daylight when placed in the carousel depicted in Fig. 1. Through this mechanism, the *active* kittens (A) transferred their own deliberate motion to the *passive* kittens (P) that were sitting in a basket. Although both types of kittens received the same visual stimuli, only the active kittens showed meaningful visually guided behavior in test situations.

Gibson [5] showed that the physical interaction further augments perceptual processing beyond what can be achieved by deliberate pose changes. In the specific experiment, human subjects had to find a reference object among a set of irregularly shaped, three-dimensional (3-D) objects (see Fig. 2). They achieved an average accuracy of 49% if these objects were shown in a single image. This accuracy increased to 72% when subjects viewed rotating versions of the objects. They achieved

Manuscript received February 20, 2017; accepted June 9, 2017. Date of publication August 11, 2017; date of current version December 14, 2017. This paper was recommended for publication by Associate Editor J. Piater and Editor C. Torras upon evaluation of the reviewers' comments. This work was supported in part by the National Science Foundation under Grant IIS-1205249, Grant IIS-1017134, and Grant EECS-0926052; in part by the Office of Naval Research; in part by the Okawa Foundation; in part by the Max Planck Society; in part by the German Science Foundation (DFG) under Grant BR 2248/3-1; and in part by the European Commission under Grant H2020-ICT-645599 on Soft Manipulation (Soma). *(Jeannette Bohg, Karol Hausman, and Bharath Sankaran contributed equally to this work.) (Corresponding author: Jeannette Bohg.)*

J. Bohg is with the Autonomous Motion Department, Max Planck Institute for Intelligent Systems, Tübingen 72076, Germany, and also with the with the Department of Computer Science, Stanford University, Stanford, CA 94305 USA (e-mail: jeannette.bohg@tuebingen.mpg.de).

K. Hausman, B. Sankaran, and G. S. Sukhatme are with the Department of Computer Science, University of Southern California, Los Angeles, CA 90089 USA (e-mail: khausman@usc.edu; bsankara@usc.edu; gsukhatme@usc.edu).

O. Brock is with the Robotics and Biology Laboratory, Technische Universität Berlin, Berlin 10623, Germany (e-mail: oliver.brock@tu-berlin.de).

D. Kragic is with the Centre for Autonomous Systems, Computational Vision and Active Perception Lab, Royal Institute for Technology (KTH), Stockholm 114 28, Sweden (e-mail: dani@kth.de).

S. Schaal is with the Department of Computer Science, University of Southern California, Los Angeles, CA 90089 USA, and also with the Autonomous Motion Department, Max Planck Institute for Intelligent Systems, Tübingen 72076, Germany (e-mail: sschaal@usc.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TRO.2017.2721939

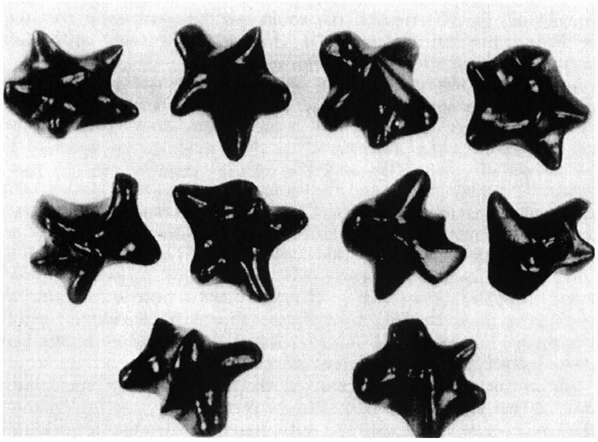


Fig. 2. Set of irregularly shaped objects among which subjects had to find a reference object. Subjects achieved near perfect performance when they could touch and rotate these objects as opposed to just looking at them in a static pose. Figure adapted from [5, p. 124] with permission.

nearly perfect performance (99%) when touching and rotating the objects in their hands.

These three examples illustrate that biological perception and perceptually guided behavior intrinsically rely on active exploration and knowledge of the relation between action and sensory response. This contradicts our introspection, as we just seem to passively *see*. In reality, visual perception is similar to haptic exploration. "*Vision is touch-like*" [3, p. 73] in that, perceptual content is not given to the observer all at once but only through skillful, active *looking*.

This stands in stark contrast to how perception problems are commonly framed in machine vision. Often, the aim is to semantically annotate a single image while relying on a minimum set of assumptions or prior knowledge. These requirements render the considered perception problems underconstrained and thereby make them very hard to solve.

The most successful approaches learn models from datasets that contain hundreds of thousands of semantically annotated *static* images, such as Pascal VOC [6], ImageNet [7], or Microsoft COCO [8]. Recently, deep learning based approaches led to substantial progress by being able to leverage these large amounts of training data. In these methods, data points provide the most important source of constraints to find a suitable solution to the considered perception problem. The success of these methods over more traditional approaches suggests that previously considered assumptions and prior knowledge did not correctly or sufficiently constrain the solution space.

Different from disembodied Computer Vision algorithms, robots are embodied agents that can move within the environment and physically interact with it. Similar to biological systems, this creates rich and more informative sensory signals that are concurrent with the actions and would otherwise not be present. There is a regular relationship between actions and their sensory response. This regularity provides the additional constraints that simplify the prediction and interpretation of these high-dimensional signals. Therefore, robots should exploit any knowledge of this regularity. Such an integrated approach to perception and action may reduce the requirement of large amounts of data and thereby provide a viable alternative to the current data-intensive approaches toward machine perception.

II. Interactive Perception

Recent approaches in robot perception are subsumed by the term *Interactive Perception* (IP). They exploit any kind of forceful interaction with the environment to simplify and enhance perception. Thereby, they enable robust perceptually guided manipulation behaviors. IP has two benefits. First, physical interaction creates a novel sensory signal that would otherwise not be present. Second, by exploiting knowledge of the regularity in the combined space of sensory data and action parameters, the prediction and interpretation of this novel signal becomes simpler and more robust. In this section, we will define what we mean by forceful interaction. Furthermore, we explain the two postulated benefits of IP in more detail.

A. Forceful Interactions

Any action that exerts a potentially time-varying force upon the environment is a forceful interaction. A common way of creating such an interaction is through physical contact that may be established for the purpose of moving the agent (e.g., in legged or wheeled locomotion), for changing the environment (e.g., to open a door or pushing objects on a table out of the way), or for exploring environment properties while leaving it unchanged (e.g., by sliding along a surface to determine its material). It may also be a contact-free interaction that is due to gravitational or magnetic forces or even lift. An interaction may only be locally applied to the scene (e.g., through pushing or pulling a specific object) or it may affect the scene globally (e.g., shaking a tray with objects standing on it). This interaction can be performed either by the agent itself or by any other entity, e.g., a teacher to be imitated or someone who demonstrates an interaction through kinesthetic teaching.

In this survey, we are interested in approaches that go beyond the mere observation of the environment toward approaches that enable its *Perceptive Manipulation*.[1] Therefore, we focus on physical interactions for the purpose of changing the environment or for exploring environment properties while leaving it unchanged. We are not concerned with interactions for locomotion and environment mapping.

B. Benefits of IP

Create Novel Signals (CNS): Forceful interactions create novel, rich sensory signals that would otherwise not be present. These signals are beneficial for estimating the quantities that are relevant to manipulation problems, such as haptic, audio, and visual data correlated over time. Relevant quantities include object weight, surface material, or rigidity.

Action Perception Regularity (APR): Forceful interactions reveal regularities in the combined space ($S \times A \times t$) of sensor information (S) and action parameters (A) over time (t). This

[1] We consider *Perceptive Manipulation* to be the equivalent term to IP. This emphasizes the blurred boundary that is traditionally drawn between manipulation and perception.

	S	A		t
		F	$\neg F$	
Sensorless Manipulation	-	✓	-	✓
Perception of Images	✓	-	-	-
Perception of Video	✓	-	-	✓
Active Perception	✓	-	✓	✓
Active Haptic Perception, Interactive Perception	✓	✓	-	✓

Fig. 3. Summary of how IP relates to other perception approaches regarding $S \times A \times t$. F stands for forceful interaction and $\neg F$ for actions that only manipulate the parameters of the sensory apparatus and not the environment.

regularity is constituted by the repeatable, multimodal sensory data that is created when executing the same action in the same environment. Not considering the space of actions amounts to marginalizing over them. The corresponding sensory signals would then possess a significantly higher degree of variation compared to the case where the regularity in $S \times A \times t$ is taken into account. Therefore, despite $S \times A \times t$ being much higher dimensional, the signal represented in this space has more structure.

Using the Regularity: Knowing this regularity corresponds to understanding the causal relationship between action and sensory response given specific environment properties. Thereby, it allows us to

1) predict the sensory signal given knowledge about the action and environment properties;

2) update the knowledge about some latent properties of the environment by comparing the prediction to the observation; and

3) infer the action that has been applied to generate the observed sensory signal given some environment properties.

These capabilities simplify perception but also enable optimal action selection.

Learning the Regularity: Learning these regularities corresponds to identifying the causal relationship between action and sensory response. This requires information about the action that produced an observed sensory effect. If the robot autonomously interacts with the environment, this information is automatically available. Information about the action can also be provided by a human demonstrator.

III. Historical Perspective

In robotics, the research field of *Active Perception* (AP) pioneered the insight that perception is active and exploratory. In this section, we relate IP to AP. Additionally, we discuss the relation of IP to other perception approaches that neglect either the sensory or action space in $S \times A \times t$. Fig. 3 summarizes this section.

A. Sensorless Manipulation

This approach to perception does not require any sensing. It aims at finding a sequence of actions that brings the system of interest from an unknown into a defined state. Therefore, after performing these actions, the system state can be considered as perceived. This kind of sensorless manipulation was demonstrated first by Erdmann and Mason [9], who used it for orienting a planar part that is randomly dropped onto a tray. The goal of the proposed algorithm is to generate a sequence of tray tilting actions that uniquely moves the part into a goal orientation without receiving sensor feedback or knowing the initial state. It uses a simple mechanical model of sliding and information on how events like collisions with walls reduce the number of possible part orientations. More recently, Dogar *et al.* [10] extend this line of thought to grasping. The authors plan for the best push-grasp such that the object of interest has a high probability of moving into the gripper while other objects are pushed away. The plan is then executed open loop without taking feedback of the actual response of the environment into account.

We argue that IP critically depends on representing a signal in the combined space of sensory information and action parameters over time. Sensorless manipulation is similar in that it also requires a model of how actions funnel the uncertainty about the system state into a smaller region in state space. However, different from the approaches in this survey, it does not require sensory feedback as it assumes that the uncertainty can be reduced to the required amount only through the actions. For complex dynamical systems, this may not always be the case or a sufficiently expressive forward model may not be available.

B. Perception of Visual Data

The vast research area of Computer Vision focuses on interpreting static images, video, or other visual data. The majority of approaches completely neglect the active and exploratory nature of human and robot perception. Nevertheless, there are examples in the Computer Vision literature that show how exploiting the regularity in $S \times A \times t$ simplifies perception problems. The first example aims at human activity recognition in video. It is somewhat obvious that this task becomes easier when observing the activity over a certain course of time. Less obvious is the result by Kjellström *et al.* [11] who showed that classifying objects is easier if they are observed while being used by a person. More recently, Cai *et al.* [12] support these results. They show that recognizing manipulation actions in single images is much easier when modeling the associated grasp and object type in a unified model.

Another example considers the problem of image restoration. Xue *et al.* [13] exploit whole image sequences to separate obstructing foreground like fences or window reflections from the main subject of the images, i.e., the background. This would be a very hard problem if only a single image were given or without the prior knowledge of the relation between optical flow and depth.

Aloimonos *et al.* [14] show how challenging vision problems, such as shape from shading or structure from motion, are easier to solve with an active than a passive observer. Given known camera motion and associated images, the particular problem can be formulated such that it has a unique solution and is linear. The case of the passive observer usually requires additional assumptions or regularization and sometimes nonlinear optimization.

C. Active Perception

In 1988, Bajcsy [15] introduced AP as the problem of intelligent control strategies applied to the data acquisition process. Ballard [16] and Aloimonos *et al.* [14] further analyzed this concept for the particular modality of vision. In this context, researchers developed artificial vision systems with many degrees of freedom [17]–[19] and models of visual attention [20], [21] that these active vision systems could use for guiding their gaze.

Recently, Bajcsy *et al.* [22] revisited AP giving an excellent historical perspective on the field and a new, broader definition of an active perceiver based on decades of research:

> *An agent is an active perceiver if it knows why it wishes to sense, and then chooses what to perceive, and determines how, when and where to achieve that perception.*

The authors identify the *why* as the central and distinguishing component to a passive observer. It requires the agent to reason about so called *Expectation-Action* tuples to select the next best action. The expected result of the action can be confirmed by its execution. Expectation-Action tuples capture the predictive power of the regularity in $S \times A \times t$ to enable optimal action selection.

1) Relation to IP: The new definition of AP is not only restricted to vision. However, the majority of approaches gathered under the term of AP consider vision as the sole modality and the manipulation of extrinsic or intrinsic camera parameters as possible actions. This is also reflected by the choice of examples in [22]. The focus on the visual sense has several implications for AP in relation to IP. First, an active perceiver with the ability to move creates a richer and more informative visual signal (e.g., from multiple viewpoints or when zooming) that would otherwise not be present. However, this may not provide all relevant information, especially not those required for manipulation problems. Natale *et al.* [23] emphasize that only through physical interaction, a robot can access object properties that otherwise would not be available (like weight, roughness, or softness).

Second, as shown in [14], we have very good understanding of multiview and perspective geometry that can be leveraged to formulate a vision problem in such a way that its solution is simple and tractable. However, when it comes to predicting the effect of physical interaction that does not only change the viewpoint of the agent on the environment, but the environment itself, we are yet to develop rich, expressive, and tractable models.

Finally, AP mainly focuses on simplifying challenging perception problems. However, a robot should also be able to manipulate the environment in a goal-directed manner. Sandini *et al.* [24, p.167] formulate this as a difference in how visual information is used: In AP, it is mainly devoted to exploration of the environment, whereas in IP, it is also used to monitor the execution of motor actions.

2) Early Examples of IP: There are a number of early approaches within the area of AP that exploit forceful interaction with the environment and are therefore early examples for IP approaches. Tsikos and Bajcsy [25], [26] propose to use a robot arm to make the scene simpler for the vision system through actions like pick, push, and shake. The specific scenario is the separation of random heaps of objects into sets of similar shapes. Bajcsy [27] and Bajcsy and Sinha [28] propose the *Looker and Feeler* system that allows to perform material recognition of potential footholds for legged locomotion. The authors hand-design specific exploration procedures of which the robot observes the outcome (visually or haptically) to determine material attributes. Salganicoff and Bajcsy [29] show how the mapping between observed attributes, actions, and rewards can be learned from training data gathered during real executions of a task. Sandini *et al.* [24, Sec. 3] propose to use optical flow analysis of the object motion while it is being pushed. The authors show that through this analysis, they can retrieve both geometrical and physical object properties that can then be used to adapt the action.

D. Active Haptic Perception

Haptic exploration of the environment relies on haptic sensing that requires contact with the environment. Interpretation of a sequence of such observations is part of IP as it requires a forceful and time-varying interaction. The interpretation of an isolated haptic *frame* without temporal information is similar to approaches in Computer Vision, such as semantic scene understanding from static images [30].

Early approaches that use touch in an active manner are applied to problems, such as reconstructing shape from touch [31], recognizing objects through tracing their surface [32], or exploring texture and material properties [31]. The complementary nature of vision and touch has been explored by Allen and Bajscy [33] in reconstructing a 3-D object shape. A more complete review of these early approaches toward active haptic perception is contained in [22] and [30].

More recent examples include [23] to learn haptic object representations, [34]–[36] for object detection and pose estimation, [37]–[39] for reconstructing the shape of objects or the environment, and [40]–[42] for texture classification or description. The most apparent difference of these recent approaches to earlier work lies in the use of machine learning techniques to either automatically find suitable exploration strategies, to learn suitable feature representations or to better estimate different quantities.

In general, active haptic perception requires deliberate contact interaction but the majority of the cases do not aim at changing the environment. Instead, for simplification, objects or the environment are often assumed to be rigid and static during contact.

IV. Applications of IP

IP methods may be applied to achieve an estimation or a manipulation goal. Currently, the vast majority of IP approaches estimate some quantity of interest through forceful interaction. Other IP approaches pursue either a grasping or manipulation goal. This means that they aim to manipulate the environment to bring it into a desired state. Usually, this includes the estimation of quantities that are relevant to the manipulation task.

Existing IP approaches can be broadly grouped into ten major application areas, as visualized in Fig. 4. In this section, we briefly describe each of these areas. For the first three applica-

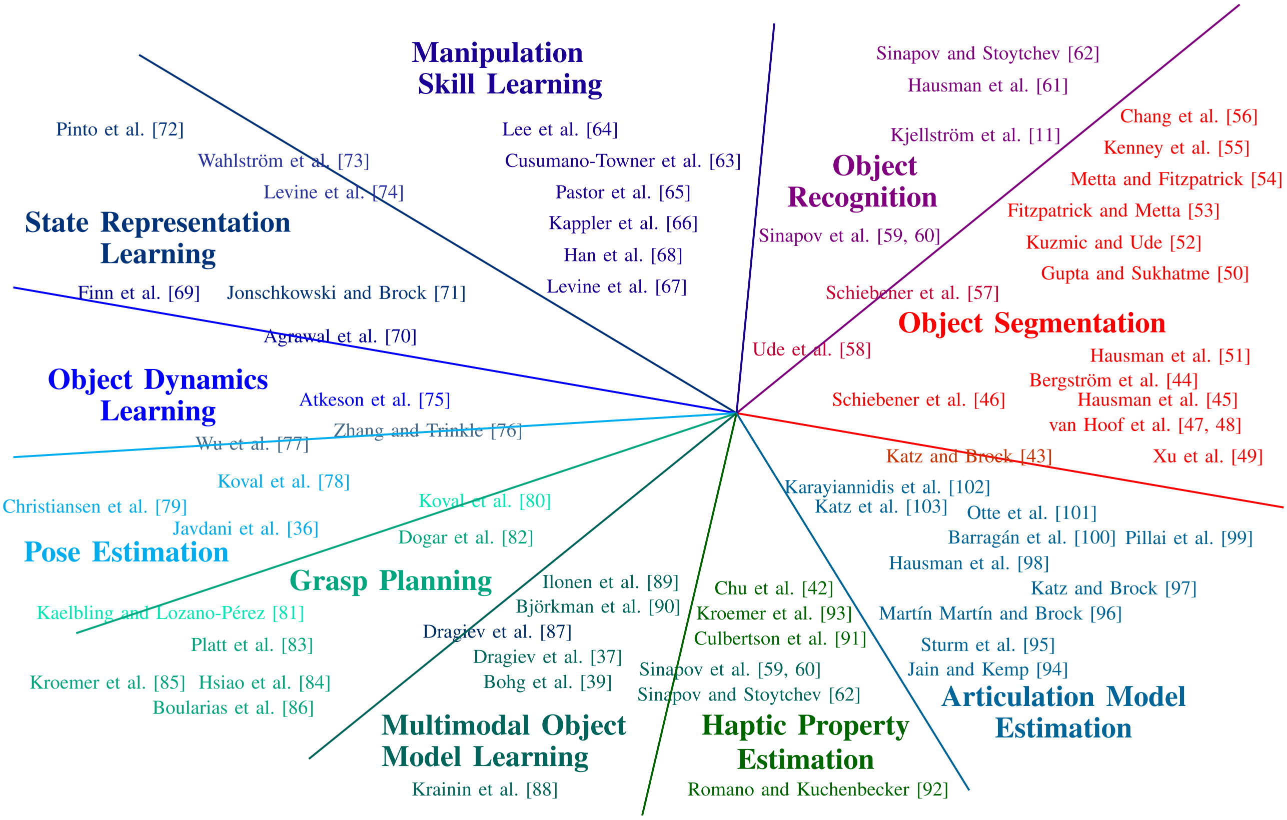


Fig. 4. Paper categorization based on application areas. Papers that address multiple application areas lie on the boundary between those applications, e.g., [57] and [81].

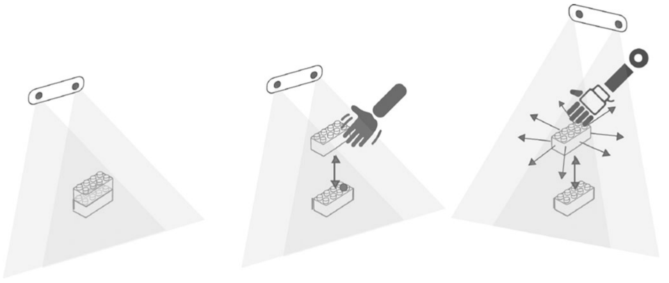


Fig. 5. Three situations in which a robot (indicated by the stereo camera and viewing cones) tries to estimate the articulation model of two Lego blocks on a table. The situations differ in the amount of information the robot has access to. [Left] The robot can only change the viewpoint to obtain more information. [Center] The robot can observe a rich sensory signal caused by a person lifting the top Lego block. [Right] The robot can interact with the scene and observe the resulting sensory signal. Therefore, it has more information about the *specific* interaction. Only in the rightmost situation, the articulation model can be reliably identified.

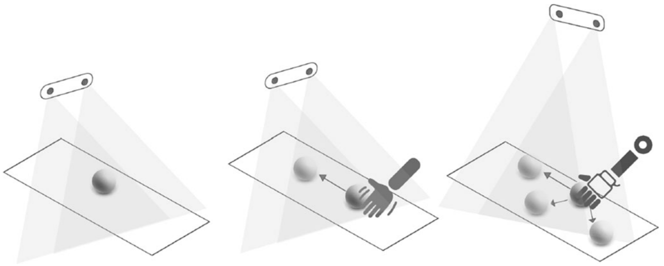


Fig. 6. Three situations in which a robot (indicated by the stereo camera and viewing cones) tries to estimate the weight of a sphere. The situations differ in the amount of information the robot has access to. [Left] The robot can only change the viewpoint to obtain more information. [Center] The robot can observe a rich sensory signal caused by a person pushing the sphere. [Right] The robot can push the sphere itself and observe the resulting sensory signal, i.e., where the sphere comes to rest. In the last situation, it has more information about the *specific* push force. Only in the rightmost situation, the weight of the sphere can be reliably unidentified.

tions (object segmentation, articulation model estimation, and object dynamics learning), we use a couple of simple examples (see Figs. 5 and 6) to allow the reader to better appreciate the benefits of IP and understand its distinction to AP.

A. *Object Segmentation*

Object segmentation is a difficult problem and, in the area of Computer Vision, it is often performed on single images [104]–[106]. To illustrate the challenges, consider the simple example scenario depicted in Fig. 5. Two Lego blocks are firmly attached to the table. The robot is supposed to estimate the number of objects on the table. When the robot is a passive observer of the scene as in Fig. 5 [Left], it would be very challenging to estimate the correct number of Lego blocks on the table without incorporating a lot of prior knowledge. The situation does not improve in this static scenario even with more sensory data from different viewpoints or after zooming in.

When the robot observes another agent interacting with the scene, as shown in Fig. 5 [Center], it will be able to segment the Lego blocks and correctly estimate the number of objects in the scene. This is an example of how forceful interactions can create rich sensory signals that would otherwise not be present (CNS). The new evidence in form of motion cues simplifies the problem of object segmentation.

The ability to interact with the scene allows a robot to also autonomously generate more informative sensory information, as visualized in Fig. 5 [Right]. Reasoning about the regularity in $S \times A \times t$ may lead to even better segmentation since the robot can select actions that are particularly well suited for reducing the segmentation uncertainty (APR).

For these reasons, object segmentation has become a very popular topic in IP. For example, Fitzpatrick and Metta [53] and Metta and Fitzpatrick [54] are able to segment the robot's arm and the objects that were moved in a scene. Gupta and Sukhatme [50] and Chang *et al.* [56] use predefined actions to segment objects in cluttered environments. van Hoof *et al.* [48] can probabilistically reason about optimal actions to segment a scene.

B. Articulation Model Estimation

Another problem that is simplified through IP is the estimation of object articulation mechanisms. The robot has to determine whether the relative movement of two objects is constrained or not. Furthermore, it has to understand whether this potential constraint is due to a prismatic or revolute articulation mechanism and what the pose of the joint axis is. Fig. 5 [Left] visualizes an example situation in which the robot has to estimate the potential articulation mechanism between two Lego blocks, given only visual observations of a static scene. This is almost impossible to estimate from single images without including a lot of prior semantic knowledge. It is also worth noting that this situation is not improved if gathering more information from multiple viewpoints of this otherwise static scene.

In Fig. 5 [Center], the robot observes an agent lifting the top-most lego block. This is another example of how forceful interactions create a novel, informative sensory signal (CNS). In this case, it is a straight-line, vertical motion of one Lego block. It provides evidence in favor of a prismatic joint in between these two objects (although, in this case, this is still incorrect).

When the robot autonomously interacts with the scene, it creates these informative sensory signals not only in the visual but also haptic sensory modality. These data are strongly correlated with a particular articulation mechanism. Fig. 5 [Right] visualizes this scenario. By leveraging knowledge of the regularity in $S \times A \times t$, the robot can also form a correct hypothesis of the articulation model (APR). The Lego blocks are rigidly attached at first, but when the robot applies enough vertical force to the top-most Lego block, there is a sensory evidence for a free body articulation model.

In the literature, there are offline estimation approaches toward this problem that either rely on fiducial markers [95] or markerless tracking [99], [103]. There are also online approaches [96] where the model is estimated during the movement. Most recent methods include reasoning about actions to actively reduce the uncertainty in the articulation model estimates [98], [100], [101].

C. Object Dynamics Learning and Haptic Property Estimation

IP has also made major inroads into the challenge of estimating haptic and inertial properties of objects. Fig. 6 shows a simple example scenario that shall serve to illustrate why IP simplifies the problem. Consider a sphere that is lying on a table. The robot is supposed to estimate the weight of the sphere given different sources of information. We assume that the robot knows the relationship between push force, distance the sphere traveled, and sphere weight. In the trivial static scene scenario illustrated in Fig. 6 [Left], the robot is not able to estimate any of the inertial properties. It encounters similar problems as in the previous example (see Fig. 5) even if it was able to change the viewpoint.

In Fig. 6 [Center], the robot can observe the motion of the sphere that is pushed by a person. Now, the robot can easily segment the ball from the table due to the additional sensory signal that was not present before (CNS). However, it remains very difficult for the robot to estimate the inertial properties of the sphere because it does not know the strength of the push. Without this information, the known regularity in $S \times A \times t$ cannot be exploited. The robot will only be able to obtain a very uncertain estimate of the sphere weight because it needs to marginalize over all the possible forces the person may have applied.

In Fig. 6 [Right], the robot interacts with the sphere. It can control the push force that is applied and observe the resulting distance at which the sphere comes to rest. Given the knowledge of the strength of the push, it can now exploit the known associations between actions and sensory responses to estimate the spheres inertial properties (APR).

There are several examples that leverage the insight that IP enables the estimation of haptic and inertial properties. For example, in [92] and [42], it has been shown that surface and material properties of objects can be more accurately estimated if the robot's haptic sensor is moved along the surface of the object.

Atkeson *et al.* [75] and Zhang and Trinkle [76] move the object to estimate its inertial properties or other parameters of object dynamics, which are otherwise unobservable.

D. Object Recognition or Categorization

Approaches to detect object instances or objects of a specific category have to learn the appearance or shape of these objects under various conditions. There are many challenges in object recognition or categorization that make this task very difficult given only a single input image. A method has to cope with occlusions, different lighting conditions, scale of the images, just to name a few. State-of-the-art approaches in Computer Vision, for e.g., [107] and [108], require enormous amounts of training data to handle these variations.

IP approaches allow a robot to move objects and hence reveal previously hidden features. Thereby, it can resolve some of the aforementioned challenges autonomously and may alleviate the need for enormous amounts of training data. Example approaches that perform object segmentation and categorization can be found in [46] and [57]. The challenge of object recognition/categorization has been tackled by Sinapov *et al.* [59] and Hausman *et al.* [61].

E. Multimodal Object Model Learning

Learning models of rigid objects (RO), articulated objects (AO), and deformable objects (DO) is a central problem in the area of Computer Vision. In the majority of the cases, the model is learned or built from multiple images of the same object or category of objects. Once the model is learned, it can be used to find the object in new, previously unseen contexts.

A robot can generate the necessary data through interaction with the environment. For example, Krainin *et al.* [88] present an approach where a robot autonomously builds an object model while holding the object in its hand. The object model is completed by executing actions informed by next best view planning. Kenney *et al.* [55] push an object on the plane and accumulate visual data to build a model of the object.

There are also approaches that build an object model from haptic sensory data, for e.g., Dragiev *et al.* [37]. Allen and Bajscy [33], Bohg *et al.* [39], Ilonen *et al.* [89], and Björkman *et al.* [90] show examples that initialize a model from visual data and then further augment it with tactile data. Sinapov *et al.* [59] present a method where a robot grasps, lifts, and shakes objects to build a multimodal object model.

F. Object Pose Estimation

IP has also been applied to the problem of object pose estimation. Related approaches focus on reducing object pose uncertainty by either touching or moving it.

Koval *et al.* [78] employ manifold particle filters for this purpose. Javdani *et al.* [36] use information-theoretic criteria, such as information gain, to actively reduce the uncertainty of the object pose. In addition to reducing uncertainty, they also provide optimality guarantees for their policy.

G. Grasp Planning

Cluttered scenes and premature object interactions used to be considered as obstacles for grasp planning that had to be avoided by all means. In contrast, IP approaches in this domain take advantage of the robot's ability to move objects out of the way or to explore them to create more successful plans even in clutter or under partial information.

Hsiao *et al.* [84] use proximity sensors to estimate the local surface orientation to select a good grasp. Dragiev *et al.* [87] devise a grasp controller for objects of unknown shape, which combines both exploration and exploitation actions. Object shape is represented by a Gaussian process implicit surface. Exploration of the shape is performed using tactile sensors on the robot hand. Once the object model is sufficiently well known, the hand does not prematurely collide with the real object during grasping attempts.

H. Manipulation Skill Learning

In some cases, the goal of IP is to accomplish a particular manipulation skill. This manipulation skill is generally a combination of some of the prespecified goals.

To learn a manipulation skill, Pastor *et al.* [65] and Kappler *et al.* [66] represent the task as a sequence of demonstrated behaviors encoded in a manipulation graph. This graph provides a strong prior on how the actions should be sequenced to accomplish the task. Lee *et al.* [64] use a set of kinesthetic demonstrations to learn the right variable-impedance control strategy. Towner *et al.* [63] propose a planning approach that uses a previously learned Hidden Markov Model to fold clothes.

The approaches discussed above can be thought of as methods that capture the regularity of complex manipulation behaviors in $S \times A \times t$ by learning them via demonstration.

I. State Representation Learning

In the majority of the IP approaches, the representation of sensory data and latent variables are prespecified based on prior knowledge about the system and task. There are, however, some approaches that learn these representations. Most notable are [71], [73], and [74]. In all of them, the authors learn some mapping from raw, high-dimensional sensory input (in this case images) to a lower dimensional state representation. All of these example approaches fix the structure of this mapping, e.g., linear mapping with task-specific regularizers [71] or *Convolutional Neural Networks* [73], [74]. The parameters of this mapping are learned from data.

V. Taxonomy of IP

In this section, we identify a number of important aspects that characterize existing IP approaches. These are additional to the two benefits of CNS and APR and independent of the specific application of an approach. We use these aspects to taxonomize and group approaches in Tables I and II. In the following, each table column is described in detail in a subsection along with example approaches. We use *paper sets* to refer to groups of similar approaches that address the same application, e.g., either object segmentation or manipulation skill learning. We split paper sets further into approaches that either exploit CNS or APR. We also list papers separately that do not pursue a unique goal, e.g., they perform both object segmentation and recognition.

A. How is the Signal in $S \times A \times t$ Leveraged?

An IP approach leverages at least one of the two aforementioned benefits: 1) it exploits a novel sensory signal that is due to some time-varying, forceful interaction (CNS) or 2) also leverages prior knowledge about the regularity in the combined space of sensory data and action parameters over time $S \times A \times t$ for predicting or interpreting this signal (APR).

1) Commonalities and Differences Between CNS and APR: Approaches that exploit the novel sensory signal (CNS) also rely on regularities in the sensory response to an interaction. In its most basic form, this regularity is usually linked to some

TABLE I
TAXONOMY OF IP APPROACHES—PART 1

Goals and Paper Set ID	Papers	How is the Signal in $S \times A \times t$ Leveraged? [a]	What Priors Are Employed? [b]	Does the Approach Perform Action Selection? [c]	What is the Objective: Perception, Manipulation, or Both?	Are multiple sensor modalities exploited? [d]	How is Uncertainty Modeled and Used? [e]
Object recognition	[11], [12], [61]	APR	RO [61], OD, AP	M [61], ✗[11], [12]	P [11], [61], [12]	✗(Vision)	No dynamics [12], SDM [11], [61], SOM [11], [61], [12], EU [11], [61]
Object Segmentation I	[44]–[46], [49]–[56], [109]	CNS	RO, PM [45], [46], [50]–[56], [109], OD [54], AP [44]–[46], [49]–[52], [56]	✗ [53], [55], M [44]–[46], [49]–[52], [54], [56], [109]	P [44]–[46], [49]–[56], M [109]	✗(Vision)	No dynamic model, DOM [45], [46], [50]–[54], [56], SOM [44], [49], [55], [109], EU [49], [109]
Object Segmentation II	[47], [48]	APR	AP	M	P	✗(Vision)	SDM, SOM, EU
Object Segmentation—Object Recognition I	[57]	CNS	RO, AP, PM	M	P	✗(Vision)	No dynamic model, DOM
Object Segmentation—Object Recognition II	[58]	APR	RO, AP	G	P	✗(Vision)	DDM, SOM, EU
Articulation Model Estimation—Object Segmentation	[43]	CNS	AO	✗	P	✗(Vision)	No dynamics, SOM
Articulation Model Estimation I	[95], [96], [99], [100], [103], [110]	CNS	AO	✗ [95], [96], [99], [103], [110], M [100]	P	✗(Vision)	No dynamics [103], SOM, SDM [95], [96], [99], [110], [100], EU [43], [95], [96], [99], [103], [110], [100]
Articulation Model Estimation II	[94], [97], [98], [101], [102]	APR	AO, PM [97]	M [98], [101], G [94], [97], [102]	P [98], [101], B [94], [97], [102]	✓(Vision, Tactile) [98], (Force/Torque or Joint Positions, Visual Odometry [94]) [94], [102], ✗(Vision) [97], [101]	SDM [98], [101], SOM [98], [101], EU [98], [101], DDM [94], [97], [102], DOM [94], [97], [102]
Pose Estimation	[36], [79], [78]	APR	RO, OD, PM [79], [78], SD [78]	✗ [78], M [36], G [79]	P [36], [78], M [79]	✗(Tactile [36], [78]), (Vision [79])	Static environment [36], SDM [79], [78], SOM [36], [78], DOM [79], EU [36], [78]
Pose Estimation—Object Dynamics Learning I	[77]	CNS	RO, PM, OD	✗	P	✗(Vision)	SDM, EU, DOM
Pose Estimation—Object Dynamics Learning II	[76]	APR	RO, PM, AP	✗	P	✓(Vision, Tactile	SDM, EU,SOM
Object Dynamics Learning	[75]	APR	RO, AP, OD	✗	P	✗(Force/Torque)	DDM, DOM
Grasp Planning I	[82], [111], [112]	CNS	RO [82], [111], OD [82], [111], AP [82], [111], DO [112]	✗[111], G [112], (dependent on the algorithm in [82])	P [82], M [111], [112]	✗(Vision)	No dynamics, DOM [111], SOM [82], [112], EU [112]

[a] CNS versus APR.

[b] A prior is a source of information that aids in the interpretation of the sensor signal by rejecting noise, possibly by projecting the signal into a lower dimensional space. RO—rigid objects, AP—action primitives, PM—planar motion, OD—object database, SD—simple dynamics, AO—articulated objects, and DO—deformable objects.

[c] Alternatively, it can rely on some hard-coded action or just interpret/exploit the interaction induced by something/someone else. M = myopic/greedy, pH = variable planning horizon, G = global policy.

[d] Does the approach use multiple modalities? If so, which ones?

[e] Is uncertainty explicitly represented? How is it used? DDM—deterministic dynamics model, SDM—stochastic dynamics model, DOM—deterministic observation model, SOM—stochastic observation model, EU—estimates uncertainty.

TABLE II
TAXONOMY OF IP APPROACHES—PART 2

Goals and Paper Set ID	Papers	How is the Signal in $S \times A \times t$ Leveraged? [a]	What priors are employed? [b]	Does the Approach Perform Action Selection? [c]	What is the Objective: Perception, Manipulation, or Both?	Are Multiple Sensor Modalities Exploited? [d]	How is Uncertainty Modeled and Used? [e]
Grasp Planning II	[29], [80], [83]–[86], [113]–[116]	APR	RO, AP [29], [83], [86], [114]–[116], PM [80], [113], [115], OD [29], [80], [85], [113], SD [80], [113]	✗[115], G [80], [84], [85], M [29], [86], [114], PH [83], [113], [116]	B [84], M [29], [83], [85], [86], [114], [115]	✗(Proximity Sensors [84]), (Vision [86], [115], [116], [114]), (Tactile [80], [113]), ✓(Vision, Tactile) [83], [29], (Vision, Joint Ang.) [85]	No dynamic model [29], [84], [85], [114], [115], DOM [86], [114], [115], SOM[29], [80], [83], [85], [113], [116], SDM [80], [83], [86], [113], [116], EU [29], [80], [83]–[86], [113], [114], [116]
Grasp Planning—Pose Estimation	[81]	APR	RO, AP	PH	M	✓(Vision, Tactile)	SDM, SOM, EU
Haptic Property Estimation I	[42], [91], [92]	CNS	RO [91], [92], OD, PM, AP [42]	✗	P	✗(Tactile)	No dynamics, DOM [91], [92], SOM [42]
Haptic Property Estimation II	[41]	APR	PM, AP, OD	M	P	✗(Tactile)	No dynamics model, SOM, EU
Multimodal Object Model Learning I	[23], [89], [93]	CNS	RO, AP, PM [93]	✗	P	✓(Vision, Tactile) [23], [89], [93]	No dynamic model [93], [23], SDM [23], [93], SOM [23], [89], [93], EU [89]
Multimodal Object Model Learning II	[37], [39], [88], [90], [117]–[120], [121]	APR	RO, AP [39], [90], [117], [120], [121], OD [118], [119], PM [39], [90]	G [37], M [88], [118], [119], ✗ [39], [90], [117], [120], [121]	P [39], [88], [117], [118], [120], [121], B [37], [119]	✓(Tactile, Vision [37], [39], [90], [117]), ✗(Vision [88], [119], [121]), ✓(Vision, Joint Ang. [118]), ✓(Audio, FT [117]), (Tactile, Audio) [120]	Static environment [37], no dynamics [39], [88], [90], [118]–[120], [121], DDM [117], DOM [120] SOM [37], [39], [88], [90], [119], EU [37], [39], [88], [90], [118], No observation model or EU [119]
Multimodal Object Model Learning—Object Recognition	[59], [60], [62], [122] [123]	CNS	RO, AP [59], [60], [62], [122]	✗	P	✓(Vision, Audio, Tactile), ✗(Vision [122], [123])	No dynamics [59], [60], [62], [122], SDM [123], SOM [59], [60], [62], [122], [123], EU [59], [60], [62], [122]
Multimodal Object Model Learning—Grasp Planning	[87]	APR	RO	G	M	✓(Tactile, Vision)	static environment, SOM, EU
Manipulation Skill Learning	[63]–[68]	APR	OD, RO [65]–[68], DO [63], [64], AP [63]	G [64]–[68], M [63]	M	✗(Vision) [63], ✓(Vision, Internal torques) [64], ✓(Vision, Tactile) [65], [66], (Joint Pos.) [67], [68]	No dynamics model [65], [66], SDM [63], [64], [67], [68], SOM [63], [64], [66], DOM [65], [67], [68], EU [63], [64]
Manipulation Skill Learning—State Representation Learning	[73], [74]	APR	OD, RO	G, PH (MPC) [73]	M	✗(Vision) [73], (Vision, Joint Pos.) [74]	DDM [73], SDM [74], DOM
State Representation Learning	[71], [72]	APR	SD [71], AP, PM, RO	✗	P	✗(Vision) [71], ✓(Vision, Tactile) [72]	DDM, DOM
State Representation Learning—Object Dynamics Learning	[69], [70]	APR	AP, PM, RO, OD [70]	M [70], ✗ [69]	M [70],P [69]	✗(Vision)	DDM, DOM

[a] CNS versus APR.

[b] A prior is a source of information that aids in the interpretation of the sensor signal by rejecting noise, possibly by projecting the signal into a lower dimensional space. RO—rigid objects, AP—action primitives, PM—planar motion, OD—object database, SD—simple dynamics, AO—articulated objects, DO—deformable objects.

[c] Does the approach use multiple modalities? If so, which ones?

[d] Alternatively, it can rely on some hard-coded action or just interpret/exploit the interaction induced by something/someone else. M = myopic/greedy, pH = variable planning horizon, G = global policy.

[e] Is uncertainty explicitly represented? How is it used? DDM—deterministic dynamics model, SDM—stochastic dynamics model, DOM—deterministic observation model, SOM—stochastic observation model, EU—estimates uncertainty.

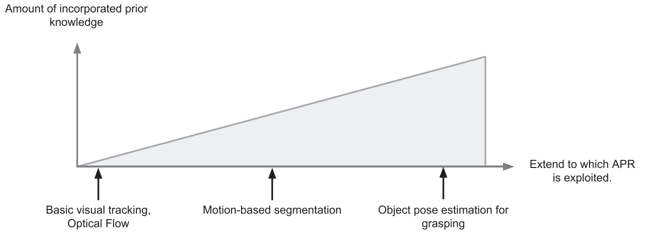


Fig. 7. Spectrum of the extent to which knowledge about the APR is exploited by IP approaches. Example problems are plotted along the x-axis. Their placement depends on how much prior knowledge about the interaction in the environment is commonly used in existing approaches toward them. For example, approaches toward basic visual object tracking or optical flow use very weak priors to regularize the solution space without incorporating knowledge about the specific interaction that caused the novel sensory response (CNS). Similar to that, approaches toward motion-based object segmentation often rely on interpreting a novel sensory response caused by an arbitrary interaction (CNS). Approaches toward object pose estimation often choose an action that is expected to provide the most informative sensory signal (APR). The wedge shape of the graph indicates that the more an approach exploits APR, the more prior knowledge it relies on. It also indicates that a strict classification of an approach may not be possible in every case.

assumed characteristic of the environment that thereby restricts the expected response of the world to an *arbitrary* action. Even more useful to robust perception and manipulation is to also include prior knowledge about the response to a *specific* interaction (APR).

Existing approaches toward IP cover a broad spectrum of how the possibilities afforded by this combined space $S \times A \times t$ are leveraged, as visualized in Fig. 7. On the one end of the spectrum, there are approaches, such as visual tracking or optical flow, that use very weak priors to regularize the solution space while maintaining generality (e.g., brightness constancy or local motion). In the middle, there are approaches that heavily rely on the regularity in the sensory response to an *arbitrary* interaction (e.g., rigid body dynamics, motion restricted to a plane or smooth motion). At the very end of the spectrum, there are approaches that leverage both assumptions about environmental constraints and knowledge about the *specific* interaction, to robustly interpret the resulting sensory signal and enable perceptually guided behavior. While using this kind of prior knowledge loses generality, it may result in more robust and efficient estimation in a robotics scenario due to a simplification of the solution space. If an approach leverages APR, then it also automatically leverages CNS.

2) Example Approaches: We start with approaches that exploit the informative sensory signal that is due to some forceful interaction (CNS). For instance, Fitzpatrick and Metta [53] and Kenney *et al.* [55] ease the task of visual segmentation and object model learning by making some general assumptions about the environment and thereby about the possible responses to an arbitrary interaction performed by the robot. Example assumptions are that only RO are present in the scene and that motion is restricted to a plane. Although the interaction is carried out by a robot, the available proprioceptive information is not used in the interpretation of the signal. Katz *et al.* [103], Sturm *et al.* [95], Pillai *et al.* [99], and Martín Martín and Brock [96] aim at understanding the structure of AO by observing their motion when they are interacted with. While objects are not restricted to be rigid or to only move in a plane, they are restricted to be piecewise rigid and to move according to some limited set of articulation mechanisms. Approaches by Bergström *et al.* [44], Chang *et al.* [56], Gupta and Sukhatme [50], Hausman *et al.* [51], [45], Kuzmic and Ude [52], and Schiebener *et al.* [57] devise different heuristics for selecting actions that generate informative sensory signals. These are used to ease perceptual tasks, such as object segmentation or object model learning. Similar to the above, none of the potentially available knowledge about interaction parameters is used to predict their effect.

The aforementioned approaches use vision as the source for informative sensory signals. Chu *et al.* [42] and Culbertson *et al.* [91] demonstrate how either unconstrained interactions in a plane or fixed interaction primitives lead to novel *haptic* sensory signals to ease the learning of material properties.

Other approaches utilize the regularity in $S \times A \times t$ to a much larger extent for easing perception and/or manipulation (APR). For example, Atkeson *et al.* [75] estimate the dynamics parameters of a robotic arm and the load at the end effector. This requires a sufficient amount of arm motion, measurements of joint torques, angles, velocities, and acceleration as well as knowledge of the arm kinematics. We can only learn the appropriate model that predicts arm motion from input motor torques if given this prior information on the structure of the space $S \times A \times t$ and data from interaction. Sinapov *et al.* [59], [60], and Sinapov and Stoytchev [62] let a robot interact with a set of objects that are characterized by different attributes, such as rigid or deformable, heavy or light, and slippery or not. Features computed on the different sensor modalities serve as the basis to learn object similarity. The authors show that this task is eased when the learning process is conditioned on joint torques and the different interaction behaviors. They also use the knowledge of the interaction in [60] and [62] to correlate various sensor modalities in the $S \times A \times t$.

Zhang and Trinkle [76] and Koval *et al.* [78] track object pose using visual and tactile data while a robot is pushing this object on a plane. Zhang and Trinkle [76] solve a nonlinear complementarity problem within their dynamics model to predict object motion given the control input. At the same time, they use observations of the object during interaction for estimating parameters of this model, such as the friction parameters. Koval *et al.* [78] assume knowledge of a lower dimensional manifold that describes the different contact states between a specific object and hand during a push motion. Hypotheses about future object poses are constrained to lie on this manifold. Hausman *et al.* [98] and Hsiao *et al.* [113] condition on the action to drive the estimation process. Hsiao *et al.* [113] estimate the belief state by conditioning the observations on the expected action outcomes. Hausman *et al.* [98] adopt a similar approach to estimate the distribution of possible articulation models based on action outcomes.

B. What Priors are Employed?

To devise an IP system means to interpret and/or deliberately generate a signal in the $S \times A \times t$. The regularity of this signal can be programmed into the system as a prior incorporating knowledge of the task; it can be learned from scratch or the system can pick up these regularities using a mixture of both priors and learning. Therefore, an important component of any IP system is this regularity and how it is encoded and exploited for performing a perception and/or manipulation task.

1) Priors on the Dynamics: IP requires knowledge of how actions change the state of the environment. Encoding this kind of regularity can be done in a dynamics model, i.e., the model for predicting the evolution of the environment after a certain action has been applied. Dependent on the number of objects in the environment, this prediction may be very costly to compute. Furthermore, due to uncertainty and noise in robot–object and object–object interactions, the effects of the interactions are stochastic.

a) Given/Specified/Engineered Priors: There are many approaches that rely on priors, which simplify the dynamics model and thereby make it less costly to predict the effect of an action. Examples of commonly used priors are the occurrence of only RO, of AO with a discrete set of links, or of only DO. Another prior includes the availability of a set of action primitives (AP), such as push, pull, grasp, etc. These action primitives are assumed to be accurately executed without failure. Many approaches assume that object motion is restricted to a plane (PM) or other simplifications of the scene dynamics (SD), e.g., quasi-static motion during multicontact interaction between objects. In this section, each prior will be explained in more detail by using one or several example approaches that exploit them.

Of the highlighted priors, some are more commonly used than others. For instance, apart from papers in paper set (Object Segmentation II, Object Segmentation—Object Recognition II, and Haptic Property Estimation II) almost all other approaches make assumptions about the nature of objects in the environment, i.e., they assume that all objects present in the environment belong exclusively to one of three classes: rigid, articulated, or deformable.

The majority of approaches in IP assume that the objects are rigid (RO). Only approaches concerned with estimating an articulation model assume the existence of AO. Similarly, [112], [63], and [64] in paper set Manipulation Skill Learning are unique in that they are the only ones that deal with the manipulation of DO.

Many approaches in the paper set Object Segmentation I utilize the planar motion prior (PM). In instances, such as [50], this prior is used for scene segmentation, where all the objects in the scene are assumed to lie on a table plane. In other approaches, e.g., [45], [46], [51], [52], and [56] in Object Segmentation I, [93] in Multimodal Object Model Learning I and [39], [90] in Multimodal Object Model Learning II, the planar motion assumption is used not only for scene segmentation, but also to track the movement of objects in the environment.

Then, there are approaches that make additional simplifying assumptions about the dynamics of the system (SD). For instance, Koval *et al.* [80] assume that the object being manipulated has quasi-static dynamics and moves only on a plane (PM). Such an assumption becomes particularly useful in cases where action selection is performed via a multistep planning procedure because it simplifies the forward prediction of object motion.

b) Learned Priors: There are approaches that learn a dynamics model of the environment given an action. Some of these let the robot learn this autonomously through trial and error. Early approaches toward this are given by Christiansen *et al.* [79], and Metta and Fitzpatrick [54], where they learn a simple mapping from the current state and action to a most likely outcome. Christiansen *et al.* [79] demonstrate this in a tray-tilting task for bringing the object lying on this tray into a desired configuration. Metta and Fitzpatrick [54] demonstrate their approach in an object pushing behavior and learn the response of an object to a certain push direction. Both of them model the nondeterminism of the response of the object to an action. More recent approaches are presented by Levine *et al.* [67], Han *et al.* [68], and Wahlström *et al.* [73], where they learn the mapping from current state to next best action in a policy search framework. Lee *et al.* [64], Pastor *et al.* [65], and Kappler *et al.* [66] bootstrap the search process through trial and error by demonstrating actions.

2) Priors on the Observations: Regularities can also be encoded in the observation model that relates the state of the system to the raw sensory signals. Thereby, it can predict the observation given the current state estimate. Only if this relationship is known, an IP robot can gain information from observations. This information may be about some quantity of interest that needs to be either estimated or directly provide the distance to some goal state.

a) Given/Specified/Engineered Observation Models: Traditionally, the relationship between the state and raw sensory signals is hand designed based on some expert knowledge. One example is models of multiview or perspective geometry for camera sensors [14], [124]. Often, approaches also assume access to an object database (OD) that allows them to predict how the objects will be observed through a given sensor, e.g., Chu *et al.* [42].

b) Learned State Representations: More recently, we see more approaches that learn a suitable, task-specific state representation directly from observations. Examples include [71], [73], and [74], where the authors use raw pixel values as input and learn the lower dimensional representation jointly with the policy that maps these learned states to actions. Jonschkowski and Brock [71] achieve this by introducing a set of hand-defined priors in a loss function that is minimal if the state representation best matches these priors. The mapping from raw pixels to the lower dimensional representation is linear. Levine *et al.* [74] map the raw pixel values through a nonlinear convolutional neural network (CNN) to a set of feature locations in the image and initialize the weights for an object pose estimation task. Both the type of function approximator (CNN encoding receptive fields) and the data for initialization can be seen as a type of prior. Wahlström *et al.* [73] use an autoencoder framework, where they not only minimize the reconstruction error from the low-dimensional space back to the original space, but also optimize the consistency in the latent, low-dimensional space.

In the case, where the mapping between state and observation is hand designed, the state usually refers to some physical quantity. In the case where the state representation is learned, it is not so easily interpretable.

C. Does the Approach Perform Action Selection?

Knowledge about the structure of $S \times A \times t$ can also be exploited to select appropriate actions. A good action will reveal as much information as possible and at the same time bring the system as close as possible to the manipulation goal. If we know something about the structure of $S \times A \times t$, we can perform action selection so as to make the resulting sensor information as meaningful as possible. The agent must balance between exploration (performing an action to improve perception as much as possible) and exploitation (performing an action that maximizes progress toward the manipulation goal).

1) Problem Formulation: For optimal action selection, the IP agent needs to know a policy that given the current state estimate returns the optimal action or sequence of actions to take. Here, optimal means that the selected actions yield a maximum expected *reward* to the IP agent. The specific definition of the reward function heavily depends on the particular task of the robot. If it is a purely perceptual task, actions are often rewarded when they reduce the uncertainty about the current estimate (exploration), e.g., van Hoof *et al.* [47]. If the task is a manipulation task, actions may be rewarded that bring you closer to a goal (exploitation), e.g., Levine *et al.* [67].

Finding this policy is one of the core problems for action selection. Its formalization depends on whether the state of the dynamical system is directly observable or whether it needs to be estimated from noisy observations. It also depends on whether the dynamics model is deterministic or stochastic.

2) Dynamics Model: Knowing the dynamics model is even more important for action selection than for improving perception. It allows us to predict the effect of an action on the quantity of interest and thereby the expected reward. A common way to find the optimal sequence of actions that maximizes reward under deterministic dynamics is forward or backward value iteration [125].

As mentioned earlier, a realistic dynamics model should be stochastic to account for uncertainty in sensing and execution. In this case, to find the optimal sequence of actions, the agent has to form an expectation over all the possible future outcomes of an action. The dynamical system can then be modeled as an *Markov Decision Process (MDP)*. Finding the optimal sequence of actions can be achieved through approaches, such as *value* or *policy* iteration [126].

In an MDP, we assume that the state of the system is directly observable. However, in a realistic scenario, the robot can only observe its environment through noisy sensors. This can be modeled with a *Partially Observable Markov Decision Process (POMDP)* where the agent has to maintain a probability distribution over the possible states, i.e., the belief, based on an observation model. For most real-world problems, it is intractable to find the optimal policy of the corresponding POMDP. Therefore, there exist many methods that find approximate solutions to this problem [127].

Predictive State Representations (PSRs) are another formalism for action selection. Here, the system dynamics are represented directly by observable quantities in the form of a set of tests instead of over some latent state representation as in POMDPs [128]–[130].

3) Planning Horizon (PH): Action selection methods can be categorized based on the number of steps they look ahead in time. There are approaches that have a single-step look ahead, which are called myopic or greedy (M). Here, the agent's actions are optimized for rewards in the next time step, given the current state of the system. Most approaches to IP which exploit the knowledge of the outcome of an action in $S \times A \times t$ are myopic (M). Myopic approaches do not have to cope with the evolution of complex system dynamics or observation models beyond a single step. Hence, this considerably reduces the size of the possible solution space. Examples of such approaches can be seen in paper sets Object Segmentation II [98], [101], in Articulation Model Estimation II [36] and in the paper set Pose Estimation.

Then, there are approaches that look multiple steps ahead in time to inform their action selection process. These multistep look-ahead solutions decide an optimal course of action also based on the current state of the system. The time horizon for these multistep look aheads can either be fixed or variable. In either case, the time horizons are generally dictated by a budget, examples of which include computational resources, uncertainty about the current state, costs associated with the system, etc. For instance, a popular multistep look-ahead approach relies on the assumption that the *maximum-likelihood estimate* (MLE) observation will be obtained in the future. This way, one can predict the behavior of the system within the time horizon and use it to select an action. Overall, we label such approaches to action selection as PH approaches. Examples of these approaches include [81] and [83].

Another set of methods tries to find global policies (GP) that specify the action that should be applied at any point in time for any state. We categorize such approaches as methods that have GP. Among these, there are approaches that take into account all possible distributions over the state space (*beliefs*) and offer globally optimal policies. These policies account for stochastic belief system dynamics, i.e., they maintain probabilities over the possible current states and probable outcomes given an action. Such methods are often solved by formulating them as POMDPs. In practice, the solution to such problems is intractable and are often solved by approximate offline methods. Javdani *et al.* [36] and Koval *et al.* [80] demonstrate such an approach to action selection for IP. Another way of finding GP uses reinforcement learning that provides a methodology to improve a policy over time. An example of a specific policy search method is presented in [67], [68], and [74].

Apart from planning-based approaches that perform action selection, there are approaches that focus on low-level control. In these approaches, the control input is computed online for the next cycle based on a global control law. We also classify these methods as GP approaches as they compute the next control input based on control law that is global, e.g., the feedback matrix in linear Gaussian controllers. The actions are generated at a high frequency and operate on low-level control commands.

Examples of these approaches include [37], [84], [87], [94], and [102].

4) Granularity of Actions: Action selection can be performed at various granularities. For example, a method may either select the next best control input or an entire high-level action. The next best controls can be low-level motor torques that are sent to the robot in the next control cycle. The corresponding action selection loop is executed at a very high frequency and is dependent on the immediate feedback from different sensors [37], [84], [87], [94], [102].

High-level action primitives are generally used in approaches that do not require reasoning about fine motor control, such as pushing or grasping actions that are represented by motion primitives. In such cases, reasoning about observations is purely dependent on the outcome of high-level actions. There are numerous approaches that utilize high-level actions for IP. Examples are [100] and the following in paper set Object Segmentation I: [44], [53], [54], [55], [95], [99], and [96].

D. What is the Objective: Perception, Manipulation, or Both?

Approaches to IP may pursue a perception or a manipulation goal and in some cases both (see Fig. 4). Object segmentation, recognition and pose estimation, multimodal object model learning, and articulation model estimation are examples of areas where IP is utilized to service perception.

Then, there are IP approaches whose primary objective is to achieve a manipulation goal (e.g., grasping or learning manipulation skills). For instance, Pastor *et al.* [65] and Kappler *et al.* [66] exploit regularities in $S \times A \times t$ to enable better action selection. The robot compares the observed perceptual signal with the expected perceptual signal given the current manipulation primitive. It then picks controls that drive the system toward the expected signal. Similarly, Koval *et al.* [78], Kaelbling and Lozano-Pérez [81], and Platt *et al.* [83] exploit the regularities in $S \times A \times t$ to facilitate task oriented grasping, i.e., locate and grasp an object of interest.

The final thread of IP approaches include a combination of both perception and manipulation. For instance, Koval *et al.* [80] and Dragiev *et al.* [87] simultaneously improve perception (object model reconstruction or pose estimation, respectively) and select better actions under uncertainty (efficient grasping). In [94], [102] in paper set Articulation Model Estimation II, the knowledge about the regularity in both the observations and dynamics in $S \times A \times t$ is used to improve the articulation model estimation as well as to enable better control. In the case of [102], the control input is directly incorporated into the state estimation procedure. In contrast, Jain and Kemp [94] use the position of the end effector in the articulation mechanism estimation. The manipulation goal in both these approaches is to enable a robot to open doors and drawers.

E. Are Multiple Sensor Modalities Exploited?

Some approaches exploit multiple modalities in the $S \times A \times t$ space, whereas other approaches restrict themselves to a single informative modality. The various sensing modalities can be broadly categorized into contact and noncontact sensing. Examples of noncontact sensing include vision, proximity sensors, sonar, etc. Contact sensing is primarily realized via tactile sensors and force–torque sensors. Approaches that only use tactile sensing include the works of Chu *et al.* [42], Koval *et al.* [78], [80], and Javdani *et al.* [36]. There are also approaches that use both contact and noncontact sensing to inform the signal in the $S \times A \times t$ space. These include some of the works listed in paper sets Articulation Model Estimation II, Pose Estimation—Object Dynamics Learning II, Multimodal Object Model Learning I and II, and Manipulation Skill Learning in Tables I and II.

F. How is Uncertainty Modeled and Used?

In IP tasks, there are many sources of uncertainty about the quantity of interest. One of them is the noisy sensors through which an agent can only partially observe the current state of the world. Another is the dynamics of the environment in response to an interaction. Some approaches toward IP model this uncertainty in either their observations and/or the dynamics model of the system. Depending on their choice, there are a wide variety of options for estimating the quantity of interest from a signal in $S \times A \times t$. For updating the current estimate, some approaches use *recursive state estimation* and maintain a full posterior distribution over the variable of interest, e.g., [76], [78], and [96]. Others frame their problem in terms of energy minimization in a graphical model and only maintain the *maximum a posteriori* (MAP) solution from frame to frame, e.g., [44]. An MLE of the variable of interest is computed in approaches that do not maintain a distribution over possible states. Examples are clustering methods that assign fixed labels [45], [50], [51], [56] to the variable of interest. More recently, *nonparametric approaches* have also been utilized. For instance, Boularias *et al.* [86] use *kernel density estimation*.

Methods that model uncertainty of the variable(s) of interest can cope better with noisy observations or dynamics, but they become slower to compute as the size of the solution space grows. This creates a natural tradeoff between modeling uncertainty and computational speed. The above-mentioned choices also have implications for action selection. If we maintain a full distribution over the quantity of interest, then computing a policy that takes the stochasticity in the dynamics and observation models into account is generally intractable [125]. If an approach assumes a known state, the dynamical system can also be modeled by an MDP with stochastic dynamics given an action. The least computationally demanding model for action selection is the one that neglects any noise in the observations or dynamics. However, it might also be the least robust depending on the true variance in the real dynamical system that the agent tries to control.

Based on the above, we propose four labels for IP approaches with respect to their way of modeling and incorporating uncertainty in estimation and manipulation tasks. Approaches that assume deterministic dynamics are labeled (DDM), stochastic dynamics (SDM), deterministic observations (DOM), stochastic observations (SOM), and approaches that estimate uncertainty are labeled (EU).

Fitzpatrick and Metta [53], Metta and Fitzpatrick [54], and Kenney *et al.* [55] propose example approaches that assume no stochasticity in the system, and model both the dynamics and observations deterministically. Then, there are approaches that assume deterministic observations but do not model the dynamics at all. These are listed in paper set Object Segmentation I that include the works of Chang *et al.* [56], Gupta and Sukhatme [50], Hausman *et al.* [51], [45], Kuzmic and Ude [52], and Schiebener *et al.* [46]. Then, there are approaches that model only stochastic observations but no dynamics because they assume that the environment is static upon interaction, e.g., [84]. Most approaches that assume both stochastic dynamics and observations have some form of uncertainty estimation technique implemented to account for the stochasticity in the system. An approach that assumes stochasticity in its observations but does not estimate uncertainty is given by Chu *et al.* [42]. Here, the authors train a max-margin classifier to assign labels to stochastic observations.

VI. Discussion and Open Questions

A. Remaining Challenges

If IP is about merging perception and manipulation into a single activity then the natural question arises of how to balance these components. When have manipulation actions (that are in service of perception) elicited sufficient information about the world such that manipulation actions can succeed that are in service of a manipulation goal? This question bears significant similarities with the exploration/exploitation tradeoff encountered in reinforcement learning. One can further ask: How can manipulation actions be found that combine these two objectives—achieving a goal and obtaining information—in such a way that desirable criteria about the resulting sequence of actions (time, effort, risk, etc.) are optimized?

When performing manipulation tasks, humans aptly combine different sources of information, including prior knowledge about the world and the task, visual information, haptic feedback, and acoustic signals. Research in IP is currently mostly concerned with visual information. New algorithms are necessary to extend IP toward a multimodal framework, where modalities are selected and balanced so as to maximally inform manipulation with the least amount of effort, while achieving a desired degree of certainty. Furthermore, for every sensory channel, one might differentiate between passively (e.g., just look), actively (e.g., change vantage point to look), and interactively (e.g., observe interaction with the world) acquired information. Each of these is associated with a different cost but also with a different expected information gain. In addition to adequately mining information from multiple modalities, IP must be able to decide in which of these different ways the modality should be leveraged.

Also, at the lower levels of perception significant changes might be required. It is conceivable that existing representations of sensory data are not ideal for IP. Given the focus on dynamic scenes with multiple moving objects, occlusions, lighting changes, and new objects appearing and old ones disappearing—does it make sense to tailor visual features and corresponding tracking methods to the requirements of IP? Are there fundamental processing steps, similar to edge or corner detection, that are highly relevant in the context of IP but have not seen a significant need in other applications of Computer Vision? The same for haptic or acoustic feedback: When combined with other modalities in the context of IP, what might be the right features or representations we should focus on?

B. Framework for IP?

All of the aforementioned arguments indicate that IP might require a departure from existing perception frameworks, as they can be found in applications outside of robotics, such as surveillance, image retrieval, etc. In IP, manipulation is an integral component of perception. The perceptual process must continuously tradeoff multiple sensor modalities that might each be passive, active, or interactive. There is no stand-alone perceptual process and not only a single aspect of the environment that must be extracted from the sensor stream as the optimization objectives may change when the robot faces different tasks over its lifetime.

After the review of existing work in the field, we conclude that there is yet no framework that can address all the challenges in IP. There are, however, candidates that represent the regularity in $S \times A \times t$ in a way that caters to a particular challenge encountered in IP. For instance, Krüger *et al.* [131] present a concept that allows us to symbolically represent continuous sensory-motor experience: *Object-Action Complexes*. The concept's current instantiations through the examples in [131] are focused on learning and detecting *affordances* [1], which describe the relationship between a certain situation (often including an object) and the action that it allows.

Other popular formalisms lend themselves particularly well to the problem of optimal action selection (see Section V-C). Examples include MDPs, POMDPS, PSRs, or multiarmed bandits. They rely on different assumptions (e.g., Markov assumptions, observable state) and make different algorithmic choices (e.g., probabilistic modeling). Approaches that rely on these decision-making frameworks often assume the availability of transition, observation, and reward functions and the possibility to analytically compute the optimal action.

For complex real-world problems this is often not the case and information about the world can only be collected through interaction. The data collected in this way are then used to update the relevant models. The problem of selecting the next best action may be based on submodularity [36], the variance in a Gaussian Process [37], [39], or the Bhattacharyya coefficient between two normal distributions [40], [41].

Reinforcement learning [126] is also a common choice to learn a policy for action selection under these complex conditions. Many approaches assume the availability of some reliable state estimator (e.g., by using motion capture or marker-based systems) where the state is of relatively low dimension and hand designed. Particularly relevant to IP are recent approaches that directly learn a state representation from data and employ reinforcement learning on this learned state representation [70], [71], [74].

All these formalisms have been used to solve particular subproblems encountered in the context of IP. We do not claim that this list is complete. However, the wealth of very different approaches suggests that there is currently not one framework for IP that can address all the relevant challenges. It is an open question what such a framework would be and how it could enable coordinated progress by developing adequate subcomponents.

C. New Application Areas

The majority of the work that is included in this survey is concerned with IP for manipulating and grasping objects in the environment. In the context of the recent *Darpa Robotics Challenge*, we have also seen a need to bridge the gap between perception and action in whole-body, multicontact motion planning and control. The ability to physically explore unstructured environments (such as those encountered in disaster sites) is of utmost importance for the safety and robustness of a robot. Probing and poking not only with your hands but also your legs can also help extract more information. Currently, these robots extensively rely on teleoperation and carefully designed user interfaces [131], [132]. We argue that they can achieve a much higher degree of autonomy if they rely on IP.

VII. Summary

This survey paper provides an overview on the current state of the art in IP research. In addition to presenting the benefits of IP, we discuss various criteria for categorizing existing work. We also include a set of problems, such as object segmentation, manipulation skills, and object dynamics learning, that are commonly eased using concepts of IP.

We identify and define the two main aspects of IP. 1) Any type of forceful interaction with the environment creates a new type of informative sensory signal that would otherwise not be present, and 2) any prior knowledge about the nature of the interaction supports the interpretation of the signal in the Cartesian product space of $S \times A \times t$. We use these two crucial aspects of IP as criteria to include a paper as related or not. Furthermore, we compare IP with existing perception approaches and named a few formalisms that allow us to capture an IP problem.

We hope that this taxonomy helps us to establish benchmarks for comparing various approaches and to identify open problems.

Acknowledgment

The authors would like to thank the anonymous reviewers for their insightful comments, all the cited authors who provided feedback upon our request, and Aleksandra Waltos for providing the visuals in Figs. 5 and 6. The authors would also like to thank the Swedish Research Council and Swedish Foundation for Strategic Research. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding organizations.

References

[1] J. J. Gibson, *The Ecological Approach to Visual Perception*. Boston, MA, USA: Houghton Mifflin, 1979.

[2] J. K. O'Regan and A. Noë, "A sensorimotor account of vision and visual consciousness," *Behav. Brain Sci.*, vol. 24, pp. 939–973, 2001.

[3] A. Noë, *Action in Perception*. Cambridge, MA, USA: MIT Press, 2004.

[4] R. Held and A. Hein, "Movement-produced stimulation in the development of visually guided behaviour," *J. Comparative Physiol. Psychol.*, vol. 56, pp. 872–876, Oct. 1963.

[5] J. J. Gibson, *The Senses Considered as Perceptual Systems*. Boston, MA, USA: Houghton Mifflin, 1966.

[6] M. Everingham, L. Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[7] O. Russakovsky *et al.*,"ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[8] T.-Y. Lin *et al.*,"Microsoft COCO: Common objects in context," in *Proc. 13th Eur. Conf. Comput. Vis.*, Zurich, Switzerland, Sep. 6–12, 2014, pp. 740–755.

[9] M. A. Erdmann and M. T. Mason, "An exploration of sensorless manipulation," *IEEE J. Robot. Autom.*, vol. 4, no. 4, pp. 369–379, Aug. 1988.

[10] M. R. Dogar, K. Hsiao, M. T. Ciocarlie, and S. S. Srinivasa, "Physics-based grasp planning through clutter," in *Proc. Robot., Sci. Syst. Conf.*, 2012.

[11] H. Kjellström, J. Romero, and D. Kragic, "Visual object-action recognition: Inferring object affordances from human demonstration," *Comput. Vis. Image Understanding*, vol. 115, pp. 81–90, 2010.

[12] M. Cai, K. M. Kitani, and Y. Sato, "Understanding hand-object manipulation with grasp types and object attributes," in *Proc. Robot., Sci. Syst. Conf.*, Ann Arbor, MI, USA, Jun. 2016.

[13] T. Xue, M. Rubinstein, C. Liu, and W. T. Freeman, "A computational approach for obstruction-free photography," *ACM Trans. Graph.*, vol. 34, no. 4, pp. 79:1–79:11, Jul. 2015.

[14] J. Aloimonos, I. Weiss, and A. Bandyopadhyay, "Active vision," *Int. J. Comput. Vis.*, vol. 1, pp. 333–356, 1988.

[15] R. Bajcsy, "Active perception," *Proc. IEEE.*, vol. 76, no. 8, pp. 966–1005, Aug. 1988.

[16] D. H. Ballard, "Animate vision," *Artif. Intell.*, vol. 48, no. 1, pp. 57–86, 1991.

[17] K. Pahlavan, T. Uhlin, and J.-O. Eklundh, "Active vision as a methodology," in *Active Perception*. Hillsdale, NJ, USA: Lawrence Erlbaum Asso., 1993, pp. 19–46.

[18] P. M. Sharkey, D. W. Murray, P. F. McLauchlan, and J. P. Brooker, "Hardware development of the Yorick series of active vision systems," *Microprocessors Microsyst.*, vol. 21, no. 6, pp. 363–375, Mar. 1998.

[19] T. Asfour, K. Welke, P. Azad, A. Ude, and R. Dillmann, "The Karlsruhe humanoid head," in *Proc. IEEE/RAS Int. Conf. Humanoid Robots*, Dec. 2008, pp. 447–453.

[20] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature Rev. Neurosci.*, vol. 2, no. 3, pp. 194–203, Mar. 2001.

[21] J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. Lai, N. Davis, and F. Nuflo, "Modeling visual attention via selective tuning," *Artif. Intell.*, vol. 78, no. 12, pp. 507–545, 1995.

[22] R. Bajcsy, Y. Aloimonos, and J. K. Tsotsos, "Revisiting active perception," *CoRR*, vol. abs/1603.02729, 2016.

[23] L. Natale, G. Metta, and G. Sandini, "Learning haptic representation of objects," in *Proc. Int. Conf. Intell. Manipulation Grasping*, Jul. 2004.

[24] G. Sandini, F. Gandolfo, E. Grosso, and M. Tistarelli, *Vision During Action*. Trenton NJ, USA: Lawrence Erlbaum Assoc., 1993, ch. 4.

[25] C. J. Tsikos and R. K. Bajcsy, "Segmentation via manipulation," Tech. Rep. MS-CIS-88-42, Dept. Comput. Inf. Sci., Univ. Pennsylvania, Philadelphia, PA, USA, 1988.

[26] C. J. Tsikos and R. K. Bajcsy, "Segmentation via manipulation," *IEEE Trans. Robot. Autom.*, vol. 7, no. 3, pp. 306–319, Jun. 1991.

[27] R. Bajcsy, "Active perception and exploratory robotics," *Robots and Biological Systems: Towards a New Bionics?*, Berlin, Germany: Springer, pp. 3–20, 1993, doi: 10.1007/978-3-642-58069-7_1.

[28] R. Bajcsy and P. R. Sinha, "Exploration of surfaces for robot mobility," in *Proceedings of the Fourth International Conference on CAD, CAM, Robotics and Factories of the Future*. New York, NY, USA: McGraw-Hill, Dec. 1989.

[29] M. Salganicoff and R. Bajcsy, "Sensorimotor learning using active perception in continuous domains," in *Proc. AAAI Fall Symp. Sensory Aspects Robot Intell.*, Nov. 1991.

[30] H. R. Nicholls and M. H. Lee, "A survey of robot tactile sensing technology," *Int. J. Robot. Res.*, vol. 8, no. 3, pp. 3–30, 1989.

[31] R. Bajscy, "Shape from touch," in *Advances in Automation and Robotics*. Greenwich, CT, USA: JAI Press, 1988.

[32] K. I. Goldberg and R. Bajscy, "Active touch and robot perception," *Cognit. Brain Theory*, vol. 7, no. 2, p. 199–214, 1984.

[33] P. Allen and R. Bajscy, "Two sensors are better than one: Example of integration of vision and touch," in *Proc. Int. Symp. Robot. Res.*, France, Oct. 1985.

[34] A. Petrovskaya and O. Khatib, "Global localization of objects via touch," *IEEE Trans. Robot.*, vol. 27, no. 3, pp. 569–585, Jun. 2011.

[35] P. Hebert, T. Howard, N. Hudson, J. Ma, and J. Burdick, "The next best touch for model-based localization," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2013, pp. 99–106.

[36] S. Javdani, M. Klingensmith, J. A. Bagnell, N. S. Pollard, and S. S. Srinivasa, "Efficient touch based localization through submodularity," in *Proc. IEEE Int. Conf. Robot. Autom.,* 2013, pp. 1828–1835.

[37] S. Dragiev, M. Toussaint, and M. Gienger, "Gaussian process implicit surfaces for shape estimation and grasping," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2011, pp. 2845–2850.

[38] M. Kim *et al.*," Exploration of unknown object by active touch of robot hand," *Int. J. Control, Autom. Syst.*, vol. 12, no. 2, pp. 406–414, 2014.

[39] J. Bohg, M. Johnson-Roberson, M. Björkman, and D. Kragic, "Strategies for multi-modal scene exploration," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2010, pp. 4509–4515.

[40] J. Fishel and G. Loeb, "Bayesian exploration for intelligent identification of textures," *Frontiers Neurorobot.*, vol. 6, 2012, doi: 10.3389/fnbot.2012.00004.

[41] G. E. Loeb and J. A. Fishel, "Bayesian action & perception: Representing the world in the brain," *Frontiers Neurosci.*, vol. 8, 2014, doi: 10.3389/fnins.2014.00341.

[42] V. Chu *et al.*, "Robotic learning of haptic adjectives through physical interaction," *Robot. Auton. Syst.*, vol. 63, pp. 279–292, 2015.

[43] D. Katz and O. Brock, "Interactive segmentation of articulated objects in 3D," in *Proc. Workshop Mobile Manipulation Int. Conf. Robot. Autom.*, 2011.

[44] N. Bergström, C. H. Ek, M. Björkman, and D. Kragic, "Scene understanding through autonomous interactive perception," in *Proc. Int. Conf. Comput. Vis. Syst.*, 2011, pp. 153–162.

[45] K. Hausman *et al.*, "Interactive segmentation of textured and textureless objects," in *Handling Uncertainty and Networked Structure in Robot Control*. New York, NY, USA: Springer, 2015, pp. 237–262.

[46] D. Schiebener, A. Ude, and T. Asfour, "Physical interaction for segmentation of textured and non-textured unknown rigid objects," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2014, pp. 4959–4966.

[47] H. van Hoof, O. Kroemer, H. B. Amor, and J. Peters, "Maximally informative interaction learning for scene exploration," in *Proc. Int. Conf. Intell. Robots Syst.*, 2012, pp. 5152–5158.

[48] H. van Hoof, O. Kroemer, and J. Peters, "Probabilistic segmentation and targeted exploration of objects in cluttered environments," *IEEE Trans. Robot.*, vol. 30, no. 5, pp. 1198–1209, Oct. 2014.

[49] K. Xu *et al.*, "Autoscanning for coupled scene reconstruction and proactive object analysis," *ACM Trans. Graph.*, vol. 34, no. 6, 2015, Art. no. 177.

[50] M. Gupta and G. S. Sukhatme, "Using manipulation primitives for brick sorting in clutter," in *Proc. Int. Conf. Robot. Autom.*, May 2012, pp. 3883–3889.

[51] K. Hausman *et al.*, "Tracking-based interactive segmentation of textureless objects," in *Proc. IEEE Int. Conf. Robot. Autom.*, Karlsruhe, Germany, May 6–10, 2013, pp. 1122–1129.

[52] E. S. Kuzmic and A. Ude, "Object segmentation and learning through feature grouping and manipulation," in *Proc. 10th IEEE-RAS Int. Conf. Humanoid Robots*, 2010, pp. 371–378.

[53] P. Fitzpatrick and G. Metta, "Towards manipulation-driven vision," in *Proc. IEEE/RSJ Intell. Robots Syst.*, 2002, vol. 1, pp. 43–48.

[54] G. Metta and P. Fitzpatrick, "Early integration of vision and manipulation," in *Proc. Proc. Int. Joint Conf. Neural Netw.,* Jul. 2003, vol. 4, pp. 2703.

[55] J. Kenney, T. Buckley, and O. Brock, "Interactive segmentation for manipulation in unstructured environments," in *Proc. Int. Conf. Robot. Autom.*, 2009, pp. 1377–1382.

[56] L. Y. Chang, J. R. Smith, and D. Fox, "Interactive singulation of objects from a pile," in *Proc. Int. Conf. Robot. Autom.*, 2012, pp. 3875–3882.

[57] D. Schiebener, J. Morimoto, T. Asfour, and A. Ude, "Integrating visual perception and manipulation for autonomous learning of object representations," *Adapt. Behav.*, vol. 21, no. 5, pp. 328–345, 2013.

[58] A. Ude, D. Omrcen, and G. Cheng, "Making object learning and recognition an active process," *Int. J. Humanoid Robot.*, vol. 5, no. 2, pp. 267–286, 2008.

[59] J. Sinapov, C. Schenck, and A. Stoytchev, "Learning relational object categories using behavioral exploration and multimodal perception," in *Proc. IEEE Int. Conf. Robot. Autom.*, Hongkong, China, May 31–Jun. 7, 2014, pp. 5691–5698.

[60] J. Sinapov, C. Schenck, K. Staley, V. Sukhoy, and A. Stoytchev, "Grounding semantic categories in behavioral interactions: Experiments with 100 objects," *Robot. Auton. Syst.*, vol. 62, no. 5, pp. 632–645, 2014.

[61] K. Hausman, C. Corcos, J. Müller, F. Sha, and G. Sukhatme, "Towards interactive object recognition," in *Proc. IROS 2014 Workshop Robots Clutter, Perception Interaction Clutter*, Chicago, IL, USA, Sep. 2014.

[62] J. Sinapov and A. Stoytchev, "Grounded object individuation by a humanoid robot," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2013, pp. 4981–4988.

[63] M. Cusumano-Towner, A. Singh, S. Miller, J. F. O'Brien, and P. Abbeel, "Bringing clothing into desired configurations with limited perception," in *IEEE Int. Con. Robot. Autom.*, Shanghai, China, May 9–13, 2011, pp. 3893–3900.

[64] A. X. Lee, H. Lu, A. Gupta, S. Levine, and P. Abbeel, "Learning force-based manipulation of deformable objects from multiple demonstrations," in *Proc. IEEE Int. Conf. Robot. Autom.*, Seattle, WA, USA, May 26–30, 2015, pp. 177–184.

[65] P. Pastor, L. Righetti, M. Kalakrishnan, and S. Schaal, "Online movement adaptation based on previous sensor experiences," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.,* 2011, pp. 365–371.

[66] D. Kappler, P. Pastor, M. Kalakrishnan, M. Wuthrich, and S. Schaal, "Data-driven online decision making for autonomous manipulation," in *Proc. Robot., Sci. Syst.*, Rome, Italy, 2015.

[67] S. Levine, N. Wagener, and P. Abbeel, "Learning contact-rich manipulation skills with guided policy search," in *Proc. IEEE Int. Conf. Robot. Autom.,* Seattle, WA, USA, May 26–30, 2015, pp. 156–163.

[68] W. Han, S. Levine, and P. Abbeel, "Learning compound multi-step controllers under unknown dynamics," in *Proc. 2015 IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Hamburg, Germany, Sep. 28–Oct. 2, 2015, pp. 6435–6442.

[69] C. Finn, I. Goodfellow, and S. Levine, "Unsupervised learning for physical interaction through video prediction," in *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds. Red Hook, NY, USA: Curran Assoc., 2016, pp. 64–72.

[70] P. Agrawal, A. Nair, P. Abbeel, J. Malik, and S. Levine, "Learning to poke by poking: Experiential learning of intuitive physics," in *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds. Red Hook, NY, USA: Curran Assoc., 2016.

[71] R. Jonschkowski and O. Brock, "Learning state representations with robotic priors," *Auton. Robots*, vol. 39, no. 3, pp. 407–428, 2015.

[72] L. Pinto, D. Gandhi, Y. Han, Y. Park, and A. Gupta, "The curious robot: Learning visual representations via physical interactions," in *Proc. 14th Eur. Conf. Comput. Vis.,* Amsterdam, The Netherlands, Oct. 11–14, 2016, pp. 3–18.

[73] N. Wahlström, T. B. Schön, and M. P. Deisenroth, "Learning deep dynamical models from image pixels," *IFAC-PapersOnLine*, vol. 48, no. 28, pp. 1059–1064, 2015.

[74] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," *J. Mach. Learn. Res.*, vol. 17, no. 39, pp. 1–40, 2016.

[75] C. G. Atkeson, C. H. An, and J. M. Hollerbach, "Estimation of inertial parameters of manipulator loads and links," *Int. J. Robot. Res.*, vol. 5, no. 3, pp. 101–119, 1986.

[76] L. Zhang and J. Trinkle, "The application of particle filtering to grasping acquisition with visual occlusion and tactile sensing," in *Proc. IEEE. Int. Conf. Robot. Autom.*, May 2012, pp. 3805–3812.

[77] J. Wu, I. Yildirim, J. J. Lim, B. Freeman, and J. Tenenbaum, "Galileo: Perceiving physical object properties by integrating a physics engine with deep learning," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Red Hook, NY, USA: Curran Assoc., 2015, pp. 127–135.

[78] M. C. Koval, N. S. Pollard, and S. S. Srinivasa,"Pose estimation for planar contact manipulation with manifold particle filters," *Int. J. Robot. Res.*, vol. 34, no. 7, pp. 922–945, 2015.

[79] A. Christiansen, M. T. Mason, and T. Mitchell, "Learning reliable manipulation strategies without initial physical models," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 1990, vol. 2, pp. 1224–1230.

[80] M. C. Koval, N. S. Pollard, and S. S. Srinivasa, "Pre-and post-contact policy decomposition for planar contact manipulation under uncertainty," *Int. J. Robot. Res.*, vol. 35, no. 1–3, pp. 244–264, 2016.

[81] L. P. Kaelbling and T. Lozano-Pérez, "Unifying perception, estimation and action for mobile manipulation via belief space planning," in *Proc. IEEE Conf. Robot. Autom.*, 2012, pp. 2952–2959.

[82] M. Dogar, M. Koval, A. Tallavajhula, and S. Srinivasa, "Object search by manipulation," *Auton. Robots*, vol. 36, no. 1/2, pp. 153–167, 2014.

[83] R. Platt, L. P. Kaelbling, T. Lozano-Pérez, and R. Tedrake, "Efficient planning in non-gaussian belief spaces and its application to robot grasping," in *Proc. Int. Symp. Robot. Res.*, 2011.

[84] K. Hsiao, P. Nangeroni, M. Huber, A. Saxena, and A. Y. Ng, "Reactive grasping using optical proximity sensors," in *Proceedings of the 2009 IEEE International Conference on Robotics Automation.* Piscataway, NJ, USA: IEEE Press, 2009, pp. 4230–4237.

[85] O. Kroemer, R. Detry, J. Piater, and J. Peters, "Combining active learning and reactive control for robot grasping," *Robot. Auton. Syst.*, vol. 58, no. 9, pp. 1105–1116, 2010.

[86] A. Boularias, J. A. D. Bagnell, and A. T. Stentz, "Learning to manipulate unknown objects in clutter by reinforcement," in *Proc. 29th AAAI Conf. Artif. Intell.*, Jan. 2015, pp. 1336–1342.

[87] S. Dragiev, M. Toussaint, and M. Gienger, "Uncertainty-aware grasping and tactile exploration," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2013, pp. 113–119.

[88] M. Krainin, B. Curless, and D. Fox, "Autonomous generation of complete 3D object models using next best view manipulation planning," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2011, pp. 5031–5037.

[89] J. Ilonen, J. Bohg, and V. Kyrki, "Three-dimensional object reconstruction of symmetric objects by fusing visual and tactile sensing," *Int. J. Robot. Res.*, vol. 33, no. 2, pp. 321–341, 2014.

[90] M. Björkman, Y. Bekiroglu, V. Hogman, and D. Kragic, "Enhancing visual perception of shape through tactile glances," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.,* Tokyo, Japan, Nov. 3–7, 2013, pp. 3180–3186.

[91] H. Culbertson, J. Unwin, and K. J. Kuchenbecker, "Modeling and rendering realistic textures from unconstrained tool-surface interactions," *IEEE Trans. Haptics*, vol. 7, no. 3, pp. 381–393, Jul.–Sep. 2014.

[92] J. Romano and K. Kuchenbecker, "Creating realistic virtual textures from contact acceleration data," *IEEE Trans. Haptics*, vol. 5, no. 2, pp. 109–119, Apr. 2012.

[93] O. Kroemer, C. H. Lampert, and J. Peters, "Learning dynamic tactile sensing with robust vision-based training," *IEEE Trans. Robot.*, vol. 27, no. 3, pp. 545–557, Jun. 2011.

[94] A. Jain and C. C. Kemp, "Pulling open doors and drawers: Coordinating an omni-directional base and a compliant arm with equilibrium point control," in *Proc. IEEE Int. Conf. Robot. Autom.,* 2010, pp. 1807–1814.

[95] J. Sturm, C. Stachniss, and W. Burgard, "A probabilistic framework for learning kinematic models of articulated objects," *J. Artif. Intell. Res.*, vol. 41, pp. 477–526, 2011.

[96] R. M. Martín and O. Brock, "Online interactive perception of articulated objects with multi-level recursive estimation based on task-specific priors," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2014, pp. 2494–2501.

[97] D. Katz and O. Brock, "A factorization approach to manipulation in unstructured environments," in *Robotics Research.* New York, NY, USA: Springer, 2011, pp. 285–300.

[98] K. Hausman, S. Niekum, S. Osentoski, and G. S. Sukhatme, "Active articulation model estimation through interactive perception," in *Proc. Int. Conf. Robot. Autom.*, May 2015, pp. 3305–3312.

[99] S. Pillai, M. Walter, and S. Teller, "Learning articulated motions from visual demonstration," in *Proc. Robot., Sci. Syst.*, Berkeley, CA, USA, Jul. 2014.

[100] P. R. Barragán, L. P. Kaelbling, and T. Lozano-Pérez, "Interactive Bayesian identification of kinematic mechanisms," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2014, pp. 2013–2020.

[101] S. Otte, J. Kulick, M. Toussaint, and O. Brock, "Entropy-based strategies for physical exploration of the environments degrees of freedom," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2014, pp. 615–622.

[102] Y. Karayiannidis, C. Smith, F. Vina, P. Ögren, and D. Kragic, "Model-free robot manipulation of doors and drawers by means of fixed-grasps," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2013, pp. 4470–4477.

[103] D. Katz, A. Orthey, and O. Brock, *Interactive Perception of Articulated Objects* (ser. Springer Tracts in Advanced Robotics). New York, NY, USA: Springer, 2014, vol. 79, pp. 301–315.

[104] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 898–916, May 2011.

[105] R. Xiaofeng and L. Bo, "Discriminatively trained sparse code gradients for contour detection," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds. Red Hook, NY, USA: Curran Assoc., 2012, pp. 584–592.

[106] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, Jun. 7–12, 2015, pp. 3431–3440.

[107] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Represent. Learn.*, 2015.

[108] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Red Hook, NY, USA: Curran Assoc., 2015, pp. 91–99.

[109] J. Pajarinen and V. Kyrki, "Decision making under uncertain segmentations," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2015, pp. 1303–1309.

[110] R. Martín-Martín, S. Höfer, and O. Brock, "An integrated approach to visual perception of articulated objects," in *Proc. IEEE Int. Conf. Robot.*, 2016, pp. 5091–5097.

[111] L. Natale and E. Torres-Jara, "A sensitive approach to grasping," in *Proc. 6th Int. Workshop Epigenetic Robot.*, 2006, pp. 87–94.

[112] S. Levine, P. Pastor, A. Krizhevsky, and D. Quillen, "Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection," in *Proc. Int. Symp. Exp. Robot.*, Tokyo, Japan, 2016.

[113] K. Hsiao, L. P. Kaelbling, and T. Lozano-Pérez, "Robust grasping under object pose uncertainty," *Auton. Robots*, vol. 31, no. 2–3, pp. 253–268, 2011.

[114] L. Pinto and A. Gupta, "Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours," in *Proc. 2016 IEEE Int. Conf. Robot. and Autom.*, May 2016, pp. 3406–3413.

[115] T. Mar, V. Tikhanoff, G. Metta, and L. Natale, "Self-supervised learning of grasp dependent tool affordances on the iCub humanoid robot," in *Proc. IEEE Int. Conf. Robot. Autom*., 2015, pp. 3200–3206.

[116] J. Pajarinen and V. Kyrki, "Robotic manipulation in object composition space," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2014, pp. 1–6.

[117] L. Natale, F. Orabona, G. Metta, and G. Sandini, "Sensorimotor coordination in a baby robot: Learning about objects through grasping," in *From Action to Cognition* (ser. Progress in Brain Research), vol. 164, C. von Hofsten and K. Rosander, Eds. Amsterdam, The Netherlands: Elsevier, 2007, pp. 403–424.

[118] B. Browatzki, V. Tikhanoff, G. Metta, H. H. Blthoff, and C. Wallraven, "Active in-hand object recognition on a humanoid robot," *IEEE Trans. Robot.*, vol. 30, no. 5, pp. 1260–1269, Oct. 2014.

[119] A. Tsuda, Y. Kakiuchi, S. Nozawa, R. Ueda, K. Okada, and M. Inaba, "Grasp, motion, view planning on dual-arm humanoid for manipulating in-hand object," in *Proc. IEEE Workshop Adv. Robot. Social Impacts*, Oct. 2011, pp. 54–57.

[120] E. Torres-Jara, L. Natale, and P. Fitzpatrick, "Tapping into touch," *Proceedings of the 5th Int. Workshop Epigenetic Robot.*, Osaka, Japan, 2005, pp. 79–86.

[121] D. Kraft *et al.*, "Birth of the object: Detection of objectness and extraction of object shape through object-action complexes," *Int. J. Humanoid Robot.*, vol. 5, no. 2, pp. 247–265, 2008.

[122] D. Omrcen, A. Ude, K. Welke, T. Asfour, and R. Dillmann, "Sensorimotor processes for learning object representations," in *Proc. 7th IEEE-RAS Int. Conf. Humanoid Robots,* Pittsburgh, PA, USA, Nov. 29–Dec. 1, 2007, pp. 143–150.

[123] D. Michel, X. Zabulis, and A. A. Argyros, "Shape from interaction," *Mach. Vis. Appl.*, vol. 25, no. 4, pp. 1077–1087, May 2014.

[124] A. Byravan and D. Fox, "SE3-Nets: Learning rigid body motion using deep neural networks," *Proc. IEEE Int. Conf. Robotics Automat.*, May 2016.

[125] S. M. LaValle, *Planning Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2006. [Online]. Available: http://planning.cs.uiuc.edu/

[126] J. Kober, J. A. D. Bagnell, and J. Peters, "Reinforcement learning in robotics: A survey," *Int. J. Robot. Res.*, Jul. 2013, pp 579–610.

[127] G. Shani, J. Pineau, and R. Kaplow, "A survey of point-based POMDP solvers," *Auton. Agents Multi-Agent Syst.*, vol. 27, no. 1, pp. 1–51, Jul. 2013.

[128] S. Singh, M. R. James, and M. R. Rudary, "Predictive state representations: A new theory for modeling dynamical systems," in *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*. Arlington, VA, USA: AUAI Press, 2004, pp. 512–519.

[129] B. Boots, S. M. Siddiqi, and G. J. Gordon, "Closing the learning-planning loop with predictive state representations," *Int. J. Robot. Res.*, vol. 30, no. 7, pp. 954–966, Jun. 2011.

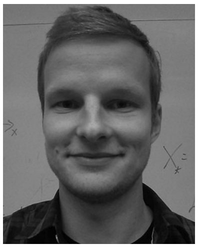
[130] J. A. Stork, C. H. Ek, Y. Bekiroglu, and D. Kragic, "Learning predictive state representation for in-hand manipulation," in *Proc. 2015 IEEE Int. Conf. Robot. Autom.*, May 2015, pp. 3207–3214.

[131] C. Atkeson *et al.*, "No falls, no resets: Reliable humanoid behaviour in the DARPA robotics challenge," in *Proc. 15th IEEE/RAS Int. Conf. Humanoid Robots*, 2015, pp. 623–630.

[132] M. Johnson *et al.*, "Team IHMC's lessons learned from the DARPA robotics challenge trials," *J. Field Robot.*, vol. 32, no. 2, pp. 192–208, 2015.

Jeannette Bohg (M'16) received the Diploma degree in computer science from Technical University Dresden, Dresden, Germany; the M.Sc. degree in applied information technology from Chalmers University of Technology, Göteborg, Sweden; and the Ph.D. degree in robotics and computer vision from Royal Institute of Technology (KTH), Stockholm, Sweden, in 2012.

She is currently an Assistant Professor of robotics in the Department of Computer Science, Stanford University, Stanford, CA, USA. She is also a Guest Researcher in the Autonomous Motion Department, Max Planck Institute for Intelligent Systems, Tübingen, Germany, where she was a Research Group Leader until fall 2017. Her research interest include the intersection between computer vision, robotic manipulation, and machine learning. Specifically, she analyzes how continuous feedback from multiple sensor modalities helps to improve autonomous manipulation capabilities of a robot.



Karol Hausman (S'15) received the M.E. degree in mechatronics from Warsaw University of Technology, Warsaw, Poland, in 2012 and the M.Sc. degree in robotics, cognition, and intelligence from Technical University Munich, Munich, Germany, in 2013. He is currently working toward the Ph.D. degree in computer science at University of Southern California, Los Angeles, CA, USA.

His research interests include active state estimation, control generation, and machine learning for robotics. More specifically, he is concentrating on model-based and learning-based approaches that address aspects of closing perception–action loops.

Bharath Sankaran (S'15) received the B.E degree in mechanical engineering from Anna University, Chennai, India, in 2006; the M.E. degree in aerospace engineering from University of Maryland, College Park, MD, USA, in 2008; the M.S. degree in robotics from University of Pennsylvania, Philadelphia, PA, USA, in 2012; and the M.S. degree in computer science in 2015 from University of Southern California, Los Angeles, CA, USA, where he is currently working toward the Ph.D. degree in computer science

His research interests include applying statistical learning techniques to perception and control problems in robotics. Here, he primarily focuses on treating traditional computer vision problems as problems of active and interactive perception.

Oliver Brock (SM'15) received the Diploma degree in computer science from Technische Universität Berlin, Berlin, Germany, in 1993 and the Master's and Ph.D. degrees in computer science from Stanford University, Stanford, CA, USA, in 1994 and 2000, respectively.

He is currently the Alexander von Humboldt Professor of Robotics in the School of Electrical Engineering and Computer Science, Technische Universität Berlin, Berlin, Germany. He held Postdoctoral positions with Rice University and Stanford University. In 2002, he was an Assistant Professor and an Associate Professor in the Department of Computer Science, University of Massachusetts Amherst, before moving back to Technische Universität Berlin in 2009. The research of his lab, the Robotics and Biology Laboratory, focuses on autonomous mobile manipulation, interactive perception, grasping, manipulation, soft hands, interactive learning, motion generation, and the application of algorithms and concepts from robotics to computational problems in structural molecular biology.

Dr. Brock is the President of the Robotics: Science and Systems Foundation.

Danica Kragic (F'16) received the M.Sc. degree in mechanical engineering from Technical University of Rijeka, Rijeka, Croatia, in 1995 and the Ph.D. degree in computer science from Royal Institute of Technology, KTH, Stockholm, Sweden, in 2001.

She is currently a Professor in the School of Computer Science and Communication, Royal Institute of Technology, KTH. She had been a Visiting Researcher with Columbia University, Johns Hopkins University, and INRIA Rennes. She is the Director of the Centre for Autonomous Systems. Her research interests include the areas of robotics, computer vision, and machine learning.

Dr. Kragic received the 2007 IEEE Robotics and Automation Society Early Academic Career Award. She is a member of the Royal Swedish Academy of Sciences and Young Academy of Sweden. She holds an Honorary Doctorate from Lappeenranta University of Technology, Lappeenranta, Finland. She chaired the IEEE RAS Technical Committee on Computer and Robot Vision and served as an IEEE RAS AdCom member. In 2012, she received an ERC Starting Grant.

Stefan Schaal (F'16) received the Ph.D. degree in artificial intelligence from Technical University of Munich in 1991. He is currently a Professor of computer science, neuroscience, and biomedical engineering with University of Southern California, Los Angeles, CA, USA, and a Founding Director of the Max Planck Institute for Intelligent Systems, Tübingen, Germany. In the past, he held positions as a Postdoctoral Fellow in the Department of Brain and Cognitive Sciences and the Artificial Intelligence Laboratory at MIT, as an Invited Researcher with the ATR Human Information Processing Research Laboratories in Japan, as an Adjunct Assistant Professor with the Georgia Institute of Technology, as an Adjunct Assistant Professor in the Department of Kinesiology, Pennsylvania State University, and as a Group Leader of an ERATO Project, Japan. His research interests include statistical and machine learning, neural networks, computational neuroscience, functional brain imaging, nonlinear dynamics, nonlinear control theory, and biomimetic robotics. He applies his research to problems of autonomous systems, artificial and biological motor control, and motor learning, focusing on both theoretical investigations and experiments with human subjects and anthropomorphic robot equipment.

Dr. Schaal is a cofounder of the IEEE/RAS International Conference and Humanoid Robotics, and a cofounder of Robotics Science and Systems Conference, a highly selective new conference featuring the best work in robotics every year. He served as the Program Chair at these conferences and he was the Program Chair of the International Conference on Simulated and Adaptive Behavior (2004) and the IEEE/RAS International Conference on Robotics and Automation (2008), the largest robotics conference in the world. He has also been an Area Chair at Neural Information Processing Systems and served as the Program Committee Member of the International Conference on Machine Learning. He serves on the editorial board of *Neural Networks*, *International Journal of Humanoid Robotics*, and *Frontiers in Neurorobotics*. He is a Member of the German National Academic Foundation (Studienstiftung des Deutschen Volkes), the Max Planck Society, the Alexander von Humboldt Foundation, the Society for Neuroscience, the Society for Neural Control of Movement, and AAAS.

Gaurav S. Sukhatme (F'10) received the Undergraduate degree in computer science and engineering from Indian Institute of Technology Bombay, Mumbai, India, in 1991, and the M.S. and Ph.D. degrees in computer science from University of Southern California (USC), Los Angeles, CA, USA, in 1993 and 1997, respectively.

He is the Fletcher Jones Professor of computer science (joint appointment in electrical engineering) with USC, and the Executive Vice Dean of the USC Viterbi School of Engineering. He is the Codirector of the USC Robotics Research Laboratory and the Director of the USC Robotic Embedded Systems Laboratory, which he founded in 2000. His research interests include perception, planning, and robot networks with applications to environmental monitoring. He has published extensively in these and related areas.

Dr. Sukhatme has served as a Principal Investigator on numerous NSF, DARPA and NASA grants. He was a Co-Principal Investigator of the Center for Embedded Networked Sensing, an NSF Science and Technology Center. He received the NSF CAREER Award and the Okawa Foundation Research Award. He is one of the founders of the Robotics: Science and Systems Conference. He was the Program Chair of the 2008 IEEE International Conference on Robotics and Automation and the 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems. He is the Editor-in-Chief of *Autonomous Robots* and has served as an Associate Editor of IEEE TRANSACTIONS ON ROBOTICS AND AUTOMATION, IEEE TRANSACTIONS ON MOBILE COMPUTING, and on the Editorial Board of IEEE PERVASIVE COMPUTING.