Building Kinematic and Dynamic Models of Articulated Objects with Multi-Modal Interactive Perception

Roberto Martín-Martín and **Oliver Brock**
Robotics and Biology Laboratory
Technische Universität Berlin, Germany*

Abstract

Interactive perception methods exploit the correlation between robot actions and changes in the sensor signals to extract task-relevant information from the sensor stream. These interactions can have effects in one or multiple sensor modalities. In this work we propose an interactive perception system to build kinematic and dynamic models of articulated objects from multi-modal sensor streams. Our system leverages two properties of the multi-modal signals: 1) consequences of actions are frequently best perceived in the combined multi-modal sensor-action stream, and 2) properties perceived in one modality may depend on properties perceived in other modalities. To fully exploit these properties we propose a perceptual architecture based on two main components: a set of (possibly recursive) estimation processes and an interconnection mechanism between them. The processes extract patterns of information from the multi-modal sensor-action stream by enriching it with prior knowledge; the interconnection mechanism leverages their inter-dependencies by using estimations of one process as measurements, priors or forward models of other processes. We show experimentally that our system integrates and balances between different modalities according to their uncertainty in order to overcome limitations of uni-modal perception.

Introduction

Interactive perception is a way to approach perceptual problems that makes interaction part of the solution by leveraging it as source of perceptual information (Bohg et al. 2016). Interactive perception methods have been successfully applied to segment the visual field, perceive shape, kinematic structures and dynamic properties, by exploiting the interaction capabilities of the robot's embodiment.

Traditionally, interactive perception methods have been based on one single sensor modality, most frequently vision (Fitzpatrick 2003; Kenney, Buckley, and Brock 2009; Hausman et al. 2013; Katz et al. 2013; van Hoof et al. 2012), force-torque sensor signals (Atkeson, An, and Hollerbach 1986; Karayiannidis et al. 2013; Endres, Trinkle, and Burgard 2013) or tactile feedback (Dragiev, Toussaint, and Gienger 2013). These uni-modal methods fail when the sensor modality does not contain the necessary perceptual information. These failures can be caused by changes in the environmental conditions (e.g. lights go off), in the characteristics of the interaction (e.g. self occlusions or grasp/contact loss), or in the perceptual task.

While some authors realized that these failures could be alleviated by integrating multiple sensor modalities (Hausman et al. 2015; Sukhoy and Stoytchev 2010; Bohg et al. 2010; Illonen, Bohg, and Kyrki 2013; Park et al. 2016), there is no generic approach to merge multiple sensor modalities and information about the interaction in a multi-purpose perceptual system. In this work we present a multi-modal interactive perception system that integrates information from

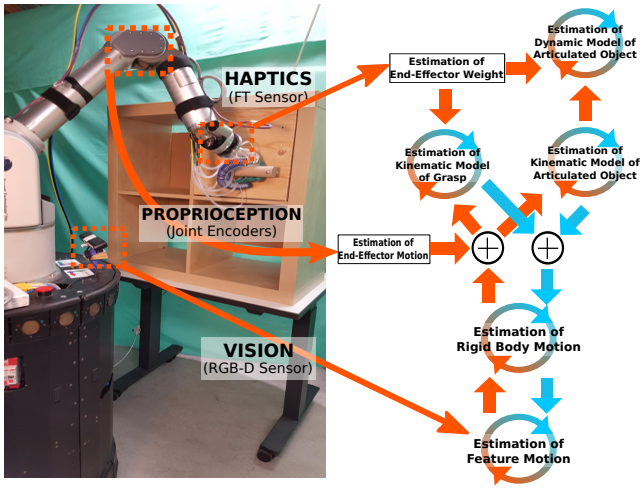

Figure 1: Our robot interacting with an articulated object and schematic representation of our multi-modal interactive perception system; the system is composed of several interconnected estimation processes; two-colored loops represent recursive processes; orange arrows indicate inputs to estimation processes (sensor streams or estimations from other processes); blue arrows indicate predictions

*All authors are with the Robotics and Biology Laboratory, Technische Universität Berlin, Germany. We gratefully acknowledge the funding provided by the Alexander von Humboldt foundation and the Federal Ministry of Education and Research (BMBF), by the European Commision (EC, SOMA, H2020-ICT-645599) and the German Research Foundation (DFG, Exploration Challenge, BR 2248/3-1).

multiple sensor modalities to extract information more robustly in dynamically changing environments. Our proposed system is in essence an interconnected network of (possibly recursive) estimation processes. Each process extracts different perceptual patterns based on priors about 1) the temporal evolution of the environment and the signals, and 2) its consequences in the multi-modal sensor-action stream. The inter-communication between processes allows them to bootstrap and feed information into each other and to reuse priors at different levels, effectively leveraging the correlation between modalities.

We apply our system to the perception of articulated objects from multi-modal sensor-action signals. Our goal is to perceive in an online manner the kinematic and dynamic properties of the articulated objects by integrating information from vision, force-torque sensing and proprioception. Such an online perceptual system supports and enables successful reactive autonomous interaction with unknown articulated objects.

In the following we present our online multi-modal interactive perception system to perceive articulated objects from interactions and show some preliminary results.

Online Multi-Modal Interactive Perception

We propose to realize online multi-modal interactive perception for autonomous robots as a network of interconnected estimation processes. Most of these processes are recursive processes (Thrun, Burgard, and Fox 2005). Recursive estimation is well suited for interactive perception because 1) it leverages the temporal coherence of the multi-modal sensor-action stream, and 2) it extracts patterns of information (e.g. motion) by interpreting a (possibly multi-modal) sensor-action stream as evidences of an underlying dynamic process.

In order to extract patterns of information a recursive estimation process makes use of prior knowledge in two forms: in the form of a process model, to predict the dynamics of the process (possibly affected by the interaction), and in the form of a measurement model, to explain the correlations between the process and the input stream. Several process models and measurement models are possible within a single estimation process. If the estimation process is not recursive, we assume that there is no process model and the prior knowledge is entirely encoded in the measurement model.

Recursive estimation can be applied to different types of input streams. First, the input stream can include information about interactions, making recursive estimation suited for Interactive Perception. Second, the input stream can include multiple modalities, making recursive estimation a good solution for multi-modal perception. Third, the input stream can contain patterns of information extracted by other processes, allowing for a bottom-up hierarchical perceptual process.

This last property defines one of the intercommunication flows that defines our network of estimation processes. The second intercommunication flow is composed of predictions (states, measurements), that can be used as alternative process models and/or as measurements by other processes. This top-down communication allows to reuse high level priors and to inject high level information into lower level perceptual processes, effectively restricting their space of possible states.

Building Kinematic and Dynamic Models of Articulated Objects

Based on the aforementioned components, we developed a system that perceives articulated objects from interactions. The system integrates visual information (RGB-D images), haptics (force-torque sensor measurements) and proprioception (joint encoder measurements) to infer the kinematic and dynamic properties of the interacted object.

This system of this work is a multi-modal extension of our system presented in (Martín-Martín and Brock 2014). In our previous system, we used only visual information provided by an RGB-D sensor and factorized the original problem (the perception of kinematic structures) into three estimation subproblems. The first subproblem is to extract patterns of local motion from the visual stream. We address this problem by estimating recursively the motion of a set salient 3D point features. The second subproblem is the extraction of patters of rigid body motion from the local patterns of motion. We tackle this problem with a set of recursive filters (extended Kalman filters) that interpret the motion of the features as motion of rigid bodies. The last subproblem is the extraction of patterns of motion of pairs of bodies, which define the kinematic model. For this problem we instantiate a set of recursive filters (again extended Kalman filters) per joint that infer the joint parameters and joint state from the motion of the rigid bodies. The combined result of these filters defines the kinematic structure and state of the interacted articulated object.

Our original system presents the limitations of a unimodal perceptual system: it fails if the visual stream does not contain the relevant information for the perceptual task due to characteristics of the environment or the task (e.g. occlusions). Clearly, when the robot interacts with an articulated object, much richer patterns of information can be extracted from the multi-modal sensor-action that includes force-torque sensor signals and proprioceptive information. In this work we propose to overcome the limitations of our original system by exploiting this richer information of the multi-modal stream (see Fig. 1).

The proprioception information (raw joint encoder values) is first merged with prior information about the kinematics of the robot manipulator to obtain an estimation of the pose of the end-effector. This is not a recursive process because the rich available prior information (the kinematic model of the robot) is enough to uniquely estimate the pose of the end-effector (forward kinematics) and does not require to exploit the temporal correlation of the proprioceptory sensor stream.

The outcome of this first estimation process is integrated with the estimations obtained from the visual stream to obtained a combined estimation of the pose of the end-effector. Both sources are integrated in a probabilistic fashion, weighting their contribution to the multi-modal estima-

tion based on their relative uncertainty. The information from proprioception is more reliable than the visual information and therefore dominates the final estimate. However, by merging both modalities we can explain the motion patterns of the visual field as evidences of the motion of the end-effector, and overcome situations that decrease the reliability or the amount of information contained in the visual stream, e.g. due to occlusions or due to an abrupt change in the lighting conditions.

We extended our previous solution for the third subproblem, the extraction of patterns of motion of pairs of rigid bodies, to fully exploit the multi-modal information of the sensor stream. Additionally to the estimation of the kinematic structure of the articulated object we estimate the grasping interface between the end-effector and the rigid body under direct interaction. We create four new models that define this kinematic relationship: perfect grasp, cylindrical grasp with rotation around the main axis, cylindrical grasp with rotation and translation along the main axis and ungrasped. These four models cover the most frequent types of grasp performed by humanoid hands (like the one we use in our experiments, see next Section) when actuating articulated objects. They constrain 6, 5, 4 or none of the 6 DoF of relative motion between the hand and the object, respectively.

In order to select the most likely grasp model we integrate information from vision, proprioception and haptics (from a force-torque sensor mounted on robot's wrist). We define a multi-modal measurement model for each new grasp model and use them to generate predictions about the expected multi-modal signals. The model that best predicts the multi-modal measurements at each step is selected (see see Sec. Experiments). The grasp models predicted force-torque sensor signals due to the kinematic constraint between hand and interacted body, but not due to the dynamics of the hand (weight). Therefore, the raw force-torque signal is first merged with prior information about the hand (mass and center of mass) and with the estimation of the current end-effector pose to obtain the force-torque signal without the weight of the manipulator.

Once we have identified the most likely grasp model, we can revert this model and use it as source of information to perceive the motion of the object under interaction based on the motion of the end-effector. This new estimation of the motion of the interacted body and the estimation obtained from visual information are merged based on their relative uncertainty (weighted average), similarly to how we integrated multiple evidences of the motion of the end-effector. However, in this integration both estimations are of comparable uncertainty. The result of the integration will be balanced between both inputs depending on the number and reliability of 3D features perceived on the interacted rigid body. The uncertainty of the 3D features depends in turn on the depth of the feature (the accuracy of the RGB-D camera decreases with the depth) and the lightness of the color images (the noise in the location of the 3D features increases in dark environments).

Finally, we extended our system with a new process to perceive the dynamic properties of articulated objects. This new process integrates information about the kinematics with the estimated force-torque signal without the weight of the end effector. The process generates a dynamic profile that characterizes the interaction with the articulated object (see Sec. Experiments). In a nutshell, the process uses the estimated kinematic model to project the force signal exerted by the robot (without the weight of the hand) into the tangential and the normal directions of the kinematically allowed dimension of motion for the interacted joint. Only the tangential force causes motion of the mechanism while the normal force is absorbed by the kinematic constraints of the articulation.

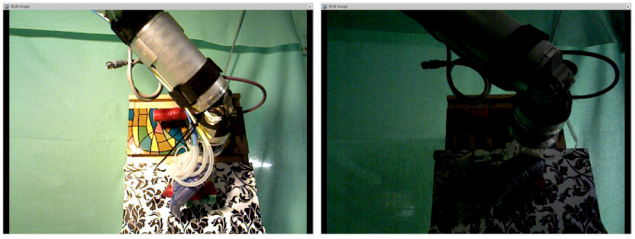

Figure 2: Robot view at two time-steps of the experiment; (left) robot view with normal lighting conditions; (right) robot view after the lights go off (most of the visual features are lost); the robot balances the uncertainty between the visual and the proprioceptive modalities and perceives correctly the kinematic properties of the object in both conditions

Experiments

We tested our system for the task of building a kinematic and dynamic model of an unknown articulated object, a cabinet with a drawer, from interactions. Our robot is a WAM 7-DoF arm and an RBO-2 soft-hand (Deimel and Brock 2015) as end-effector. The robot has an Asus RGB-D sensor mounted on the base of the arm and a ATI 6-DoF force-torque sensor mounted on the wrist.

To trigger the interaction we guide the robot's hand to one of the handles in the environment, command it to close the hand and start a compliant pulling interaction that exploits the force-torque sensor readings to guide the manipulation to the allowed dimensions of motion of the object. We use a velocity-force controller that tries to keep a certain linear velocity for the end-effector, increasing or decreasing the applied force if the velocity is under or over the desired value (1 cm/s). The orientation of the end-effector is adapted compliantly to the allowed dimension of motion of the articulated object.

Robustness against changing environmental and task conditions

In the first experiment, we evaluate the robustness of our multi-modal system to changes in the environmental and task conditions, by switching on and off the lights (Fig. 2) and causing the robot to lose the grasp. In the first phase the robot observes its own interaction with an articulated object

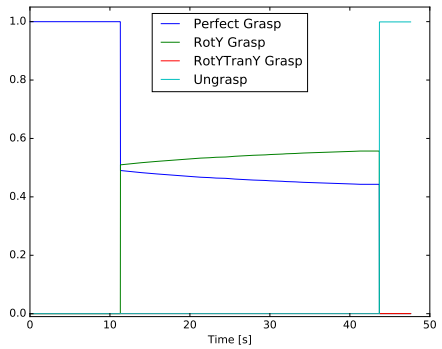

Figure 3: Probability of the different grasping models between the end-effector and the articulated object; initially the most probable joint type is perfect grasp; after enough visual and proprioceptive information has been acquired, the most probable grasp model is cylindrical grasp with allowed rotation around the y axis; at the end of the interaction the robot perceives that the grasp was lost based on the force-torque measurements and correctly assigns the highest probability to the ungrasp model

(a drawer) and determines the most probable grasp model integrating the evidences from vision, proprioception and haptic information. At some point of the interaction the lights go off and the robot uses proprioception (through the estimated grasp model) as most reliable source of information to perceive and update the kinematic state. Eventually, the robot perceives that the grasp was lost and stops using proprioception to update the kinematic state of the object (Fig. 3). Figure 4 depicts the result of the grasp model selection. The final configuration of the mechanism is 21.0 cm while the robot perceived 21.9 cm. Our perceptual system combines and balances the reliability of the different sources of information and is unaffected by abrupt environmental changes.

Perceiving dynamic properties

In the second experiment, we evaluate the integration of vision, proprioception and haptics to generate a dynamic model of the articulated object. The plot at the bottom of Fig. 5 depicts the dynamic profile perceived from the interaction with the articulated object. The new process integrates the kinematic structure perceived by another process based on the multi-modal vision-proprioception stream with haptic information, and separates the force exerted by the robot into forces in the normal and the tangential directions to the kinematically allowed dimension of motion.

The plot at the top of Fig. 5 depicts the joint state and the joint velocity associated to the interaction. The *saw-like* profile of the force plot and the *stair-like* profile of the joint state is the effect of the velocity-force compliant controller and the stiction of the mechanism: under certain tangential force (around 2.5 N of tangential force) stiction dominates the interaction and there is no motion. This causes the desired force of the controller to increase until overcomes stiction, which triggers a fast joint motion of the object. The velocity generated is over the desired value for the controller

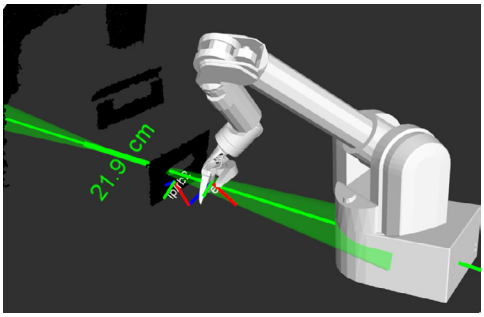

Figure 4: 3-D visualization of the final step of the interaction with an articulated object; the black region is the received colored point cloud when the lights are off, which is the visual input of our algorithm; the green line indicates the perceived prismatic joint (translucent cone indicates uncertainty); the gray model is a model of our robot placed using the joint encoder values; note that the grasp was lost at the end of the experiment and the hand is away from the articulated object

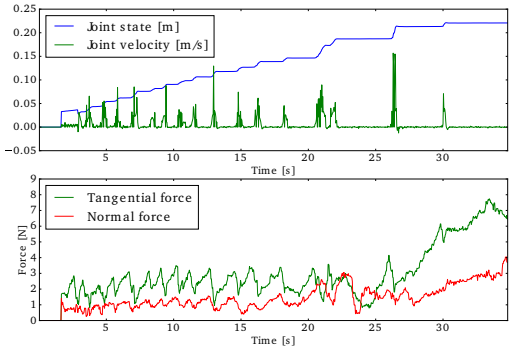

Figure 5: Dynamic profile of an articulated object perceived with multi-modal Interactive Perception; (top) joint state and joint velocity of the articulated object; (bottom) force applied by the robot on the mechanism, separated into tangential and normal forces; only normal forces cause motion

and causes the reference force to decrease under the stiction threshold. At the end of the interaction the tangential force increases drastically because the joint limit was reached.

This preliminary experiment indicates that is possible to perceive dynamic properties of the articulated object in an online manner and that they contain crucial information to improve ongoing interactions.

References

[Atkeson, An, and Hollerbach 1986] Atkeson, C. G.; An, C. H.; and Hollerbach, J. M. 1986. Estimation of Inertial Parameters of Manipulator Loads and Links. *The International Journal of Robotics Research (IJRR)* 5(3):101–119.

[Bohg et al. 2010] Bohg, J.; Johnson-Roberson, M.; Björkman, M.; and Kragic, D. 2010. Strategies for multi-modal scene exploration. In *Intelligent Robots and*

Systems (IROS), 2010 IEEE/RSJ International Conference on, 4509–4515.

[Bohg et al. 2016] Bohg, J.; Hausman, K.; Sankaran, B.; Brock, O.; Kragic, D.; Schaal, S.; and Sukhatme, G. S. 2016. Interactive perception: Leveraging action in perception and perception in action. *CoRR* abs/1604.03670.

[Deimel and Brock 2015] Deimel, R., and Brock, O. 2015. A novel type of compliant and underactuated robotic hand for dexterous grasping. *The International Journal of Robotics Research*.

[Dragiev, Toussaint, and Gienger 2013] Dragiev, S.; Toussaint, M.; and Gienger, M. 2013. Uncertainty aware grasping and tactile exploration. In *Proceedings of the IEEE International Conference on Robotics and Automotion (ICRA)*.

[Endres, Trinkle, and Burgard 2013] Endres, F.; Trinkle, J.; and Burgard, W. 2013. Learning the Dynamics of Doors for Robotic Manipulation. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*.

[Fitzpatrick 2003] Fitzpatrick, P. 2003. First contact: an active vision approach to segmentation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, volume 3, 2161–2166.

[Hausman et al. 2013] Hausman, K.; Balint-Benczedi, F.; Pangercic, D.; Marton, Z.-C.; Ueda, R.; Okada, K.; and Beetz, M. 2013. Tracking-based interactive segmentation of textureless objects. In *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, 1122–1129.

[Hausman et al. 2015] Hausman, K.; Niekum, S.; Osentoski, S.; and Sukhatme, G. S. 2015. Active Articulation Model Estimation through Interactive Perception. In *International Conference on Robotics and Automotion*.

[Illonen, Bohg, and Kyrki 2013] Illonen, J.; Bohg, J.; and Kyrki, V. 2013. 3-d object reconstruction of symmetric objects by fusing visual and tactile sensing. *The International Journal of Robotics Research* 33(2):321–341.

[Karayiannidis et al. 2013] Karayiannidis, Y.; Smith, C.; Vina, F. E.; Ogren, P.; and Kragic, D. 2013. Model-free robot manipulation of doors and drawers by means of fixed-grasps. In *Proceedings of the IEEE International Conference on Robotics and Automotion (ICRA)*.

[Katz et al. 2013] Katz, D.; Kazemi, M.; Bagnell, J.; and Stentz, A. 2013. Interactive segmentation, tracking, and kinematic modeling of unknown 3d articulated objects. In *IEEE International Conference on Robotics and Automation*, 5003–5010.

[Kenney, Buckley, and Brock 2009] Kenney, J.; Buckley, T.; and Brock, O. 2009. Interactive segmentation for manipulation in unstructured environments. In *IEEE International Conference on Robotics and Automation*, 1377–1382.

[Martín-Martín and Brock 2014] Martín-Martín, R., and Brock, O. 2014. Online Interactive Perception of Articulated Objects with Multi-Level Recursive Estimation Based on Task-Specific Priors. In *International Conference on Intelligent Robots and Systems*.

[Park et al. 2016] Park, D.; Erickson, Z. M.; Bhattacharjee, T.; and Kemp, C. C. 2016. Multimodal execution monitoring for anomaly detection during robot manipulation. In *2016 IEEE International Conference on Robotics and Automation, ICRA 2016, Stockholm, Sweden, May 16-21, 2016*, 407–414.

[Sukhoy and Stoytchev 2010] Sukhoy, V., and Stoytchev, A. 2010. Learning to detect the functional components of doorbell buttons using active exploration and multimodal correlation. In *2010 10th IEEE-RAS International Conference on Humanoid Robots*, 572–579.

[Thrun, Burgard, and Fox 2005] Thrun, S.; Burgard, W.; and Fox, D. 2005. *Probabilistic Robotics (Intelligent Robotics and Autonomous Agents)*. The MIT Press.

[van Hoof et al. 2012] van Hoof, H.; Kroemer, O.; Amor, H. B.; and Peters, J. 2012. Maximally Informative Interaction Learning for Scene Exploration. In *Proc. of Int. Conf. on Intelligent Robots and Systems (IROS)*, 5152–5158.