Cross-Modal Interpretation of Multi-Modal Sensor Streams in Interactive Perception Based on Coupled Recursion

Roberto Martín-Martín Oliver Brock

Abstract—We present an online system to perceive kinematic properties of articulated objects from multi-modal sensor streams. The novelty of our system is that it leverages multi-modal information in a cross-modal manner: instead of simply fusing information from different modalities, sensor streams are interpreted by leveraging information from another modality. We realize each cross-modal information extraction process using recursive estimation, with each process addressing a perceptual subproblem. Several estimators are then coupled in a cross-modal network, leading to efficient and robust online perception. We demonstrate experimentally that our cross-modal system improves over its uni-modal counterparts, increasing the variability of environments and task conditions in which the robot can robustly perceive the articulated objects. We further demonstrate that these perceptual abilities are sufficiently fast to provide feedback during manipulation actions and sufficiently comprehensive to allow the generation of new manipulation actions.

I. INTRODUCTION

We propose a system for the cross-modal perception of kinematic properties of articulated objects. Cross-modal perception is a form of multi-modal perception that leverages the information obtained in one modality to facilitate the interpretation of another modality (and possibly vice-versa). Cross-modal integration is able to leverage regularities in the combined multi-modal signal space, whereas traditional multi-modal perception interprets the individual sensor signals independently and then combines the results.

An example of cross-modal perception is the McGurk effect [1]. In this perceptual illusion, a subject watches a video of a person pronouncing syllables. The subject is convinced to hear different utterances, such as ba-ba or ga-ga, when in fact the sounds are identical. The subjects misjudges the sound because the video in fact shows the person saying different syllables but the sound has been altered to play the identical syllable. The illusion occurs because the visual cue produced by the facial motions influences the perception of the sound. As a result, identical sounds are perceived as being different. This illusion demonstrates that visual cues affect hearing. This cross-modal interpretation of multiple modalities is necessary for robust perception of speech [2].

To leverage cross-modality in robot perception, we propose to realize the robot’s perceptual system as a network of coupled recursive estimation filters. Each estimation filter addresses a perceptual subproblem. The coupling of these components allows us to use the estimated value of one recursive estimation loop as a prior for others, even across different modalities. For example, our system predicts motion from proprioception and uses it to interpret visual perception. The system also combines proprioception and vision to perceive the type of grasp achieved by the robot hand, then using this information as a prior to disambiguate proprioceptive signals. These examples illustrate that information from multiple modalities and multiple perceptual subproblems propagates through the network, leading to robust, online perception.

We evaluate the proposed cross-modal interactive perception system on different articulated objects while varying the task and environmental conditions. Our method increases robustness and accuracy of perception by extracting more information from the multi-modal stream. We also demonstrate that the information perceived during an interaction can be used to monitor manipulation and also to generate new interactions (Fig. 1).

II. RELATED WORK

The work we present here is a) a new interactive perception system, and b) a multi-modal perceptual system leveraging cross-modal information. On our path towards...
a multi-modal system, we will also propose c) a novel perceptual approach to perceive kinematic structures based only on proprioception. We will now discuss these three areas of related work.

a) Interactive Perception: Interactive perception methods exploit correlations between actions and changes in the sensor signals for robot perception [3]. They enhance robot perception by exploiting interaction as an additional source of information for the perceptual process. This strategy has proved successful for object segmentation and recognition [4, 5], shape reconstruction [6, 7], and the perception of dynamic [8, 9] and kinematic properties [10, 11] of articulated objects. However, these methods are based on a single modality or use multiple sensor modalities but they apply one independently to each perceptual sub-task. This neglects the benefits of a tighter integration and exploitation of the correlations between sub-tasks. The multi-modal interactive perception system we present here improves robustness and versatility over previous approaches by using cross-modal communication to acquire from one modality priors for the interpretation of the other.

b) Multi-Modal Perception: Multi-modality has been applied previously in recursive filters to overcome limitations of uni-modal robotic perceptual systems. The common methodology is to estimate a correction by fusing the multimodal signal into a single estimate [12, 13]. This approach does not leverage information from one modality to help interpret the other. Differently, we exploit the results from one recursive filter as priors in the others to obtain more information. This cross-modal exploitation was applied successfully by Garcia Cifuentes et al. [14] to track a robot arm and an object from a multi-modal stream. However, their method requires models of the arm and the object and cannot be applied to perceive previously unseen articulated objects.

c) Perceiving Kinematic Models from Proprioception: Previous approaches showed that the kinematic properties of an articulated object can be perceived from end-effector trajectories [15] and applied wrenches [16] during interaction. These methods are based on two assumptions that limit their applicability: 1) there is only one moving part connected with a joint to the static environment, and 2) there is no translation between the end-effector and the moving part during the interaction. We leverage information from vision to correctly interpret proprioception and to overcome these limitations, estimating the correct grasp model and perceiving more complex kinematic structures.

III. VISION-BASED PERCEPTION OF KINEMATIC PROPERTIES

Our goal is to integrate vision and proprioception into a single multi-modal system that exploits cross-modal information. In this section, we will summarize our previously presented perceptual system based only on vision. In the next section, we present a novel perceptual system based on proprioception. Then, in Section V, we will explain how to integrate both systems leveraging cross-modal information.

![Diagram of system components]

Fig. 2: Our proposed system for interactive perception of articulated objects based on cross-modal information between coupled recursive filters; bottom: input sensor signals; arrows: information flow between filters and across modalities (blue: input measurements, red: alternative predictions)

The integrated system with its most relevant recursive filters is depicted in Fig. 2.

The interactive perception system we presented in [10] perceives kinematic models (structure and state) of unknown objects in an online manner from RGB-D streams. The key to achieve robust online perception was 1) to factorize the original problem into subproblems that can be solved robustly and efficiently using recursive estimation based on task-specific priors, and 2) to reuse priors of one recursive filter in the others by communicating estimations and predictions between the estimators.

Before we summarize the recursive filters used in that work [10], let us review the most important elements of recursive state estimation. In recursive estimation, the goal is to infer the current state of a dynamic system, $x_t$, based on the previous state, $x_{t-1}$, an observation, $z_t$, and a control input, $u_t$. When state and observation are stochastic processes, the recursive estimation filter can be implemented as a recursive Bayesian filter. In this case, the filter estimates the posterior $p(x_t|z_{1:t}, u_{1:t})$ over the state, based on the prior state $p(x_t|x_{t-1}, u_{t-1})$ [17]. Implementations require a forward process model for predicting the next state, $p(x_t|x_{t-1}, u_t)$, and a measurement model, $p(z|x)$, to be able to merge observations and predictions. These models can be defined based on domain knowledge (i.e. task-specific priors), or learned from regularities in the sensor-action space (e.g. [18] or the alternative models in [10]).

Our prior work factorizes the perception of articulated objects into three recursive estimation subproblems: 1) the estimation of the location of $N$ 3D point features ($x^{m}_{n} \in \mathbb{R}^{3n}$) from a tracking process in RGB-D images leveraging motion continuity, 2) the estimation of the motion (pose and velocity) of $M$ moving rigid bodies $(x_{rm} = (p, v), p \in$
SE(3), \( v \in \mathbb{R}^6, m \in \{0, \ldots, M - 1\} \)\(^1\) from assigned subsets of moving features leveraging rigid body physics, and 3) the estimation of motion constraints and degrees of freedom between pairs of moving bodies leveraging kinematics of articulated objects. The solution to this last subproblem is the kinematic model.

To estimate a joint of the kinematic model we maintain four joint filters, one for each perceivable joint type: prismatic, revolute, rigidly connected, and disconnected. The filters estimate the distribution of continuous random variables over the joint parameters and state, \( x_{\text{joint}} \), that is different for each type. At each time step the perceived kinematic model is composed of the most likely joint filter estimate between each pair of bodies, given the observed relative motion between bodies. See [10] for a more in depth explanation.

The way our system reuses priors from one recursive filter in the others is by passing 1) their estimated state to the filter above (see left part of Fig. 2, blue arrows) as measurements, and 2) their predicted measurements to the filter below as alternative predicted state (red arrows). In the first case, the estimated state from the filter \( i \) is used as (virtual) measurement by the filter \( j \), \( z'_j = x'_i \), using the appropriate measurement model, \( p(z'_j | v') \). We used this communication pattern to "inject" more priors at each filter until we solve the original perceptual problem.

In the second case, predictions from filter \( j - \) state predictions, \( \hat{x}'_j \) that lead to measurement predictions, \( z'_j \) are used as state predictions by the filter \( i \), \( \hat{x}'_i = z'_j \). To resolve for multiple predictions at one filter, we select the one that best explains the measurements based on maximum-likelihood model selection: \( \hat{x}_i = \arg \max_{d \in \mathbb{R}^6} p(z_i | \hat{x}'_i) \). Exploiting this communication pattern we restrict the space of possible solutions of one subproblem using the other processes (and their priors) as alternative forward and measurement models. We will exploit similar intercommunication patterns in the perceptual system based on proprioception and to exploit cross-modal information in the multi-modal system.

**Limitations:** This system fails if the interaction cannot be observed visually, e.g. due to poor lighting conditions, occlusions, or insufficient texture. Later, we will leverage information from proprioception through cross-modal integration to improve the interpretation of the visual stream and to overcome these limitations.

**IV. PROPRIOCEPTION-BASED PERCEPTION OF KINEMATIC PROPERTIES**

Proprioception refers to sensory information about the configuration of the robot's own body (kinaesthetic) and the forces it exerts (haptics). Our robot obtains proprioceptive signals from a force-torque sensor on its wrist, from air-pressure sensors monitoring the chambers of its soft hand, and from measurements of the configuration of the joints of its arm. The goal is to use these signals to perceive the motion of the object the robot is interacting with as well as its motion constraints, leading to the object's kinematic model.

The motion of the interacted body and the robot’s end-effector are coupled, as their relative motion is constrained by their contact. Because our robot uses a soft hand for the interaction, the relative motion between the hand and the object depends on the deformation of the hand and on the remaining degrees of freedom of the contact interaction (grasp).

We factorize the perception of articulated bodies into the following five subproblems: the estimation of A) the motion of the end-effector, B) the bending state of the soft-hand, C) the kinematic model of the grasp, D) the motion of the interacted body, and E) the constraints in the motion of the interacted body. Fig. 2 depicts the recursive filters addressing these subproblems, together with the filters of the vision-based system. The estimation of motion of the interacted body is subsumed with the estimation of other bodies from vision in the box “Rigid Body Motion”. In the following, we will explain how we solve the subproblems of the proprioception-based system using coupled recursion estimation (the last subproblem is solved the same way as in the previous section).

**A. Estimation of End-Effector Motion**

The first recursive filter estimates the motion of the end-effector. The state of the end-effector is represented by its pose and velocity, \( x^{ee} = (p^{ee}, v^{ee}) \). To predict the next state, we use a velocity-based kinematic update: \( \hat{p}^{ee}_t = \exp(v^{ee} \Delta t) p^{ee}_{t-1}, \hat{v}^{ee}_t = v^{ee}_{t-1} \). The measurements for the estimation of the end-effector motion are the pose and velocity of each robot joint provided by robot’s joint encoders, \( z^{ee} = (q_k, \dot{q}_k), k \in \{0, \ldots, N - 1\}, N = \text{robot's number of joints} \). We combine them with prior knowledge about the robot’s embodiment and forward kinematics and obtain a direct measurement of the end-effector’s pose and velocity that we integrate recursively: \( z^{ee} = (p^{ee}, v^{ee}) \).

**B. Estimation of Hand Bending**

When the robot interacts with an object, the soft hand deforms. This changes their relative pose (see Fig. 3). In

---

\(^1\)We changed the representation of rigid body poses to avoid the singularities of the exponential coordinates at \( 2 \pi \) rotations. Poses are now elements of the Special Euclidean group \( p \in SE(3) \) affected by a small random perturbation in the tangential Lie space \( \xi \in \mathbb{R}^5 \) so that their composition \( p = \exp(\xi) p \) is also a random variable [19]

---

![Fig. 3: Effect of the deformation of the soft-hand; left: hand in the nominal state; middle and right: hand in the bent state after a motion of the end-effector without motion of the interacted body (a door handle)](image-url)
the second recursive filter, we estimate the consequences of this bending effect. We represent the bending state of the soft hand as the relative transformation between the nominal end-effector pose (estimated by the filter described above) and the pose of a virtual body we call bent end-effector (defining the hand’s physical pose), $x^{\text{bent}} = p^{\text{ee}} \ominus p^{\text{bee}}$, where $\ominus$ is the inverse composition of elements of $SE(3)$. We assume that the bending state remains constant between consecutive time steps, $\dot{x}_{\text{bent}} = x_{\text{bent}}$. We use as measurements the signals of the proprioceptive stream that correlate to the bending of the hand. These are the wrenches measured at robot’s wrist and the pressure values in the four air chambers of the soft-hand: $z^{\text{bent}} = (w, P)$, where the wrenches $w \in \mathbb{R}^6$ and the air pressure signals $P \in \mathbb{R}^4$. However, defining an explicit measurement model relating bending and proprioceptive signals for a complex soft-manipulator as the RBO Hand 2 [20] is a difficult problem [21]. We will adopt a data-driven approach and learn from experiences a model that transforms the proprioceptive signals into direct observations of the bending state: $f(w, P) = z^{\text{bent}} = p^{\text{ee}} \ominus p^{\text{bee}}$.

We approximate $f$ using an artificial neural network. To obtain labeled data to train the model, we execute 15 interactions of the robot grasping an object that is rigidly attached to the environment. We recorded the wrenches and pressure signals at different relative poses of the bent soft hand with respect to the nominal pose during these interactions. We then trained a multi-layered perceptron regressor (MLPR1) to map from wrenches and pressure signals to the 6D relative pose observations.

To integrate the observations recursively, we also need to learn their uncertainty. Following the approach proposed in [22], we trained several partial MLPRs, leaving out groups of two trials, and computing the standard deviation between predictions from these partial MLPRs and the fully trained MLPR. We then train a second MLPR (MLPR2), mapping wrenches and pressure signals to standard deviation of the regressor. With this procedure, the second MLPR learns the difficulty of the transformation problem for each input signal and allows us to filter proprioceptive signals into a robust estimate of the hand bending state.

C. Estimation of Grasp Model

In the third recursive filter, we estimate a kinematic model of the grasp. The grasp model explains the kinematic constraints between the motion of the bent end-effector and the interacted body. We maintain and estimate independently the parameters of four filters for grasp models, one for each type of grasp that our anthropomorphic soft-hand can perform: (i) perfect grasp (no relative motion), (ii) revolute grasp (allowing rotation around the grasping axis), (iii) cylindrical grasp (allowing rotation around and translation along the grasping axis), and (iv) failed grasp (no motion constraint). For revolute and cylindrical grasps, the estimated parameters are the position and orientation of the grasping axis: $x^{gr} = x^{\text{rv}} \ominus (a_p, \hat{a}_o), a_p, \hat{a}_o \in \mathbb{R}^3$ and $|\hat{a}_o| = 1$. For the perfect grasp, the parameters are the fixed relative pose between the bent end-effector and the interacted body: $x^{gr} = p^{\text{bee}} \ominus p^{\text{ib}}$.

The failed grasp does not impose any motion constraints and therefore does not have any parameters to estimate, $x^{gr} = \emptyset$. We initialize these parameters based on the morphology of the hand.

The estimation of the grasp model leverages the coupling between filters to obtain measurements. The estimates of the pose of the bent hand (from the previous two filters) and the interacted body (from the next filter) are combined to generate a measurement, $z^{gr} = f(x^{\text{rv}}, x^{\text{bent}}, x^{ib}) = p^{\text{ee}} \ominus p^{\text{ee}} \ominus p^{\text{ib}} \ominus p^{\text{ib}}$. The estimation of the parameters and the most likely type are performed similarly to the estimation of joint parameters of a kinematic model in [10].

D. Estimation of Interacted Body Motion

The fourth recursive filter estimates the motion of the body the robot interacts with. The state of the interacted body is represented by its pose, $x^{ib} = p^{ib}$. The prediction of its next state leverages also the coupling between filters: the change in pose depends on the motion of the end-effector, corrected with the bending effect and propagated through the grasping model, $p_{\text{ib}}^t = \exp(x^{\text{ee}} \text{Ad}_{\text{pib}}} p_{\text{ib}}^{t-1}$, where $x^{\text{ee}}$ is a $6 \times 6$ matrix representation of the kinematic constraints of the grasping model and $\text{Ad}_{\text{pib}}$ is the adjoint transformation associated to the bending effect. None of the proprioceptive signals can be used as observations of the motion of the interacted body, and thus the prediction becomes the belief state.

Limitations: The first limitation of the system based only on proprioception is due to the mutual dependency between the estimation of the interacted body motion and the grasp model. This effectively reiterates the initial prior distribution over the grasp model. The accuracy of the estimated interacted body motion depends on this prior.

The second limitation is that the proprioceptive signals, because of their limited range, only provide measurements about the state of the robot and the responses from the interacted body. The system can only perceive a single body connected by a joint to the environment defining the kinematic model. Overcoming both limitations will require additional prior knowledge that our integrated system will obtain from vision by leveraging cross-modal information.

V. INTEGRATION OF VISION AND PROPRIOCEPTION

Once we have explained how to extract information from each modality, we will explain how to leverage information from one modality to help interpret the other. The proposed multi-modal system exploits cross-modal information to overcome the limitations of uni-modal perception.

Predictions about the motion of the interacted body from proprioception are leveraged to assign correctly visual point features, even under challenging visual conditions. The features can be used as the missing observations to correct the proprioceptive predictions, $z^{ib} = x^{\text{vib|ib}}$, where $x^{\text{vib|ib}}$ are the visual point features assigned to the interacted body. The cross-modal predictions from proprioception to vision and corrections from vision to proprioception lead to a
new estimate that breaks the mutual dependency of the proprioception-only system, \( x^{ib} = p^{ib} \).

Using the interacted body motion perceived from cross-modal information, our system can correctly interpret the constraints in the bent end-effector motion perceived from proprioception, and that define the kinematic grasp model, \( x^{vi} \). The type and parameters of the grasp model are inferred from the relative motion between bent end-effector and cross-modal estimates of interacted body motion (section IV-C): \( x^{vi} = p^{ib} \in p^{ib} \).

The system can use grasp model estimates from cross-modal information as prior to further interpret proprioceptive signals when the visual modality degenerates (e.g. the object goes out of the field of view, is occluded, or due to extremely bad lighting conditions or not enough visual texture). The prior obtained from cross-modal information is sufficient to estimate the kinematic model of the interacted body using only proprioceptive signals.

Finally, the integrated system interprets correctly the constraints in the motion of the interacted body perceived from proprioception leveraging information from vision. The system perceives from vision the motion of other bodies apart from the directly interacted and uses this prior to analyze the motion constraints of the interacted body from proprioception. The integrated system based on cross-modal information can perceive complex kinematic models with multiple joints or when the interacted body is not connected to the static environment, \( x^{joint} \).

VI. ROBOT MOTION GENERATION AND CONTROL

Generating a multi-modal stream rich on information depends on the strategy to control robot’s interaction. Our goal is to generate motion in the dimensions allowed by the (initially unknown) kinematic structure. This adaptive behavior can be achieved using a compliant controller based on the force-torque signals [23]. We use an operational space impedance controller on the Lie Group \( SE(3) \) [24] to adapt a desired trajectory of the 6D pose of the end-effector, \( p^{eed}_{des} (i) \), based on the signals from the force-torque sensor. The behavior of this controller is parametrized by three \( 6 \times 6 \) matrices –stiffness \((K)\), damping \((D)\) and mass \((M)\)– that transform virtually the end-effector into a spring-mass damped system with different reactive behavior for each dimension.

The aforementioned controller can adapt an initial exploratory trajectory. To generate such a trajectory for articulated objects we propose a velocity-based controller that sets at each step a new goal for the end-effector pose, \( p^{eed}_{des} \), based on the error between the measured and desired velocity twists: \( \dot{p}^{eed}_{des} = \exp(k_p(v^{eed}_{meas} - v^{eed}_{des})) \dot{p}^{eed}_{des} \). Combining both controllers the robot can explore articulated objects with different dynamic properties and create rich multi-modal signals for perception.

Once the kinematic structure has been revealed and perceived leveraging cross-modal information, the robot should be able to use this information to improve the interaction or generate new manipulations. We implemented this skill as an online trajectory generator that computes an end-effector operational space trajectory to achieve a manipulation task (i.e. reaching a desired object configuration). The trajectory generator uses the perceived kinematic model of the object and interpolates its joint configurations towards the goal configuration. Then it computes the trajectory of the interacted body, and from that, the trajectory of the end-effector that will generate the desired change in the object’s configuration.

VII. EXPERIMENTS

We conducted two sets of experiments. In the first set, we evaluated quantitatively the performance of our system when perceiving different articulated objects and compare the use of only vision, proprioception, or the multi-modal stream leveraging cross-modal information. We measure the robustness, accuracy, and convergence of the kinematic model estimation by comparing to ground truth for the joint parameters and state. In the second set, we made use of the online information from the cross-modal system to control robot’s motion and fulfill a manipulation task. The robot explores an articulated object until it discovers a joint and perceives that it reaches a desired joint configuration. Then, the robot exploits the perceived information to plan a new trajectory to return the object to its initial configuration. We measure the accuracy of the execution (final joint state) of both the explorative and the exploitative interactions.

A. Experimental Setup

In our experiments, we use a robot manipulator composed of a Barrett WAM arm and a RBO Soft Hand 2 [20]. The joint configurations of the arm are measured at 200 Hz by encoders placed at the motors controlling the cables. The stretching of the cables introduces uncertainty about the end-effector’s pose that we model with a covariance of 1 cm and 3° in the end-effector pose measurements, resulting from an offset calibration. The visual input is an RGB-D stream (640×480 pixels at 30 Hz) provided by a Carmine sensor rigidly attached and registered to the robot’s base. The force-torque signals are provided by an ATI 6-DoF sensor mounted on robot’s wrist and delivering signals at 100 Hz. Air pressure in the chambers of the soft-hand are delivered at 100 Hz. To compensate for the disparity in sensor frequencies we accumulate signals and process them at 15 Hz. This estimation rate can be maintained on an Intel Xeon E5520 PC at 2.27 GHz.

The estimated states of each filter are assumed to be Gaussian distributions. Both process and measurement models are of the form \( \delta_t = g(x_{t-1}, u_t) + \epsilon_t \) and \( z_t = h(x_t) + \delta_t \), where \( g \) and \( h \) are possibly non-linear (but linearizable) functions, and \( \epsilon \) and \( \delta \) are process and measurement additive Gaussian noise. This allows us to implement the recursive estimation filters as Kalman filters or their variant for non-linear models, extended Kalman filters.

The neural network regressors (MLPR) have a topology of three layers with 10-10-10 fully connected neurons. This topology was selected in a hyperparameter search by a leave-one-out cross-validation process, selecting between 1
and 100 neurons per layer in networks of one, two, or three layers. The vision-based system tracks \( N = 200 \) point features. To focus the attention on the estimation of the kinematic model of the articulated object and not on the robot’s arm, we project a model of the robot on the camera plane and subtract this part from the visual analysis.

In our experiments, we parametrized the controller to be compliant in all dimensions (main diagonal elements of \( K = 0.1 \), \( D = 1 \), and \( M = 1 \)) except in the pulling direction of the end-effector (main diagonal elements of \( K = 400 \), \( D = 200 \), and \( M = 1 \)). These parameters perform well in all objects we evaluated. To generate an exploratory behavior we guide the robot’s end-effector into a grasping distance of the object, command the robot to close the hand and command a desired velocity, \( v_{\text{desired}} \), of 2 cm s\(^{-1}\) in the pulling direction (\( k_p = 0.1 \)). As a result, the robot reveals the kinematic structure by pulling with increasing force and adapting to the dimension of allowed motion of the articulated object.

We evaluate our system on articulated objects with different types of joints, size, color, and surface properties. We did not add artificial visual markers that could facilitate the visual perception. The objects are placed at different pose with respect to the robot and sensors. In some experiments we also change abruptly the lighting conditions to evaluate the robustness of the perceptual systems. To obtain the ground truth for the joints, we manually measured the joint parameters and the final joint state.

### B. Results

**Uni-Modal vs Cross-Modal Perception:** We evaluate the accuracy and convergence of the kinematic model estimates from the three perceptual systems: only vision, only proprioception and cross-modal integration. Figure 4 shows three images from the RGB-D sensor (initial, middle, final steps) and graphs of the estimation error to ground truth over time.

In the first experiment the robot interacts with a drawer. After 6.5 s (indicated with a vertical line in the plot) we change abruptly the lighting conditions by switching off the lights. The vision-only system stops perceiving the object while the proprioception-only and the cross-modal system continue the estimation. The final joint state estimated by the cross-modal system is the most accurate (22.3 cm, ground truth 22.5 cm), followed by the proprioception-only (23 cm). The vision-only system stops tracking at 9.8 cm. The cross-modal system achieves the best performance because it leverages vision to estimate a more accurate grasping model, which lead to more accurate body motion estimates and robustness against vision failures from the interpretation of proprioception.

In the second experiment, the robot interacts with a door that rotates around a revolute joint. The robot almost completely occludes the object during the first 25 s of interaction (indicated with a vertical bar in the plot). The proprioception-only and the cross-modal system perceive the object during the entire interaction. The vision-only system perceives the interacted object only when it becomes clearly visible. The final joint state estimation from the cross-modal system (80°,
TABLE I: Robot Interaction Based on Online Multi-Modal Perception

<table>
<thead>
<tr>
<th></th>
<th>Exploration</th>
<th>Exploitation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sliding Door</td>
<td>2.2 cm ± 1.6 cm</td>
<td>1.8 cm ± 1.6 cm</td>
</tr>
<tr>
<td>Camera Tripod</td>
<td>7.8 ± 2.3 cm</td>
<td>2.6 ± 2.24 cm</td>
</tr>
<tr>
<td>Glass Door</td>
<td>1.3 ± 0.13 cm</td>
<td>0.6 ± 0.5 cm</td>
</tr>
</tbody>
</table>

ground truth 85°) is the most accurate, followed by the proprioception-only (78°). The final estimation of the visual system (43°) is affected by the delayed start. The cross-modal system achieves the best performance because it uses the proprioceptive signals to interpret the visible motion in the small non-occluded parts of the object.

In the third experiment, the robot interacts with a cardboard box and closes one of its lids. As a result from the explorative interaction, the entire box translates. We focus on the analysis of the estimation of the relative revolute joint between the box and the lid. Both uni-modal systems fail to detect this joint. The vision system only perceives the lower part of the box, while the proprioception system detects only the motion of the lid and interprets it as a revolute joint with respect to the environment. The cross-modal system correctly perceives the relative joint between the box and the lid because it uses the motion of the box perceived from vision to correctly interpret the motion constraints of the lid perceived from proprioception. The final joint state estimate from multi-modality is 90° (ground truth 100°).

d) Controlling Interaction with Online Interactive Perception: We tested our cross-modal perceptual system and online trajectory generator for the manipulation of three previously unseen objects (see objects in Fig.1): opening a glass door (GD) 20°, turning camera tripod (CT) 45° and opening a sliding door (SD) 30 cm. These objects are challenging because they do not present strong textured surfaces and because the hand cannot grasp them perfectly. We repeated the interactions 5 times on each object with different initial robot-object pose. The results (mean and standard deviation on the error to ground truth) are depicted in Table I. The interaction succeeded in the 15 trials (see video attachment) indicating that the information perceived online can be used to generate new successful trajectories. Our proposed perceptual system leverages information between vision and proprioception to estimate accurately the joint state at the turning point and the end of the manipulation, which indicates that the system can be applied to monitor ongoing interactions.

VIII. Conclusion

We presented a perceptual system for online multi-modal interactive perception of articulated objects based on coupled recursive estimation. The key novelty of our system is the concept of cross-modality, when information extracted from one sensor modality is leveraged to extract information from the sensor stream of another modality. This approach to multi-modal perception is present in human perception [1] and as we show also benefits robot perception. We demonstrate this in the context of perceiving articulated objects cross-modally, using information from vision (RGB-D sensor) and proprioception (from joint encoders, a force-torque sensor, and air-pressure sensors). Our cross-modal approach increases robustness and versatility of interactive perception under varying environmental conditions and overcomes limitations of uni-modal systems. We also show that our system can be used to monitor ongoing interactions and to generate new manipulation actions.

REFERENCES