A Computational Framework for Visual Perception of Inertial Affordances

Walter A. Talbott
Cognitive Science
UC San Diego
Email: wtalbott@ucsd.edu

Javier Movellan
Machine Perception Lab
UC San Diego

***Abstract*—We present a Model Predictive Control approach to visual affordances. Under this framework, visual features are used to estimate the inertial parameters of internal models. These models are used to generate and update motor control policies. Finally, the accelerations observed while applying the control policies are used to refine the mapping between visual features and inertial model parameters. The proposed approach allows to generalize visual affordances to novel objects, and novel tasks. Computer simulations show that the proposed approach models results from existing behavioral experiments, suggests and makes predictions for new experiments, and is amenable for implementation in humanoid robots.**

I. Introduction

Consider the two objects in Figure 1. Most people have not encountered these objects before, yet they have strong intuitions about which object is better for mashing potatoes and about how to grasp it for doing so. This is an example of a visual affordance, our intuition of how to interact with a novel object based on visual inspection. In a classic experiment, Mounoud and Bower [9] showed that by 9 months of age infants estimated inertial properties of objects based on visual information. Rods of different sizes were given to infants at arm's length, so that when the experimenter released the object the infant's arm either dropped, raised, or remained stable. Infants soon learned to predict the weight of the rod from its length, as shown by the fact that the initial arm-drop decreased as new rods were presented. After this learning had occurred, infants were given a hollow decoy rod that broke the learned relationship between length and weight. The infants' arms lifted up dramatically. This is an indication that the infants had learned a visual affordance: prior experience taught them to modulated the forces to be applied to a new object in response to the visual perception of that object.

In this paper, we explore a computational approach for how visual affordances may be developed and applied in the context of motor planning and motor control. We focus on perception of the inertial properties of objects (e.g., weight, center of mass, moment of inertia). We propose an approach that can model the phenomena from Mounoud and Bower [9], suggest and predict the results of novel experiments, and be implemented in physical robots.

The approach we propose has three main components:

1) A visual system that predicts inertial properties of objects based on their visual features.
2) A proprioceptive system that estimates inertial properties of objects from observed joint torques and resulting accelerations. The proprioceptive system is used to update the predictions of the visual system.
3) A model predictive controller that can both run internal simulations of object dynamics for anticipatory choices, and control behavior in the physical world.

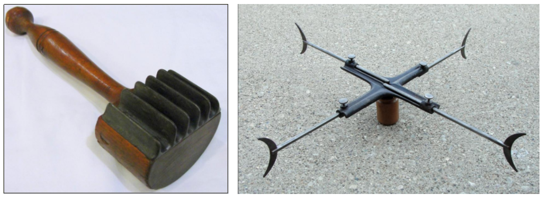

Fig. 1. Two uncommon objects.

A. Prior Work

Fitzpatrick et al. [4] presented one of the pioneering approaches to robotic affordance learning. In their approach, a robot poked objects and observed whether or not they rolled. This allowed the robot to learn to predict whether new objects will roll based on visual information alone. Other approaches focused on learning to choose from a discrete set of actions based on the results of these actions on visually similar objects [14] [13] [7] [8]. For example, in [14], a robot performs predefined movements with color-coded tools to move a puck on a flat surface. The results of each movement are recorded, and the previous results are used to select behaviors to move the puck to desired locations. One limitation of these approaches is that they do not generalize past the actions in their predefined sets.

Hermans et al. [5] argues for an approach similar to ours: rather than classifying a discrete set of affordances directly from visual features, the approach first makes inferences about intermediate object properties, such as material or size. Then, the presence or absence of a set of affordances (pushable, liftable, *etc.*) is classified using the intermediate features. Our approach differs from this previous proposal in that, instead of predicting specific affordance labels, we learn to map visual features into inertial properties of objects. These inertial properties can then be used within a Model Predictive Control

(MPC) approach to infer the response of the object to a wide range of tasks and situations.

There is evidence that humans use internal models of their own bodies and of external objects in order to formulate motor control plans [12]. There is also evidence that the cerebellum plays an important role in the formation adaptation and real time use of these internal models. For example Imamizu et al. [6] had subjects in an fMRI scanner learn to track a target with the cursor of a computer mouse with a rotated coordinate frame. The cerebellar activations showed one pattern that was proportional to the error between the mouse cursor and a tracking target, and another pattern that showed increased activation even after the error decreased to baseline levels. They propose that this activation is evidence of an internal model being learned and remaining active when the task is performed.

In the next sections, we formalize the problem we are trying to solve, specify the proposed approach, and run computational experiments to explore how the approach behaves in different conditions of interest.

II. Problem Formalization

The goal of our approach is to explain phenomena like the one illustrated in Figure 1, in which people can make a wide range of predictions for how to choose and use novel objects for a wide range of tasks. We formalize this problem from the point of view of MPC: we aim to develop closed loop control policies for articulated bodies (robots) that use rigid objects (e.g., tools) to achieve goals. These control policies are developed by using internal models that predict the future consequences of actions. Our goal is for the visual appearance of the objects to modulate the resulting control policies in an intelligent manner, *i.e.*, a manner that is optimal with respect to a well defined function.

When the robot is not attached to an external object, its dynamics adhere to the standard equations of motion for articulated bodies

$$M(\theta_t)\ddot{\theta}_t = \tau_t + N(\theta_t, \dot{\theta}_t) \tag{1}$$

where θ_t is the vector of joint angles at time t, $\dot{\theta}_t$ the angular velocities, $\ddot{\theta}_t$ the angular accelerations, $M(\theta_t,)$ is the moment of inertia matrix of the entire articulated body, τ_t the vector torques applied by the rotational joint actuators, and $N(\theta_t, \dot{\theta}_t)$ is the vector of gravitational, friction and Coriolis/Centripetal forces. We assume the motion dynamics of the robot are known, so they can be used in a model predictive control framework to develop control policies for tasks such as trajectory tracking.

When the robot grasps an object, it alters the robot itself. An object attached to the robot by a grasp changes the robot's geometry, inertial properties and equation of motion. Here, we use the parameter λ to represent the dependency of the robot dynamics on the inertial properties of the grasped object

$$M(\theta_t, \lambda)\ddot{\theta}_t = \tau_t + N(\theta_t, \dot{\theta}_t, \lambda) \tag{2}$$

Formally, a control policy c is a mapping between robot states (angles, and angular velocities) and actions (torques applied to each joint), i.e.,

$$\tau_t = c(\theta_t, \dot{\theta}_t) \tag{3}$$

Goals are formulated in terms of a scalar function that captures the expected reward resulting from using a control policy over a finite period of time $[0, T]$

$$\rho(c) = \int_0^T E[R_t|c]dt \tag{4}$$

where R_t is the reward rate and T is the terminal time. The reward rate R_t is simply a function that expresses the desirability of achieving a robot's state at a particular point in time t. For example, R_t could be the Euclidean distance, at time t, between a target location and the tip of a robotic finger.

The problem is to find a control policy $\hat{c}$ that minimizes $\rho(\hat{c})$ subject to the dynamics in (2), where the inertial properties λ of the object are not completely known. Moreover we want for the control policy to be modulated by the visual appearance of the object, in a principled manner.

III. Proposed Approach

We assume that while the inertial properties of the original robot are known, the inertial properties of the object are not fully known. We formalize the lack of knowledge of the object's inertial properties by using a probability distribution, $p(\lambda_t \mid v_t, s_t)$, over inertial parameters of the object (weight, center of mass, moment of inertia). Here λ_t represents the inertial properties of the object observed at time t. v_t represents the visual information observed while manipulating objects up to time t, and s_t is the proprioceptive information (joint torques and angular accelerations) obtained while manipulating objects up to time t. This probability distribution is the mathematical expression for the visual perception and learning of inertial affordances.

Suppose at time t the robot is presented a new object. Prior experience observing and manipulating objects (v_t, s_t) allows the robot to predict probable inertial parameters for the current object: $p(\lambda_t \mid v_t, s_t)$. Based on this prediction, the robot formulates a control policy to achieve a goal using the object. The control policy is implemented, and as a result of it the robot grasps the object and manipulates it. During the manipulation, the robot keeps track of the joint torques and resulting joint accelerations. This new data is used to update the probability distribution of inertial parameters given visual features. Figure 2 shows the different components of the proposed approach: (1) A visual system generates a probability distribution of object inertial properties based on their visual properties (*e.g.* color, texture, shape). Given a new image, the system generates a probability distribution over object geometry and density. This is then used to compute a probability distribution over probable inertial object properties (weight, center of mass, moment of inertia) given their observable visual features. (2) A proprioceptive system that infers the

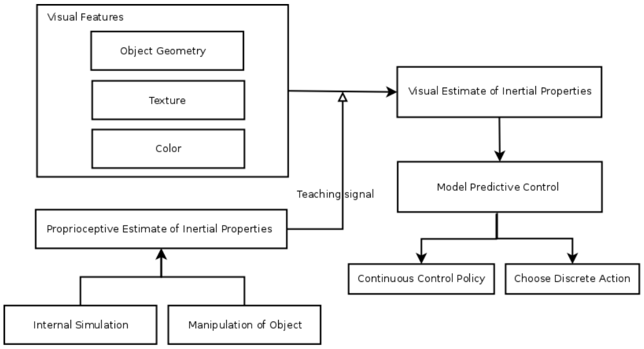

Fig. 2. Diagram of the proposed approach.

posterior distribution over the inertial properties of objects given: a prior distribution over inertial object parameters, observed torques applied to the robot joints, and the resulting angular accelerations observed on the joints. This is used as a teaching signal to update the probability distribution generated by the visual system. (3) A control system that generates control policies to achieve goals given the current probability distribution over object inertial parameters. The control system uses an MPC approach, i.e., it utilizes internal models with the known inertial parameters of the robot, plus the available distribution of inertial object parameters to formulate control policies. These control policies are then applied in the physical world. The results are used to update the internal model.

Here we explore a Minimum Angular Acceleration approach to MPC of articulated bodies [11]. This approach has the advantage that it generates human-like trajectories, adheres to a well-defined optimality framework, is non-iterative, and is computationally efficient. In order to handle the uncertainty over the inertial parameters of grasped objects, we sample from the available distribution of probable parameter values, and compute the optimal control policy given the sampled parameter. The control policy is applied to the physical world and the results are used by the proprioceptive system to improve the probability distribution of inertial parameters from the visual component.

IV. Computer Simulations

We run 4 computer simulations of the proposed approach. The goal of the simulations was to gain insights about how the approach behaves before we implement it in a physical robot. The first simulation focuses on the Mounoud and Bower (1974) study previously described [9]. The second simulation describes an additional experiment suggested by the approach and specifies the predictions made by the model. The third and fourth simulations show that the affordances can generalize to novel objects and tasks.

For the simulations, we used a 7 degree-of-freedom model of the human arm. The first joint (shoulder) had 3 degrees of freedom, the second joint (elbow) had 2 degrees of freedom and the third joint (wrist) had 2 degrees of freedom. The links were simulated as ellipsoids with the density of ice. Gravitational forces used the Earth surface standard. The simulator was implemented in Matlab using the Gaussian mechanics approach to articulated bodies [10]. The equations of motion were integrated using a 4th order implicit Runge-Kutta method. Our implementation was validated using the Matlab Robotics Toolbox [1].

A. Simulation I: Modeling Mounoud and Bower's 1974 study

Mounoud and Bower [9] showed that infants older than 9 months use visual information to adapt their motor behavior to compensate for the weight of novel objects. Here we simulated their experiment using the following procedure: On each trial an ellipsoidal object of a given length and material was presented to the robot arm. Each material had a distribution of possible densities, and was indicated by color. The robot reached out horizontally and was then given the object (the object was attached to the hand). The desired behavior was to hold the object fixed at the same height where it was given. The first 4 objects were all of same density, but varied in length. The last object, which we call the decoy, had a much lower density. The first four objects were presented in order from shortest to longest. We used the color, size, and axis lengths of the ellipsoids as visual features. There are well known approaches from the computer vision literature to estimate geometric properties of 3D ellipsoidal objects from 2D projections, *e.g.* [2]. Here, we assume that these or other algorithms were used to estimate these 3D geometric properties.

On each trial, the proprioceptive system uses the observed torques and accelerations measured while holding the object to estimate the inertial properties of the object. This estimate is used to update the probability distribution of object density given its observed color. On the next trial, a new object is presented. The most probable density given the observed visual features was used to estimate the object's inertial parameters. This estimate was used by the controller to generate a policy to keep this new object as level as possible. Figure 3 shows the magnitude of the drop in arm height for each of the first 4 objects. As observed in the Mounoud and Bower experiment, we found a decrease in the magnitude of arm drop between trials. This is due to the fact that on each trial, the model improves its estimates of the object's weights based on the observed visual properties (see Figure 3).

Figure 4 shows the heights of the object through time as the arm holds each object. The trajectory of the decoy object is displayed as a magenta line. As observed in the Mounoud and Bower [9] experiment, the arm lifts the decoy object much higher than the target level. This result confirms that the computational approach outlined here can reproduce behavior that is analogous to human behavior.

B. Simulation II: Center of Mass

The Monoud and Bower experiment was designed to test visual perception of one inertial property of objects (their weights). Here, we explore a generalization of their original experiment designed to test visual perception of a different

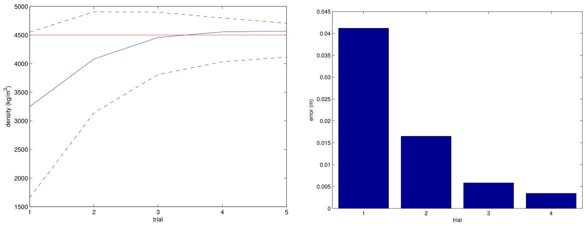

Fig. 3. (Left) Estimate of material density across trials. Dashed lines show 25th and 75th percentile of estimate distribution. Horizontal line shows true density. (Right) Magnitude of arm drop for each trial. The drop is reduced from trial to trial, even though the mass of the objects increases.

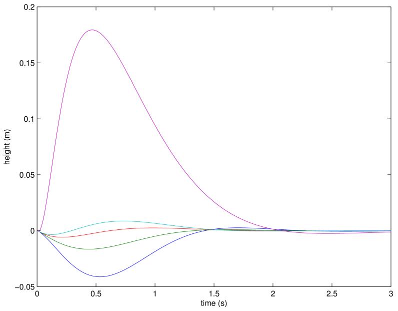

Fig. 4. The vertical position of the objects over 3 seconds. The first object (blue line) presented dips lower than the goal of 0 meters because the density estimate is lower than the actual density. As the affordance is learned, the trajectory becomes flatter (green, red, cyan lines). The decoy object (magenta line), which is much less dense than its appearance suggests, is mistakenly lifted above the desired height.

inertial property (center of mass). We show the proposed experiment and run computer simulations to specify the predictions made by our model. For this experiment, we used 3 different objects, each of which is made of 3 aligned ellipsoidal components (see Figure 5). The 3 objects weigh the same, however they have different centers of mass. The first object has two components made of wood (low density) near the hand, and one component made of steel (high density) farthest from the hand. The second object is the reverse, i.e., the steel is near the hand, and the wood is farther from the hand. The third object, the decoy in this experiment, has the appearance of the first object, but the density distribution of the second. Because its density and appearance are inconsistent with the learned affordance, we expect behavior similar to the decoy response from the first experiment.

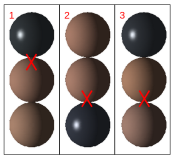

Fig. 5. The three objects used for Simulation II. Object center of mass is indicated by an X. Object 3 is the decoy object: it has the appearance of object 1, but the inertial properties of object 2. All objects have the same mass.

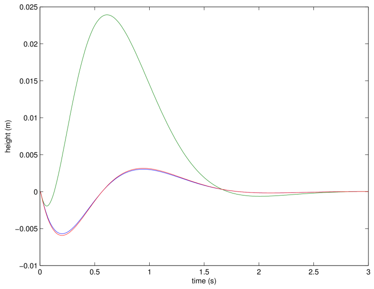

Fig. 6. Height above the desired holding position for each of three objects. The red and blue lines correspond to Objects 1 and 2, for which the visual affordances are accurate. The green line shows the response to the decoy object (Object 3).

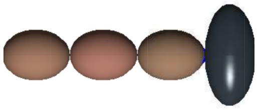

Fig. 7. The novel object used for simulations three and four. The object is made of the same materials as the object from simulation two, but in a different shape.

As in Simulation I, we present each object to the robot when the arm is horizontal, and observe the drop or lift relative to the initial height. The density estimate of each material is updated between each trial.

Figure 6 shows how the robot arm behaves in response to the three objects. The results are similar to Simulation I, where an exaggerated lift of the decoy object is observed.

The exaggerated response to the decoy is observed even though the objects are all the same weight. We also see that the moment of inertia matrix can be estimated from the visual features of the object, and used to plan behavior. The inertial properties of objects can give the robot important information about how an object will behave in particular situations. The final two experiments explore how this information can help the robot generalize its expertise to novel tasks with the same object.

C. Simulation III: Choosing a Grip for Reaching with a Tool

In this case, the task is to move an object so that the end point of the object tracks a specified trajectory in space.

According to our model, the robot internally compares the energy required to move a novel object, shown in Figure 7, through a trajectory for different grasp configurations. This comparison is made by running MPC using the estimated inertial model of the novel object. Two grasps are compared: one where the steel component is closest to the hand, and one where the wooden component is. Once the energy has been estimated for each configuration, the robot chooses whichever grasp requires the lowest energy.

For this problem the optimal solution is to grasp the object by the steel component. We examine the effect of experience with the objects from Simulation II on the proportion of grasps

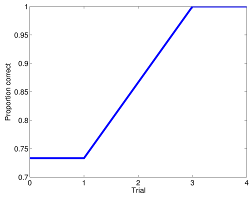

Fig. 8. The proportion of correct grasp choices for how to move a novel test object efficiently as a function of amount of experience with different objects.

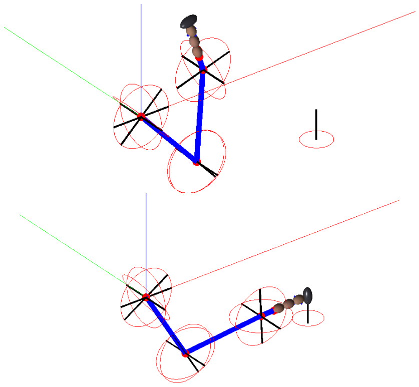

Fig. 9. Illustration of the hammering behavior. The top panel shows the start point. The bottom panel shows the hammer striking the simulated nail.

chosen correctly for the novel object. A trial consists of the following steps:

1) The robot is given an object from Simulation II and must hold it as level as possible.
2) The robot's belief about material density is updated.
3) The robot is asked to choose the best grasp for moving the novel object through a specified trajectory. The robot is not allowed to actually grasp the object and track the trajectory.

After step 2, the robot has an updated distribution of each material's density. The robot choice is made using the MPC with parameters sampled from the probability distribution of material density. The experiment is repeated multiple times so we can obtain the proportion of correct choices averaged across experiments. Figure 8 shows that the proportion of correct choices increases as the robot gains experience with similar objects. Note that the robot is never allowed to interact with the test object or to actually perform the tracking task. The amount of experience, refers only to how often the robot has interacted with different objects made of the same material.

D. Simulation IV: Choosing a Grip for Hitting with a Tool

In this simulation, the goal is to use a novel object to efficiently hit a nail with sufficient force. The MPC component uses the estimated object model to determine the required energy for generating a desired force at the object's end point.

Similar to Simulation III, the robot makes choices between grasps based on running its estimate of the object's inertial properties through its internal model. The two grasp positions are the same as before, one grasping the steel end and the other grasping the wood end.

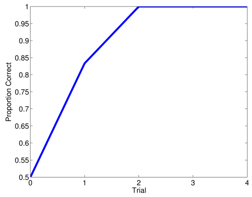

Fig. 10. The proportion of correct grasp choices for how to hammer with a novel test object efficiently as a function of amount of experience with different objects.

A trial is defined as in Simulation III:

1) The robot is given an object from Simulation II and must hold it as level as possible.
2) The robot's belief about material density is updated.
3) The robot chooses best grasp for hammering with the novel object. The robot is not allowed to perform the task.

To make a choice between the grasps, the MPC component generates simulated hammering trajectories (Figure 9), and chooses the grasp that requires the least energy for a given contact force. For this task the optimal choice is to grasp the wooden side of the object.

Figure 10 shows the probability of a correct choice as a function of trial number. At the start of the experiment the robot is at chance. After the first trial, the robot chooses the correct grasp 65 % of the times. After the second trial it chooses the correct grasp all the time.

V. Conclusion

We proposed a computational framework for visual perception of inertial affordances. The approach combines three modules: (1) a visual system that predicts inertial properties from visual information. (2) a proprioceptive system that uses observed forces and accelerations to teach the visual system. (3) A Model Predictive Controller that uses internal models to generate control policies.

The proposed approach is designed to reproduce humans' ability to use visual information to plan how to use novel objects in novel ways. Our approach is based on the use of internal models of our own bodies and of external tools. Model-based approaches have become popular for solving complex tasks in humanoid robots [3]. In addition, there is mounting evidence that internal models are part of the machinery for motor control used by the brain [12].

Our simulations show that the proposed approach can replicate behaviors observed in human experiments, and make predictions for new experiments. The approach can also be implemented in robots to use visually perceived affordances in a manner similar to the way humans do.

Acknowledgments

This research was supported by NSF IIS 0968573 and NSF IIS 0808767.

References

[1] P.I. Corke. A robotics toolbox for matlab. *IEEE Robotics and Automation Magazine*, 3(1):24–32, 1996.

[2] G. Cross and A. Zisserman. Quadric reconstruction from dual-space geometry. In *Computer Vision, 1998. Sixth International Conference on*, pages 25–31, Jan 1998. doi: 10.1109/ICCV.1998.710697.

[3] Tom Erez, Kendall Lowrey, Yuval Tassa, Vikash Kumar, Svetoslav Kolev, and Emanuel Todorov. An integrated system for real-time model predictive control of humanoid robots. *International Conference on Humanoid Robots*, 2013.

[4] P. Fitzpatrick, G. Metta, L. Natale, S. Rao, and G. Sandini. Learning about objects through action - initial steps towards artificial cognition. In *Robotics and Automation, 2003. Proceedings. ICRA '03. IEEE International Conference on*, volume 3, pages 3140 – 3145 vol.3, sept. 2003.

[5] Tucker Hermans, James M Rehg, and Aaron Bobick. Affordance prediction via learned object attributes. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA): Workshop on Semantic Perception, Mapping, and Exploration*. Citeseer, 2011.

[6] Hiroshi Imamizu, Satoru Miyauchi, Tomoe Tamada, Yuka Sasaki, Ryousuke Takino, Benno PuÈtz, Toshinori Yoshioka, and Mitsuo Kawato. Human cerebellar activity reflecting an acquired internal model of a new tool. *Nature*, 403(6766):192–195, 2000.

[7] R. Jain and T. Inamura. Learning of tool affordances for autonomous tool manipulation. In *System Integration (SII), 2011 IEEE/SICE International Symposium on*, pages 814 –819, dec. 2011.

[8] L. Montesano, M. Lopes, A. Bernardino, and J. Santos-Victor. Learning object affordances: From sensory–motor coordination to imitation. *IEEE Journal of Robotics*, 24 (1):15–26, 2008.

[9] Pierre Mounoud and T.G.R. Bower. Conservation of weight in infants. *Cognition*, 3(1):29 – 40, 1974. URL http://www.sciencedirect.com/science/article/pii/0010027774900213.

[10] J. R. Movellan. *Physics for Roboticds and Animation: A Gaussian Approach*. MPLab, UCSD, 2011.

[11] J.R. Movellan. Minimum angular acceleration control of articulated body dynamics. In *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*, pages 5006–5011, Oct 2012.

[12] Giovanni Pezzulo, Matteo Candidi, Haris Dindo, and Laura Barca. Action simulation in the human brain: Twelve questions. *New Ideas in Psychology*, (0):–, 2013. URL http://www.sciencedirect.com/science/article/pii/S0732118X13000263.

[13] J. Sinapov and A. Stoytchev. Learning and generalization of behavior-grounded tool affordances. In *Development and Learning, 2007. ICDL 2007. IEEE 6th International Conference on*, pages 19 –24, july 2007.

[14] Alexander Stoytchev. Learning the affordances of tools using a behavior-grounded approach. In Erich Rome, Joachim Hertzberg, and Georg Dorffner, editors, *Towards Affordance-Based Robot Control*, volume 4760 of *Lecture Notes in Computer Science*, pages 140–158. Springer Berlin / Heidelberg, 2008.