Capturing and Reproducing Hand-Object Interactions Through Vision-Based Force Sensing

Tu-Hoa Pham, Abderrahmane Kheddar, Ammar Qammaz, Antonis Argyros

Capturing and Reproducing Hand-Object Interactions Through Vision-Based Force Sensing

Tu-Hoa Pham[1,2], Abderrahmane Kheddar[1,2], Ammar Qammaz[3], Antonis A. Argyros[3,4]
[1]CNRS-AIST Joint Robotics Laboratory. [2]CNRS-UM LIRMM. [3]Institute of Computer Science, FORTH. [4]Computer Science Department, University of Crete.

Capturing and reproducing hand-objects interactions would open considerable possibilities in computer vision, human-computer interfaces, robotics, animation and rehabilitation. Recently, we witnessed impressive vision-based hand tracking solutions that can potentially be used for such purposes. Yet, a challenging question is: to what extent can vision also capture haptic interactions? These induce motions and constraints that are key for learning and understanding tasks, such as dexterous grasping, manipulation and assembly, as well as enabling their reproduction from either virtual characters or physical embodiments. Contact forces are traditionally measured by means of haptic technologies such as force transducers, whose major drawback lies in their intrusiveness, with respect to the manipulated objects (impacting their physical properties) and the operator's hands (obstructing the human haptic senses). Others include their extensive need for calibration, time-varying accuracy and cost. In this paper, we present the force sensing from vision framework [6] to capture haptic interaction by means of a cheap and simple set-up (e.g., a single RGB-D camera). We then illustrate its use as an implicit force model improving the reproduction of hand-object manipulation scenarios even in poor performance visual tracking conditions.

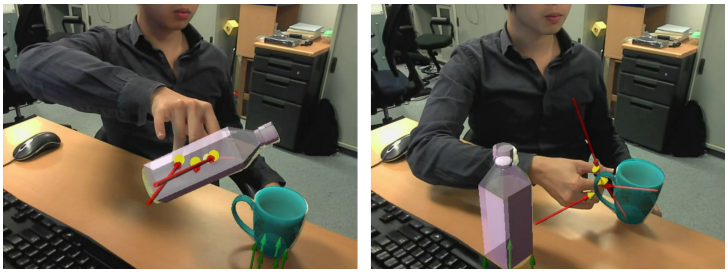


Figure 1: Using a single RGB-D camera, it is possible to infer with high accuracy the forces that occur in real hand-object manipulation tasks based on markerless tracking, kinematics, and considerations on human grasping.

Towards force sensing from vision

Previous work on correlating fingernail coloration changes to the touch force applied at fingertips [4] and estimating whole body contact forces and internal joint torques from contact dynamics and human kinematics [1] showed that vision could help infer such information up to a certain extent. Conversely, manipulation forces are the basis for physics-based hand tracking methods [3, 7]. However, these simulated forces are constructed to be compatible with 'visual' observations rather than matching the actual forces humans apply. Recent work on force sensing from vision (FSV) demonstrated that it is possible to estimate the interaction forces occurring in hand-object manipulation scenarios using a single RGB-D camera [6], provided that human hand geometry and object properties (shape, contact friction μ, mass m and inertia $\mathbf{J_q}$) are known. First, both the hand's and the object's motions are monitored using model-based 3D tracking. Numerical differentiation of the object's positional data over time then yields its kinematics, i.e. translational/rotational velocity $(\mathbf{v}, \omega)$ and acceleration $(\mathbf{a}, \alpha)$. With $\mathcal{F}_\text{d}$ and τ_d known non-contact force and torque (e.g., gravity), it can be inferred that the following net contact force $\mathcal{F}_\text{c}$ and torque τ_c are exerted by the hand:

$$\begin{cases} \mathcal{F}_\text{c} = m\mathbf{a} - \mathcal{F}_\text{d} \\ \tau_\text{c} = \mathbf{J_q}\cdot\alpha + \omega\times(\mathbf{J_q}\cdot\omega) - \tau_\text{d}. \end{cases} \tag{1}$$

Once hand-object-(environment) contact points have been estimated from tracking by proximity detection, manipulation forces can be computed as solutions of a second-order cone program (SOCP) enforcing Eq. (1) as equality constraints, as well as Coulomb's friction model as inequality constraints. Minimizing the force distribution's L^2 norm as cost function yields nominal forces that are physically valid and explain the observed kinematics. However, humans typically manipulate objects using internal forces that secure the object with a firmer grip than what is required from the Newton-Euler dynamics. This statical indeterminacy is addressed by decomposing each finger force $\mathbf{F}_k$ into a nominal $\mathbf{F}_k^{(n)}$ and an internal $\mathbf{F}_k^{(i)}$ components:

$$\mathbf{F}_k = \mathbf{F}_k^{(n)} + \mathbf{F}_k^{(i)} \quad \text{with} \quad \begin{cases} \mathbf{F}_k^{(n)} = f_k^{(n)}\mathbf{n}_k + g_k^{(n)}\mathbf{t}_k^x + h_k^{(n)}\mathbf{t}_k^y \\ \mathbf{F}_k^{(i)} = f_k^{(i)}\mathbf{n}_k + g_k^{(i)}\mathbf{t}_k^x + h_k^{(i)}\mathbf{t}_k^y, \end{cases} \tag{2}$$

with $(\mathbf{t}_k^x, \mathbf{t}_k^y, \mathbf{n}_k)$ being a local frame at finger k. Nominal forces are responsible for the object's motion through the Newton-Euler equations and internal forces are neutral regarding its state of equilibrium:

$$\begin{cases} \sum_{k\in\mathcal{F}} \mathbf{F}_k^{(n)} = \mathcal{F}_\text{c}, & \sum_{k\in\mathcal{F}} \overrightarrow{\mathbf{CP}_k}\times\mathbf{F}_k^{(n)} = \tau_\text{c} \\ \sum_{k\in\mathcal{F}} \mathbf{F}_k^{(i)} = \mathbf{0}, & \sum_{k\in\mathcal{F}} \overrightarrow{\mathbf{CP}_k}\times\mathbf{F}_k^{(i)} = \mathbf{0}. \end{cases} \tag{3}$$

A dataset of reference nominal-internal decompositions was collected by monitoring 160 manipulation experiments performed by different subjects on various contact and mass distribution configurations, using 1D tactile sensors providing ground-truth force measurements. Nominal-internal force decompositions occurring in real manipulation tasks were extracted by incorporating Eq. (3) into an SOCP that computes the internal normal forces $f_k^{(i)}$ that best bridge the gap between the nominal normal forces $f_k^{(n)}$ and tactile sensor measurements $(\widetilde{f_k})$ without perturbing the object's observed kinematics, using a new objective function:

$$\mathcal{C}_\text{decomp}(\mathbf{x}) = \sum_{k\in\mathcal{F}} \left[\left\| \mathbf{F}_k^{(n)} \right\|_2^2 + \left(f_k^{(n)} + f_k^{(i)} - \widetilde{f_k} \right)^2 \right], \tag{4}$$

with $\mathbf{x}$ denoting the vector of the unknown contact force components.

The FSV framework was completed by training multilayer perceptrons to correlate the amount and distribution of internal forces to only grasp and kinematics features, allowing to predict and reconstruct full manipulation forces on new experiments from vision alone. Experimental results obtained on datasets annotated with ground-truth data from tactile sensors and an inertial measurement unit showed the potential of the proposed method to infer hand-object contact forces that are both physically realistic and in agreement with the actual forces exerted by humans during grasping (see Fig. 2).

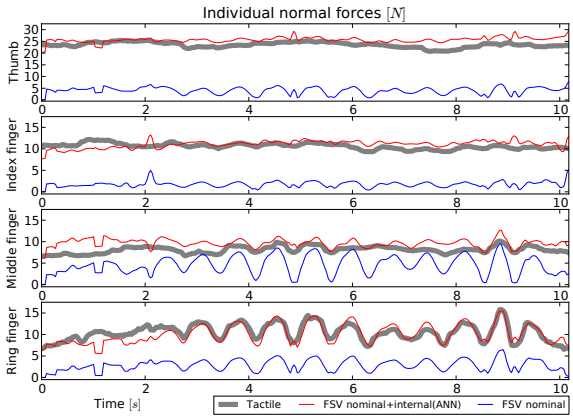
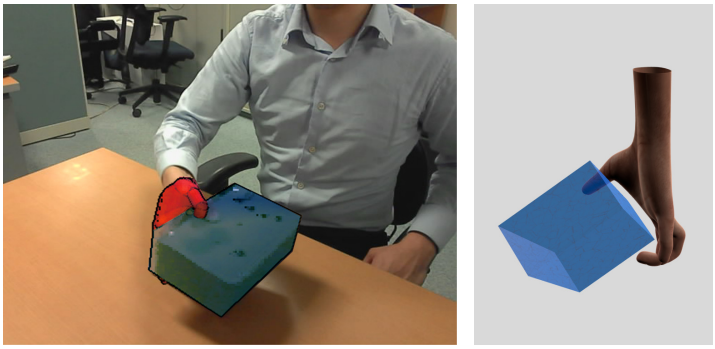
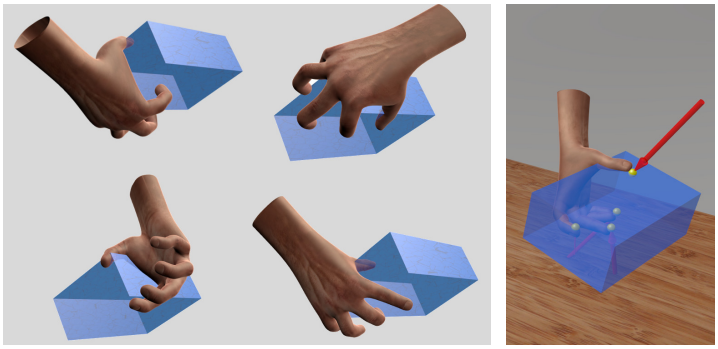


Figure 2: Artificial neural networks used in conjunction with cone programming successfully predict force distributions that explain the observed motion in agreement with human force patterns.



(a) (b)

Figure 3: The markerless tracking of a hand interacting with an object with a single RGB-D camera (a) may result in a hand-object pose that is physically inconsistent (b) due to occlusions or other visual tracking imperfections.



(a) (b)

Figure 4: We compute alternative hand poses in the vicinity of the erroneous initial hypothesis using existing grasp taxonomies (a) and look for the configuration that best explains the object's kinodynamics (b).

Reconstructing natural grasps from incomplete observations

Incomplete observation is a major challenge for motion tracking. The case of object manipulation tasks in particular is inherently subject not only to self-occlusions of each of the involved entities, but also mutual occlusions between the object and the hand. For this reason, hand tracking approaches that do not enforce physical consistency are often likely to produce unrealistic grasps, that match the observable features while being physically incorrect, especially in the case of a single camera (see Fig. 3). While multi-camera setups may partially alleviate occlusion issues, those are impractical, and are not immune to severe occlusion cases. Incorporating physics has resulted in successful approaches to improve tracking realism, both in multiple [7] and single [3] camera configurations. Such methods consider the tracked bodies' motions jointly with their underlying causes, e.g., a given object motion is only plausible if the hand pose is such that there are contact points that are able to produce a compatible force distribution. While physically valid, these simulated forces are generally not constructed to resemble the actual forces exerted by humans during grasping.

In this work, we show that the FSV framework can be extended beyond the direct estimation of contact forces and, conversely, be used as an implicit force model for physics-based tracking, to reconstruct hand poses that are both physically plausible and in agreement with the way humans naturally grasp objects. Our methodology is as follows. First, while our approach relies on the good tracking of the object due to its unarticulated, lower-dimensionality nature, we consider the hand tracking data as an initial, possibly erroneous guess rather than an absolute reference. We then initialize our search space by constructing a set of possible contacting hand poses using existing grasp taxonomies [2] in the vicinity of the initial hypothesis. At this stage, we have a set of hand poses that represent a wide variety of ways a human operator could grasp the object. The goal is now to reconstruct the actual hand pose the human operator uses.

To do so, we postulate that humans spontaneously manipulate objects using grasps that are optimal in some sense with regard to the task they try to execute. For instance, previous studies showed that humans notably regulate their grip strength to prevent muscle fatigue [5]. Given the observed object's kinematics and using the FSV framework, it is possible to predict the actual force distributions that would be incurred by each of the hand pose candidates. We then search for the actual hand pose by exploring the search space with Particle Swarm Optimization (PSO) and generate a final grasp that minimizes the force distribution's intensity, which we take as a measure of the perceived effort for the human operator. To each hand pose candidate $\mathbf{H}$, we associate the following cost based on the resulting nominal and internal force decomposition:

$$\mathcal{C}_{intensity}(\mathbf{H}) = \left\| \left(\mathbf{F}_k^{(n)} + \mathbf{F}_k^{(i)} \right)_{k \in \mathcal{F}} \right\|_2 \quad (5)$$

In order to assess the validity of our approach, we consider a standard unconstrained manipulation experiment involving rapid motions. Except for the initial frames, the hand quickly becomes at least partially unobservable, either because of self occlusions between the palm and the fingers, or external occlusions from the object. While the hand tracking is unreliable, when in contrast the object tracking is fairly accurate, we can rely on it to compute the object's kinematics over time. Generating initial hand pose candidates by combining grasp taxonomies and tracking data yields various plausible grasps (see Fig. 4(a)). We then assess each of these candidates by computing the corresponding force distribution and optimizing the cost function described in Eq. (5) with PSO. Experimental results show that focusing on reducing the intensity of the force distribution yields grasps that may be substantially different from the initial hypothesis. We therefore augment the PSO objective function with a term penalizing visual discrepancy between hand pose hypothesis and observation as in [3], therefore allowing the reconstruction of physically realistic grasps that match the actual observation despite inaccurate tracking data (see Fig. 4(b)). As such, we showed that the FSV framework can be used as a valuable implicit force model for physics-based tracking, motion editing and retargeting, as human-like forces can augment the pose search with biomechanical considerations such as muscle fatigue or energy expenditure.

Discussion and future work

Towards estimating contact forces from vision, the issue of static indeterminacy was tackled by applying machine learning techniques to internal forces. An alternative approach would be to formulate the evolution of the full contact forces following various objects and grasp taxonomies as an inverse optimal control problem. If invariants are found, they could be used to refine the formulation of the optimization problem and possibly result in a better understanding of human grasping. Extending the ground truth force measurement setup with embedded three-axis or force-torque miniature sensors would also benefit both learning and optimal control approaches. The FSV framework could also expand to the robotics field for human activities monitoring, serving various purposes such as task segmentation and learning from demonstration.

[1] M. A. Brubaker, L. Sigal, and D. J. Fleet. Estimating Contact Dynamics. In *ICCV*, 2009.

[2] T. Feix, H.-B. Schmiedmayer, J. Romero, and D. Kragić. A comprehensive grasp taxonomy. In *RSS Conference: Workshop on Understanding the Human Hand for Advancing Robotic Manipulation*, 2009.

[3] N. Kyriazis and A. Argyros. Physically plausible 3d scene tracking: The single actor hypothesis. In *CVPR*, 2013.

[4] S. A. Mascaro and H. H. Asada. Photoplethysmograph fingernail sensors for measuring finger forces without haptic obstruction. *IEEE Trans. on Robotics and Automation*, 17(5):698–708, October 2001.

[5] J. Park, T. Singh, V. M. Zatsiorsky, and M. L. Latash. Optimality versus variability: effect of fatigue in multi-finger redundant tasks. *Experimental brain research*, 216(4):591–607, 2012.

[6] T.-H. Pham, A. Kheddar, A. Qammaz, and A. A. Argyros. Towards force sensing from vision: Observing hand-object interactions to infer manipulation forces. In *CVPR*, 2015.

[7] Y. Wang, J. Min, J. Zhang, Y. Liu, F. Xu, Q. Dai, and J. Chai. Video-based hand manipulation capture through composite motion control. *ACM Trans. on Graphics*, 32(4):43, 2013.