Towards Force Sensing from Vision: Observing Hand-Object Interactions to Infer Manipulation Forces

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.

Contact forces are traditionally measured by means of haptic technologies such as force transducers. A major drawback of such technologies lies in their intrusiveness, as they require to be mounted onto the manipulated objects (thus impacting their shape or other physical properties) or onto the operator's hands (thus obstructing the human haptic senses and limiting the natural range of motion). Others include their extensive need for calibration, time-varying accuracy and cost. Reliably capturing and reproducing human haptic interaction with surrounding objects by means of a cheap and simple set-up (e.g., a single RGB-D camera) would open considerable possibilities in computer vision, robotics, graphics, and rehabilitation. Computer vision research has resulted in several successful methods for capturing motion information. A challenging question is: to what extent can vision also capture haptic interaction? The latter is key for learning and understanding tasks, such as holding an object, pushing a chair or table, as well as enabling its reproduction from either virtual characters or physical embodiments.

A vision-based alternative to pressure sensors consists in correlating fingernail coloration changes to the touch force applied at fingertips [2]. The combination of marker-based tracking with tactile sensors was also explored to estimate hand joint compliance and synthesize interaction animations [3]. An inspiring use of motion tracking for force estimation, linked contact dynamics and human kinematics to estimate whole body contact forces and internal joint torques [1]. Physics-based tracking methods also involve the computation of manipulation forces [4, 7], yet in an implicit fashion, i.e., the manipulation forces are constructed to be compatible with visual observations without aiming at matching the actual forces humans apply, as happens with the method we propose in this work.

We demonstrate that, under the assumption of accurate markerless tracking with a single RGB-D camera, it is possible to compute interaction forces occurring in hand-object manipulation scenarios where object properties such as shape, contact friction μ, mass m and inertia $\mathbf{J_q}$ are known, along with human hand geometry. First, we monitor both the hand and the object motions by using model-based 3D tracking. As the observation of manipulation tasks may be subject to hand-object occlusions, we rely on a variant of the method proposed in [5] that is tailored to our needs and assume that contact points do not change significantly while they cannot be observed. From the tracking data, we then estimate hand-object contact points through proximity detection. Algebraic filtering [6] computes the object's kinematics, i.e. translational/rotational velocity $(\mathbf{v}, \omega)$ and acceleration $(\mathbf{a}, \alpha)$, resulting in target net force $\mathcal{F}_c$ and torque τ_c:

$$\begin{cases} \mathcal{F}_c = m\mathbf{a} - \mathcal{F}_d \\ \tau_c = \mathbf{J_q}\cdot\alpha + \omega\times(\mathbf{J_q}\cdot\omega) - \tau_d, \end{cases} \tag{1}$$

with $\mathcal{F}_d$ and τ_d known non-contact force and torque (e.g., gravity). Nominal contact force distributions are then computed by solving a second-order cone program (SOCP) that explains the observed kinematics through Eq. (1), using Coulomb's friction model, and minimizing the grasp's overall L^2 norm. However, when manipulating objects, humans typically apply more (internal) forces than what is required from the Newton-Euler dynamics. We address this statical indeterminacy by decomposing each finger force $\mathbf{F}_k$ into a nominal component $\mathbf{F}_k^{(n)}$ and an internal component $\mathbf{F}_k^{(i)}$:

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

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

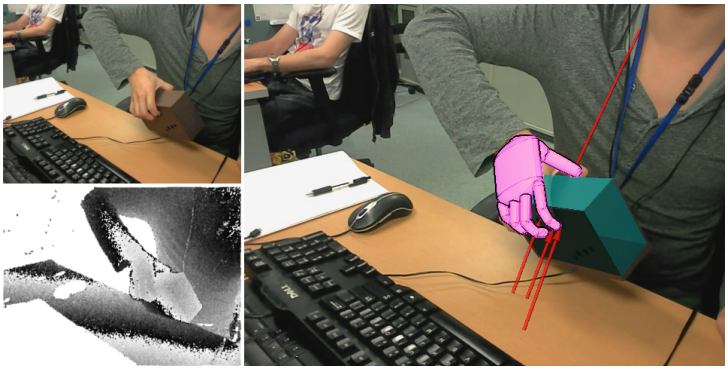

Figure 1: Using a single RGB-D camera, we track markerless hand-object manipulation tasks and estimate with high accuracy contact forces that are applied by human grasping throughout the motion.

are neutral regarding its state of equilibrium:

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

These constraints are incorporated into a new SOCP that computes the nominal-internal decompositions that best match the tactile sensors' measurements $(\widetilde{f_k})$, using a new objective function:

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

We then improve our approach by using machine learning techniques to estimate the amount and distribution of internal forces among the fingers in contact, based on grasping features that can be measured from sole vision, allowing force prediction on new experiments with a third SOCP variant.

The experimental results obtained on datasets annotated with ground-truth data from additional sensors show 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. To the best of our knowledge, this is the first time that this problem is addressed and solved based solely on markerless visual observations.

[1] Marcus A. Brubaker, Leonid Sigal, and David J. Fleet. Estimating Contact Dynamics. In *ICCV*, 2009.

[2] Thomas Grieve, John M. Hollerbach, and Stephen A. Mascaro. Force prediction by fingernail imaging using active appearance models. In *World Haptics*, pages 181–186. IEEE, 2013.

[3] Paul G. Kry and Dinesh K. Pai. Interaction capture and synthesis. *ACM Trans. on Graphics*, 25(3):872–880, 2006. ISSN 0730-0301. doi: http://doi.acm.org/10.1145/1141911.1141969.

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

[5] Nikolaos Kyriazis and Antonis A. Argyros. Scalable 3d tracking of multiple interacting objects. In *IEEE CVPR*, pages 3430–3437. IEEE, 2014.

[6] Mamadou Mboup, Cédric Join, and Michel Fliess. Numerical differentiation with annihilators in noisy environment. *Numerical Algorithms*, 50(4):439–467, 2009.

[7] Yangang Wang, Jianyuan Min, Jianjie Zhang, Yebin Liu, Feng Xu, Qionghai Dai, and Jinxiang Chai. Video-based hand manipulation capture through composite motion control. *ACM Trans. on Graphics*, 32 (4):43, 2013.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.