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, Japan [2] CNRS-UM LIRMM, IDH, France
[3] Institute of Computer Science, FORTH, Greece [4] Computer Science Department, University of Crete

Abstract

We present a novel, non-intrusive approach for estimating contact forces during hand-object interactions relying solely on visual input provided by a single RGB-D camera. We consider a manipulated object with known geometrical and physical properties. First, we rely on model-based visual tracking to estimate the object's pose together with that of the hand manipulating it throughout the motion. Following this, we compute the object's first and second order kinematics using a new class of numerical differentiation operators. The estimated kinematics is then instantly fed into a second-order cone program that returns a minimal force distribution explaining the observed motion. However, humans typically apply more forces than mechanically required when manipulating objects. Thus, we complete our estimation method by learning these excessive forces and their distribution among the fingers in contact. We provide a full validity analysis of the proposed method by evaluating it based on ground truth data from additional sensors such as accelerometers, gyroscopes and pressure sensors. Experimental results show that force sensing from vision (FSV) is indeed feasible.

1. Introduction

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 (e.g., robotic) embodiments.

Contact forces are usually measured by means of haptic technologies such as force transducers. The main drawback of such technologies is that they are obtrusive. Computer vision techniques would therefore be an ideal alternative to circumvent this issue. Yet, is it possible to estimate forces from visual observation? There is evidence that haptic perception can be induced through illusion and substitution dominated by vision, e.g. [24]. We aim at exploring computer vision to infer the forces exerted by humans on surrounding objects. In particular, we consider hand-object grasping and manipulation. The problem is extremely complex. Indeed, establishing that a hand-object contact has occurred is difficult because of occlusions and tracking inaccuracies. Nevertheless, the detection of events like an object being lifted or discontinuities in body motion may provide useful hints towards disambiguating discrete events. Additionally, even if contact positions can be determined efficiently, the estimation of the applied forces is still challenging because of the inherent multiplicity of solutions.

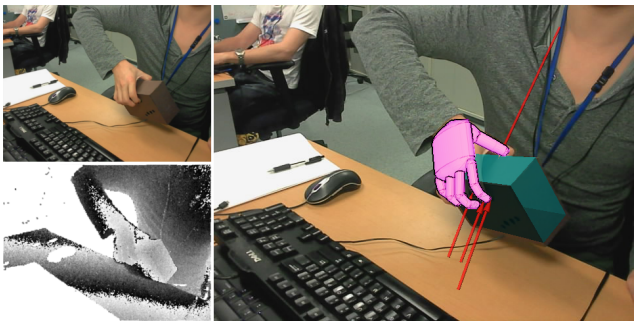

Figure 1: Using a single RGB-D camera, we track markerless hand-object manipulation tasks and estimate with high accuracy contact forces that are applied by human grasping throughout the motion.

We demonstrate that, by solely using computer vision, it is possible to compute interaction forces occurring in hand-object manipulation scenarios where object properties such as shape, contact friction, mass and inertia are known, along with human hand geometry. First, we monitor both the hand

and the object motions by using model-based 3D tracking (other visual tracking techniques can also be used if they meet performance requirements). From the tracking data, we estimate hand-object contact points through proximity detection. Algebraic filtering computes the object's kinematics, i.e. velocity and acceleration. Contact force distributions explaining the kinematic observations are then resolved using conic optimization. When manipulating objects, humans typically apply more (internal) forces than what is required from the Newton-Euler dynamics. Thus, we improve our estimation method by using neural networks to learn the amount and distribution of these internal forces among the fingers in contact. The experimental results obtained on datasets annotated with sensors' ground-truth 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.

2. Related Work

Applications of force sensing in robotics, virtual reality and human physiological studies result in challenging requirements for transduction technologies [7, 8]. Common drawbacks of mechatronic force sensing devices reside in their extensive need for calibration, time-varying accuracy (e.g., hysteresis) and cost. Additionally, they are required to be mounted onto the manipulated objects, thus impacting their shape or other physical properties, or onto the operator (e.g., force sensing gloves), thus obstructing the human haptic senses and limiting the natural range of motion.

An alternative to pressure sensors proposed in [26] consists in instrumenting fingers with miniature LEDs and photodetectors to measure changes in fingernails coloration, that are then correlated to the touch force applied at fingertips. Later, this technology evolved to predict normal and shear forces, and even changes in posture, that appear to have different blood volume patterns [27]. In [43], fingernail color and surrounding skin changes are monitored and processed using an external camera system to estimate contact fingertip forces [44, 16, 47]. Conversely, computer graphics models were developed to simulate fingertip appearance changes based on simulated forces [2]. This result illustrates that computer vision can be used effectively to measure touch forces on fingertips. This approach is, however, limited to fingertip contacts and requires extensive calibration for each individual user. It is also limited by the necessity of having fingernails visible at all time and at high resolution, hence requiring appropriately mounted miniature cameras.

Used in conjunction with force sensing technologies, motion tracking can provide information on body dynamics to explain how humans interact with their environment. A setup combining marker-based motion capture and mechanical force sensors was used in [21] to estimate hand joint compliance and synthesize interaction animations. Although precise, marker-based motion capture is invasive and hardly applicable to real-world contexts.

Alternative tracking techniques were therefore developed to circumvent markers' intrusiveness. The markerless vision-based tracking of a hand was first treated in [38] and has lately received renewed attention [32, 9, 20, 36, 45, 46]. While these works show impressive results for tracking a hand in isolation, they are not applicable to our scenario, mostly due to severe mutual occlusions from strong hand-object interaction, resulting in missing observations of both. For the 3D tracking of a hand in interaction with object(s), the existing methods can be classified into mostly bottom up, mostly top-down or hybrid, depending on the relative strength of the discriminative and the generative processes involved. The method in [39] performed 3D tracking of a hand-object interaction by synthesizing its appearance based on a non-parametric discriminative model and by treating hand pose estimation as a classification problem.

In general, top-down methods do not treat occlusions as a distractor but rather as a source of information [33, 3, 34, 41]. A generative process creates hypotheses of a joint hand-object model that are then evaluated against actual observations by solving a multi-parameter optimization problem. In the representative approach presented in [33] the availability of a joint 3D model facilitates the consideration of possible hand-object occlusions. Additionally, it makes possible the incorporation of strong priors in the optimization process, e.g., the fact that two different objects cannot share the same physical space. In this manner, even more complex problems such as tracking two strongly interacting hands [34] or tracking two hands in interaction with an object [3] have been successfully addressed. To cope with hand-object occlusions, even stronger priors resulting from a physics-based simulation of the scene have also been incorporated [22].

The scalability issues of generative methods are dealt with in [23], where it is shown that the 3D tracking of the interaction of two hands with several rigid objects can be performed accurately and efficiently. The approach taken in our work to support visual tracking of hand-object interactions is based on a variant of the method presented in [23]. A representative hybrid method is proposed in [17]. In that work, the 3D tracking of a hand manipulating an object is achieved by using bottom-up evidence to directly guide (rather than simply support) the formulation of hand pose hypotheses that is performed by generative means.

An inspiring use of motion tracking for force estimation is presented in [6], where whole body contact forces and internal joint torques are estimated by solving an optimiza-

tion problem linking contact dynamics and human kinematics. The goal of the work in [49] is the realistic synthesis of detailed hand manipulations. Towards this direction, motion capture data is used to formulate the synthesis problem taking into account contact points and contact forces. This method would benefit a lot from the unobtrusive vision-based estimation of actual contact forces we propose. Also aiming at realistic synthesis and motion reconstruction, the work in [51, 48] estimates the motion of a hand in interaction with an object together with contact points and exerted forces. Though this method successfully tracks challenging manipulation scenarios, the manipulation forces are only constructed to be compatible with visual observations, without aiming at matching the forces humans actually apply, as happens with the method we propose in this work.

3. Force Sensing from Vision (FSV)

Let $\mathcal{S}$ be a rigid body with mass m and inertia matrix $\mathbf{J}$ relative to its center of mass $\mathbf{C}$. For any element e of its contacting environment (e.g. human hand, table), let μ_e denote the corresponding Coulomb friction coefficient. We assume these quantities to be known as they can be obtained from the object of interest's CAD model or existing identification techniques [40]. Interestingly, it has been shown that aspects of such information (e.g., mass distribution) can also be estimated by visual means [30]. We then consider a scenario where $\mathcal{S}$ is grasped and manipulated by a hand, with possible contacts with the environment. Observing the scene with a sole RGB-D camera, which we suppose calibrated with the vertical direction known, our goal is to estimate the interaction forces between $\mathcal{S}$ and the user's hand, and between $\mathcal{S}$ and its environment when applicable. We decompose FSV into four subproblems as follows:

1. Track $\mathcal{S}$ and the hand and perform, for each step, vision-based proximity or collision detection to identify contacting fingers and corresponding contact points (Section 3.1).
2. Let $\mathbf{X}_i$ be the estimated pose for $\mathcal{S}$ at instant i. Based on sequence $(\mathbf{X}_i)_{i\in[0,N]}$, estimate for each frame the body's first and second-order kinematics, i.e. translational (resp. rotational) velocity $\mathbf{v}_i$ (resp. $\boldsymbol{\omega}_i$) and acceleration $\mathbf{a}_i$ (resp. $\boldsymbol{\alpha}_i$) (Section 3.2).
3. Compute a force distribution explaining the object's state computed at step 2 following Newton-Euler's laws of motion and Coulomb's friction model, using the contact points identified at step 1 (Section 3.4).
4. Learn and reproduce human internal force distributions among the fingers in contact (Section 3.6).

Each of these subproblems presents a number of challenges. First, the observation of manipulation tasks may be subject to hand-object occlusions. To overcome this issue, we address step 1 by means of model-based tracking as inspired by [23]. Second, the limited camera acquisition frequency along with tracking errors can make the differentiation process of step 2 unstable. We tackle this issue by estimating derivatives using algebraic filtering derived from [29]. Algebraic filtering was chosen for the sake of robustness, as it relies on no statistical assumption on the signal's noise. We then address step 3 by computing minimal force closure distributions as solutions of a second-order cone program. Finally, step 4 stems from the fact that in contrast with [43] where multiple photodetectors monitor each fingernail's blood flow individually, such microscopic features cannot be observed by a single RGB-D camera observing the whole scene. The object may indeed be grasped with more or less intensity without this being visible at a macroscopic scale. We tackle this statical indeterminacy with machine learning on usual human grasping practices.

3.1. Hand-object tracking

Our approach requires a good 3D pose estimate of the manipulated object together with that of the user's hand. To achieve this, we rely on a variant of the method proposed in [23] that is tailored to our needs. In [23], the model-based hand-object 3D tracking is formulated as an optimization problem, which seeks out the 3D object(s) pose and hand configuration that minimizes the discrepancy between hypotheses and actual observations. The optimization problem is solved based on PSO [10].

Since this method estimates the generalized pose of a hand interacting with an object, it is straightforward to compute the 3D positions of the estimated fingertips in relation to the object's surface (i.e., contact points). Still, in our implementation of [23], we have incorporated one important modification. The original 3D hand-object tracking framework provides solutions that are compatible with visual observations and are physically plausible in the sense that the hand and the object do not share the same physical space (i.e., the hand does not penetrate the modeled volume of the object). However, occluded fingers may have different poses that respect the above constraints, making the estimation of contact points an under-constrained problem. To overcome this issue, we assume that contact points do not change significantly when they cannot be observed. Time and space coherency is thus enforced by penalizing solutions in which hidden contact points are far from their last observed position.

3.2. Numerical differentiation for kinematics

In theory, velocity and acceleration can be estimated by numerical differentiation of poses obtained from tracking. However, this process is highly dependent on two factors: (a) the acquisition frequency of the RGB-D frames, and (b) the quality of the motion tracking. First, even a perfect tracking would result in poor velocity and acceleration

estimates if performed over time steps far apart from each other, also depending on the way the hand moves. However, this is not a freely controllable parameter, as most commercial RGB-D cameras offer acquisition frame-rates capped between 30 and 60 fps. We present our results on a 30 fps SoftKinetic DepthSense 325 camera. Second, acceleration profiles occurring in manipulation tasks are naturally spiky (see for example Fig. 4). Therefore, numerical differentiation is challenging in that while the number of samples used for each derivative estimate must be sufficient to alleviate tracking errors, it must also be kept minimal to discern the sudden variations that acceleration profiles are subject to.

As an alternative to existing numerical differentiation methods, algebraic parameter estimation approaches [12] led to a new class of derivative estimators called algebraic numerical differentiators [29]. The tracking errors resulting from the employed model-based tracking framework seem to follow a Gaussian distribution, yet they are not independent of one another, which rules out the white noise formalism. Subsequently, and in order to keep the kinematics estimation process unbiased by the use of a particular tracking method, we implement the so-called minimal (κ, μ) algebraic numerical differentiators, which do not assume prior knowledge of the signal errors' statistical properties.

3.3. From kinematics to dynamics

We suppose the rigid body $\mathcal{S}$ subject to n_d non-contact forces $(\mathbf{F}_k^d)_{k\in[1,n_d]}$ applied at points $(\mathbf{P}_k^d)_{k\in\{1,\dots,n_d\}}$ (e.g., gravitation, electromagnetism). We consider them fully known based on the object's properties. We seek to estimate n_c contact forces $(\mathbf{F}_k^c)_{k\in[1,n_c]}$ applied at contact points with the hand or the environment $(\mathbf{P}_k^c)_{k\in[1,n_c]}$ that are obtained from tracking (Section 3.1). Using the object's kinematics as estimated in Section 3.2, its motion is governed by Newton-Euler equations. Therefore, the resulting net force $\boldsymbol{\mathcal{F}}_\mathrm{c}$ and torque $\boldsymbol{\tau}_c$ due to the contact forces are such that:

$$\begin{cases} \boldsymbol{\mathcal{F}}_\mathrm{c} = m\mathbf{a} - \boldsymbol{\mathcal{F}}_\mathrm{d} \\ \boldsymbol{\tau}_\mathrm{c} = \mathbf{J}_\mathbf{q}\cdot\boldsymbol{\alpha} + \boldsymbol{\omega}\times(\mathbf{J}_\mathbf{q}\cdot\boldsymbol{\omega}) - \boldsymbol{\tau}_\mathrm{d}, \end{cases} \tag{1}$$

with $\boldsymbol{\mathcal{F}}_\mathrm{d}$ and $\boldsymbol{\tau}_d$ the resulting force and torque due to non-contact forces, and $\mathbf{J}_\mathbf{q}$ the inertia matrix with orientation $\mathbf{q}$.

The contact forces are subject to friction, which we model using Coulomb's law. Let $\mathbf{n}_k$ be the unit contact normal oriented inwards $\mathcal{S}$ at contact point $\mathbf{P}_k^c$. Let then $\mathbf{t}_k^x$ and $\mathbf{t}_k^y$ be two unit vectors orthogonal to each other and to the normal $\mathbf{n}_k$, thus defining the tangent plane. Each contact force $\mathbf{F}_k^c$ is decomposed as follows:

$$\mathbf{F}_k^c = f_k\mathbf{n}_k + g_k\mathbf{t}_k^x + h_k\mathbf{t}_k^y, \tag{2}$$

With μ_k the friction coefficient at $\mathbf{P}_k^c$, Coulomb's law reads:

$$\|g_k\mathbf{t}_k^x + h_k\mathbf{t}_k^y\|_2 \leq \mu_k f_k, \tag{3}$$

which is a strict equality in the case of dynamic friction.

3.4. Nominal forces from cone programming

We address the estimation of the minimal contact forces responsible for the observed motion (i.e., nominal forces) as a second-order cone program (SOCP) [25, 4, 5]:

$$\begin{aligned} \min\quad & \mathcal{C}(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{P}\mathbf{x} + \mathbf{r}^T\mathbf{x} \\ \text{s. t.}\quad & \|\mathbf{A}_j\mathbf{x} + \mathbf{b}_j\|_2 \leq \mathbf{c}^T\mathbf{x} + \mathbf{d}_j, \quad j = 1,\dots,m \\ \text{and}\quad & \mathbf{F}\mathbf{x} = \mathbf{g}. \end{aligned} \tag{4}$$

As we track the object and the user's hand, we can determine, at each timeframe, newly established and broken contacts, and also those that remain still and those that slide. Therefore, we are explicitly considering static and kinetic (i.e. dynamic) friction in the constraints formulation. With $n_{c,s}$ and $n_{c,k}$ the respective numbers of friction forces and n_c their sum, we choose the optimization vector as follows:

$$\begin{aligned} \mathbf{x} = (&f_1, g_1, h_1, \dots, f_{n_{c,s}}, g_{n_{c,s}}, h_{n_{c,s}}, \\ &f_{n_{c,s}+1}, f_{n_{c,s}+2}, \dots, f_{n_{c,s}+n_{c,k}})^T \end{aligned} \tag{5}$$

The SOCP formulation in Eq. (4) then allows the direct handling of Coulomb static friction as inequality constraints and kinetic friction as equality constraints. Moreover, having each normal vector $\mathbf{n}_k$ oriented inwards $\mathcal{S}$, n_c additional positivity constraints are added such that $f_k \geq 0$. Equality constraints enforcing that the resulting contact force distribution explains the observed kinematics stem from Newton-Euler's equations, as combining Eq. (1) with contact force expressions from Eq. (2) directly yields linear equations in $\mathbf{x}$. We complete the SOCP with the objective function:

$$\mathcal{C}_\mathrm{SoS}(\mathbf{x}) = \sum_{k\in\mathcal{F}}\left[f_k^2 + g_k^2 + h_k^2\right] = \sum_{k\in\mathcal{F}}\|\mathbf{F}_k^c\|_2^2, \tag{6}$$

where $\mathcal{F}$ is the set of contacting fingers. As stated earlier in Section 3, there exists an infinity of possible force distributions for a given kinematics and set of contact points. Using the contact forces' sum of squares as an indicator on the grasp intensity (i.e., its L^2 norm), the objective function $\mathcal{C}_\mathrm{SoS}$ allows to search for the optimal grasp in that sense. Numerical resolution is performed using CVXOPT [1].

3.5. Reproducing human grasping forces

Humans do not manipulate objects using nominal closures (i.e. minimal grasp forces). They tend to "over-grasp" and produce workless internal forces, i.e. firmer grasps than mechanically required. This human grasp property is described by considering finger forces as two sets [19, 50]: nominal forces responsible for the object's motion, and internal forces that secure the object through a firm grip but do not affect the object's kinematics as they cancel each other out [28, 31]. Studies showed that humans apply internal forces to prevent slip [18, 11] and control their magnitude

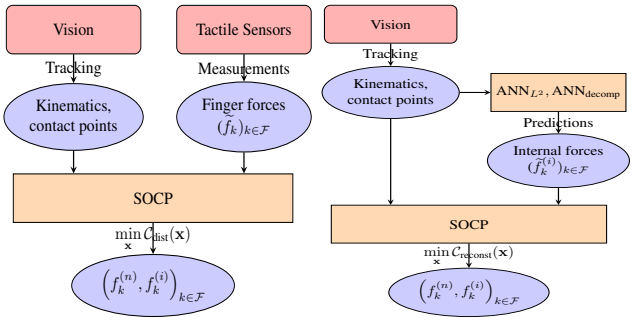

Figure 2: (a) Measurements from tactile sensors are used to estimate nominal and internal force decompositions from vision. (b) Full contact forces are reconstructed by feeding ANN predictions into a third SOCP variant.

to avoid muscle fatigue or damaging fragile objects [15, 35]. We extend the formulation of the SOCP to this decomposition and learn from tactile sensors' measurements how humans apply internal forces when manipulating objects.

Each finger force $\mathbf{F}_k$ is decomposed into a nominal component $\mathbf{F}_k^{(n)}$ and an internal component $\mathbf{F}_k^{(i)}$:

$$\begin{aligned} &\mathbf{F}_k = \mathbf{F}_k^{(n)} + \mathbf{F}_k^{(i)} \\ &\text{with} \quad \begin{cases} \mathbf{F}_k^{(n)} = f_k^{(n)}\mathbf{n}_k + g_k^{(n)}\mathbf{t}_k^x + h_k^{(n)}\mathbf{t}_k^y \\ \mathbf{F}_k^{(i)} = f_k^{(i)}\mathbf{n}_k + g_k^{(i)}\mathbf{t}_k^x + h_k^{(i)}\mathbf{t}_k^y. \end{cases} \end{aligned} \tag{7}$$

It is to mention that although both forces are decomposed along the same contact frame $(\mathbf{n}_k, \mathbf{t}_k^x, \mathbf{t}_k^y)$ as in Eq. (2), nothing constraints them to be similarly oriented. We subsequently gather the nominal and internal components into a new optimization vector $\mathbf{x}$ as in Eq. (5).

By definition, nominal forces are responsible for the object's motion through the Newton-Euler equations while internal forces are neutral regarding its state of equilibrium:

$$\begin{cases} \sum_{k\in\mathcal{F}} \mathbf{F}_k^{(n)} = \boldsymbol{\mathcal{F}}_{\text{c}}, & \sum_{k\in\mathcal{F}} \overrightarrow{\mathbf{CP}_k} \times \mathbf{F}_k^{(n)} = \boldsymbol{\tau}_{\text{c}} \\ \sum_{k\in\mathcal{F}} \mathbf{F}_k^{(i)} = \mathbf{0}, & \sum_{k\in\mathcal{F}} \overrightarrow{\mathbf{CP}_k} \times \mathbf{F}_k^{(i)} = \mathbf{0}. \end{cases} \tag{8}$$

Equation (8) provides a new set of constraints that we integrate into the SOCP of Section 3.4. Ensuring that the resulting distribution still obeys Coulomb's law of friction, we finally compute the distribution of nominal and internal forces that best match the tactile sensors' measurements $(\widetilde{f}_k)_{k\in\mathcal{F}}$, using a new objective function:

$$\mathcal{C}_{\text{dist}}(\mathbf{x}) = \sum_{k\in\mathcal{F}} \left[\left\| \mathbf{F}_k^{(n)} \right\|_2^2 + \left(f_k^{(n)} + f_k^{(i)} - \widetilde{f}_k \right)^2 \right]. \tag{9}$$

The reason why we do not directly identify internal forces as the differences between the measurements $\widetilde{f}_k$ and the minimal forces resulting from the initial SOCP of Section 3.4 is that possible sensor measurement errors may lead them not to compensate each other. By integrating their computation into the SOCP, we ensure that the resulting internal forces $f_k^{(i)}$ bridge the gap between $f_k^{(n)}$ and measurements $\widetilde{f}_k$ without perturbing the object's observed kinematics. We illustrate the decomposition process in Fig. 2(a).

3.6. Learning internal force distributions

Recent studies attempted to build mathematical models correlating grasp forces to kinematic data, yet limited to cyclic movement patterns and two-finger grasps [14, 42], hence concealing the issue of statical indeterminacy. In contrast, our approach learns how humans apply internal forces by means of artificial neural networks (ANN). We first construct an experimental dataset by having human operators manipulate an instrumented box (see Section 4) over tasks such as pick-and-place, lift and release, rotations, and unguided compositions of these. Experiments were conducted over a pool of three female, including one left-handed, and three male operators using their preferred hand on different contact and object mass configurations. Executing 160 manipulation experiments of approximately 10 s duration, we perform motion tracking and record the tactile sensors' measurements to compute the best-matching decompositions $\left(f_k^{(n)}, f_k^{(i)}\right)$ following the SOCP of Section 3.5.

The next step is to learn the variations of internal forces $f_k^{(i)}$ with motion and grasping features. We select the learning parameters as those that directly impact the force distributions through the Newton-Euler equations. Contact forces vary with the object's mass and acceleration, or more accurately including the contribution of gravity $\boldsymbol{\mathcal{F}}_{\text{c}} = m \cdot (\mathbf{a} - \mathbf{g})$. We can consider this dependence as twofold: on the magnitude of $\boldsymbol{\mathcal{F}}_{\text{c}}$ itself, and on the relative orientation of $\boldsymbol{\mathcal{F}}_{\text{c}}$ with the contact normals, as in [14]:

$$p_1 = \|\boldsymbol{\mathcal{F}}_{\text{c}}\|_2 \tag{10}$$

$$p_{2,k} = \mathbf{n}_k \cdot \mathbf{u}_{\boldsymbol{\mathcal{F}}_{\text{c}}}, \quad \text{with} \quad \mathbf{u}_{\boldsymbol{\mathcal{F}}_{\text{c}}} = \frac{\boldsymbol{\mathcal{F}}_{\text{c}}}{\|\boldsymbol{\mathcal{F}}_{\text{c}}\|_2}. \tag{11}$$

Similarly, we consider the case of rotational kinematics through the magnitude of the contact torque $\boldsymbol{\tau}_{\text{c}}$ of Eq. (1) and the individual torques each finger is able to generate:

$$p_3 = \|\boldsymbol{\tau}_{\text{c}}\|_2 \tag{12}$$

$$p_{4,k} = \left(\overrightarrow{\mathbf{CP}_k} \times \mathbf{n}_k \right) \cdot \mathbf{u}_{\boldsymbol{\tau}_{\text{c}}}, \quad \text{with} \quad \mathbf{u}_{\boldsymbol{\tau}_{\text{c}}} = \frac{\boldsymbol{\tau}_{\text{c}}}{\|\boldsymbol{\tau}_{\text{c}}\|_2}. \tag{13}$$

Finally, we learn internal forces as a function of kinematics and grasp parameters $(p_1, (p_{2,k})_{k\in\mathcal{F}}, p_3, (p_{4,k})_{k\in\mathcal{F}})$ using two ANNs: a first network, ANN_{L^2}, estimates the

amount of internal forces applied, quantified as the overall L^2 norm, while a second network, $\text{ANN}_{\text{decomp}}$, jointly estimates the relative participation of each finger in the grasp's intensity. The outputs of $\text{ANN}_{\text{decomp}}$ are percentages constructed as the individual forces normalized with the overall L^2 norm. Note that, as that similar motions can stem from different force distributions, using a single ANN would mean linking similar inputs to highly varying individual forces. Yet, we observed that different grasp intensities still tend to be similarly shared among fingers, hence two ANNs to account for natural intensity variance but consistent decompositions. In order to avoid samples where FSV or tactile sensors are not reliable, we only use as learning data those where the resulting net forces are within a specified threshold from each other. In our experiments, setting this threshold to $1.5N$ yields a final dataset of 8200 samples, which we partition into training and validation datasets to construct and assess different ANN configurations by cross-validation. Performing numerical resolution with the neuralnet package for statistical analysis software R [13, 37], we choose ANN_{L^2} and $\text{ANN}_{\text{decomp}}$ with logistic neurons trained with resilient backpropagation and two hidden layers, with respectively 6 and 8 neurons in the first hidden layer, and 7 and 13 neurons in the second.

4. Experiments

In order to assess our approach, we perform manipulation experiments on a rectangular cuboid of dimensions 171mm $\times$ 111mm $\times$ 60mm. The simplified shape of this ground-truth object is chosen to meet sensing instrumentation constraints and offer several grasping possibilities. We instrument the box with two types of sensors. The first is an Xsens MTi-300 attitude and heading reference system (AHRS) motion sensor measuring reference rotational velocities and translational accelerations. Its purpose is to validate the numerical differentiation of tracking data by algebraic filtering, see Section 3.2. The second consists of five Honeywell FSG020WNPB piezoresistive one-axis force sensors that can be positioned at different predefined grasp spots on the box. We finally evaluate the contact forces estimated from the SOCP in Section 3.4 with the force sensors' measurements in terms of: (i) normal forces per finger, (ii) resulting net force, and (iii) sum of squares. We summarize the validation protocol in Fig. 3.

4.1. Kinematics from vision vs AHRS

We assess the validity of our approach by executing motions emphasizing each of the three coordinates of both translation accelerations and rotational velocities, and comparing the kinematics estimated from vision to measurements from the Xsens MTi-300 AHRS. Statistical analysis of the estimation errors shows that algebraic numerical differentiation is well suited for kinematics estimation (see

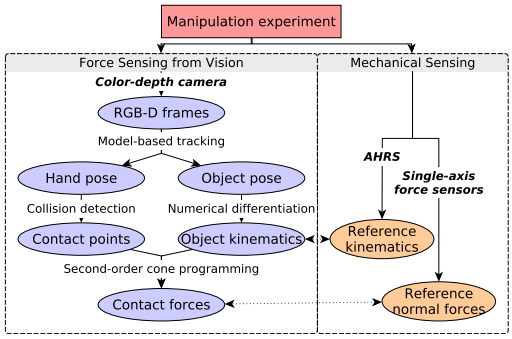

Figure 3: Validation protocol.

		Central	Gaussian	Algebraic
Trans. acc. $[m \cdot s^{-2}]$	Avg.	-0.029	$\mathbf{-0.022}$	-0.024
	St.d.	1.686	1.627	**0.904**
Rot. vel. $[rad \cdot s^{-1}]$	Avg.	0.084	0.070	**0.052**
	St.d.	1.559	1.294	**1.241**

Table 1: Kinematics estimation errors for central finite difference, Gaussian filtering, and algebraic filtering.

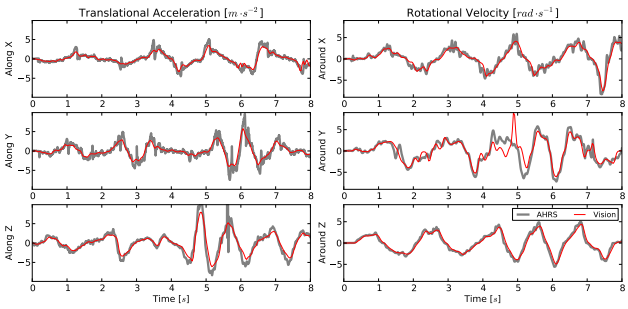

Figure 4: Comparison between vision-based kinematics and AHRS-embedded accelerometer and gyroscope.

Table 1). Though on translational acceleration the error is slightly higher than with Gaussian filtering, its variance is also considerably lower. Its performance on rotational kinematics is also the best of all three tested approaches. We summarize the results of the six-axis experiments in Fig. 4.

4.2. Nominal forces from vision-based kinematics

We now validate our vision-based force estimation framework using Honeywell FSG020WNPB sensors placed at pre-specified positions over the instrumented box. As a first validation step, contact points obtained from vision were compared to the expected contact points based on the sensors' locations and resulted in estimation errors of mean $-1.55mm$ and standard deviation $6.13mm$. Furthermore, we assessed the sensitivity of FSV to these uncertainties by comparing the force distributions obtained using either the

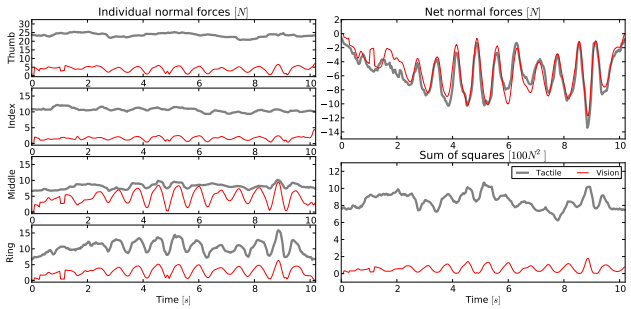

Figure 5: Contact forces from vision based on L^2 criterion are individually lower than tactile sensors' measurements but result in the same net force.

contact points from vision or the tactile sensors' positions. We found that FSV is relatively robust to such estimation errors, resulting in force uncertainties of mean $0.216N$ and standard deviation $1.548N$. Therefore, we rely solely on vision-based kinematics and contact points for the rest of this work. When performing experiments, we also observed that the force applied by the pinky finger was consistently below the sensitivity threshold of our force sensors, hence we present our results on four-finger experiments. Still, as the force distribution problem introduced in Section 3 becomes statically indeterminate from three fingers, using four fingers preserve the grasps' complexity and does not impact the generality of our results. We represent the force sensors' measurements along with FSV's outputs in Fig. 5.

As mentioned in Section 3.4, the comparison of the normal components from vision and from tactile sensors shows that the latter's measurements are overall greater. This illustrates the fact that humans seize objects harder than the required force closure, in contrast with the L^2-optimal grasp estimated from vision, which is visible in the sum of squares plot. Still, the resulting net forces are matching well, which demonstrates that FSV can successfully capture the object's motion characteristics and compute a reliable force distribution explaining the observed kinematics.

4.3. Reconstructing full contact force distributions

By recording new manipulation experiments, we extract as in Section 3.6 the kinematics and grasping parameters $(p_1, (p_{2,k})_{k\in\mathcal{F}}, p_3, (p_{4,k})_{k\in\mathcal{F}})$ over time and use the trained ANNs to predict the internal forces the human operator most likely applied over the experiment, $(\widetilde{f}_k^{(i)})_{k\in\mathcal{F}}$. We finally construct the final contact force distributions using the variant of the SOCP described in Section 3.5 that features an objective function which aims at matching not the full contact forces but only their internal components:

$$\mathcal{C}_{\text{reconst}}(\mathbf{x}) = \sum_{k\in\mathcal{F}} \left[\left\| \mathbf{F}_k^{(n)} \right\|_2^2 + \left(f_k^{(i)} - \widetilde{f}_k^{(i)} \right)^2 \right]. \quad (14)$$

User	Mass	Grasp	Part. training Avg. [%]	Part. training St.d. [%]	Full training Avg. [%]	Full training St.d. [%]
◯	◯	◯	10.7	12.4	9.71	12.0
×	◯	◯	10.9	12.3	10.3	11.8
◯	×	◯	10.8	11.3	10.4	12.4
◯	◯	×	14.6	14.5	10.9	11.3
×	×	×	14.9	14.8	9.94	12.6

Table 2: Relative force estimation errors based on the exhaustivity of the training dataset. ◯ and × indicate features that respectively appear or not in the partial training dataset.

We illustrate the final estimation process in Fig. 2(b). By feeding the internal ANNs' predictions into the SOCP, we ensure that the final internal forces $\left(\mathbf{F}_k^{(i)}\right)_{k\in\mathcal{F}}$ are not only consistent with natural grasping patterns but also physically correct and do not impact the object's observed kinematics through the resulting net force, as shown in Fig. 6.

4.4. Robustness analysis

We investigate the robustness of our approach to features that do not appear in the training dataset. To this end, we train another instance of the ANNs described in Section 3.6, not over the entire dataset but on a partial subset relative to a single operator, on a single grasp pose, and a single mass configuration. We then evaluate the resulting ANNs on datasets obtained with another user, another grasp, and/or a 10% mass increase. We report the relative errors with respect to the tactile sensors' measurements in Table 2 along with reference results from fully-trained ANNs.

First, it appears that ANNs trained over a single operator may be generalized to other users with no significant performance decrease, which suggests that humans tend to apply internal forces following similar patterns. Second, reasonable changes in mass do not seem to significantly impact the estimation accuracy either. This is allowed by the fact that in our problem formulation, mass is not a training variable by itself but is implicitly taken into account through the product $\boldsymbol{\mathcal{F}}_c = m \cdot (\mathbf{a} - \mathbf{g})$. Under this formalism, manipulating a heavy object with a given kinematics is analogous to manipulating a lighter object with a higher acceleration. Therefore, the ANNs may accomodate mass changes provided that they were trained over a sufficient variety of kinematics. In the end, ANNs seem most sensitive to grasp pose changes. This may be explained by the fact that placing fingers differently may substantially change their synergies. Still, the performance decrease remains reasonable while force distributions, by construction, still explain the observed motion. Eventually, the main sensitivity to grasp poses is comforted by the fact that also changing user and mass does not decrease the estimation accuracy further.

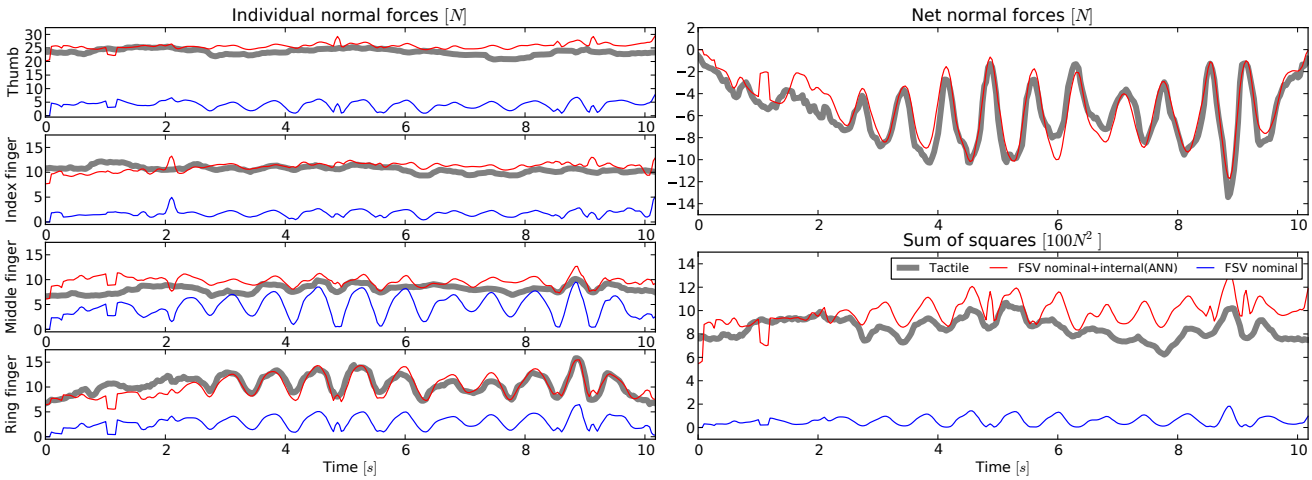

Figure 6: Artificial neural networks used in conjunction with cone programming successfully predict force distributions that both explain the observed motion and follow natural human force distribution patterns.

5. Summary and Discussion

Our work establishes that a single RGB-D camera can be used to capture interaction forces occurring in rigid object manipulation by a human hand without the need for visual markers or dedicated force sensing devices. Force sensing from vision is a novel and important contribution since it circumvents the intrusive instrumentation of object, environment and hands. Its exploitation can expand to the robotics field for daily on-line human activities monitoring, serving various purposes such as imitation learning.

Our method is validated with several experiments based on ground truth data and is able to estimate fairly accurately the force distributions applied during actual manipulation experiments. Although we confirmed that tracking noise is well mitigated by algebraic filtering, which produces truthful pose derivative estimates, guessing the hand-object contact points under strong occlusions remains a challenging, open problem in computer vision. We achieved this by using a state-of-the-art model-based tracking method under the somewhat practical assumption that occluded fingers remain at their last observed position until they are visible again. While this assumption is fairly valid in numerous interesting cases, it is not true when considering tasks such as dexterous manipulation with finger repositioning or sliding. Still, this limitation does not call into question the force estimation framework per se, and could be alleviated by extending the markerless tracking method to multi-camera inputs, which would remain non-intrusive and keep an edge over tactile sensors regarding usability and cost.

With respect to computational performance, SOCP and internal force predictions are performed in real-time, and only hand-object tracking is computationally expensive. Given the recent developments on GPGPU implementations of hand-object tracking [23], our framework could be employed in real-time applications. This, combined with our reliance on a single camera, makes FSV suitable for daily observation and learning. Still, our approach is generic enough to accommodate any advance to the topic of 3D hand tracking and could be seamlessly extended to other methods, for instance when non real-time performance and a heavier setup are possible. Conversely, our framework could also be used as an implicit force model for physics-based tracking and motion editing, as human-like forces could augment the pose search with biomechanical considerations such as muscle fatigue or energy expenditure.

Towards estimating contact forces from vision, we tackled the issue of static indeterminacy by applying machine learning techniques to internal forces. Rather than predicting new force distributions based on past observations, 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 cost function, which could result in more reliable contact forces than the nominal distributions computed by minimization of the grasp's L^2-norm. 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. Further work could also address the case of surface contact models in place of point contacts (as the fingertip is deforming), namely for dexterous manipulations, or make use of synergy properties of the hand for bimanual tasks. Finally, combining our approach with visual SLAM or automated camera calibration methods would allow it to be deployed in unknown, varying environments, e.g. on mobile robots.

Acknowledgements This work was partially supported by the FP7 EU RoboHow.Cog project.

References

[1] M. Andersen, J. Dahl, and L. Vandenberghe. Cvxopt: A python package for convex optimization, version 1.1.6. `abel.ee.ucla.edu/cvxopt`, 2013.

[2] S. Andrews, M. Jarvis, and P. Kry. Data-driven fingertip appearance for interactive hand simulation. *ACM SIGGRAPH conference on Motion in Games (MIG)*, 2013.

[3] L. Ballan, A. Taneja, J. Gall, L. Van Gool, and M. Pollefeys. Motion capture of hands in action using discriminative salient points. In *ECCV*, 2012.

[4] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004.

[5] S. P. Boyd and B. Wegbreit. Fast computation of optimal contact forces. *IEEE Trans. on Robotics*, 23(6):1117–1132, 2007.

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

[7] M. R. Cutkosky, R. D. Howe, and W. R. Provancher. *Springer Handbook of Robotics*, chapter Force and Tactile Sensors, pages 455–476. Springer, Berlin, Heidelberg, 2008.

[8] R. S. Dahiya, G. Metta, M. Valle, and G. Sandini. Tactile sensing - from humans to humanoids. *IEEE Trans. on Robotics*, 26(1):1–20, 2010.

[9] M. de La Gorce, D. J. Fleet, and N. Paragios. Model-based 3d hand pose estimation from monocular video. *IEEE Trans. on PAMI*, 33(9):1793–1805, 2011.

[10] R. C. Eberhart, Y. Shi, and J. Kennedy. *Swarm Intelligence (The Morgan Kaufmann Series in Evolutionary Computation)*. Morgan Kaufmann, 1 edition, Apr. 2001.

[11] J. R. Flanagan and R. S. Johansson. Hand movements. *Encyclopedia of the human brain*, 2:399–414, 2002.

[12] M. Fliess and H. Sira-Ramírez. An algebraic framework for linear identification. *ESAIM: Control, Optimisation and Calculus of Variations*, 9:151–168, 7 2003.

[13] S. Fritsch, F. Guenther, and M. Suling. Package neuralnet, version 1.32. `http://CRAN.R-project.org/package=neuralnet`, 2012.

[14] F. Gao, M. L. Latash, and V. M. Zatsiorsky. Internal forces during object manipulation. *Experimental brain research*, 165(1):69–83, 2005.

[15] S. L. Gorniak, V. M. Zatsiorsky, and M. L. Latash. Manipulation of a fragile object. *Experimental brain research*, 202(2):413–430, 2010.

[16] T. Grieve, J. M. Hollerbach, and S. A. Mascaro. Force prediction by fingernail imaging using active appearance models. In *World Haptics*, pages 181–186. IEEE, 2013.

[17] H. Hamer, K. Schindler, E. Koller-Meier, and L. Van Gool. Tracking a hand manipulating an object. In *ICCV*, 2009.

[18] R. Johansson and G. Westling. Roles of glabrous skin receptors and sensorimotor memory in automatic control of precision grip when lifting rougher or more slippery objects. *Experimental Brain Research*, 56(3):550–564, 1984.

[19] J. Kerr and B. Roth. Analysis of multifingered hands. *IJRR*, 4(4):3–17, 1986.

[20] C. Keskin, F. Kıraç, Y. E. Kara, and L. Akarun. Hand pose estimation and hand shape classification using multi-layered randomized decision forests. In *-ECCV*, pages 852–863. Springer, 2012.

[21] P. G. Kry and D. K. Pai. Interaction capture and synthesis. *ACM Trans. on Graphics*, 25(3):872–880, 2006.

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

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

[24] A. Lécuyer, S. Coquillart, A. Kheddar, P. Richard, and P. Coiffet. Pseudo-haptic feedback: Can isometric input devices simulate force feedback? In *IEEE Virtual Reality Conference*, pages 83–90, New Brunswick, NJ, 18-22 March 2000. IEEE Computer Society.

[25] M. S. Lobo, L. Vandenberghe, S. Boyd, and H. Lebret. Applications of second-order cone programming. *Linear Algebra and its Applications*, 284:193–228, 1998.

[26] 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.

[27] S. A. Mascaro and H. H. Asada. Measurement of finger posture and three-axis fingertip touch force using fingernail sensors. *IEEE Trans. on Robotics and Automation*, 20(1):26–35, February 2004.

[28] M. T. Mason and J. K. Salisbury. *Robot Hands and the Mechanics of Manipulation*. MIT Press, Cambridge, MA, 1985.

[29] M. Mboup, C. Join, and M. Fliess. Numerical differentiation with annihilators in noisy environment. *Numerical Algorithms*, 50(4):439–467, 2009.

[30] A. Mkhitaryan and D. Burschka. Visual estimation of object density distribution through observation of its impulse response. In *VISAPP (1)*, pages 586–595, 2013.

[31] R. M. Murray, S. S. Sastry, and L. Zexiang. *A Mathematical Introduction to Robotic Manipulation*. CRC Press, Inc., Boca Raton, FL, USA, 1st edition, 1994.

[32] I. Oikonomidis, N. Kyriazis, and A. Argyros. Efficient model-based 3d tracking of hand articulations using kinect. In *BMVC*. BMVA, 2011.

[33] I. Oikonomidis, N. Kyriazis, and A. Argyros. Full dof tracking of a hand interacting with an object by modeling occlusions and physical constraints. In *ICCV*. IEEE, 2011. Oral presentation.

[34] I. Oikonomidis, N. Kyriazis, and A. Argyros. Tracking the articulated motion of two strongly interacting hands. In *CVPR*. IEEE, June 2012.

[35] 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.

[36] C. Qian, X. Sun, Y. Wei, X. Tang, and J. Sun. Realtime and robust hand tracking from depth. In *IEEE CVPR*, 2014.

[37] R Core Team. R: A language and environment for statistical computing. R-project.org, 2014.

[38] J. M. Rehg and T. Kanade. Visual tracking of high dof articulated structures: an application to human hand tracking. In *ECCV*, pages 35–46. Springer Berlin Heidelberg, 1994.

[39] J. Romero, H. Kjellström, C. Ek, and D. Kragic. Non-parametric hand pose estimation with object context. *Image and Vision Computing*, 2013.

[40] C. Schedlinski and M. Link. A survay of current inertia parameter identification methods. *Mechanical Systems and Signal Processing*, 15(1):189 – 211, 2001.

[41] T. Schmidt, R. Newcombe, and D. Fox. Dart: Dense articulated real-time tracking. *RSS*, 2014.

[42] G. Slota, M. Latash, and V. Zatsiorsky. Grip forces during object manipulation: experiment, mathematical model, and validation. *Experimental Brain Research*, 213(1):125–139, 2011.

[43] Y. Sun, J. M. Hollerbach, and S. A. Mascaro. Predicting fingertip forces by imaging coloration changes in the fingernail and surrounding skin. *IEEE Trans. on Biomedical Engineering*, 55(10):2363–2371, October 2008.

[44] Y. Sun, J. M. Hollerbach, and S. A. Mascaro. Estimation of fingertip force direction with computer vision. *IEEE Trans. on Robotics*, 25(6):1356–1369, December 2009.

[45] D. Tang, H. J. Chang, A. Tejani, and T.-K. Kim. Latent regression forest: Structured estimation of 3d articulated hand posture. In *IEEE CVPR*, 2014.

[46] J. Tompson, M. Stein, Y. Lecun, and K. Perlin. Real-time continuous pose recovery of human hands using convolutional networks. *ACM Trans. on Graphics*, 33(5):169, 2014.

[47] S. Urban, J. Bayer, C. Osendorfer, G. Westling, B. B. Edin, and P. van der Smagt. Computing grip force and torque from finger nail images using gaussian processes. In *IROS*, pages 4034–4039. IEEE, 2013.

[48] 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.

[49] Y. Ye and C. K. Liu. Synthesis of detailed hand manipulations using contact sampling. *ACM Tran. on Graphics*, 31(4):41, 2012.

[50] T. Yoshikawa and K. Nagai. Manipulating and grasping forces in manipulation by multifingered robot hands. *IEEE Trans. on Robotics and Automation*, 7(1):67–77, Feb 1991.

[51] W. Zhao, J. Zhang, J. Min, and J. Chai. Robust realtime physics-based motion control for human grasping. *ACM Trans. on Graphics*, 32(6):207:1–207:12, Nov. 2013.