Estimating 3D Motion and Forces of Person-Object Interactions from Monocular Video

Zongmian Li\textsuperscript{1,2} Jiri Sedlar\textsuperscript{3} Justin Carpentier\textsuperscript{1,2} Ivan Laptev\textsuperscript{1,2} Nicolas Mansard\textsuperscript{4} Josef Sivic\textsuperscript{1,2,3}

\textsuperscript{1}DI ENS PSL \textsuperscript{2}Inria \textsuperscript{3}CIIRC, CTU in Prague \textsuperscript{4}LAAS-CNRS

Abstract

In this paper, we introduce a method to automatically reconstruct the 3D motion of a person interacting with an object from a single RGB video. Our method estimates the 3D poses of the person and the object, contact positions, and forces and torques actuated by the human limbs. The main contributions of this work are three-fold. First, we introduce an approach to jointly estimate the motion and the actuation forces of the person on the manipulated object by modeling contacts and the dynamics of their interactions. This is cast as a large-scale trajectory optimization problem. Second, we develop a method to automatically recognize from the input video the position and timing of contacts between the person and the object or the ground, thereby significantly simplifying the complexity of the optimization. Third, we validate our approach on a recent MoCap dataset with ground truth contact forces and demonstrate its performance on a new dataset of Internet videos showing people manipulating a variety of tools in unconstrained environments.

1. Introduction

People can easily learn how to break concrete with a sledgehammer or cut hay using a scythe by observing other people performing such tasks in instructional videos, for example. They can also easily perform the same task in a different context. This involves advanced visual intelligence capabilities such as recognizing and interpreting complex person-object interactions that achieve a specific goal. Understanding such complex interactions is a key to building autonomous machines that learn to interact with the physical world by observing people.

Figure 1: Our method automatically reconstructs in 3D the object manipulation action captured by a single RGB video. \textbf{Top:} Frames from the input video. \textbf{Bottom:} the output human and object 3D motion including the recovered contact forces (yellow arrows) and moments (white arrows).

This work makes a step in this direction and describes a method to estimate the 3D motion and actuation forces of a person manipulating an object given a single unconstrained video as input, as shown in Figure 1. This is an extremely challenging task. First, there are inherent ambiguities in the 2D-to-3D mapping from a single view: multiple 3D human poses correspond to the same 2D input. Second, human-object interactions often involve contacts, resulting in discontinuities in the motion of the object and the human body part in contact. For example, one must place a hand on the hammer handle before picking the hammer up. The contact motion strongly depends on the physical quantities such as the mass of the object and the contact forces exerted by the hand, which renders modeling of contacts a very difficult task. Finally, the tools we consider in this work, such as hammer, scythe, or spade, are particularly difficult to recognize due to their thin structure, lack of texture, and frequent occlusions by hands and other human parts.

To address these challenges, we propose a method to jointly estimate the 3D trajectory of both the person and the object by visually recognizing contacts in the video.
and modeling the dynamics of the interactions. We focus on rigid stick-like hand tools (e.g. hammer, barbell, spade, scythe) with no articulation and approximate them as 3D line segments. Our key idea is that, when a human joint is in contact with an object, the object can be integrated as a constraint on the movement of the human limb. For example, the hammer in Figure 1 provides a constraint on the relative depth between the person’s two hands. Conversely, 3D positions of the hands in contact with the hammer provide a constraint on the hammer’s depth and 3D rotation. To deal with contact forces, we integrate physics in the estimation by modeling dynamics of the person and the object. Inspired by recent progress in humanoid locomotion research [16], we formulate person-object trajectory estimation as an optimal control problem given the contact state of each human joint. We show that contact states can be automatically recognized from the input video using a deep neural network.

2. Related work

Here we review the key areas of related work in both computer vision and robotics literature.

Single-view 3D pose estimation aims to recover the 3D joint configuration of the person from the input image. Recent human 3D pose estimators either attempt to build a direct mapping from image pixels to the 3D joints of the human body or break down the task into two stages: estimating pixel coordinates of the joints in the input image and then lifting the 2D skeleton to 3D. Existing direct approaches either rely on generative models to search the state space for a plausible 3D skeleton that aligns with the image evidence [59, 26, 25] or, more recently, extract deep features from images and learn a discriminative regressor from the 2D image to the 3D pose [37, 49, 53, 62]. Building on the recent progress in 2D human pose estimation [51, 50, 34, 14], two-stage methods have been shown to be very effective [5, 71, 9, 19] and achieve state-of-the-art results [47] on 3D human pose benchmarks [35]. To deal with depth ambiguities, these estimators rely on good pose priors, which are either hand-crafted or learnt from large-scale MoCap data [71, 9, 37]. However, unlike our work, these methods do not consider explicit models for 3D person-object interactions with contacts.

Understanding human-object interactions involves both recognition of actions and modeling of interactions. In action recognition, most existing approaches that model human-object interactions do not consider 3D, instead model interactions and contacts in the 2D image space [29, 20, 70, 55]. Recent works in scene understanding [36, 24] consider interactions in 3D but have focused on static scene elements rather than manipulated objects as we do in this work. Tracking 3D poses of people interacting with the environment has been demonstrated for bipedal walk-
Estimation stage

1. We give details of the formulation as an optimal control problem with contact and dynamics constraints. In this paper, trajectories are denoted as underlined variables, e.g., $\tilde{x}$, $u$, or $c$.

2. In the estimation stage, these image measurements are then fused in order to estimate the 3D motion, 3D contacts, and the controlling forces of both the person and the object. The person and object trajectories, contact positions, and contact forces are constrained jointly by our carefully designed contact motion model, force model, and dynamics equations. Finally, the reconstructed object manipulation sequence can be applied to control a humanoid robot via behavior cloning.

3. In the following, we start in Section 4 by describing the estimation stage giving details of the formulation as an optimal control problem. Then, in Section 5 we give details of the recognition stage including 2D human pose estimation, contact recognition, and object 2D endpoint estimation. Finally, we describe results in Section 6.

4. Estimating person-object trajectory under contact and dynamics constraints

We assume that we are provided with a video clip of duration $T$ depicting a human subject manipulating an object. We encode the 3D poses of the human and the object, including joint translations and rotations, in the configuration vectors $q^h$ and $q^o$, for the human and the object respectively. We define a constant set of $K$ contact points between the human body and the object (or the ground plane). Each contact point corresponds to a human joint, and is activated whenever that human joint is recognized as in contact. At each contact point, we define a contact force $f_k$, whose value is non-zero whenever the contact point $k$ is active. The state of the complete dynamical system is then obtained by concatenating the human and the object joint configurations $q$ and velocities $\dot{q}$ as $x := (q^h, q^o, \dot{q}^h, \dot{q}^o)$. Let $\tau^h_k$ be the joint torque vector describing the actuation by human muscles. This is a $n_q = 6$ dimensional vector where $n_q$ is the dimension of the human body configuration vector. We define the control variable $u$ as the combination of the joint torque vector together with contact forces at the $K$ contact joints, $u := (\tau^h_k, f_k, k = 1, ..., K)$. To deal with sliding contacts, we further define a contact state $c$ that consists of the relative positions of all the contact points with respect to the object (or ground) in the 3D space.

Our goal is two-fold. We wish to (i) estimate smooth and consistent human-object and contact trajectories $x$ and $c$, while (ii) recovering the control $u$ which gives rise to the observed motion. This is achieved by jointly optimizing the 3D trajectory $\tilde{x}$, contacts $c$ and control $u$ given the measurements (2D positions of human joints and object endpoints together with contact states of human joints) obtained from the input video. The intuition is that the human and the object’s 3D poses should match their respective projections in the image while their 3D motion is linked together by the recognized contact points and the corresponding contact forces. In detail, we formulate person-object interaction estimation as an optimal control problem with contact and dynamics constraints.
dynamics constraints:

\[
\begin{align*}
\text{minimize} & \quad \sum_{e \in \{h,o\}} \int_0^T l_e(x, u, c) \, dt, \\
\text{subject to} & \quad \kappa(x, c) = 0 \quad (\text{contact motion model}), \\
& \quad \dot{x} = f(x, u, c) \quad (\text{full-body dynamics}), \\
& \quad u \in \mathcal{U} \quad (\text{force model}),
\end{align*}
\]

where \( e \) denotes either ‘h’ (human) or ‘o’ (object), and the constraints (2)-(4) must hold for all \( t \in [0, T] \). The loss function \( l_e \) is a weighted sum of multiple costs capturing (i) the data term measuring the discrepancy between the observed and re-projected 2D joint and object endpoint positions, (ii) the prior on the estimated 3D poses, (iii) the physical plausibility of the motion and (iv) the smoothness of the trajectory. Next, we in turn describe these cost terms as well as the insights leading to their design choices. For simplicity, we ignore the superscript \( e \) when introducing a cost term that exists for both the human \( l^h \) and the object \( l^o \) component of the loss. We describe the individual terms using continuous time notation as used in the overall problem formulation (1). A discrete version of the problem as well as the optimization and implementation details are relegated to Section 4.5.

4.1. Data term: 2D re-projection error

We wish to minimize the re-projection error of the estimated 3D human joints and 3D object endpoints with respect to the 2D measurements obtained in each video frame. In detail, let \( j = 1, \ldots, N \) be human joints or object endpoints and \( \mathcal{P}^D_j \) their 2D position observed in the image. We aim to minimize the following data term

\[
l_{\text{data}} = \sum_j \rho \left( \mathcal{P}^D_j - P_{\text{cam}}(P_j(q)) \right),
\]

where \( P_{\text{cam}} \) is the camera projection matrix and \( P_j \) the 3D position of joint or object endpoint \( j \) induced by the person-object configuration vector \( q \). To deal with outliers, we use the robust Huber loss, denoted by \( \rho \).

4.2. Prior on 3D human poses

A single 2D skeleton can be a projection of multiple 3D poses, many of which are unnatural or impossible exceeding the human joint limits. To resolve this, we incorporate into the human loss function \( l^h \) a pose prior similar to [9]. The pose prior is obtained by fitting the SMPL human model [43] to the CMU MoCap data [1] using MoSh [44] and fitting a Gaussian Mixture Model (GMM) to the resulting SMPL 3D poses. We map our human configuration vector \( q^h \) to a SMPL pose vector \( \theta \) and compute the likelihood under the pre-trained GMM

\[
l^h_{\text{pose}} = -\log \left( p(q^h; \text{GMM}) \right). \tag{6}
\]

During optimization, \( l^h_{\text{pose}} \) is minimized in order to favor more plausible human poses against rare or impossible ones.

4.3. Physical plausibility of the motion

Human-object interactions involve contacts coupled with interaction forces, which are not included in the data-driven cost terms (5) and (6). Modeling contacts and physics is thus important to reconstruct object manipulation actions from the input video. Next, we outline models for describing the motion of the contacts and the forces at the contact points. Finally, the contact motions and forces, together with the system state \( x \) are linked by the laws of mechanics via the dynamics equations, which constrain the estimated person-object interaction. This full body dynamics constraint is detailed at the end of this subsection.

Contact motions. In the recognition stage, our contact recognizer predicts, given a human joint (for example, left hand, denoted by \( j \)), a sequence of contact states \( \delta_j : t \rightarrow \{0,1\} \). Similarly to [16], we call a contact phase any time segment in which \( j \) is in contact, i.e., \( \delta_j = 1 \). Our key idea is that the 3D distance between human joint \( j \) and the active contact point on the object (denoted by \( k \)) should remain zero during a contact phase:

\[
\|p^h_j(q^h) - p^o_k(x, c)\| = 0 \quad (\text{point contact}), \tag{7}
\]

where \( p^h_j \) and \( p^o_k \) are the 3D positions of joint \( j \) and object contact point \( k \), respectively. Note that position of the object contact point \( p^o_k(x, c) \) depends on the state vector \( x \) describing the human-object configuration and the relative position \( c \) of the contact along the object. The position of contact \( p^o_k \) is subject to a feasible range denoted by \( \mathcal{C} \). For stick-like objects such as hammer, \( \mathcal{C} \) is approximately the 3D line segment representing the handle. For the ground, the feasible range \( \mathcal{C} \) is a 3D plane. In practice, we implement \( p^o_k \in \mathcal{C} \) by putting a constraint on the trajectory of relative contact positions \( c \).

Equation (7) applies to most common cases where the contact area can be modeled as a point. Examples include the hand-handle contact and the knee-ground contact. To model the planar contact between the human sole and ground, we approximate each sole surface as a planar polygon with four vertices, and apply the point contact model at each vertex. In our human model, each sole is attached to its parent ankle joint, and therefore the four vertex contact points of the sole are active when \( \delta_{\text{ankle}} = 1 \).

The resulting overall contact motion function \( \kappa \) in problem (1) is obtained by unifying the point and the planar contact models:

\[
\kappa(x, c) = \sum_j \sum_{k \in \phi(j)} \delta_j \left\| T^{(kj)} \left( p^h_j(q^h) \right) - p^o_k(x, c) \right\|, \tag{8}
\]
where the external sum is over all human joints. The internal sum is over the set of active object contact points mapped to their corresponding human joint \( j \) by mapping \( \phi(j) \). The mapping \( T^{(k)} \) translates the position of an ankle joint \( j \) to its corresponding \( k \)-th sole vertex; it is an identity mapping for non-ankle joints.

**Contact forces.** During a contact phase of the human joint \( j \), the environment exerts a contact force \( f_k \) on each of the active contact points in \( \phi(j) \). \( f_k \) is always expressed in contact point \( k \)'s local coordinate frame. We distinguish two types of contact forces: (i) 6D spatial forces exerted by objects and (ii) 3D linear forces due to ground friction. In the case of object contact, \( f_k \) is an unconstrained 6D spatial force with 3D linear force and 3D moment. In the case of ground friction, \( f_k \) is constrained to lie inside a 3D friction cone \( K^3 \) (also known as the quadratic Lorentz "ice-cream" cone [16]) characterized by a positive friction coefficient \( \mu \). In practice, we approximate \( K^3 \) by a 3D pyramid spanned by a basis of \( N = 4 \) generators, which allows us to represent \( f_k \) as the convex combination \( f_k = \sum_{n=1}^N \lambda_{kn} g_n^{(3)} \), where \( \lambda_{kn} \geq 0 \) and \( g_n^{(3)} \) with \( n = 1, 2, 3, 4 \) are the 3D generators of the contact force. We sum the contact forces induced by the four sole-ground contact points and express a unified contact force in the ankle’s frame:

\[
f_j = \sum_{k=1}^4 \left( \frac{f_k}{p_k \times f_k} \right) = \sum_{k=1}^4 \sum_{n=1}^N \lambda_{jkn} g_n^{(6)} ,
\]

where \( p_k \) is the position of contact point \( k \) expressed in joint \( j \)'s (left/right ankle) frame, \( \times \) is the cross product operator, \( \lambda_{jkn} \geq 0 \), and \( g_n^{(6)} \) are the 6D generators of \( f_j \). Please see the appendix in the extended version of this paper available at [40] for additional details including the expressions of \( g_n^{(3)} \) and \( g_n^{(6)} \).

**Full body dynamics.** The full-body movement of the person and the manipulated object is described by the Lagrange dynamics equation:

\[
M(q)\ddot{q} + b(q, \dot{q}) = \tau + \tau^h,
\]

where \( M \) is the generalized mass matrix, \( b \) covers the centrifugal and Coriolis effects, \( g \) is the generalized gravity vector and \( \tau \) represents the joint torque contributions. \( \ddot{q} \) and \( \dot{q} \) are the joint velocities and joint accelerations, respectively. Note that (10) is a unified equation which applies to both human and object dynamics, hence we drop the superscript \( e \) here. Only the expression of the joint torque \( \tau \) differs between the human and the object and we give the two expressions next.

For human, it is the sum of two contributions: the first one corresponds to the internal joint torques (exerted by the muscles for instance) and the second one comes from the contact forces:

\[
\tau^h = \begin{pmatrix} 0_6 \\ \tau^h_m \end{pmatrix} + \sum_{k=1}^K (J^h_k)^T f_k ,
\]

where \( \tau^h_m \) is the human joint torque exerted by muscles, \( f_k \) is the contact force at contact point \( k \) and \( J^h_k \) is the Jacobian mapping human joint velocities \( q^h \) to the Cartesian velocity of contact point \( k \) expressed in \( k \)'s local frame. Let \( n^h_o \) denote the dimension of \( q^h, \dot{q}^h \) and \( \ddot{q}^h \), then \( \tau^h_m \) and \( J^h_k \) are of dimension \( n^h_o - 6 \) and \( 3 \times n^h_o \), respectively. We model the human body and the object as free-floating base systems. In the case of the human body, the six first entries in the configuration vector \( q \) correspond to the 6D pose of the free-floating base (translation + orientation), which is not actuated by any internal actuators such as human muscles. This constraint is taken into consideration by adding the zeros in Eq. (11).

In the case of the manipulated object, there is no actuation other than the contact forces exerted by the human. Therefore, the object torque is expressed as

\[
\tau^o = - \sum_{k \text{ object contact point}} (J^o_k)^T f_k ,
\]

where the sum is over the object contact points, \( f_k \) is the contact force, and \( J^o_k \) denotes the object Jacobian, which maps from the object joint velocities \( \dot{q}^o \) to the Cartesian velocity of the object contact point \( k \) expressed in \( k \)'s local frame. \( J^o_k \) is a \( 3 \times n^o_o \) matrix where \( n^o_o \) is the dimension of object configuration vectors \( q^o, \dot{q}^o \) and \( \ddot{q}^o \).

We concatenate the dynamics equations of both human and object to form the overall dynamics in Eq. (3) in problem (1), and include a muscle torque term \( \tau^h_{\text{torque}} = \| \tau^h_m \|^2 \) in the overall cost. Minimizing the muscle torque acts as a regularization over the energy consumption of the human body.

### 4.4. Enforcing the trajectory smoothness

**Regularizing human and object motion.** Taking advantage of the temporal continuity of video, we minimize the sum of squared 3D joint velocities and accelerations to improve the smoothness of the person and object motion and to remove incorrect 2D poses. We include the following motion smoothing term to the human and object loss in (1):

\[
l_{\text{smooth}} = \sum_j \left( \| \nu_j(q, \dot{q}) \|^2 + \| \alpha_j(q, \dot{q}, \ddot{q}) \|^2 \right),
\]

where \( \nu_j \) and \( \alpha_j \) are the spatial velocity and the spatial acceleration\(^2\) of joint \( j \), respectively. In the case of object, \( j \) represents an endpoint on the object. By minimizing \( l_{\text{smooth}} \), both the linear and angular movements of each joint/endpoint are smoothed simultaneously.

\(^2\)Spatial velocities (accelerations) are minimal and unified representations of linear and angular velocities (accelerations) of a rigid body [23]. They are of dimension 6.
Regularizing contact motion and forces. In addition to regularizing the motion of the joints, we also regularize the contact states and control by minimizing the velocity of the contact points and the temporal variation of the contact force. This is implemented by including the following contact smoothing term in the cost function in problem (1):

$$I_c^{\text{smooth}} = \sum_j \sum_{k \in \theta(j)} \delta_j \left( \omega_k \| \dot{c}_k \|^2 + \gamma_k \| \dot{f}_k \|^2 \right) dt, \quad (14)$$

where $\dot{c}_k$ and $\dot{f}_k$ represent respectively the temporal variation of the position and the contact force at contact point $k$. $\omega_k$ and $\gamma_k$ are scalar weights of the regularization terms $\dot{c}_k$ and $\dot{f}_k$. Note that some contact points, for example the four contact points of the human sole during the sole-ground contact, should remain fixed with respect to the object or the ground during the contact phase. To tackle this, we adjust $\omega_k$ to prevent contact point $k$ from sliding while being in contact.

4.5. Optimization

Conversion to a numerical optimization problem. We convert the continuous problem (1) into a discrete nonlinear optimization problem using the collocation approach [8]. All trajectories are discretized and constraints (2), (3), (4) are only enforced on the “collocation” nodes of a time grid matching the discrete sequence of video frames. The optimization variables are the sequence of human and object poses $[x_0...x_T]$, torque and force controls $[u_1...u_T]$, contact locations $[c_0...c_T]$, and the scene parameters (ground plane and camera matrix). The resulting problem is nonlinear, constrained and sparse (due to the sequential structure of trajectory optimization). We rely on the Ceres solver [4], which is dedicated to solving sparse estimation problems (e.g. bundle adjustment [65]), and on the Pinocchio software [17, 18] for the efficient computation of kinematic and dynamic quantities and their derivatives [15]. Additional details are given in the appendix available in the extended version of this paper [40].

Initialization. Correctly initializing the solver is key to escape from poor local minima. We warm-start the optimization by inferring the initial configuration vector $q_0$ at each frame using the human body estimator HMR [37] that estimates the 3D joint angles from a single RGB image.

5. Extracting 2D measurements from video

In this section, we describe how 2D measurements are extracted from the input video frames during the first, recognition stage of our system. In particular, we extract the 2D human joint positions, the 2D object endpoint positions and the contact states of human joints.

Estimating 2D positions of human joints. We use the state-of-the-art Openpose [14] human 2D pose estimator, which achieved excellent performance on the MPII Multi-Person benchmark [7]. Taking a pretrained Openpose model, we do a forward pass on the input video in a frame-by-frame manner to obtain an estimate of the 2D trajectory of human joints, $p_{ij}^{2D}$. Recognizing contacts. We wish to recognize and localize contact points between the person and the manipulated object or the ground. This is a challenging task due to the large appearance variation of the contact events in the video. However, we demonstrate here that a good performance can be achieved by training a contact recognition CNN module from manually annotated contact data that combine both still images and videos harvested from the Internet. In detail, the contact recognizer operates on the 2D human joints predicted by Openpose. Given 2D joints at video frame $i$, we crop fixed-size image patches around a set of joints of interest, which may be in contact with an object or ground. Based on the type of human joint, we feed each image patch to the corresponding CNN to predict whether the joint appearing in the patch is in contact or not. The output of the contact recognizer is a sequence $\delta_{ij}$ encoding the contact states of human joint $j$ at video frame $i$, i.e. $\delta_{ij} = 1$ if joint $j$ is in contact at frame $i$ and zero otherwise. Note that $\delta_{ij}$ is the discretized version of the contact state trajectory $\tilde{\delta}_j$ presented in Sec. 4.

Our contact recognition CNNs are built by replacing the last layer of an ImageNet pre-trained Resnet model [31] with a fully connected layer that has a binary output. We have trained separate models for five types of joints: hands, knees, foot soles, toes, and neck. To construct the training data, we collect still images of people manipulating tools using Google image search. We also collect short video clips of people manipulating tools from Youtube in order to also have non-contact examples. We run Openpose pose estimator on this data, crop patches around the 2D joints, and annotate the resulting dataset with contact states.

Estimating 2D object pose. The objective is to estimate the 2D position of the manipulated object in each video frame. To achieve this, we build on instance segmentation obtained by Mask R-CNN [30]. We train the network on shapes of object models from different viewpoints and apply the trained network on the test videos. The output masks and bounding boxes are used to estimate object endpoints in each frame. The resulting 2D endpoints are used as input to the trajectory optimizer. Details are given next.
mented by 2D geometric transformations (translation, rotation, scale) to handle the changes in shapes of tool instances in the videos. In addition, domain randomization [42, 63] is applied to handle the variance of instances and changes in appearance in the videos caused by illumination: the geometrically transformed shape is filled with pixels from a random image (foreground) and pasted on another random image (background). We utilized random images from the MS COCO dataset [41] for this purpose. We use a Mask R-CNN (implementation [3]) model pre-trained on the MS COCO dataset and re-train the head layers for each tool.

At test time, masks and bounding boxes obtained by the re-trained Mask R-CNN are used to estimate the coordinates of tool endpoints. Proximity to coordinates of estimated wrist joints is used to select the mask and bounding box in case multiple candidates are available in the frame. To estimate the main axis of the object, a line is fitted through the output binary mask. The endpoints are calculated as the intersection of the fitted line and boundaries of the bounding box. Using the combination of the output mask and the bounding box compensates for errors in the segmentation mask caused by occlusions. The relative orientation of the tool (i.e. the head vs. the handle of the tool) is determined by spatial location of endpoints in the video frames as well as by their proximity to the estimated wrist joints.

6. Experiments

In this section we present quantitative and qualitative evaluation of the reconstructed 3D person-object interactions. Since we recover not only human poses but also object poses and contact forces, evaluating our results is difficult due to the lack of ground truth forces and 3D object poses in standard 3D pose benchmarks such as [35]. Consequently, we evaluate our motion and force estimation quantitatively on a recent Biomechanics video/MoCap dataset capturing challenging dynamic parkour motions [45]. In addition, we report joint errors and show qualitative results on our newly collected dataset of videos depicting handtool manipulation actions.

6.1. Parkour dataset

This dataset contains videos capturing human subjects performing four typical parkour actions: two-hand jump, moving-up, pull-up and a single-hand hop. These are highly dynamic motions with rich contact interactions with the environment. The ground truth 3D motion and contact forces are captured with a Vicon motion capture system and several force plates. The 3D motion and forces are reconstructed with frame rates of 400Hz and 2200Hz, respectively, whereas the RGB videos are captured in a relatively lower rate of 25Hz, making this dataset a challenge for computer vision algorithms due to motion blur.

<table>
<thead>
<tr>
<th>Method</th>
<th>Jump</th>
<th>Move-up</th>
<th>Pull-up</th>
<th>Hop</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMPLify [9]</td>
<td>121.75</td>
<td>147.41</td>
<td>120.48</td>
<td>169.36</td>
<td>139.69</td>
</tr>
<tr>
<td>HMR [37]</td>
<td>111.36</td>
<td>140.16</td>
<td>132.44</td>
<td>149.64</td>
<td>135.65</td>
</tr>
<tr>
<td>Ours</td>
<td>98.42</td>
<td>125.21</td>
<td>119.92</td>
<td>138.45</td>
<td>122.11</td>
</tr>
</tbody>
</table>

Table 1: Mean per joint position error (in mm) of the recovered 3D motion for each action on the Parkour dataset.

<table>
<thead>
<tr>
<th>L. Sole</th>
<th>R. Sole</th>
<th>L. Hand</th>
<th>R. Hand</th>
</tr>
</thead>
<tbody>
<tr>
<td>Force (N)</td>
<td>144.23</td>
<td>138.21</td>
<td>107.91</td>
</tr>
<tr>
<td>Moment (N·m)</td>
<td>23.71</td>
<td>22.32</td>
<td>131.13</td>
</tr>
</tbody>
</table>

Table 2: Estimation errors of the contact forces exerted on soles and hands on the Parkour dataset.

Evaluation set-up. We evaluate both the estimated human 3D motion and the contact forces. For evaluating the accuracy of the recovered 3D human poses, we follow the common approach of computing the mean per joint position error (MPJPE) of the estimated 3D pose with respect to the ground truth after rigid alignment [27]. We evaluate contact forces without any alignment: we express both the estimated and the ground truth 6D forces at the position of the contact aligned with the world coordinate frame provided in the dataset. We split the 6D forces into linear and moment components and report the average Euclidean distance of the linear force and the moment with respect to the ground truth.

Results. We report joint errors for different actions in Table 1 and compare the results with the HMR 3D human pose estimator [37], which is used to warm-start our method. To make it a fair comparison, we use the same Openpose 2D joints as input. In addition, we evaluate the recent SMPLify [9] 3D pose estimation method. Our method outperforms both baselines by more than 10mm on average on this challenging data. Finally, Table 2 summarizes the force estimation results. To estimate the forces we assume a generic human physical model of mass 74.6 kg for all the subjects. Despite the systematic error due to the generic human mass assumption, the results in Table 2 validate the quality of our force estimation at the soles and the hands during walking and jumping. We observe higher errors of the estimated moments at hands, which we believe is due to the challenging nature of the Parkour sequences where the entire person’s body is often supported by hands. In this case, the hand may exert significant force and torque to support the body, and a minor shift in the force direction may lead to significant errors.

6.2. Handtool dataset

In addition to the Parkour data captured in a controlled set-up, we would like to demonstrate generalization of our approach to the “in the wild” Internet instructional videos. For this purpose, we have collected a new dataset of object manipulation videos, which we refer to as the Handtool dataset. The dataset contains videos of people manipulating four types of tools: barbell, hammer, scythe, and spade. For each type of tool, we chose among the top videos re-
Figure 3: Example qualitative results on the Handtool dataset. Each example shows the input frame (left) and two different views of the output 3D pose of the person and the object (middle, right). The yellow and the white arrows in the output show the contact forces and moments, respectively. Note how the proposed approach recovers from these challenging unconstrained videos the 3D configuration of the person-object interaction together with the contact forces and moments.

For additional video results please see the project webpage [2].

We then cropped short clips from each video showing the whole human body and the tool.

**Evaluation of 3D human poses.** For each video in the Handtool dataset, we have annotated the 3D positions of the person’s left and right shoulders, elbows, wrist, hips, knees, and ankles, for the first, the middle, and the last frame. We evaluate the accuracy of the recovered 3D human poses by computing their MPJPE after rigid alignment. Quantitative evaluation of the recovered 3D poses is shown table 3. On average, our method outperforms the strong HMR [37] and SMPLify [9] baselines. However, the differences between the methods are reaching the limits of the accuracy of the manually provided 3D human pose annotations on this dataset. Videos available at project website [2] demonstrate that our model produces smooth 3D motion, respecting person-object contacts and capturing the interaction of the person with the tool. This is not the case for the HMR [37] and SMPLify [9] baselines that are applied to individual frames and do not model the interaction between the person and the tool. Example results for our method are shown in Figure 3. For additional results including examples of the main failure modes please see the appendix in the extended version of this paper available at [40].

<table>
<thead>
<tr>
<th>Method</th>
<th>Barbell</th>
<th>Spade</th>
<th>Hammer</th>
<th>Scythe</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMPLify [9]</td>
<td>130.69</td>
<td>135.03</td>
<td>93.43</td>
<td>112.93</td>
<td>118.02</td>
</tr>
<tr>
<td>HMR [37]</td>
<td>105.04</td>
<td>97.18</td>
<td>96.34</td>
<td>115.42</td>
<td>103.49</td>
</tr>
<tr>
<td>Ours</td>
<td>104.23</td>
<td>95.21</td>
<td>95.87</td>
<td>114.22</td>
<td>102.02</td>
</tr>
</tbody>
</table>

Table 3: Mean per joint position error (in mm) of the recovered 3D human poses for each tool type on the Handtool dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>Barbell</th>
<th>Spade</th>
<th>Hammer</th>
<th>Scythe</th>
<th>25</th>
<th>50</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mask R-CNN</td>
<td>33/42/54</td>
<td>54/79/93</td>
<td>35/44/45</td>
<td>63/72/76</td>
<td>69/88/98</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4: The percentage of endpoints for which the estimated 2D location lies within 25/50/100 pixels (in 600×400 pixel image) from the manually annotated ground truth location.

Evaluation of 2D object poses. To evaluate the quality of estimated object poses, we manually annotated 2D object endpoints in every 5th frame of each video in the Handtool dataset and calculated the 2D Euclidean distance (in pixels) between each manually annotated endpoint and its estimated 2D location provided by our method. The 2D location is obtained by projecting the estimated 3D tool position back to the image plane. We compare our results to the output of the Mask R-CNN instance segmentation baseline [30] (which provides initialization for our person-object interaction model). In Table 4 we report for both methods the percentage of endpoints for which the estimated endpoint location lies within 25, 50, and 100 pixels from the anno-
References


