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

Authors: Zongmian Li, Jiri Sedlar, Justin Carpentier, Ivan Laptev, Nicolas Mansard, Josef Sivic (CVPR-2019)

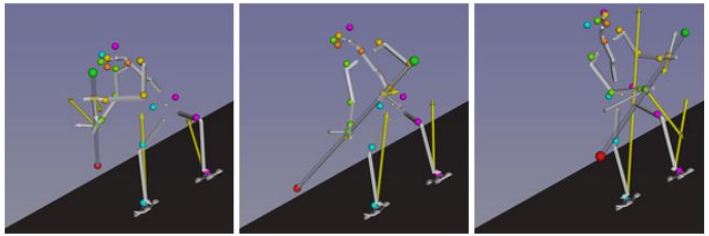
Sahana Eshwaran, Alicia Alarie, Simba Nyatsanga

UCDAVIS
UNIVERSITY OF CALIFORNIA

Problem Statement

Input: A monocular RGB video

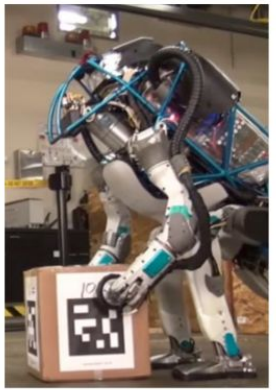
Output 3D Motion and Contact Forces



Given an input video, reconstruct 3D motion and forces of a person interacting with an object from a single RGB video

Motivation

- Large-scale learning of human-object interactions from internet videos, speed up the pace of learning
- Advanced visual intelligence capabilities to recognize and interpret complex person-to-object interactions



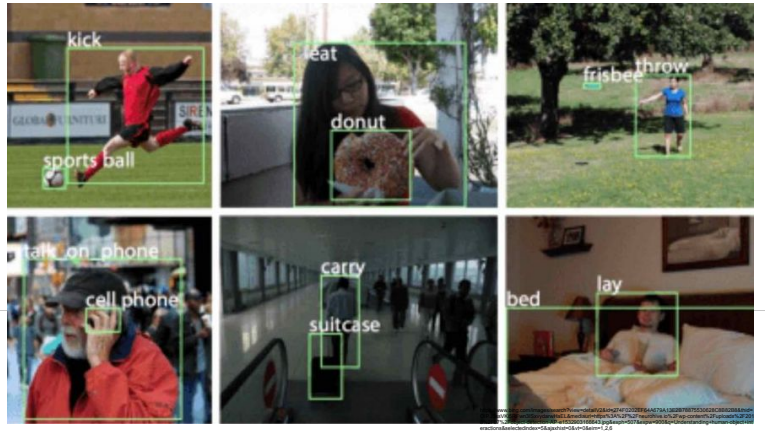
"Atlas" by Boston Dynamics

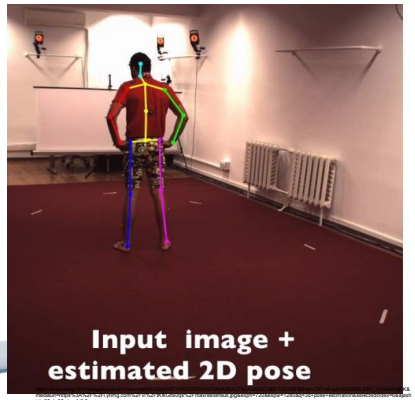
UCDAVIS

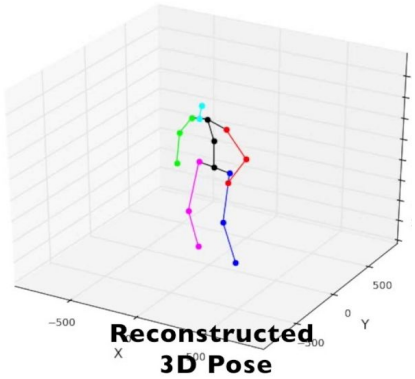
https://www.di.ens.fr/willow/research/motionforcesfromvideo/research/li19mfv/li2019motionforces_slides.pdf

Related Work

- Single-View 3D pose estimation
- Human- Object Interactions
- Object 3D pose estimation
- Learning from instructional videos





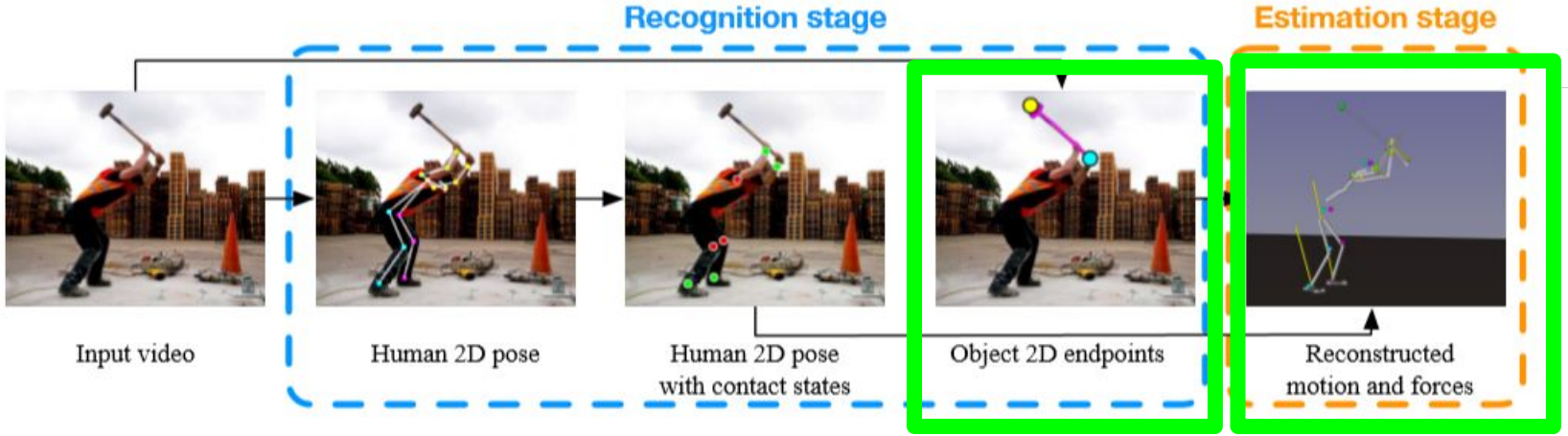


UCDAVIS

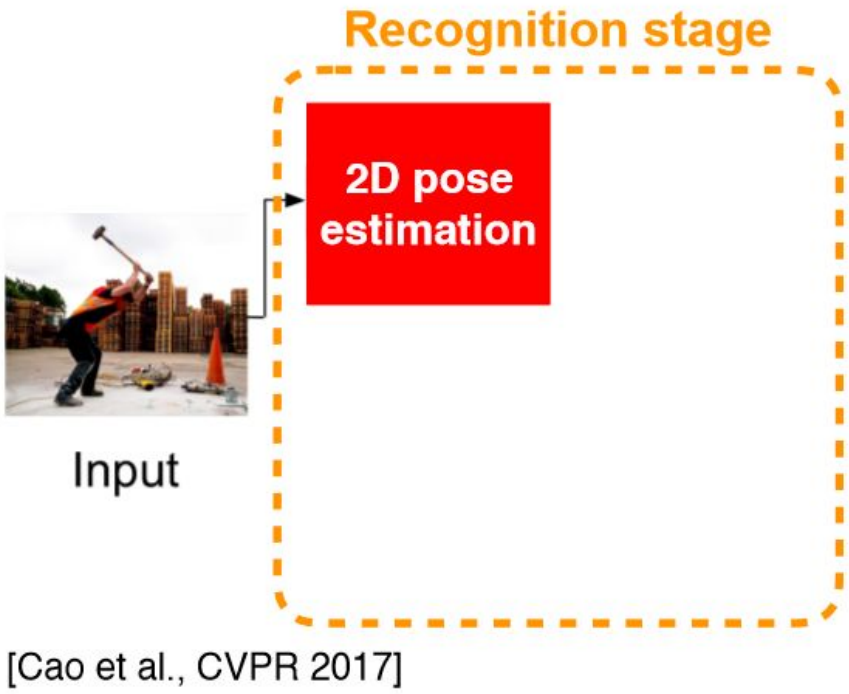
Work borrowed from Person-object interactions in robotics.

- They use the concept of rigid body models introduced in robotics and estimate the 3D person-object interactions from monocular videos using optimal control problem under contact constraints.
- Then identify the contact irregularities by identifying contact states from visual input & localize the contact points in 3D via trajectory estimator.

Key Contributions



Method: A two-stage approach

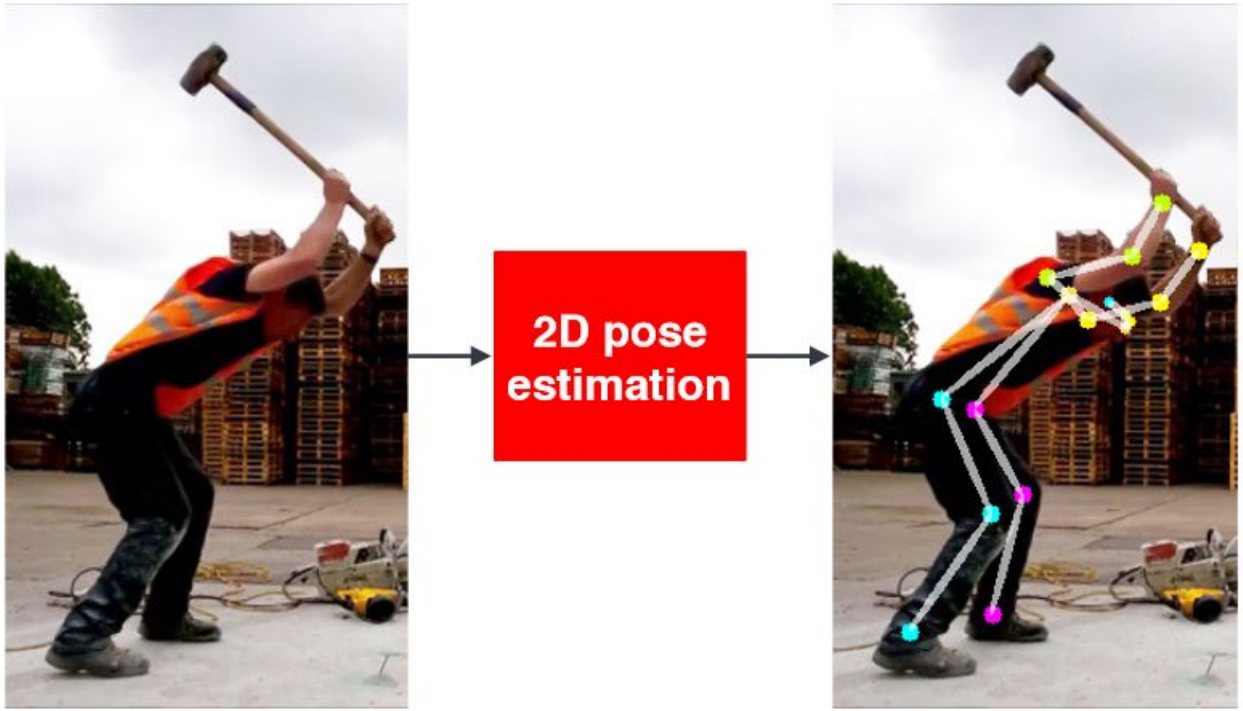


[Cao et al., CVPR 2017]

UCDAVIS

https://www.di.ens.fr/willow/research/motionforcesfromvideo/research/li19mfv/li2019motionforces_slides.pdf

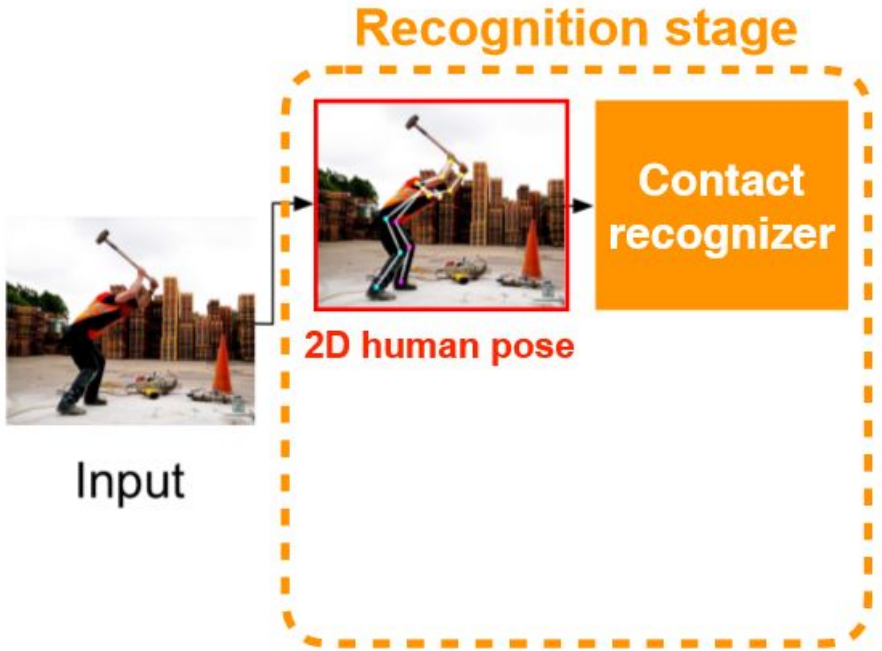
Method: A two-stage approach



UCDAVIS

https://www.di.ens.fr/willow/research/motionforcesfromvideo/research/li19mfv/li2019motionforces_slides.pdf

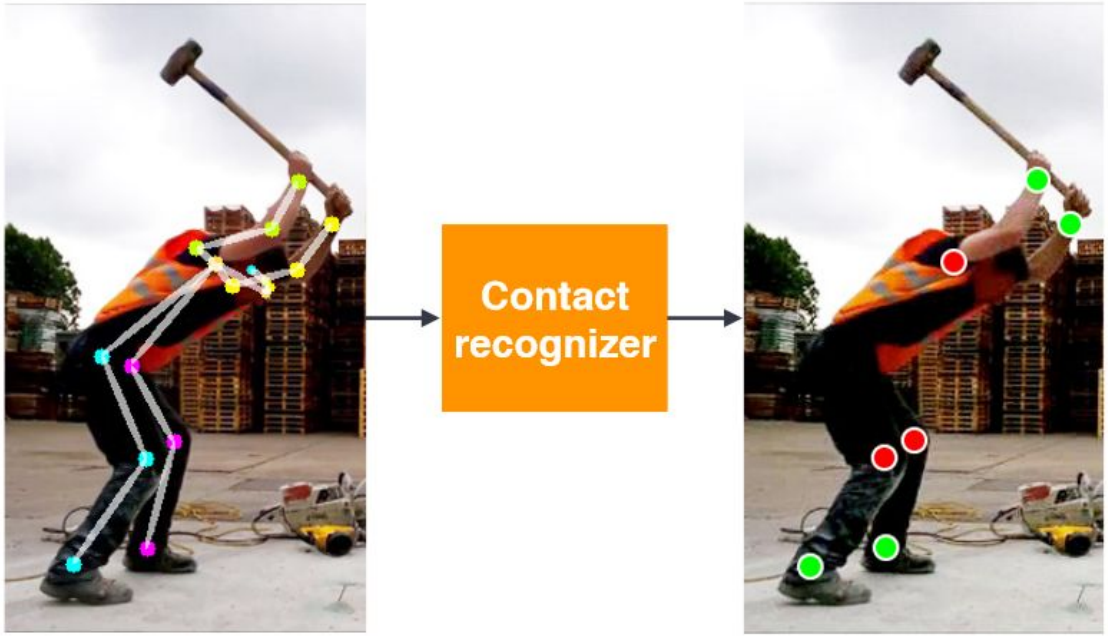
Method: A two-stage approach



[He et al., CVPR 2016]

https://www.di.ens.fr/willow/research/motionforcesfromvideo/research/li19mfv/li2019motionforces_slides.pdf

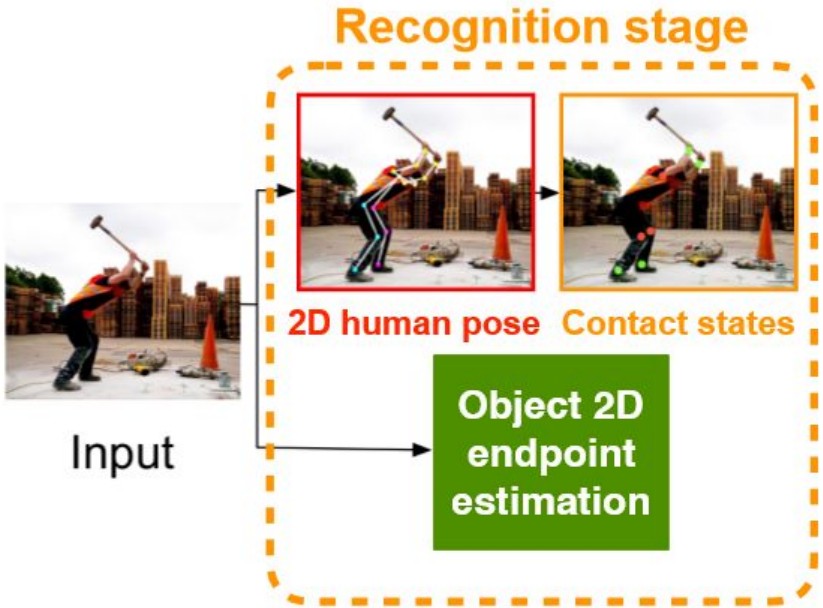
Method: A two-stage approach



UCDAVIS

https://www.di.ens.fr/willow/research/motionforcesfromvideo/research/li19mfv/li2019motionforces_slides.pdf

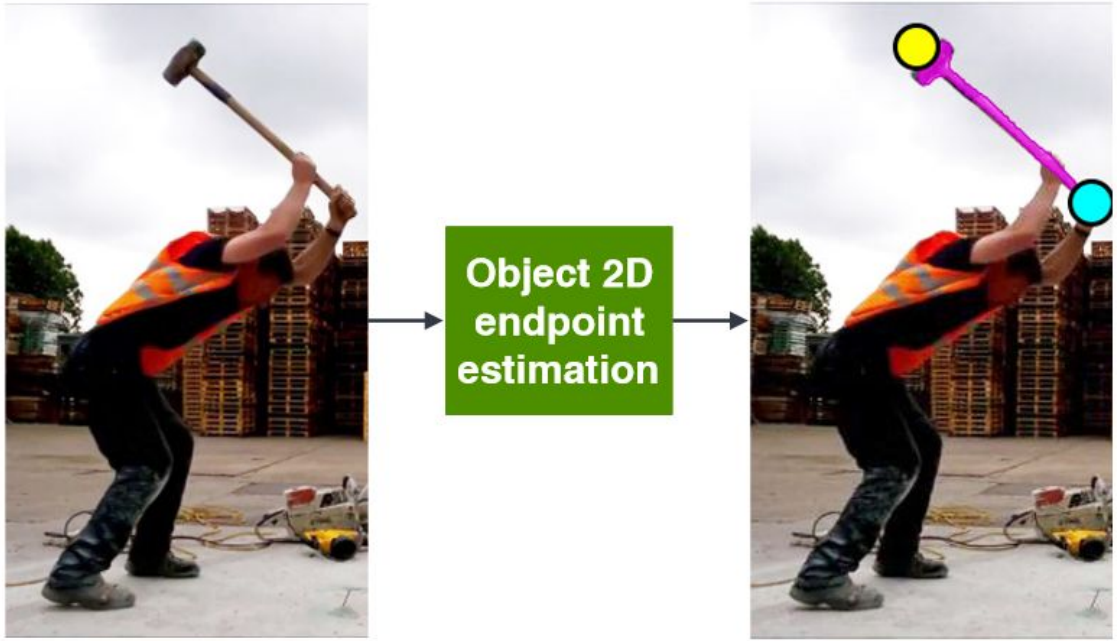
Method: a two-stage approach



[He et al., ICCV 2017]

https://www.di.ens.fr/willow/research/motionforcesfromvideo/research/li19mfv/li2019motionforces_slides.pdf

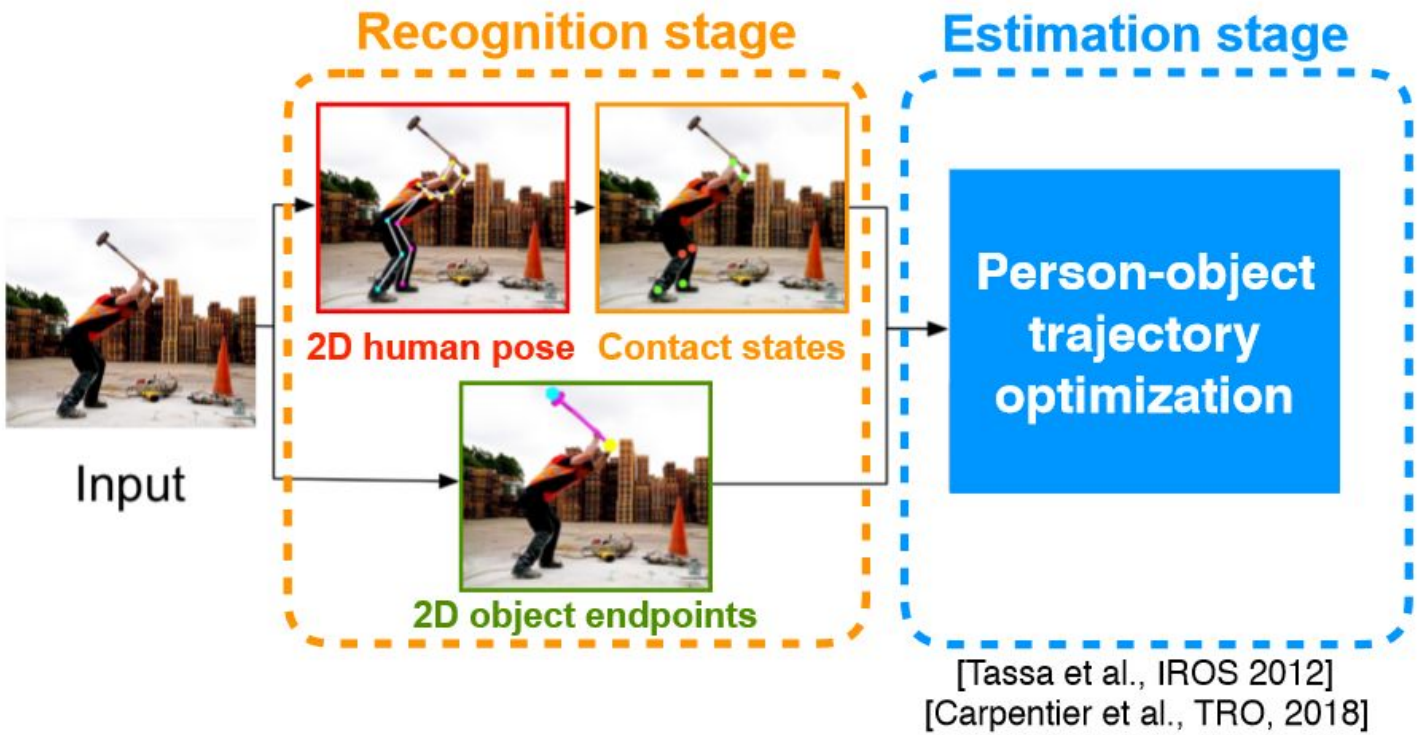
Method: a two-stage approach



UCDAVIS

https://www.di.ens.fr/willow/research/motionforcesfromvideo/research/li19mfv/li2019motionforces_slides.pdf

Method: a two-stage approach



https://www.di.ens.fr/willow/research/motionforcesfromvideo/research/li19mfv/li2019motionforces_slides.pdf

Estimation Stage

- Assume a video clip of duration T depicting a human manipulating an object.
- Encode the 3D poses of the human and object using configuration vectors (translation and rotational displacement).

$$x = (q^h, q^o, \dot{q}^h, \dot{q}^o)$$

- K is constant set of possible contact points between human, object and/or ground plane.

Estimation Stage

- Model the contact force exerted at a contact point (e.g. grip, Newton's 3rd law)

$$f_k \; where \; k = 1, \; \ldots, K$$

- Model the joint torque vector describing actuation by human muscles

$$\tau_m^h$$

- Control variable combining joint torque vector and contact forces at the K contact joints

$$u \; = \; \left(\tau_m^h \, , \; f_k \, , \; k = 1, \; \ldots, K\right)$$

Estimation Stage

- For sliding contacts, define a constant c consisting of relative positions of all contact points w.r.t the object/ground
- **Objective 1** - estimate smooth and consistent human-object and contact trajectories: $\underline{x}$ $\underline{c}$
- **Objective 2** - recover the control value $\underline{u}$ that causes the observed motion
- **Intuition** - human and object 3D poses = respective projections in the image

Estimation Stage

- Formulate person-object interaction estimation as a optimal control problem with contact and dynamic constraints:

$$\underset{\underline{x},\,\underline{u},\,\underline{c}}{\text{minimise}} \sum_{e \in \{h,\,o\}} \int_0^T l^e\,(x,\,u,\,c)\,dt$$

- Constrained by

$$\kappa(x,\,c) = 0$$

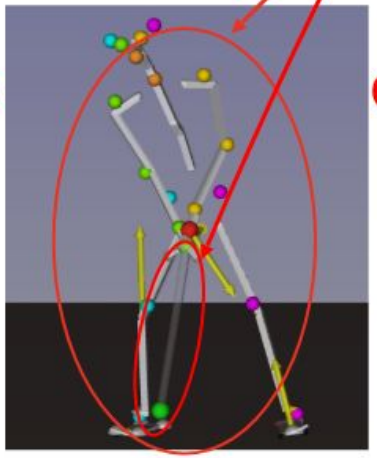
$$\dot{x} = f(x,\,c,\,u)$$

$$u \in \mathcal{U}$$

Estimation Stage

$$\underset{\boldsymbol{x},\underline{u},\underline{c}}{\text{minimize}} \quad \sum_{e\in\{\text{h,o}\}} \int_0^T l^e\left(\boldsymbol{x}, u, c\right) \mathrm{d}t$$

Person 3D poses



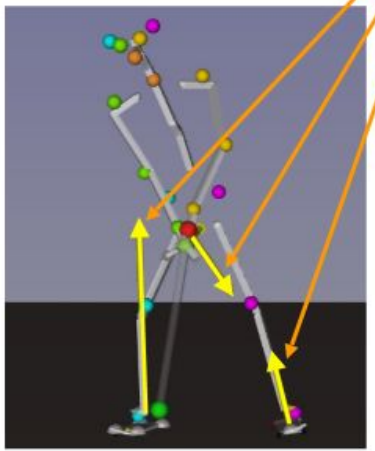
Object 3D poses

UCDAVIS

https://www.di.ens.fr/willow/research/motionforcesfromvideo/research/li19mfv/li2019motionforces_slides.pdf

Estimation Stage

$$\underset{\underline{x},\boldsymbol{u},\underline{c}}{\text{minimize}} \quad \sum_{e\in\{\text{h,o}\}} \int_0^T l^e\left(x, \boldsymbol{u}, c\right) \mathrm{d}t$$



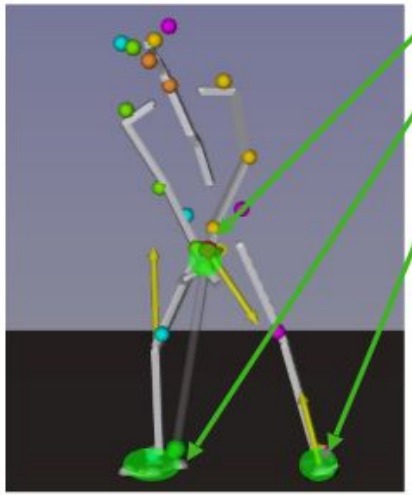
Person-object
person-ground
contact forces

UCDAVIS

https://www.di.ens.fr/willow/research/motionforcesfromvideo/research/li19mfv/li2019motionforces_slides.pdf

Estimation Stage

$$\underset{\underline{x},\underline{u},\boldsymbol{c}}{\text{minimize}} \quad \sum_{e\in\{\text{h,o}\}} \int_0^T l^e\left(x, u, \boldsymbol{c}\right) \mathrm{d}t$$



Contact positions

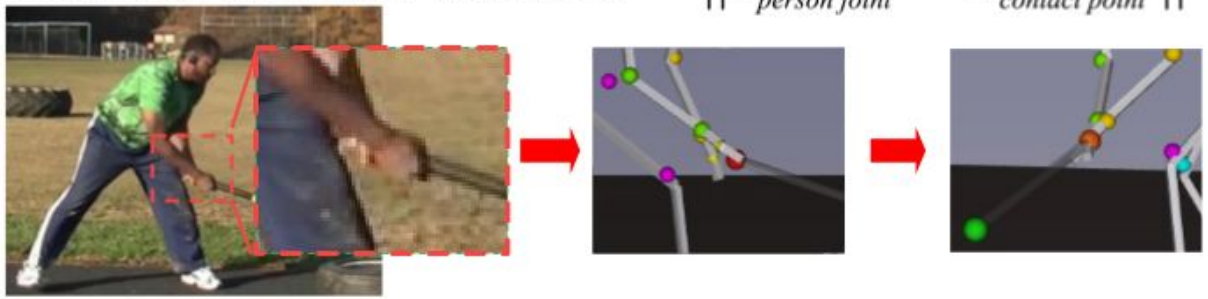
https://www.di.ens.fr/willow/research/motionforcesfromvideo/research/li19mfv/li2019motionforces_slides.pdf

Estimation Stage

$$\underset{\underline{x},\underline{u},\underline{c}}{\text{minimize}} \quad \sum_{e\in\{\text{h,o}\}} \int_0^T l^e\left(x,u,c\right)\mathrm{d}t$$

Subject to

1. Contact motion model: $\left\| p_{\textit{person joint}} - p_{\textit{contact point}} \right\| = 0$



UCDAVIS

https://www.di.ens.fr/willow/research/motionforcesfromvideo/research/li19mfv/li2019motionforces_slides.pdf

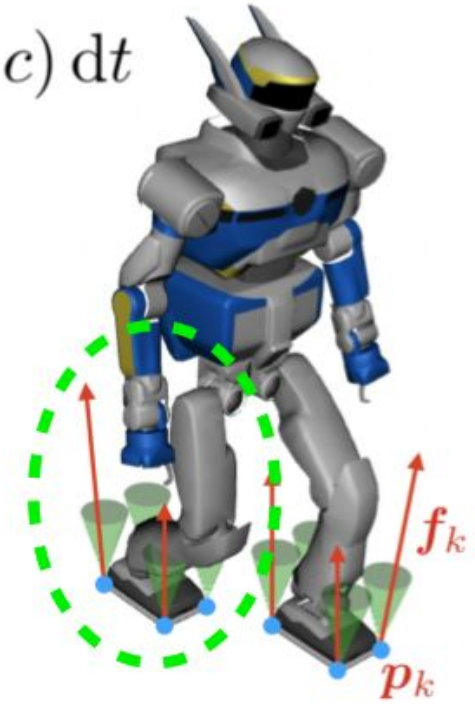
Estimation Stage

$$\underset{\underline{x},\underline{u},\underline{c}}{\text{minimize}} \quad \sum_{e\in\{\text{h,o}\}} \int_0^T l^e\left(x,u,c\right)\mathrm{d}t$$

Subject to:

2. Contact force constraints:

- Prevent the feet from sliding



https://www.di.ens.fr/willow/research/motionforcesfromvideo/research/li19mfv/li2019motionforces_slides.pdf

Estimation Stage

$$\underset{\underline{x},\underline{u},\underline{c}}{\text{minimize}} \quad \sum_{e\in\{\text{h,o}\}} \int_0^T l^e\left(x,u,c\right)\mathrm{d}t$$

Subject to:

3. Lagrangian dynamics equation

$$M(x)\ddot{x} + b(x,\dot{x}) = \tau(u,c)$$

[Carpentier, et al. Pinocchio. https://stack-of-tasks.github.io/pinocchio. 2015-2019]

Composite Loss Function

- Loss functio l^e = weighted sum of multiple costs
- **Data term** - observed vs re-projected 2D joint and obj endpoint positions.
- **Pose prior term** - encourage only physically plausible poses.
- **Physical plausibility term** - contact motions, contact forces, full body dynamics.
- **Trajectory smoothness term**

Data Term

- Minimise the re-projection error of the estimated 3D human joints and 3D objection endpoints w.r.t 2D measurement in the video frame.
- Given **j = 1, ..., N** joints or object endpoints

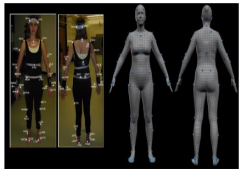
$$l_{data} = \sum_{j} \rho \left(p_j^{2D} - P_{cam}(p_j(q)) \right)$$

Pose Prior Term

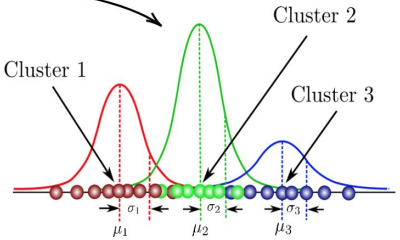
- **Motivation** - Single 2D skeleton = projection of multiple 3D poses; some unnatural or impossible exceeding human joint limits.
- **Solution** - Limit to only plausible pose priors.

Prior Pose Term (Obtaining pose priors)

$$MoSh = \quad = SMPL(MoSh(\quad))$$



CMU MoCap



CMU Mocap data, http : // mocap. cs. cmu. edu /

Loper et. al "MoSh : Motion and Shape Capture from Sparse Markers" SIGGRAPH Asia 2014"

Bogo et. al, "Keep it SMPL : Automatic Estimation of 3DHuman Pose and Shape from a Single Image, ECCV 2016"

UCDAVIS

Pose Prior Term

- Map q^h to SMPL pose vector and compute likelihood under pre-trained GMM
- Minimise to favor more plausible poses vs impossible ones

$$l^h_{pose} = -log\left(p\left(q^h; GMM\right)\right)$$

Physical Plausibility Term

- **Contact motions** - minimise distance between active joints and contact points on object.
- **Contact forces** - model constraints induced by physical forces (friction, Newton's 3rd Law).
- **Full body dynamics** - model trajectory of body-object system using Lagrangian mechanics.

Contact Motions

- For an active joint minimise distance between joint and active contact point

$$\left\| p_j^{\mathrm{h}}(q^{\mathrm{h}}) - p_k^{\mathrm{c}}(x, c) \right\| = 0$$

- For planar contact motion

$$\kappa(x, c) = \sum_j \sum_{k \in \phi(j)} \delta_j \left\| T^{(kj)} \left(p_j^{\mathrm{h}}(q^{\mathrm{h}}) \right) - p_k^{\mathrm{c}}(x, c) \right\|$$

Contact Forces

- Environment exerts a contact force f_k on each active contact point.
- Two distinct contact forces.
- 6D spatial forces exerted by objects - 3D linear force and 3D moment.
- 3D linear forces due to friction - constrained inside 3D “cone of friction” approximated as 3D pyramid.

$$f_k = \sum_{n=1}^{N} \lambda_{kn} g_n^{(3)} \qquad f_j = \sum_{k=1}^{4} \begin{pmatrix} f_k \\ p_k \times f_k \end{pmatrix} = \sum_{k=1}^{4} \sum_{n=1}^{N} \lambda_{jkn} g_{kn}^{(6)}$$

Full Body Dynamics

- Fully body movements of person and manipulated object described by Lagrangian dynamics equation.

$$M(q)\ddot{q} + b(q, \dot{q}) = g(q) + \tau$$

Trajectory Smoothness Term

- Regularise human and object motion.
- Leverage temporal continuity in the video to improve smoothness of human and object motion.
- Minimise spatial velocities and accelerations for smoothness and removing incorrect 2D poses.

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

- Both linear and angular movements are smoothed simultaneously.

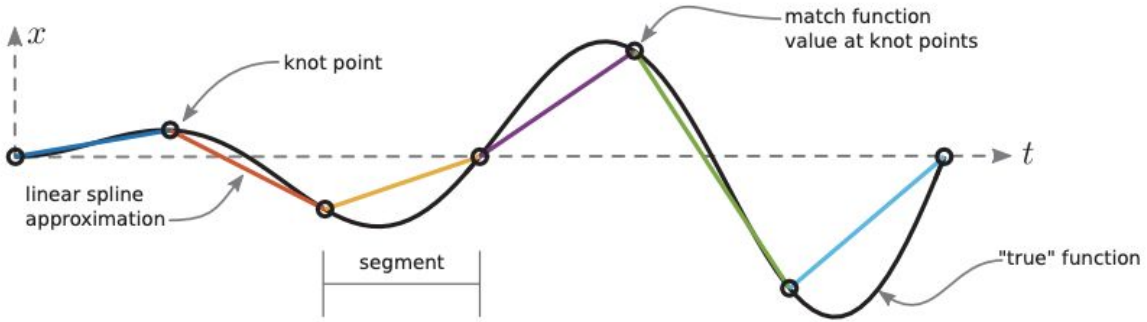
Trajectory Smoothness Term

- Regularise contact motion and forces.
- Minimise the velocity of the contact points and temporal variation of contact force.

$$l^{\mathrm{c}}_{\mathrm{smooth}} = \sum_{j} \sum_{k \in \phi(j)} \delta_j \left(\omega_k \|\dot{c}_k\|^2 + \gamma_k \|\dot{f}_k\|^2 \right) \mathrm{d}t$$

Optimisation

- Convert continuous problem to discrete non-linear optimisation problem using collocation method.
- Discretise all trajectories - $\underline{x}\ \underline{c}\ \underline{u}$

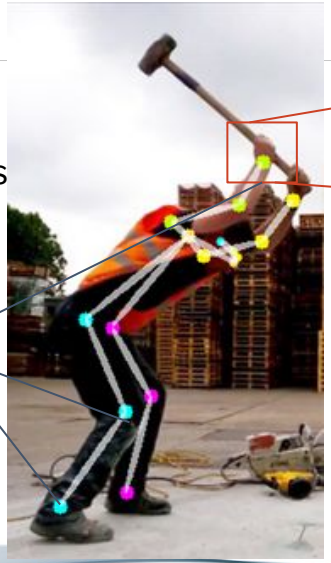


Extracting 2D Measurements from Video

- **Estimating 2D position of joints**: OpenPose [Zhe Cao et al., Berkeley AI Research]

Extracting 2D Measurements: Contact Points

- Training dataset: manually annotated contact data (Google Image Search & youtube videos) harvested from the Internet



separate models for five types of joints: hands, knees, foot soles, toes, and neck

Image patch

Contact Recognition CNN

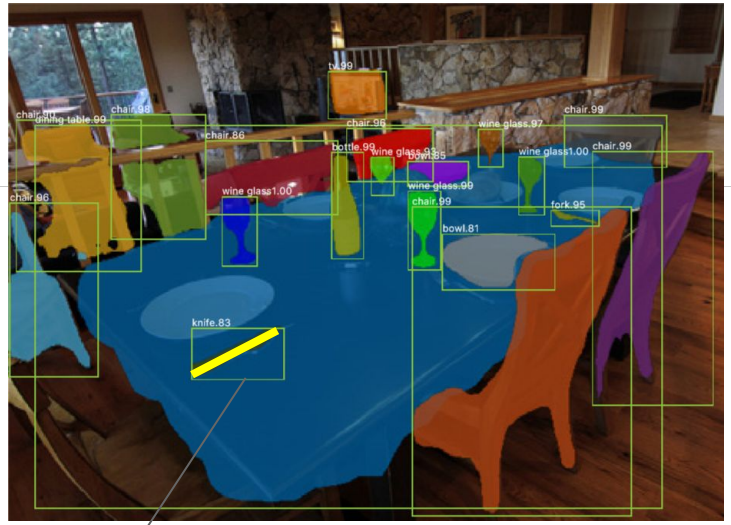
Replaces the last layer of an ImageNet pre-trained Resnet model with a fully connected layer that has a binary output

$$\delta_{ji} = 1$$

Predictions for contact at joint j, frame i (possible states 0 or 1)

Extracting 2D Measurements: Object Pose

- Mask R-CNN
  - Pre-trained on MS COCO dataset
  - Retrained with head layers for each tool type.
  - Shapes of 3D models or 2D annotated images of each specific tool (spade, hammer etc) are used as a training dataset with 2D geometric transformations & domain randomization (illumination)



The endpoints are calculated as the intersection of the fitted line and boundaries of the bounding box

UCDAVIS

Experimental Results

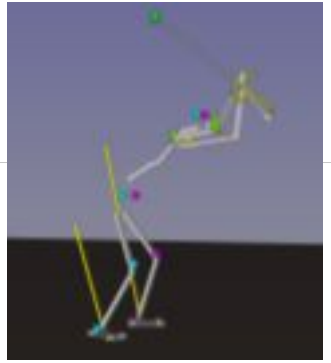
Parkour Dataset

Mean per Joint Position Error (in mm)

Method	Jump	Move-up	Pull-up	Hop	Avg
SMPLify [9]	121.75	147.41	120.48	169.36	139.69
HMR [37]	111.36	140.16	132.44	149.64	135.65
Ours	**98.42**	**125.21**	**119.92**	**138.45**	**122.11**

Improvement over state-of-the-art for Barbell, Spade, and on Average over all tool types.

Parkour Dataset



Errors of the contact forces for soles and hands

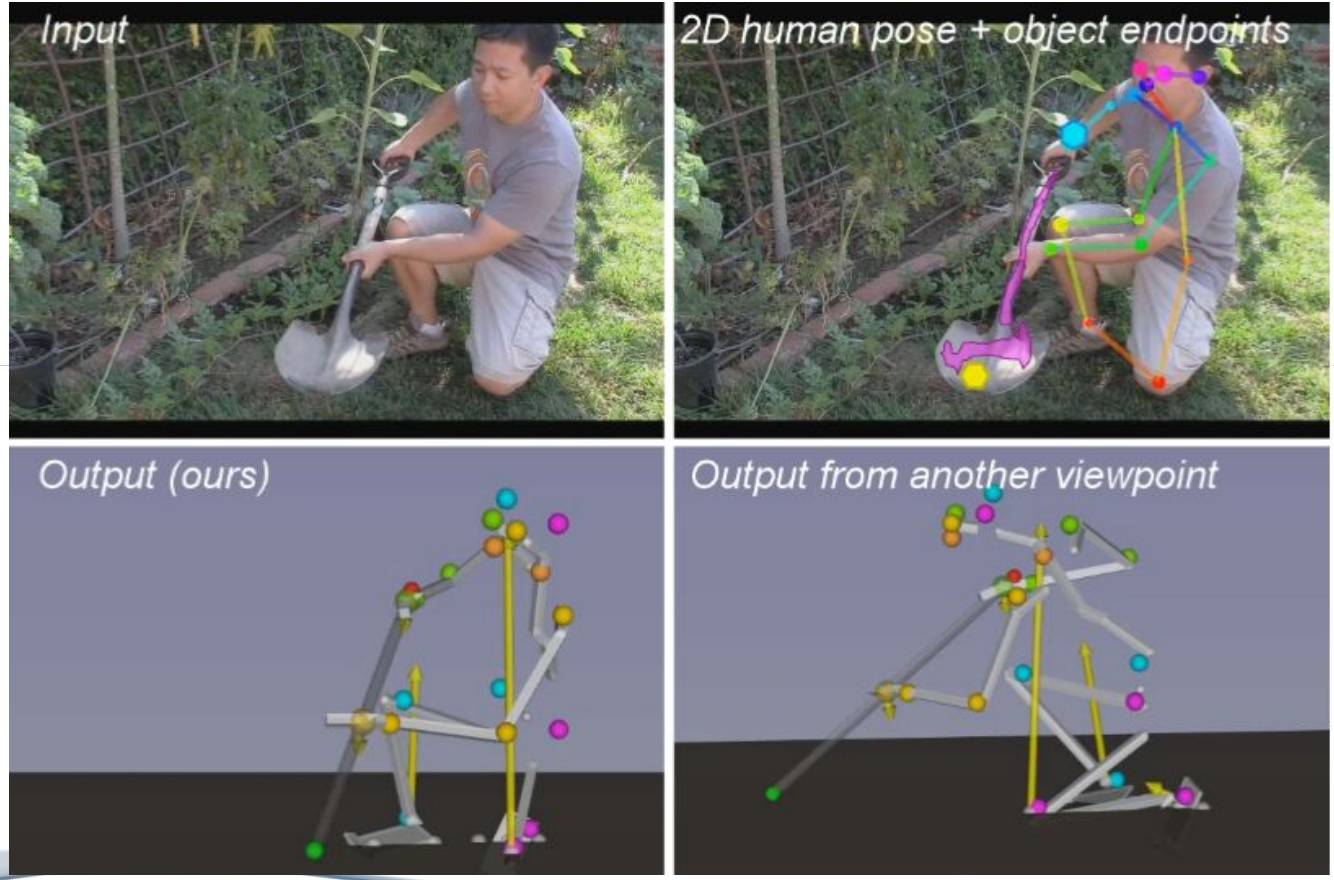
	L. Sole	R. Sole	L. Hand	R. Hand
Force (N)	144.23	138.21	107.91	113.42
Moment (N·m)	23.71	22.32	131.13	134.21

To estimate the forces we assume a generic human physical model of mass 74.6kg for all the subjects

A New Dataset: Handtool Videos

UCDAVIS

https://www.di.ens.fr/willow/research/motionforcesfromvideo/research/li19mfv/li2019motionforces_slides.pdf



UCDAVIS

https://www.di.ens.fr/willow/research/motionforcesfromvideo/research/li19mfv/li2019motionforces_slides.pdf

Results: Video

Video

Results: 3D Pose Estimation on Handtool Dataset

Mean per Joint Position Error (in mm)

Method	Barbell	Spade	Hammer	Scythe	Avg
SMPLify	130.69	135.03	93.43	112.93	118.02
HMR	105.04	97.18	96.34	115.42	103.49
Ours	104.23	95.21	95.87	114.22	102.02

Improvement over state-of-the-art for Barbell, Spade, and on Average over all tool types.

Results: 2D Endpoint Estimation

% of Endpoints within 25/50/100 pixels from ground truth

Method	Barbell	Spade	Hammer	Scythe
Mask-R-CNN	33/42/54	54/79/93	35/44/45	63/72/76
Ours	38/71/98	57/86/99	61/91/99	69/88/98

Improvement over state-of-the-art for all tool types.

Strengths & Weaknesses

- + Inducing physics priors to trajectory estimation.
- + Recovers both human-object interaction and forces that give rise to the motion.
- + Able to generalise to the Parkour dataset videos of 400-2200Hz although trained on ~25Hz videos.
- - More visual illustrations for the physical constraints would have helped.
- - Formulation of 6D force generators is difficult to understand (even in the Appendix).

How can this work be extended?

- Articulating tools.
- Apply to robots.
- Marker-less motion capture data from videos
- What other data-driven/DL problems can benefit for infusion of common sense priors (e.g. physics, optics) as part of the optimisation stage?

TGIF!!:)