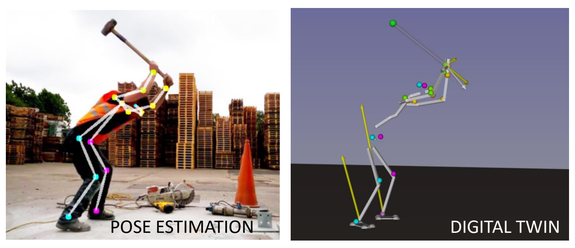
Estimating 3D Motion and Forces of Person-Object Interactions from Monocular Video

Reviewed by Matthieu Dinot, Balthazar Neveu

MATHÉMATIQUES VISION APPRENTISSAGE



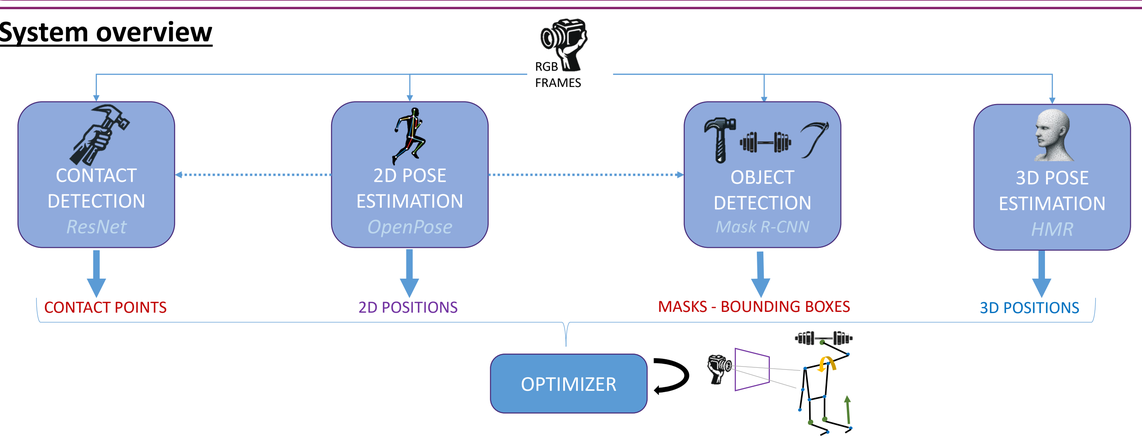
Estimating 3D Motion and Forces of Person-Object Interactions from Monocular Video
Zongmian Li, Jiri Sedlar, Justin Carpentier, Ivan Laptev, Nicolas Mansard and Josef Sivic

Goal : reconstruct a digital twin of a human without a motion capture system from non-stereo videos in the wild

- 3D body motion
- body torques
- contact forces

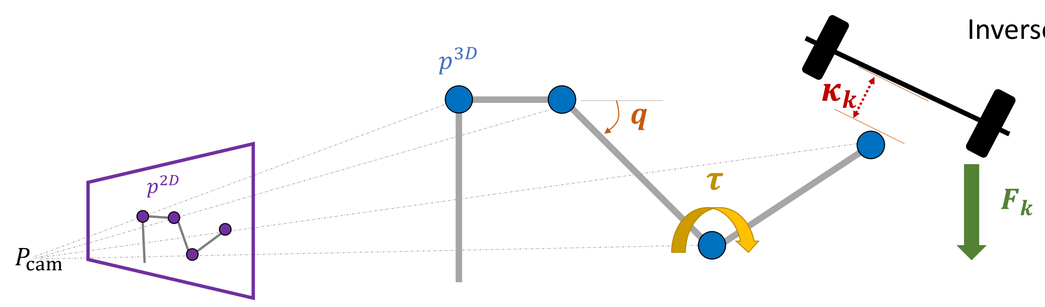
→ in order to do behavior cloning on robots

System overview



Optimizer

Inverse dynamics



$$min_{\underline{q},\underline{\tau},\underline{F_k},\underline{c}} \int_0^T l^{\text{Human}}(q,\tau,F_k,c_k) + l^{\text{Object}}(q,\tau,F_k,c_k)$$

Minimize *costs over States, Torques , Forces, Contacts , Camera pose*

Subject to

(1) Dynamics constraints (q , τ, F_k)
(2) Contact motion model $\kappa(q, c_k)$
(3) Force model $F_k \in \mathcal{F}$

Data fidelity

2D reprojection error

$$l_{2D} = \sum_j \rho\left[p_j^{2D} - P_{\text{cam}}(p_j(q))\right]$$

3D forward kinematics error

$$l_{3D} = \sum_j \rho\left[p_j^{3D} - p_j(q)\right]$$

Motion priors

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

Linear velocity *Linear acceleration*

$$l_{\text{pose}} = -\log(p(q;\text{GMM}))$$

Negative log likelihood of the pose

$$l_{\text{torque}} = \|\tau\|^2$$

Torque norm

Constraints

(1) Lagrange dynamics constraints

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

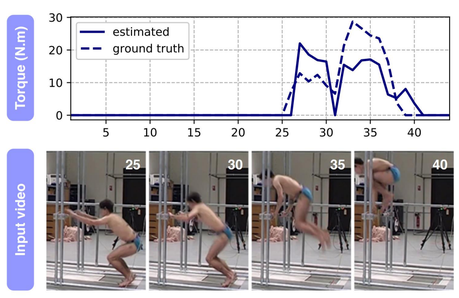
(2) $\kappa(q, c_k)$ contact constraints
- *||contact point object – contact point hand||*
- *soles stay on the ground.*

(3) $F_k \in \mathcal{F}$ ($\mathcal{F}$: force model) *e.g. respect friction cone*

Evaluation

Parkour dataset

Vicon / Force sensors



Results

Dynamics regularization improves pose estimations

	Jump	Move-up	Pull-up	Hop	Avg
SMPLify [2]	121.75	147.41	120.48	169.36	139.69
HMR [3]	111.36	140.16	132.44	149.64	135.65
Ours	**98.42**	**125.21**	**119.92**	**138.45**	**122.11**

3D joints error [mm]

Novel task

	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

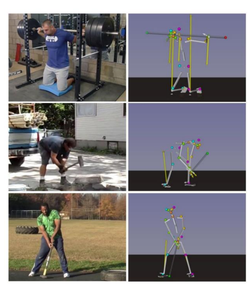
External Forces [N] Torques [N.m]

New manually annotated tool dataset

Method	Barbell	Hammer	Scythe	Spade
Mask R-CNN (He et al., 2017)	33/42/54	35/44/45	**63**/72/76	54/79/93
Ours (generic model)	**47/72/96**	**63/91/98**	51/**87/98**	**56/85/99**

Correct Object localization [% in @25/50/100px]

Limitations

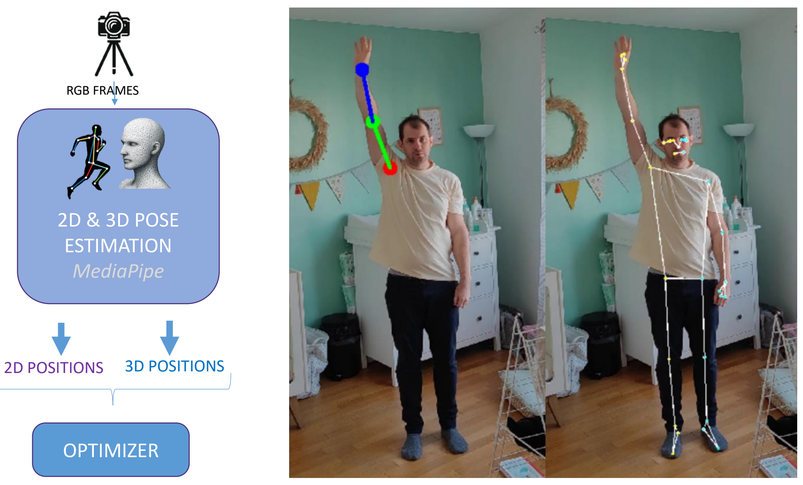


- Confidence on the forces and torques estimation
- High complexity method and code.
- Tedious tuning of loss coefficients
- Novel and difficult benchmark
  - no simulation, motion capture only
- Variability with tool weight and body dimensions
- Vision pipeline with multiple inter-dependencies.

Simplifications

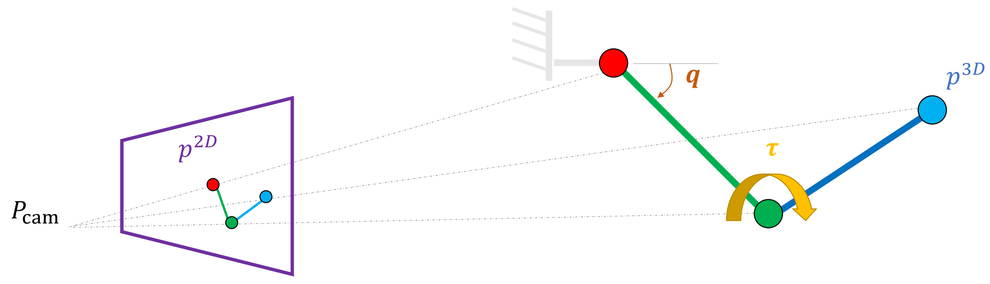
Arm model, No objects, Static calibrated camera

System Overview



Optimizer

Minimize *costs over States q , Torques τ*

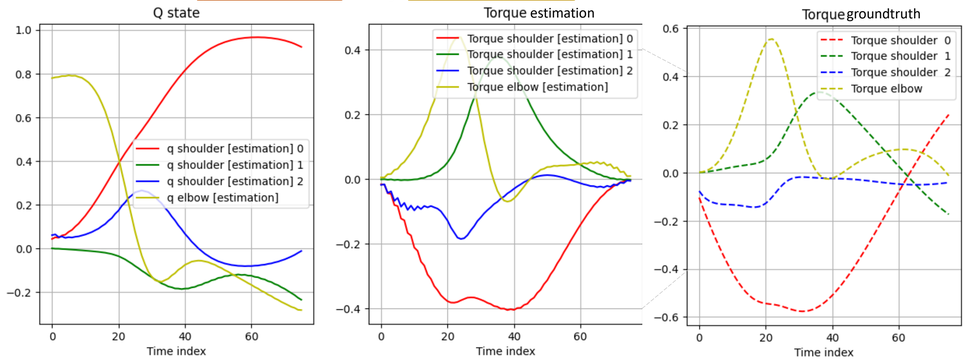


$$min_{\underline{q},\underline{\tau}} \sum_t \left[\underbrace{\sum_j \rho\left[p_j^{3D} - p_j(q)\right]}_{l_{3D}} + \underbrace{\sum_j \|v_j(q,\dot{q})\|^2 + \|\alpha_j(q,\dot{q},\ddot{q})\|^2}_{l_{smooth}} + \underbrace{\|\tau\|^2}_{l_{torque}} + \underbrace{\|M(q)\ddot{q} + b(q,\dot{q}) - g(q) - \tau\|^2}_{\textit{Relaxed dynamics [2]}} \right]$$

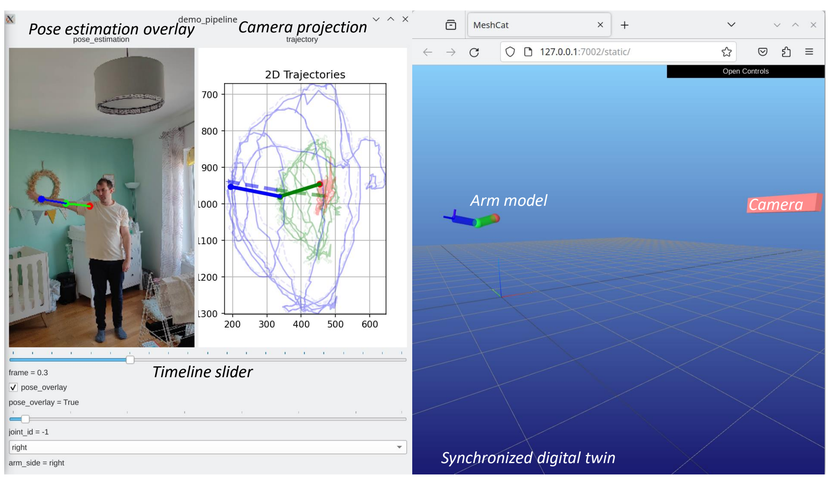
Linear velocity *Linear acceleration*

Simulation of arm in free fall with friction

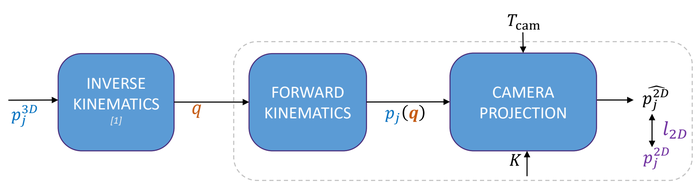
Results, average pose error 3.72e-03m, average torque error 1.67e-01N.m



Demo



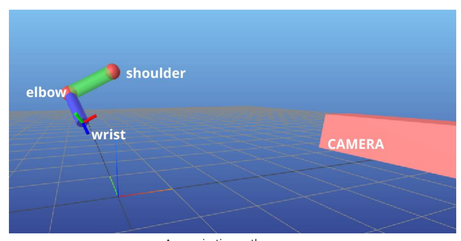
Python – Pinocchio – Meshcat visualization – Scipy solver – Interactive pipe – OpenCV – Mediapipe

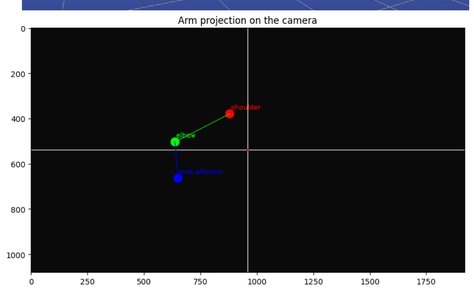


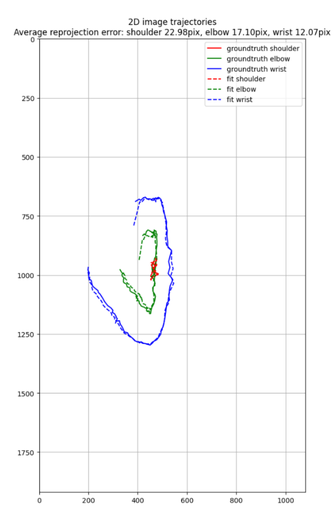
$$\min_{\underline{q},T_{\text{cam}}} \sum_j \left\| p_j^{2D} - K.[Q_{\text{cam}} \mid T_{\text{cam}}] p_j(q) \right\|^2$$

- *3*2 equations < 3 angles + 3 translations + 4 arm states*
- *Pre-calibrate [3] intrinsic matrix K*
- *Force Q_{cam}= identity*
- *Fit simplified camera pose T_{cam}*
- *Smoothness term $\| \dot{T}_{\text{cam}} \|$*
- *Ideally: Inverse kinematics on q with 2D reprojection error*

Camera pose estimation







[1] 3D arm lengths are standardized [2] $\dot{q}$, $\ddot{q}$ estimated using finite differences [3] A Flexible New Technique for Camera Calibration, Zhengyou Zhang