3D Tracking of Hands Interacting with Several Objects

Nikolaos Kyriazis
kyriazis@ics.forth.gr
Antonis A. Argyros
argyros@ics.forth.gr

Institute of Computer Science, FORTH,
N. Plastira 100, Vassilika Vouton,
GR70013, Heraklion, Crete, Greece

Humans and robots may acquire knowledge by observing demonstrations of object manipulation in scenarios ranging from everyday tasks, such as tieing laces and executing a recipe, to critical operations like surgery, electronics (dis-)assembly, *etc*. As the corpus of related videos is enlarged and the knowledge extraction becomes mission critical, the automation of the knowledge extraction process becomes increasingly important.

Estimating the detailed configuration of hands and objects in 3D space and across time may be of fundamental importance towards achieving high level understanding of such hand-object interactions. Most relevant work in the literature regards the presented context with identifiable limits. There is plenty of work on the problem of tracking a single hand in 3D. For a review the reader is referred to [8]. There are also a few approaches to the problem of 3D tracking of a single hand and a single object [2, 12], two hands [11] and two hands and a single object [1]. However, in the aforementioned demonstrations it is rarely the case that only a single object is being manipulated. On the contrary, usual interaction scenarios involve several objects, with instances even involving concurrent interaction among sizeable sub-groups (*e.g*. (dis-)assembly, surgery, *etc*.). Effective handling of the intrinsically large complexity of these cases constitutes a challenging and interesting goal.

We propose that the detailed computation of an entire scene's 3D configuration should be achieved through top-down model-based 3D tracking. Evidently, top-down methods present straightforward generalization paths, as indicated in the transition from the 3D tracking of a single hand based on RGBD input [10] to single hand and single object 3D tracking from multi-camera input [9]. Parametric 3D models of hands and objects can either be designed or acquired [5]. The most important aspect of interaction, *i.e*. dynamics, can also be modelled in parametric forms [5, 7, 9]. To simplify the incurred optimization problems temporal continuity is assumed, which in turn gives rise to tracking.

Here, we present two distinct and complementary approaches, which fall within the same computational framework and push 3D tracking beyond the boundary of two hands and an object. One approach emphasizes on optimization and makes it practical for large scene scales by exploiting the structure of the tracking problem. The other approach focuses on modelling and shows that the exploitation of physics as a prior can lower the dimensionality of the tracking problem.

1 The computational framework

The main goal in this work is to track the 3D state of multiple interacting entities based on visual input. The adopted approach [4] is model based, *i.e*. the expected variability in appearance and interaction of the entities is captured in predefined parametric models. The entire parametric model is connected to observations through an objective function, which acts as a compatibility measure between hypothesized model instantiations and actual observations. Then, tracking amounts to identifying the most compatible model instantiation, *i.e*. the set of parameters, which configures the model so that it best reproduces the observations.

In notation, let a scene comprise N entities whose entire and combined state is given by the parameter vector $x \in X$. Let also $\mathbf{M}$ be a forward model that maps such states into a feature space F:

$$f = \mathbf{M}(x), f \in F, x \in X. \tag{1}$$

Given that there exists a procedure $\mathbf{P}$ that maps actual observations $o \in O$ into F, a prior term $\mathbf{L}$ that reflects how likely hypotheses originally are, regardless of the observations and the solution history h for the previous frames, we can formulate the following error function:

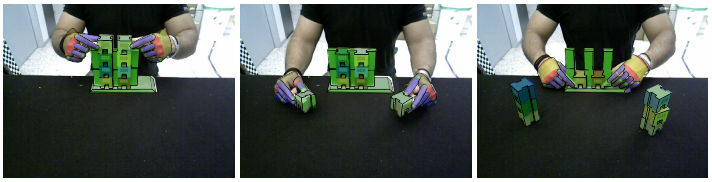


Figure 1: Two hands and 15 blocks are tracked, in 3D, in a 159-D problem. The appearance of the scene constituents is similar and in some cases it is identical. The objects are also tightly packed. The ECT approach (see section 2) is not confused in such a difficult and large problem due to the joint yet computationally scalable consideration of the entire scene, during tracking. Images were extracted from `https://youtu.be/SCOtBdhDMKg`.

$$\mathbf{E}(x,o,h) = \|\mathbf{M}(x,h) - \mathbf{P}(o)\| + \lambda \mathbf{L}(x,h) \tag{2}$$

which quantifies the discrepancy between actual observations o and a hypothesized scene state x. Then, the problem of estimating the state of the scene for the current frame reduces to finding the minimizer s of the parameters x of $\mathbf{E}$ for given observations o:

$$s \overset{\Delta}{=} \arg\min_{x} \mathbf{E}(x,o,h). \tag{3}$$

We exploit temporal continuity between adjacent input frames and perform optimization locally, in the vicinity of solutions for the previous frames. By initializing search near last known solutions and by modulating computations so as to impose a preference over smooth estimation transitions between frames, we can effectively substitute global optimization for local optimization and become, at the same time, effective and fast in the state estimation task, as long as the temporal continuity assumption holds. A shortcoming of the presented approach is visually ambiguous situations.

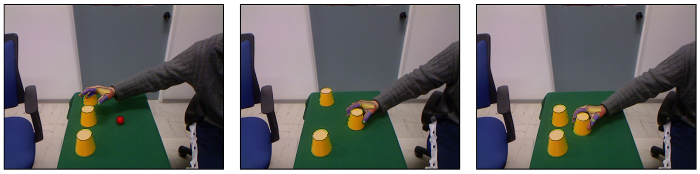
The computation of eq. (3) is performed by employing Particle Swarm Optimization (PSO) [3]. In what is presented, $\mathbf{M}$ corresponds to 3D rendering, $\mathbf{P}$ corresponds to foreground extraction and $\mathbf{L}$ amounts to penalization of collisions. From eq. (3) and various instantiations of $\mathbf{M}$, $\mathbf{P}$ and $\mathbf{L}$, a GPU-accelerated computational framework has been devised, which uses parallel search heuristics, such as PSO, to solve the tracking problem [4].

2 Scalable optimization by exploiting structure

A single hand can be tracked by the process described in section 1, implemented in [10], by employing the computational framework of [4]. Two independently moving, non-overlapping hands could be tracked by employing two distinct instances of [10]. In [6] this method is referred to as a Set of Independent Trackers (SIT). For tracking to succeed in the case where the hands interacted closely it would be required to solve the joint tracking problem, as in [11], or the unaccounted interactions would lead to tracking failure. In [6], this method is referred to as Joint Tracking (JT).

For tracking two hands and 15 objects (see, fig. 1) where every object potentially interacts with every other object, both in projective 2D space, through overlap, and in 3D space, through collisions, JT is most appropriate but impractical [6]. This is due to the large search space and the lack of a scalable exploratory strategy. To solve such problems which involve many articulated and rigid objects, we propose an Ensemble of Collaborative Trackers (ECT) [6].

ECT acts as a middle ground between SIT and JT. As in SIT, it incorporates a tracker for each distinct tracked object. Nevertheless, the trackers are not entirely independent, since inter-communication is introduced



(a) (b) (c)

Figure 2: The *single actor hypothesis* (see section 3) distinguishes the constituents of a scene into being active (hands) or passive (objects). The state of the entire scene is invariably determined by the state of the active objects alone. Thus the amount of degrees of freedom to be tracked are significantly decreased. *E.g.*, a cup or a ball have no associated degrees of freedom, given that it is only because of the hand that they might be moved. This is straightforward to represent in a physics simulation, which is parameterized over hand motion alone but can generate expectations for the evolution of a scene of arbitrary size. Employing physics can also help in reasoning over the state of the ball even when it disappears as it is covered by a cup. Images were extracted from `https://youtu.be/0RCsQPXeHRQ`.

to approximate joint consideration as in JT. In more detail, each tracker is assigned to a distinct object. Still, each tracker regards the entire state x (including other objects) during optimization, and thus, $\mathbf{M}$ is evaluated on the entire state x (see eq. (2)), so that all interactions are accounted for. However, optimization in each tracker only regards a small subset of x: the subset which corresponds to the parameterization of the corresponding object alone. The rest of the parameters in x are bound to the optimal state for the rest of the objects, as computed by other trackers during the parallel optimization of all.

The decomposition of the problem, combined with the tolerance of each individual tracker against delayed state updates by other trackers, allows for: (a) a computational profile which closely approximates the simple and scalable SIT approach and (b) an approximation of ideal JT application, which is such that yields an overall better performance for ECT, compared to JT, not only in high-dimensional problems but also in low-dimensional problems.

3 Scalable modelling through the use of physics

In this approach the focus lies in modelling. In the discussed scenario of hands manipulating several objects (*e.g.* see fig. 2), a distinction can be made, whose exploitation can lead to physics-powered scalable 3D tracking [5]. The objects can be distinguished between being active or passive. The state of the passive parts of a scene is invariably determined by the state of the active parts of the scene. The reason for which an active object is moved is unknown and is corresponded to a parameter set of unknown variables, i.e. the degrees of freedom of the object. Passive objects are those whose state can be altered only as a consequence of the motion of an active object. In the hand(s)-object(s) interaction scenario, the passive building blocks of fig. 1 and the passive cups and ball of fig. 2 are only moved, directly or indirectly, due to the motion of the active hand(s).

The aforementioned observation has a straightforward representation in dynamics simulation. Each object, articulated or rigid, is corresponded to a set of bodies with each body having a 3D shape and physical traits which describe interactive behaviour, *e.g.* mass, friction, restitution, *etc*. [5]. Active bodies are accelerated by some unknown forces. Passive bodies are only accelerated through collision with other bodies. A hand(s)-object(s) scenario can be represented after a simulation process $\mathbf{S}$, with active bodies for hands and passive bodies for the rest of the objects (table, building blocks, cups, ball, *etc*.). This process is parameterized over the motion of the hand(s) alone, but can generate expectations which regard the whole scene. Back-projecting such expectations to actual observations yields an objective function which quantifies the likelihood of hand(s) motion by also incorporating consequences. From this it follows that by replacing $\mathbf{M}$ in eq. (2) with $\mathbf{M}' = \mathbf{M}(\mathbf{S}(\mathbf{x}))$ one can eliminate from x the parameters which correspond to the the state of the passive objects, as they are now directly computed as a consequence rather than being hypothesized as independent variables.

Incorporating physics does not only help isolate the true degrees of freedom of the tracking problem. It is true that in the scenario of fig. 2, the problem of tracking a hand and 4 objects is reduced from a 56-D problem to a 27-D one. However, what is also striking is the ability that is given to tracking to account for total occlusions. The ball is at some point entirely occluded by a covering cup which is moved around. Humans know the ball has to travel with the cup. Physics-powered 3D tracking can also successfully invoke this kind of common sense as it is a direct implication of the fact that, even unobserved, the series of events cannot have the ball escape the space between the table and the covering cup. More equally motivating examples are given in [5].

4 Discussion

We have presented two distinct approaches that deal with the scalability problem in 3D tracking of hand(s)-object(s) interaction scenes. One approach deals directly with optimization and the other deals directly with modelling. Both approaches have had thorough validation, through experimentation, as presented in [5, 6]. Also, both approaches have been implemented as instantiations of the same computational framework [4]. We find exciting the possibility of applying both in order to help track large scenes, in 3D, by confronting simultaneously the modelling and optimization scalability issues. The last remark constitutes the object of future work.

Acknowledgements

This work was partially supported by projects FP7-IP-288533 Robohow and FP7-ICT-2011-9 WEARHAP.

[1] Luca Ballan, Aparna Taneja, Jürgen Gall, Luc Van Gool, and Marc Pollefeys. Motion capture of hands in action using discriminative salient points. In *Computer Vision–ECCV 2012*, pages 640–653. Springer, 2012.

[2] Henning Hamer, Konrad Schindler, Esther Koller-Meier, and Luc Van Gool. Tracking a hand manipulating an object. In *CVPR*, pages 1475–1482. IEEE, 2009.

[3] James Kennedy. Particle swarm optimization. In *Encyclopedia of Machine Learning*, pages 760–766. Springer, 2010.

[4] Nikolaos Kyriazis. *A computational framework for observing and understanding the interaction of humans with objects of their environment*. PhD thesis, University of Crete, 2014.

[5] Nikolaos Kyriazis and Antonis Argyros. Physically plausible 3d scene tracking: The single actor hypothesis. In *CVPR*, pages 9–16. IEEE, 2013.

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

[7] Nikolaos Kyriazis, Iason Oikonomidis, and Antonis A Argyros. Binding vision to physics based simulation: The case study of a bouncing ball. In *BMVC*. BMVA, 2011.

[8] Alexandros Makris, Nikolaos Kyriazis, and Antonis Argyros. Hierarchical particle filtering for 3d hand tracking. In *CVPRW*, pages 8–17, 2015.

[9] Iason Oikonomidis, Nikolaos Kyriazis, Antonis Argyros, et al. Full dof tracking of a hand interacting with an object by modeling occlusions and physical constraints. In *ICCV*, pages 2088–2095. IEEE, 2011.

[10] Iason Oikonomidis, Nikolaos Kyriazis, and Antonis A Argyros. Efficient model-based 3d tracking of hand articulations using kinect. In *BMVC*, volume 1, page 3, 2011.

[11] Iasonas Oikonomidis, Nikolaos Kyriazis, and Antonis A Argyros. Tracking the articulated motion of two strongly interacting hands. In *CVPR*, pages 1862–1869. IEEE, 2012.

[12] Javier Romero, Hedvig Kjellström, and Danica Kragic. Hands in action: real-time 3d reconstruction of hands in interaction with objects. In *ICRA*, pages 458–463. IEEE, 2010.