

Predicting Physical Object Properties from Video

Martin Link

Autonomous Intelligent Systems
University of Bonn
Bonn, Germany
mlink@uni-bonn.de

Max Schwarz

Autonomous Intelligent Systems
University of Bonn
Bonn, Germany
schwarz@ais.uni-bonn.de

Sven Behnke

Autonomous Intelligent Systems
University of Bonn
Bonn, Germany
behnke@cs.uni-bonn.de

Abstract—We present a novel approach to estimating physical properties of objects from video. Our approach consists of a physics engine and a correction estimator. Starting from the initial observed state, object behavior is simulated forward in time. Based on the simulated and observed behavior, the correction estimator then determines refined physical parameters for each object. The method can be iterated for increased precision. Our approach is generic, as it allows for the use of an arbitrary—not necessarily differentiable—physics engine and correction estimator. For the latter, we evaluate both gradient-free hyperparameter optimization and a deep convolutional neural network. We demonstrate faster and more robust convergence of the learned method in several simulated 2D scenarios focusing on bin situations.

Index Terms—System identification, physics simulation, physical parameters, iterative refinement

I. INTRODUCTION

Many tasks in robotics and autonomous systems require a reliable model of the world that is surrounding the agent. Notably, this includes a model of the physical properties of foreign objects: Mass, surface friction, elasticity, moment-of-inertia, and density distribution/center-of-mass can all play a crucial role and—when guessed incorrectly—lead to failure cases. One prominent example application is bin picking [1], [2], where a robotic agent has to detect and manipulate objects.

We explore ways to improve the knowledge of physical parameters of objects, as illustrated in Fig. 1. One way to achieve this is the use of a physics engine which in itself is differentiable and therefore allows for backpropagation of observed errors between simulation and reality *through* the engine, in order to update the physical parameters of a simulation. In this work, we investigate a more flexible approach: We propose a framework that uses a correction estimator—e.g. a neural network—to refine the physical parameters in a simulation iteratively, by comparing an observation to the simulation. Since we treat the physics engine as a black box, we remove the need for differentiability and instead rely on well-understood neural network and machine learning techniques. This simplification allows for the use of commercially available physics engines that are already optimized for speed and efficiency but are not necessarily differentiable.

Our contributions include 1) a general iterative framework for optimization of physical object parameters and 2) a thorough evaluation of learning-based and task-agnostic approaches for the correction estimator.

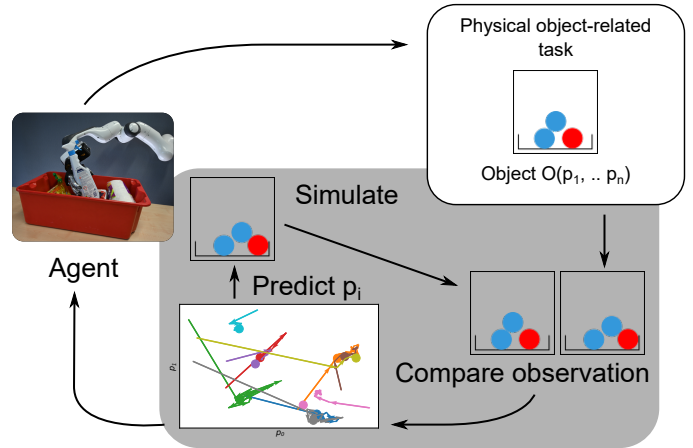


Fig. 1. Overview. An agent needs to complete a general task related to a physical object O that is described by parameters p_i . Our approach (grey box) allows the agent to learn these parameters by observing the behavior of the objects and comparing it to a simulated scene that uses the current best estimation of p_i . Comparing both observation and simulation yields a refined set of parameters that can either be used to attempt completing the task or re-iterated with the current observation.

II. RELATED WORK

The task of predicting physical parameters of objects has been investigated before, for example from short video sequences in an unsupervised fashion [4]. Often, properties are also predicted from single pictures, for example via micro-CT pictures for porous media [5], pictures of stacked crops [6], or pictures of liquid crystals [7]. In a more robotics-centric context, neural networks have been used to predict the hardness of objects with a GelSight sensor [8] and to identify parameters of unmanned aerial vehicles [9]. In contrast to these earlier works, our dynamic approach utilizes iteration over observed scenes, which allows for constant refinement of the parameters by taking into account new information.

One application of our technique is bin picking. Here, the success of planned operations is critically dependent on how well a robotic agent can be controlled. Often, controllers are optimized in simulated environments with reinforcement learning, particle swarms, or genetic algorithms, completely free of any derivatives [10]–[12]. As a consequence, the robot is often treated as black box, which prevents the use of efficient gradient-based deep learning methods. Physics

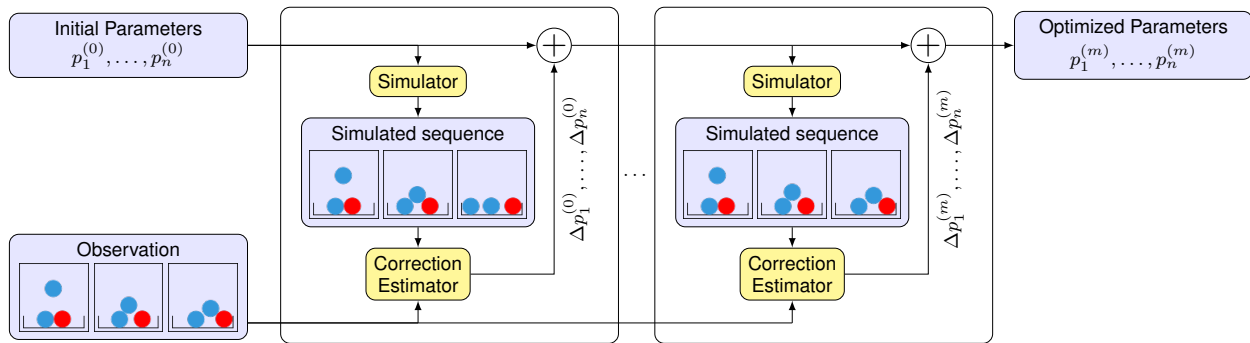


Fig. 2. Overview of our approach. The process starts with an initial guess of physical parameters $p_1^{(0)}, \dots, p_n^{(0)}$ and the observed video sequence. In each iteration, the scene is simulated using the current parameter estimate. The (learned) correction estimator then produces a parameter update, which results in a better parameter estimate for the next iteration. Figure inspired by [3].

engines like MuJoCo [13] allow for the evaluation of gradients between states and actions through a robot, however, they do not allow backpropagation to the initial model parameters. More recent work investigated special physics engines that are end-to-end differentiable [14], [15], however, when testing the use of such engines, we found that they were much slower and less adaptable to new scenarios than commercial engines like NVidia PhysX¹. Stillleben is a PhysX-based framework for the creation of simulated data to train agents in bin picking scenarios [16]. Recently, Werling *et al.* [17] introduced nimblephysics, a differentiable fork of the Dart physics engine. While nimblephysics allows backpropagation of object parameters, this is currently limited to the object mass—highlighting the versatility of our approach, as we enable optimization of any parameter as long as it influences the simulation result in a learnable manner.

In this work, we present a framework that allows efficient estimation of physical properties in an online fashion. This includes keeping our framework efficient in order to be able to actively improve the scene knowledge at runtime. To this end, we build on an iterative approach that is used to estimate the 6D pose of objects [3]. There, a simulated pose is compared to an observed object in order to determine changes in the pose. This is done iteratively, in order to refine the predictions and allow for incremental improvements. In our case, we *observe* time series of frames to allow for efficient interpretation of the shown dynamics. To estimate the relevant parameters, we compare the observed scene to a simulated scene. The comparison of both then allows to determine corrections to parameters of the simulation that correspond to physical properties of the involved objects. The neural network used as a correction estimator is based on the ResNet architecture [18], but modified for the use of 3D convolutions to account for the time dimension.

III. METHOD

Our approach to predicting the physical parameters of objects is inspired by earlier work on iterative prediction of pose estimation [3]. In summary, we observe the scene as a

sequence of video frames and try to estimate the physical properties by producing a simulated scene with guessed parameters. Importantly, this approach separates the simulation from the task of predicting the input parameters. In comparison to approaches where the simulation itself has to be differentiable, we can independently choose which physics engine we use for simulation and how to optimize the parameters. This principle is illustrated in Fig. 2. Starting from an initial guess of the parameters $p_1^{(0)}, \dots, p_n^{(0)}$, we simulate the scene. Note that we assume some sort of pose estimation that gives us the capability to match the observable system state, so that we can start the simulation in the same configuration as seen in the first observation frame.

Both sequences of pictures, observed and simulated, are then passed to the correction estimator, which outputs a set of parameter changes $\Delta p_i^{(t)}$ for all n parameters entering the simulator. These parameter changes are then added to the initial parameters to yield the updated simulation parameters.

The process can be iterated as often as needed: The updated parameters can be used for the next simulation. In summary, each iteration consists of one forward pass through the simulator and one pass through the correction estimator. One important question is how our approach should treat different numbers of objects in the scene. When predicting physical parameter changes, the naive approach would be to predict all parameter updates at the same time. However, since the shape of the output is fixed at runtime, this would mean the network is only usable for a fixed number of objects. In our case, we decided to gain flexibility by predicting parameter changes for objects one at a time. We select the object of interest on the input by marking it with a different color, see Fig. 3. In the bin picking use case, this would amount to an initial segmentation task that can either be performed separately, or by the correction estimator itself. Comparing the single-object and multi-object approaches, no significant difference in performance was observed. An alternative approach to color-marking would be to place the objects themselves in separate channels, e.g. by instance segmentation.

¹<https://developer.nvidia.com/physx-sdk>

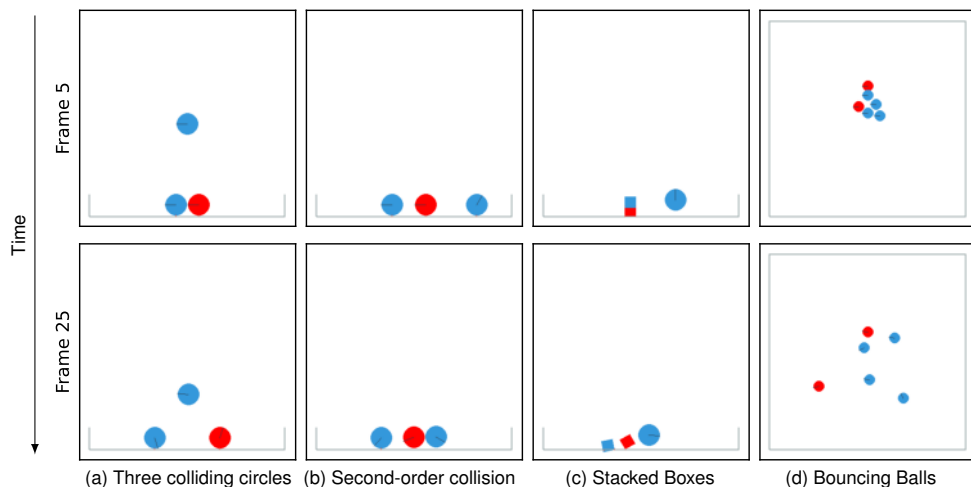


Fig. 3. Experimental scenes.

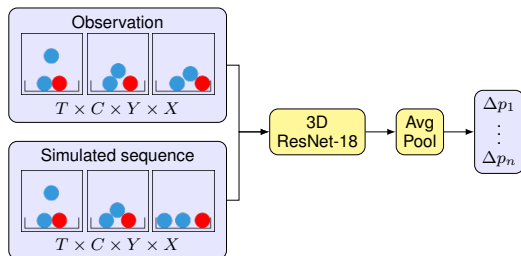


Fig. 4. Architecture of the learned correction estimator. The batch dimension is omitted.

A. Simulator Module

In principle, with our approach, the choice of the physics engine is arbitrary. For testing in a two-dimensional environment, we chose the Python-based physics library Pymunk², which is built on the physics engine Chipmunk³. The library allows for efficient simulation of two-dimensional rigid-body physics. Objects can be defined with arbitrary (two-dimensional) shapes, represented through polygons. Other physical properties like the mass, center-of-mass, moment of inertia, elasticity, and friction can be set as well, which allows for several different parameters to be predicted by our framework. Forces can be added to objects, which allows for dynamical behavior. Fixed (immovable) objects can be added to the scene that can act as barriers. Pymunk includes helper functions for visualization with pygame⁴, which allows for prototyping and visualization of our test scenes.

B. Correction Estimator

One possible choice for the correction estimator is to use generic gradient-free optimization techniques. In our case, we

employ the *hyperopt* library [19] for Python, which implements a Tree of Parzen Estimators (TPE) approach to find the optimal set of parameters for a given objective function [20]. The objective function has to take the parameters as input and provide a measure for the accuracy of the output. As a criterion for accuracy, we choose the mean squared error between the ground truth and simulated time series in image-space. Hyperopt logs arbitrary measures during the optimization cycle and provides the best set of parameters after optimization. We expected that Hyperopt gives a reasonable baseline for optimization.

An alternative choice for the correction estimator is to use a neural network, which learns the task of predicting optimal parameter updates. The expectation is that a learned estimator can exploit the characteristics of the underlying system much better than a generic optimizer.

Image-like representations and their time series are well-suited as an input representation, since they are readily available in our application domain (e.g. from semantic segmentation of the scene) and contain the necessary details, such as object position and contact information. Therefore, our network directly operates on this representation. The network architecture closely resembles a ResNet-18 architecture [18], however, we use 3D convolutional layers to accommodate for the time series collection of pictures (see Fig. 4). The input to our network has the shape $(B \times 2T \times C \times Y \times X)$, where B is the batch size (10 or 20, with no discernable difference in performance), $2T$ the length of the concatenated time series (usually $2 \times 30 = 60$), C the number of color / object channels and $Y \times X$ the resolution of the frames. We furthermore found that in our case, dropout regularization leads to more stable results as compared to the usually employed batch normalization. The network is trained to predict correction estimates between randomly chosen parameters (p and p' for observation and guess time series, respectively) for a randomly chosen object in the scene. The loss function is the mean

²<http://www.pymunk.org>

³<http://chipmunk-physics.net>

⁴<https://www.pygame.org>

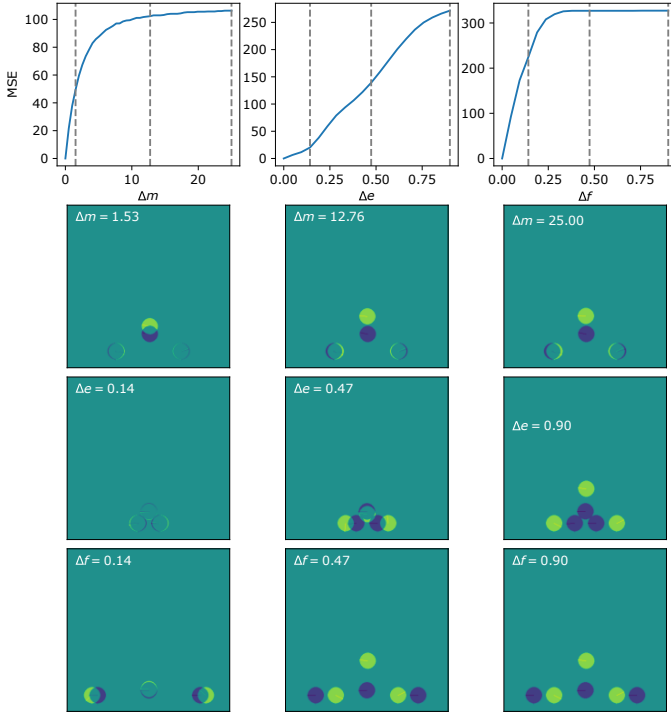


Fig. 5. Image-space MSE between two full time series with different physical parameters. The parameter under investigation is fixed to a specific value ($m_1 = 5$ and $f_1 = e_1 = 0.1$ for mass, friction, and elasticity) for one of the time series and varied for the other ($p_2 = p_1 + \Delta p$ for $p \in (m, e, f)$). The pictures in the lower three rows show the difference between the last frame of each time series, respectively. The grey dashed lines in the top three plots indicate the values of the pictures in the bottom three rows (each panel in the first row from left to right represents a row from top to bottom).

squared error between the ground truth parameter correction $\Delta p_{\text{gt}} = p - p'$, and guessed parameter correction by the correction estimator. In comparison to the Hyperopt approach, the network predicts a correction of the parameters that go into the simulator Δp_i (as compared to the absolute value p_i). The updated value can then be calculated by incrementing the old value with the parameter update $p_i^{(t+1)} = p_i^{(t)} + \Delta p_i^{(t)}$.

C. Multiple Objects

To keep our approach flexible with regards to the number of objects present in the scene, we highlight one of the objects by color. The correction estimator then predicts parameter updates for the highlighted object. Alternative approaches are to place the investigated object in a different image channel, or fix the output dimension to allow for simultaneous prediction of several objects' properties. Both approaches have been tested and were found to yield similar performance. While highlighting single objects requires N_{obj} more forward passes through the prediction module, it allows for more flexibility. We note that the required forward passes may be performed in parallel.

IV. EVALUATION

We evaluate our approach in two different scenarios with different combinations of predicted parameters and object

configurations—a simple bin configuration and a more complex pool table setup. We compare the performance of Hyperopt and the neural network as correction estimator. Our measure of performance is the minimum achieved mean squared error between guessed and real parameters. To be able to compare different parameters with different number ranges, each parameter is normalized by its maximum achievable value. In case of unknown ground truth values, the minimum achieved error can be found by comparing the mean squared error between the time series, similar to the objective function of the Hyperopt approach. Throughout our testing, runtime was dominated by physics simulation, not correction estimation. As a consequence, the number of iterations needed to find good parameters directly measures the performance of the approach for the correction estimator.

The learning-based estimator was trained using the Adam optimizer (learning rate $5e-6$) on 130000 simulated scenes. We note that it does not need to be trained on a specific scene configuration. The required training time can thus be done before deployment in a particular application.

A. Objects in a Bin

The first investigated scene is inspired by bin picking. It consists of a two-dimensional box formed by immovable lines, with three objects placed in it. One of the observed objects is a test object with constant physical parameters throughout all scene iterations. This object fixes the numerical value for the other objects, as the collisions dynamics between objects only depends on the ratio between their physical parameters. All objects are subject to gravitational force, and the test object is additionally accelerated towards the other objects. In this scenario, we investigate three different scenes: 1) Three circles, where the test object is accelerated towards the two unknown object from above, 2) second-order collisions, where the test object is accelerated from one direction towards one of the unknown objects, which then interacts with the other unknown object and 3) stacked boxes, with two square-shaped objects stacked on top of each other, where the circular shaped test object is accelerated towards the stacked objects. These scenes are depicted in Fig. 3 (a) to (c).

To verify predictiveness of the different physical parameters, we investigate how the mean squared error between the raw time series changes for different values of the predicted parameters, see Fig. 5, topmost row. This analysis shows, that for both mass and elasticity, a gradient towards small parameter changes always exists (albeit smaller for mass). For friction on the other hand, the curve looks flat above a certain threshold. This can be explained by physical considerations: To first order the collisional dynamics of rigid circles does not depend on the surface friction, therefore, changing the friction does not influence the trajectory of the objects. Fig. 5 also shows the difference between the final frames of the compared timeseries for different parameter values. This highlights the necessity for expressive dynamics, that yields information about the physical parameters under investigation.

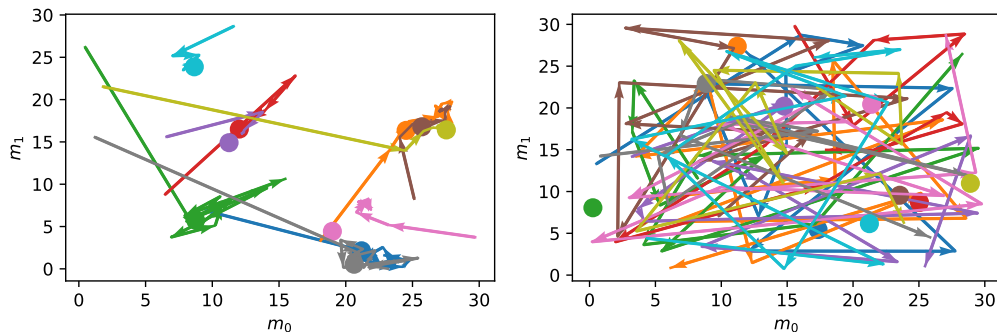


Fig. 6. Parameter space trajectories for two unknown masses m_0 and m_1 for the three circle scene (Fig. 3a). Shown are a neural network (left) and Hyperopt (right) as correction estimator. Each color represents a new scene instance, with optimization iterations shown as arrows and the ground truth parameters as a filled circle.

TABLE I
ESTIMATION ERRORS

Configuration	Mass	Elasticity	Friction	Mass + Elasticity
3 circles	$(2.6 \pm 0.9) \times 10^{-3}$	$(8.7 \pm 1.9) \times 10^{-5}$	-	$(1.4 \pm 0.5) \times 10^{-2}$
3 circles (Hyperopt)	$(1.1 \pm 0.3) \times 10^{-2}$	-	-	$(5.2 \pm 0.7) \times 10^{-2}$
second order	$(3 \pm 0.7) \times 10^{-4}$	$(4 \pm 2) \times 10^{-3}$	-	-
box scene	$(2 \pm 0.5) \times 10^{-3}$	-	$(3.9 \pm 1.2) \times 10^{-3}$	-
bouncing balls	$(6.3 \pm 0.8) \times 10^{-3}$	-	-	-

We show the normalized minimum achieved mean squared errors between predicted and true values after 11 iterations. Each mean is taken over 10 runs with random starting parameters. We also show standard deviations for each error.

We train the network on each scene separately on randomly chosen parameters to predict parameter changes. To evaluate the performance, we generate new scenes with random parameters and predict parameter updates for each mass starting with a random guess. We then resimulate and iterate this procedure. We find that our approach generally converges towards its final value after one or two iterations, see Fig. 7. For comparison, we test generic hyperparameter optimization to predict the parameters, which finds comparably good values after an order of magnitude more iterations, see Fig. 7 (bottom). This is also highlighted in parameter space, where the Hyperopt approach performs much more exploration, see Fig. 6. The best value after 11 iterations for each tested scene / parameter combination is shown in Table I. In agreement with the results of Fig. 5, we found that for some combinations of scenes and parameters, the performance of the NN drops drastically, for example when predicting the friction in scene 1). However, in a more expressive scene with regard to friction, i.e. scene 3), we find comparable performance. From Fig. 5, we also find that the image-space MSE has a favorable shape for the elasticity in the three circle scene, which is reflected in the superior performance when predicting elasticity.

We also investigate the question whether the network can handle scene observations of different nature than the sharp segmentations provided by the simulator. For this purpose, we apply a gaussian filter with $\sigma = 3$ pixels to the observations, so that we obtain a more fuzzy observation. The simulated scenes are, however, supplied in their original, sharp version. In our experiments we did not observe significant changes in the performance of the neural network as a correction estimator.

B. Bouncing Balls

The second scenario investigates the role of a variable number of objects in the scene. We use the well-known "bouncing balls" scene (a two-dimensional representation of a "pool table"), see Fig. 3d. We place one test object with fixed mass on a pool table that is bounded by four rigid lines. The test mass is accelerated towards a variable number of balls with unknown mass that are placed in a triangular pattern, such that all balls collide with at least one other object. We predict each objects mass separately by marking it with a different color. We randomly place between two and six balls with random mass on the table. To evaluate the performance of our approach, we update each objects mass as predicted by the network and then iterate over the updated guesses. Even in this more complex state space, we find only slightly reduced accuracy after several iterations. However, in comparison to fewer objects, it takes about three to four iterations for the values to converge. We furthermore find that the network is able to predict a scene with just one unknown object with high accuracy, although this setup was not part of the training process, see Fig. 7a, green line.

V. LIMITATIONS

Our method makes several assumptions, which will be discussed here. For instance, the system is currently limited to 2D scenes, but this is not an inherent constraint. Furthermore, we assume that the observed state can be easily compared to the simulated one, in our case by rendering a similar view from the simulation data. As (visual) scene complexity increases, this assumption may not hold anymore.

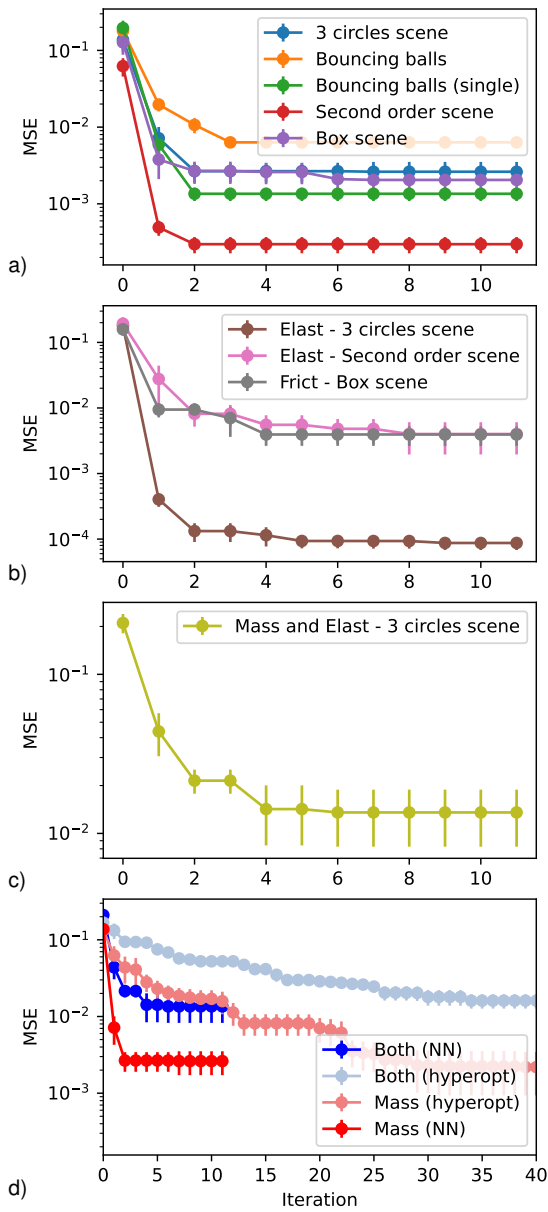


Fig. 7. Minimum normalized Mean Squared Error between predicted and true parameters for different scene and parameter combinations: a) mass, b) elasticity & friction, c) mass & elasticity. d) shows a comparison between hyperopt and the learned network for mass and mass & elasticity. Note that we show the best parameters until each iteration, with the image-space MSE used to determine the so-far-found optimum.

VI. CONCLUSION

In conclusion, we find that our approach of predicting physical object properties from video allows for much more rapid refinement of parameters in comparison to generic gradient-free parameter optimization techniques. This allows for a fast determination of physical parameters from a video input. The accuracy on all parameters is 10% or better. For higher accuracy, it may be necessary to either finetune the used neural network, or use a generic optimization approach

or a differentiable physics engine as a second step in the optimization process. Due to the iterative nature and fast convergence of our approach, it allows for online refinement of physical parameters. In any case, the accuracy of the estimation strongly depends on the expressiveness of the scene with regard to the investigated parameters.

ACKNOWLEDGMENTS

This work was funded by grant BE 2556/18-2 (Research Unit FOR 2535 Anticipating Human Behavior) of the German Research Foundation (DFG).

REFERENCES

- [1] D. Morrison, A. W. Tow, M. McTaggart, R. Smith, N. Kelly-Boxall, S. Wade-McCue, J. Erskine, R. Grinover, A. Gurman, T. Hunn, D. Lee, A. Milan, T. Pham, G. Rallos, A. Razijigaev, T. Rowntree, Z. Zhuang, C. Lehnert, I. Reid, P. Corke, and J. Leitner, “Cartman: The low-cost cartesian manipulator that won the Amazon Robotics Challenge,” in *International Conference on Robotics and Automation (ICRA)*, 2018, pp. 7757–7764.
- [2] M. Schwarz, C. Lenz, G. M. Garcia, S. Koo, A. S. Periyasamy, M. Schreiber, and S. Behnke, “Fast object learning and dual-arm coordination for cluttered stowing, picking, and packing,” in *International Conference on Robotics and Automation (ICRA)*, 2018.
- [3] Y. Li, G. Wang, X. Ji, Y. Xiang, and D. Fox, “DeepIM: Deep Iterative Matching for 6D Pose Estimation,” *International Journal of Computer Vision*, vol. 128, no. 3, pp. 657–678, 2020.
- [4] J. Wu, J. Lim, H. Zhang, J. Tenenbaum, and W. Freeman, “Physics 101: Learning physical object properties from unlabeled videos,” in *British Machine Vision Conference (BMVC)*, 2016, pp. 39.1–39.12.
- [5] N. Alqahtani, R. T. Armstrong, and P. Mostaghimi, “Deep learning convolutional neural networks to predict porous media properties,” in *SPE Asia Pacific oil and gas conference and exhibition*, 2018.
- [6] M. Hamdan, D. Rover, M. Darr, and J. Just, “Mass estimation from images using deep neural network and sparse ground truth,” in *International Conference On Machine Learning And Applications (ICMLA)*, 2019, pp. 1987–1992.
- [7] H. Y. Sigaki, E. K. Lenzi, R. S. Zola, M. Perc, and H. V. Ribeiro, “Learning physical properties of liquid crystals with deep convolutional neural networks,” *Nature scientific reports*, vol. 10, no. 1, pp. 1–10, 2020.
- [8] W. Yuan, C. Zhu, A. Owens, M. A. Srinivasan, and E. H. Adelson, “Shape-independent hardness estimation using deep learning and a GelSight tactile sensor,” in *International Conference on Robotics and Automation (ICRA)*, 2017, pp. 951–958.

- [9] A. Ayyad, M. Chehadeh, M. I. Awad, and Y. Zweiri, "Real-time system identification using deep learning for linear processes with application to unmanned aerial vehicles," *IEEE Access*, vol. 8, pp. 122 539–122 553, 2020.
- [10] C. Aguilar-Ibañez, "Stabilization of the PVTOL aircraft based on a sliding mode and a saturation function," *International Journal of Robust and Nonlinear Control*, vol. 27, no. 5, pp. 843–859, 2017.
- [11] J. d. J. Rubio, E. Garcia, G. Aquino, C. Aguilar-Ibañez, J. Pacheco, and A. Zacarias, "Learning of operator hand movements via least angle regression to be taught in a manipulator," *Evolving Systems*, vol. 11, no. 2, pp. 317–332, 2020.
- [12] J. Meda, "Estimation of complex systems with parametric uncertainties using a jssf heuristically adjusted," *IEEE Latin America Transactions*, vol. 16, no. 2, pp. 350–357, 2018.
- [13] E. Todorov, T. Erez, and Y. Tassa, "MuJoCo: A physics engine for model-based control," in *International Conference on Intelligent Robots and Systems (IROS)*, 2012, pp. 5026–5033.
- [14] J. Degraeve, M. Hermans, J. Dambre, *et al.*, "A differentiable physics engine for deep learning in robotics," *Frontiers in neurorobotics*, vol. 13, 2019.
- [15] F. de Avila Belbute-Peres, K. Smith, K. Allen, J. Tenenbaum, and J. Z. Kolter, "End-to-end differentiable physics for learning and control," *Advances in neural information processing systems*, vol. 31, 2018.
- [16] M. Schwarz and S. Behnke, "Stilleben: Realistic scene synthesis for deep learning in robotics," in *International Conference on Robotics and Automation (ICRA)*, 2020, pp. 10 502–10 508.
- [17] K. Werling, D. Omens, J. Lee, I. Exarchos, and C. K. Liu, "Fast and feature-complete differentiable physics for articulated rigid bodies with contact," *arXiv preprint arXiv:2103.16021*, 2021.
- [18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
- [19] J. Bergstra, D. Yamins, and D. Cox, "Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures," in *International conference on machine learning (ICML)*, 2013, pp. 115–123.
- [20] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for hyper-parameter optimization," *Advances in neural information processing systems*, vol. 24, 2011.