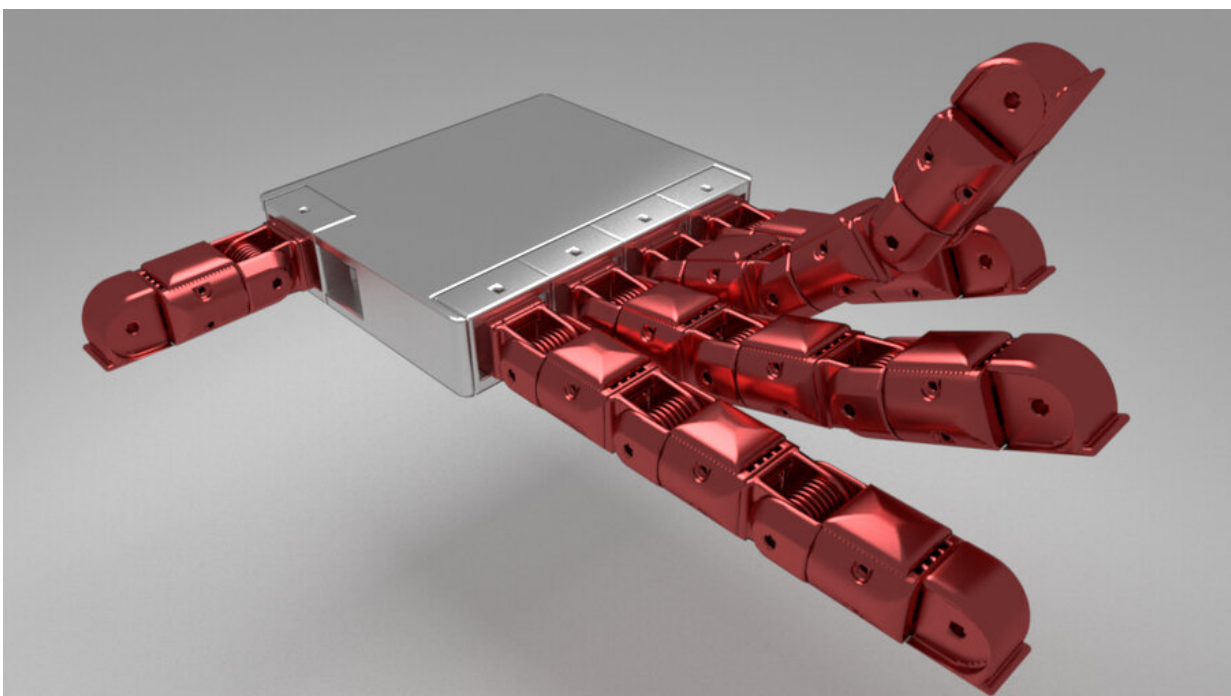


Trained neural network pipeline simulates physical systems of rigid and deformable bodies and environmental conditions

May 3 2022, by Lauren Hinkel



MIT researchers used the RISP method to predict the action sequence, joint stiffness, or movement of an articulated hand, like this one, from a target image or video. Credit: Massachusetts Institute of Technology. Credit: Massachusetts Institute of Technology

From "Star Wars" to "Happy Feet," many beloved films contain scenes

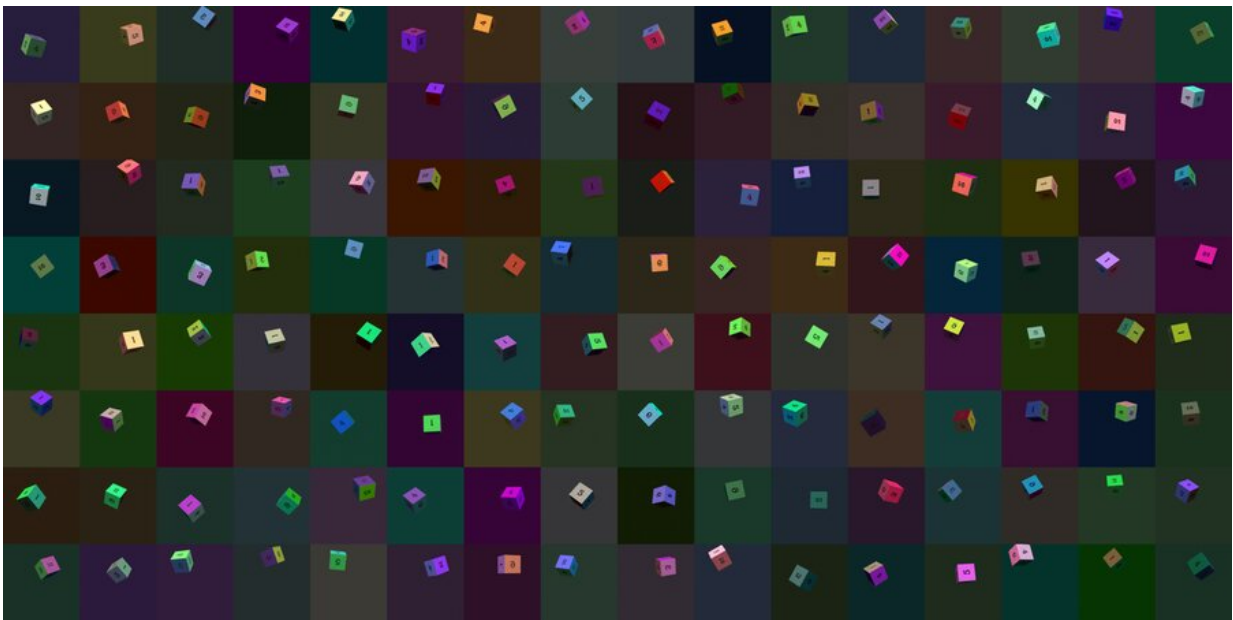
that were made possible by motion capture technology, which records movement of objects or people through video. Further, applications for this tracking, which involve complicated interactions between physics, geometry, and perception, extend beyond Hollywood to the military, sports training, medical fields, and computer vision and robotics, allowing engineers to understand and simulate action happening within real-world environments.

As this can be a complex and costly process—often requiring markers placed on objects or people and recording the action sequence—researchers are working to shift the burden to [neural networks](#), which could acquire this data from a simple video and reproduce it in a model. Work in physics simulations and rendering shows promise to make this more widely used, since it can characterize realistic, continuous, dynamic motion from images and transform back and forth between a 2D render and 3D scene in the world. However, to do so, current techniques require precise knowledge of the environmental conditions where the action is taking place, and the choice of renderer, both of which are often unavailable.

Now, a team of researchers from MIT and IBM has developed a trained neural network pipeline that avoids this issue, with the ability to infer the state of the environment and the actions happening, the physical characteristics of the object or person of interest (system), and its control parameters. When tested, the technique can outperform other methods in simulations of four [physical systems](#) of rigid and deformable bodies, which illustrate different types of dynamics and interactions, under various environmental conditions. Further, the methodology allows for imitation learning—predicting and reproducing the trajectory of a real-world, flying quadrotor from a video.

"The high-level research problem this paper deals with is how to reconstruct a digital twin from a video of a dynamic system," says Tao

Du Ph.D. '21, a postdoc in the Department of Electrical Engineering and Computer Science (EECS), a member of Computer Science and Artificial Intelligence Laboratory (CSAIL), and a member of the research team. In order to do this, Du says, "we need to ignore the rendering variances from the video clips and try to grasp of the core information about the dynamic system or the dynamic motion."



Caption: This training set was used to train the RISP pipeline to see how differences in rendering can affect texture, light, and background. Credit: Massachusetts Institute of Technology

Du's co-authors include lead author Pingchuan Ma, a graduate student in EECS and a member of CSAIL; Josh Tenenbaum, the Paul E. Newton Career Development Professor of Cognitive Science and Computation in the Department of Brain and Cognitive Sciences and a member of CSAIL; Wojciech Matusik, professor of electrical engineering and

[computer science](#) and CSAIL member; and MIT-IBM Watson AI Lab principal research staff member Chuang Gan. This work was presented this week the International Conference on Learning Representations.

While capturing videos of characters, robots, or dynamic systems to infer dynamic movement makes this information more accessible, it also brings a new challenge. "The images or videos [and how they are rendered] depend largely on the on the lighting conditions, on the background info, on the texture information, on the material information of your environment, and these are not necessarily measurable in a real-world scenario," says Du. Without this rendering configuration information or knowledge of which renderer is used, it's presently difficult to glean dynamic information and predict behavior of the subject of the video. Even if the renderer is known, current neural network approaches still require large sets of training data. However, with their new approach, this can become a moot point. "If you take a video of a leopard running in the morning and in the evening, of course, you'll get visually different video clips because the lighting conditions are quite different. But what you really care about is the dynamic motion: the joint angles of the leopard—not if they look light or dark," Du says.

In order to take rendering domains and image differences out of the issue, the team developed a pipeline system containing a neural network, dubbed "rendering invariant state-prediction (RISP)" network. RISP transforms differences in images (pixels) to differences in states of the system—i.e., the environment of action—making their method generalizable and agnostic to rendering configurations. RISP is trained using random rendering parameters and states, which are fed into a differentiable renderer, a type of renderer that measures the sensitivity of pixels with respect to rendering configurations, e.g., lighting or material colors. This generates a set of varied images and video from known ground-truth parameters, which will later allow RISP to reverse

that process, predicting the environment state from the input video. The team additionally minimized RISP's rendering gradients, so that its predictions were less sensitive to changes in rendering configurations, allowing it to learn to forget about visual appearances and focus on learning dynamical states. This is made possible by a differentiable renderer.

The method then uses two similar pipelines, run in parallel. One is for the source domain, with known variables. Here, system parameters and actions are entered into a differentiable simulation. The generated simulation's states are combined with different rendering configurations into a differentiable renderer to generate images, which are fed into RISP. RISP then outputs predictions about the environmental states. At the same time, a similar target domain pipeline is run with unknown variables. RISP in this pipeline is fed these output images, generating a predicted state. When the predicted states from the source and target domains are compared, a new loss is produced; this difference is used to adjust and optimize some of the parameters in the source domain pipeline. This process can then be iterated on, further reducing the loss between the pipelines.

To determine the success of their method, the team tested it in four simulated systems: a quadrotor (a flying rigid body that doesn't have any physical contact), a cube (a rigid body that interacts with its environment, like a die), an articulated hand, and a rod (deformable body that can move like a snake). The tasks included estimating the state of a system from an image, identifying the system parameters and action control signals from a video, and discovering the control signals from a target image that direct the system to the desired state. Additionally, they created baselines and an oracle, comparing the novel RISP process in these systems to similar methods that, for example, lack the rendering gradient loss, don't train a neural network with any loss, or lack the RISP neural network altogether. The team also looked at how the gradient loss

impacted the state prediction model's performance over time. Finally, the researchers deployed their RISP system to infer the motion of a real-world quadrotor, which has complex dynamics, from video. They compared the performance to other techniques that lacked a loss function and used pixel differences, or one that included manual tuning of a renderer's configuration.

In nearly all of the experiments, the RISP procedure outperformed similar or the state-of-the-art methods available, imitating or reproducing the desired parameters or motion, and proving to be a data-efficient and generalizable competitor to current motion capture approaches.

For this work, the researchers made two important assumptions: that information about the camera is known, such as its position and settings, as well as the geometry and physics governing the object or person that is being tracked. Future work is planned to address this.

"I think the biggest problem we're solving here is to reconstruct the information in one domain to another, without very expensive equipment," says Ma. Such an approach should be "useful for [applications such as the] metaverse, which aims to reconstruct the physical world in a virtual environment," adds Gan. "It is basically an everyday, available solution, that's neat and simple, to cross domain reconstruction or the inverse dynamics problem," says Ma.

More information: RISP: Rendering-Invariant State Predictor with Differentiable Simulation and Rendering for Cross-Domain Parameter Estimation. openreview.net/forum?id=uSE03demja

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT

research, innovation and teaching.

Provided by Massachusetts Institute of Technology

Citation: Trained neural network pipeline simulates physical systems of rigid and deformable bodies and environmental conditions (2022, May 3) retrieved 19 April 2024 from <https://techxplore.com/news/2022-05-neural-network-pipeline-simulates-physical.html>

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.