Supplementary Material for VRGym: A Virtual Testbed for Physical and Interactive AI

Xu Xie Hangxin Liu Zhenliang Zhang Yuxing Qiu Feng Gao Siyuan Qi Yixin Zhu Song-Chun Zhu
UCLA Center for Vision, Cognition, Learning, and Autonomy
Los Angeles, CA 90095
{xuxie,hx.liu,zlz,yx.qiu,f.gao,syq,yixin.zhu}@ucla.edu,sczhu@stat.ucla.edu

1 RELATED WORK

1.1 Passive Dataset

Large-scale labelled datasets play an important role in current development of artificial intelligence (AI) and machine learning, e.g., ImageNet [7] and Microsoft COCO [27] have greatly facilitated the advancements in both object detection and classification. More recently, some datasets are beyond categorization tasks, e.g., Visual Genome [23] focuses on learning the relationship. However, manually labeling dataset is proven to be tedious and error-prone, limiting both its quantity and accuracy.

Synthetic image datasets have recently been a source of training data for object detection and correspondence matching [8, 11, 12, 29, 30, 34, 59], single-view reconstruction [17], view-point estimation [49], human pose estimation [10, 31, 41, 45, 46, 52, 57, 60], depth prediction [48], pedestrian detection [16, 28, 32, 53], action recognition [37, 38, 40], semantic segmentation [39], scene understanding [5, 13, 14, 47], and in benchmark data sets [15, 18]. Previously, synthetic imagery, generated on the fly, online, had been used in visual surveillance [36] and active vision / sensorimotor control [50]. However, these datasets can only afford passive observation or very limited interactions, thereby difficult to generalize to scenarios where an AI agent can interact with a human.

1.2 Simulation Platform

Table 1 summarizes detailed comparisons against similar simulation platforms, showing the uniqueness of VRGym.

Robotics simulation platforms originated from the Robotics Operating System (ROS) (e.g., Gazebo [21] and V-Rep [42]) have been playing an important role in robotics development. However, as these platforms focus on the robotics application, they rarely provide means for either the virtual reality (VR) integration or the human interactions.

Virtual training platforms including OpenAI Gym [3] and MuJoCo [51] are designed to evaluate and benchmark machine learning algorithms. Although these platforms allow easy-setup and fast training by providing interfaces with popular packages and games, they still lack sufficient levels of interactions, especially for physical agents.

Physics-based simulation platforms leverage the content developed by the game industry and incorporate sophisticated physics-based simulation. For instance, CARLA [9] is an open-source simulator for autonomous driving, whereas AirSim [44] provides a photorealistic rendering of outdoor scenes for drone navigation. They are, however, platforms for specific tasks, e.g., vehicle or drone navigation. More general-purpose platforms are also available, including AI2THOR [22] and Gibson [55]. Specifically, AI2THOR provides detailed 3D indoor scenes where AI agents can navigate in the scenes and interact with objects to perform tasks; however, there is no human embodiment, and most of the interactions are symbolic-level. Although virtual agents in Gibson can receive a constant stream of visual observations and a human can be represented as a Mujoco humanoid, it still lacks a sufficient level of manipulations and interactions for the human embodiment.

1.3 Domain Adaptation

Although the presented work does not directly involve domain adaptation, this plays a vital role in learning from virtual environments, as the goal of using virtual training is to transfer the learned model and apply it to real-world scenarios. A review of existing work in Table 1: Comparison with existing 3D virtual environments. Scale: Contains a large number of scenes. Physics: Supports physics-based simulation on agents and objects. Real: Provides a life-like rendering. Action: Object states can be changed by actions. Fine-grained: Enables fine-grained actions and simulates plausible object state changes. Human: Humanoid agents. Multi: Supports a multi-agent setting.

<table>
<thead>
<tr>
<th>Environment</th>
<th>Scale</th>
<th>Physics</th>
<th>Real</th>
<th>Action</th>
<th>Fine-grained</th>
<th>Human</th>
<th>Multi</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUNCG [47]</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Matterport3D [5]</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Malmo [19]</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>DeepMind Lab [2]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VizDoom [20]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>MINOS [43]</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HoME [4]</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gibson [55]</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>House3D [56]</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AI2-THOR [22]</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VirtualHome [33]</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>VRGym</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ACM TURC 2019, May 17–19, 2019, Chengdu, China
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-7158-2/19/05 . . $15.00
https://doi.org/10.1145/3321408.3322633
this area is beyond the scope of this paper; we refer the reader to a recent comprehensive survey [6].

1.4 VR for Cognitive Studies and Machine Learning

VR is capable of offering versatile settings for human training and testing, providing a convenient way for cognitive studies to quickly set up a specific scenario without building a costly physical apparatus. There are many successful cases; some recent work includes studying the deceptive behaviors [1], examining human physical judgments in abnormal environments [58], improving driving habits [24], designing the game level automatically [56], and training for earthquake [26].

VR is also a fast means to collect data and train for machine learning. Towards this goal, some researchers have built plugins for game engines, such as UETorch [25] and UnrealCV [35]. However, to date, such plugins only offer APIs to control game state and record data, requiring additional packages to train virtual agents.

In contrast, VRGym is capable of providing detailed logging of the data generated inside VR environment, as well as supporting training virtual agents directly for various tasks, subsuming the functions provided in previous virtual environments.

2 SYSTEM PERFORMANCE

VRGym runs in real-time (30fps) on a modern PC with an Intel 8700K CPU, a set of DDR4 memory totaling 64GB, and an EVGA GTX 1080 Ti GPU. Figure 1 shows the system performance running the physics-based simulation of complex manipulation using the human input devices, together with the software and hardware interface we develop. The results indicate that the VRGym is efficient on CPU and memory utilization. As the real-time physics-based simulation relies heavily on parallel computing, it requires relatively more GPU power. In general, VRGym can be supported by modern computers without requiring special setups or dependencies.

3 EVALUATION OF COMMUNICATION BANDWIDTH

A profile of the performance is shown in Figure 2, in which 20 packages are sent individually for each concurrent connection. The communication latency for the VRGym-ROS bridge increases from 0.04sec to 0.41sec. Linear regression is fitted to the mean of the latency $t_9 = 2.9025, p = 0.01, r^2 = 0.9998$, indicating a strong linear trend with respect to the increase of the concurrent connections in the VRGym-ROS bridge.
REFERENCES


