The Benefits of Immersive Demonstrations for Teaching Robots

Astrid Jackson[1,2], Brandon D. Northcutt[1], Gita Sukthankar[2]
[1]Toyota Research Institute, Los Altos, California
[2]Department of Computer Science, University of Central Florida, Orlando, Florida
{astrid.jackson, brandon.northcutt}@tri.global, gitars@eecs.ucf.edu

***Abstract*—One of the advantages of teaching robots by demonstration is that it can be more intuitive for users to demonstrate rather than describe the desired robot behavior. However, when the human demonstrates the task through an interface, the training data may inadvertently acquire artifacts unique to the interface, not the desired execution of the task. Being able to use one's own body usually leads to more natural demonstrations, but those examples can be more difficult to translate to robot control policies.**

This paper quantifies the benefits of using a virtual reality system that allows human demonstrators to use their own body to perform complex manipulation tasks. We show that our system generates superior demonstrations for a deep neural network without introducing a correspondence problem. The effectiveness of this approach is validated by comparing the learned policy to that of a policy learned from data collected via a conventional gaming system, where the user views the environment on a monitor screen, using a Sony Play Station 3 (PS3) DualShock 3 wireless controller as input.

***Index Terms*—Learning from Demonstration; Imitation Learning; Robot Manipulation; Virtual Reality; User Study**

I. Introduction

Learning from demonstration is a powerful and versatile framework for robot skill acquisition; however more research attention has been devoted to improving the learning methodology, without considering the quality of the demonstration data. Popular choices of acquiring demonstrations are teleoperation [1] and kinesthetic teaching [2]–[5] since these transfer quite naturally between teacher and robot. Chernova and Thomaz assert that "in many situations it is more effective or natural for the teacher to perform the task using their own body" [6]. However, there is an inherent trade off between the naturalness of the interface for the user and the ease of translation to the robot; depending on the robot's kinematics, a direct mapping from the teacher's state and actions to those of the robot rarely exists. Despite this limitation, our hypothesis is that more natural demonstrations are superior for training deep neural network architectures that can learn complex transfer embeddings.

We propose the use of a mainstream Virtual Reality (VR) system, to collect high-quality demonstrations for manipulation tasks. A human operator uses a VR headset to step inside the 3D environment and perceives the VR controller as the robot's gripper. Thus the robot's gripper becomes a natural extension of the human operator's arm. This setup combines advantages of both teleoperation and use of one's own body since recordings come directly from the robot's sensors, i.e. moving the robot's gripper via motion-tracked VR controller allows for intuitive and natural movements.

In this paper, we evaluate the effects of utilizing the HTC Vive VR system to gather demonstrations for robot learning. These demonstrations are then used to train a deep neural network robot control policy. We compare this policy with a policy learned using the same deep neural network but with data collected using a conventional gaming system, where the user views the environment on a monitor screen and uses a PS3 controller instead. We hypothesize the quality of the demonstrations in VR is higher and thus leads to a policy that is more effective than the policy learned utilizing data acquired using a conventional gaming system. Furthermore, we believe that not only do users need less training in order to provide high-quality data when demonstrating the task in VR, but as the complexity of the task increases performance becomes unwieldy if not impossible when executing the task in a conventional gaming system while it is still manageable in VR. For videos and supplemental materials please see our website[1].

The next section presents an overview of prior work on data collection for robot learning from demonstration, before introducing the smoothness score which is used for trajectory evaluation. Then we describe the experimental setup and our user study quantifying the benefits of VR for task execution. The following section presents an evaluation of policies learned from this data. We conclude with a summary of our findings and future work.

II. Related Work

Collecting data suitable for learning robot manipulation tasks from demonstrations is difficult. Using one's own body to demonstrate a task is probably the most intuitive form of collecting demonstration data. The motion is usually recorded by equipping the teacher with wearable sensors, such as motion capture systems and inertial sensors [7]–[10]. Those devices provide high quality in the observations but due to cost find only restricted use outside of laboratories. In addition, it can be difficult to find a mapping between the human demonstrator and the robot [11].

[1] https://sites.google.com/view/immersive-demonstrations

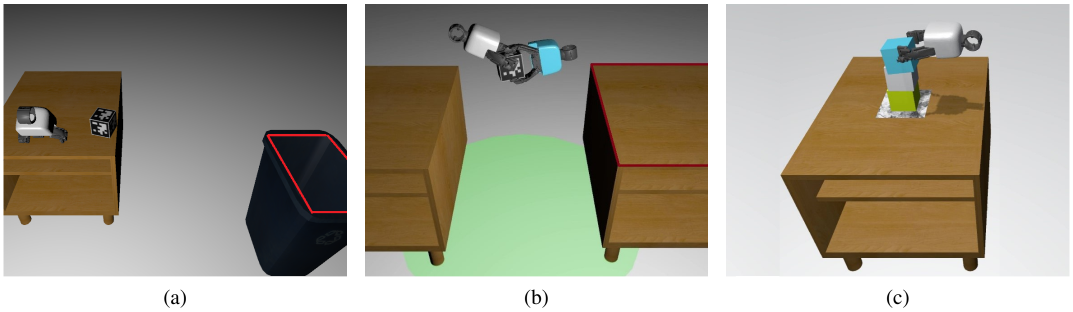


Fig. 1. An overview of the three manipulation tasks. Participants of the user study had to perform each task in conventional gaming system (on a monitor screen and controlling a PS3 controller) and within a mainstream virtual reality system (a) Clean the room by dropping the box in the trash bin. (b) Hand the box from one gripper to another and place it on the other table. (c) Stack the boxes in a specific order.

Kinestatic teaching avoids this problem by recording the trajectories from the robot's own sensors while manipulating the joints through physical force [2]–[4]. However, this method of teaching is not always intuitive and can introduce unwanted artifacts. Teleoperating the robot via a joystick is another form of recording trajectories from the robot's own sensors [1], [5], [12]–[14], but can similarly result in collecting unwanted artifacts unique to the interface.

Operating the robot in virtual reality benefits from the natural motions of the operator without introducing a correspondence problem. Animation performance research investigates the use of virtual reality in order to control a character via one's own body [15]–[19]. The focus of this research is the real-time mapping between the human motion to that of the virtual character. Rather than puppeteering a character, we are concerned with learning a policy that allows the robot to perform the task without further human input. Furthermore, the animation community has not looked at quantifying the benefits of using virtual reality. Zhang et al. [20] collect RGBD data while teleoperating a robot in virtual reality to train a convolutional neural network for solving a wide variety of manipulation tasks. Prior to this work, virtual reality was mostly limited to collecting waypoints of low-dimensional robot states [21], [22].

The benefits and disadvantages of demonstrating a task in a virtual environment rather than in a real world setting are investigated in [23]. The task is demonstrated with the use of a dataglove which tracks the body motion in virtual reality. However, the authors are not concerned with how the data would affect the learned policy nor did they compare the data of different input controls. Whitney et al. [24] compare the user experience of teleoperating a robot to perform a cup-stacking task using different modalities including the use of keyboard and monitor and a virtual reality system. No policy is learned from the data captured. The ease of use is also evaluated by Gadre et al. [25] who performed a user study on a visual programming task using an AR (HoloLense) system to program a robot for a pick-and-place task. A policy was learned from the waypoints set via the AR system using LfD algorithms, though neither the data nor the learned policies are analyzed. Koganti et al. [26] learn a visual attention system in virtual reality by simulating attention (i.e., glimpse) via VR vision by using two cameras which simulate narrow field of view (FOV) and higher visual acuity (resolution per degree). They provide a brief comparison of the game performance, showing that the performance of the player increases with the use of virtual reality over the use of a joystick or keyboard. However, the authors do not provide an evaluation of the acquired data or the performance of the resulting model. In [27] users can control a virtual robot amongst real entities using augmented reality. In this setting the user manipulates the end effector of the robot in virtual reality to set start and intermediate goal points from which a collision-free path can be generated. To evaluate the quality of their system the output curves (i.e., trajectories) are being compared against known desired curves. No user study is performed to evaluate the ease of use of the system.

III. Background

A. Normalized Jerk

This paper utilizes normalized jerk as the quantitative measure of trajectory smoothness. The calculation of normalized jerk was formalized by Teulings, Contreras-Vidal, and Stelmach [28] when investigating movement coordination in Parkinson's disease subjects. However, it translates to any kind of motion quite naturally.

Jerk is the rate of change in acceleration, and thus the third time derivative of position. The smoother the movement, the smaller the time-integrated squared jerk. Since jerk is influenced by time, it must be normalized for different trajectory durations. Formally,

$$\text{Normalized Jerk} = \sqrt{1/2 \int j^2(t) \frac{\Delta T^5}{s^2} dt} \qquad (1)$$

where ΔT and s are the duration and distance of the demonstration, respectively.

B. Mixture Density Networks

Mixture density networks (MDN) as described in [29] belong to a class of neural networks allowing for multi-valued outputs. They consist of a feed-forward neural network whose outputs determine the parameters of a probability density

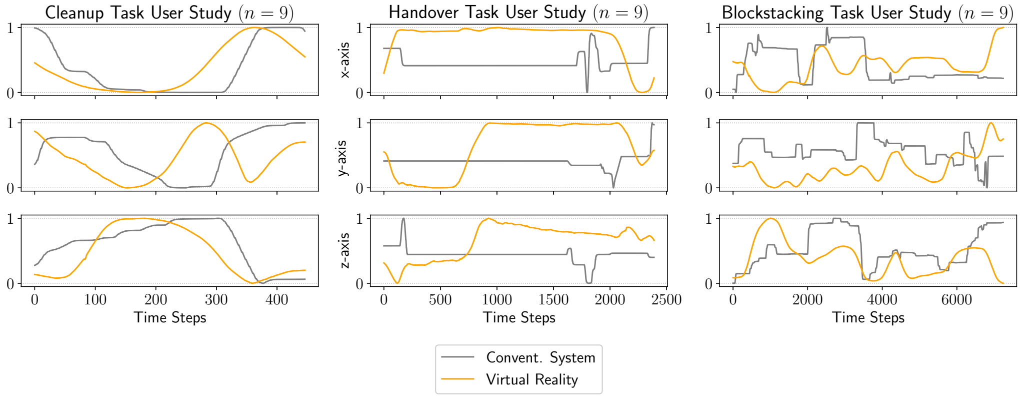


Fig. 2. Visual comparison of randomly selected trajectories generated by user study participants. The trajectories were collected at 30 Hz and normalized to account for differences in trajectory lengths and axis values. Trajectories gathered by participants within the virtual reality system are significantly smoother and with more continuous motion than those gathered via a conventional gaming system.

function, which is represented as a linear combination of kernel functions in the form

$$p(y|x) = \sum_{i=1}^{m} \alpha_i(x)\phi_i(y|x) \tag{2}$$

where m is the number of components in the mixture, $\alpha_i(x)$ are the mixing coefficients which can be considered the prior probabilities, and $\phi_i(y|x)$ are multivariate Gaussian functions. The Gaussian kernel function considered in this paper is of the form

$$\phi_i(y|x) = \frac{1}{(2\pi)^{c/2}\sigma_i(x)^c} exp\left\{ - \frac{||y - \mu_i(x)||^2}{2\sigma_i(x)^2} \right\} \tag{3}$$

where the vector $\mu_i(x)$ represents the center of the i^{th} kernel.

The total number of network outputs is given by $(c+2) \times m$. The centers μ_i embody *location* parameters and are represented directly by $c \times m$ network outputs. The variances σ_i represent *scale* parameters and are given in terms of the exponential of the m corresponding network outputs. To satisfy the constraint $\sum_{i=1}^{m} \alpha_i(x) = 1$ the m network outputs corresponding to α_i are passed through a softmax function.

The loss function used for training is the negative logarithm of the likelihood in the form

$$\mathcal{L} = -\ln \left\{ \sum_{i=1}^{m} \alpha_i(x)\phi_i(y|x) \right\} \tag{4}$$

with $\phi_i(y|x)$ given by Equation 3.

IV. Experimental Setup

All experiments were performed in a virtual environment that we created with the Unity3D game engine for performing virtual manipulation tasks. Our experiments focus on three successively harder tasks.

In the *cleanup task* (see Figure 1a), the user picks up a box from a side table and drops it into a trash bin adjacent to the table. The box is randomly placed and oriented on the table. To complete this task the gripper must be moved from its initial random location to a position close to the box, opened and rotated such that the fingers fit around the box, and then closed. With the box firmly held by the gripper, the gripper must now be moved to the trash bin until the box lines up with its opening. Finally, the gripper must be opened to release the box. The time limit for the execution of this sequence is one minute.

The second task is the *handover task* (see Figure 1b) in which the user controls two grippers to pick up a box from the side table on the left with one gripper, hand it over to the other gripper, and then place it on the table on the right. Similar to the previous task, the box is placed and rotated randomly on the table. Once the gripper has successfully picked up the box, as described in the previous task, it must be moved to a position and rotation suitable for handing over the box to the other gripper. The second gripper must now be moved to a position close to the box which is held by the first gripper, opened and rotated such that the fingers fit around the box but without colliding with the gripper holding the box, and then closed around the box. Once the box is firmly held by the second gripper, the first gripper must open to release the box, and move out of the way. Finally, the second gripper can move the box to the table on the right and open to release the box. The time limit to execute the specified sequence is one minute.

The final task is *blockstacking* (see Figure 1c) in which the user is presented with three boxes that must be stacked in a predetermined order. All three boxes are placed in random locations and rotations on the table. First the gripper must pick up the yellow box and place it on the silver sheet located in the middle of the table. Next the gripper must pick up the white box, line it up with the yellow box and place it carefully on top of the yellow box. Finally, the gripper must pick up the blue box, carefully line it up with the top of the white box, and open the gripper to release the box. The blocks tower must remain standing for one second to be considered a successful execution of the task. Two minutes are allocated for performing this sequence.

There are two major sources of difficulty in demonstrating these tasks. The first issue is the precision required to grasp a box. Since the size of the box is only slightly smaller than the widest span of the open gripper, great care must be taken in orienting the gripper around the box to perform a stable grasp. The other difficulty is the possibility of the box to enter an unreachable state. The box may become unreachable by being pushed or knocked off the table.

The failure criteria in all described tasks is twofold:

1) The box does not reach its target location before the allocated time has run out, or
2) Any of the boxes drops on the ground, either by knocking it off the table or while carrying it to its target location.

In order to be able to perform the previously described sequences with a conventional gaming system, the PS3 controller is configured to allow for both linear (left and right stick) and angular (D-Pad and left and right bumper) motion of the gripper. Furthermore, the user needs to be able to open and close the gripper (right trigger button) and, during the handover task, switch between grippers (Y-button).

The controls for the VR are much simplified since linear and angular motion are controlled by the user's hand motion and during the handover task the user is able to use both hands to control a gripper. The only button mapping required for opening and closing the gripper was achieved by the trigger button on the VR controller.

V. User Study: Setup and Analysis

Our central hypothesis is that a VR system makes it significantly easier to capture high-quality data, which in turn positively affects the learned control policy. To test our hypotheses we conducted a user study in which the participants performed each task using both systems, i.e. the VR system and the conventional gaming system. To control for learning effects, the order of task and system presentation was randomized. Each task-system pair was performed for 5 minutes, resulting in a total of 30 minutes each participant was engaged with performing tasks.

Dependent measures. Trajectory quality is determined by the normalized jerk (Equation 1) as smoothness score. Performance is measured using the dependent factors of success rate and time for successful task completion in minutes. As a subjective measure we asked each participant to give feedback on their experience for each task to gauge their perceived differences in the system. We collected this feedback via a questionnaire. For each question the subjects were asked to rate their experience on a 7-point Likert scale, where 1-Strongly Disagree, 4-Neither Agree or Disagree, and 7-Strongly Agree.

Subject allocation. We recruited two female and seven male participants primarily ranging from 35-44 years old. As determined by a preliminary questionnaire, one of the participants did not have prior experience with a VR system. All others reported familiarity in both systems. Each participant was provided with a brief outline of the tasks and a short practice period, approximately one minute, before each task and system to familiarize themselves with the task and controls.

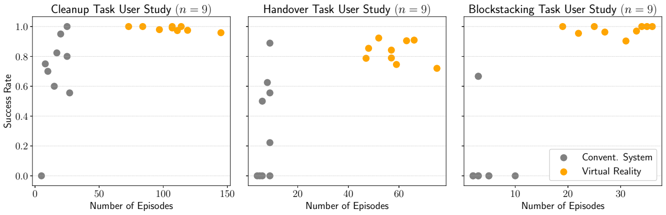


Fig. 3. Evaluation of human subjects performance on manipulation tasks of various difficulty. Success rates averaged over the number of episodes performed over a 5 minute period by each subject. As the difficulty of the task increases the benefit of using the virtual reality system becomes more pronounced as is evidence in the success rate and the number of performed episodes.

TABLE I
Evaluation of the demonstrated trajectories smoothness. The trajectories were generated by 9 participants during task execution. A Mann-Whitney U test was applied to comparatively assess the two groups. Medians for the normalized jerk are reported below.

Task	U, z	p-value	Conv. Sys.	VR
Cleanup	$5494, -18.33$	<0.0005	0.091	**0.031**
Handover	$5174, -9.16$	<0.0005	0.057	**0.015**
Blockstacking	$1503, -6.63$	<0.0005	0.113	**0.025**

Analysis of the trajectory smoothness. We predicted that the motion of the gripper captured within the virtual reality would be smoother and more continuous as compared to the data generated when handling a PS3 controller. Figure 2 provides a visual comparison of trajectories generated during the user study. From the pool of successfully executed episodes the trajectories were selected randomly and normalized to account for different lengths and axis values. Using the PS3 controller users favored moving along a single degree of freedom at a time, as apparent by the discontinuity of the trajectory, whereas they felt comfortable moving along multiple degrees of freedom simultaneously using the VR system. This is likely due to the motion of the VR controller being much closer to natural motion of the hand combined with the depth perception afforded by the stereoscopic view in the VR headset.

As a quantitative measure for smoothness we calculated the normalized jerk, including the trajectories of all nine participants, and ran a Mann-Whitney U test to determine if there were differences in the smoothness score between the trajectories gathered in VR and the conventional gaming system. As can be seen in Table I, we found that in all three scenarios the smoothness score was statistically significantly different (p<0.0005) with the VR trajectories being smoother than those from the conventional gaming system. Thus, we were able to validate our visual observation.

A. Analysis of the Cleanup task

Hypothesis. We hypothesize that it is easier to learn the controls required to perform the task well in VR than it is with a conventional gaming system.

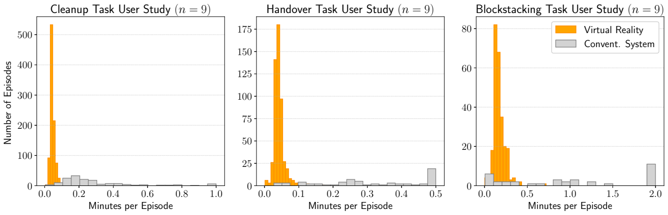


Fig. 4. Time in minutes to perform the task. Participants are able to perform the task within the virtual reality system more quickly than using a conventional gaming system, enabling them to gather more samples in the same amount of time.

Analysis. A one-way repeated measures ANOVA was conducted to determine whether there was a statistically significant difference in success rate and time spent to complete the cleanup task. Performing the task in VR elicited statistically significant changes in both success rate, $F(1,8) = 8.701, p = 0.018$ and task completion time, $F(1,8) = 15.340, p = 0.004$. As can be seen in Figure 3 participants succeeded notably more often when immersed in the environment through a VR system with an average success rate of 0.986 ± 0.0154 as opposed to 0.687 ± 0.296 using a conventional gaming system. Furthermore, participants were able to perform significantly more task repetitions (i.e., episodes). This can be explained by the fact that they spend less time executing an episode (see Figure 4). The average time to complete the cleanup task increased from 0.05 ± 0.010 minutes when using the VR system to 0.39 ± 0.270 minutes when using a PS3 controller.

Prior to the user study, participants rated their proficiency level for each system on a 4-point proficiency scale, where 1-Novice, 2-Intermediate, 3-Advanced, and 4-Expert. A one-way repeated measures ANOVA revealed a statistically significant difference in the reported proficiency level, $f(1,7) = 5.727, p<0.048$ with the participants presenting a higher proficiency level in the use of a conventional gaming system (2.50 ± 1.069) than in the use of a virtual reality system (1.75 ± 0.463).

To gain insight as to why participants performed better with the VR system though their proficiency level was rated lower, we evaluated the subjective measure, rated on a 7-point Likert-scale, where 1-Strongly Disagree, 4-Neither Agree or Disagree, and 7-Strongly Agree. A one-way repeated measures ANOVA revealed that participants felt they intuitively understood the mechanics of the controls, $F(1,8) = 64.222, p = 0.008$ when using the VR controller (6.33 ± 0.323) while that was not the case for the PS3 controller (2.56 ± 2.128). In addition, participants stated that the controls were easy to handle, $F(1,8) = 26.597, p = 0.001$ within the virtual reality (6.56 ± 0.726) but did not agree this was the case when handling the PS3 controller (3.00 ± 2.179). No statistically significant difference in the feeling of fatigue, nausea, or disorientation between the two system was found.

B. Analysis of the Handover task

Hypothesis. We hypothesize that it is easier to learn the controls required to perform the task well in VR than it is with a conventional gaming system. In addition we hypothesize that users prefer performing the task in VR.

Analysis. Figure 3 shows a clear increase in difficulty in performing the handover task. Four out of nine participants did not once complete the task successfully with the conventional gaming system. Though the success rates were not quite as high as for the cleanup task participants still performed the handover task significantly better using the VR system. This observation was supported by a one-way repeated measures ANOVA showing a statistically significant difference in the success rate, $F(1,8) = 24.346$, $p<0.001$ with the success rate increasing from 0.31 ± 0.340 using the conventional gaming system to 0.83 ± 0.074 with the VR system. The time to perform the task with the conventional gaming system (0.68 ± 0.18) was also significantly longer, $F(1,8) = 88.753$, $p<0.0005$ than in VR (0.087 ± 0.013).

Evaluation of the subjective measures on a 7-point Likert scale with a one-way repeated measures ANOVA showed that similar to the cleanup task the participants had an intuitive understanding of the controller mechanics, $F(1,8) = 23.583, p = 0.001$ when using the VR controller $(5.67 \pm 2,121)$ but not when using the PS3 controller (2.22 ± 1.394). With a statistically significant difference of $F(1,8) = 6.983, p = 0.030$, they also felt that the PS3 controller was harder to handle (2.78 ± 2.167) than the VR controller (5.56 ± 1.878). In addition they reported that they enjoyed performing the task, $F(1,8) = 8.909, p = 0.017$ more when performing it in VR (5.44 ± 1.33) as opposed to with the conventional gaming system (3.11 ± 2.224). No statistically significant difference in the feeling of fatigue, nausea, or disorientation between the two system was found.

With the conventional gaming system, in order to switch from one gripper to the other, the gripper that was currently not under user control needed to be suspended in air. This stabilized the uncontrolled gripper to some extent. This was not the case in VR, where both gripper were constantly in motion based on the user's hand movements. Therefore, for the handover to be successful, the user had to be careful to hold his hands as still as possible. However, this disadvantage was not enough to lessen the impact greater depth perception and easier controls had on the end result.

C. Analysis of the Blockstacking task

Hypothesis. We hypothesize that it is easier to learn the controls required to perform the task well in VR than it is with a conventional gaming system. Additionally, we hypothesize that with increasing complexity of the task, execution of the task becomes unwieldy if not impossible with the conventional gaming system, while still manageable in VR.

Analysis. The difficulty of performing the blockstacking task with the conventional gaming system as compared to the VR system is limited depth perception and lack of fine motor control. Without the presence of either plus the complicated controls users were unable to successfully execute the task (see Figure 3). For a quantitative evaluation we conducted a one-way repeated ANOVA and found a statistically significant

TABLE II
PERFORMANCE OF THE LEARNED POLICY. SUCCESS RATES AND MINUTES PER TRIAL OF THE LEARNED POLICY AVERAGED OVER 5 SESSIONS OF 100 TRIALS. p-VALUES WERE DETERMINED BY A PAIRED-SAMPLES T-TEST.

	Cleanup Task			Handover Task		
	p-value	Convent. System	Virtual Reality	p-value	Convent. System	Virtual Reality
Success Rate	0.002	0.05 ± 0.021	$\mathbf{0.18 \pm 0.031}$	0.035	0.0 ± 0.0	$\mathbf{0.02 \pm 0.012}$
Minutes Per Trial	<0.001	0.35 ± 0.325	$\mathbf{0.22 \pm 0.211}$	0.002	0.29 ± 0.207	$\mathbf{0.19 \pm 0.169}$

difference in the success rate, $F(1,8) = 72.742, p < 0.0005$ with an increase in the rate of success from 0.15 ± 0.284 when using the conventional gaming system to 0.98 ± 0.034 with the VR system. A statistically significant difference was also noted for the episode completion time, $F(1,8) = 40.194, p < 0.0005$, which decreased from 1.25 ± 0.512 minutes with the conventional gaming system to 0.18 ± 0.045 minutes with the VR. Figure 4 shows that with the conventional gaming system participants mostly failed immediately by knocking boxes off the table or they timed out. In VR task completion times remain consistent.

Evaluation of the subjective measures with a one-way repeated measures ANOVA revealed that the enjoyment in performing the task, $F(1,8) = 14.593, p = 0.005$ increased from 3.33 ± 2.297 with the conventional gaming system to 5.89 ± 1.537 in VR. This is probably in strong correlation with the level of frustration, $F = (1,8) = 14.089, p = 0.006$ which decreased from 5.00 ± 2.398 using the PS3 controller to 2.00 ± 1.581 using the VR controller. Furthermore, we discovered that the participants felt their performance was not improving over time, $F = (1,8) = 10.557, p = 0.012$ with the conventional gaming system (4.11 ± 2.205) however, they did feel they improved with the VR system (5.89 ± 1.054). Similar to the other tasks they found the control mechanics more intuitive, $F(1,8) = 23.69, p = 0.001$ for the VR system (6.67 ± 0.707) than for the conventional gaming system (3.44 ± 2.007) and in general believed the VR controller (6.56 ± 0.726) easier to handle, $F(1,8) = 28.000, p = 0.001$ than the PS3 controller (2.67 ± 2.121). No statistically significant difference in the feeling of fatigue, nausea, or disorientation between the two system was found.

User study participants struggled immensely to line up the boxes when performing the task with the conventional gaming system and commented on the fact that the game physics of the blocks were not conducive for stacking the boxes. However, they did not have any problems with the task when immersed in the environment with the VR system. As a matter of fact, participants remarked on how easy it is. Considering that the blocks physics is the same in both systems, it can be deduced that task environments are more forgiving in terms of physics and other game engine specific mechanics when used with the VR system as opposed to the conventional system. That means, less time can be spent on fine tuning an environment without sacrificing the quality of the demonstration data.

VI. Evaluation of the Learned Policies

The effects of superior demonstrations on the learned policy were evaluated by passing the demonstration trajectories collected during the user study to a neural network for training. Demonstrations were recorded at 30 Hz and collected independently for both the conventional gaming system and the VR system. Prior to training the trajectories were sub-sampled at 4 Hz. We used the staggered sub-sampling technique to generate multiple low-frequency trajectories from each demonstration. Failed demonstrations were excluded from the training data.

The input $x_t = (p_t^{1:n}, s_t^{1:n}, q_t^{1:k})$ at time t to the neural network includes (a) the absolute pose p_t of the $n \in [1,2]$ grippers in the simulated environment, (b) the open/close state s_t of each gripper, and (c) the absolute pose q_t of the $k \in [1,3]$ objects being manipulated.

The neural network architecture closely follows [12]. The observation x_t is first passed through three consecutive LSTM layers capable of extracting changes in the objects poses over time. This is followed by an MDN for predicting the target pose and state of the $n \in [1,2]$ grippers.

The output of the neural network is a probability density function of the target pose and state of the $n \in [1,2]$ grippers.

Depending on the average demonstration sequence length in each category the network was unrolled for different time steps. The network was optimized using RMSProp [30] with default learning rate of 1e-3 and decay rate of 0.999. Initial values of network parameters were uniformly sampled from [-0.08, 0.08]. Once the validation error remained unchanged for 20 epochs training stopped and the best performing network was used for policy evaluation.

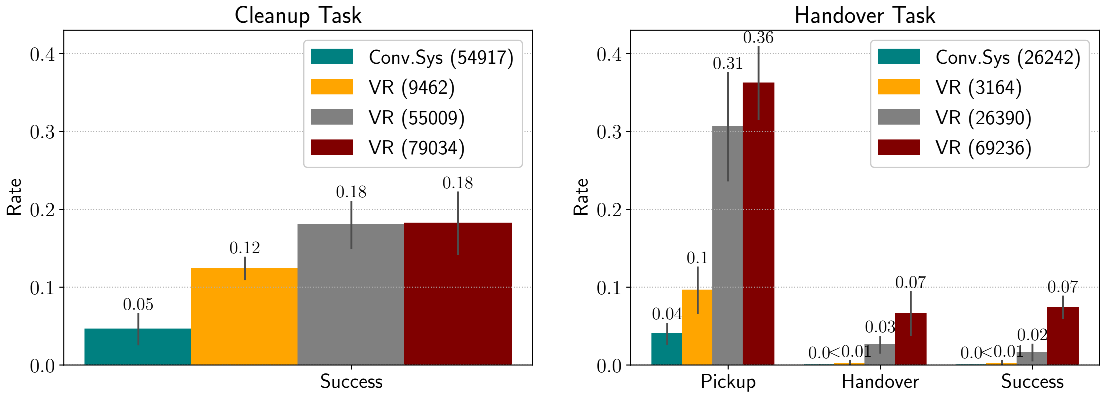
During evaluation at each time step t the successive target pose and state of the $n \in [1,2]$ grippers was drawn from the probability density function derived by the neural network.

A. Evaluation of Policy Performance

Hypothesis We hypothesize that higher quality in demonstrations directly translates into better performance of the learned policy.

Analysis. Table II show a clear qualitative and quantitative benefit of collecting demonstrations with a VR system[2]. We conducted a paired-samples t-test to identify if there were statistically significant differences in the success rate and the

[2] Since not enough data points were collected for the blockstacking task to learn a reasonable policy we refrain from evaluating the performance of the policies for this task.



(a) (b)

Fig. 5. Comparison of the learned policies. Results averaged over 5 sessions of 100 trials. Given in parenthesis (see legend) are the number of data points used for learning. The highest number represents the total number of data points collected at 30 Hz, the second highest number approximates the number of data points to the number of data points collected using a conventional gaming system, and the lowest number is the data points remaining when matching the data collected using the virtual reality system to the same number of successful episodes achieved with the conventional gaming system. Even with far less data points feeding into the machine learning system, the policy learned from data collected within the virtual reality system outperforms the policy learned from data collected with the conventional gaming system. (a) Success and failure rates for the cleanup task. (b) Success and failure rates and partial successes for the handover task.

task execution time. We found that in both the cleanup task and the handover task the success rate of the learned policy is statistically significantly higher for the policy learned from VR demonstrations than that of the policy learned from the PS3 controller demonstrations. For the cleanup task the percent change is +260%. This indicates that in general the learned policy from the VR demonstrations is capable of maneuvering the gripper more precisely. Also notable is the fact that the task can generally be executed in significantly less time when the data was collected with the VR system. For example, each demonstration in the cleanup tasks takes on average only 0.22 minutes as opposed to 0.35 minutes when the policy was learned using demonstrations collected with the PS3 controller. A similar comparison can be made for the execution time of the handover task.

PS3 controller demonstrations resulted in more failures and generally took longer to execute. Therefore, PS3 controller training data resulted in less data points than VR training data. To compare the success rates of policies learned from PS3 controller demonstrations to policies learned from VR demonstrations we consider learning policies by passing different number of data points to the neural network for training (see Figure 5). In particular, we were interested in the comparison of the performance when matching the VR training data to (a) the number of data points and (b) to the number of episodes of the PS3 controller training data. As can be seen in Figure 5a the success rate of the cleanup task did not change significantly when approximating the VR training data to the same number of data points. After matching to the same number of episodes, the VR training data consisted of only a sixth of the PS3 controller training data. Nevertheless, it outperformed the policy from the PS3 controller demonstrations by 0.07%. That means that in the cleanup task with a success rate of 98.6% and an average task execution time of 0.05 minutes experts only have to spend 5.84 minutes of task demonstrations to outperform the policy that was learned from using all PS3 controller demonstrations – a time commitment of 45 minutes. Similar results, though not as pronounced, can be seen for the handover task.

To gain further insights into the performance of each policy we consider partial successes. In the context of the cleanup task we were interested whether the gripper is able to pick up the box before time runs out or the box is dropped on the ground. We found that all policies outlined in Figure 5a never dropped the box once it was successfully picked up. This highlights the importance of the policy being able to precisely position the gripper.

In context of the handover task we considered the following partial successes:

1) The gripper is able to pick up the box but drops it on the ground or time runs out before the box is successfully handed over.
2) The gripper is able to handover the box to the other gripper, but the box drops on the ground or time runs out before the box is successfully placed on the side table.

Figure 5b shows that the policy *Conv. Sys (26242)* was unable to perform a successful handover, though the gripper was able to pick up the box $(4.0 \pm 1.4)\%$ of the time. Comparing this to the results of the policy *VR (26390)*, which approximates the number of data points to the PS3 controller training set, we find that the gripper is able to handover the box a total of $(4.2 \pm 1.6)\%$ of the time. Before failing the task the gripper is able to pick up the box with a rate of $(30.6 \pm 7.0)\%$. *VR (3164)*, which matches the VR training set to the number of episodes of the PS3 controller training data, outperforms the policy learned from PS3 demonstrations with

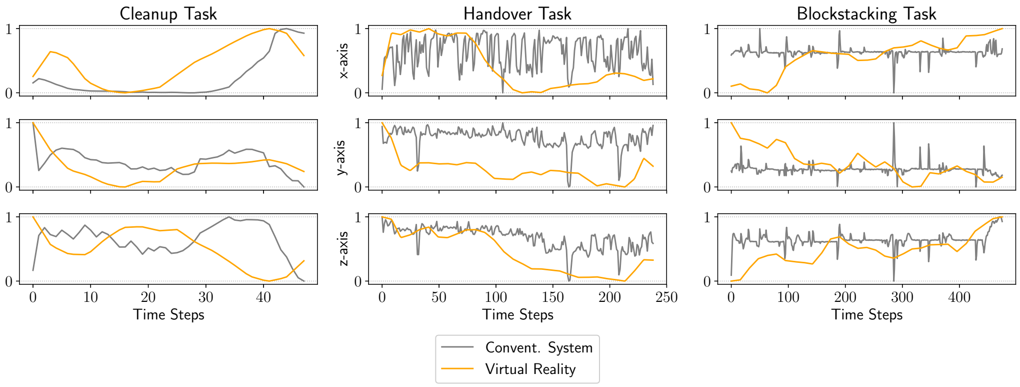


Fig. 6. Visual comparison of randomly selected trajectories generated by the learned policy. The trajectories were sampled at 4 Hz and normalized to account for differences in trajectory lengths and axis values. Smoothness and more continuous input trajectories correlate with smoother and more continuous output trajectories.

a total handover rate of $(0.4 \pm 0.05)\%$ and a pickup rate of $(9.6 \pm 3.0)\%$.

B. *Evaluation of Trajectory Smoothness*

Hypothesis. We hypothesized that the smoothness and continuity of the demonstrated trajectories leads to smoother and more continuous sampled trajectories.

Analysis. A visual comparison of the resulting trajectories is depicted in Figure 6. The trajectories were selected randomly from the pool of sampled trajectories generated by the learned policies and normalized to account for differences in lengths and positions. Visually, the trajectories for the handover task and the blockstacking task appear smoother for the policy learned from the VR data than for the policy learned from the conventional gaming system. It is inconclusive whether there is a difference in the trajectories for the cleanup task from visual observation alone.

TABLE III
ASSESSMENT OF THE SAMPLED TRAJECTORY SMOOTHNESS. THE NORMALIZED JERK WAS CALCULATED FOR 500 TRIALS AND COMPARATIVELY EVALUATED BY A MANN-WHITNEY U TEST. MEDIANS FOR THE NORMALIZED JERK ARE REPORTED BELOW.

Task	U, z	p-value	Conv. Sys.	VR
Cleanup	$88239, -35.87$	<0.0005	**0.546**	0.638
Handover	$32820, -47.82$	<0.0005	1.258	**0.591**
Blockstacking	$30563, -48.05$	<0.0005	1.269	**0.645**

A Mann-Whitney U test on the calculated normalized jerk (Equation 1) of the trajectories revealed that there is a statistically significant difference in the smoothness score in all three scenarios. Interestingly, the results in Table III indicate that for the cleanup task the trajectories generated by the policy learned from the conventional gaming system data are smoother than those generated by the policy learned from the VR data. Considering the number of failed trials and that trials failed within 0.01 ± 0.325 minutes we concluded that this could be explained by many short trajectories. Therefore, we evaluated the normalized jerk considering only successful trajectories. Median smoothness score for the policy derived from conventional gaming system data (0.639) and the policy learned from VR data (0.632) was not statistically significantly different, $U = 976, z = -7.80, p = 0.3383$. However, evaluation of the normalized jerk confirmed that the trajectories generated from policies learned from VR data are smoother in both the handover task and the blockstacking task.

VII. CONCLUSION

The results show that the VR environment provides an intuitive and easy means to demonstrate tasks of increasing complexity with a shallow learning curve. This is supported by the fact that the demonstrations acquired in VR produce more continuous and natural motions than those acquired via the PS3 controller. Users are less frustrated by the controls and enjoy performing the task more in the VR environment. That means it is easier to recruit and motivate expert demonstrators. Furthermore, less time can be devoted on setting up a task environment when used with the VR system since the mechanics are much more forgiving.

The VR demonstrations result in learned policies that are more efficient both in execution time and success rate. In particular, we were able to show that with VR demonstrations significantly fewer data points are required for training. As the complexity of the task increases, a policy could be learned from VR demonstrations when other methods of data collection failed.

We believe that these findings remain true even if traditional machine learning algorithms are employed for policy learning. We are of the opinion that a learned policy can only be as good as the data it derived from. Therefore, we conjecture that the superior VR data will always lead to better performing policies.

In future work we will transfer the learned policy onto the physical robot to prove that these findings are still valid when performed in the real world. We also plan to teleoperate the robot during the demonstration acquisition phase in VR to increase the accuracy of the demonstrations. Furthermore, we will investigate tasks that require obstacle avoidance.

References

[1] P. Abbeel, A. Coates, M. Quigley, and A. Y. Ng, "An application of reinforcement learning to aerobatic helicopter flight," pp. 1–8, 2006.

[2] S. Calinon and A. Billard, "What is the Teacher's Role in Robot Programming by Demonstration? - Toward Benchmarks for Improved Learning," *Interaction Studies. Special Issue on Psychological Benchmarks in Human-Robot Interaction*, vol. 8, no. 3, pp. 441–464, 2007.

[3] B. D. Argall, E. L. Sauser, and A. G. Billard, "Tactile Guidance for Policy Adaptation," *Foundations and Trends in Robotics*, vol. 1, no. 2, pp. 79–133, 2010.

[4] B. Akgun, M. Cakmak, K. Jiang, and A. L. Thomaz, "Keyframe-based Learning from Demonstration," *International Journal of Social Robotics*, vol. 4, no. 4, pp. 343–355, November 2012.

[5] A. D. Dragan, S. Siddhartha Srinivasa, and K. Kenton Lee, "Teleoperation with Intelligent and Customizable Interfaces," *Journal of Human-Robot Interaction*, vol. 2, no. 2, pp. 33–57, June 2013.

[6] S. Chernova and A. L. Thomaz, "Robot Learning from Human Teachers," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 8, no. 3, pp. 1–121, April 2014.

[7] J. Aleotti and S. Caselli, "Grasp recognition in virtual reality for robot pregrasp planning by demonstration," in *Proceedings 2006 IEEE International Conference on Robotics and Automation*. IEEE, 2006, pp. 2801–2806.

[8] A. Billard, Y. Epars, G. Cheng, and S. Schaal, "Discovering imitation strategies through categorization of multi-dimensional data," in *Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems*, vol. 3. IEEE, 2003, pp. 2398–2403.

[9] J. Koenemann, F. Burget, and M. Bennewitz, "Real-time imitation of human whole-body motions by humanoids," in *2014 IEEE International Conference on Robotics and Automation*. IEEE, May 2014, pp. 2806–2812.

[10] R. Rahmatizadeh, P. Abolghasemi, L. Bölöni, and S. Levine, "Vision-Based Multi-Task Manipulation for Inexpensive Robots Using End-To-End Learning from Demonstration," Tech. Rep., 2017.

[11] B. D. Argall, S. Chernova, M. Veloso, and B. Browning, "A survey of robot learning from demonstration," *Robotics and Autonomous Systems*, vol. 57, no. 5, pp. 469–483, May 2009.

[12] R. Rahmatizadeh, P. Abolghasemi, A. Behal, and L. Bölöni, "From virtual demonstration to real-world manipulation using LSTM and MDN," March 2016. [Online]. Available: http://arxiv.org/abs/1603.03833

[13] S. Reddy, A. D. Dragan, and S. Levine, "Shared Autonomy via Deep Reinforcement Learning," Tech. Rep., 2018.

[14] J. D. Sweeney and R. Grupen, "A model of shared grasp affordances from demonstration," in *2007 7th IEEE-RAS International Conference on Humanoid Robots*. IEEE, November 2007, pp. 27–35.

[15] H. Rhodin, J. Tompkin, K. I. Kim, K. Varanasi, H.-P. Seidel, and C. Theobalt, "Interactive Motion Mapping for Real-time Character Control," *Computer Graphics Forum*, vol. 33, no. 2, 2014.

[16] H. Rhodin, J. Tompkin, K. I. Kim, E. De Aguiar, H. Pfister, H.-P. Seidel, and C. Theobalt, "Generalizing Wave Gestures from Sparse Examples for Real-time Character Control," in *ACM Transactions on Graphics (TOG)*, vol. 34, no. 6, 2015, p. 181.

[17] Y. Cui and C. Mousas, "Master of Puppets: An Animation-by-Demonstration Computer Puppetry Authoring Framework," *3D Research*, vol. 9, no. 1, p. 5, 2018.

[18] C. Mousas, "Performance-Driven Dance Motion Control of a Virtual Partner Character," in *IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, 2018.

[19] Y. Seol, C. O'Sullivan, and J. Lee, "Creature Features: Online motion puppetry for non-human characters," in *Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, 2013.

[20] T. Zhang, Z. McCarthy, O. Jow, D. Lee, X. Chen, K. Goldberg, and P. Abbeel, "Deep Imitation Learning for Complex Manipulation Tasks from Virtual Reality Teleoperation," October 2017. [Online]. Available: http://arxiv.org/abs/1710.04615

[21] V. Kumar, A. Gupta, E. Todorov, and S. Levine, "Learning Dexterous Manipulation Policies from Experience and Imitation," November 2016. [Online]. Available: http://arxiv.org/abs/1611.05095

[22] X. Yan, J. Hsu, M. Khansari, Y. Bai, A. Pathak, A. Gupta, J. Davidson, and H. Lee, "Learning 6-DOF Grasping Interaction via Deep Geometry-aware 3D Representations," Tech. Rep., 2017.

[23] J. Aleotti, S. Caselli, and M. Reggiani, "Leveraging on a virtual environment for robot programming by demonstration," *Robotics and Autonomous Systems*, vol. 47, no. 2-3, pp. 153–161, 2004.

[24] D. Whitney, E. Rosen, E. Phillips, G. Konidaris, and S. Tellex, "Comparing Robot Grasping Teleoperation across Desktop and Virtual Reality with ROS Reality," in *Proceedings of the International Symposium on Robotics Research*, 2017.

[25] S. Y. Gadre, E. Rosen, G. Chien, E. Phillips, S. Tellex, and G. Konidaris, "End-User Robot Programming Using Mixed Reality."

[26] N. Koganti, K. Nakayama, A. Rahman, Y. Matsuo, and Y. Iwasawa, "Virtual Reality as a User-friendly Interface for Learning from Demonstrations," in *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018.

[27] S. K. Ong, J. W. S. Chong, and A. Y. C. Nee, "Methodologies for immersive robot programming in an augmented reality environment," in *Proceedings of the 4th international conference on computer graphics and interactive techniques in Australasia and Southeast Asia*, 2006.

[28] H.-L. Teulings, J. L. Contreras-Vidal, G. E. Stelmach, and C. H. Adler, "Parkinsonism Reduces Coordination of Fingers, Wrist, and Arm in Fine Motor Control," *Experimental neurology*, vol. 146, no. 1, pp. 159–170, 1997.

[29] C. M. Bishop, "Mixture Density Networks," Aston University, Birmingham, UK, Tech. Rep., 1994.

[30] Tijmen Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," pp. 26–31, 2012.