MuJoCo HAPTIX: A Virtual Reality System for Hand Manipulation

Vikash Kumar, Emanuel Todorov

***Abstract*— Data-driven methods have lead to advances in multiple fields including robotics. These methods however have had limited impact on dexterous hand manipulation, partly due to lack of rich and physically-consistent dataset as well as technology able to collect them. To fill this gap, we developed a virtual reality system combining real-time motion capture, physics simulation and stereoscopic visualization. The system enables a user wearing a CyberGlove to "reach-in" the simulation, and manipulate virtual objects through contacts with a tele-operated virtual hand. The system is evaluated on a subset of tasks in the Southampton Hand Assessment Procedure – which is a clinically validated test of hand function. The system is also being used by performer teams in the DARPA Hand Proprioception & Touch Interfaces program to develop neural control interfaces in simulation. The software is freely available at `www.mujoco.org`**

I. INTRODUCTION

Dexterous hand manipulation lags behind other areas of robotics such as kinematic motion planning or legged locomotion. While there are multiple reasons for this, here we focus on the challenges specific to data-driven approaches. Unlike full body movements, hand manipulation behaviors unfold in a compact region of space co-inhabited by the objects being manipulated. This makes motion capture difficult, due to occlusions as well as marker confusion in the case of passive systems. Manipulation also involves large numbers of contacts, including dynamic phenomena such as rolling, sliding, stick-slip, deformations and soft contacts. The human hand can take advantage of these rich dynamics, but recording the data and interpreting it with regard to well-defined physics models is challenging and has not been done in a systematic way.

The solution we propose is to leverage the adaptation abilities of the human brain, and move the data collection from the physical world to a physically-realistic simulation. The simulation is based on the MuJoCo physics engine we have developed [1]. We have recently shown [2] that MuJoCo outperforms a number of alternative simulators in terms of both speed and accuracy on modelling systems relevant to robotics, especially simulated hands grasping objects. In this work, we augment the simulator with real-time motion capture of arm and hand movements, and stereoscopic visualization using OpenGL projection from the viewpoint of the user's head (which is also tracked via motion capture.) The resulting system has empirically-validated end-to-end latency of 42 msec. It creates a sense of realism which is sufficient for untrained human users to interact with virtual objects in a natural way, and perform tasks selected from the Southampton Hand Assessment Procedure (SHAP).

The authors are with the University of Washington and Roboti LLC. E-mail: {vikash, todorov}@cs.washington.edu

The system is called MuJoCo HAPTIX. We have developed it for DARPA, with the goal of facilitating research in the ongoing Hand Proprioception & Touch Interfaces (HAPTIX) program. A number of performer teams are already using it to explore novel neural interfaces for prosthetic hands. In this paper we present our own tests of the system's latency and usability, and show that humans can indeed perform manipulation tasks with virtual objects. This clears the way to collecting rich and physically-consistent dataset of hand-object interactions. Since the interaction happens in simulation, we can record every aspect of it – including joint kinematics and dynamics, contact interactions, simulated sensor readings etc. There are no sensor technologies available today that could record such rich dataset from hand-object interactions in the physical world. Furthermore, since the interaction is based on our simulation model of multi-joint and contact dynamics, the dataset is by definition physically-consistent. Systematic collection of large dataset is left for future work. Here we focus on describing the system and demonstrating its capabilities.

II. VR HARDWARE SELECTION

Recent years have seen significant developments in 3D visualization devices including the Oculus Rift [3], Google Cardboard [4], Microsoft Hololens [5], Sony HMZ-T3W head-mounted display [6]. These devices have wide viewing angles for immersive experience, and some of them have integrated head tracking. In the context of hand manipulation, however, it is not clear that immersion is necessary or even desirable. The alternative – which we adopt here – is to use a stereoscopic monitor (BenQ) and combine it with head-tracking to create the impression of a virtual workspace that is glued to the monitor and can be viewed from different angles. The ZSpace system [7] is a commercial product based on this approach. Here we achieve the same effect using LCD shutter glasses (NVidia 3D Vision 2) tracked by an infrared motion capture system (OptiTrack V120:Trio.) The advantage of keeping the projection surface fixed in space is that image correction for head rotation becomes unnecessary. In contrast, such correction is essential when using head-mounted displays, and is difficult to implement at low-enough latencies to fool the human visual system into perceiving stable images. Furthermore, fixed monitors avoid any optical distortions and provide high resolution over the relevant workspace.

Another recent technological development are consumer depth cameras such as the Kinect [8]. They have enabled rapid progress in the area of activity recognition. While Kinect-style cameras focus on medium range sensing for

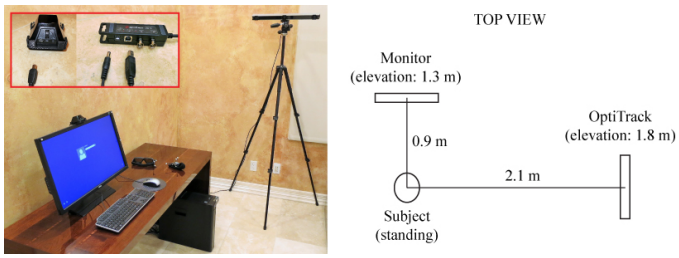


Fig. 1: System overview. Left: Overall system. Right: Schematic representation of the VR system with relative distances.

full body tracking, other devices such as PrimeSense and LeapMotion [9] can be deployed for close-range tracking of hands [10]. This technology however suffers from occlusions, and appears to be better suited for recognizing a small set of predefined gestures than unconstrained finger tracking. This is why we have adopted an older but more functional approach, which is to combine an infrared motion capture system for head and forearm tracking, with a CyberGlove [11] for wrist and finger tracking. The resulting system is considerably more expensive compared to recent consumer devices, but if we are to build an interactive simulation environment where users can indeed perform manipulation tasks, we do not see a viable alternative for the time being.

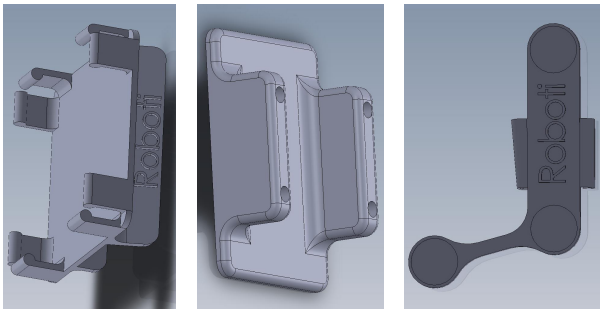
III. MuJoCo HAPTIX

A. Hardware

MuJoCo HAPTIX uses standard hardware components, except for the 3D-printed attachments for the motion capture markers and the glasses emitter (Fig 3). The default computer system is a Dell Precision T5810 workstation with Intel Xeon E5-1650 v3 processor, NVidia Quadro K4200 video card, 8GB of 2133MHz DDR4 RAM, 256GB SSD. The software relies on quad-buffered OpenGL for stereoscopic 3D rendering. It is not memory or I/O intensive and the CPU is mostly idle – which is because the MuJoCo physics engine can run realistic simulations much faster than real-time. NVidia 3D Vision 2 glasses are used to stereoscopic visualization, together with a BenQ GTG XL2720Z stereo monitor. MoCap markers are attached to the glasses for head tracking using 3D printed attachment (Fig 3c). The emitter is fitted in a 3D printed holder (Fig3a) and placed on top of the monitor. This avoids blanking of the stereo glasses due to occlusion of the emitter via hand. MoCap markers are also attached to the monitor and used to define the frame for the motion capture data; in this way the infrared cameras can be moved during a session without any effect on the simulation and rendering.

Motion capture is based on the OptiTrack Trio:V120 system [12]. This device has tracking speed and accuracy comparable to devices that cost substantially more. Its main limitation is that all three cameras are mounted in one elongated bar. This is sufficient to achieve stereo vision, but since all three views are quite similar, the system cannot track the hand in situations where markers are occluded or overlap.

Fig. 2: Motion capture markers



(a) Emitter holder (b) Hand attachment (c) 3D-Glasses

Fig. 3: 3D printed marker attachments

For head and monitor tracking this is not an issue, and the hand tracking workspace is sufficient for object manipulation tasks. Finally, we use a CyberGlove for wrist and finger tracking [11].

B. Markers and attachments

The MoCap markers and their attachments are shown in figure 2 and 3 respectively. For the head and the hand tracking we use custom 3D-printed parts to which the markers are glued. There are three markers per rigid body. The hand-tracking body uses 7/16" markers from Natural-Point (makers of the OptiTrack), while the rest are 9mm removable-base markers from MoCap Solutions. The head-tracking attachment (Fig 3c) slides on the stereo glasses (Fig2-left pane) to track the head position of the user. The hand-tracking attachment (Fig 3b) has Velcro on the bottom, and attaches to a mating velcro strap that goes over the wrist (Fig 2-mid pane) for tracking the base of the hand. Note the orientation of the hand-tracking body. Attaching it in the wrong orientation significantly reduces the usable workspace in terms of forearm pronation-supination. The monitor markers are attached directly to the monitor bezel. The positioning of these markers (as shown in Fig 2-right pane) is important, because the software uses them to compute the position, orientation and size of the LCD panel – which in turn is needed for rendering from a head camera perspective.

C. Tracking

MuJoCo HAPTIX expects a real-time stream with information about the 6D (3D position plus 3D orientation) orientation of the monitor, the users's head and the user's hand. This is provided by the OptiTrack library which is loaded in MuJoCo HAPTIX. For hand tracking a combination of data stream from OptiTrack and CyberGlove is used.

1) Head tracking: The user's head and the screen are tracked via the markers attached to the screen and glasses. The markers attached to the monitor are also used to measure the physical dimensions of the screen, which are then taken into account to create appropriate projection. The 6D head orientation is used to render the scene from the physical location of the eyes using oblique projections to create the impression that the virtual world is glued to the monitor.

2) Hand tracking: The hand-tracking body attached to the user's wrist controls the base of the simulated hand, but this control is not direct. "Direct control" would involve setting the position and orientation of the virtual hand base equal to the motion capture data. This has undesirable effects in terms of physics simulation and sensor modeling. Instead we use the motion capture data to set the pose of a dummy body, and connect this body to the base of the virtual hand with a soft equality constraint. The constraint is enforced by the MuJoCo solvers that compute the contact forces and the joint friction and the joint limit forces. By adjusting the relative softness of the different constraints, we can set their priority. For example, if the user attempts to move the hand into the table, there is a tradeoff between assigning priority to the non-penetration constraints (in which case the dummy body will move into the table but the virtual hand will remain on the surface), and priority to tracking the dummy body as faithfully as possible.

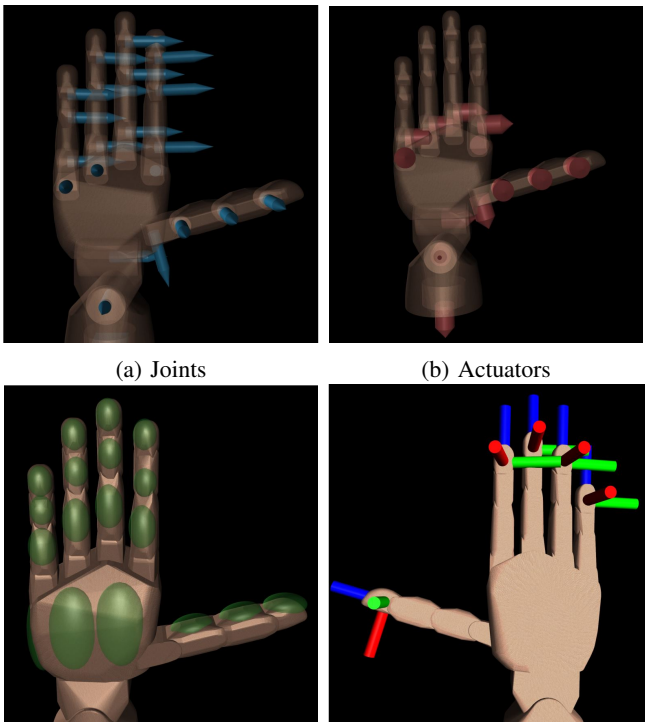
A CyberGlove, calibrated for the hand model in use, is used to set the position offset of the PD controller driving the finger joints. The OptiTrack data stream fused with the CyberGlove data stream enables hand tracking.

IV. Modelling

We modelled the Modular Prosthetic Limb (MPL) [13] and ADROIT hand [14] (Fig 5) for testing purposes. The two hand models are expressed in MuJoCo's model definition language called MJCF. These models can be populated inside diverse Virtual Environments, also expressed in MJCF.

MPL model consists of 22 dof (Fig 4a) (19 in the fingers and 3 in the wrist) and 13 actuated dof (Fig 4b). Finger flexion and extension are coupled using differential arrangement and actuated using a single actuator. Ring and little finger's adduction-abduction are coupled together and actuated using a single actuator. MuJoCo's equality constraints are leveraged to model all couplings. Index finger MCP adduction-abduction, all thumb joints and all wrist joints have independent actuation. Convex meshes derived from the original CAD models are used for collision purposes. In accordance with the sensing capabilities of the real hand, the state of the system is exposed to the users via MuJoCo's simulated sensors. Sensory capabilities include - joint position and velocity sensors on all 22 joints, actuator position, velocity and force sensors on all 13 actuators, touch sensors (Fig 4c) and inertial measurement units on all 5 finger tips(Fig 4d).

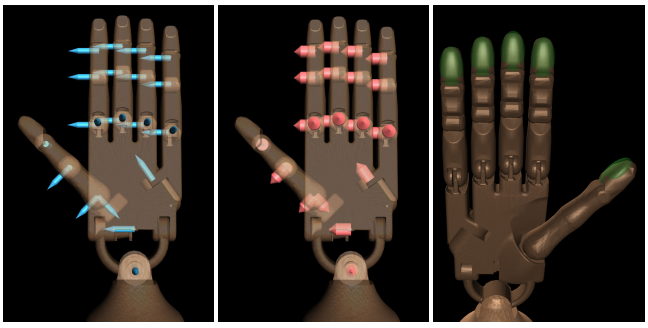
ADROIT model consists of 24 dof (Fig 5a) with 20 independently actuated dof (5b). Finger's distal joints are coupled with the proximal joints using MuJoCo's tendon constraints. Convex meshes derived from CAD models are used for collision detection. Sensing capabilities include – joint sensors on all 24 joints, position and force sensors on all 20 actuators, and touch sensors on the distal segments of all fingers (Fig 5c). For space restrictions, rest of the paper focuses on MPL hand, but similar results are available for ADROIT hand as well.



(a) Joints (b) Actuators

(c) Touch sensors (d) IMU sensors

Fig. 4: MPL sensor configurations



(a) Joints (b) Actuators (c) Touch sensors

Fig. 5: Adroit sensor configurations

Virtual environments (Fig 8 & 9) were primarily modelled using geometric shape primitives for fast and well behaved collisions. Note that MuJoCo doesn't support fluid modelling. Multiple small and smooth balls were instead used as a replacement.

V. Evaluation and Results

A. Latency-tests

The latency of a virtual environment is an important factor affecting human sensorimotor performance. End-to-

end latency, from movement of the physical MoCap marker to change of the pixels on the monitor, of our VR system was established to be between 42 and 45 msec. This was determined empirically in two different ways.

The first – the last two marker positions obtained from the MoCap was used to extrapolate its position some time into the future (45 msec in this case) under constant velocity assumption. Both the current (white) and extrapolated (green) positions were rendered (Fig 6), using projection to the surface of the monitor. Hence marker movement closer to the monitor allowed a visual comparison of the physical and the rendered marker positions. The prediction interval was adjusted such that the extrapolated (green) marker neither leads nor lags the physical marker, but instead was aligned with it on average. Extrapolation was done under a constant velocity assumption, thus non-zero acceleration would affect the measurement, but hand acceleration is zero on average (it changes direction) hence the result is unbiased. In figure 6 the hand is moving and the green marker cannot be seen because it is exactly under the physical marker. In the right panel the hand is stationary, thus the white and green markers are on top of each other. They appear above the physical marker because it is some distance away from the monitor, and the camera is above the hand.

The second - a tap was applied to the hand tracking body with a pen (figure 6 right panel), recorded with a 120Hz camera, and the video frames separating the physical contact event and the change in virtual marker speed (which was printed on the screen) was counted. Similarly, we also moved the physical marker up and down across a horizontal line, and counted the number of video frames between the physical and virtual markers crossing the line. These tests showed around 42 msec overall latency, and the frame counts were very consistent (5 frames in almost all the cases).

The total amount of time that motion capture data spends in the VR pipeline - from the time it is delivered by the motion capture library to the time the video driver reports that rendering is finished, is between 6 msec and 12 msec, because the relative timing of the motion capture and video card fluctuates. It includes processing of the motion capture data, simulation and rendering. The rest is due to latency in the hardware devices in use.

Overall, the virtual environment is very responsive and usable.

B. SHAP tests

To evaluate the manipulation capabilities, we modelled Southampton Hand Assessment Procedure (SHAP) [15] in our virtual environment. SHAP is a clinically validated hand function tests developed by Colin Light, Paul Chappell and Peter Kyberd in 2002 at the University of Southampton. Originally developed to assess the effective functionality (Fig 7) of upper limb prostheses, the SHAP has now been applied for assessments of musculoskeletal and unimpaired participants too. The SHAP is made up of 6 abstract objects (AO) and 14 Activities of Daily Living (ADL). Each task is timed by the participant, so there is no interference or reliability on the reaction times of the observer or clinician.

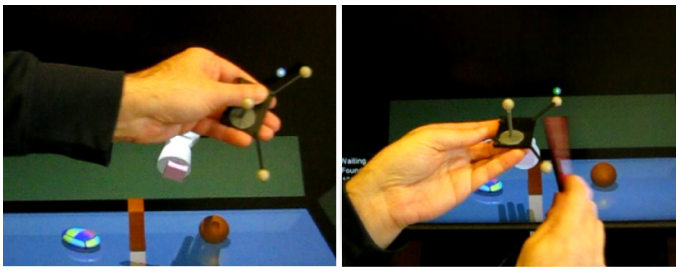


Fig. 6: Latency experiments. (Left) - The hand is moving. The green marker is exactly below the physical marker. (Right) - The hand is stationary. The green and the white marker are on top of each other.

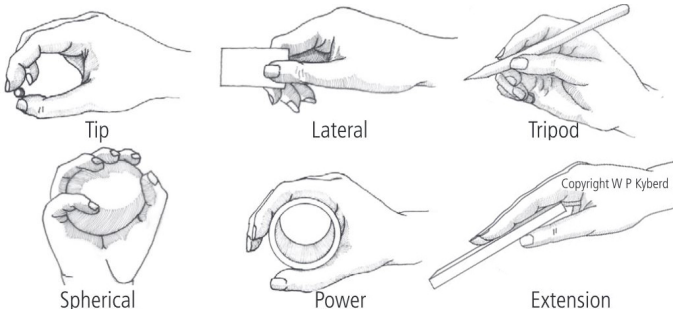
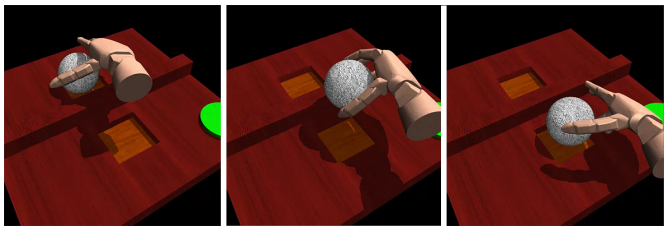


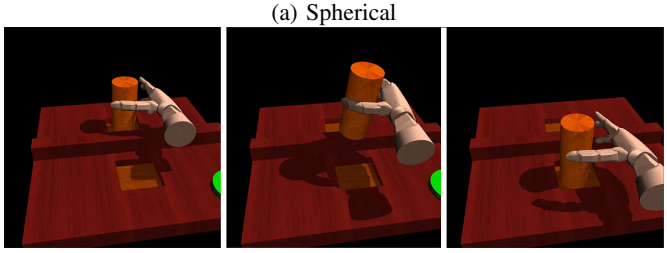
Fig. 7: Six grip classifications used in Southampton Hand Assessment Procedure assessment (courtesy- W P Kyberd).

Listed below are the SHAP procedures we tested against. Note that we ignored the procedures that can't be modelled using rigid body dynamics.

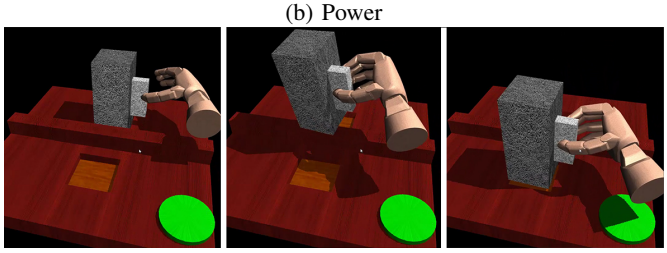
a) AO-Spherical (Fig 8a) evaluates spherical gripping capabilities as the user moves a spherical object over a barrier to a slot in front.
b) AO-Power (Fig 8b) evaluates power grasping capability as the user moves a cylindrical object over a barrier to a slot in front.
c) AO-Lateral (Fig 8c) evaluates lateral grasping capability as the user repositions an object with a handle over a barrier.
d) AO-Tip (Fig 8d) evaluates grasping capabilities as the user deploys trip grasp to reposition an object over a barrier.
e) AO-Tripod (Fig 8e) evaluates tripod grip by asking the user to move an triangular pipe shaped object over a barrier to a slot in front.
f) ADL-Pick Coins (Fig 9a) uses tripod or tip grip to drop a sequence of coins from the table into the jar.
g) ADL-Pouring (Fig 9b) uses lateral or power grip to pour liquid (roll small rigid balls in this case) from one jar to the other.
h) ADL-Key (Fig 9c) evaluates rotation of a key by 90 degrees
i) ADL-Screw (Fig 9d) evaluates rotation of a screw by 90 degrees using screwdriver (power grasp)
j) ADL-Handle (Fig 9e) evaluates rotating door handle using power grasp.



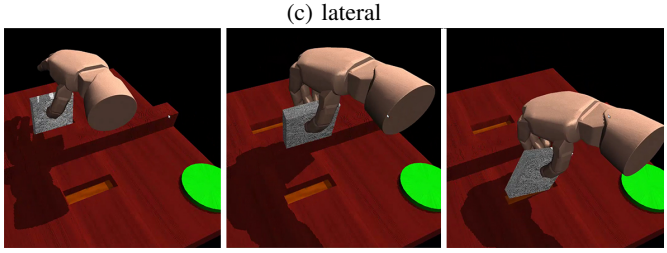
(a) Spherical



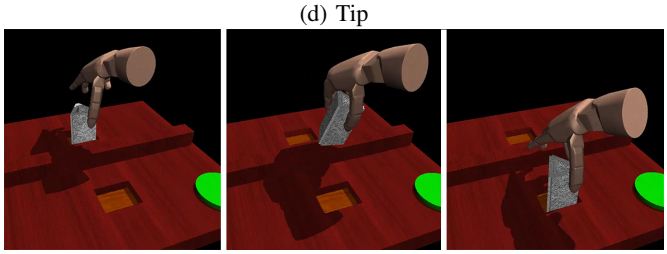
(b) Power



(c) lateral

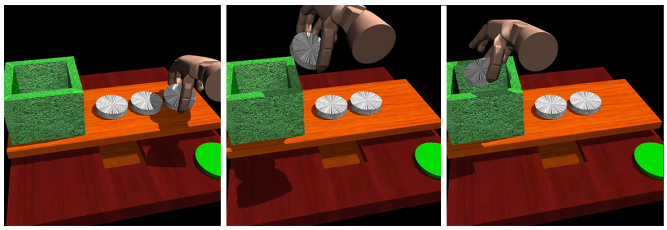


(d) Tip

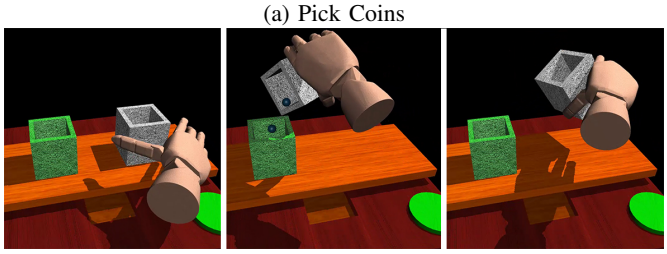


(e) Tripod

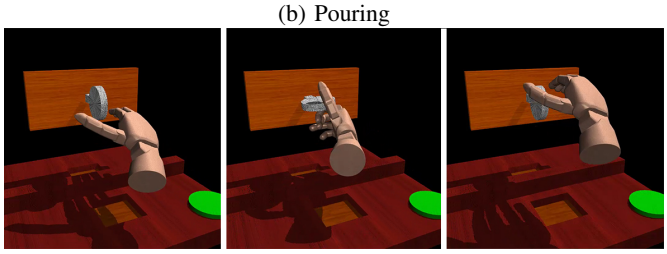
Fig. 8: Abstract Objects SHAP test sequence



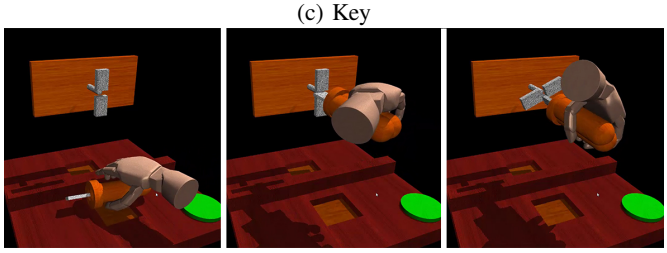
(a) Pick Coins



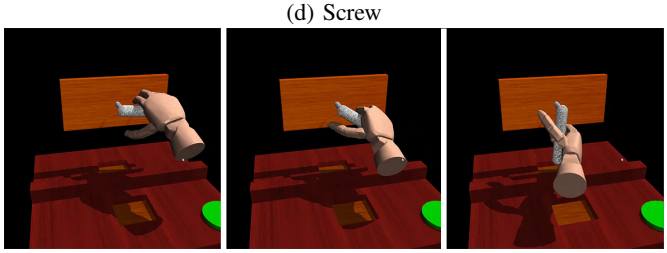
(b) Pouring



(c) Key



(d) Screw



(e) Handle

Fig. 9: Activity for daily learning SHAP test sequence

Our user group comprised of 7 subjects between the age of 22-28. Each subject starts by calibrating the CyberGlove against the hand model in use. She is then allowed some exploration time before the trial starts. She begins by pressing the button to start the timer, performs the task and then presses the button again to stop the timer. Each SHAP procedure is repeated 10 time. Each user performs the evaluation for two different version of the hand under study - fully actuated and with coupled actuation. The resulting 1400 user demonstrations across 10 tasks provide convincing results to back the claims made in paper regarding usability of the system and its ability to support manipulation. An extensive user study with users from different demographics is being pursued as future work to investigate various effects of age, demographics, feedback choices, visualization techniques, task intensity etc [16].

The average of the time taken by a user on any particular trial across all SHAP procedures serves as metric to evaluate the usability of the system. Lower the average time, more usable the system. The usability metric plotted over all the trial (Fig 10) reveals the learning curve of our system. The curve ramps down slowly and plateaus after 4 trial indicating the low learning curve of our system.

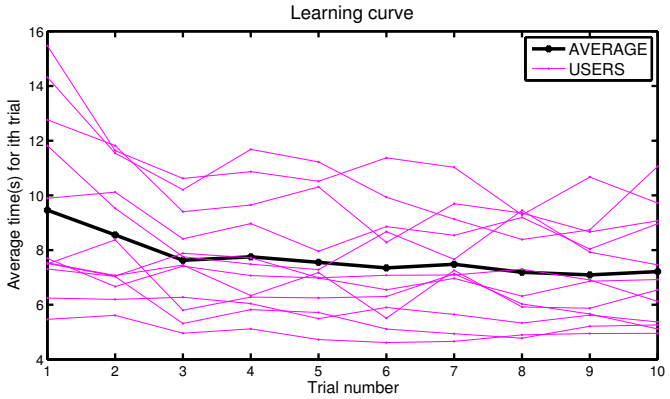


Fig. 10: Learning curve of the MuJoCo HAPTIX system

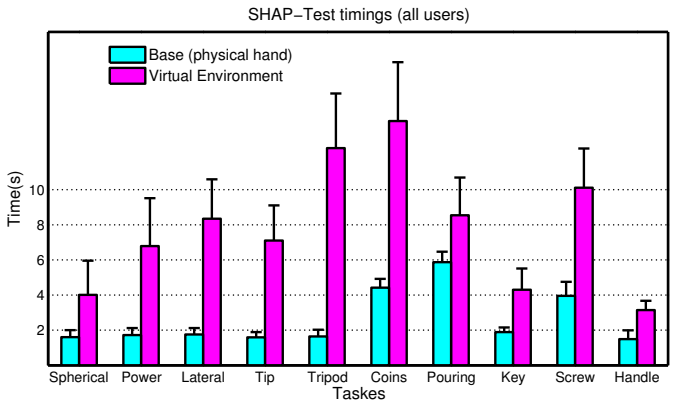


Fig. 11: SHAP timing comparison between physical tasks and tasks performed in MuJoCo HAPTIX, by healthy control groups

Figure 11 summarises the SHAP timing from MuJoCo HAPTIX, along with the results originally reported in [15] (base-line) using physical objects from a healthy control group of twenty-four volunteers selected on the basis of optimum hand function using these criteria: age (range, 18-25y), and no adverse hand trauma, neurologic condition, or disabling effects of the upper limb. All the users were able to finish all tasks well within the time limit. Based on the timing results, we establish that our system supports rich manipulation capabilities to fulfil SHAP within reasonable time frame. On average a user took 3 times more time than the base-line. We attribute this slowdown to a number of factors - primarily to the lack of haptic feedback. All results reported in this work are without any user feedback (except visual feedback).

C. Hand design evaluation

SHAP provides a quantifiable assessment of hand capabilities. These assessment can further be treated as a measure of hand's dexterity under the engaged controller. By changing the hand designs without changing the controller (i.e. the user) we obtain a novel mechanism of relative evaluation of hand designs. Given the plasticity and expertise of the human brain in controlling human hands, we assume that we have engaged a rather smart controller for the hand design under evaluation. Plasticity of the assessment pose a significant advantage. Random perturbations or physical artifacts (like wind, low friction coefficient) can be easily added to increase the severity of the task. Alternatively, evaluation intensity can be reduced by providing help clues (like grasp metrics etc) to the users. SHAP can not only help with design evaluation but also provide assessments like – exceptional power and spherical grip, or impaired ability to perform finer manipulations with tip and tripod grasps. In addition, a 'SHAP Index of Function score' can also be generated, which is one number that provides an overall assessment of hand function under the engaged controller.

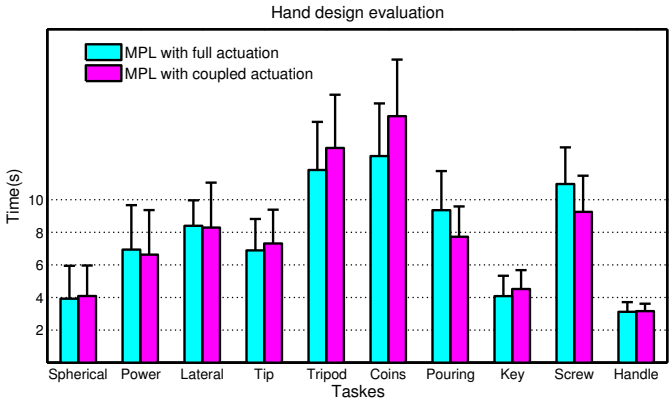


Fig. 12: Design evaluation of MPL hand - with full and with coupled actuation

A fully actuated system is maximally capable of exploiting the dexterity of a hardware. However, it can be hard to manufacture due to cost and design constraints, specially in case of anthropomorphic hand designs due to high dof density in the workspace. Couplings and transmissions help solve some of these constraints at the cost of reduced dexterity. Couplings morph the workspace, which can either help the hardware in achieving some tasks better by creating favourable subspaces or can hinder some functionality by constraining dof of the system. One can use above mentioned pipeline to evaluate and iterate over the coupling options that least affects the functionally of the overall system. The pipeline facilitates fast evaluation before manufacturing, resulting in huge cost and time advantage.

Table 12 presents the design evaluation result between two MPL hands – one with the coupled actuation as described in [13] and the other with full actuation. Users were asked to perform SHAP tests with both the hands without information about which one they were using. While MPL with coupling performs at par (or slightly better) in most SHAP tests, increase in timings for tip, tripod, coins and key highlight reduction in ability to perform fine manipulation, essential for these SHAP-tests.

VI. Future Works

Although figure 11 provides convincing proof that MuJoCo HAPTIX is capable of supporting manipulation, it also

highlights a large divide between demonstrations performed in reality vs demonstrations done in virtual environment. A number of factors can be attributed to explain the divide - including but not limited to the lack of haptic feedback, off-calibration of CyberGlove etc. Exploration of haptic feedback techniques is our primary point of focus in future. We are also exploring non-standard feedback modalities like visual cues, audio cues etc to convey information about contact events, rolling, sliding, stick-slip etc.

Diverse physically consistent dataset (concealing rich information about phenomena like contacts, friction, stiction, deformation, sliding, rolling etc) can be leveraged by data driven machine leaning techniques and inverse optimal control techniques for understanding movements and synthesising behaviours. Medical robotics, rehabilitation and prosthesis can further use MuJoCo HAPTIX for training and testing purposes. Analysis of the resulting dataset can help understand the effectiveness of the mechanism and its effects on the users. MuJoCo HAPTIX is fast with negligible latency. Access to physical entities can be used for real-time feedback. To further this endeavour, MuJoCo HAPTIX is being used for Darpa's ongoing Hand Proprioception & Touch Interfaces (HAPTIX) program for investigation of minimum sensory motor feedback required in order to close the loop between a prosthesis and its human counterpart.

VII. Acknowledgement

This work was supported by DARPA and the NSF. Authors also want to thank Visak Kumar, Mechanical Engineering, University of Washington for his help with the user studies.

References

[1] E. Todorov, T. Erez, and Y. Tassa, "Mujoco: A physics engine for model-based control," in *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*. IEEE, 2012.

[2] T. Erez, Y. Tassa, and E. Todorov, "Simulation tools for model-based robotics: Comparison of bullet, havok, mujoco, ode and physx." in *IEEE International Conference on Robotics and Automation, (ICRA'15)*.

[3] Oculus Rift, http://oculusrift.com.

[4] Google cardboard, www.google.com/get/cardboard/.

[5] Microsoft HoloLens, http://www.microsoft.com/microsoft-hololens.

[6] Sony Head mounted display, www.sony.co.uk/electronics/head-mounted-display/t/head-mounted-display.

[7] Zspace, http://zspace.com/.

[8] Microsoft Kinect, www.microsoft.com/en-us/kinectforwindows/.

[9] Leap Motion, www.leapmotion.com.

[10] T. Schmidt, R. Newcombe, and D. Fox, "Dart: Dense articulated real-time tracking," *Proceedings of Robotics: Science and Systems, Berkeley, USA*, vol. 2, 2014.

[11] Cyber Glove Systems, www.cyberglovesystems.com/.

[12] OptiTrack, www.optitrack.com.

[13] Modular Prosthetic Limb, Johns Hopkins Applied Physics Lab, http://www.jhuapl.edu/prosthetics/scientists/mpl.asp.

[14] V. Kumar, Y. Tassa, T. Erez, and E. Todorov, "Real-time behaviour synthesis for dynamic hand-manipulation," in *Robotics and Automation (ICRA), 2014 IEEE International Conference on*. IEEE, 2014, pp. 6808–6815.

[15] C. M. Light, P. H. Chappell, and P. J. Kyberd, "Establishing a standardized clinical assessment tool of pathologic and prosthetic hand function: Normative data, reliability, and validity," *Archive of Physical Medicine and Rehabilitation*, 2002, http://eprints.soton.ac.uk/256475/.

[16] MuJoCo HAPITX Studies, www.cs.washington.edu/homes/vikash/P_HAPTIX.html.