Robotics as a Tool for Training and Assessment of Surgical Skill

Marcia K O'Malley PhD, Ozkan Celik PhD, Joel C Huegel PhD, Michael D Byrne PhD, Jean Bismuth MD, Brian J Dunkin MD, Alvin C Goh MD, and Brian J Miles MD

Abstract Technological advances have enabled new paradigms for skill training using virtual reality and robotics. We present three recent research advances in the field of virtual reality and human-robot interaction for training. First, skill assessment in these systems is discussed, with an emphasis on the derivation of meaningful and objective quantitative performance metrics from motion data acquired through sensors on the robotic devices. We show how such quantitative measures derived for the robotic stroke rehabilitation domain correlate strongly to clinical measures of motor impairment. For virtual-reality based task training, we present task analysis and motion-based performance metrics for a manual control task. Lastly, we describe specific challenges in the surgical domain, with a focus on the development of tasks for skills assessment in surgical robotics.

1 Introduction

A primary purpose of virtual environment technology is to enable a medium for safe and affordable practice of a broad range of manual tasks. Virtual training can be designed either to provide a virtual practice environment that matches the targeted physical environment as closely as possible, or to provide virtual assistance intended to improve training effectiveness. Regardless of the approach, the aim of training in VEs is to transfer what is learned in the simulated environment to the equivalent real world task. Caution should be taken when using virtual environments for training, since it has been shown in the literature that intuitive training schemes in computationally mediated environments with visual and auditory feedback may not result in positive transfer effects and can even lead to negative transfer [1, 2, 3, 4]. Negative transfer effects are attributed mainly to limitations in the fidelity of the virtual task compared to the real task due to simplifications required for rendering. Negative transfer effects may also be attributed to the augmentation of task dynamics due to the presence of virtual guidance [5, 6, 7, 8]. In contrast to these findings, VEs have been shown to be effective for training in navigation and simple sensorimotor tasks [9, 10], and multi-modal environments for surgical skill training are under development [e.g., 11, 12].

Devices such as instrumented joysticks for simulator systems, robotic surgery platforms, or robotics rehabilitation devices offer precision data acquisition during task execution, training, and therapy. Access to such data drives the need to identify adequate performance measures that accurately measure skill. Such measures can be task-dependent, such as time to completion, or success and/or failure rates. Skill measures can also be task independent and based on movement characteristics such as trajectory error [13], force [14], input frequency [15, 16], movement smoothness [17, 18, 19], and more. The use of motion-based data to assess performance is increasing with the accessibility of new sensing technologies, ranging from high-end multi-camera and reflective marker based systems to handheld video gaming devices.

In this paper, we discuss current research on skills assessment in physical human-robot interaction (HRI) systems. We discuss how motion-based performance metrics in the rehabilitation robotics domain correlate to clinical measures of motor impairment. Then, in virtual reality-based tasks, we describe our efforts to determine key

Marcia K O'Malley, PhD
Rice University, Houston TX

Ozkan Celik, PhD
San Francisco State University, San Francisco CA

Joel C Huegel, PhD
Tecnologico de Monterrey-Campus Guadalajara, Guadalajara, Mexico

Michael D Byrne, PhD
Rice University, Houston TX

Jean Bismuth, MD
The Methodist Hospital, Houston TX

Brian J Dunkin, MD
The Methodist Hospital, Houston TX

Alvin C Goh, MD
USC Institute of Urology, Los Angeles CA

Brian J Miles, MD
The Methodist Hospital, Houston, TX

strategies that enable high performance on tasks, and associated motion-based metrics that quantify these strategies. Finally, we describe efforts to design appropriate tasks for skills assessment in surgical robotics.

2 Skills assessment in physical-HRI systems

We have extensively studied skill acquisition in virtual environments with a number of input devices ranging from low-cost gaming controllers (Nintendo, Sony) to research-grade haptic joysticks providing multi-modal feedback and augmented guidance (IE2000, Immersion and custom haptic enabled virtual environment with dynamic target hitting task, see Figure 1). Key to these studies was access to quantitative movement data from the robotic devices, which allowed objective analysis of performance, and correlation of motion-based performance metrics to outcome-based measures. Our findings suggest that underlying movement characteristics can differentiate skill level (experts versus novices) in novel motor tasks, and can differentiate motor impairment severity in stroke populations.

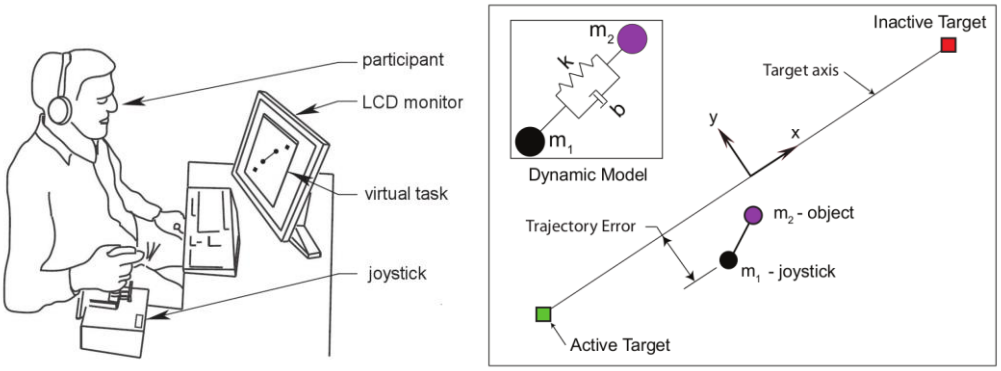


Fig. 1 Target hitting task: Subjects control location of m_1 (joystick) in order to cause m_2 (object) to hit the desired target. Inset shows virtual underactuated system. The user controls the system by applying forces to mass m1 through a force feedback joystick interface. Performance is measured by number of target hits and by instantaneous error, defined as the deviation of m_2 (object) from the target axis. (© MIT Press, [20], reprinted with permission)

2.1 Motor impairment assessment in robotic rehabilitation

Although there have been numerous studies on the design and testing of novel therapeutic robots, an effective method for objective assessment and comparison of such devices is yet to be determined. The potential prospects of robotic rehabilitation include home-based rehabilitation systems, remote supervision by therapists, and automated adaptive rehabilitation programs. For all of these opportunities to be embraced, a unified set of robotic motor recovery measures with known correlation to clinical measures is highly desirable.

Performance measures for conventional and robotic rehabilitation are typically determined via clinical assessment scales, where specific activities are conducted by the physical or occupational therapist with the patient, or by surveys indicated amount and quality of use of the impaired limb(s). Moreover, quantitative metrics of human performance can also be derived from the data collected by robotic devices used to administer therapy. We identified key aspects for a set of unified robotic motor recovery measures by analyzing the motor function improvement scores of nine chronic stroke patients who underwent a hybrid robotic plus conventional therapy program. We used quantitative motion data to generate performance metrics that correlate significantly to functional impairment measures used clinically to assess stroke severity [19].

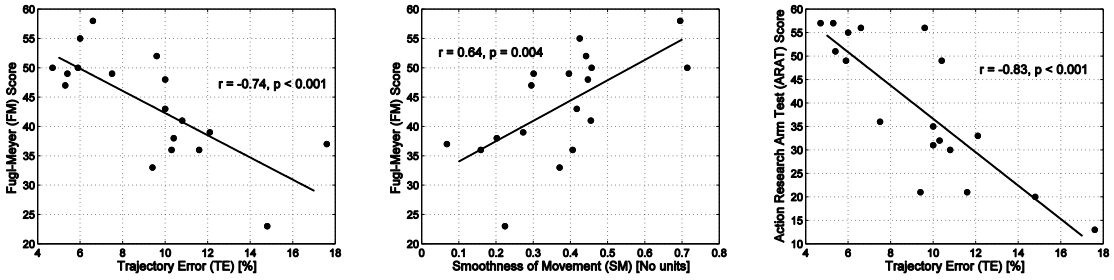
Specifically, we analyzed the motor function improvement scores of nine chronic stroke patients, utilizing four clinical measures (Fugl-Meyer (FM), Action Research Arm Test (ARAT), Jebsen-Taylor Hand Function Test (JT) and Motor Activity Log (MAL)) and four robotic measures (smoothness of movement (SM), trajectory error (TE), target hits per minute (HPM), and mean tangential speed (MTS)). We used our clinical data to compute correlations

between robotic and clinical measures and furthermore indicate important properties that such measures should exhibit for strong correlation with clinical measures.

Smoothness of movement and trajectory error, temporally and spatially normalized measures of movement quality defined for point-to-point movements, were found to have significant moderate to strong correlations with all four of the clinical measures (see Table 1). Our measures that quantify movement quality, TE and SM, demonstrated significant and moderate to strong correlations with all clinical measures (see Figure 2). In contrast, correlations of movement speed based measures, HPM and MTS, with clinical measures mostly failed to show significance, and correlations ranged from none at all (MTS-ARAT) to moderate (HPM-FM). Therefore, we conclude that one key feature in order for a robotic measure to have strong correlation with clinical measures is a focus on movement quality rather than on speed. The strong correlations suggest that smoothness of movement and trajectory error may be used to compare outcomes of different rehabilitation protocols and devices effectively, provide improved resolution for tracking patient progress compared to only pre- and post-treatment measurements, enable accurate adaptation of therapy based on patient progress, and deliver immediate and useful feedback to the patient and therapist.

Table 1 Results of the correlation analyses of FM, ARAT, JT, and MAL measures on TE, SM, HPM, and MTS measures (see text for full versions of abbreviations). Correlation coefficient (Pearson's) *r* is listed. * denotes significant correlation ($p<0.05$) (© IEEE, [19], reprinted with permission)

	TE	SM	HPM	MTS
FM	-0.74*	0.64*	0.54*	0.22
ARAT	-0.83*	0.51*	0.37	0.00
JT	0.63*	-0.49*	-0.53*	-0.32
MAL	-0.49*	0.57*	0.46	0.21



(a) (b) (c)

Fig. 2 Regression plots for clinical measures FM, ARAT, and robotic measures TE, SM. Correlation coefficients between two types of measures and the p value of the correlation coefficients are given. Each patient is represented by two points (pre- and post- treatment scores). (a) Strong and significant correlation exists between FM and TE measures. (b) There is a moderate and significant correlation between FM and SM measures. (c) There is a very strong and significant correlation between ARAT and TE measures. (© IEEE, [19], reprinted with permission)

2.2 Motor skill acquisition in virtual environments

We have used similar techniques to compare motion-based performance metrics to objective task performance (time to completion, success rates) [20]. Specifically, we have investigated performance during reaching movements where a virtual object is manipulated while haptic feedback corresponding to the task dynamics is provided, as shown in Figure 1 [5, 6, 7, 8].

Learning in this dynamic target hitting task, where the subject controls the black mass (m_1) via the haptic joystick to control the movement of the green object (m_2) in order to hit targets, is not uniform across subjects. The dynamics of the task are similar to that of a yo-yo or paddle-ball game, where the two masses are connected via a spring and

damper and can oscillate and swing around the planar environment. The task is scored according to the number of target hits that the operator can score in a 20-second interval. This is a challenging task for many subjects. However, with practice, most (but not all) subjects become fairly proficient with the task and can generate slightly more than 1 target hit per second. In fact, one of the things that makes this task interesting is precisely the issue of learning. Some subjects start out poorly and improve only a modest amount across multiple experimental sessions. Still other subjects start out doing well and show a similar modest improvement, generating strong scores across all trials. Finally, a third group of subjects starts out doing poorly, but learns rapidly and ends up doing about as well as subjects who started out strong. Figure 3 presents data from Huegel et al. [20] showing this breakdown. High performers are defined as subjects whose initial hit count performance is more than one standard deviation above the mean. Low performers are defined as subjects whose final hit count performance is more than one standard deviation below the mean. The third group consists of all other subjects; they transition from performing like low performers into performing like high performers.

The distinct shapes of the learning curves in Figure 3 suggest that different subjects learn dramatically different things over the course of the experiment. More detailed analyses of the raw motion data have provided certain critical insights. Expertise, as determined by the outcome-based measure (hit count, as presented in Figure 3), correlates with measures computed via kinematic and kinetic sensor data acquired via the robotic device [20]. In particular, performance on this task appears to be a function of two major components of the motion data: (1) Off-axis error: as off-axis movement decreases, scores increase; and (2) Movement frequency: power spectra analysis shows that as movement frequency approaches the natural frequency of the system (approximately 1 Hz), scores increase. Interestingly, these two measures of movement performance are only weakly correlated. Thus, doing well at the task requires that subjects master both aspects of the fundamental movements. These results suggest that training for such tasks should not be based simply on the outcome-based measures, but truly effective training requires examination of performance at a deeper level as well as feedback not simply at the outcome level, but at the level of basic movements.

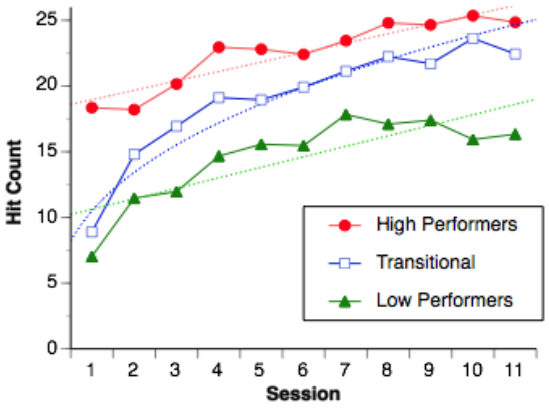


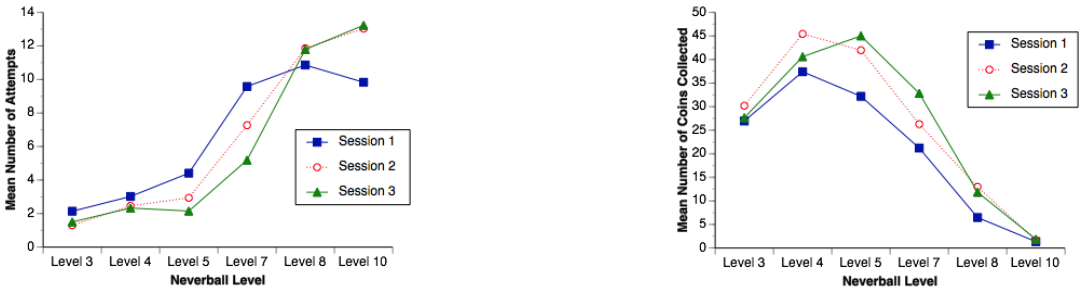
Fig. 3 Mean hit count per trial across sessions for the three subject groups in the target-hitting task.

2.3 Human performance for novel virtual environment tasks

Based on the insight gained via the virtual target hitting task, we have expanded our study to new tasks and computer interface devices. We have recently analyzed results from a multi-session training experiment using Neverball, an open source game in which the player controls a platform on which a ball rolls. Neverball was chosen because the motion dynamics involved resemble the spring-mass-damper system of the target-hitting task with which we were familiar, but has a richer set of outcome measures and requires a wider variety of movements. The first key result from this research is that there were no performance differences associated with using different physical controllers requiring different styles of movement to control the game (e.g. the Nintendo Wii remote, the Sony Sixaxis controller, and the Novint Falcon).

In a second experiment, we made use of only one controller (the Wiimote, for purely practical reasons), and we were concerned with learning over a more extended time frame. Preliminary results are compelling, with evidence of substantial learning across sessions, and clear differences in the raw movement profiles between subjects early in learning and late in learning. We have found strong evidence of learning (see Figure 4), though the amount of

learning depended on which level of the game was being played. Generally, subjects showed improvements in the levels of moderate difficulty, taking fewer attempts each session to complete each level while simultaneously collecting more coins on each level. The obvious exception here is level 10, which is the 3D "bowl" level designed to elicit movement patterns similar to the prior target-hitting task. Subjects clearly did not show improvement here.



(a) (b)

Fig. 4 (a) Mean number of attempts per level across sessions. Fewer attempts meant subjects were able to complete the levels more quickly (i.e., lower is better). (b) Mean number of coins collected per level across sessions. More coins is better.

While we see evidence of learning in the outcome measures of the Neverball task (e.g. number of attempts, coins collected), analysis of motion data collected from the low-cost gaming controller (Wiimote in this case) also shows evidence of learning and strategy development. Consider the "half-pipe" level of the Neverball task, where the subject must maneuver the half-pipe environment via the gaming controller to collect coins and proceed to the goal. The raw acceleration traces (shown in Figure 5) provide evidence of the subject's convergence to a specific motion pattern for this high performing subject. The plots show x-axis accelerations versus z-axis accelerations across the three training sessions (for a subset of trials which are representative of trends across our high-performing subjects).

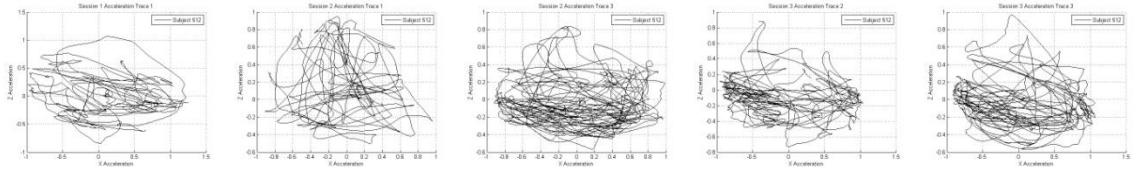


Fig. 5 Acceleration traces (x-axis versus z-axis) collected via Wii remote over three sessions of training in the half-pipe Neverball task environment. Note convergence to side-to-side (+/- z) movements with slight progression forward (+ x) towards the goal

3 Challenges in the surgical domain

In many surgical domains, caseload is used to determine adequacy of training [21, 22]. This is problematic, especially in fields like vascular surgery, which is distinguished by being a low-volume/high-complexity specialty, where rigorous assessment of technical skill is vital. The traditional apprenticeship model introduced by Halsted of "learning by doing" may not be valid in the modern practice of vascular surgery. The model is often criticized for being somewhat unstructured, as a resident's experience is based on what ""comes through the door." Therefore, virtual environments for surgical skill training offer the potential to increase caseload. Simulator-based training and testing offers a crucial and standardized methodology to evaluate a trainee's proficiency. Skills acquisition and maintenance, however, are controversial issues, as is the transfer of skills gained in simulation environments to real-world cases [23].

Other challenges are noted in the field of robotic surgery. It is estimated that more than 70% of radical prostatectomies are performed with robotic assistance [24], and that a significant learning curve exists when surgeons acquire technical skills using robotic platforms [25]. Standard laparoscopic tasks cannot distinguish skill level in the robotic setting and show rapid reduction in novice times in only a few trials [26, 27]. Therefore, there is a need to identify validated tasks to assess proficiency and optimize training on the robotic platform.

We have used expert robotic surgeons to deconstruct the robot-assisted prostatectomy to identify technical skills essential to robotic surgical performance. Our objectives were to develop inanimate tasks that can accurately measure robotic technical skill, to demonstrate the effectiveness of this model, and to establish face and construct validity. A series of structured inanimate tasks were designed to progressively emphasize key cognitive and motor abilities, including elements singular to the robotic platform like clutch and camera control (see Figure 6).

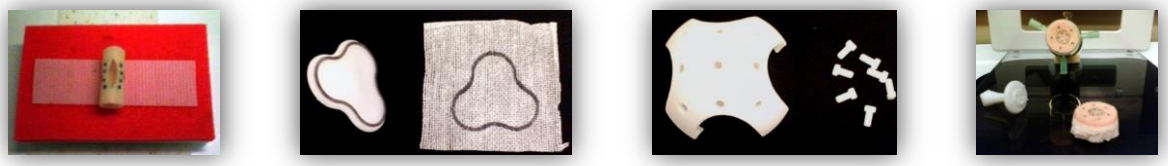


Fig. 6 Inanimate tasks, from left to right: horizontal mattress suture, clover pattern cut, dome and peg placement, circular target needle placement

Thirteen subjects, including two experts (involved in >30 robotic cases) and eleven novice surgeons, completed four training tasks. Overall performance and learning curves were measured using a scoring system developed to evaluate accuracy and efficiency. Mean expert and novice performance scores were significantly different for each inanimate training task (p<0.01), as shown in Figure 7. Experts consistently scored better than novices overall and following each successive trial (p<0.01). While improvement in performance was observed in the novice group with repetition, expert level was not reached (p<0.01). Expert performance remained stable over time. Subjects agreed the tasks were appropriately challenging and incorporated technical skills needed in robotic surgery. These findings establish face and construct validity for a series of structured inanimate training tasks. Performance of inanimate tasks can be a marker for robotic technical skills, and our model may be useful in robot-assisted surgical training and evaluation.

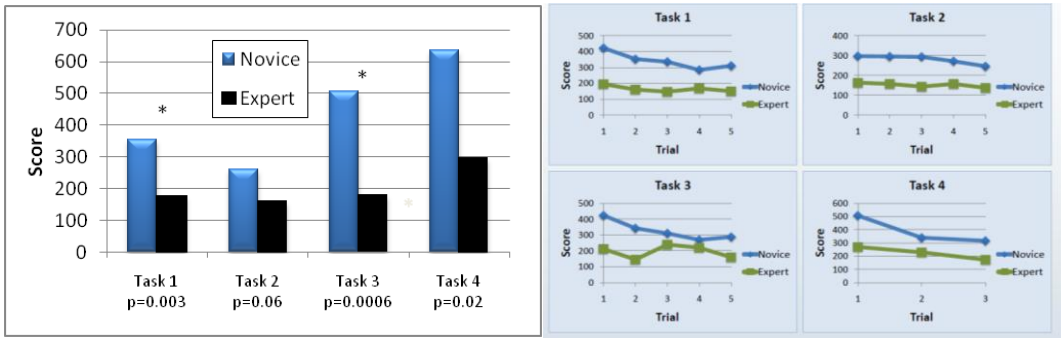


Fig. 7 (left) Mean scores of novices versus experts for each of the four tasks, (right) learning curves for novice and expert performance for each task

4 Conclusions

Virtual reality and robotics offer effective mechanisms for skill training at low cost, and the sensors inherent to the physical systems provide quantitative data upon which objective measures of task performance can be derived. We have studied skill acquisition in virtual environments with a number of input devices ranging from low-cost gaming controllers to research-grade robotic devices providing multi-modal feedback and augmented guidance. Our findings suggest that underlying movement characteristics can differentiate skill level (experts versus novices) in novel motor tasks, and can differentiate motor impairment severity in stroke populations. There is the potential for this motion-based analysis of performance to carry over to the surgical domain, where virtual reality is becoming a prominent feature of educational, residency, and certification programs. To objectively assess performance, robust tasks must be identified that can differentiate performance, an especially challenging task in the domain of robotic surgery, where advanced visualization and teleoperation technologies significantly enhance the performance of the

surgeons on fundamental tasks compared to the use of laparoscopic instruments. Such tasks are necessary for proper validation of simulator systems in order to verify retention and transfer of skill to the operating room.

Acknowledgements

Portions of this work have been support in part by grants from the National Science Foundation (IIS-0448341 and IIS-0812569) and Mission Connect, a project of the TIRR Foundation.

References

[1] Kozak, J. J., Hancock, P. A., Arthur, E. J., and Chrysler, S. T. (1993). Transfer of training from virtual reality. *Ergonomics 36,* 1, 777–784.

[2] Lintern, G. (1991). An informational perspective on skill transfer in human-machine systems. *Human Factors 33,* 3, 251–266.

[3] Lintern, G. and Roscoe, S.N. (1980). *Visual cue augmentation in contact flight simulation*. Aviation psychology. Iowa State University Press.

[4] Gamberini, L. (2000). Virtual reality as a new research tool for the study of human memory. *CyberPsychology and Behavior 3,* 3, 337–342.

[5] O'Malley M.K., Gupta, A., Gen, M., & Li, Y. (2006) Shared control in haptic systems for performance enhancement and training. *ASME Journal of Dynamic Systems, Measurement and Control, 128(1),* 75–85.

[6] Li, Y., Huegel, J., Patoglu, V. & O'Malley, M. K. (2009) Progressive Shared Control for Training in Virtual Environments, *Proceedings of the International Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems and the Third Joint World Haptics Conference (HAPTICS)*, pp. 332-337.

[7] Li, Y., Patoglu, V., & O'Malley, M.K. (2009). Negative efficacy of fixed gain error reducing shared control for training in virtual environments. *ACM Transactions on Applied Perception*. 6(1): 3-1 – 3-21.

[8] Huegel, J. C., & O'Malley, M.K. (2010). Visual versus haptic progressive guidance for training in a virtual dynamic task, *Proceedings of the International Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems and the Third Joint World Haptics Conference (HAPTICS)*, 399-400.

[9] Waller, D., Hunt, E., and Knapp, D. (1998). The Transfer of Spatial Knowledge in Virtual Environment Training. Presence: Teleoper. Virtual Environ. 7, 2, 129-143.

[10] Rose FD, Attree EA, Brooks BM, Parslow DM, Penn PR, Ambihaipahan N. (2000). Training in virtual environments: transfer to real world tasks and equivalence to real task training. Ergonomics. 43(4):494-511.

[11] Tendick, F., Downes, M., Goktekin, T., Cavusoglu, M. C., Feygin, D., Wu, X., Eyal, R., Hegarty, M., and Way, L. W. 2000. A Virtual Environment Testbed for Training Laparoscopic Surgical Skills. Presence: Teleoper. Virtual Environ. 9, 3 (Jun. 2000), 236-255.

[12] Basdogan, C., Ho, C.-H., and Srinivasan, M.A. (2001) Virtual environments for medical training: graphical and haptic simulation of laparoscopic common bile duct exploration. IEEE/ASME Transactions on Mechatronics, 6(3): 269-285.

[13] Li, Y., Patoglu, V., & O'Malley, M. K. (2006). Shared control for training in virtual environments: Learning through demonstration? In *Proceedings of EuroHaptics,* 93–99.

[14] Morris, D., Tan, H., Barbagli, F., Chang, T., & Salisbury, K. (2007). Haptic training enhances force skill learning. In *Proceedings of the IEEE Second Joint EuroHaptics Conference and Symposium on Haptic Interfaces for Virtual Environments and Teleoperator Systems* (*WHC '07*), 21–26.

[15] Huang, F. C., Gillespie, R. B., & Kuo, A. D. (2007). Visual and haptic feedback contribute to tuning and online control during object manipulation. *Journal of Motor Behavior, 39*(3), 179–193.

[16] Israr, A., Kapson, H., Patoglu, V., & O'Malley, M. K. (2009). Effects of magnitude and phase cues on human motor adaptation. In Proceedings of the IEEE Third Joint EuroHaptics Conference and Symposium on Haptic Interfaces for Virtual Environments and Teleoperator Systems (WHC '09), 344–349.

[17] Flash, T., & Hogan, N. (1985). The coordination of arm movements: An experimentally confirmed mathematical model. *Journal of Neuroscience, 5*(7), 1688–1703.

[18] Svinin, M., Goncharenko, I., Zhi-Wei, L., & Hosoe, S. (2006). Reaching movements in dynamic environments: How do we move flexible objects? *IEEE Transactions on Robotics, 22*(4), 724–739.

[19] Celik, O., O'Malley, M.K., Boake, C., Levin, H., Yozbatiran, N., & Reistetter, T. (2010). Normalized movement quality measures for therapeutic robots strongly correlate with clinical motor impairment measures, *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 18(4): 433-444.

[20] Huegel, J., Celik, O., Israr, A., & O'Malley, M.K. (2009). Expertise- based performance measures in a virtual training environment. Presence: Teleoperators and Virtual Environments, 18(6), 449-467.

[21] Cronenwett JL. Vascular surgery training: is there enough case material? *Semin Vasc Surg.* Dec 2006;19(4):187-190.

[22] Schanzer A, Steppacher R, Eslami M, Arous E, Messina L, Belkin M. Vascular surgery training trends from 2001-2007: A substantial increase in total procedure volume is driven by escalating endovascular procedure volume and stable open procedure volume. *J Vasc Surg*. May 2009;49(5):1339-1344.
[23] Bismuth, J., M.A. Donovan, M.K. O'Malley, H.F. El Sayed, J.J. Naoum, E.K .Peden, M.G. Davies, and A.B. Lumsden. (2010) Incorporating simulation in Vascular Surgery, Journal of Vascular Surgery, 52(4): 1072-80.
[24] Lepor H (2009) Status of radical prostatectomy in 2009: is there medical evidence to justify the robotic approach? Rev Urol. 11: 61-70.
[25] Samadi D, Levinson A, Hakimi A, Shabsigh R, Benson MC (2007) From proficiency to expert, when does the learning curve for robotic-assisted prostatectomies plateau? The Columbia University experience. World J Urol 25(1): 105-110.
[26] Judkins TN, Oleynikov D, Stergiou N (2009) Objective evaluation of expert and novice performance during robotic surgical training tasks. Surg Endosc 23(3): 590-597.
[27] Narazaki K, Oleynikov D, Stergiou N (2007) Objective assessment of proficiency with bimanual inanimate tasks in robotic laparoscopy. J Laparoendosc Adv Surg Tech A 17(1): 47-52.