Orthographic Vision-based Interface for Robot Arm Teleoperation

Waleed Uddin, Maram Sakr, Camilo Perez Quintero
and H.F. Machiel Van der Loos

***Abstract*— Robot teleoperation is crucial for many hazardous situations like handling radioactive materials, undersea exploration and firefighting. Visual feedback is essential to accurately teleoperate a robot. Existing solutions to improve teleoperation involve the use of multiple cameras, expensive sensors, depth cameras or VR/AR headsets. These systems, however, have some limitations including: safety hazards, complexity, cost, and inadaptability. Contrary to the existing work, we provide a simple, cost-effective and intuitive teleoperation system by visualizing the remote environment in an effective way to provide depth information using only one inexpensive webcam. To validate our system we perform a pilot study where users teleoperate 6-DOF arm and gripper to complete a pick and place task. We compare our proposed interface with a binocular camera setup. In addition, we test three input modalities with our interface: joystick, keyboard and Leap Motion. We use completion time and object manipulation accuracy as evaluation metrics. Results from the pilot study suggest that our interface in comparison with the binocular camera visualization improves completion time by 64%, 43% and 41% for the joystick, keyboard and Leap Motion, respectively. Furthermore, the number of errors declined using our vision system regardless of the control modality used.**

I. INTRODUCTION

Teleoperation refers to a task done by a robot which is remotely controlled by a human operator. Over the years, the use of teleoperation has become popular in several areas such as military [1], space exploration [2], underwater exploration [3] and tele-surgery [4]. The teleoperation system's performance is directly affected by the sensory information, visualization of the remote environment, control interface and operator capabilities. Teleoperating a robot is a cumbersome task for non-experts if the system is unintuitive. Humans prefer natural communication and control interfaces with the robot. To get such a system in a teleoperation setup, an informative and simple visual interface with an intuitive control system is crucial.

Thus, our paper focuses on these two key components of the teleoperation system. Conventional input modalities such as a joystick are considered non-intuitive and are detrimental to overall task efficiency [5]. With the advent of motion-sensing devices, HRI researchers have shifted focus on devising more intuitive teleoperation interfaces [6]. Quintero et al. [7] recently developed a novel semi-autonomous means to control a robotic arm. Using the Kinect skeleton tracker, they mapped the human arm joints to robot joints, providing fast but coarse positioning of the robot arm. To mitigate this effect they introduce a visual servoing interface for fine positioning of the arm. Kim et al. [8] presented a master-slave direct control interface for an excavator using sensors placed on the operator's hand. The results of that study in comparison with joystick control were promising but not conclusive. These input strategies have not successfully been put into commercial use primarily due to safety concerns, design complexity, equipment costs and inadaptability.

In addition to the input modality adapted for teleoperation, the level of enrichment and information provided in a feedback modality is crucial. A camera is typically used to provide visual feedback of the teleoperation environment to the operator. However, the 2D video seen on a screen fails to provide depth information for the environment. Without depth perception, the operator is likely to make errors during teleoperation [9]. To solve this problem, people have tested teleoperation with multiple cameras placed at different locations to acquire depth perception for the environment and improve task efficiency [10], [11]. However, the option of using multiple cameras is costly, adds complexity to the system and requires additional space. Moreover, using another camera can also cause object occlusion [12]. Vision-based high-fidelity depth cameras like the Microsoft Kinect in teleoperation systems have also been proposed [7] along with recent advances in AR and VR devices, which provide immersion and telepresence to the operator. Peppoloni et al. [13] implemented a 3D augmented reality-based visual feedback system to teleoperate a Baxter robot.

These efforts to solve the problem of depth perception are commendable; however, the efficacy of these systems relies on expensive sensors or cameras, and the design itself may pose additional complexity. Furthermore, these systems may require special training to gain familiarity. Prolonged use of VR or AR headsets in teleoperation systems may also cause VR sickness [14].

Building on this existing work, we provide a simple, cost-effective and intuitive teleoperation system by focusing our efforts on visualizing the remote environment in an effective way to provide depth information using only one inexpensive webcam. In addition, we provide a comparison between three control modalities: joystick, keyboard and the Leap Motion. Thus we aimed for finding the best modality combined with our proposed visual system to achieve a balance in task completion time and object manipulation accuracy. Our main contributions in this paper are the following:

- Providing depth information along with visual feedback in teleoperation system using only one inexpensive ordinary camera.
- Comparing between different input modalities for robot

Collaborative Advanced Robotics and Intelligent Systems (CARIS) Laboratory, Mechanical Engineering, University of British Columbia.

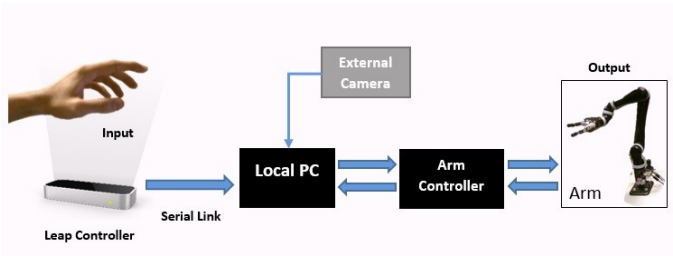

Fig. 1. Simplified description of our system with Leap device as input

arm teleoperation.

In the rest of this paper, we describe the proposed system in section II. Then, we present the pilot user study that we conducted to evaluate our system and compare between different modalities in section III. Section IV provides the results of the study, then a discussion of the results is provided in section V. Lastly, we conclude the paper and propose some future directions in section VI.

II. SYSTEM DESCRIPTION

A. Overview

We propose a direct unilateral and cartesian-based control of a 6-DOF robot in real-time. The operator can control the arm using one of three modalities: joystick, keyboard or Leap Motion [15] (hand motion tracking device). Fig. 1 shows our system in its simplified form when Leap Motion controller is being used as input method. This Leap Motion controller is connected to a local computer using a serial link. An external camera is responsible for providing a view of the remote environment and sending it to the local computer. Using computer-vision techniques, depth information of the remote location is added to the camera view to make it easier for the user to teleoperate the robot.

Fig. 2 shows the physical setup our teleoperation system. In the remote location, a robotic arm is used to accomplish a pick and place task. The object to be picked is placed on top of a box that has a QR code to help in capturing its position with the camera. While in the operator's location, a computer is used to process the camera views and to handle the communication between the robot and the control interface. Distances between camera and the arm as well as the arms end-effector with respect to its origin are shown in Fig. 3. The camera is fixed at a distance such that its field of view captures the robot and its surrounding environment. These distances are crucial for creating an effective visualization.

Our system is evaluated on a 6-DOF Kinova Jaco2 arm. Actuators are geared DC servomotors, which operate at 24VDC and have built-in encoders for sensing joint angles. For our research, we operate it in Cartesian control mode.

Communication with the Kinova arm is done through a USB port. Commanded Cartesian commands are sent to the robot. The robot controller calculates and rotates respective joints through its inverse kinematics module to achieve the commanded Cartesian pose. The arms internal joint encoders provide actual pose information back to the computer, where

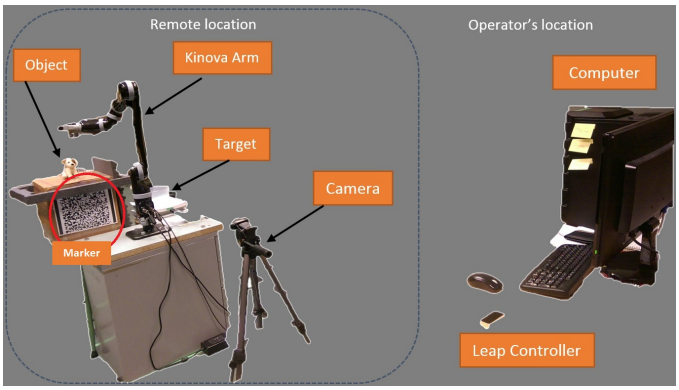

Fig. 2. Physical Setup of the teleoperation system

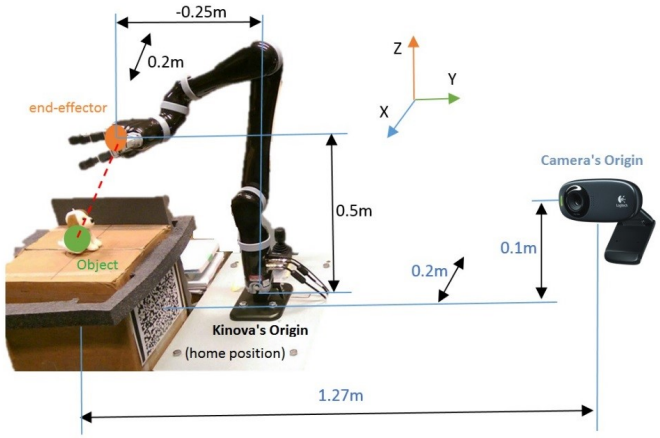

Fig. 3. Physical Distances in our system

it can be visualized along with the commanded pose. Our software is composed of four modules that interact with each other simultaneously, as shown in Fig. 4.

The main communications with the robot's DSP using the Kinova API are functions that send a desired Cartesian trajectory and get an actual Cartesian pose. Parameters that are passed are in the form of a data structure. The arrays "ActualPose" and "CommandedPose" hold Cartesian information about actual and commanded pose. They contain the following float variables:

$$ActualPose = [x_p, y_p, z_p, \theta_{xp}, \theta_{yp}, \theta_{zp}, f_{1p}, f_{2p}, f_{3p}] \quad (1)$$

$$CommandedPose = [x_t, y_t, z_t, \theta_{xt}, \theta_{yt}, \theta_{zt}, f_{1t}, f_{2t}, f_{3t}] \quad (2)$$

These structures are passed to the Jaco2's API functions. The f variables represent fingers of the arm. Variables x,y and z are the Cartesian coordinates while θ variables are wrist coordinates.

B. Input Modalities

We control the arm in Cartesian mode, i.e., moving the end-effector in x, y and z directions directly. Using our teleoperation system, arm can be controlled via joystick, keyboard or the Leap Motion controller as follows:

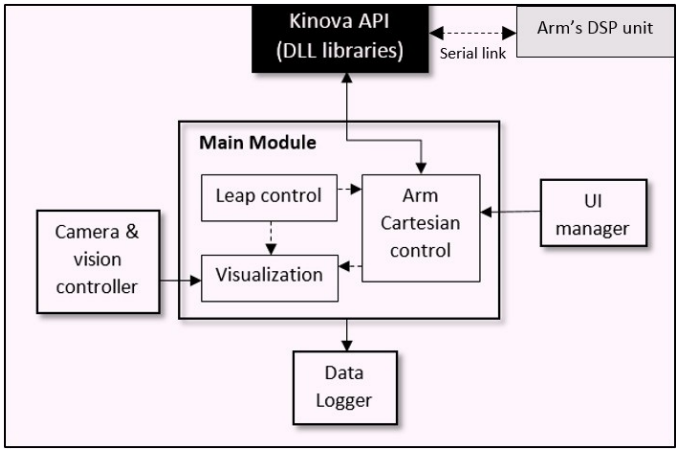

Fig. 4. Program modules for system

Joystick: A joystick is available from the manufacturer to control the robot. Rotating it left or right moves the robot's end-effector sideways in the x-axis while forward and backward moves it in the y-axis. For movement along z-axis (up and down), the joystick handle is rotated clockwise or counter-clockwise.

Hand Motion Tracking: We use the Leap Motion controller for real-time marker-less hand motion tracking. This controller is geared to small-scale VR development applications. Fig. 1 shows the Leap Motion in our system. Hand movements are captured by the Leap Motion controller at a rate of approximately 115 Hz. Hand coordinates x, y, z in meters are sent to a local computer. A program script in C# handles the incoming data from the Leap Motion and communicates with the Kinova arm through its API over a separate USB link. The processed x, y, z values are finally sent to the robot controller in the form of a Cartesian command data structure.

Keyboard: The robot end-effector moves forward with the W key, backward with the S key, left and right with the A and D keys, and up and down with the E and Q keys. The Space key is to grasp an object, i.e., the gripper closes, while the X key opens the gripper to drop the object. We chose these keys because the mapping has been used extensively with video games, making it easier for users to memorize.

C. Visualization

Our visual interface consists of an orthographic visualization of the scene as shown in Fig. 5. The environment is set up using the Unity 3D engine. The normal front view is provided to the user along with a vision-based top view. This top view is created using vision-based depth information. A black ball represents the robot end-effector's Cartesian movement in x, y, z directions in real-time in both views. A green ball represents the operator's hand movements in x, y, z directions if the Leap Motion device is being used as the input modality. The green ball serves as the Cartesian command and the black ball as the follower. The blue cube in the top view is the representation of the physical box upon which the toy object is placed at center, calculated by the

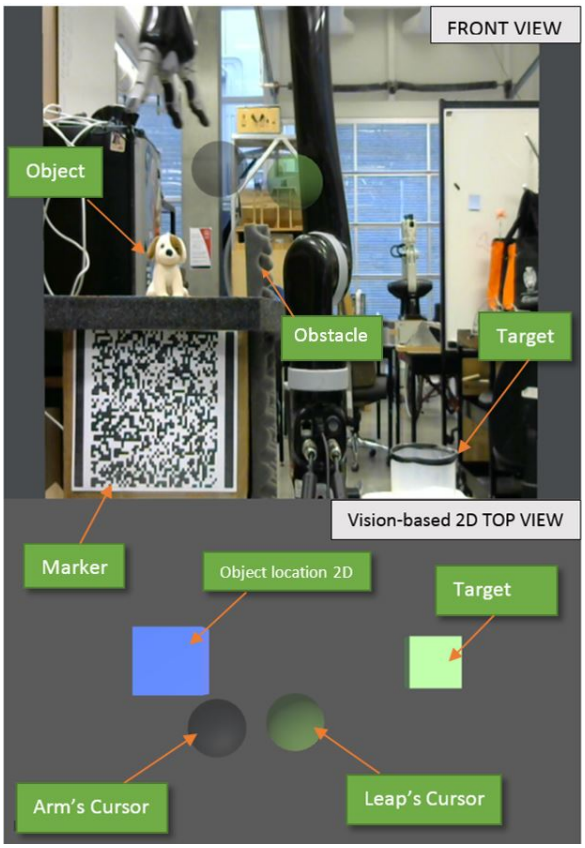

Fig. 5. Orthographic visualization of the remote location

vision-based camera. The blue cube is not the representation of the toy rather it is the representation of its physical location. If the box is moved, the 2D representation of the box moves accordingly. Similarly, the green cube represents the target where the object is to be dropped. The problem of depth perception is solved using the marker-based vision system described as follows.

In our system, a single webcam is placed at a set distance from the robot as shown in Fig. 3 earlier. The camera has two purposes:

1) To provide visual feedback: a normal 2D frontal view of the robot and environment in x and z axis.
2) To use computer-vision for depth calculation: as shown in Fig. 6, vision-based camera computes distance of marker attached to the box, relative to itself based on a natural phenomenon of object perception. This depth information can be visualized on the screen and the operator can then control the arm accurately with perceptive information about all three dimensions

The apparent size of the marker depends upon the visual angle experienced by the camera and not the actual size. As this visual angle seems to be proportional to the apparent size [16], we can detect changes in size of the marker.

As the real size of the marker is constant and known, any change in the apparent size of the marker would mean that the marker is either coming closer or moving away from the camera. If we move the marker towards camera, the apparent

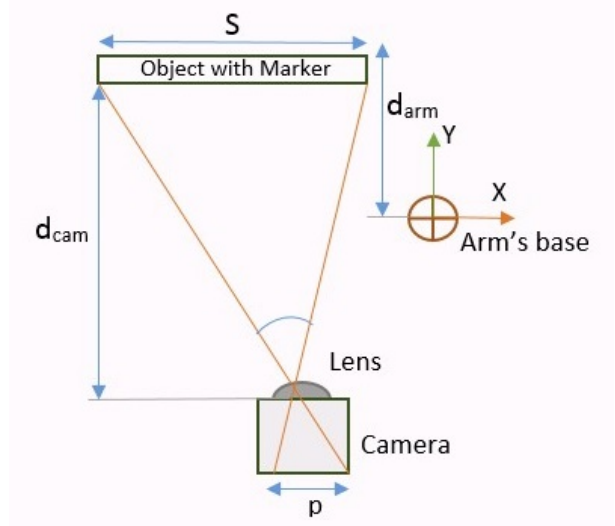

Fig. 6. Representation of marker-based depth measurement (top view)

size of the marker experienced by camera gets bigger. As the distance correlates with the apparent size [17], we can use it to compute the actual distance of object from the camera. Fig. 6 shows the described concept.

Distance of the marker relative to the camera is computed through computer-vision based API inside our 3D visualization software that calculates distance using the following function:

$$d_{cam} = f(S, p) \tag{3}$$

where, S is width of marker which is 25 cm, attached to a box. d_{arm} and d_{cam} are distances of marker relative to the arm's base origin and the camera origin. p is the marker image size perceived by the camera. The camera and arm bases are fixed. d_{cam} is calculated first using camera vision API functions then d_{arm} is simply calculated as an offset.

The depth information d_{arm} is used to visualize the distance of the object in the y-axis with respect to the robot's base location as shown in the top view of Fig. 5. By presenting this top view using single camera along with a normal frontal view to the operator, we provide perceptive information regarding all three dimensions to the operator.

Although researches involving marker-based object detection through computer vision have been conducted for Augmented-Reality (AR) systems in past [18], yet provision of depth perception utilizing such approach specifically for teleoperation systems is something that has not been explored before to the best of our knowledge.

III. User Study

In order to assess the overall efficacy of our vision-based teleoperation system, we compared our system against a conventional teleoperation system consisting of two cameras displaying front and top view lacking any visualization aids or computer vision [19]. In addition, we compare between three input modalities: joystick, keyboard and leap motion using the two visual systems. We recruited three participants (two males, one female) from the University of British Columbia who had no prior experience with robot teleoperation. A consent form approved by the ethics committee was provided to each participant for signature prior running the experiment which included details of experiment. The study lasted 50-60 minutes. The participants were asked to complete a pick and place task which consists of picking up a small toy in a gentle way and moving towards the destination while avoiding obstacle and then dropping the toy in the target container. The obstacle is a wall, placed in the way to the target position. In addition, the toy was placed in the beginning on a box that can be opened if the participant pressed hard on it while picking the object up. This allows us to judge if the participants pick up the object gently or not. Each participant went through the following steps to accomplish the task:

A. Procedure

- Participants were introduced to the experimental setup via verbal briefing. Then, each participant was provided five minutes to familiarize himself with the system which includes the three input modalities and two visual interfaces. The participant was only allowed to look into the screen that showed either the top and front view using standard vision system or one camera view with marker vision-based top view using our proposed system.
- Using the standard vision system, participants were asked to complete the pick and place task using Joystick, Leap controller and Keyboard three times each, with total of nine trials, provided in a random order.
- Then using our vision-based system, participants completed the same task using the three input modalities for nine trials in a random order to fairly compare between the three modalities.
- At the end, participants were asked to complete a survey involving qualitative metrics to rate the performance of each input modality subjectively.

B. Performance metrics

To evaluate our system, we used the following as performance metrics:

- Task Completion time: this is the time taken by the participant to complete the task. We recorded the completion time in each trial for all three input modalities and the two visual interfaces.
- Number of errors: the errors we refer to in this context are the cases in which the participant pressed the box underneath the object to pick up (tough pickup). We also counted the attempts to pick up the object, if the participant did not successfully pick it up the first time. Also, when the participant hit the wall between the source and destination. Lastly, the case in which the participant did not drop the object in its target position.

Using the above metrics, we compared between the three modalities using both standard vision system (two cameras) and our proposed orthographic vision system. In addition,

we measured the performance improvement rate for each modality.

IV. RESULTS

In this section, we present results for task completion time and total number of errors collected from all 3 participants during pilot experiments. Fig. 7 shows the completion time averaged across all participants and trials for the three modalities: joystick, keyboard, and leap motion using the standard vision system and our proposed system. It is clearly shown that, our proposed vision system outperforms the standard system regardless the input modality used. Our system shows a significant decrease in task completion time with 64%, 43%, and 41% compared to the standard system using joystick, keyboard, and leap motion, respectively. In addition, with our proposed vision system using leap motion shows the lowest completion time with an average of 28.67 $\pm$ 6.58 compared to 39.78 $\pm$ 6.28 using keyboard and 48.33 $\pm$ 12.59 using joystick. Using the same performance metric, we also noted the general trend of the performance improvement within the training trials of each input modality with both visual interfaces as shown in Fig. 8. Overall, the task completion time decreased with the trials increased except with the case of using joystick with the standard vision system. In addition, it is clearly shown that joystick improvement rate is higher than the rate of both keyboard and Leap Motion. Also, Leap Motion achieves the lowest completion time in all trials compared to the other two modalities.

In terms of committed errors as shown in Fig. 9, the number of errors is comparable among the three modalities. Similar to completion time, our proposed vision system shows higher performance than using the standard system among all modalities. Using our vision system has average number of errors around 1 $\pm$ 1, 0 $\pm$ 1, and 1 $\pm$ 1 using joystick, keyboard, and leap motion, respectively. Compared to 4 $\pm$ 4, 2 $\pm$ 1, and 2 $\pm$ 1 using the same modalities, respectively with the standard vision system.

To assess the system performance in terms of perceived cognitive and physical workload, the participants were asked to fill out a NASA-TLX Load index form after the experiment had finished. Regarding satisfaction of task completion using each input modality, two participants rated Leap Motion controller above keyboard while the third participant rated keyboard first following the joystick commenting that even though Leap Motion-based control was intuitive yet it was physically more demanding than keyboard or joystick specially if the task is too long. Leap Motion outperforms other input modalities in terms of mental demand for all participants.

V. DISCUSSION

Results from the pilot study suggests that our orthographic vision-based system outperforms the conventional teleoperation system with binocular camera setup. From experiments, we learnt that the inferior performance of the standard system as manifested in results, was mainly due to poor depth perception and object occlusion visually experienced at the operator side. Although there are two cameras that cover the front and top views, the participants found that it is really hard to accurately reach the target position. In addition, the participants perceived the binocular camera system as more mentally-demanding than ours because they need to switch between the two views to accomplish the task. While using the marker-based camera view provided a sort of guidelines that helped to finish the task easier. Furthermore, using the Leap Motion controller with the marker-based vision system leverage its capabilities and helped the user to finish the task with higher performance. Rodriguez et al. [20] used the Leap Motion for teleoperating a WAM robot and they found that

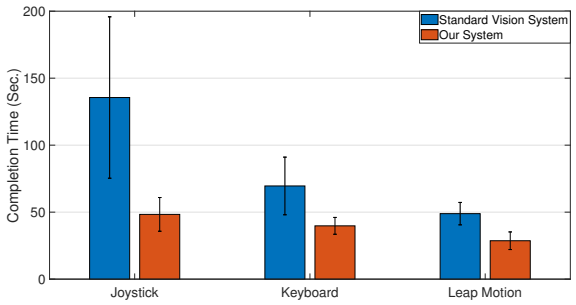

Fig. 7. The average completion time using the three modalities with both the standard vision system and our proposed system

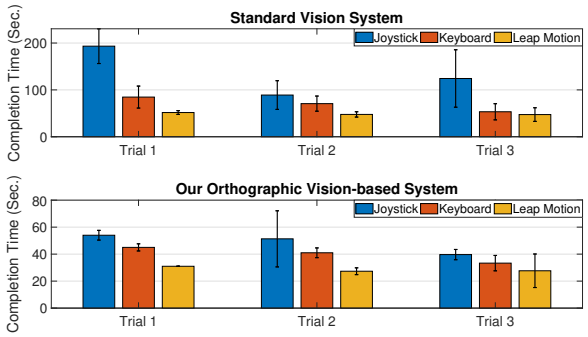

Fig. 8. The improvement rate of the completion time across three trials of the control modalities using two vision system

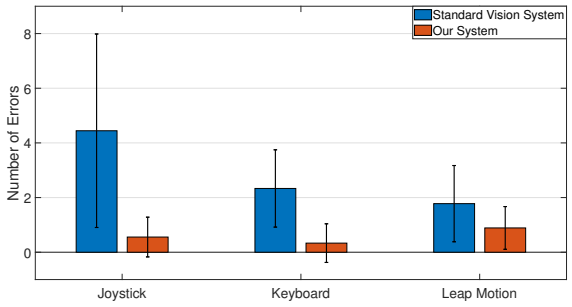

Fig. 9. The average number of errors using the three modalities with both the standard vision system and our proposed system

the Leap Motion is faster than the other interfaces but has poor precision. This is because the lack of visual feedback about the hand position which was proven to be effective with our experiments. In comparison between various input modalities, we found that joystick seemed to perform poorly due to its discrete movements and non-intuitive control while the Leap Motion controller seemed to perform slightly better than keyboard providing a continuous control to the operator. In addition, the task completion time improvement rate among the three trials using joystick is higher than the other two modalities. This suggests that joystick needs more time for training and it is not intuitive for non-experts to use. While keyboard and Leap Motion showed a slight improvement which suggested that the users can easily use them. Regarding our single-camera system, one drawback is its inability to construct the top-view if something is blocking the camera's view.

Conventional approaches in teleoperation such as binocular camera system, are more mentally-demanding than ours because they need to do mental calculations to perceive the depth compared to our system that presents the depth information in a simple way. This was proved from what Marble et al. [21] concluded from their study, in which most of their participants indicated a desire to have visual feedback in the teleoperation system presented with depth indicators rather than to have to deduce the depth from the interface.

This pilot study compellingly verifies the possibility of using hand motion-capture system coupled with a simple yet effective orthographic vision-based interface to greatly enhance the efficacy of teleoperation tasks.

VI. CONCLUSION AND FUTURE WORK

In this paper, we addressed the fundamental problems of perception and control experienced by the operator related to teleoperation systems. We put efforts on providing a simple, cost-effective and intuitive teleoperation system of a 6-DOF robot arm in Cartesian mode. We focused on visualizing the remote environment in an effective way by providing depth information using only one inexpensive webcam. In addition, we provided a comparison between three control modalities: joystick, keyboard and Leap Motion. Our pilot study involved three participants and consisted of a 'pick and place' task. Experiments involved comparison of our vision-based camera system with a conventional system consisting of two cameras that provide visual feedback to the operator. We tested task performance of both systems using joystick, keyboard and Leap controller. Results from Pilot studies showed that our vision-based system outperforms the conventional teleoperation system in terms of efficiency and accuracy. Among three input modalities, Leap Motion controller slightly outperformed the keyboard while joystick performed poorly. Since our pilot study results favor our proposed interface, we therefore in future, plan to conduct an extensive user study by involving more participants and evaluate our visual system and control modalities with different tasks. In addition, we plan to address problems related to the Leap Motion interface in terms of its physical demand in the long tasks.

REFERENCES

[1] T. Kot and P. Novák. Application of virtual reality in teleoperation of the military mobile robotic system taros. *International Journal of Advanced Robotic Systems*, 15(1):1729881417751545, 2018.

[2] G. Liu, X. Geng, et al. Haptic based teleoperation with master-slave motion mapping and haptic rendering for space exploration. *Chinese Journal of Aeronautics*, 2018.

[3] R. Saltaren, R. Aracil, et al. Field and service applications-exploring deep sea by teleoperated robot-an underwater parallel robot with high navigation capabilities. *IEEE Robotics & Automation Magazine*, 14(3):65–75, 2007.

[4] J. Marescaux, J. Leroy, et al. Transatlantic robot-assisted telesurgery. *Nature*, 413(6854):379, 2001.

[5] N. Mavridis, G. Pierris, et al. Subjective difficulty and indicators of performance of joystick-based robot arm teleoperation with auditory feedback. In *Advanced Robotics (ICAR), 2015 International Conference on*, pages 91–98. IEEE, 2015.

[6] B. Fang, F. Sun, et al. A novel data glove for fingers motion capture using inertial and magnetic measurement units. In *Robotics and Biomimetics (ROBIO), 2016 IEEE International Conference on*, pages 2099–2104. IEEE, 2016.

[7] C. P. Quintero, R. T. Fomena, et al. Interactive teleoperation interface for semi-autonomous control of robot arms. In *2014 Canadian Conference on Computer and Robot Vision (CRV)*, pages 357–363. IEEE, 2014.

[8] D. Kim, J. Kim, et al. Excavator tele-operation system using a human arm. *Automation in construction*, 18(2):173–182, 2009.

[9] J. S. Tittle, A. Roesler, et al. The remote perception problem. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 46, pages 260–264. SAGE Publications Sage CA: Los Angeles, CA, 2002.

[10] M. Fischer and D. Henrich. Surveillance of robots using multiple colour or depth cameras with distributed processing. In *Distributed Smart Cameras, 2009. ICDSC 2009. Third ACM/IEEE International Conference On*, pages 1–8. IEEE, 2009.

[11] D. Saakes, V. Choudhary, et al. A teleoperating interface for ground vehicles using autonomous flying cameras. In *2013 23rd International Conference on Artificial Reality and Telexistence (ICAT)*, pages 13–19. IEEE, 2013.

[12] E. Hsiao and M. Hebert. Occlusion reasoning for object detectionunder arbitrary viewpoint. *IEEE transactions on pattern analysis and machine intelligence*, 36(9):1803–1815, 2014.

[13] L. Peppoloni, F. Brizzi, et al. Augmented reality-aided tele-presence system for robot manipulation in industrial manufacturing. In *Proceedings of the 21st ACM Symposium on Virtual Reality Software and Technology*, pages 237–240. ACM, 2015.

[14] D. M. Romano. Virtual reality therapy. *Developmental medicine and child neurology*, 47(9):580–580, 2005.

[15] L. Motion. Leap motion controller. *URl: https://www. leapmotion. com*, 2015.

[16] T. Konkle and A. Oliva. Canonical visual size for real-world objects. *Journal of Experimental Psychology: human perception and performance*, 37(1):23, 2011.

[17] W. Epstein. The known-size-apparent-distance hypothesis. *The American Journal of Psychology*, 74(3):333–346, 1961.

[18] Y. S. Villegas-Hernandez and F. Guedea-Elizalde. Markers position estimation under uncontrolled environment for augmented reality. *International Journal on Interactive Design and Manufacturing (IJIDeM)*, 11(3):727–735, 2017.

[19] T. Fong and C. Thorpe. Vehicle teleoperation interfaces. *Autonomous robots*, 11(1):9–18, 2001.

[20] D. Rodriguez, C. Perez, et al. A comparison of smartphone interfaces for teleoperation of robot arms. In *Computer Conference (CLEI), 2017 XLIII Latin American*, pages 1–8. IEEE, 2017.

[21] J. L. Marble, D. J. Bruemmer, et al. Lessons learned from usability tests with a collaborative cognitive workspace for human-robot teams. In *Systems, Man and Cybernetics, 2003. IEEE International Conference on*, volume 1, pages 448–453. IEEE, 2003.