Int J Soc Robot (2012) 4:5–18
DOI 10.1007/s12369-011-0117-8

A Novel Evaluation Framework for Teleoperation and a Case Study on Natural Human-Arm-Imitation Through Motion Capture

Nikolaos Mavridis · Nikolas Giakoumidis · Emerson Lopes Machado

Accepted: 1 October 2011 / Published online: 19 November 2011

Abstract Although tele-operation has a long history, when it comes to tuning, comparison, and evaluation of tele-operation systems, no standard framework exists which can fulfill desiderata such as: concisely modeling multiple aspects of the system as a whole, i.e. timing, accuracy, and event transitions, while also providing for separation of user-, feedback-, as well as learning-dependent components. On the other hand, real-time remote tele-operation of robotic arms, either industrial or humanoid, is highly suitable for a number of applications, especially in difficult or inaccessible environment, and thus such an evaluation framework would be desirable. Usually, teleoperation is driven by buttons, joysticks, haptic controllers, or slave-arms, providing an interface which can be quite cumbersome and unnatural, especially when operating robots with multiple degrees of freedom. Thus, in thus paper, we present a two-fold contribution: (a) a task-based teleoperation evaluation framework which can achieve the desiderata described above, as well as (b) a system for teleoperation of an industrial arm commanded through human-arm motion capture, which is used as a case study, and also serves to illustrate the effectiveness of the evaluation framework that we are introducing. In our system the desired trajectory of a remote robotic arm is easily and naturally controlled through imitation of simple movements of the operator's physical arm, obtained through motion capture. Furthermore, an extensive real-world evaluation is provided, based on our proposed probabilistic framework, which contains an inter-subject quantitative study with 23 subjects, a longitudinal study with 6 subjects, as well as opinions and attitudes towards tele-operation study. The results provided illustrate the strengths of the proposed evaluation framework—by enabling the quick production of multiple task-, user-, system-, as well as learning-centric results, as well as the benefits of our natural imitation-based approach towards teleoperation. Furthermore, an interesting ordering of preferences towards different potential application areas of teleoperation is indicated by our data. Finally, after illustrating their effectiveness, we discuss how both our evaluation framework as well as teleoperation system presented are not only applicable in a wide variety of teleoperation domains, but are also directly extensible in many beneficial ways.

Keywords Tele-operation · Robots · Motion capture · Imitation · Evaluation · Learning

1 Introduction

Numerous application domains of robotics make the physical co-presence of human operators nearby the robot difficult, for example, hazardous or radioactive environments, space, etc. Furthermore, towards full-body android telepresence, teleoperation is implicated as one of the key supporting technologies. Quite some research on teleoperation has take place [4], but most systems rely on unnatural controllers, such as joysticks, which require previous training. Notable exceptions do exist, for example a demonstration of the benefits of using human natural arm movement for controlling an excavator [5]. In that paper, the authors claim to solve two problems usually related with excavators: high risk involved in the operation, and difficulty inherent of manipulation by joysticks. The authors use a combination of orientation sensor, rotary encoder, and inclinometer to read

N. Mavridis (✉) · N. Giakoumidis · E.L. Machado
Interactive Robots and Media Lab, CIT, UAE University,
17551 Al Ain, UAE
e-mail: nmav@alum.mit.edu

the human arm and hand movement and transmit the data to a computer through bluetooth, which then controls the excavator. In another related work [6], the authors used optical motion caption to copy the operator's arm and head movement to an android. Their intention was to create a natural human-like movement on the android as a way of improving the interaction between it and humans. In our system, we use real-time motion capture, for easy and intuitive teleoperation of an industrial arm, by imitation of human arms movements towards completing pre-specified tasks. Regarding the important problem of correspondence choice between imitator and imitated (robot and human in our case), the reader is primarily referred to the extensive analysis in Alissandrakis et al. [2], as well as to Alissandrakis et al. [1]. Apart from the design and implementation of teleoperation systems, another important aspect is the evaluation of the complete human-machine system. Although time-delay as well as limited spatial aspects of the performance of such systems has been reported, to the best of our knowledge, no task-based evaluation of the reaction of unskilled people during teleoperation of an industrial robotic arm has yet taken place.

2 System and Methods

In this section, we first introduce the evaluation framework, and then describe the specifics of our system, the task chosen for the evaluation, as well as a second subjective questionnaire-based evaluation that was carried out in parallel to our quantitative framework, in order to assess not only performance-related aspects, but also opinions and attitudes regarding teleoperation.

2.1 The Evaluation Framework

The secondary purpose of our evaluation framework was to investigate the effectiveness of design choices for the teleoperation system presented in this paper, through experiments with subjects that operated the system. However, foremost and most importantly, our primary purpose was to create and test an initial version of a more general task transcription and modeling framework which can be used to provide a firm basis for investigating effect of design choices, user variability, as well as aspects of user adaptation and fatigue, for a multitude of teleoperation systems and tasks.

2.1.1 The Framework

At the heart of our framework, there is a probabilistic model M made up from three components, as we shall see. In order to apply the framework, one needs to first choose a specific system S under evaluation, and then devise a task T, with:

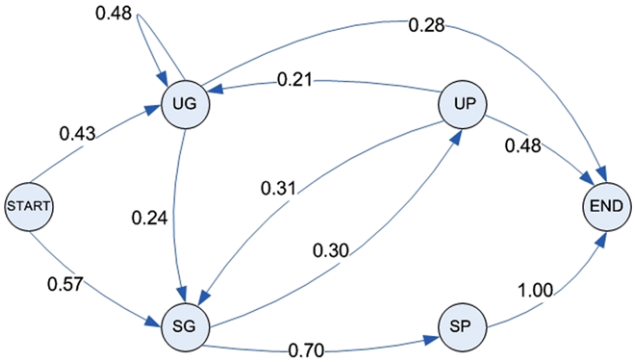

Fig. 1 Probabilistic Markovian model of task, based on Tij

($TC1$) Discrete Events (task start/end, ball grasp attempts/placements, vehicles reaching lines etc.)
($TC2$) Time Intervals between and/or within events
($TC3$) Result Metrics (as an example, placement accuracy measures for placement tasks)

Once the task is devised and appropriate $TC1-3$ are chosen, our goal is to estimate an appropriate probabilistic model $M(S, T, TC1-3)$, and a number of derived overall metrics Mi.

The resulting model probabilistic model M contains $\{Tij, P(\Delta\tau \mid SiSj), P(\alpha)\}$, i.e.:

T_{ij}: Transition Probabilities between the Discrete Events $TC1$
$P(\Delta\tau \mid SiSj)$: Time Interval distributions for $TC2$
$P(a)$: Result Metric distribution for $TC3$

As a first approximation, transitions between the discrete events can be assumed to be Markovian, an assumption that can be later justified if required on the basis of empirical data. As an illustration, the tri-partite model that was derived from our observations in the system described in this paper can be seen in Figs. 1, 2 and 3. Transition probabilities (Fig. 1) were calculated from the transition matrix resulting from our observations ($69 = 3$ balls $\times$ 23 subjects). Histograms of the transition time distributions are in Fig. 2, and of the placement accuracy distributions in Fig. 3. Extensive commentary/explanations on these figures is provided in the results section.

2.1.2 Overall Metrics Augmenting the Framework

The tri-partite model is also augmented with a set of overall metrics, which are reported as cumulative parametric statistics (in our case, mean, standard deviation, and median) which are:

($M1$) Task success rate (Success/Failure states of task)
($M2$) Total task time (first event to last event)
($M3$) Result Metric average (rel. to $TC3$)
($M4$) Number of states average

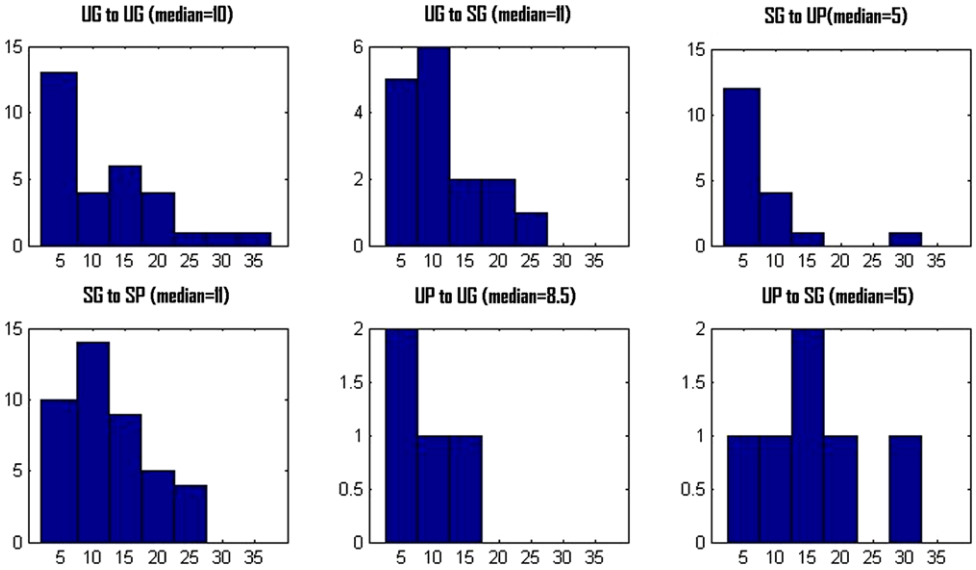

Fig. 2 Time interval distributions for state transitions, $P(\Delta\tau \mid SiSj)$

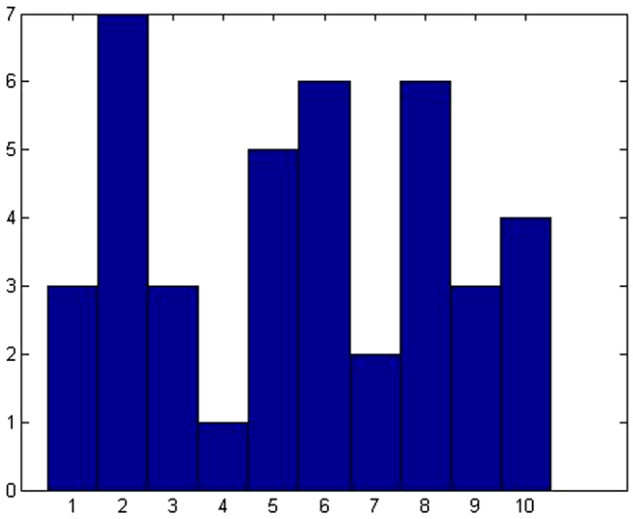

Fig. 3 Placement accuracy distribution for successful placements, $P(\alpha)$

As can be seen, overall metrics $M1$ and $M4$ are related to $TC1$, $M2$ to $TC2$, and $M3$ to $TC3$. These four overall metrics, can also be augmented with other such metrics for each specific task case, including: (a) metrics of the mix and ratios of states visited, as well as (b) metrics on possible scoring functions and components of scores that might have been given as maximization targets to the experimental subjects. For our specific case study (Fig. 4), there were six more such metrics, ($M5-10$), including three state mix metrics and three score-related metrics, as we shall see below.

2.2 The System

The tele-operation system consists of five major subsystems: Motion Capture, CyberGlove, TeleOp Controller, Robotic Arm, and User Feedback (Fig. 5).

2.2.1 Motion Capture Subsystem

The motion capture subsystem, consists of cameras operating at VGA resolution (640×480) supporting up to 200 fps (Standard Deviation brand). The cameras have infrared LED rings around them, and are placed at a height of 2.62 m on the corners as well as the short-side midpoints of a rectangle with size 6 m by 4.80 m. The effective capture area thus has a footprint of roughly 3 m diameter. The human is wearing a special suit on which 19 reflective ball markers of diameter 2.5 cm are placed. Three types of suits were used (Fig. 6): either strap-based for western-dressed humans, white traditional Emirati dresses for men, and black traditional Emirati dresses for women. The software API of the mocap system exposes a number of methods in C++, which enable the quasi-realtime readout of the 3D positions of the tracked markers.

2.2.2 CyberGlove Subsystem

The 5DT Ultra Series 14 gloves are used, which can produce 14 finger measurements plus 2 accelerometer readings. The gloves provide triggers for controlling the gripper of the robotic arm.

2.2.3 TeleOperation Controller Subsystem

The controller reads out the marker position timeseries from the motion capture, performs coordinate mapping and correspondence (Fig. 7), checks limits of movement, and issues appropriate commands to the robotic arm.

Correspondence Choice and Coordinate Mapping We have limited our choice of what the robotic arm should imitate to a simple hand position on a Cartesian space. Therefore, our system captures the human subject's right hand

Fig. 4 Table of overall metrics Mi augmenting the tri-partite model $\{Tij, P(\Delta\tau \mid SiSj), P(\alpha)\}$

M1) Task time per ball (first event to last event): Median 16sec, Mean 22.7sec, Std 20.7sec
M2) Task final states: Success 60.8%, Fail 39.2% (UP 11.6%, UG 27.6%)
M3) Accuracy for successful placements: Median 6cm, Mean 5.5cm, Std 2.9cm
M4) Number of states per trial: Median 2, Mean 2.73, Std 1.56
M5) Ratio of overall success. to uns. events: SG:UG = 0.92, SP:UP = 2.27 (UGs often repetitive)
M6) Ratio of probability of Success/Failure of Grip: P(SG|Start):P(UG|Start) = 1.32
M7) Ratio of probability of Success/Fail Placement: P(SP|SG):P(UP|SG) = 2.33
M8) Accuracy Comp. of Score (0=UP, 10=1cm Acc): Median 8, Mean 7.8, Std 5.6 (across 3 balls)
*M9) Time Comp. of Score ((600-Ttotsec)/600*30):* Median 23.2, Mean 22.0, Std 3.46
M10) Total Score (accuracy + time components): Median 29.85, Mean 29.8, Std 7.37

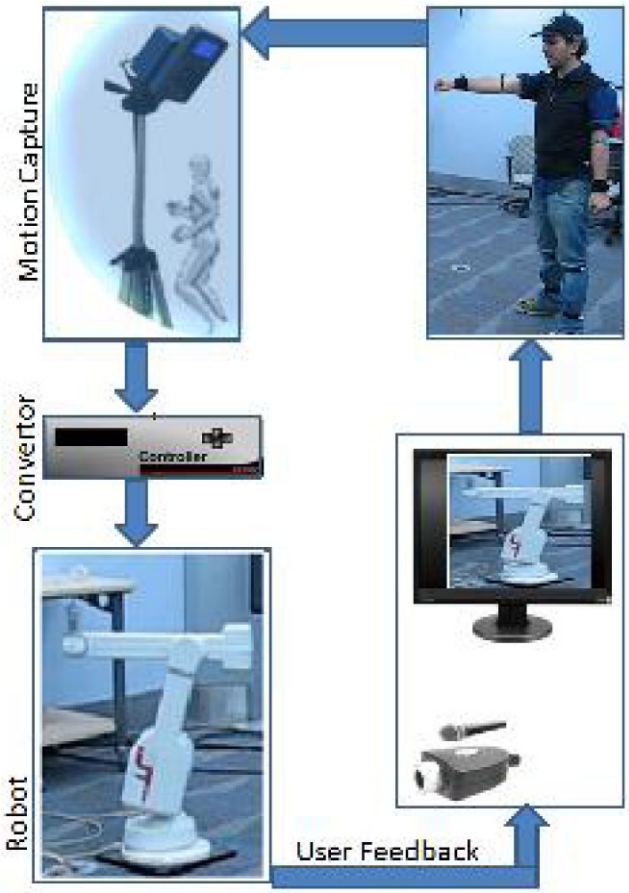

Fig. 5 Teleoperation system block diagram

position relative to his own arm position and maps it to the robotic arm's hand coordinates. The robotic arm controller takes care of doing the inverse kinematics and moving its motors in a way its hand goes to the requested position. Two markers on the subject are necessary to perform this operation: the right wrist marker (RW) and the right shoulder marker (RS). The 3D coordinates of the subject's hand are computed by subtracting the coordinates of the RW marker from the ones of the RS marker. As the robotic arm and the human arm are not of the same size, and even different among various subjects, we empirically computed a scale factor in order to match the subjects' fully extended arm to the robot' fully extended arm. Therefore, the coordinates acquired from the motion capture system are multiplied by this factor before being sent to the robot.

The robotic arm also has the ability of moving its hand relative to its wrist. We have chosen to map this movement to the subject left arm. Therefore, we have made a correspondence between the subject's left elbow angle to the direction of the robotic arm's hand. When the subject's left arm is fully extended, the robotic arm's hand points downward and when it is fully contracted, the robotic arm's hand points upward. To compute this angle, we used three markers: left wrist (LW), left elbow (LE), and left shoulder (LS). These markers can be seen as a triangle and, thus, the elbow marker can be easily computed using the cosine rule.

Limits of Movement Due to security reasons, we have limited the robotic arm's movement to 180 degrees on the X coordinate. This means that the subject can move its arm to points where the robot can't, but the robotic arm only goes where it would not crash into its surrounding objects and break itself. Furthermore, the robot's gripper is not allowed to go below the floor level, up to a safety margin.

Temporal and Software Aspects Another limitation of the robotic arm is on the data rate it can receive. Its controller ignores commands sent when it is performing a movement and thus we implemented a synchronized communication between it and the teleoperation controller. The tele-operation controller software was developed in Java 6 and integrated with the C++ API of the mocap subsystem, using Java Native Interface (JNI).

2.2.4 Robotic Arm Subsystem

The ST Robotics ST 17 manipulator arm is used, which has 5 degrees of freedom on the body, plus one for the gripper. Communication to the robot is achieved through a virtual serial port fed by IP, in the form of RoboForth messages. The workspace of the robot is contained within a hemisphere of one meter radius.

Fig. 6 MoCap suits with markers: Western (L), Emirati Women (M), Emirati Men (R). Notice the CyberGlove on the left hand in L

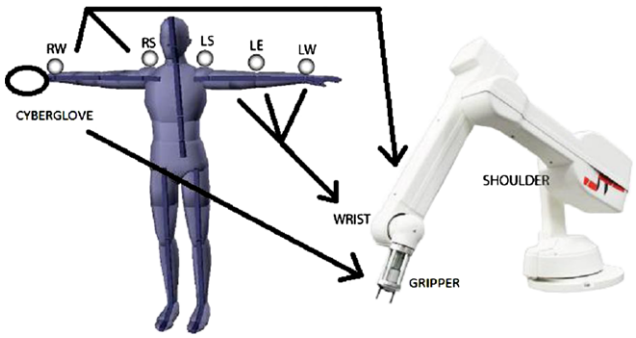

Fig. 7 Correspondence choice: Human R hand controls robot up to wrist, R palm controls gripper, L hand controls robot wrist

2.2.5 *User Feedback Subsystem*

User Feedback is provided through three channels: two visual, and one auditory. The visual channels are video feeds from two cameras placed in the robot's location: one on the gripper, and one on a tripod behind the robot, overlooking it from an angle. The video feeds are shown on two 40″ LCD screens in the room of the operator. Auditory feedback is provided through a microphone in the robot location driving a speaker system in the operator location. The feeds are delivered through proprietary camera software, and the VLC player has also been used in the past.

2.3 Task-Based Evaluation

The purpose of our evaluation was to provide real-world experience to our experimental subjects in tele-operation, to investigate the effectiveness of the design choices for our system, and also to create a task transcription and modeling framework which can be used to provide a firm basis for investigating effect of design choices, user variability, as well as aspects of user adaptation and fatigue.

2.3.1 *The Task*

The task chosen was to move three balls from their fixed home positions to their target positions. The task was designed so that it had intermediate difficulty, so that we can get meaningful results, without being neither impossible nor trivial. The layout of the workspace for the task is shown in (Fig. 8). The home positions had a 1.5 cm elevation from the floor, while the target positions had a paper underneath them with concentric circles marked with 1 cm–10 cm signs corresponding to the accuracy of the placement.

2.3.2 *Administering the Task*

The subjects were first exposed to a 5-minute introduction of system usage by the person who was administering the task (competent user). Then, the goal of their trial was made explicit: to try to move all the balls to the targets with maximum placement accuracy, as fast as possible, but within 10 minutes. An explicit scoring function was also given to them, in order to remove the arbitrariness of subjective weighing of the two components of the goal: first, the number of balls successfully placed (n), taking into account the accuracy of placements ($a1, a2, a3$) in cm and second, the total time (t) in minutes. I.e. the subjects were told that they had a maximum of 60 points, out of which $30 - (t \cdot 3)$ points for total time, and $10 \cdot (10 - a)$ points for the accuracy of each ball (which would default to 0 if the ball was not successfully placed). The main purpose of this function was to direct equal importance to both components of the goal for the subjects, so that they don't concentrate more on one of the components, and thus introduce bias. After the goal was made explicit, the subjects were given 5 minutes to play with the system, and then their up to trial time started (maximum allowed duration 10 minutes), during which video recordings as well as system log files were kept.

2.3.3 *Transcription and Modelling*

Each trial was analyzed on the basis of six different types of events: Start, Unsuccessful Grip (UG), Successful Grip (SG), Unsuccessful Placement (UP), Successful Placement (SP), and End. Each trial was thus transcribed as a sequence starting with a Start event, containing a number of UG, SG, UP and SP, and finishing with End. These event sequences were also augmented with the time intervals between the events. Transcription was done by humans on the basis of the video recordings. The chosen underlying model for the

Fig. 8 Task setup: robot, balls in home positions, and targets

observed data was a probabilistic automaton with 6 states, corresponding to the 6 events. The transition probabilities as well as the transition durations for this automaton were thus estimated on the basis of the observed data, as we shall see.

2.3.4 Repetitive Trials

While most of our subjects only had one trial on our system, we chose to perform repetitive trials for a subset of our subjects in order to start investigating learning and fatigue effects, as we shall see in the results section.

2.4 Opinions and Attitudes Evaluation

The purpose of this evaluation was to: (a) illuminate opinions and attitudes towards the use of tele-operation in different application domains, (b) assess the estimated emotional reaction of people in the subject's social circle towards the system, (c) to see whether the system demo stimulated subjects to learn more about robotics and tele-operation, (d) to gather comments for system improvements.

2.4.1 The Questionnaire

The questionnaire [7] had the form of a single-sided A4 sheet, and was available in two languages: Arabic and English, which the subjects could choose. It was partitioned in five parts: demographic questions, opinions and attitudes towards applications, estimated emotional responses, wanting to learn more, and suggestions/comments. The questionnaire can be seen in Fig. 9.

The demographic questions queried country of birth, age, sex, college education. Regarding (a) and (b), a 4-point modified likert scale was used (forced choice), with strongly disagree (1), slightly disagree (2), slightly agree (3) and strongly agree (4) boxes. Regarding (a) the seven application areas queried were: medical, workplace, child instruction, games, communication with people, dangerous environments, and space. Regarding (b) the four emotional responses queried were happy, comfortable, angry and afraid.

2.4.2 Administering the Questionnaire

The subjects were given the questionnaire in our lab after going through a standard five-minute introduction to teleoperation, during which a video of our system was shown, as well as a video of android teleoperation, and the benefits of the technology were explained. Most of the subjects that completed the questionnaire also tried out the system themselves.

3 Results

In this section, we present demographics for our subjects, as well as results for the task-based evaluation, as well as the extra questionnaire-based evaluation for opinions and attitudes.

3.1 Demographics

29 subjects completed the questionnaire, out of which 23 also tried out the system themselves.

Of the 29 subjects, 18 (62%) chose to complete the questionnaire in Arabic, and 11 (38%) in English. 24 of the 29 subjects, 24 (82%) were UAE nationals, 2 Iranians, as well as 1 Palestinian, 1 Greek, and 1 citizen of the USA. Their age ranged between 17, . . . , 43, while 23 out of the 29 subjects

Fig. 9 The questionnaire that was used for the subjective evaluation of opinions and attitudes towards teleoperation

Thanks you for agreeing to take our survey. Your participation is voluntary and your answers will be anonymous. We are asking these questions for research purposes, to help improve the design of novel Brain-Computer Interfaces and robot tele-presence applications.

1. Country of Birth :_____ 2. Age: ____ 3. Sex:____ Male ____Female

4. Do you have a college degree? : ____ YES _____ NO

5. Please circle the number that represents your agreement or disagreement with each of:

	Strongly Disagree	Slightly Disagree	Slightly Agree	Strongly Agree
I wouldn't mind if a tele-operated robot, helped me / treated me at the hospital	1	2	3	4
I wouldn't mind if a tele-operated robot, was in my workplace.	1	2	3	4
I wouldn't mind if my child was instructed by a tele-operated robot.	1	2	3	4
I wouldn't mind using robotic tele-operation technology to play games.	1	2	3	4
I wouldn't mind using robotic telepresence to communicate with other people.	1	2	3	4
I wouldn't mind if tele-operated robots were used in dangerous environments.	1	2	3	4
I wouldn't mind if tele-operated robots were used in space.	1	2	3	4
Many people in my social circle would feel happy if they saw this demo today.	1	2	3	4
Many people in my social circle would feel comfortable if they saw this demo today.	1	2	3	4
Many people in my social circle would feel angry if the saw this demo today.	1	2	3	4
Many people in my social circle would feel afraid if they saw this demo today.	1	2	3	4

7. After having seen this presentation today, do you want to learn more about robotics technology _____ Yes _____ No

8. After having seen this presentation today, do you want to learn more about tele-operation and telepresence _____ Yes _____ No

9. How would you want us to improve the system you saw: (use back of the page)

10. Other Comments?

were UAEU students aged between 17, . . . , 22 years old. The female to male ratio was 12:17, i.e. approximately 3:4.

3.2 Questionnaire-Based Evaluation

Here we present results from the extra questionnaire-based evaluation, which was performed in addition to the task-based evaluation. We start by presenting findings regarding attitudes towards different application areas, followed by estimated emotional stances of peers, expressed willingness to learn more about such systems, as well as further comments and feedback that was provided by the subjects that took the questionnaire. In the next section, we present the results of the task-based evaluation, in terms of the tri-partite probabilistic model that was introduced earlier.

3.2.1 Attitudes Towards Application Areas

The results of the attitudes towards the seven application areas are presented in (Fig. 10). One can observe that: (note that here, by agree we refer to the sum of slightly

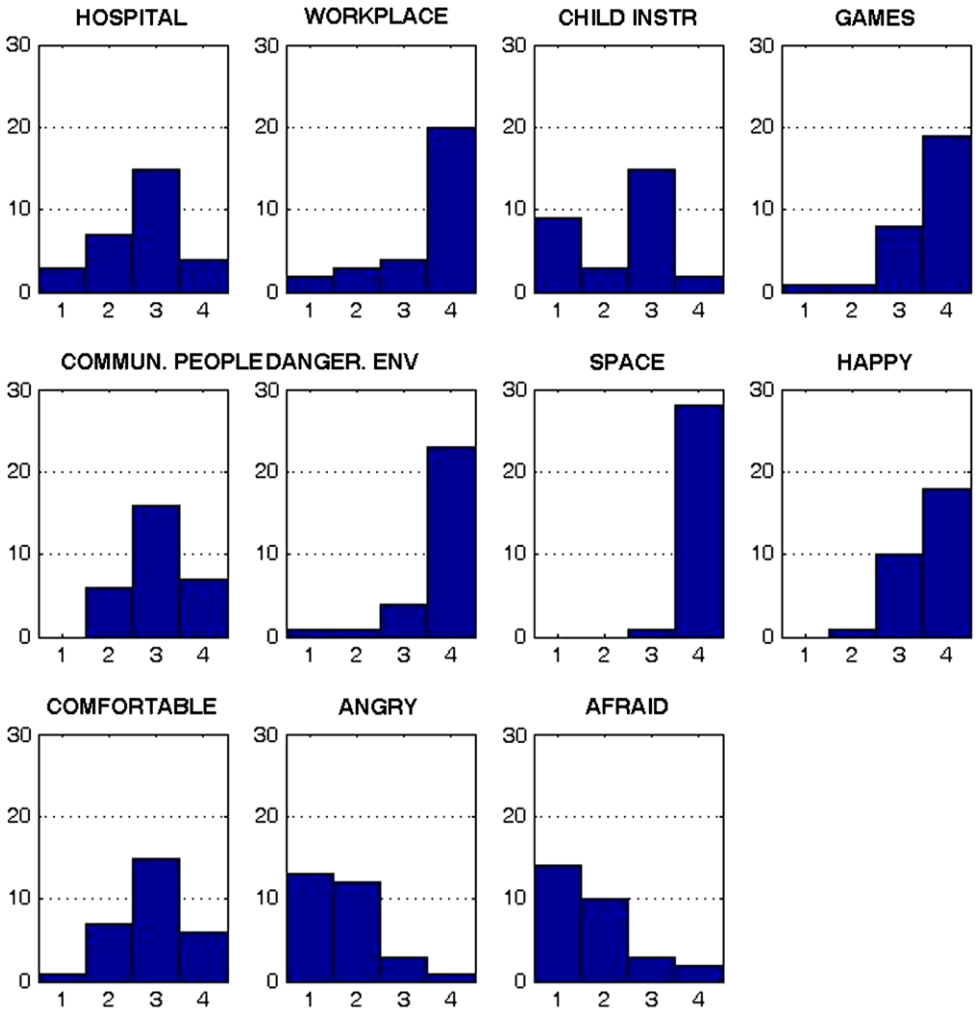

Fig. 10 Questionnaire results in histogram form for the seven application domains, and the four estimated emotions of peers

and strongly agree, and by disagree to the sum of slightly/strongly disagree, rounded to 1%)

(A1) Hospital:
66% agree, Most: slight agree
(A2) Workplace:
83% agree, Most: strong agree
(A3) Child Instruction:
31% strong disagr, 52% slight agree (bimodal)
(A4) Games:
93% agree, Most: strong agree
(A5) Communication with Humans:
79% agree, Most: slight agree
(A6) Dangerous Environments:
93% agree, Most: strong agree
(A7) Space:
100% agree, Most: strong agree

One can conjecture the following preference ordering for teleoperation applications (order of decreasing preference):

I. Strong Agree:
Space, Dangerous environments, Games
II. Slight Agree:
Workplace, Remote comms, Hospital
III. Bimodal Slight Agree/Strong Disagree:
Child Instruction

After a quick investigation, it was found that the sex (male/female) or age group $(17, \ldots, 22$ vs. $23, \ldots)$ of the subjects could not predict the two categories (slight agree or strong disagree) apparent in the bimodality of the attitudes towards the use of robots for child instruction.

3.2.2 Estimated Emotions of Peers

The estimated emotions of peers questions were querying four descriptors of affective states: happy, comfortable, angry, and afraid. The first two have positive valence, the second two negative. From the results in (Fig. 9), one can see that (see comments of subsection above):

Happy: 97% agree, Most: strong agree
Comfortable: 72% agree, Most: slight agree

Angry: 86% disagree, Most: strong disagree
Afraid: 83% disagree, Most: strong disagree

Thus, one can conjecture that subjects estimate that their peers (belonging to their social circle) would generally feel happy if they saw the demo. However, the subjects would only slightly agree that their peers would feel comfortable. In contrast, the subjects estimate that their peers would generally not feel angry or afraid.

3.2.3 Willingness to Learn More

Twenty five out of 29 subjects answered the two "willingness to learn more after demo" questions. All 25 (100%) answered positive to the question whether they wanted to learn more about robotics, while 2 out of 25 answered No regarding whether they wanted to learn more about teleoperation, and 23 out of 25 (92%) answered Yes.

3.2.4 Suggestions and Comments

Nine out of 29 subjects provided suggestions and comments, 6 of which contained suggestions, regarding speed, smoothness, delay, and size:

- Make it more smooth at move,
- I think it be smaller to be easy to use,
- Shorter delay more accurate movements etc.,

and 3 of which were congratulatory:

- It was great,
- I like it very much etc.

3.3 Task-Based Evaluation

In this section, we present results from the task-based evaluation. We start by presenting the overall metrics Mi, and then proceed to the derived tri-partite model $\{Tij, P(\Delta\tau \mid SiSj), P(\alpha)\}$.

3.3.1 Task-Based Evaluation: Overall Metrics

As mentioned above, the task was chosen in order to have intermediate difficulty, situated between the trivial and the impossible. For example, our gripper design and the soft balls used often resulted in unsuccessful grips, if the grip position was not precise enough.

Across 69 trials, we evaluated various overall metrics, which can be seen in Fig. 4.

3.3.2 Task-Based Evaluation: Derived Model

The probabilistic finite state machine that was derived from our observations can be seen in Fig. 1. Transition probabilities were calculated from the transition matrix resulting from our observations ($69 = 3$ balls $\times$ 23 subjects). Histograms of the transition time distributions are in Fig. 2, and of the placement accuracy distributions in Fig. 3.

3.3.3 Observations on Derived Model

The derived model, which can be packaged in the form of a transition matrix T, together with the six transition time distributions $P(\Delta t|Si, Sj)$, and the placement accuracy distribution $P(r)$, provides for a compact description of the performance of the system across users for a single trial, and overall metrics can generally be derived by it. Various observations follow directly: first, according to the score distribution, indeed we have a task which is neither trivial nor impossible; median and mean scores are very near the midpoint of the scale, with a decent amount of variance. Second, upon further analysis, lots of interesting patterns exist in the data: for example, have a look at Fig. 2: Following a successful grip (SG), there are two possible next events—a successful placement (SP) and an unsuccessful placement (UP). The time interval between SG and the next event though is a pretty good predictor of whether it will be successful or not: intervals above 7.5 sec most often lead to success—and whoever rushes often fails—as the time distributions of SG to UP vs. SG to SP seem to indicate. More such patterns remain to be explored, and can be quantitatively supported using probabilistic argument, given more empirical data. The most important observation though has to do with the possible semantics of the $\{T, P(\Delta t|Si, Sj), P(r)\}$ description given alternative experimental settings. One can thus ask: how can one try to deconvolve the effects of correspondence choice, user feedback channel, operator ability, learning, and fatigue through such models?

3.3.4 Toward insights on Learning

Towards investigating the previous question, and aspects of learning in particular, 7 out of the 23 first trial subjects were chosen to continue upon a longer-term study, which took place over 16 days, and was comprised of 4 sessions for each subject. The main goal in this experiments was to investigate the effect of learning, by examining the model $\{T, P(\Delta t|Si, Sj), P(r)\}$ and the resulting trajectories of the metrics Mi across the four sessions $T1-T4$. Graphs of the results follow below, together with relevant commentary. Where not otherwise noted, thick lines correspond to mean values, thin to median, and dotted lines to one standard deviation above and below the mean.

Figures 11 and 12 depict the trajectories of the total task time per ball (sec), as well as the number of states per ball, across the four sessions. From the figures, it becomes clear that as learning takes place, both the total task time per ball, as well as the number of states, decrease. The minimum number of states per ball is 2 (one SG followed by one SP); and indeed, while the mean number of states is still decreasing, quickly the median converges to 2. The total task duration per ball seems to have pretty much leveled

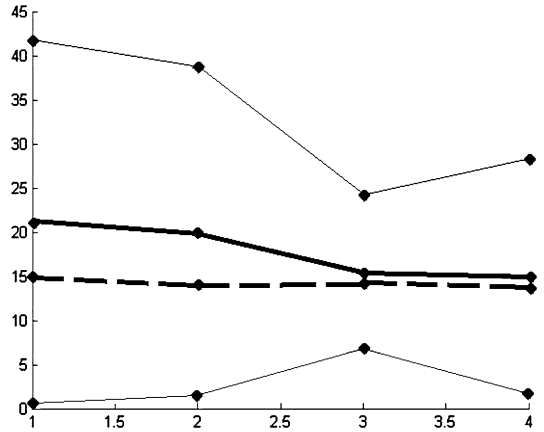

Fig. 11 Task time per ball (sec), across 4 sessions

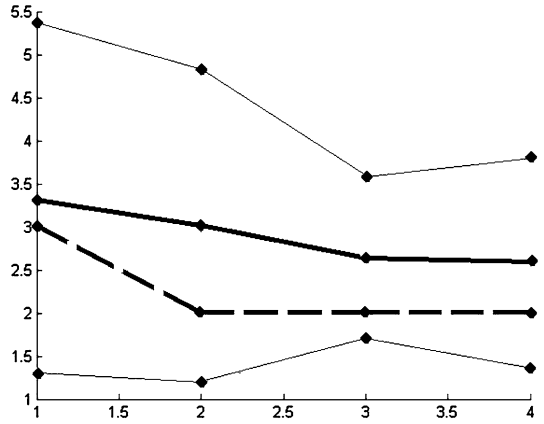

Fig. 12 Number of states per ball, across 4 sessions

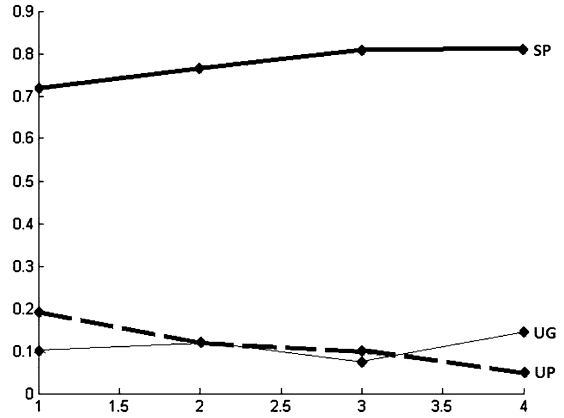

Fig. 13 Percentage of final states: *thick* = SP, *dotted* = UP, *thin* = UG

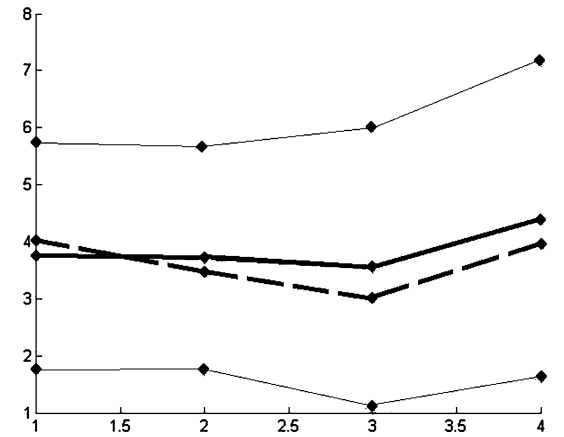

Fig. 14 Placement accuracy (cm), across 4 sessions

off within four learning sessions, to 16 seconds or so for our experimental setting. There is an apparent increase in variance between the 3rd and the 4th session; however no statistical significance for this effect can be claimed given our data. If such an effect does gather significance given future data, one possible explanation though is that after reaching best performance at the third session, there is a certain "relaxation/decrease of attention" or "fatigue" that each player experiences in subsequent sessions, thus creating a slight increase to the task completion time. These hypothesis could be examined given more data.

In Fig. 13, the resulting percentages for the possible final states of each ball are depicted. Initially, roughly 70% of all balls ended up in a successful placement (SP); at the fourth session, this percentage had leveled off to approximately 80%. At the same time, balls ending with unsuccessful placements (UP) start off at almost 20%, with 10% of attempts finished at UG, while at the fourth session, end-state UP's fall down substantially to 5%.

Interestingly enough, although the number of end-state SP's increases noticeably, the mean as well as median placement accuracy does not change significantly, as can be seen by Fig. 14, and even seems to become a little worse. In terms of net effect, this is compensated by the fact that now a much smaller number of placements are unsuccessful; so, even with a small decrease in overall accuracy, this is an overall positive trend.

Thus, so far: total task time as well as number of states seem to decrease during learning, and did level off within four sessions. Placement accuracy did not seem to decrease; but the number of balls ending with successful placements did increase, and leveled off too. Now, let us provide yet one more peak into learning trajectories for this system and task, as illustrated by Fig. 15 below, which was derived by the proposed $\{T, P(\Delta t|Si, Sj), P(r)\}$ model:

In this figure, the trajectories of the overall ratio of SP to UP, and SG to UG is depicted. Overall, the number of successful placements as compared to unsuccessful increased significantly across the four sessions. On the other hand, the ratio of successful grips as compared to unsuccessful grips, has increased from 0.71 to 1.15, albeit not in a monotonic fashion, and not in such a dramatic way.

Thus, the use of the $\{T, P(\Delta t|Si, Sj), P(r)\}$ model and the investigated derived metrics, as measured across four

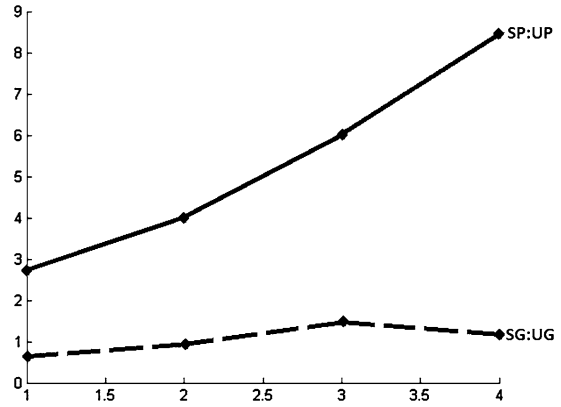

Fig. 15 Overall ratio of SP:UP (*thick*) and SG:UG (*dotted*)

learning sessions, was able to illuminate the following observations, which hold for our system and task:

(O1) There is a clear learning effect regarding total task time per ball, with a decrease of roughly 25% as compared to the initial time, as well as a decrease of variance. This effect levels off after four sessions.
(O2) The number of states per ball also decreases, with the mean falling from roughly 3.5 to under 3. There is a leveling effect in four sessions, but a small further decrease might be possible.
(O3) The final states of each ball, have a 10% increase of successful states (SP), starting off at 70% and leveling off at 80%. At the same time, the unsuccessful states are originally dominated by unsuccessful placements (20%) which fall off to 5%, leaving the rest to abandoned attempts after unsuccessful grips (UG).
(O4) Placement accuracy does not increase; and actually seems to be marginally decreasing over the four sessions; however, this is compensated by the big decrease of unsuccessful placement attempts, which are now turned to successful, albeit with not a very good accuracy.
(O5) Operator dexterity regarding successful placements increases considerably, as shown by the fast increasing SP:UP ratio. There is also an improvement regarding successful gripping (SG:UG), albeit not as large.

In summary, the above analysis indicates, for our system and task, that in 4 sessions:

– Total task time and states decrease (roughly 25% and 16%) and level off,
– Success rate starts at 70% increases and levels off at 80%,
– Placement accuracy does not improve, but
– Operator dexterity regarding grip success increases by 50%, and
– Operator dexterity regarding placement success increases considerably, by 300%.

3.3.5 Toward Insights on Fatigue

In order to get a quantitative overview of fatigue effects, two of the seven long-term trial subjects went through an extra long section, where they were instructed to keep on gripping and placing balls till they think they cannot operate the system anymore. These two sessions took place after all previous sessions had taken place, and they had a duration of 7 minutes 31 seconds and 8 minutes 7 seconds respectively, with $3 \times 6 = 18$ balls total. The initial gathered data indicated no discoverable patterns, which might well be so due to the small size of the sample regarding investigation of fatigue. Across the 18 balls, windowed as single balls or averaged across larger units, no clear trend regarding time, accuracy, or state-distribution patterns could be found. Thus, pending further data, no predictive patterns towards fatigue were observed.

4 Discussion

Many possible avenues for future extensions exist. Currently, we are pursuing an extension of the population taking part in our evaluations, and mainly the longer-term multi-session multi-trial evaluations towards insights on fatigue described above. Another avenue that we plan to pursue is the investigation of the effectiveness of our design choices of human- to robot- correspondence.

Initially, we had experimented with a glove-less system, in which a different degree of freedom of the human left hand was utilized for controlling the gripper. However, this was found to be highly confusing and difficult to learn for pilot subjects. Still, it is not clear that the current correspondence choice is by any means optimal; so further choices could be potentially investigated. Furthermore, it was noticed that the current two-camera setting for user feedback often does not provide an accurate perception of depth when approaching the ball, which results in misestimation and grip failures. Thus, we plan to investigate alternative camera placements for better results: for example, it was observed that placement of a camera sideways of the gripper, while connected to the gripper, could potentially enable a much easier estimation of depth, which would make gripping easier and more effective.

Quite importantly, we also plan to use the evaluation framework described in this paper towards a thorough cross-comparison of joystick-based vs. whole-arm motion-capture based control of humanoid arms. It might well be the case that there exist aspects in which joystick-based control is preferable; the important point is that our evaluation framework will enable us to discern in which aspects of performance (for example, task completion time, positioning accuracy, types of unsuccessful transitory states, learning curve

and speed) joystick-based control can be superior to motion capture. We also plan to augment this study with a subjective questionnaire addressing the important issue of perceived naturalness; do users really feel that motion-capture based control is in any sense more "natural" than joysticks? And do they feel so when they start using it, or after a number of sessions, and if so, how many sessions does it take them to feel "naturalness"? And for which kinds of tasks are joysticks preferable, as compared to motion capture? These are very interesting and important questions that we plan to provide answers for, through a thorough empirical multi-user study, and fortunately the evaluation framework presented here can be really helpful towards providing strong quantitative answers to these questions.

Another important issue, apart from "naturalness", is trying to find ways to provide clear distinctions between perceived "naturalness" and perceived "intuitiveness" as well as "attractiveness". For example, subjective intuitiveness could well correlate with the learning time required to reach peak performance; and subjective attractiveness could well correlate with a challenging predicted difficulty level—which does not usually go along with "naturalness". Thus, the generally positive outlook that users had towards our system, as illustrated from the questionnaire results, could also be partially explained by the "challenging" nature of our chosen task, which had intermediate difficulty (neither trivial nor impossible; but fun to try!).

Also, many interesting possibilities for pattern recognition and prediction problems based on our task model exist: for example, one could try to predict the score of an individual on the basis of the first 10 seconds or the first ball of his trial; and one could even try to investigate if recognition of an individual through his task-signature is possible, for suitably modified tasks. One could even envision the investigation of possible correlations between individual traits and characteristic features exposed during task performance: for example, persistence as illustrated by multiple attempts towards a successful grip following an unsuccessful grip.

Yet another direction which we have started pursuing is migrating our teleoperation system to our conversational Arabic-speaking robot Ibn Sina [8, 9] (also see online videos available at youtube channel irmluaeu); in which case it will be used for motion training as well as embodied robotic telepresence, and will cover two hands as well as facial expression imitation. Ibn Sina is part of an interactive theatre, in which various modes of tele-participation are supported, including human-robot interaction through avatars in virtual worlds, remote brain-computer interfacing teleoperation [3] etc.

Finally, it is worth commenting upon the generality of the proposed $\{T, P(\Delta t|Si, Sj), P(r)\}$ model for evaluating many different types of teleoperation or other such human-machine tasks, as long as one can find a suitable set of discrete events describing the task, as well as a "resulting accuracy" measure at the end of each trial, and as long as time intervals between events can be measured, and the Markovian assumption implied by our model is warranted, given the nature of the described task. Thus, the methodology illustrated above can easily be transferred to any such task, and many such tasks exist: for example, let us consider an application scenario for teleoperation of Unmanned Aerial Vehicles, and for a task where the UAV must pass successfully (with minimal retries) through a number of narrow openings or perform docking rendevouz with other vehicles (events), before dropping certain carried packages and landing accurately at a target. Again, we have a model with events (start flight, success. pass/dock and drop/land with accuracies etc.), transition times between events, and accuracy metrics). Thus, we can easily create a $\{T, P(\Delta t|Si, Sj), P(r)\}$ model as well as the associated overall metrics, and our evaluation framework nicely applies, also in this case.

Similar examples can be found in other domains, too: for example in the medical domain, and in teleoperation of mobile robots doing construction/object rearrangement tasks etc. Also, the framework can even be used to assess improvements in robot learning by demonstration. Thus, the scope of the generality of the evaluation framework is wide, and its benefits important.

5 Conclusion

In this paper we presented: (a) a task-based Evaluation Framework for teleoperation, as well as (b) a designed and implemented System for Teleoperation of an industrial arm commanded through human-arm motion capture, which is used as a case study, and also serves to illustrate the effectiveness of the evaluation framework that we are introducing.

The motivation for our research was two-fold: first, the lack of suitable task-based evaluation frameworks for cross-comparing and improving the design of teleoperation systems, while enabling a rich description of operator-, system-, feedback, as well as learning and fatigue-effects, motivated our proposed framework. Second, although real-time remote teleoperation of robotic arms, either industrial or humanoid, is highly desirable for a number of applications, especially in difficult or inaccessible environments, it is the case that usually, teleoperation is driven by buttons, joysticks, haptic controllers, or slave-arms set—and this motivated our choice of motion capture as a sensing modality for our implemented teleoperation system, which also served as a testbed for our evaluation system: In our system the desired trajectory of the arm is naturally controlled through imitation of simple movements of the operator's physical arm, obtained through motion capture. Apart from a detailed description of our system and the design choices made, we presented an extensive

evaluation of the performance of our system, based on our framework, and containing: first, task-related measurables for a fixed task performed by numerous previously untrained subjects, in short-term as well as four-session longitudinal (learning) settings (i.e. a direct application of our evaluation framework to the system and task at hand); and second, user-opinion/attitude data obtained through a questionnaire administered to experimental subjects.

During the task-based evaluation, which was tuned in order to be neither trivial nor impossible, the probabilistic tripartite model that is at the heart of our evaluation framework was derived: a finite-state Markovian task model, augmented with transition time distributions as well as placement accuracy distributions. This tripartite model is in essence, a compact triad representing the performance of the coupled system-user pair. The applicability of this compact triad towards investigating the deconvolution of user, feedback, learning, and other components of performance was discussed, and future extensions presented. Furthermore, interesting results arose not only from the task-based but also from the questionnaire-based evaluation: for example, an ordering of desirability of a number of application areas for tele-operation arose as a conjecture—showing that most people strongly agree on the application of tele-operation for space or dangerous environments, but that there is potentially strong disagreement for a group of people regarding child instruction through tele-operated robots. Thus, we now know much more regarding opinions of people towards teleoperation, and its multiple possible domains of application.

Many future extensions of this work are directly visible, and we have started working on them. First, we would like to use the tripartite evaluation framework described in this paper towards investigating the effects of correspondence choice (in which way should the degrees of freedom of the human arm map to the robot arm) as well as of feedback choice (how does the placement of cameras, and the possibility of auditory and haptic feedback affect performance). Also, very importantly, we plan to carry our an extensive cross-comparison of joystick-based vs. motion capture-based control of the teleoperation system—and our evaluation framework provides the right tools for such a comparison, which could also be augmented with assessment of the subjective naturalness of the two types of controllers. Furthermore, we plan to apply our evaluation framework not only to arm-control pick-and-place tasks, but also to UAV and medical teleoperation domains, where it can readily be applied too.

In conclusion, through the presentation of a novel framework for modeling and task-based evaluation of teleoperation systems with a human-in-the-loop, and through a real-world system as well which also served to illustrate application of our task-based evaluation framework, and also through an extra questionnaire-based evaluation study, valuable results and insights were derived, ultimately towards the wider beneficial application of tele-operated robotics by untrained humans in an increasing range of real-world application areas.

Acknowledgements The authors would like to thank Nikos Batalas, Iman Al Shebli, Eida Al Ameri, Fatima Al Neyadi, Alya Al Neyadi, Amr Elfaham, Walid Soulakis, Aziz El Kayoumi, Saif Al Ketbi, and all of UAEU students and IRML friends for their invaluable help and support.

References

1. Alissandrakis A, Nehaniv CL, Dautenhahn K (2007) Correspondence mapping induced state and action metrics for robotic imitation. IEEE Trans Syst Man Cybern, Part B, Cybern 37(2):299–307. Special issue on robot learning by observation, demonstration and imitation
2. Alissandrakis A, Otero N, Saunders J, Dautenhahn K, Nehaniv CL (2009) Helping robots imitate—metrics and computational solutions inspired by human-robot interaction studies. In: Gray J, Nefti-Meziani S (eds) Advances in cognitive systems. IET Press, Stevenage
3. Christoforou C, Mavridis N, Machado EL, Spanoudis G (2010) Android tele-operation through Brain-Computer Interfacing: a real-world demo with non-expert users. In: Proceedings of international symposium on robotics and intelligent sensors IRIS 2010
4. Cui J et al (2006) A review of teleoperation system control. In: Proceedings of the 2006 Florida conference recent advances in robotics (FCRAR), Florida Atlantic University, FL
5. Kim D, Kim J, Lee K, Park C, Song J, Kang D (2009) Excavator tele-operation system using a human arm. Autom Constr 18:173–182
6. Matsui D, Minato T, MacDorman KF, Ishiguro H (2007) Generating natural motion in an android by mapping human motion, humanoid robots, human-like machines
7. Questionnaire available at the Interactive Robots and Media Lab website. http://irml.uaeu.ac.ae (last retrieved November 2010)
8. Videos of the Interactive Robots and Media Lab projects online at: http://www.youtube.com_channel_irmluaeu (last retrieved November 2010)
9. Mavridis N, Hanson D (2009) The IbnSina center: an augmented reality theater with intelligent robotic and virtual characters. In: Proceedings of the 18th IEEE symposium on human and robot interactive communication (Ro-MAN)

Nikolaos Mavridis received his Ph.D. from the Massachusetts Institute of Technology Media Labporatory in 2007, and is currently serving as Asst. Professor of Intelligent Systems at the UAE University, where he has founded and is directing the Interactive Robots and Media Laboratory. His research interests include Human-Robot Interaction and Social Robots, Teleoperation and Telepresence, and Cognitive Systems.

Nikolas Giakoumidis is graduating from the Technical University of Piraeus in June 2011, and has taken his internship at the Interactive Robots and Media Laboratory of the UAE University under the supervision of Dr. Mavridis. His research interests include tele-operation as well as unmanned vehicles.

Emerson Lopes Machado is a Robotics Engineer, and has received the Master of Informatics from the University of Brasilia in 2007 and the Bachelor in Computer Science in 2003. His research interests include machine learning, as well as neuroscience and brain-computer interfacing applied to robotic control.