2013 IEEE/RSJ International Conference on
Intelligent Robots and Systems (IROS)
November 3-7, 2013. Tokyo, Japan

Comparative Usability and Performance Evaluation of Surgeon Interfaces in Laser Phonomicrosurgery

Giacinto Barresi[1], Nikhil Deshpande[1], Leonardo S. Mattos[1], Andrea Brogni[1],
Luca Guastini[2], Giorgio Peretti[2], and Darwin G. Caldwell[1]

[1]Department of Advanced Robotics, Istituto Italiano di Tecnologia, Genova, Italy

[2]Department of Otorhinolaryngology, Università degli Studi di Genova, Italy

Email: {giacinto.barresi, nikhil.deshpande, leonardo.demattos, andrea.brogni, darwin.caldwell}@iit.it,
luca@guastini.eu, g.peretti@tin.it

***Abstract*—Robot-assisted surgical procedures, such as Laser Phonomicrosurgery (LP), suffer from susceptibility to variation in surgeon skill and equipment characteristics. Ergonomic and human-centered approaches acquire increased importance in the design of surgeon-machine interfaces. This paper proposes a protocol for comparative evaluation of surgeon-machine interfaces based on two criteria: (i) the subjective evaluation of their usability using questionnaires, and (ii) the objective evaluation of their performance using an imaging-based feature extraction method. Two interfaces in LP, the traditional ("AcuBlade") interface and the novel ("Virtual Scalpel") interface, were evaluated to demonstrate the effectiveness of the proposed scheme. A series of experimental trials were conducted using the interfaces in surgery-like tasks in a controlled environment. The subjective evaluation pointed to the superiority of the Virtual Scalpel interface (score: 83.06) in terms of confidence and ease of use, and learnability, over the AcuBlade interface (score: 65.56). The objective evaluation showed the Virtual Scalpel interface having an overall score (55.96) significantly superior to the AcuBlade (51.37). It is thus shown that the multidimensional evaluation approach allowed to clearly distinguish between levels of perceived usability and effective performance of surgeon-machine interfaces from a user-centered perspective.**

***Index Terms*—Surgeon performance, ergonomic evaluation, unified rating, usability, SUS.**

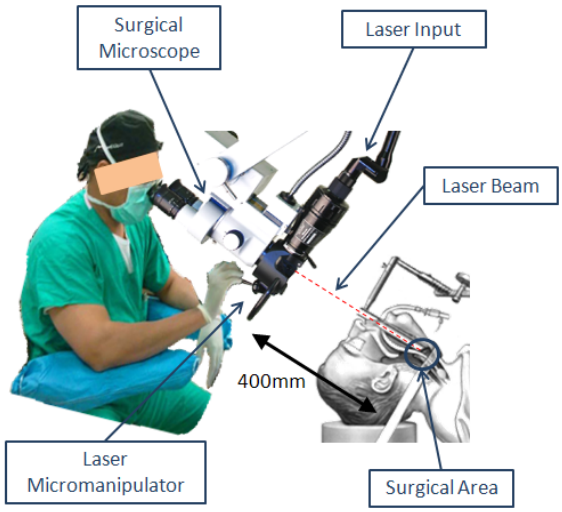
I. Introduction

As robot-assisted surgical systems become prevalent in the operating room [1], their evaluation in terms of surgical outcomes and ergonomic features becomes increasingly important. Surgical interventions require a high degree of effectiveness (capability to perform a task), efficiency (performing the task with least resources) and safety, for both the patient and the surgeon. Robot-assisted surgical systems have the potential to provide significant advantages in all these three aspects of surgery through increased task precision, reduced tremor in gestures, timely execution of repetitive tasks, among other features. Clearly, the quality and efficiency of the surgical outcome depends also on the characteristics of the robotic surgical equipment [2]. There are currently major concerns with equipment usability and performance in the case of laser phonomicrosurgery (LP), where a combination of poor ergonomics, sub-optimal visualization, difficult surgical-site access, and surgeon discomfort can affect the surgical outcome [3]. Therefore, the ergonomic and human-centered approaches acquire increased importance in the design of surgeon-machine interfaces in LP.

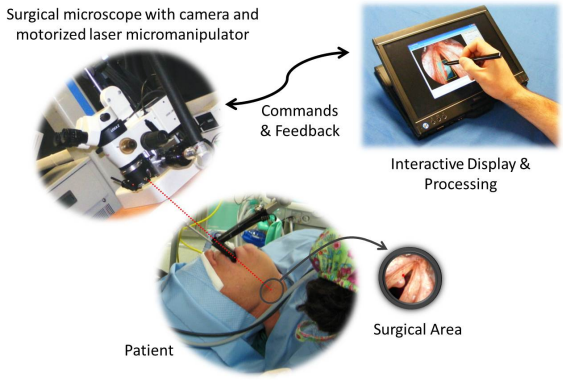
LP is a state-of-the-art procedure within the domain of non-invasive, trans-oral surgeries, for the treatment of abnormalities in the vocal cords, such as, tumors and cysts. The traditional system currently used in the operating room employs a CO_2 surgical laser, coupled with a surgical microscope [4]. The laser beam, used to either ablate or remove the abnormality, is aimed at the surgical area (the vocal folds) from a distance of 400 mm, by maneuvering a mechanical micromanipulator. The laser is activated using a separate foot-switch. The setup, called the Digital AcuBlade system, can be seen in Figure 1(a). Evidently, the surgeon is directly impacted by the poor ergonomics - specific posture during surgery, the need for high psychomotor skills for surgical gestures and maneuvers, hand-foot coordination for laser activation [5]. LP is the focus of new technologies and research in the context of the European project - μRALP, at the Istituto Italiano di Tecnologia (IIT). Mattos et al. [3] presented a novel surgeon interface design, called the "Virtual Scalpel" system. This system replaces the manual micromanipulator interface with a motorized one, which is controlled through a graphics stylus and a touch-screen tablet with live video of the surgical area. Figure 1(b) depicts the setup. The immediate advantage is that both aiming and activation of the laser are controlled by the stylus. This novel solution can provide greater precision along with better ergonomics over the traditional laser micromanipulator [5], [3], [6].

However, to provide the required significance to any technological improvements, a comprehensive user-centered evaluation with respect to the surgical outcomes, is required. This paper introduces a protocol for such evaluation of surgeon interfaces - a combination of the subjective (self-reported) and objective (quantitative performance) assessment for a comparative evaluation. The remainder of the paper: (i) introduces the evaluation methodology and the experimental procedure (Section III); (ii) presents results and discusses the analysis of the experimental trials (Section IV); and (iii) concludes the paper with a reference to future work (Section V).

978-1-4673-6357-0/13/$31.00 ©2013 IEEE



(a) Traditional LP System



(b) Virtual Scalpel System

Figure 1: Laser Phonomicrosurgery equipment and surgical setup.

II. Related Work

Any approach towards introducing robotic technology in the surgical environment has to necessarily involve a task-specific, human-centered development model. O'Toole et al. [7] postulated that the acceptability of new technologies in surgery is dependent on a three-tier hierarchy: (i) clinical need, (ii) effectiveness and safety, and (iii) compatibility, cost and usability. Their assessment of existing surgical equipments demonstrated a lack of conformity to the top-tier criteria, i.e., involving usability, and this hampered their wide-spread acceptance. Funda et al. [8] demonstrated the ergonomic viability of their 7-axis surgical robot arm using a series of trajectory following tasks. This also allowed to demonstrate the dependence of ease-of-use of the robot on the available degrees of freedom. Das et al. [9] used the analysis of variance (ANOVA) method to evaluate the usability of a telemanipulator for robot-assisted microsurgery. Their subjects included surgery students as well as engineers to understand the impact of background on the system's usability. Fujii et al. [10] presented a 3-DOF forceps control robot in the context of laparoscopic surgery with enhanced user interface design. The authors adapted the NASA-TLX [11] method for scoring the workload required by the task based on the subjective evaluation of the users. Martelli et al. [1] developed their own set of questionnaires aimed at analyzing the subjective feeling of the surgeons in using the system, and objective aspects of the system in total-knee replacement surgery.

In LP, this research is still in its nascent stages. The techniques introduced in this paper can be adapted to the larger domain of laser surgeries, leading to a significantly enhanced capacity for evaluating robot-assisted surgical interfaces.

III. Evaluation Methodology

A. Subjective Evaluation

Ergonomics can be defined as the study of the interactions between technology and humans, the environment in which the technology is being used, and the problems and benefits it presents [12]. The human factor–oriented approach of ergonomics permits the analysis of the limits of user interfaces in terms of aspects such as usability and mental workload. Mental workload represents the amount of cognitive resources used in order to accomplish a task [13], while usability is defined as "the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use." (classic definition in [14]). For the evaluations, the System Usability Scale (SUS) questionnaire was utilized [15] (Table I). SUS is a questionnaire composed of ten items that allows the subjective assessment of usability and provides a global view on its main aspects.

Table I: SUS Questionnaire Items

1. I think that I would like to use this system frequently.
2. I found the system unnecessarily complex.
3. I felt very confident using the system.
4. I found the various functions in this system were well integrated.
5. I thought the system was easy to use.
6. I thought there was too much inconsistency in this system.
7. I found the system very cumbersome to use.
8. I would imagine that most people would learn to use this system very quickly.
9. I think that I would need the support of a technical person to be able to use this system.
10. I needed to learn a lot of things before I could get going with this system.

Upon receiving the SUS questionnaire, a subject must read each statement and evaluate whether and how much s/he agrees with it. This evaluation is expressed by marking a point on an oriented 5- or 7-points scale. Each point of the scale represents a different level of agreement, from "strongly disagree" for point 1 to "strongly agree" for point 5 (or point 7).

B. Objective Evaluation

The performance assessment of the trials was done based on the metrics introduced in [16]. The metrics are described in brief in Table II. The significance of these imaging-based metrics and the robustness of the associated *unified rating* based classification was established in [16].

Table II: Imaging-based Metrics

Metric	Description
1. Area Ratio (AR)	Ratio of pixel-count in laser-traced and desired shapes.
2. Perimeter Ratio (PR)	Ratio of pixel-count in boundaries of laser-traced and desired shapes.
3. Aspect Ratio Measure (ARM)	Ratio of aspect ratios of laser-traced and desired shapes.
4. Orientation Measure (OM)	Absolute difference in orientations of laser-traced and desired shapes.
5. Shape Measure (SM)	The similarity (or lack thereof) of the laser-traced and desired shapes.
6. Path Following Error (RMSE)	Root-mean-square-error method for distance between the laser-traced and desired shapes.
Rating (f)	$f_{AR} + f_{PR} + f_{ARM} + f_{OM} + f_{SM} + f_{RMSE}$
7. Trial Time	Time taken in laser-tracing the given shape.

The *unified rating* is then obtained as the weighted sum of these individual metrics, taking into account the natural variation in any human-operated equipment [16]. The *Trial Time* metric was used for the comparative assessment of actual and perceived times as noted in section IV-C.

C. Experimental Procedure

The experiments for the user trials (introduced in [16]) included sets of trajectory following exercises, where the subjects performed surgical maneuvers to follow preset random shapes. The shapes, including straight lines, C-curves, and S-curves, are representative of real surgical actions. They were stamped on small plaster blocks, as illustrated in Figure 2, with each target block having 12 shapes, featuring one of five different randomized sequences of shapes and shape orientations. This method of trials offers an unambiguous task definition and also facilitates easy task randomization and evaluation, since the blocks are clearly marked by the CO_2 laser. These artificial precision target blocks were placed on a holding structure 400 mm from the surgical microscope, typical of LP.

The experiments were designed to evaluate two conditions of laser control in LP: (i) the AcuBlade condition, and (ii) the Virtual Scalpel condition. The trials were performed at the San Martino Hospital (Genova, Italy). The complete control of the environmental and social context inside the experimental room helped avoid any disturbance. The experimental data was collected in three different ways here: (i) through the laser traces on the target blocks; (ii) through a CCD camera installed co-axially with the microscope, capturing images of the scene observed by the subjects; and (iii) through an external video camera that records potentially relevant behaviors of the subjects (Figure 2). Figure 3 shows a random representative trial from the use of the two surgeon interfaces.

D. Subjects and Groups

A series of field trials were conducted involving a sample of potential end users of the surgical devices. Each subject received the instructions for the trials with the LP system, and an 'informed consent form' describing the rights, responsibilities and risks involved in participating in the trials. A questionnaire gathered personal data from the subjects before the experimental trials. The subjects performed the trials in two sessions of two plaster blocks each, with, at least, a break of 10 minutes between the two. After the experimental session, the subjects filled out the SUS questionnaire. Additionally, a question regarding the subjective evaluation of the perceived time spent performing the trial was included.

The subjects were an elective population to study the effects on two aspects: (i) surgery-like practice perception, since part of the undergraduate sub-group and the whole graduate sub-group already assist in medical interventions; (ii) and surgery-related learning, since both sub-groups are in different stages of their medical education. The subjects were divided into two groups, one per condition, AcuBlade group vs. Virtual Scalpel group, with 12 subjects in each. The groups were composed in order to balance the study level (undergraduate and graduate) and the gender of the subjects.

IV. Data Analysis

The experiments involved the manipulation of one independent variable, the laser control interface, with 2 conditions: AcuBlade vs. Virtual Scalpel. Each condition was associated with one group of the subjects, and each subject performed the trials in only one condition, at present. The dependent variables are:

1) The scores of the SUS questionnaire: Information about the subject's perception of the tool's usability and its

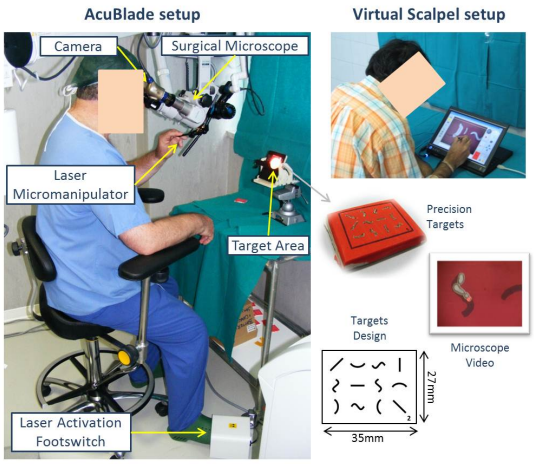


Figure 2: Experimental setup, video snapshot from a trial, and details of the precision targets stamped on a plaster block.

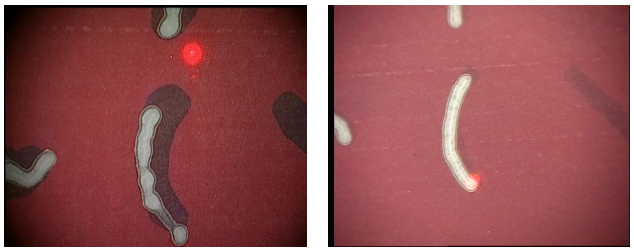


Figure 3: Images showing the laser-traced shapes overlapped on the desired shapes (Representative Trials).

Left: AcuBlade Condition Right: Virtual Scalpel Condition

dimensions.

2) The objective values of performance processed through the imaging-based metrics algorithm developed in [16].
3) The subjective self-evaluation of the total time spent to perform the trials. This information is intrinsically related to the mental effort of the users. Errors in time estimation increase as a function of the amount of attentional resources needed for concurrent tasks [17].

A. *Subjective Evaluation of Usability*

Differences evidenced by the analysis demonstrate the different level of usability of each interface (AcuBlade vs. Virtual Scalpel). The detailed list of the comparisons of SUS scores is represented in Table III.

Table III: Comparison of SUS scores*

	AcuBlade condition mean (%)	Virtual Scalpel condition mean (%)	% variation for Virt. Scal. over AcuBlade
Global Score	**65.56**	**83.06**	**26.69**
Sub-scale 1	75.00	77.78	3.71
Sub-scale 2	77.78	90.28	16.07
Sub-scale 3	37.50	72.22	92.59
Sub-scale 4	70.83	76.93	7.84
Sub-scale 5	61.11	84.72	38.64
Sub-scale 6	87.50	80.56	-7.94
Sub-scale 7	83.33	91.67	10.00
Sub-scale 8	56.94	93.06	63.41
Sub-scale 9	47.22	77.78	64.71
Sub-scale 10	58.33	86.11	47.62

* The 10 sub-scales correspond to the 10 questions in Table I. The table shows normalized scores (0 - minimum; 100 - maximum).

The difference between the SUS global scores for the two conditions was statistically significant according to the Student's t-test (t = 2.83, p = 0.009). In particular, the global score of subjective usability of Virtual Scalpel (m = 83.06, sd = 12.49) is higher than the score for AcuBlade (m = 65.56, sd = 17.37). The sub-scales are considered for an explorative data analysis, because they do not satisfy the normality assumption for the Student's t-test [18]. According to this qualitative exploration, 5 specific sub-scales of the SUS contribute most significantly to the improved scores for the Virtual Scalpel over the AcuBlade, as represented in Figure 4.

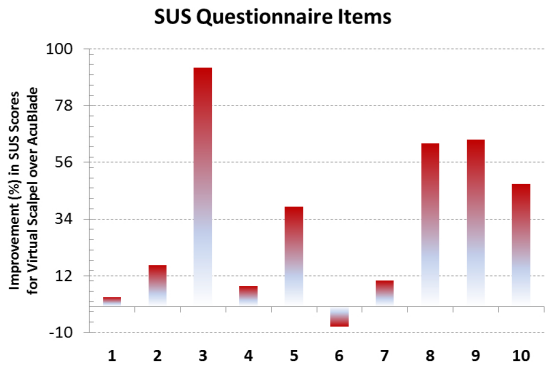


Figure 4: Differences between the SUS sub-scales scores for the two interfaces.

The change in the SUS sub-scale scores for the Virtual Scalpel condition over the AcuBlade condition.

The following observations are made based on this explorative analysis:

1) Sub-scale 3 (93% better): The users of Virtual Scalpel feel more confident than the users of AcuBlade during the tasks.
2) Sub-scale 5 (39% better): The Virtual Scalpel interface is easier to use than the AcuBlade interface.
3) Sub-scale 8 (63% better): The Virtual Scalpel interface is easier to learn than the AcuBlade interface.
4) Sub-scale 9 (65% better): The users of Virtual Scalpel would require less support by an expert than the users of AcuBlade.
5) Sub-scale 10 (48% better): The users of Virtual Scalpel would require to learn less processes than the users of AcuBlade.

B. *Objective Evaluation of Performance*

Considering the number of subjects (12) and the number of shapes traced by each (48), there are 576 data points for each of the metrics in each condition (AcuBlade and Virtual Scalpel). The ideal value for the unified rating is 60.

Table IV summarizes the t-test values for the metric ratings and the unified rating, averaged over the 12 subjects for each condition.

Table IV: Comparison of average values of the Metrics and Unified Rating scores

	AcuBlade condition		Virtual Scalpel condition		t (for f)	p (for f)
	Raw	f	Raw	f		
Unified Rating	**–**	**51.37**	**–**	**55.96**	**4.259**	**4e-4**
Area Ratio	0.88	9.38	0.80	9.34	0.194	0.848
Perimeter Ratio	0.87	9.44	0.98	9.93	5.191	3e-4
Aspect Ratio	1.14	8.10	0.83	8.85	2.892	0.009
Orientation	2.39	8.42	1.11	9.43	3.831	0.001
Shape	2.60	8.71	2.55	8.83	0.375	0.711
Path Foll. Error	0.51	7.33	0.25	9.58	7.212	1e-5

The two conditions show a statistically significant difference in performance based on the unified rating score, with t = 4.26 and p = 4e-4 with the Student's t-test. The Virtual Scalpel condition (m = 55.96, sd = 2.03) allows a significantly better performance than the AcuBlade condition (m = 51.37, sd = 3.14). Among the individual metrics, the following show a significant difference in performance between the two conditions[1]:

- Perimeter and Aspect Ratio: The laser-traced shapes under the Virtual Scalpel condition conform to the length and thickness of the desired shapes, significantly better than in the AcuBlade condition.
- Orientation Measure: The laser-traced shapes under the Virtual Scalpel condition are better aligned with the desired shapes than in the AcuBlade condition.
- Path Following Error: The Virtual Scalpel condition permits easier trajectory following with the laser than the AcuBlade condition.

The Perimeter Ratio and the Aspect Ratio metrics signify the usage of the laser during tracing of the desired shape. A ratio value greater than '1' means that the laser-trace is bigger than the desired trace, indicating excess usage of the laser. On contrary, a low value of the ratios indicates that not enough of the desired trace has been covered. A deeper analysis of this excess area showed that the AcuBlade interface suffered more significantly from excess usage of the laser than the Virtual Scalpel interface. Less trials in the Virtual Scalpel condition had excess laser usage than the AcuBlade condition, pointing to its greater overall efficiency.

The Orientation Measure and the Path Following Error are directly related with the operational safety of the interfaces. A low rating for the Orientation Measure indicates that the majority of the laser-traced shape is not aligned with the desired shape. Similarly, a low value for the Path Following Error gives an account of how far the laser-traced shape is from the desired shape. Both these states mean that the laser is active in non-desired areas. In case of real surgeries, this would mean that the laser is ablating more than the desired area of the tissue, possibly healthy tissue.

C. Subjective and Objective Assessment of Time

A t-test was also performed to understand the difference between the perceived and actual times for trial completion effect in the two conditions. There is a significant difference between the subjective evaluations of the trial times in the two conditions (t = 3.07, p = 0.005). The assumptions of normality and homoscedasticity are checked for the t-test. It was noted that the subjects in the AcuBlade condition estimated a longer time taken to complete the trials (m = 20.31, sd = 6.57) than the subjects in the Virtual Scalpel condition (m = 12.09, sd = 6.58). The quantitative analysis is contrary to the subjective evaluation. It shows a significant difference (t = 2.69, p = 0.014), but the AcuBlade interface (m = 4.39, sd = 1.68) takes a much shorter time to complete than the Virtual Scalpel condition (m = 6.39, sd = 1.93). Please refer to Figure 5. This is an interesting result, since it allows to infer the level of mental workload required in each condition. The actual times spent in tasks for Virtual Scalpel is greater while the estimated time taken for the tasks with it is lesser. The subjects in the Virtual Scalpel condition perceive to have spent less time performing the trials than the subjects in the AcuBlade condition. Therefore, the Virtual Scalpel condition seems to induce less mental workload than the AcuBlade condition. This is advantageous for the acceptance and usability of the Virtual Scalpel condition since it is perceived as a low mental workload interface [17].

[1]The Student's t-test was used for the comparison of the unified rating scores. The Welch's t-test was used for the individual metrics since the homoscedasticity assumption for the variances was not satisfied [19].

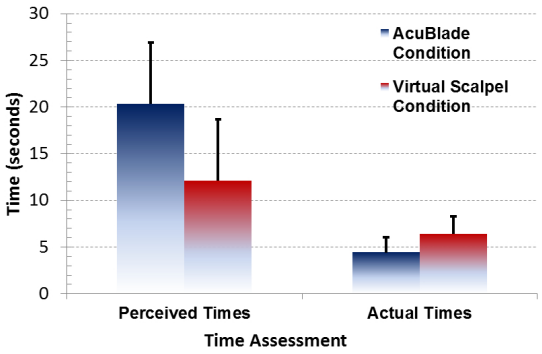


Figure 5: Time Assessment Comparison.
The means and the standard deviations are shown.

V. Conclusions and Future Work

In this paper, an integrative protocol was proposed for a comprehensive user-centered evaluation of surgeon-machine interfaces. The combination of the subjective SUS-based evaluation and the quantitative metrics-based evaluation provides for a clear classification of surgeon-machine interfaces from the usability and performance perspective. The results (Figure 6) from the trials conducted with the two examined interfaces are summarized as follows:

1) The Virtual Scalpel interface shows a higher score of usability than the AcuBlade interface, and this is demonstrated through the *global usability* score of the SUS questionnaire.
2) The objective evaluation points to the clear advantage of the Virtual Scalpel interface over the AcuBlade interface, according to the *unified rating* score.
3) Although the actual time for task performance is greater with the Virtual Scalpel interface, subjects perceive that it takes significantly less time than the AcuBlade interface. This implies that the Virtual Scalpel interface requires a much lower mental workload on the part of the user.
4) Taken together, the subjective and objective evaluations classify the Virtual Scalpel as having superior usability and performance capability as a surgeon interface.

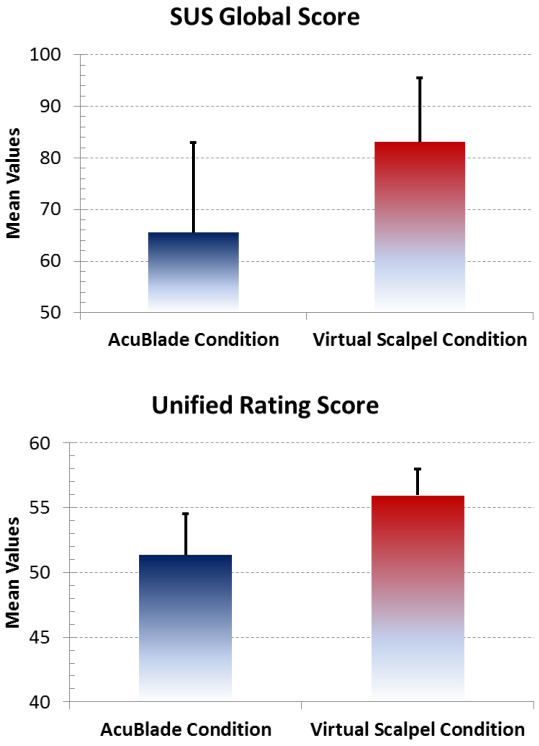


Figure 6: The *Unified Rating* and the *SUS Global* scores for the two conditions.

The means and the standard deviations are shown. '60' and '100' are the respective maximum values.

This comprehensive assessment approach can provide evidence for a clear and unbiased comparison among different interfaces for delicate surgical procedures like LP. In the extension of this research, further studies are planned to better understand the usability preferences by considering the factor of experience and background with larger groups of subjects. The observations derived from the explorative data analysis of the SUS sub-scales shall be investigated further with this larger group. The implementation of this ergonomic methodology shall be explored in the development of a novel tool dedicated for the integration of subjective and objective analyses in the evaluation of user interfaces in broader surgical applications.

ACKNOWLEDGMENT

The research leading to these results has received funding from the European Union Seventh Framework Programme FP7/2007-2013 – Challenge 2 – Cognitive Systems, Interaction, Robotics – under grant agreement μRALP - n°288233.

REFERENCES

[1] S. Martelli, L. Nofrini, P. Vendruscolo, and A. Visani, "Criteria of Interface Evaluation for Computer assisted Surgery Systems," *International Journal of Medical Informatics*, vol. 72, no. 1-3, pp. 35–45, 2003.

[2] S. Serefoglou, W. Lauer, A. Perneczky, T. Lutze, and K. Radermacher, "Multimodal User Interface for a Semi-Robotic Visual Assistance System for Image Guided Neurosurgery," in *Proc. Computer Aided Radiology and Surgery, (CARS 2005)*, vol. 1281, 2005, pp. 624–629.

[3] L. S. Mattos, G. Dagnino, G. Becattini, M. Dellepiane, and D. G. Caldwell, "A Virtual Scalpel System for Computer-assisted Laser Microsurgery," in *Proc. IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems, (IROS 2011)*, Sep. 2011, pp. 1359–1365.

[4] Digital AcuBlade System. Lumenis Inc. Israel. [Online] Available: http://www.surgical.lumenis.com. Accessed on 6-Sept-2012.

[5] G. Dagnino, L. S. Mattos, and D. G. Caldwell, "New Software Tools for Enhanced Precision in Robot-Assisted Laser Phonomicrosurgery," in *Proc. 34^{th} Intl. Conf. of IEEE Engineering in Medicine and Biology Society, (EMBC 2012)*, 2012, pp. 2804–2807.

[6] L. S. Mattos, M. Dellepiane, and D. G. Caldwell, "Next-generation Micromanipulator for Computer-assisted Laser Phonomicrosurgery," in *Proc. 33^{rd} Intl. Conf. of IEEE Engineering in Medicine and Biology Society, (EMBC 2011)*, Sep. 2011, pp. 4555–4559.

[7] M. D. O'Toole, K. Bouazza-Marouf, D. Kerr, M. Gooroochurn, and M. Vloeberghs, "A Methodology for Design and Appraisal of Surgical Robotic Systems," *Robotica*, vol. 28, pp. 297–310, 2010.

[8] J. Funda, K. Gruben, B. Eldridget, S. Gonioryt, and R. Taylor, "Control and Evaluation of a 7-axis Surgical Robot for Laparoscopy," in *Proc. Intl. Conf. on Robotics and Automation, (ICRA 1995)*, 1995.

[9] H. Das, H. Zak, J. Johnson, J. Crouch, and D. Frambach, "Evaluation of a Telerobotic System to Assist Surgeons in Microsurgery," *Computer Aided Surgery*, vol. 4, no. 1, pp. 15–25, 1999.

[10] M. Fujii, K. Fukushima, N. Sugita, T. Ishimaru, T. Iwanaka, and M. Mitsuishi, "Design of Intuitive User Interface for Multi-DOF Forceps for Laparoscopic Surgery," in *Proc. Intl. Conf. on Robotics and Automation, (ICRA 2011)*, May 2011.

[11] S. Hart and L. Staveland, "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research," in *Human Mental Workload*, ser. Advances in Psychology, P. A. Hancock and N. Meshkati, Eds. ©Elsevier, 1988.

[12] M. Helander, *A Guide to Human Factors and Ergonomics*, 2^{nd} ed. ©CRC Press, 2005.

[13] T. E. Nygren, "Psychometric Properties of Subjective Workload Measurement Techniques: Implications for their Use in the Assessment of Perceived Mental Workload," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 33, no. 1, pp. 17–33, 1991.

[14] *Ergonomic requirements for office work with visual display terminals (VDTs), Part 11: Guidance on Usability.*, ISO Std. 9241-11, 1998.

[15] J. Brooke, "SUS: A "Quick and Dirty" Usability Scale," *Usability Evaluation in Industry*, 1995.

[16] N. Deshpande, L. S. Mattos, G. Barresi, A. Brogni, G. Dagnino, L. Guastini, G. Peretti, and D. G. Caldwell, "Imaging based Metrics for Performance Assessment in Laser Phonomicrosurgery," in *Proc. Intl. Conf. on Robotics and Automation, (ICRA 2013)*, 2013.

[17] M. Lind and H. Sundvall, "Time estimation as a measure of mental workload," in *Engineering Psychology and Cognitive Ergonomics*. Springer, 2007, pp. 359–365.

[18] A. C. Boneau, "The effects of violations of assumptions underlying the t-test," *Psychological Bulletin*, vol. 57, no. 1, pp. 49–64, 1960.

[19] B. L. Welch, "The generalization of "Student's" problem when several different population variances are involved," *Biometrika*, vol. 34, no. 1-2, pp. 28–35, 1947.