www.elsevier.com/locate/ijhcs

Evaluation of a Surgical Interface for Robotic Cryoablation Task using an Eye-Tracking System

Alper Açık, Duygun Erol Barkana, Gökhan Akgün, Asım Evren Yantaç, Çağla Aydın

PII: S1071-5819(16)30080-5
DOI: http://dx.doi.org/10.1016/j.ijhcs.2016.07.004
Reference: YIJHC2049

To appear in: *Journal of Human Computer Studies*

Received date: 7 August 2015
Revised date: 2 July 2016
Accepted date: 5 July 2016

Cite this article as: Alper Açık, Duygun Erol Barkana, Gökhan Akgün, Asım Evren Yantaç and Çağla Aydın, Evaluation of a Surgical Interface for Robotic Cryoablation Task using an Eye-Tracking System, *Journal of Human Compute Studies,* http://dx.doi.org/10.1016/j.ijhcs.2016.07.004

Evaluation of a Surgical Interface for Robotic Cryoablation Task using an Eye-Tracking System

Alper Açık*, Duygun Erol Barkana, Gökhan Akgün, Asım Evren Yantaç, Çağla Aydın

Özyeğin University, Psychology Dept., Çekmeköy Campus, Nişantepe District, Orman Street, 34794, Çekmeköy – İSTANBUL

*Corresponding author. Tel.: +902165649388. alper.acik@ozyegin.edu.tr

Abstract

Computer-assisted navigation systems coupled with surgical interfaces (SIs) are providing doctors with tools that are safer for patients compared to traditional methods. Usability analysis of the SIs that guides their development is hence important. In this study, we record the eye movements of doctors and other people with no medical expertise during interaction with an SI that directs a simulated cryoablation task. There are two different arrangements for the layout of the same SI, and the goal is to evaluate whether one of these arrangements is ergonomically better than the other. We use several gaze related statistics some of which are employed in an SI design context for the first time. Even though the performance and gaze related analysis reveals that the two arrangements are comparable in many respects, there are also differences. Specifically, one arrangement leads to more saccades along the vertical and horizontal directions, lower saccade amplitudes in the crucial phase of the task, more locally clustered and yet globally spread viewing. Accordingly, that arrangement is selected for future use. The present study provides a proof of concept for the integration of novel gaze analysis tools developed for scene perception studies into the interface development process.

Keywords: User Interface Design, Surgical Interfaces, Eye Movements

1. Introduction

Computer-assisted navigation is a surgical system used to perform surgical tasks. These systems help surgeons to navigate through the 3D representations of the patient's body and surgical devices (Wegner, 1998; Mezger et al., 2013; Münzer et al., 2006). Surgeons access volumetric, functional, and navigation-assisting data through the Surgical Interface (SI), thanks to the visualization tools integrated on these interfaces (Wegner, 1998; Mezger et al., 2013; Münzer et al., 2006). SIs are already in use in clinical practices to avoid risks, and handle real-time situations (Wegner, 1998). Such SIs are able to improve the safety of surgical interventions by introducing less invasive procedures (Mezger et al., 2013). Various SIs have been designed to allow surgeons to plan and simulate surgical interventions by deploying various sources of information such as Computed Tomography, Magnetic Resonance Imaging, 3D Ultrasound images, and details of the operation area (Peterhans et al., 2011; Fasquel et al., 2008). However, the complexity of the SIs, problems related with the repeatability of the surgeons' actions, and suspicions about the accuracy of the presented information prevent surgeons from using these SIs to the maximum advantage (Martelli et al.,

2007). In order to circumvent these problems and promote SI use, recently human factors and usability are taken into consideration in designing SIs to increase the accuracy of the surgical outcome and decrease invasiveness (Yang et al., 2012).

Human factors research explores how much and what sorts of information a person can use effectively while interacting with a system, and how the information about the system should be organized and presented in order to make the interaction optimal (Klatzky et al., 1996). The main objectives of human factors research are the maximization of efficiency, improvement in system performance, and increase in ease of use and safety (Salvendy, 2012). In the recent past, human factors studies started to provide design solutions for the disciplines of medicine and psychology with a special emphasis on human-machine interactions (Salvendy, 2012). Researchers make use of ergonomics principles (Martin et al., 2008; Weinger et al., 2010; Shrivastava et al., 2014; Vincent et al., 2014) and employ methods such as semi-structured interviews with people during the design process (Privitera et al., 2009), analysis of user group differences in terms of "personas" (Vincent and Blandford, 2014), and usability analysis based on visual design principles (Sawyer et al., 1996). The layout and the arrangement of a surgical interface are critical because information required for the surgical operation must be extracted without attentional effort. Correct settings for visual symbol size, contrast, color, display depth, and shape coding facilitate the rapid identification of information from the user interfaces (Sawyer et al., 1996). Internal consistency and clear hierarchy of the elements in the interface are also very important in order to reduce uncertainty and misunderstanding for the expert user (Altaboli and Lin, 2011). Another critical issue is the correct alignment of visual elements to reduce the visual load of the user, and help the user with understanding the information structure, which can be achieved with balancing the relative scale of the elements according to their functions (Schlatter and Levinson, 2013). Thus, the appearance of the medical interfaces can be improved by incorporating visual design principles and asking users about their subjective interaction experiences. Accordingly, investigation of the usability and functionality factors in SI development will produce improvements in the designs of these interfaces and the quality of the surgical intervention (Calisir et al., 2014).

During the last two decades there has been an increase in studies employing eye movement analysis in usability research so that visual aspects of interface designs can be addressed with objective data (Bergstrom and Schall, 2014; Goldberg and Wichansky, 2003; Pannasch et al., 2008; Poole and Ball, 2005). Since eye movements reflect how observers serially shift their attention from one part of the visual stimulus to the next, usability researchers profit from their analysis while addressing the efficiency and aesthetics of the systems used by different populations (Halverson and Hornof, 2011). During active viewing, saccades, the ballistic movements of the eyes, are overt shifts of our attention from place to place and fixations between saccades correspond to intervals of visual information acquisition (Kowler, 2011). Nevertheless, eye-movement based usability for medical technologies is still considered in its infancy. Recently, (Asan and Yang, 2015) performed an extensive search for articles published after 2004 that addressed usability evaluation based on eye-movement data for health information systems. Whereas only nine such studies were identified, among those only the report (Erol Barkana et al., 2014) that inspired the current study investigated eye-movements in the context of a surgical operation. The remaining studies investigated computerized tools that provide general medical information to patients or doctors for treatment or health awareness purposes (Asan and Yang, 2015). Thus, despite the rise in usability studies featuring eye-movement analysis for several technologies, the field of surgical interface development is a notable exception.

Which measures derived from eye-tracking data will be informative within the usability context? In the literature it is most common to investigate fixation locations and especially the amount of fixations within areas of interest that characterize task-related regions of the system that is in use (for a review see Poole and Ball, 2005)). Moreover, many studies analyze fixation durations to address ease of local information acquisition (reviewed by Jacob and Karn, 2003). However, compared to the tools of vision science that investigates attention allocation mechanisms during viewing of natural scenes (e.g. Foulsham et al., 2008; Smith and Henderson, 2009; Wilming et al., 2013), these measures capture a very limited set of natural viewing characteristics. For instance, distributions of saccade directions display an abundance of eye-movements along the horizontal and vertical axes, compared to other oblique directions (Foulsham et al., 2008). The vertical and horizontal contour content of the scene is able to explain this abundance only partially (Foulsham and Kingstone, 2010), which suggests that making horizontal and vertical saccades is more natural in a default viewing mode. Amplitudes of saccades can be analyzed (Dorr et al., 2010) to reveal whether important locations on an interface that are fixated successively are close to each other. An extension of this investigation, which captures the locations of successive fixations together with the departure and landing points of saccades in between (Smith and Henderson, 2009; Wilming et al., 2013), is telling in terms of whether users' fixations remain in a limited region within a short time window. There are also measures that quantify the overall spread of fixations during the whole trial by quantifying the uniformity of the spatial fixation distributions (Judd et al., 2011). Given that eye movement patterns change drastically with the task of the user (Yarbus, 1967), it remains to be explored which of these measures will be informative for the analysis of the task-guided gaze behavior during SI use. Accordingly, fine-grained analyses of eye-tracking data that goes beyond fixation duration and locations, and takes into account saccade dynamics and the serial nature of fixation location selection has the potential to reveal the efficiency, and ease of human interaction with medical interfaces.

The contribution of advanced medical expertise to the perceptual and other cognitive abilities of doctors in a medical context is well documented and a comprehensive summary can be found in Reingold and Sheridan's recent review (2011). The authors list more than twenty studies that explicitly compare the gaze behavior of people with different levels of medical expertise. These include comparisons between laypeople and radiology experts searching for abnormalities in mammograms (e.g. Nodine, Kundel, Lauver, and Toto, 1996) or lung nodules in radiographs (e.g. Donovan, Manning, and Crowford, 2008). Wilson and colleagues (Wilson, McGrath, Vine, Brewer, Defriend and Masters, 2011) have compared the eye movements of doctors who were either experts or novices in laparoscopic surgery and found that during crucial phases of a virtual task, the experts kept their eyes fixed for longer. Given that the studies on radiological expertise and eye movements' relationship were inspired by the growing reliance on medical imaging techniques, the development of SIs for surgical operations (e.g. Wilson et al., 2011) motivates similar studies comparing the visual interactions of doctors and laypeople with such medical interfaces.

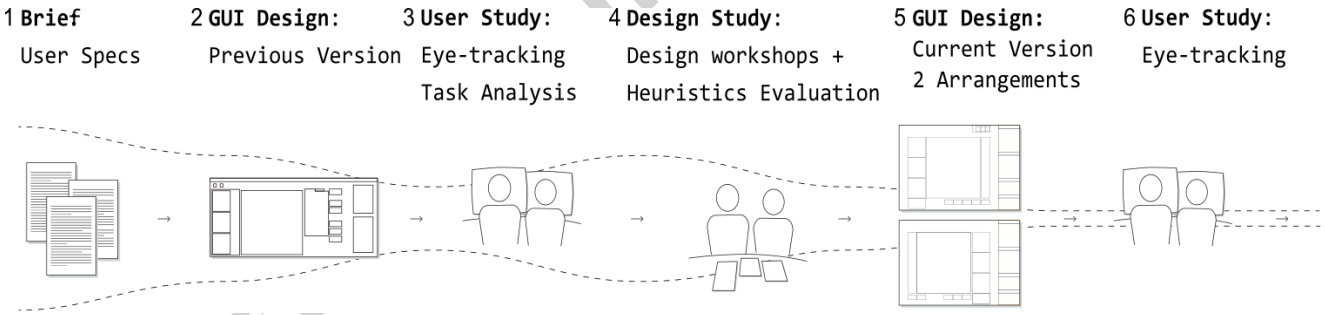
In this study we have collected eye-data from medical doctors and other people with no medical expertise while they interacted with our SI that is developed for the planning and execution of a robotized cryoablation procedure. Our aim was to develop an SI with low complexity and implement a simulation task that would guide the interactive behaviors of the user such that the procedure would be repeatable both within and between subjects. We have tested two arrangements for the SI, which contain the same information but display the information differently. Eye-movement data collected during task execution are subject to a

sophisticated analysis of fixation and saccade statistics. Our results suggest that the viewing patterns obtained with one SI arrangement reflected a more natural viewing-like and efficient interaction experience. We argue that computer-assisted navigation experience of surgeons can be improved significantly by incorporating eye-movement data during the development and evaluation of SI designs.

2. Methods

In this paper we seek the optimum arrangement of a surgical interface (SI), which is being designed for a FP7 European project titled Intelligent Surgical Robotics (I-SUR), using an eye-tracking system. The main goal of the I-SUR project is to demonstrate an autonomous robotic surgical system that can carry out a simple puncturing task, and in particular on the needle insertion for the cryoablation procedure of kidney tumors (Muradore et al., 2015). To achieve this the team members developed kidney phantoms, a robotic system, planners for the robotic system movements, and the current SI that supervises surgical actions, demonstrates the surgical task execution to the surgeon, and provides solutions to the surgeon when unexpected situations occur. As explained below, two different arrangement suggestions have been made for the visualization of the SI. We first explain briefly the design process, and then concentrate on the evaluation of the SI designs by using an eye-tracker to quantify the gaze patterns of doctors and other people with no medical expertise (i.e. non-experts) interacting with the two versions of the SI design.

Figure 1. The user centered design process we followed during the development of the SI for the cryoablation task. The user study at stage three (Erol-Barkana et al. 2014) employed the previous version of the SI. The user study at the final stage is the current report evaluating the two arrangements for the current version.



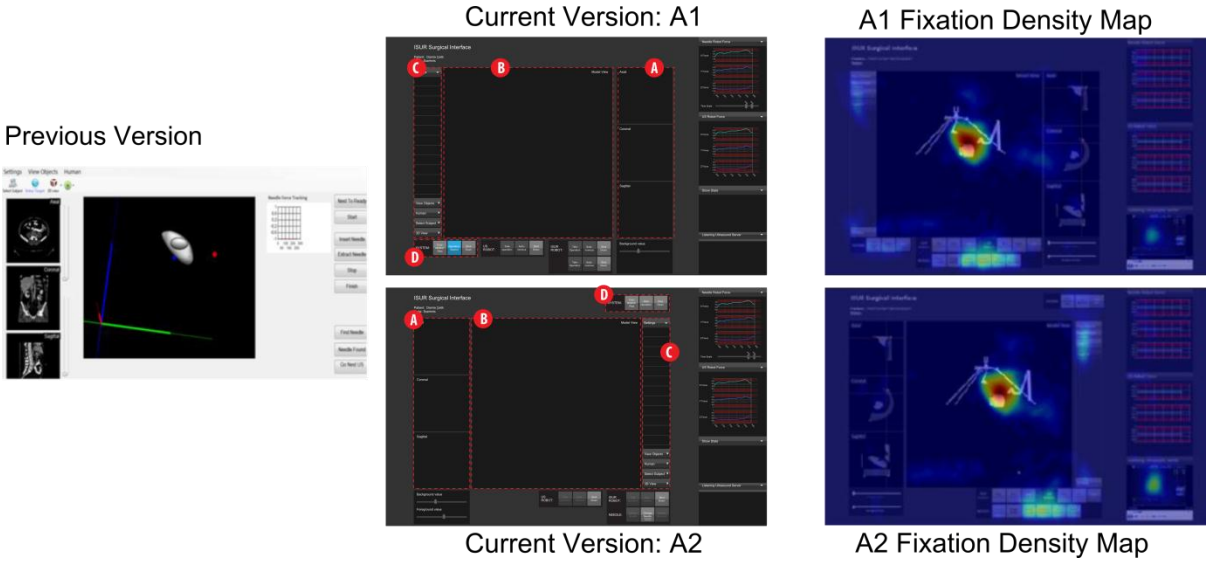
2.1. Design and Development of the SI

Here we provide only a brief summary of the SI design process (Figure 1), which is explained in full detail by Yantaç and colleagues (2014). The previous version of the SI (Figure 2) was developed and tested by Erol-Barkana and colleagues (2014). Incorporation of design heuristics (Chan et al., 2012; Nielsen and Molich, 1990), and the evaluation of the previous version of SI during design workshop meetings (Hanington, 2003) led to the development of the current version of the SI (Figure 2). The improvement from the previous to current version addressed four factors: (1) Color scheme; (2) dialog design; (3) information architecture, and (4) layout design. The overall contrast of the interface was reduced by preferring gray over black and white in order to keep the attention of the operator on the

surgical presentation. Furthermore the buttons are redesigned as active components. During the simulated surgical operation described below, some gray buttons turned to blue, yellow, or red according to whether the operator was provided with information, warnings or error messages, respectively. This change also contributed to the solution of problems caused by previous dialog design using popup screens in the center. Moreover, in the new version, care was taken to ensure that the task-related regions of the interface such as panels, buttons, and other visual representations appeared closer to each other while the size and alignment of each component is reconfigured for the better use of attention.

Apart from these changes, the most critical decision for the current version concerned the locations of the four main visual components of the SI (Figure 2): The three 2D projection view panels (A), main model view panel (B), main model settings panel (C), and the "system shutdown" buttons (D). The better ordering of these components may provide a less complex and more easy-to-use SI given the constraints of the simulation task. With this perspective, we have created two arrangements (A1 and A2, Figure 2) for the current SI version. Both arrangements were developed according to the order of actions expected during the simulated cryoablation operation that will be explained below. They contain exactly the same visual components with identical functionality, but there are changes in the locations of certain components. While A2 displays the main model view panel more into the center, A1 moves the 2D projection view panels to the middle. Another change is about the placement of the "system shutdown" buttons. Unlike the bottom location bringing all buttons together in A1, A2 locates this button set at the top right to distinguish, avoid misuse and save space for the model views on the left. Since both arrangements for the current version of the SI integrate the above mentioned design principles, we did not want to make an arbitrary choice, and hence performed the current study to evaluate the usability of the arrangements with eye-tracking.

Figure 2. Comparison of the previous SI with the two arrangements of the current SI. The panel on the left displays the previous SI and the middle panel depicts the current arrangements (A1 and A2) that are used in the present study. Note the changes in the locations of the components labeled A, B, C, and D. On the right, we superimposed on A1 and A2 the so-called heat maps that show the smoothed distribution of fixations collected from all subjects interacting with the arrangement of interest. Warmer colors indicate more fixations. As can be seen, the fixation hot spots shift according to the location of the components in the two arrangements.



2.2. Participants

We have collected data from 22 participants who volunteered for the experiment. They did not receive money, course credit, or any other incentive. Ten (three females) of those were urologists or radiologists, and we will refer to this participant group as "doctors". They were recruited by calling and visiting nearby hospitals. Twelve additional participants (3 females), which we label "non-experts", consisted of university students (engineering and social science departments) and course instructors of the Social Sciences Faculty of Sabanci University, Istanbul. The latter group is called non-experts simply because they lack medical expertise. They were included in the study to address potential medical expertise differences in SI interaction. Accordingly, we were able to compare the gaze behavior of participants with or without medical expertise while they interacted with the SI.

A questionnaire was prepared to collect information on the doctors' previous exposure to or unfamiliarity with the cryoablation task. The questionnaire consisted of demographic questions designed to solicit knowledge about gender, age, working experience, cryoablation task experience, and familiarity with medical interfaces. The average age of the doctors was 37. Forty percent had work experience for more than 15 years. Ninety percent of them had performed more than 15 cryoablation tasks over the last two years. Only two of them had used a medical interface in their operations.

Prior to coming to the laboratory, all subjects were informed about the SI that we have developed and told that the experiment would entail eye-tracking. Nevertheless, before arriving at the lab, both doctors and non-experts were still naive to the demands of the experimental task that is explained below. Participants in neither group had been previously exposed to the SIs presented in this study.

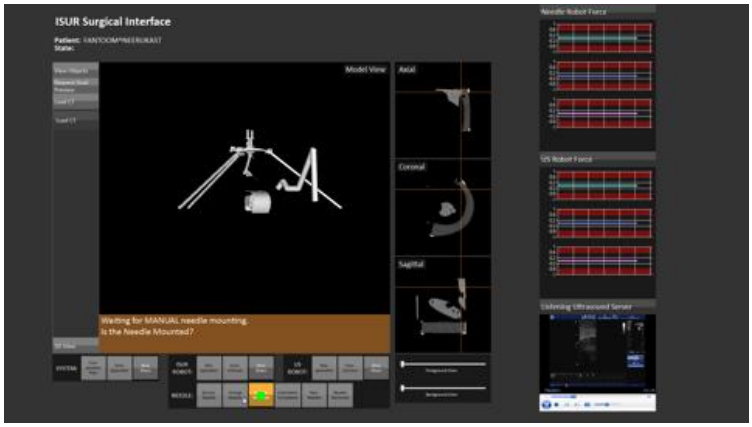
The participant sample size of the current study is relatively small (non-experts: n = 12; doctors: n = 10). Yet, most of the analysis, which is described in detail below, uses all fixations and saccades, and each participant provides hundreds of such data points. This approach is reminiscent of common psychophysics research (Anderson and Vingrys, 2001), in which few participants contribute hundreds of trials. Whereas the analyses based on trial

medians can address either the arrangement or medical expertise differences due to the sample size, using the whole set of fixation and saccade set allows the study of the interaction of these two variables.

2.3. Simulated Cryoablation Task

In order to record and analyze eye movements while users interact with SI, it is important that all participants perform the same task that has a clear beginning state and a final goal. A common task for all users ensures that we can pool the data assuming similar motivational and attentional states. We have developed a simulation task consisting of visually guided mouse clicks to perform the experiments. Cryoablations usually last around 20 minutes and the task here only concerns the needle insertion stage of the procedure. In each trial (see Movie 1 for the screen recording of one complete trial with fixations of a single subject overlaid and descriptive text superimposed), after the SI was loaded and made visible on the screen, the first step was to turn on the three robot visualizations with mouse clicks on the respective buttons. Next, the CT scan images are turned on by clicking the "Load CT" button. Upon the successful completion of this stage, the "Cryoablation Plan" button turned from gray to blue informing the subject that the puncturing plan could be called. We refer to the portion of the trial up the point where this button is clicked is as Phase 1 - Initiation. After the button was clicked, a new window appeared showing a 3D image of the tumor together with the needle insertion trajectories for cryoablation. The user closed the cryoablation plan window after learning how many needles are needed. A smaller window appearing at the center of the screen repeated the number of needles to be used so that it was clear that the user did not miss the information. Upon user confirmation by a mouse click, the window disappeared and the "Start Operation" button changed to blue showing that it was active. The user pressed this button, and the simulated robots started to move. The portion of the task between the end of Phase 1 and clicking of the Start button is called Phase 2 - Plan. In the remainder of the trial, that is, in Phase 3 – Operation, one of subtasks of the user was to click on "Needle Mounted" and "Needle Removed" buttons whenever they turned to yellow. There were two additional aspects of the simulated task that required focused attention. The first was to keep track of how many needles are already inserted. After the insertion of each needle the user was asked to make a choice by the color change of two buttons simultaneously. One button was labeled "New Needle" and the other “Insertion Completed". If the user made the wrong choice, a warning prompt informed the user that more needles are still needed, or that the insertions are already complete. The second task of the user was to react to a system error. In each trial, during the insertion of one needle, an error prompt in red declared that the force limit was reached, which meant that the current needle has to be removed and a new needle was to be inserted. This part of the scenario required the subject to reassess the number of needles still needed. After all needles were inserted, the user was informed that it was now possible to shut down the system and the "Shut Down" button became blue. Phase 3 - Operation and thus the trial ended after the user clicked on that button. Depending on the number of needles, in most cases the trail was completed between 90 and 150 seconds. Thus, our simulation task with multiple stages required the focused attention of the subject and was designed such that we could analyze the speed of reaction to task events.

Movie 1. The simulated cryoablation task as performed by one of the doctors using the SI with A1 layout to insert two needles. Green squares correspond to the fixations of the doctor. Note the congruence of fixations and mouse cursor locations (5 frames per second).



2.4. Procedure

After admission into the lab, the participant first filled a consent form and was informed about the purpose of the experiment. After the instructions were verbally explained, the trials began. Each subject received two training trials and two experimental trials, and all four were slightly different from each other, since the trial type had a 2 by 2 design. One factor was the arrangement of the SI, which had two versions as explained above, and we will refer to them as A1 and A2. The second factor was the number of needles to insert that could be either two or three, which we label N2 and N3. Due to the serial nature of this task, N3 is always expected to last longer than N2. During training, if one trial was A1N2, the other was A2N3. If it was A2N2, then the other was A1N3. After the training the remaining trial conditions were used in the experiment. This ensured that the user has seen both arrangements and both needle amounts during the training and the experiment separately. Thus, half of each subject group - 6 out of 12 for non-experts and 4 out of 10 for doctors - provided A1N2 and A2N3 trials, and the other half provided the A1N3 and A2N2 trials. The order of the two trials inside the training and experiment was randomized. During the first training trial, the experimenter performed the task while verbally describing each stage and the participant watched and listened. The participant was encouraged to interact with the experimenter by asking questions. During the second training trial, the participant performed the task while the experimenter was next to her. Again, the participant was expected to talk to the experimenter by explaining each stage and asking questions. During the experimental sessions, the experimenter was outside of the visual field of the participant. The participant knew that for the next two trials she was expected to perform the task in silence, nevertheless she was told that in case of confusion brief questions could be asked. The total duration of the experiment including training never exceeded twenty minutes. The study conformed to the Declaration of Helsinki and national guidelines for human research, and received approval from the ethics committee of the Sabanci University, Istanbul.

2.5. Eye-tracking & Analysis

Eye-tracking was performed with the Tobii TX300 eye-tracker with a sampling rate of 120 Hz. Eye movement data contains fast, ballistic movements of the eyes labeled saccades, and

fixation periods where the eyes rest at a certain location. The eye-tracker software automatically detects fixations and saccades with an I-VT algorithm applying a velocity threshold of 30 º/s (Olsen and Matos, 2012). Thus, those parts of the gaze data where the eye-position velocity is above that threshold are labeled as saccades. The remaining valid, i.e. non-blink, sections of the data are treated as fixations. The participant was seated in front of the monitor and the distance between the eyes of the subject and our 50.8cm-wide monitor was 60cm. With this configuration the width of the monitor spanned roughly 46º of visual field angle and due to the 1920 pixels horizontal resolution of the SI, 1º of visual angle covered about 42 screen pixels. The eye-tracking was initiated after a nine-point calibration.

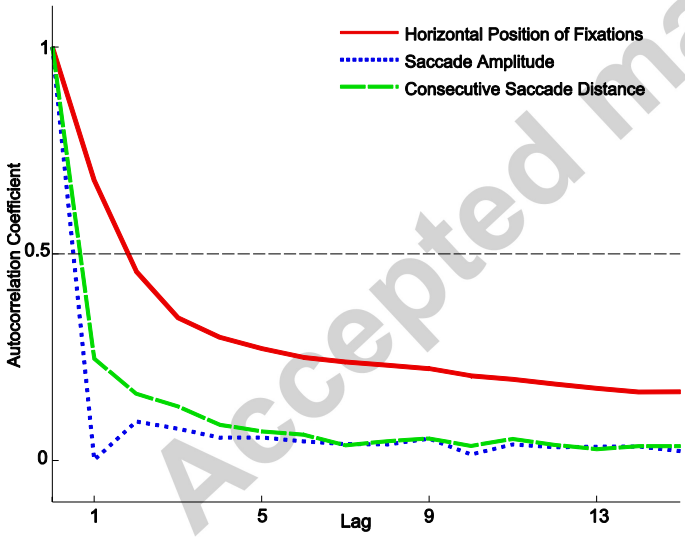
In the current study we have three factors in relation to which we can analyze the data and make comparisons between different conditions. These consist of participant type (two levels, doctor and non-expert), SI arrangement (two levels, A1 and A2), and number of needles required for the task (two levels, two and three). Since each participant completed two experimental trials only, we did not have data from each subject for each of the four SI arrangement and number of needles combinations. Accordingly, we collected four data sets for each trial type combination from doctors, and six data sets in the case of non-experts.

For variables where each trial corresponds to a single sample, such as the task completion time or average fixation duration, the amount of trials we have is not enough to obtain general statistical models that would consider each of three factors described above together with their interactions. In order to circumvent this limitation, statistical testing involved the initial pooling of the data from both needle conditions, since apart from the longer task completion times; there was no specific hypothesis considering the number of needles used in the task. After that we either collapsed the data over the participant types in order to characterize SI configuration differences, or we collapsed the data over SI configurations in order to characterize participant type differences. Bootstrap tests were used to see whether the distributions were significantly different for the different levels of a given factor. Thus, task completion times and trial averages of fixations durations were addressed in terms of configuration and participant type separately, after combining the data for the complementary factor.

The second type of analysis for the eye movement data included all fixations and saccades available in our data set. Distributions of fixations durations, saccade amplitudes and saccade directions were statistically compared across the SI configurations and participant types. Whereas Kolmogorov-Smirnoff test (KS-test) was used for linear variables (durations and amplitude), Kuiper's test was used for the testing of circular saccade direction data. Kuiper's test is similar to the KS-test but is able to compare circular distributions (Berens, 2009). As is the case with all common statistical tests, both the KS-test and the Kuiper's test make the assumption that the samples of the empirical distributions that are being tested are independent. This assumption is likely to be violated in our data since we use all the fixations in the analysis and the locations of successive fixations performed during the viewing of natural images tend to be correlated (Engmann, 2006). Moreover, recent studies reveal that making a saccade from a fixation towards the last or penultimate fixation location is more common than making a saccade in the orthogonal direction (Wilming et al., 2013; Smith and

Henderson, 2009). Even the duration of a fixation can be partially predicted by looking at the angular difference between the saccades that bring the fixation there and move it away, an effect labeled saccadic momentum (Smith and Henderson, 2009). In order to detect and control for the linear statistical dependence in our eye data samples, we have used a method first described by Einhäuser and König (2003) and generalized to fixations by Engmann (2006). For each distribution that would enter the test, we have concatenated the samples from the distribution as a vector and computed the autocorrelation of that vector. The lag at which the normalized autocorrelation value dropped to less than 0.5 was used as a corrective factor in the KS and Kuiper's tests. The degrees of freedom of the tests were manually reduced by dividing the sample size by this corrective factor. As can be seen in Figure 3, for horizontal fixation location, fixation duration, and saccade amplitude, the autocorrelation value at lag 2 is always less than 0.5. Accordingly, the effective sample size for KS and Kuiper's tests was taken as n/2. For all statistical testing, α was taken as 0.05. In summary, SI configuration and participant type differences of eye-data distributions were addressed after correcting for the linear dependencies among the samples of the distributions.

Figure 3. Autocorrelations used for sample size corrections before statistical testing. Whereas the autocorrelation of concatenated horizontal fixation position data falls under 0.5 for the first time at lag 2, for saccade amplitude and the distance between the departure and landing points of consecutive saccades, the autocorrelations are already below 0.5 at lag 1.



Our final analysis addressed the return saccades, i.e. saccades that bring the eyes to a location that was the departure point of the preceding saccade (Wilming et al., 2013). For each consecutive saccade pair, we have computed the distance between the departure point of the first one and the landing location of the second and created the empirical distributions of these distances for the two arrangements and the two subject groups. Please note that for any nth fixation, the departure point of the preceding saccade is related to the location of fixation n-1, and the landing point of the following saccade is related to the location of fixation n+1. Accordingly, the lag that is computed for fixation location data in order to correct for

correlated samples is already captured in this variable. This is confirmed by the inspection of the autocorrelation function where the coefficient assumes a value smaller than 0.5 already at lag 1 (Figure 3). In order to see whether these distributions are a result of overall viewing biases or correspond to a viewing strategy whereby saccade landing points close to the departure points of the previous saccade are preferred, we have shuffled the saccades to remove order effects and created distance distributions for the shuffled data set (Wilming et al., 2013). Any significant difference between actual and shuffled distance distributions would reveal aspects of the viewing strategy used during interaction with the SI, and this was again quantified using KS-tests. This analysis addressing the distances between fixated locations is informative about whether one SI arrangement leads to more clustered viewing.

Even if the users tend to display clustered viewing with successive saccades remaining in a local region, the overall spread of fixations may be more or less uniform. For this purpose, we have first generated fixation density maps for each arrangement and subject group. Each fixation density map was generated by creating an SI-size matrix with ones at fixations and zeros elsewhere that was then convolved with a two-dimensional Gaussian window that had a full width at half maximum of 1º. The map was normalized to obtain a probability distribution that sums up to one. A common metric employed to quantify the spread of fixations and estimate the uniformity of the fixation map is the information theoretic measure of entropy. Since it is a measure of randomness, the entropy is higher if the fixations are spread over the viewed area, and it is lower if fixations accumulate in few local regions. It is estimated using the following formula:

$$-\sum_{i=1}^{N} p(x_i)\log(p(x_i))$$

Each $p(x_i)$ corresponds to the fixation probability in a local image region. The base of the logarithm that is used is arbitrary since we are interested in comparisons across conditions. We have used base two and hence the unit of entropy is bit. The discretization of the empirical probability histogram is known to influence the estimation of the entropy, since different bin sizes will reveal different approximations of the underlying probability distribution (Wilming et al., 2011). Accordingly, we have used primarily two bin size selections that are used in previous studies: 16x16 bins, (Judd et al., 2011), and bins approximately covering the area of a circle with a diameter of 2º, thus 13x24 bins (Wilming et al., 2011). Since the total number of fixations for each participant group and arrangement combination is different, we had to ensure that the entropy estimation does not depend on sample size. In order to circumvent this problem, each fixation density map contained 2000 fixations selected with replacement from each combination. This procedure was repeated 5000 times in order to obtain bootstrap distributions for entropy estimations for each of the four participant and arrangement combination. We will report the median and the 95% percentiles for these distributions in order to allow their comparison. Thus, the information theoretical measure of entropy is estimated from the fixation probability maps to quantify the overall spread of fixations on the SI.

In order to characterize how the users respond to on-screen events that required to them to click certain buttons of SI, we have performed two additional analyses. These events (see Movie 1) comprise color changes of certain buttons and the user is expected to click the button upon color change. By doing so the user initiates the operation, shuts the system down, confirms whether a needle is inserted or removed, and decides whether more needles are needed for cryoablation or whether the task is complete. For each such event, we have first collected the latency of mouse click responses as a measure of reaction time. Accordingly, we could analyze whether factors such as arrangement or participant expertise modulates these response latencies. Second, we have measured the average eye position location before the event, relative to the location of the proceeding mouse click. This allows us to check whether the eyes were relatively close to or away from the event just before its occurrence. For this purpose, we have calculated the median eye position within a 400ms temporal window preceding the event, independent of whether the sample belonged to a saccade or fixation. We have subtracted from that the location of the event in order to compute the distance in between. Thus, by characterizing the response latencies to task relevant events and measuring the distance of the eyes just before the event, we could address the relationship between arrangement and task performance independent of accuracy.

Some of the detected fixations and saccades were removed from the data before analysis. One reason for the removal is the presence of corrective and miniature saccades that are not informative for attentional switches between different parts of the SI (Rolfs, 2009). Another problem with fixation and saccade detection using eye position data that is collected during the presentation of dynamic stimuli is the presence of smooth-pursuit movements (Larsson et al., 2014; Valsecchi et al., 2013). These relatively slow eye-movements allow keeping a visual target on the foveal region of the retina while either the perceiving agent or the target moves in the environment (Sparks, 2002). The SI that we have developed contains the dynamic representation of the surgical robots that move during the simulated cryoablation task. Visual inspection of the data revealed that it was common for the subjects to watch the movements of the robot on the screen. Nevertheless, our concern was the fixations and saccades of the subject, since the task involved switching visual attention back and forth between information panels and display buttons. Moreover, whenever the subject had to interact with the SI in order to issue commands and respond to warnings, the robot images were perfectly still. In order to reduce the amount of corrective and miniature saccades and smooth-pursuit in the data, we have removed fixations that lasted shorter than 60ms (1.6%) and saccades with amplitudes less than 1º of visual angle (30%).

3. Results

3.1. Task Completion Times

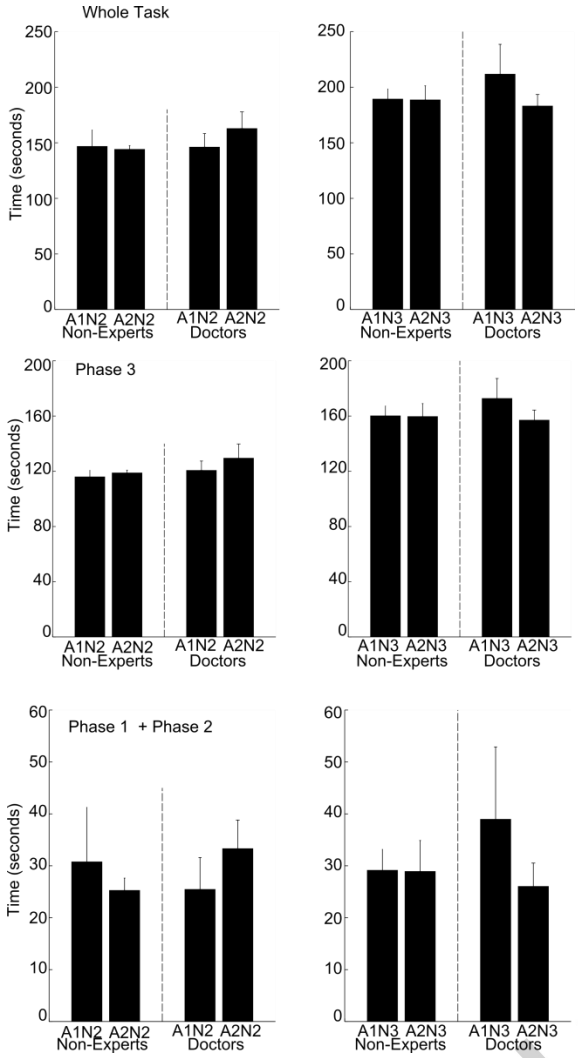
We have first addressed the time it takes to complete the simulation task and the duration of the individual phases of the task. Naturally, the task requiring more needle insertions lasted relatively longer due to the serial nature of the task (Figure 4A). This difference was related to the Phase 3 - Operation of the trial, since it is in this phase that the needles are inserted one after another (Figure 4B and Table 1). The combined total duration of Phase 1 – Initiation and Phase 2 – Plan is provided in Figure 3C. As can be seen, the differences across

arrangement and needle conditions are not more than 10 seconds. In order to compare the task and phase completion times for the two arrangements; we have collapsed the data over number of needles and subject type. Thus, we ended up with 22 completion times for each configuration. Bootstrap tests revealed no significant differences between the task and phase completion times obtained with the two configurations (total duration $p = 0.33$, Phase 1 $p = 0.11$, Phase 2 $p = 0.39$, Phase 3 $p = 0.45$). Next we have pooled over number of needles and configurations and addressed subject type differences. Accordingly, we obtained 24 trials for non-experts and 20 trials for doctors. There were no significant differences in the task and task phase completion times of the two subject groups (total duration $p = 0.15$, Phase 1 $p = 0.11$, Phase 2 $p = 0.47$, Phase 3 $p = 0.18$). Thus, task and task phase completion times revealed neither configuration nor subject type differences.

Table 1: Mean (standard deviation) task phase completion times in seconds

		A1N2	A1N3	A1N3	A2N3
Doctors					
	Phase 1	17.1 (4.5)	22.4 (3.9)	29.0 (11.1)	17.9 (3.6)
	Phase 2	8.3 (1.3)	10.9 (1.2)	10,0 (2.3)	8.1 (0.6)
	Phase 3	120.1 (6.9)	129.4 (9.2)	172.7 (12.9)	157. 0 (6.6)
Non-Experts					
	Phase 1	21.0 (6,8)	16.5 (2.0)	19.3 (2.8)	19.5 (4.3)
	Phase 2	9.7 (2,8)	8.8 (0.9)	9.7 (1.7)	9.4 (1.8)
	Phase 3	115.9 (4.3)	118.8 (1.8)	160.1 (6.5)	159.7 (8.6)

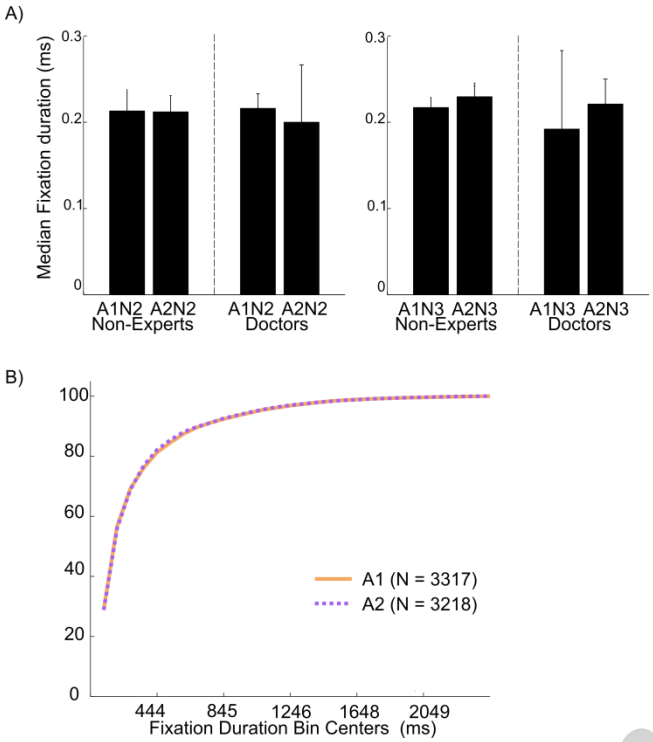
Figure 4. Participant means of task and task phase durations. A) Total task durations for each subject type, number of needles (N), and arrangement (A) combination. B) The total duration of Phase 1 - Initiation and Phase 2 – Plan combined. Error bars denote standard deviations over participants. Note the increase in task duration with the addition of one more needle in N3 trials that is purely related to the Phase 3 in which the needles are inserted. For bootstrap tests on subject type or design-collapsed data please refer to the text.



3.2. Gaze Analysis

Do the fixation durations change with SI configuration or subject type? In order to approach this question we have performed two types of analysis. Figure 5A displays the median fixation durations for different conditions and the whole trial. There were no significant differences for the median fixations between the two SI configurations after the data pooling described above (bootstrap tests, design differences: total task duration $p = 0.45$, Phase 1 $p = 0.073$, Phase 2 $p = 0.37$, Phase 3 $p = 0.26$; expertise differences: total task duration $p = 0.48$, Phase 1 $p = 0.11$, Phase 2 $p = 0.28$, Phase 3 $p = 0.41$). Next, we have performed a more fine-grained analysis, in which all fixations in the trial served as samples. For that purpose, we have compared the cumulative distribution functions of fixation durations. As explained in the Methods section, statistical testing involved KS tests with corrected degrees of freedom, whereby the effective sample size was reduced to half. As can be seen in Figure 5B, the cumulative distributions of fixation durations of the whole trial look nearly identical ($p = 0.45$). This also holds within each participant group (doctors $p = 0.50$, non-experts $p = 0.27$) and each phase of the experiment (all ps > 0.14). Thus, the median trial fixation durations did not reveal any SI arrangement differences, and the expertise of the doctors did not seem to play a role either.

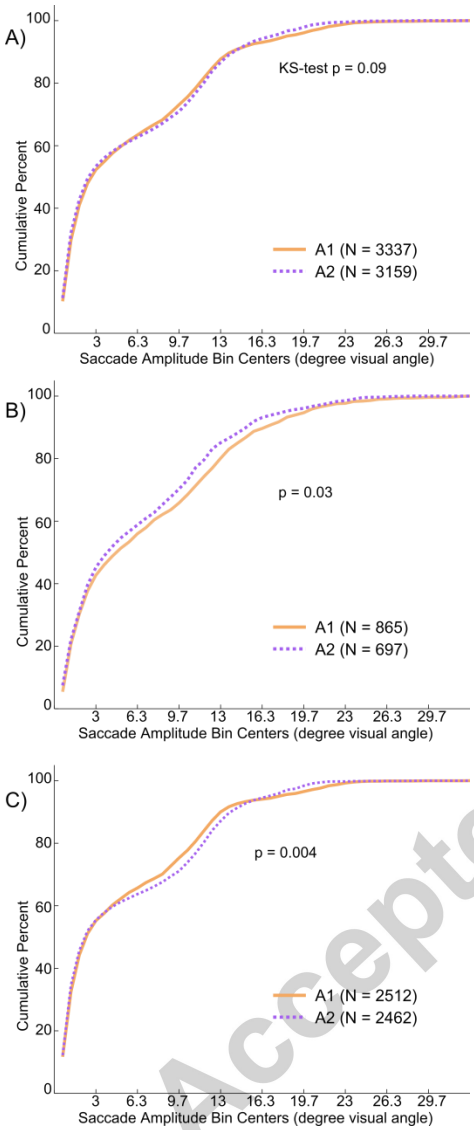
Figure 5. Fixation durations (A) Median fixation durations for the whole trial. Error bars denote standard deviations over participants. For statistical tests on subject type or design-collapsed data please refer to the text. (B) Cumulative distribution functions of fixation durations. The figure displays the fixations of all participants separately for A1 and A2. Note the curves are nearly identical.



The next analysis concerned saccade amplitudes. For the subject group, number of needles, and task phase pooled data, Figure 6A shows the cumulative distribution functions of saccade amplitudes obtained with the two configurations separately. The two distributions are not statistically different from each other (sample-size corrected KS-test, $p = 0.09$). The same holds if the analysis is repeated for each subject group individually (doctors $p = 0.34$, non-experts $p = 0.17$). However, there are task phase specific differences. Even though the two distributions for the different arrangements are still statistically indistinguishable from each other if Phase 1 and Phase 2 are considered individually, when data from both of these phases are pooled, it appears that in the case of A2, short amplitude saccades are relatively more frequent ($p = 0.03$). As can be seen in Figure 6B, during this period the participants inspect the A2 with relatively smaller amplitude saccades. The same analysis for individual subject groups do not reveal significant differences between the distributions (doctors $p = 0.19$, non-experts $p = 0.28$). However, during the last and longest phase of the experiment, where the subjects issue needle insertion and removal commands, and monitor the movements of the robots, the significant difference between the two distributions ($p = 0.004$) is in the other direction. That is, there are more low amplitude saccades for A1 compared to A2 (Figure 6C). This result was replicated within the non-experts ($p = 0.02$), but not within the doctors ($p = 0.13$). In summary, whereas during the initial encounter with the interface and the viewing of the cryoablation plan the participants view the A2 with lower amplitude saccades, such lower

amplitude saccades are more abundant during the cryoablation itself when the A1 arrangement is used.

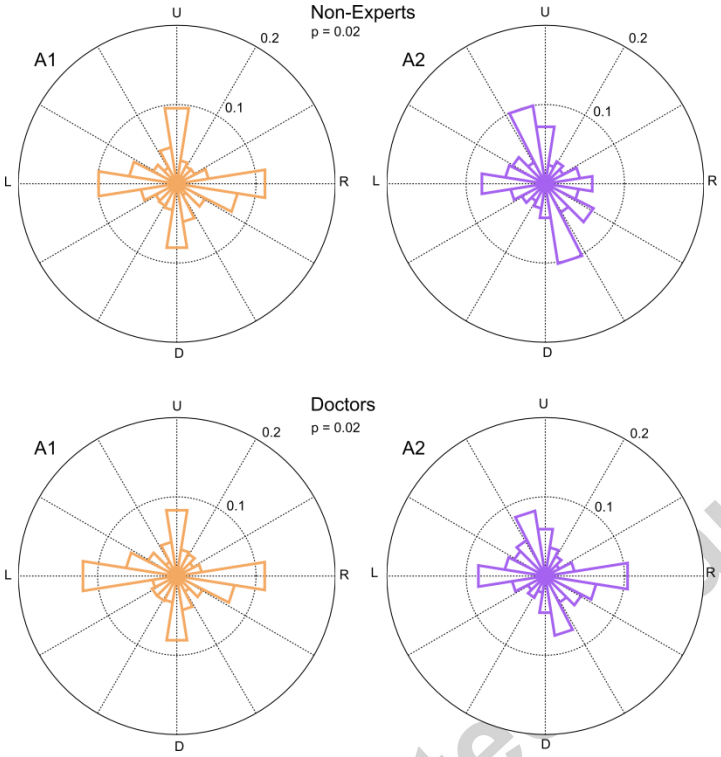
Figure 6. Cumulative distribution functions for saccade amplitudes. (A) All saccades. (B) Saccades executed during Phase 1 – Initiation and Phase 2 - Plan. C) Saccades executed during Phase 3 – Operation. The p-values are obtained with KS-tests. The distributions display subject pooled data. For participant group data, please refer to text.



While interacting with the two types of SI configuration, do the participants make saccades towards different directions? In order to answer this question, we have created circular distributions of saccade direction angles (Figure 7). The differences that are readily visible in the plots are confirmed with sample size-corrected Kuiper's tests, analogous to KS tests described above. The comparison of circular distribution functions reveals significant differences for the whole duration of the task both when all subjects are included in the analysis ($p < 0.001$), and when the configurations are compared separately for each participant group (doctors, $p = 0.02$; non-experts, $p = 0.02$). Limiting the comparison to the Phase 3 -

Operation replicates the same result (p = 0.02). Thus, interaction with A1 leads to more horizontally and vertically oriented saccades.

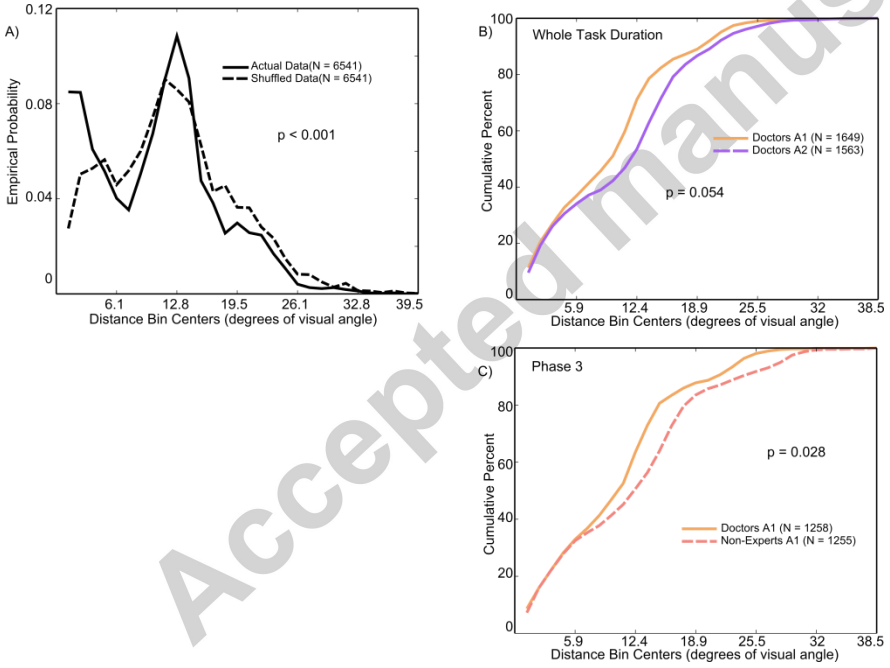
Figure 7. Polar probability distribution functions for the directions of saccades executed during the experiment. Upper panels: Non-experts data (A1 saccades N = 1705, A2 N = 1580). Lower panels: Doctors data (A1 N = 1675, A2 N = 1581). Note the abundance of saccades in cardinal directions in the panels on the left. L: Leftward saccades, R: Rightward, U: Upward, D: Downward. In each individual plot, the 0.1 and 0.2 correspond to the probability. Kuiper test p-value results are given below the participant group labels.



For successive saccades pairs, we have addressed the spatial relationship between the departure point of the first saccade and the landing point of the next saccade. Distributions were prepared for the distances between the departure and landing points of saccades that preceded and followed any given fixation point, respectively. In order to see whether the distribution of these distances is a viewing strategy or simply a result of obtained saccade amplitudes, we have compared this distribution to a shuffled baseline distance distribution. As can be clearly seen (Figure 8A), departure and landing points of two respective consecutive saccades tend to be relatively closer to each other, compared to the shuffled baseline distribution. The difference between the underlying distributions is confirmed by a KS test ($p < 0.001$). Next we checked whether there are subject type and/or design differences in the distance distributions. There were significant design differences for the distance distributions in the case of subject group pooled data ($p = 0.009$). There was a statistical trend in the case of doctors ($p = 0.054$) but not in the case of non-experts ($p = 0.310$). The observed design differences between the distance distributions stemmed from relatively close distances between the departure and landing points of successive saccades while interacting with A1 (see Figure 8B for doctors' data). Limiting the analysis to the

saccades of Phase 3 – Operation revealed significant differences for all comparisons (participant groups pooled $p < 0.001$; doctors $p = 0.014$; non-experts $p = 0.002$). In order to scrutinize the subject group differences we have compared doctors and non-experts data for the two designs separately. In Phase 3, the distance distributions for doctors and non-experts were different while interacting with A1 ($p = 0.028$, Figure 7C), but not with A2 ($p = 0.24$, for the whole task duration there was only a trend $p = 0.062$). Thus, for fixations, the departure points of the saccades preceding them and the landing points of the saccades following them are closer to each other especially for doctors interacting with A1, revealing an expertise and design interaction.

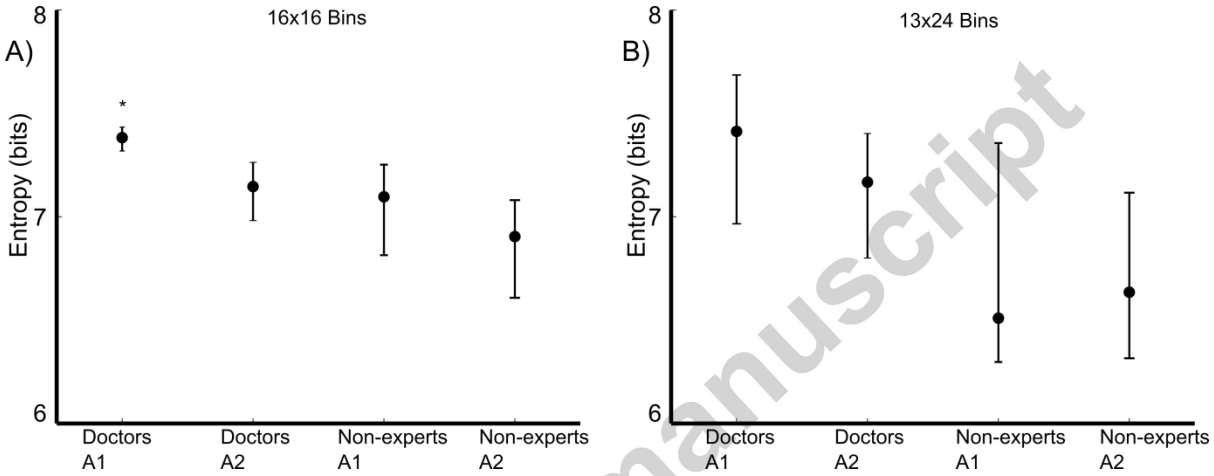
Figure 8. Analyses of the distances between the departure point of a saccade and the landing point of the next saccade. (A) Empirical probability distributions of the distances obtained with the actual ordering of fixations and fixation position shuffled data that serve as a control. It can be clearly seen that short distances are more prevalent in the actual data. (B) Cumulative distribution functions of distances obtained from doctor data with arrangements kept separate. (C) Cumulative distribution functions of distances obtained from A1 data with expertise kept separate. Thus, the solid (orange) curve is nearly identical in (B) and (C) and the difference is due to the omission of Phase 1 and 2 saccades in panel C). All p-values pertain to KS-tests. For other details please refer to text.



In order to see whether there were participant group or arrangement differences in the spread of fixations over the SI, we have estimated the entropy of the fixation probability maps. Please note that the more spatially uniform the fixation probability, the higher is the entropy. As can be seen in Figure 9, the entropy of doctors' fixations was higher for both bin sizes used in the calculation. In the case of 16x16 bins, there was no overlap in the 95% confidence intervals (CIs) of the bootstrap distributions computed for the A1-doctors data and the other

combinations. The remaining three combinations had overlapping CIs. Please note the higher variance in the computation with more bins, suggesting a more reliable probability distribution estimation with 256 bins. Indeed, the computations with slightly fewer bins replicated the observations with 16x16 bins. These results show that the doctors' fixations covered the SI more uniformly than non-experts, and that this difference is higher for doctors viewing the A1.

Figure 9: Entropy analysis for the spread of fixations. A) Entropy computed with fixation probability maps binned as 16x16. B) The same analysis with 13x24 bins. For both panels, the circular data point corresponds to the median of bootstrap distribution, and the error bars denote 95% confidence intervals (CIs). The star denotes the absence of CI overlap between the Doctors-A1 data and the other three sets of data.



Finally, we have addressed the relationship between the SI arrangement and doctors' performance by analyzing their reaction times to task relevant events displayed on the SI. As explained before, while the user executed the task, changes in the colors of the buttons informed the subject that these buttons were to be clicked. We have analyzed the latency between the button's color change and the expected mouse click. For each arrangement there were 124 mouse clicks. The median latencies were 1869 ms for A1 and 1881ms for A2 clicks. These medians were not significantly different from each other (bootstrap test, $p = 0.48$). Next, we wondered where the gaze position of the user was just before and relative to the event. For that purpose we have obtained the median gaze position on the SI in a 400 ms temporal window ending just immediately before the event. The eyes were on average 12.9º away from the upcoming event location while viewing A2. The distance was 13.2 º in the case of A1 and there was no significant difference ($p = 0.45$). Thus, the arrangement did not have an influence on reaction times to task relevant events and the gaze was at a comparable distance from these events before their onset.

4. Discussion and Conclusion

We have evaluated the arrangement alternatives proposed for our surgical interface (SI) by analyzing the eye movements of doctors and others interacting with this interface during a simulated cryoablation task. Specifically, we have checked whether one of the two SI arrangements (A1 and A2), both developed according to design principles, is ergonomically

superior to the other. For that purpose we have used statistics derived from the gaze patterns of our participants (doctors and people with no medical expertise that we label non-experts) including measures that are either commonly encountered in the field or used in a human-computer interaction context for the first time. We observed both similarities and differences in gaze behavior across the two arrangements. As we explain below, the differences obtained with several eye movement measures suggested that the usability of one SI arrangement, A1, was better than the other.

The task of the users interacting with the SI was a realistic simulated cryoablation task consisting of turning on physiological recording panels, viewing the operation plan, and needle insertions. Informal conversations with the doctor participants after the experiment revealed that they have found the operation very easy to grasp and handle. Nevertheless, there were no SI arrangement differences for task execution durations and response latencies to task relevant events on the screen. Even though it remains possible that such differences could be demonstrated with a larger participant pool, we believe it is more important to consider the similarities between the two arrangements. Both versions of the SI included the same information and visual elements such as panels, buttons and representations of the surgical robots. Moreover, they both were designed after several improvements to the earlier versions of SI (Erol Barkana et al., 2014; Yantaç et al., 2014). Thus, while interacting with the two versions of the SI that were identical to each other in terms of task-relevant information content, reactions to task relevant events and task completion times did not reveal arrangement differences.

We have addressed overt attentional usability differences for A1 and A2 first by analyzing fixation durations. In usability research, fixation durations are one of the most popular statistics since they might reveal the ease of interaction with an interface (Jacob and Karn, 2003). Long fixation durations are taken as indicators of difficulty in visual information acquisition during the fixation (Goldberg and Kotval, 1999). Yet this depends highly on the contingencies of the task. For instance, Wilson and colleagues (2011) demonstrate that while performing a simulated surgical task that requires precise tool manipulation, more experienced surgeons were able to keep their eyes more still and completed the task sooner than novices. In our data, we have not found statistically reliable differences in fixation durations while comparing the subject groups or arrangements. We believe that this is due to the relative simplicity of the task we have employed. The actions and the gaze position of the user were guided by changes in the colors of the SI buttons. Furthermore, most fixations were either related to watching the robot movements or reading the text displayed on screen. The identical movement and text content of the two arrangements and the independence of reading ability and movement tracking from medical expertise may explain the absence of arrangement and expertise differences in fixation durations.

We find the saccade direction differences observed during interaction with the two arrangements relevant for the final layout decision. Both the doctors and other participants executed more saccades along the horizontal and vertical directions while using the A1 version. An abundance of eye movements along these cardinal directions is a characteristic of natural image viewing (Foulsham et al., 2008). Nevertheless, this is related to the presence of

vertical and horizontal contours in urban and natural settings, such as the horizon line, trees and building edges. The most prominent saccade directions shift in correspondence with changes in the orientation of the presented images. However, during viewing of fractal images, horizontal saccades are more prominent than eye movements in other directions independent of image orientation, and this finding cannot be explained by the rectangular frameworks in which the visual stimuli are usually placed (Foulsham and Kingstone, 2010). This suggests an image-content independent bias for the execution of horizontal saccades, and might reflect a default viewing strategy. Moreover, in the vast majority of world languages, reading is characterized by horizontal eye movements, and vertical reading is the only exception to this rule. In interface design, the relative placement of components vertically and horizontally increases balance and improves usability and aesthetic appeal (Altaboli & Yin, 2011). Such arrangements would elicit more horizontal and vertical eye movements than other arrangements. We argue that, all other things being equal, the interface designs that lead to viewing with more saccades in horizontal and vertical directions must be preferred, since this reflects both the content of natural settings humans occupy, their image-independent viewing strategies and their encounters with symbolic material.

Saccade amplitude differences reveal nuances in the usage of the two arrangements. The initial interaction with the SI - the approximately 30 second long window including the visualization of the CT scans and the viewing of the cryoablation plan - contained relatively shorter saccades during interaction with A2. This may reflect that the panels that are used only within this part of the interaction may have been closer to each other compared to A1. However, during the longer operational phase in which the needle insertions are executed and monitored, shorter saccades were more prevalent with A1. Saccade amplitudes depend on the type of task (Goldberg et al., 2002) and whereas longer saccades are indicators of efficient and directed search, shorter saccades imply focal processing and conscious analysis of visual information (Burmistrov et al., 2015). Whereas during early encounter with the SI the users are expected to find relevant information efficiently, the operation phase requires focused attention for issuing task related commands and monitoring the consequences. Once again, note that A1 and A2 contain the very same information, and the cryoablation task is identical with each of them. Shorter saccades with A1 reflect that the functional components of the SI were adjusted such that the eyes had to travel shorter distance between fixations on those components during the crucial operational phase, revealing a more practical viewing behavior.

Recent eye movement research employing natural scenes profits from novel statistical measures to characterize viewing behavior in terms of the coverage of the viewed field both by few successive fixations and the whole fixation set. The former approach focuses on the directional and locational relationships between successive saccades and fixations (Smith and Henderson, 2009; Wilming et al., 2013). Unlike saccade amplitudes that relate to the distance between pairs of successive fixations, this type of analysis addresses the viewing patterns that evolve over a period of three fixations by quantifying spatial properties of the departure point of a saccade, the fixation followed by that saccade, and the landing point of the saccade that terminated the fixation. This type of analysis is valuable since it demonstrates whether people

tend to revisit locations that were fixated recently, and if yes, whether this depends on the information content of the region of interest (Wilming et al., 2013). Our data show that A1 induces a more locally clustered viewing pattern during task execution. This is revealed by relatively short distances between the departure point of a saccade that brings the eyes to a location and the landing point of the saccade that takes the eyes away from that location. Thus, task demands were met by the users in conjunction with nearby successive fixation locations while using the A1. This shows that the panels and buttons of the SI were better placed in A1 and the users could execute the task easily by shifting the focus of their attention between nearby regions. Moreover, during interaction with the A1, the doctors' viewing patterns were characterized by even shorter distances than the non-expert data. To sum up, fixations with two saccades and one fixation in between tend to be closer to each other during interaction with A1, and this effect is stronger for doctor users. Also considering the lower saccade amplitudes with this arrangement, the doctors interacting with the A1 were able to keep their gaze in a limited portion of the SI over sequences of at least three fixations.

That the eyes travel relatively shorter distances during interaction with A1 does not automatically imply that the whole viewing pattern in this case was limited to a few regions of the SI. Indeed, the analysis of the uniformity of the spatial fixation distributions over the whole SI reveals that it is the other way around. The entropy analysis (Judd et al., 2011, Wilming et al., 2011), which quantifies the overall spread of fixations on the viewed field, showed that doctors' fixations are more spread over the image while they interact with A1. In other words, doctors interacting with the A1 version of the SI had visited nearby locations over few fixations, and still observed the SI with more coverage. Taken together, we argue that locally clustered viewing over short temporal scales coupled with more globally spread fixation behavior over the whole course of the trial is a more efficient viewing strategy and characterizes the gaze behavior of doctors interacting with the A1.

The present findings summarized above reveal how layout principles for visual design (Schlatter and Levinson, 2013) can be confirmed by profiting from eye movement data. Furthermore, after the experiments some of our doctor participants explicitly mentioned that using A1 was easier, without being able to provide concrete reasons for their preference. Even though a more complete evaluation of the SI necessitates its combined usage with the real robotic system, and perhaps with an operation on a 3D model or even an animal, the current simulation study significantly extends our previous work (Erol-Barkana, 2014). Thus, saccades along cardinal directions, the shorter distances between locations fixated successively, and more spread fixation behavior over the whole trial are all in line with the subjective reports of the doctors stating that the A1 design for the SI is superior.

Even though we interpreted the current results as providing sufficient evidence for the final SI arrangement selection, there are many other tools that could address eye movements for interface design. In the present study we have analyzed fixation and saccade statistics independent of interface content. In the future, this type of analysis must be coupled with the analysis of low and high-level content at the center of gaze (Onat et al., 2014). Both low-level properties such as luminance contrast, color and movement, and task related higher-order information such as the function of individual buttons and panels can interact with aspects of

saccade and fixation metrics. Accordingly, in our ongoing work we are using low and high-level characterizations of visual saliency in order to reveal which regions of the SI are more attractive for the eyes and whether this is congruent with task demands.

There are several approaches to the improvement of usability of medical technologies (Goldberg and Wichansky, 2003). These include the incorporation of general usability principles and heuristics (Weinger et al., 2010), consideration of user group differences (Vincent and Blandford, 2014), and conduction of interviews with the users (Privitera et al., 2009). There are several suggestions about how to make a medial interface look better and easier to use. Nevertheless, these approaches rely either on the validity of knowledge gathered from other fields within the medical context, or, in the case of interviews, on the subjective ratings of users. Objective measures that directly quantify aspects of user behavior are usually limited to task accuracy and task completion times (e.g. (Yang et al., 2012)). Even though these two measures are vital for surgical operations, they might fail to reflect other aspects of user behavior that depend on the arrangement of the interface in use. Eye movements, on the other hand, provide a moment-to-moment reflection of the interactive experience of the user. Visual attention research shows that both the visual aspects of the stimulus such as the spatial distribution of color, contrast, movement and the task of the person interacting with the stimulus determine the regions to attend to (Açık et al., 2014; Einhäuser et al., 2008; Onat et al., 2014; Parkhurst and Niebur, 2003). Even though there has been an increase in studies addressing user gaze patterns during interaction with medical information systems (Law et al., 2004; Mello-Thoms et al., 2002; Schulz et al., 2011; Zheng et al., 2011), very few studies profit from this type of analysis for the development and evaluation of surgical interfaces in order to improve their usability (Asan and Yang, 2015).

Here, by analyzing eye movement measures that are usually employed in natural scene viewing literature, we concluded that one of the two arrangement suggestions for our SI is superior from a design perspective. SIs are becoming an indispensable tool in the operating room since they lead to more successful and less invasive procedures (Mezger et al., 2013; Münzer et al.,2006). Nevertheless, complicated and difficult-to-use systems are more likely to repel medical professionals, which could have serious consequences for patients (Martelli et al., 2007). Methodological coupling of visual design principles and eye movement analysis exemplified by the results presented here will lead to the development of more efficient and easy-to-use surgical interface systems and improve public health by helping the surgeon in the operating room. The natural scene viewing tradition, on the other hand, may find a test bed for theories of overt visual attention by considering the ecologically relevant tasks of the SI design research and the domain-specific knowledge of the medical doctors. Accordingly, more cross-talk between SI design communities and natural scene perception researchers will benefit both of the fields.

Acqknowledgements

We are grateful to Peter König for suggestions on data analysis and Esra İpek Gülergin for coding the timing of events. The research leading to these results has been funded by the European Union Seventh Framework Programme FP7/2007-2013 under grant agreement

n.270396 (Intelligent Surgical Robotics, I-SUR). This work was supported by the BAGEP Award of the Science Academy.

References

Açık, A., Bartel, A., & König, P. (2014). Real and implied motion at the center of gaze. Journal of vision, 14, 2.

Altaboli, A., & Lin, Y. (2011). Investigating effects of screen layout elements on interface and screen design aesthetics. Advances in Human-Computer Interaction, 10, 5-15.

Anderson, A. J., & Vingrys, A. J. (2001). Small samples: Does it matter? Investigative Ophthalmology & Visual Science, 42, 1411-1413.

Asan, O., & Yang, Y. (2015). Using eye trackers for usability evaluation of health information technology: A systematic literature review. JMIR Human Factors, 2, e5.

Berens, P. (2009). Circstat: a matlab toolbox for circular statistics. J Stat. Softw, 31, 1-21.

Bergstrom, J. R., & Schall, A. (2014). Eye tracking in user experience design. Boston, MA: Elsevier.

Burmistrov, I., Zlokazova, T., Izmalkova, A., & Leonova, A. (2015). Flat Design vs Traditional Design: Comparative Experimental Study. In Human-Computer Interaction–INTERACT 2015 (pp. 106-114). Springer International Publishing.

Calisir, F., Basak, E., & Erol Barkana, D. (2014). Relative importance of usability and functionality factors for computer-assisted navigation system for cryoablation of kidney tumors. In Global Conference on Healthcare Systems Engineering (GCHSE) (pp. 28-33).

Chan, A. J., Islam, M. K., Rosewall, T., Jaffray, D. A., Easty, A. C., & Cafazzo, J. A. (2012). Applying usability heuristics to radiotherapy systems. Radiotherapy and Oncology, 102 , 142-147.

Donovan, T., Manning, D. J., and Crawford. T. (2008). Performance changes in lung nodule detection following perceptual feedback of eye movements. Paper presented at the Medical Imaging 2008: Image Perception, Observer Performance, and Technology Assessment, San Diego, CA, USA.

Dorr, M., Martinetz, T., Gegenfurtner, K. R., & Barth, E. (2010). Variability of eye movements when viewing dynamic natural scenes. Journal of vision, 10 , 28.

Einhauser, W., & König, P. (2003). Does luminance-contrast contribute to a saliency map for overt visual attention? European Journal of Neuroscience, 17, 1089-1097.

Einhauser, W., Spain, M., & Perona, P. (2008). Objects predict fixations better than early saliency. Journal of Vision, 8, 18.

Engmann, S. (2006). Integration of Luminance Contrast and Colour Contrast in Directing the Human Gaze volume 2-2006. Institute of Cognitive Science.

Erol Barkana, D., Açık, A., Duru Goksel, D., & Duru, D. (2014). Improvement of design of a surgical interface using an eye tracking device. Theoretical Biology and Medical Modelling, 11 , S4.

Fasquel, J.-B., Waechter, J., Goffin, L., Nicolau, S., Agnus, V., Soler, L., & Marescaux, J. (2008). A xml based component oriented architecture for image guided surgery: illustration for the video based tracking of a surgical tool. In Insight Journal, Workshop on Systems and Architectures for Computer Assisted Interventions, 11th International Conference on Medical Image Computing and Computer Assisted Intervention.

Foulsham, T., & Kingstone, A. (2010). Asymmetries in the direction of saccades during perception of scenes and fractals: Effects of image type and image features. Vision Research, 50, 779-795.

Foulsham, T., Kingstone, A., & Underwood, G. (2008). Turning the world around: Patterns in saccade direction vary with picture orientation. Vision Research, 48, 1777-1790.

Goldberg, J., & Wichansky, A. (2003). Eye tracking in usability evaluation: A practitioner's guide. In J. Hyona, R. Radach & H. Deubel (Eds.), The Mind's Eyes: Cognitive and Applied Aspects of Eye Movements (pp. 493-516). Oxford, UK: Elsevier.

Goldberg, J. H., & Kotval, X. P. (1999). Computer interface evaluation using eye movements: methods and constructs. International Journal of Industrial Ergonomics, 24 , 631-645.

Goldberg, J. H., Stimson, M. J., Lewenstein, M., Scott, N., & Wichansky, A. M. (2002). Eye tracking in web search tasks: design implications. In Proceedings of the 2002 symposium on Eye tracking research & applications (pp. 51-58). ACM.

Halverson, T., & Hornof, A. J. (2011). A computational model of active vision for visual search in human-computer interaction. Human-Computer Interaction, 26 , 285-314.

Hanington, B. (2003). Methods in the making: A perspective on the state of human research in design. Design issues, 19 , 9-18.

Jacob, R. J., & Karn, K. S. (2003). Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. In J. Hyona, R. Radach & H. Deubel (Eds.), The Mind's Eyes: Cognitive and Applied Aspects of Eye Movements (pp. 573-605). Oxford, UK: Elsevier.

Judd, T., Durand, F., & Torralba, A. (2011). Fixations on low-resolution images. Journal of Vision, 11(4), 1-14.

Klatzky, R. L., Kober, N., Mavor, A., (1996). Safe, comfortable, attractive, and easy to use: improving the usability of home medical devices. Washington, D.C.: National Academy Press.

Kowler, E. (2011). Eye movements: The past 25years. Vision research, 51 , 1457-1483.

Larsson, L., Nystroum, M., & Stridh, M. (2014). Discrimination of fixations and smooth pursuit movements in high-speed eye-tracking data. In Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE (pp. 3797-3800). IEEE.

Law, B., Atkins, M. S., Kirkpatrick, A. E., & Lomax, A. J. (2004). Eye gaze patterns differentiate novice and experts in a virtual laparoscopic surgery training environment. In Proceedings of the 2004 symposium on Eye tracking research & applications (pp. 41-48). ACM.

Martelli, S., Zaffagnini, S., Bignozzi, S., Lopomo, N., & Marcacci, M. (2007). Description and validation of a navigation system for intra-operative evaluation of knee laxity. Computer Aided Surgery, 12 , 181-188.

Martin, J. L., Norris, B. J., Murphy, E., & Crowe, J. A. (2008). Medical device development: The challenge for ergonomics. Applied Ergonomics, 39 , 271-283.

Mello-Thoms, C., Nodine, C. F., & Kundel, H. L. (2002). What attracts the eye to the location of missed and reported breast cancers? In Proceedings of the 2002 symposium on Eye tracking research & applications (pp. 111-117). ACM.

Mezger, U., Jendrewski, C., & Bartels, M. (2013). Navigation in surgery. Langenbeck's Archives of Surgery, 398 , 501-514.

Münzer, S., Zimmer, H. D., Schwalm, M., Baus, J., & Aslan, I. (2006). Computer-assisted navigation and the acquisition of route and survey knowledge. Journal of Environmental Psychology, 26 , 300-308.

Muradore, R., Fiorini, P., Akgün, G., Erol Barkana, D., & Yantac_, A. E. (2015). Development of a cognitive robotic system for simple surgical tasks. International Journal of Advanced Robotic Systems, 12, 37 .

Nielsen, J., & Molich, R. (1990). Heuristic evaluation of user interfaces. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 249-256). ACM.

Nodine, C. F., Kundel, H.L., Lauver, S.C., & Toto, L.C. (1996). Nature of expertise in searching mammograms for breast masses. Academic Radiology, 3, 1000--1006.

Olsen, A., & Matos, R. (2012). Identifying parameter values for an I-VT fixation filter suitable for handling data sampled with various sampling frequencies. Proceedings of the Symposium on Eye Tracking Research and Applications, ETRA ‘12, 317-320

Onat, S., Açık, A., Schumann, F., & König, P. (2014). The contributions of image content and behavioral relevancy to overt attention. PloS one, 9 , e93254.

Pannasch, S., Helmert, J. R., & Velichkovsky, B. M. (2008). Eye tracking and usability research: an introduction to the special issue. Journal of Eye Movement Research, 2 , 1-4.

Parkhurst, D. J., & Niebur, E. (2003). Scene content selected by active vision. Spatial vision, 16 , 125-154.

Peterhans, M., vom Berg, A., Dagon, B., Inderbitzin, D., Baur, C., Candinas, D., & Weber, S. (2011). A navigation system for open liver surgery: design, workow and first clinical applications. The International Journal of Medical Robotics and Computer Assisted Surgery, 7 , 7-16.

Poole, A., & Ball, L. J. (2005). Eye tracking in human-computer interaction and usability research: Current status and future. In Prospects, Chapter in C. Ghaoui (Ed.): Encyclopedia of Human-Computer Interaction. Pennsylvania: Idea Group, Inc.

Privitera, M., Design, M., & Murray, D. (2009). Applied ergonomics: Determining user needs in medical device design. In Annual International Conference of the IEEE Engineering in Medicine and Biology Society (pp. 5606-5608).

Reingold, E. M., & Sheridan, H. (2011). Eye movements and visual expertise in chess and medicine. Oxford handbook on eye movements, 528-550.

Rolfs, M. (2009). Microsaccades: small steps on a long way. Vision research, 49 , 2415-2441.

Salvendy, G. (2012). Handbook of human factors and ergonomics. John Wiley & Sons.

Sawyer, D., Aziz, K., Backinger, C., Beers, E., Lowery, A., & Sykes, S. (1996). An introduction to human factors in medical devices. US Department of Health and Human Services, Public Health Service, Food and Drug Administration, Center for Devices and Radiological Health .

Schlatter, T., & Levinson, D. (2013). Visual usability: principles and practices for designing digital applications. Waltham, MA: Morgan Kaufmann.

Schulz, C., Schneider, E., Fritz, L., Vockeroth, J., Hapfelmeier, A., Wasmaier, M., Kochs, E., & Schneider, G. (2011). Eye tracking for assessment of workload: a pilot study in an anaesthesia simulator environment. British journal of anaesthesia, 106 , 44-50.

Shrivastava, S. R., Shrivastava, P. S., & Ramasamy, J. (2014). Exploring the scope of participatory ergonomics in the health care industry. J Environ Occup Sci., 3 , 196-198.

Smith, T. J., & Henderson, J. M. (2009). Facilitation of return during scene viewing. Visual Cognition, 17 , 1083-1108.

Sparks, D. L. (2002). The brainstem control of saccadic eye movements. Nature Reviews Neuroscience, 3 , 952-964.

Valsecchi, M., Gegenfurtner, K. R., & Schu□tz, A. C. (2013). Saccadic and smooth-pursuit eye movements during reading of drifting texts. Journal of vision, 13 , 8.

Vincent, C. J., & Blandford, A. (2014). The challenges of delivering validated personas for medical equipment design. Applied Ergonomics, 45 , 1097-1105.

Vincent, C. J., Li, Y., & Blandford, A. (2014). Integration of human factors and ergonomics during medical device design and development: It's all about communication. Applied Ergonomics, 45 , 413-419.

Wegner, K. (1998). Surgical navigation system and method using audio feedback. In Proceedings of the International Conference on Auditory Display. British Computer Society.

Weinger, M. B., Wiklund, M. E., & Gardner-Bonneau, D. J. (2010). Handbook of human factors in medical device design. Boca Raton, FL: Taylor & Francis Group.

Wilming, N., Betz, T., Kietzmann, T. C., & König, P. (2011). Measures and limits of models of fixation selection. PloS one, 6(9), e24038.

Wilming, N., Harst, S., Schmidt, N., & König, P. (2013). Saccadic momentum and facilitation of return saccades contribute to an optimal foraging strategy. PLoS Computational Biology, 9 , e1002871.

Wilson, M. R., McGrath, J. S., Vine, S. J., Brewer, J., Defriend, D., & Masters, R. S. (2011). Perceptual impairment and psychomotor control in virtual laparoscopic surgery. Surgical endoscopy, 25(7), 2268-2274.

Yang, X., Lee, W., Choi, Y., & You, H. (2012). Development of a user-centered virtual liver surgery planning system. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (pp. 772-776). SAGE Publications volume 56.

Yantaç A. E., Çay D., Akgün, G., & Erol-Barkana, D. (2014). Insights from user studies for the graphical user interface design of a surgical operation robot. In Proceedings of the 8th International Conference on Interfaces and Human Computer Interaction. InderScience.

Yarbus, A. L. (1967). Eye movements during perception of complex objects. Springer US.

Zheng, B., Tien, G., Atkins, S. M., Swindells, C., Tanin, H., Meneghetti, A., Qayumi, K. A., & Panton, O. N. M. (2011). Surgeon's vigilance in the operating room. The American journal of surgery, 201 , 673-677.

Highlights

- Two arrangements for a surgical interface (SI) are developed using design principles.
- Doctors & laypeople perform SI-guided simulated cryoablation while gaze is recorded.
- Saccades in cardinal directions are more likely with one arrangement.
- Spatial fixation distributions at different temporal scales support the same arrangement.
- Design and evaluation with eye movements measures produce better SIs.