POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Usability assessment a mixed reality platform for pre-operative planning in cardiac interventions.

Laurea Magistrale in Biomedical Engineering - Ingegneria Biomedica

Author: Clelia Galluzzi

Advisor: Prof. Emiliano Votta

Co-advisor: Ing. Omar Pappalardo, PhD

Academic year: 2021-2022

1. Introduction

Extended reality (XR) refers to the technologies which combine the physical world with virtual objects and environments, allowing for different degrees of user interaction with digital and real objects.
Mixed reality (MR) is a subset of XR technologies which allows the user to visualize virtual 3D objects superimposed on the real world and interact with them while maintaining physical connection with the surrounding environment. The 3D rendering is realized through a powerful workstation which processes digital images and the generated content is visible to the user by wearing a semi-transparent head-mounted display (HMD), a headset equipped with several technologies including screens, sensors (e.g., infrared cameras and gyroscopes), and a central processing unit (CPU).
MR is being introduced in various fields ranging from automotive industry to clinics, typically because it allows for examining in detail complex systems in a more intuitive way as compared to the standard visual inspection through the rendering of the system on a physical 2D screen. In clinics, MR also offers the possibility to visualize complex anatomies capturing their real 3D shape without relying solely on 2D clinical images that can be sometimes difficult to interpret or even deceiving. This feature makes MR an attractive technology to support pre-operative planning and intra-operative image-based guidance in procedures that rely on clinical imaging. In pre-operative planning, the possibility to navigate not only clinical images, but also co-registered 3D anatomical reconstructions allows for better understanding anatomical relationships, and to visualize and measure anatomical regions As a result, MR can reduce the mental workload for physicians to reconstruct the 3D features of the patient's anatomy based on 2D views, and it can improve planning accuracy and time-efficiency.
In the intra-operative phase, MR potentially allows for co-registering 3D virtual models onto the surgical field in open procedures and for guiding transcatheter interventions, thus providing guide for, e.g, puncturing the access vessels.
This study evaluates the usability of a mixed reality (MR) platform designed to support the pre-operative planning of cardiovascular percutaneous procedures, by comparing it with a traditional DICOM viewer software currently used for the same purpose. The comparison was carried

out by asking sixteen clinicians, with a sound background in cardiovascular anatomies and different levels of experience in the use of DICOM viewers, to navigate pre-operative images of a human heart and to perform three tasks with the two technologies. Their performance was evaluated and they graded the various aspects of the two technologies based on their user experience through three validated questionnaires and one ad hoc developed questionnaire. Also, they were asked for their opinion on the features the technology should have to provide a relevant added value in pre-operative planning.

2. Background

2.1. Usability

The concept of usability refers to the capability of a product to be correctly used by specified users in a specified context of use, and to the extent of effectiveness, efficiency, and satisfaction with which specific goals are achieved through the device.
Referring to medical devices, the standards IEC 62366-1 [1] and the technical report IEC TR 62366-2 [2] are the reference guideline to implement the "usability engineering process", i.e., the process the manufacturer should follow to design and develop a usable device. The process is coarsely divided into two phases: i) formative evaluation, to be performed iteratively during device development in order to identify product strengths and shortcomings, user needs, and opportunities for improvement; ii) summative evaluation, which instead is performed at the end of the development cycle to validate the safety of the user interface of the device, thus ensuring that representative users can interact with it without incurring dangerous use errors. Data from this process are collected in the "usability engineering file" and should allow the manufacturer to conclude that no further user interface improvement is needed or applicable. The new Medical Device Regulation (MDR) 2017/745 [3] expects, when applicable, the mitigation of risks derived from human errors thus implicitly requiring usability evaluation. Questionnaires are often administered during usability studies since they are a cheap and time-saving way of collecting information from many participants. Some validated questionnaires commonly used in usability studies are presented in the following paragraphs.

Surgery Task Load Index - The Surgery Task Load index (S-TLX) is a tool used to evaluate the impact of different sources of stress on the cognitive workload of healthcare operators and to compute a total score representing the entity of the workload perceived in performing a task (the higher the score, the higher the workload and stress perceived). In particular, six dimensions of workload are considered in the questionnaire: mental demand; physical demand; temporal demand; task complexity; situational stress; distractions. The first part of the questionnaire is dedicated to computing the sources of load (i.e., the weight of each dimension): the six workload sources are combined two by two into fifteen couples; for each couple, the user must select the item that contributed the most to the workload perceived during a task. The number of selections for each source (N_i) is obtained. In the second part of the questionnaire, the magnitude of each sources of workload M_i is computed on a scale ranging from 0 to 20, with unitary resolution. Finally, to each source, a stressor S_i is assigned, where $S_i = N_i M_i$. The overall workload score can be computed as $\sum_{n=1}^{6} S_i = N_i M_i$. [10]

User Experience Questionnaire - The User Experience Questionnaire (UEQ) is a validated tool designed to provide a quantitative measure of the user experience and usability of interactive products. It is composed of 26 items which are represented by two opposite adjectives divided by a 7-point scale and participants have to indicate the term that better describes the product under evaluation by placing an "X" on the point scale. The closer the symbol is to one of the two words, the better that term applies to the product in the respondent's opinion. The items can be grouped into 6 evaluation dimensions: Attractiveness; Perspicuity; Efficiency; Dependability; Stimulation; Novelty. [11]

System Usability Scale - The System Usability Scale (SUS) is a very simple and fast questionnaire composed of 10 statements designed for the assessment of the usability of a product. The statements cover different aspects

of a system's usability, such as complexity, need for training, confidence, and willingness to use it. In the end, a total usability score is computed and the higher it is, the greater is the usability perceived by the user with respect to the product under evaluation. [4]

2.2. The technology under evaluation

This study focuses on a MR platform developed by Artiness, a startup founded in 2018 by researchers and professors from the Bioengineering department of Politecnico di Milano which develops MR platforms for pre-operatory planning and intra-operatory support in the context of interventional, structural, and vascular cardiology. This study focuses on ARTICOR, a MR platform allowing the holographic visualization of patient-specific medical data derived from TC images. It aims to facilitate surgical planning and device sizing in the context of transcatheter cardiovascular interventions through the 3D rendering of the patient's anatomy, and consequently simplify clinicians' work and improve the clinical outcomes for the patient. The application is currently developed in Unity and runs on HoloLens 2, the HMD developed by Microsoft (Figure 2). Medical images in DICOM format are uploaded on a workstation dedicated to data processing: images are segmented through proprietary algorithms and a 3D model of the relevant anatomy is generated. The models created in this way are then stored on secure cloud platforms and sent, in wireless modality, to a standalone HMD (HoloLens 2), through which it is possible to interact with both the hologram and the original images without the need for additional hardware. The technology allows the user to navigate the reconstructed anatomical model and to virtually position implantable devices in the respective implantation sites.

The study herein presented focused on a specific module of the R&D version of ARTICOR, namely the one dedicated to medical imaging navigation, with the purpose of evaluating its usability: due to the novelty of the technology, in fact, the number of usability studies and related specific standards concerning clinical applications is still very limited. Thus, the acquisition of further usability data is of fundamental importance for the definition of related specific standards.

3. State of the Art

Focusing on the medical applications of XR, some examples from the literature describing the usability engineering process are reported.

In the context of electrophysiology, the Enhanced Electrophysiology Visualization and Interaction System (ELVIS) is a MR platform that provides electrophysiologists with the possibility to visualize electroanatomic maps in 3D through a HMD. Silva et al. [8] conducted formative evaluations of the system during the design and development phase, as required by IEC 62366-1, to provide feedback about the preferred method of interaction, menu legibility, and potential use errors. In addition, the final version of the platform was tested through an in-human study in which 3 physicians were asked to perform some tasks on electroanatomic mapping images of 16 patients using both a standard mapping system and ELVIS, and to answer 7 questions about the usability of the system. The questionnaires reported positive scores about comfort, ease of use, tools accessibility, and improved capability in interpreting the information obtained.

Glas et al. [6] investigated the use of a MR visualization platform for image-guided surgery (IGS) composed by a surgical navigation system (Brainlab) connected to Microsoft HoloLens. The performance and usability of the MR platform against a traditional navigation interface were evaluated by asking 12 participants to search for 3 physical landmarks and 3 trajectories, and by measuring the associated time-expense and accuracy in reaching the target. Measurements were in favour of the MR technology. After completing the tasks, participants were also asked to fill out a questionnaire to rate the usability of the system and all the participants reported a positive experience.

The feasibility of 3D MR holograms for diagnostic and morphological preoperative analysis was investigated by Brun H. et al. [5] The 3D model of the heart of a pediatric patient with a rare congenital heart defect was reconstructed from cardiac computed tomography angiogram; 36 members of the heart team visualized the hologram through HoloLens and had to recognize some anatomical landmarks to perform a diagnosis. Subsequently, they filled out a ques-

tionnaire concerning anatomy identification, diagnostic output, 3D experience of the model, and quality of the interactions with the hologram. All the participants were able to identify the selected landmarks, all but two performed the correct diagnosis, and all the ratings were close to the maximum.
The usability of a medical device software intended to assist intra-operative planning was investigated by Sternini et al. [9]. This system allows for visualizing the 3D reconstruction of a patient's anatomy and for interacting with it through a touchless user interface based on Leap Motion sensors, which are able to detect and track the hands of the user. The research group implemented the whole usability engineering process through both summative and formative evaluation. In the former, designers and usability experts defined the primary operating functions and the position of the sensor and of the screen to guarantee ergonomics for the user. Subsequently, focus groups with real users were carried out to test the outcomes of the previous stage and to identify possible additional issues. The formative evaluation led to modifying the position of the sensor and of the screen, and to introduce a tutorial section. Summative evaluation was intended to confirm the usability of the final version of the device, thus users were involved in simulations of the real use. They performed some defined tasks and filled in questionnaires concerning the usability of the platform, among which the UEQ. A decreasing rate of use errors from the first to the last task was observed, representing a steep learning curve, and usability was rated with high scores.
In the study by S. Moosburner et al., [7] a modified version of the SUS questionnaire was exploited to compare the usability of two HMDs: Microsoft HoloLens and Meta 2. To the 10 original items about usability, 5 more questions about ergonomics, uncomfortable sensations, visual clarity, field of view, and gesture control were added. 15 medical students were asked to fill out the questionnaire after interacting with both devices and visualizing a 3D model of a liver created from a CT scan. Even if the HoloLens's field of view is smaller, participants particularly appreciated the improved mobility offered by its wireless and stand-alone functioning. Thus, this study suggested HoloLens's superiority as a usable device in surgical settings.

4. Materials and methods

This comparative study has been carried out in collaboration with Artiness and with IRCCS Fondazione Cà Granda Ospedale Maggiore Policlinico di Milano.
The usability test was focused on end-user interaction with medical images and 3D anatomical reconstructions through the ARTICOR platform in the pre-procedural phase. The comparator consisted a traditional DICOM viewer software, namely RadiAnt. The two technologies were compared in terms of general experience, usability, task complexity, time required, and perceived workload. The research questions were defined as follows:

1. Report of participants' comments and opinions, and descriptions of the observed use errors and difficulties during the test;
2. Comparative evaluation of the usability of ARTICOR with respect to RadiAnt;
3. Differences in the time required to fulfill a task with the two technologies;
4. Influence of the level of experience in using DICOM viewer software on usability evaluation and time performances for both technologies;
5. Absolute evaluation of the usability of ARTICOR platform.

Users were enrolled based on two inclusion criteria: i) professional background in cardiovascular anatomy; ii) general confidence in the management of CT and echocardiographic images. As a result, a total of 16 participants were enrolled (6 cardiology residents, 1 vascular resident, 1 emergency medicine resident, 1 medical student, 1 interventional cardiologist, 2 hemodynamic cardiologists, and 4 cardiac surgeons). None of them had a significant level of experience with XR technologies. Due to the different levels of expertise in using DICOM viewers instead, they were clustered into 2 groups: users with more than 5 years of experience with DICOM viewers (group 1, n=7) and users with less than 5 years of experience (group 0, n=9).
Participants were asked to perform three ad hoc defined tasks both with ARTICOR and RadiAnt. These consisted in obtaining 3 echographic-like views of the heart of a patient: 4-chambers (T1), 3-chambers LVOT (T2), and

ventricular short axis (T3) (Figure 1). Notes of participants' comments and difficulties were taken during the whole test session and the time required to perform the 3 tasks was measured. At the end of the test, the users filled out three validated questionnaires (S-TLX, UEQ, SUS) in reference to both technologies and one questionnaire concerning ARTICOR conceived ad hoc for this work. The latter was composed of 8 questions:

- The weight of the system is a problem;
- The graphic rendering obtained with a semi-transparent image is sufficient;
- The field of view is too limited;
- The data presented in 3D and the possibility of controlling the viewing angle allow for an easier data understanding than the current standard;
- Altered depth perception is a problem;
- Procedures can be simplified and workload reduced thanks to the disintermediation of information and reduction of interaction with technical personnel;
- It has often happened not to be able to grab the image or to press a button due to the altered depth perception;
- The 3D data visualization allows for learning additional anatomical notions, especially in the case of complex anatomies.

Also, the entity of different symptoms experienced by the users while interacting with the MR platform was investigated.
The last part of the study was aimed at defining some technology-specific evaluation criteria on which the evaluation of similar technologies could be based and their relative relevance. The dimensions of evaluation were derived from the literature. Subsequently, pairwise comparison was exploited to determine the weight to be associated with each criterion. The selected criteria were combined in couples and, for each couple, respondents had to select one of the two items answering the question: "between the items of each couple, which functionality of the technology do you think is more important to be guaranteed in order to boost its effectiveness and added value?". Furthermore, the pairwise comparison was performed both before and after the test, so to investigate whether the a priori ideas the users had on the technology were consistent with the ones developed after having tried it.

For each participant, the total workload score was computed from the S-TLX questionnaire and the total usability score from the SUS and the ad hoc developed questionnaire. Data from the UEQ were analyzed through pre-set Excel files and resulted in trends relative to both the 26 single items and the 6 scales. Statistical analysis was performed to test differences in S-TLX and SUS scores between the two technologies and between the two groups of users. Considering all the participants, the S-TLX and the SUS scores related to the two technologies were compared via paired-sample sign test in the form `SUS_ARTICOR` vs `SUS_RadiAnt`; `STLX_ARTICOR` vs `STLX_RadiAnt`. Differences between the two groups of users were analyzed instead through the Mann-Whitney U test, which was applied to the following two dependent variables: **V1** = `SUS_ARTICOR - SUS_RadiAnt` , **V2** = `STLX_ARTICOR - STLX_RadiAnt`. Furthermore, the same statistical analysis was applied also to the absolute value of the two variables. In all cases, the significance level was set to 0.05.

5. Results and conclusions

When using ARTICOR, every participant managed to complete the tasks; the most relevant problem was related to altered depth perception which, in some cases, caused the failure in grabbing or moving virtual objects. The interaction with Radiant was problematic for users with limited experience in using it, and some of them did not manage to fulfill the experimental tasks. Newbies to DICOM viewer software took a shorter time (less than 1 minute per task) to perform the tasks with ARTICOR, requiring instead more than 1 minute to perform each same task with RadiAnt. Experts in using DICOM viewer software resulted in being faster in performing the tasks with RadiAnt, even though the average time required to fulfill each of them using ARTICOR was below 1 minute as well.
Considering the whole group of participants, the paired-sample sign test did not find statistical differences between the S-TLX and SUS scores associated with the two technologies. The comparison of the differential S-TLX and SUS scores between the two groups of participants yielded instead statistically significant differences: group 0 rated ARTICOR with lower S-TLX scores and higher SUS scores with re-

spect to the comparator, while group 1 perceived RadiAnt as less stressful and more usable than ARTICOR ($p<0.001$ and $p= 0.006$ respectively). Even the absolute value of the considered variables reported statistical differences between the two groups ($p=0.021$ and $p=0.024$, respectively). On average, the user experience, analyzed through the validated UEQ was rated positive for both technologies. ARTICOR, however, reported higher user experience scores with respect to RadiAnt and, when compared to a benchmark dataset, was ranked in the range of 10% best results across almost all the evaluation dimensions.

These results suggest that the level of confidence in using the gold standard tool for pre-operative planning could be very relevant to the acceptance of the new technology. Operators with sufficient experience with DICOM viewer software may be less prone to perceive the advantages of the MR platform and to use it instead of the consolidated solution. On the contrary, participants without notable experience in using DICOM viewer software nor in using MR platforms perceived ARTICOR as significantly more intuitive, informative and usable.

The ad hoc developed questionnaire referring to ARTICOR reported participants' general agreement with the positive sentences (i.e., the ones expressing strengths of the system) and general disagreement with the negative ones (those reporting weaknesses of the system). None of the considered symptoms was reported as annoying by the participants. Furthermore, even without previous knowledge and with a very short period of interaction with the MR technology, the participants correctly understood the application contexts in which it might be more efficiently exploited such as mini-invasive surgery, surgical procedures planning and simulation, percutaneous vascular interventional surgery, and didactics.

In the end, the attempt to define a methodology to be applied in the assessment of MR technologies resulted in the selection of 6 evaluation criteria: reduced weight and ergonomics, field of view width, good depth perception, rendering quality, workflow simplification, and simplicity and immediacy of use. Among them, low weight and ergonomics were rated as the least important features by participants while the dimensions relating to depth perception and rendering quality as the most relevant. Also, the opinion of the users changed after the performance of the test, leading to the conclusion that the initial ideas and expectations they had about the technology did not necessarily correspond to the needs and priorities perceived when using it.

However, it's worth stressing the short duration of the interaction of the users with the technology, which could have mitigated some symptoms, the small sample size, which could have affected the statistical significance of the results, and the simplicity of the tasks, which could have influenced participants' opinion concerning the MR technology. Even if usability is a very immediate feature to be perceived and it does not require complex tasks to be evaluated, future studies should address the same problem by increasing the sample size, the interaction time, and the tasks' complexity.

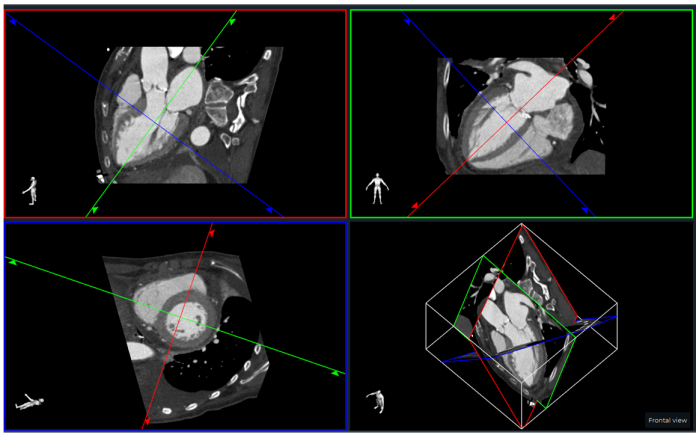

Figure 1: *Echographic-like views of the heart: 3-chambers LVOT (top left), 4-chambers (top right), ventricular short axis (bottom left).*

Figure 2: *Microsoft HoloLens 2 headset (side view).*

References

[1] *IEC 62366-1:2015 Medical devices — Part 1: Application of usability engineering to medical devices*, 1° edition, February 2015.

[2] *IEC/TR 62366-2:2016 Medical devices — Part 2: Guidance on the application of usability engineering to medical devices*, 1° edition, April 2016.

[3] Regulation 2017/745 on medical devices. *European Parliament and Council of the European Union, Official Journal of the European Union*, April 2017.

[4] John Brooke. SUS - a quick and dirty usability scale.

[5] Brun et al. Mixed reality holograms for heart surgery planning: first user experience in congenital heart disease. *European Heart Journal - Cardiovascular Imaging*, 2019.

[6] Glas et al. Augmented reality visualization for image-guided surgery: A validation study using a three-dimensional printed phantom. *Journal of Oral and Maxillofacial Surgery*, September 2021.

[7] Moosburner et al. Real world usability analysis of two augmented reality headsets in visceral surgery. *Artificial Organs*, 2019.

[8] Silva et al. Development and human factors considerations for extended reality applications in medicine: The enhanced electrophysiology visualization and interaction system (ĒLVIS). *Lecture Notes in Computer Science*, 2020.

[9] Sternini et al. Usability assessment of an intraoperative planning software. *HEALTHINF 2021 - 14th International Conference on Health Informatic*, 2021.

[10] Wilson et al. Development and validation of a surgical workload measure: the surgery task load index (SURG-TLX). *World J Surg.*, 2011.

[11] Martin Schrepp. User Experience Questionnaire handbook, 2019.