Exploring the Suitability of Using Virtual Reality and Augmented Reality for Anatomy Training

Ramiro Serrano-Vergel, Pedro Morillo, Sergio Casas-Yrurzum, and Carolina Cruz-Neira, *Senior Member, IEEE*

***Abstract*—Research on alternative ways to provide anatomy learning and training has increased over the past few years, especially since the COVID-19 pandemic. Virtual reality (VR) and augmented reality (AR) represent two promising alternatives in this regard. For this reason, in this work, we analyze the suitability of applying VR and AR for anatomy training, comparing an optical-based AR setup and a semi-immersive setup based on a VR table, using the same anatomy training software and the same interaction system. The AR-based setup uses a Magic Leap One, whereas the VR table is configured through the use of stereoscopic TV displays and a motion-capture system. This experiment builds on a previous one (Vergel et al., 2020) on which we have improved the AR-based setup and increased the complexity of one of the two tasks. The goal of this new experiment is to confirm whether the changes made in the setups modify the previous conclusions. Our hypothesis is that the improved AR-based setup will be more suitable, for anatomy training, than the VR-based setup. For this reason, we conducted an experimental research with 45 participants, comparing the use of an anatomy training software. Objective and subjective data were collected. The results show that the AR-based setup is the preferred choice. The differences in measurable performance were small but also favorable to the AR setup. In addition, participants provided better subjective ratings for the AR-based setup, confirming our initial hypothesis. Nevertheless, both setups offer a similar overall performance and provide excellent results in the subjective measures, with both systems approaching the highest possible values.**

***Index Terms*—Anatomy, augmented reality (AR), comparative study, magic leap one, training, virtual reality (VR), VR table.**

I. Introduction and Related Work

THE technologies of virtual reality (VR) and augmented reality (AR) are increasingly used in medicine. Both are based on the use of virtual interactive three-dimensional (3-D) elements and both provide a means to explore the human body in ways that would be impossible without them. They can also be supplied with real 3-D data coming from *medical diagnostic images* [2], [3], such as *computed tomography scan*, *positron emission tomography*, or *magnetic resonance imaging*. They can also work with surgical robots [4] or be used to simulate them [5]. Thus, the possibilities for using these technologies in the practice of customized medicine are almost endless. However, VR and AR are different technologies and provide different, sometimes complementary, advantages and features. Therefore, it is important to understand their differences and what is the best way in which they can convey their potential benefits to each medical area.

VR could be defined as the technology by which one or several individuals experience the sensation of belonging to an alternative reality different to the one they are really experiencing [6] (i.e., they experience only virtual content). Mixed reality (MR), on the contrary, is a technology by which it is possible to experience virtual interactive content appropriately blended with real content (typically through a view of this real content). When the amount of virtual content is smaller than the amount of real content, it is often called AR [7].

The application of VR and AR to medicine is a subject undergoing intense study with many recent works in this particular field. In this regard, the use of the HoloLens AR device is particularly common [8], [9], [10], [11], [12], [13], as this device was the first wireless holographic computer allowing hands-free AR medical applications, something that is important in medicine. Other devices, such as the magic leap one, have not been as studied, given its novelty, with just a few works reporting its use in the healthcare sector [14], [15], [16], [17]. VR, a much more mature technology, has also been used in rehabilitation [18], pain therapy [19], and psychological or psychiatric disorder treatment [20], to name a few areas.

Although the surgical use seems to be the dominant application in medicine—especially in the case of AR—several works have proposed the use of VR and AR also for anatomy training [21], [22], [23], [24], [25], [26] since the traditional way in which anatomy training is addressed—using real cadavers—usually entails many problems.

1) Cadavers are scarce [27] (at least those suitable for medical study) and thus expensive.
2) They need to be properly maintained. Otherwise, they deteriorate quickly becoming black or deformed [28]. In

Manuscript received 10 June 2022; revised 28 October 2022; accepted 12 December 2022. Date of publication 25 January 2023; date of current version 15 March 2023. This work was supported in part by Spain's Agencia Estatal de Investigación and in part by the Spanish Ministry of Science and Innovation, under the ViPRAS Project (Virtual Planning for Robotic-Assisted Surgery) under Grant PID2020-114562RA-I00 (funded by MCIN/AEI/10.13039/501100011033). This article was recommended by Associate Editor M. S. Neubert. *(Corresponding author: Sergio Casas-Yrurzum.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by UALR Institutional Review Board under Application No. 19-164-R1, and performed in line with the Code of Federal Regulations.

Ramiro Serrano-Vergel is with the Emerging Analytics Center, University of Arkansas at Little Rock, Little Rock, AR 72204 USA (e-mail: rxserranove@ualr.edu).

Pedro Morillo and Sergio Casas-Yrurzum are with the Institute on Robotics and Information and Communication Technologies, University of Valencia, 46010 Valencia, Spain (e-mail: pedro.morillo@uv.es; sergio.casas@uv.es).

Carolina Cruz-Neira is with the Department of Computer Science, University of Central Florida, Orlando, FL 32816 USA (e-mail: carolina@ucf.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/THMS.2023.3235250.

Digital Object Identifier 10.1109/THMS.2023.3235250

2168-2291 © 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Universidad de Valencia. Downloaded on May 19,2023 at 08:07:25 UTC from IEEE Xplore. Restrictions apply.

fact, this will eventually or gradually occur even if they are properly maintained.
3) Dead tissues do not behave and look like living tissues.
4) Anatomy training with real cadavers is slow since usually many students need to share a single corpse, given the shortage of cadavers. In addition, faculty staff need to coordinate the groups in order to maximize the availability of the cadaver.
5) The dissection of real cadavers raises ethical questions in many cultures.
6) The chemicals that need to be used to preserve them pose health risks. For instance, formaldehyde is classified by the International Agency for Research on Cancer as carcinogenic [29]. In addition, these chemicals make body tissues stiff, making them difficult for students to handle.

On the contrary, VR and AR are becoming more and more available. They also have the ability to portrait the anatomical structure in 3-D space, which is an improvement with respect to textbooks and 2-D written material. They can also be used to see through a real body (AR) without damaging it or to reproduce/simulate a particular anatomical deformation (VR and AR) at very little additional cost. These key advantages make them potentially helpful for anatomy training.

Although different studies on the use of VR and AR in the anatomy field have reached different conclusions, they seem to agree that these technologies can be successfully used for anatomy training and learning. In this regard, some works even claim that these technologies should substitute traditional approaches (those based on textbooks, lectures, plastic models, or even real cadavers) [23]. However, other authors declare that VR and MR technologies are inferior to physical models, whereas some other authors are more cautious [30] and advocate that they used to complement, rather than replace, traditional methods [24], [31]. The rationale behind this is that, among other problems, virtual cadavers cannot yet provide proper natural haptic feedback and students need to explore the textures of real tissues. In any case, although it is likely that real cadavers will continue to be used in medical schools, it is important to explore other ways for anatomy training, given the problems that cadavers pose.

AR focuses on the "here-and-now," while VR orbits around the concept of presence and the idea of "being somewhere else." This makes AR more suitable for experiential learning and VR more suitable for subjective experiences [32] and simulation. This can be translated to clinical uses as well. AR, for instance, is a promising tool for the treatment of phobias, while VRs ability to modify the sensation of presence makes it ideal for the treatment of pain, for instance [32]. In other words, VR has the potential to be a transformative technology although it may be sometimes perceived as unreal, while AR is a more context-aware technology, although it can be less transformative. Nevertheless, both technologies can be used in similar contexts.

In the case of anatomy training, AR increases the tangibility of the application, making what is experienced seems real, while VR can evoke greater subjectivity, which can help achieve greater engagement in the learning process. In addition, unlike VR, most current AR technologies are not yet able to provide wide fields of view (FoV) for the virtual content, potentially reducing immersion and increasing the visual information density [33]. Thus, it is important to research about the suitability of using these two technologies for anatomy training.

For this reason, in a recently published paper [1], we preliminarily analyzed the suitability of applying VR and AR for anatomy training. For that purpose, we compared an optical-based AR setup, implemented with a Microsoft HoloLens device, and a semi-immersive setup based on a VR table, using the same anatomy training software application. The user interface (UI) of these two setups was substantially different, whereas the AR-based setup was built upon the *metaphoric* [34], [35], [36] hand-tracking capabilities of the HoloLens, the UI of the VR-based setup used a handheld controller. We chose the HoloLens because it is one of the preferred devices in the medical field [37], whereas the VR table was chosen because it follows the *forensic table metaphor*, which seems a natural way to display a virtual cadaver. By forensic table metaphor, we mean that the trainee works with a device that is arranged horizontally as a forensic table. A total of 82 people participated in the experiment by completing two anatomy-related tasks with the aforementioned setups. From the results of that study, we concluded that the VR-based setup was significantly more suitable for anatomy training than the AR-based setup. However, these results raised also some questions since most of the participants complained about the limited FoV of the HoloLens device and about the UI of the AR-based setup. The former question suggested that we look for a newer device with improved display capabilities, such as the magic leap one. Regarding the latter, we faced a dilemma: should we change the UI of the AR-based setup to mimic the UI of the VR table or should we instead try to improve the natural hand interaction of the AR setup by using a more realistic *isomorphic* [34], [35], [36] approach?

For this reason, we conducted a second experiment, published in [36], designed to analyze the differences in natural hand interaction between a HoloLens device and a magic leap one device for two pick-and-place and drag-and-drop tasks (unrelated to anatomy). The goal of this experiment was to understand the differences (in general accuracy and performance) between the metaphoric hand-gesture-tracking paradigm represented by the HoloLens and the isomorphic paradigm represented by the Magic Leap One. We found only very small differences in the use of hand interaction between these two devices, although the Magic Leap One did allow a faster completion of one of the tasks. Despite this is a substantially different research than the one we present here, this second experiment helped us decide about the dilemma of how to modify the UI of the AR-based setup presented in [1].

With all this information at hand, we concluded the following points.

1) There was a substantial improvement potential in the AR-based setup based on the comments of the participants of [1].
2) Some of the differences found in [1] were caused by the use of the metaphoric gesture-based interaction system

Authorized licensed use limited to: Universidad de Valencia. Downloaded on May 19,2023 at 08:07:25 UTC from IEEE Xplore. Restrictions apply.

used in the AR-based setup, but most likely will not be solved using an isomorphic approach.

3) A more recent AR device, such as the Magic Leap One, could provide some additional value because it also provides a larger FoV.

Thus, we decided to perform a third experiment, which we present in this article, to compare, in the use of anatomy training, a VR-based setup similar to the one, as shown in [1], to an improved AR-based setup. In the latter setup, among other changes, the HoloLens is substituted by a Magic Leap One and the UI is changed so that it now uses a handheld controller instead of using any type of natural hand interaction.

Since the research leading to the publication of the article presented in [1], where we also reviewed the state-of-the-art of this matter, the world has suffered, and still does, the consequences of the worst pandemic in a century, the COVID-19 pandemic. As a result, funding programs and calls for research in medical areas have multiplied. Another important consequence is that online learning, based on computer-supported technologies, has been a necessity in many universities. Anatomy learning has been no exception and alternatives to face-to-face lessons have been explored [38], [39], [40], [41]. This is reflected in the academic literature, where dozens of research articles have been published, just in one year, about the use of virtual environments in the anatomy field. Many researchers have published review works or meta-analysis about this issue in recent months [42], [43], [44], [45], [46], [47], [48], [49], [50], reflecting the huge interest in the topic. The general conclusion is that VR and AR are viable alternatives and increasingly useful for teaching anatomy, given the latest advances in display and interaction technology in these areas. Other authors focus on presenting 3-D-/AR-/VR-/MR-based methods for improving anatomy understanding [51], [52], [53], [54]. Other authors analyze the conditions upon which these technologies are useful for anatomy lessons [55], [56], [57] or their effectiveness [58], [59], [60].

Our approach is different since we do not intend to study the learning outcomes provided by these technologies. Our research is currently focused on understanding what is the best way to utilize and configure these technologies for anatomy training. To the best of our knowledge, there are not research articles performing the kind of analysis that we present here.

Given our previous results and the amount of interest in the question, we analyze in this article, in line with what we stated in the future work section of the article presented in [1], if the setup based on the VR table still holds an advantage—for anatomy training with virtual cadavers—over the improved optical AR-based setup. Our hypothesis is that these improvements will make the AR-based setup more suitable for anatomy training than the VR table in terms of these three dimensions: subjective perception; objective performance; and explicit two-choice recommendation. The improvements are described throughout the article.

The rest of this article is organized as follows. Section II describes the materials and methods utilized to perform the experiments. Section III details this new experimental study. In Section IV, the results of the experiments are presented and discussed. Finally, Section V concludes this article.

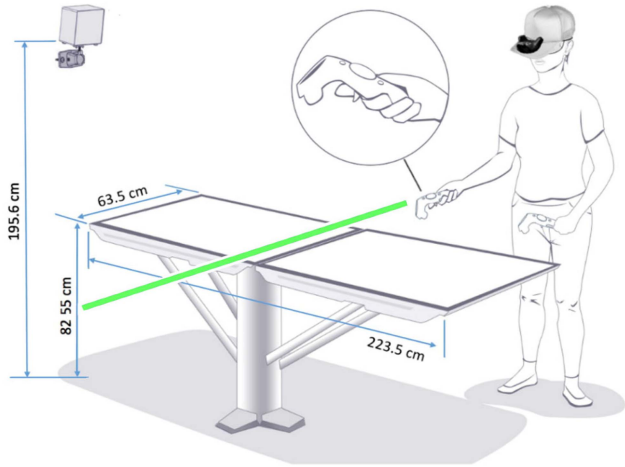

Fig. 1. Schema of the VR-based setup, showing two 3-D TVs, the HTC VIVE-tracking system (base station + tracker), and the VIVE controller.

Fig. 2. User practicing with the VR-based setup. The image is enhanced with a virtual light ray that is not visible from the position of the external observer but is visible from the position of user.

II. Materials and Methods

Two different hardware setups were used to perform this research and evaluate the anatomy training application. The first one, which we will call *Setup A*, was based on a 3-D TV with stereoscopy. The second one, which we will call *Setup B*, used a Magic Leap One as its main device.

Setup A implemented a semi-immersive VR paradigm. It used two *Sony Bravia* 50" stereoscopic TVs and an HTC Vive-tracking system to recognize the user's position and the interaction events. The tracking system was composed of a base station and a tracker, which was attached to a cap (see Figs. 1 and 2). The two TVs were arranged side-to-side and placed with the display on the horizontal plane facing up so that the setup provided a forensic table metaphor. This setup was described in [1] and the hardware did not change with respect to that. There were only some software changes that will be described later.

Authorized licensed use limited to: Universidad de Valencia. Downloaded on May 19,2023 at 08:07:25 UTC from IEEE Xplore. Restrictions apply.

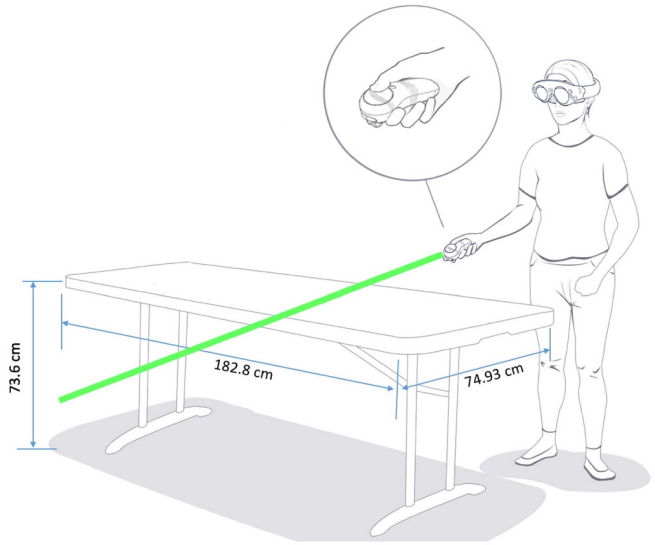

Fig. 3. Schema of the AR-based setup, showing a table—on top of which the virtual cadaver is placed—and a user wearing a Magic Leap One device handling the 6-DoF controller.

Fig. 4. User practicing with the AR-based setup. The image is enhanced with virtual elements that are not visible from the position of the external observer but are visible from the position of user.

Setup B was a semi-immersive AR-based setup. This new setup substituted the HoloLens-based setup used in [1] and it was built using a Magic Leap One AR device (see Figs. 3 and 4). This is a three-piece system. It includes a headset called *lightwear*, a small wearable computer called the *lightpack*, where the software is stored and run, and a handheld controller with six degrees of freedom (DoF). The headset is a state-of-the-art see-through head-mounted display (HMD) with head-tracking and inward-facing eye-tracking cameras.

The Magic Leap One display technology is based on the *virtual retinal display* technology, which draws a raster display directly onto the retina of the eye. It also provides pupil-tracking technology, although we did not use this feature in our experiments. The device provides an FoV of 40° × 30°, a display with

TABLE I
AGE DISTRIBUTION OF THE PARTICIPANTS OF THE EXPERIMENT

Age	*< 20*	*20–29*	*30–39*	*40–49*	*> 50*	*Total*
Group A	*4*	*8*	*6*	*3*	*2*	*23*
Group B	*1*	*12*	*4*	*4*	*1*	*22*
Total	*5*	*20*	*10*	*7*	*3*	*45*
%	*11.1*	*44.4*	*22.2*	*15.5*	*6.7*	*100.0*

1280 × 960 pixels per eye, and a two hand-gesture isomorphic recognition system, including finger tracking (with three joints per finger), which we also did not use in the experiments, for reasons explained earlier.

There are two main differences between the old and the new AR-based setups. First, the new setup substitutes the metaphoric gesture-based interaction (that the HoloLens provided) used in [1] by a controller-based laser–pointer interaction metaphor (i.e., the controller is used as a laser–pointer so that the objects are picked using ray casting [61]). We wanted to avoid the use of gestures for the selection and movement of objects in the scene, since from the results of previous experiments, we have concluded that the use of gestures involves a cognitive load that challenges the performance and preference of users. Instead, we propose to use the 6-DoF Magic Leap handheld controller in a similar way that the Vive Controller is used in setup A. Using this solution, we also remove one possible source of statistical variation and focus on the visualization and conceptualization paradigm.

The other main problem previously identified for the AR-based setup was the FoV of the HoloLens. For this reason, we have based this new setup on the Magic Leap One device, which provides an improved FoV over the HoloLens. The Magic Leap's FoV, although still limited, represents an important improvement with respect to the AR setup in [1], which offered an FoV of about 30 × 17 degrees and a resolution of 1268 × 720 pixels per eye.

Regarding the software application, the same software developed in [1] was used in these new experiments with some changes. The software is a unity 3-D-based application in which a virtual cadaver is depicted on top of a forensic table. This table is virtual in the case of the VR table but real in the case of the AR-based setup. Some of the elements of the cadaver are labeled and can be individually selected and moved from/to a forensic tray so that the participants can identify, locate, and move different parts of the human body.

Regarding the changes made with respect to the article presented in [1], first, the *InputManager* module has been changed and now the AR-based setup uses the same type of interaction technique used in the VR-based setup. The interaction is based on a laser–pointer paradigm. This way, the controller casts a laser-like green light ray (see Figs. 1–4) and whenever the laser passes through the objects, the selection of the desired organ is performed taking into account object-to-user distance. The selected object can then be grabbed by pressing the trigger button. Once grabbed, it can be moved by moving the controller and released by releasing the trigger button. This software improvement has been applied to both setups (A and B). Second,

Authorized licensed use limited to: Universidad de Valencia. Downloaded on May 19,2023 at 08:07:25 UTC from IEEE Xplore. Restrictions apply.

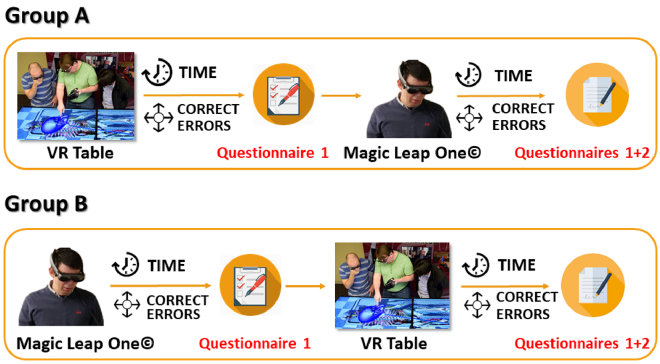

Fig. 5. Schematic view of the counterbalanced process followed in the evaluation protocol of the anatomy training application.

the *CameraController* module has been changed to adapt to the Magic Leap display. Finally, the *TaskManager* module has been updated to include more selectable objects so that the tasks can be more complex.

III. Experimental Study

A. Experimental Design

An experimental design similar to the one used in [1] was employed in this new research. We recruited students and physicians who did not have any previous experience using AR or VR technologies (for anatomy training) and who did not previously participate in the experimental research shown in [1]. We ran these new experiments in the Emerging Analytics Center of the University of Arkansas at Little Rock. However, due to the limitations imposed by the COVID-19 pandemic, the number of subjects was reduced to 45 people.

Of the 45 participants, 19 were women (42.22%) and 26 men (57.78%), with ages ranging from 18 to 73 (mean 30.62 $\pm$ 12.45). The age distribution is shown in Table I.

The goal of the experiment was to identify differences—with respect to the three dimensions stated in the research hypothesis—in the use of the anatomy training application between the two setups previously described. The experiment was designed so that the users would perform the same two tasks, which will be explained later, using both setups. Keeping that in mind, users were divided into two different groups of 23 (Group A) and 22 people (Group B), respectively, as other similar works propose [62], [63], [64] so that a counterbalanced measures design could be applied. The users of Group A started the experiment with the VR table and then used the AR-based setup. Group B included the users who tried the AR-based setup first and then the VR table. With this design, it is possible to analyze the results of each group separately (performing a within-subjects' analysis for each group) or in combination (performing a between-groups comparison between those people who tested each setup first).

B. Experimental Protocol

The experimental procedure followed an eight-step protocol, which is described next and shown in Fig. 5.

Step 1—Presentation and description: Before the start of the experiment itself, users were provided with a short description (5 min) about the software, the two setups and the tasks they had to complete. They were informed about the maximum time they had to complete the experiment (40 min in total, including the questionnaires). Then, users were required to sign an informed consent and a short questionnaire to provide basic information (gender, age, and profession). Finally, they were informed that the application recorded performance data and that the experiment was completely anonymous.

Step 2—Instruction and practice: Before the start of the experiment, users received a short briefing on how to use either of the setups (depending on which setup they would have to test first). In both cases, a free practice of 5 min was carried out on three main actions: select, move/drag, and drop.

Step 3—Experiment: Once the participants were familiar with the setup, the experiment begun. As previously explained, it consisted of two different tasks. Each of the tasks had to be completed within 5 min.

Step 4—Evaluation: After users finished the two tasks using the first setup, they were prompted to complete Questionnaire 1. Table II lists the questions asked in this questionnaire, which were presented as seven-point Likert-scale questions with the usual meanings: 1 = strongly disagree, 2 = disagree, 3 = somewhat disagree, 4 = neutral, 5 = somewhat agree, 6 = agree, and 7 = strongly agree. As performed in [1], instead of analyzing the results of each question individually, the questions were grouped in six factors: sensory factors (SF), control factors (CF), distraction factors (DF), ergonomic factors (EF), realism factors (RF), and other factors (OF). These factors were adapted from the work described in [65]. Questionnaire 1 included also three questions about depth perception, usefulness, and a global score, which summarizes the overall subjective experience. These last three questions had a different meaning: 1 = poor, 2 = bad, 3 = somewhat bad, 4 = neutral, 5 = positive, 6 = good, and 7 = excellent. This questionnaire addressed the first dimension of our hypothesis (subjective perception).

Steps 5–7—Once a participant finished the experiment with the first setup, steps 2–4 were repeated using the other setup (the presentation step did not need to be repeated).

Step 8—Final comparative evaluation: When the tasks carried out using the second setup were also completed and users had completed Questionnaire 1 for both setups, they were also asked to fill Questionnaire 2, as shown in Table III, about user preference and recommendation regarding these two setups. This two-choice questionnaire included also an open-ended question "Additional comments and explanations." Although these comments (and the responses to the *Why?* questions) cannot be analyzed in an objective way, they can be very useful for the researchers in terms of qualitative information since they allow us to explore the users' impressions about the two setups. In some occasions, they can even provide insights into the ultimate reasons explaining the results obtained in the experiments. This questionnaire addressed the last dimension of our hypothesis.

Authorized licensed use limited to: Universidad de Valencia. Downloaded on May 19,2023 at 08:07:25 UTC from IEEE Xplore. Restrictions apply.

TABLE II
QUESTIONNAIRE 1

Question (factor)
The information displayed on the device was adequate. (SF – sensory factors)
It was easy to handle the VR/AR application. (CF – control factors)
The information displayed on the device was easy to read. (SF)
The information displayed on the VR/AR device was clear. (SF)
It was easy to handle the device and its accessories. (CF)
I did not have to strive to recognize the instructional elements as 3-D elements. (RF – realism factors)
The 3-D virtual elements looked like real. (RF)
The handling of the device and its accessories was simple and without complications. (CF)
The system responded to my actions adequately. (CF)
The handling of the system and its accessories was natural. (CF)
I did not feel delays between my actions with the device and the expected results. (CF)
The control mechanisms of the VR/AR (glasses, lights, surface, etc.) did not distract me. (DF – distraction factors)
I got used to the VR/AR application for medical purposes and the device. (CF)
The device and the application were easy to use. (EF – ergonomic factors)
I found very useful the information provided by the VR/AR application to complete the actions. (RF)
I had the impression that the aid elements appeared in 3-D on the device. (RF)
The VR/AR application helped me to complete the require actions on the virtual human body. (RF)
I found the instructional elements to be useful. (RF)
I had the impression that the 3-D labels, required for the task, were a part of the scene. (RF)
There were moments that I thought that the elements that appeared on the device were real. (RF)
I did not pay attention to differences between the instructional elements and the actual device. (RF)
I had the impression that I could have touched the items that appear in the VR/AR. (RF)
I liked the visual aids to help me complete the task. (RF)
I liked how virtual elements correlate with the actual device. (RF)
I have not felt any kind of discomfort during the experience (dizziness). (EF)
I have felt the sensation of going in motion with this system. (OF – other factors)
I would like to use this technology with other uses. (OF)
The use of the system and its accessories was comfortable for my legs and arms. (EF)
I liked the experience of using a VR/AR application with a virtual human body. (OF)
The use of the system did not require a great effort from the legs or arms. (EF)
The use of the system did not require a great mental effort. (SF)
I have focused on the actions I had to do and not on the system or the environment. (DF)
My arms and legs are not tired after the experiment. (DF)
I felt involved during the experience. (OF)
At the end of the experience, I was an expert in the management of the system. (CF)
Rate the feeling of 3-D (depth perception).
Rate the system as a device.
Rate the usefulness of the system as a system to train medical students, new surgeons, or help medical professionals refresh their skills.

TABLE III
QUESTIONNAIRE 2

#	*Question*
Q1	Which system did you find most useful?
Q2	Why?
Q3	Which system did you like the most?
Q4	Why
Q5	What system would you recommend to be used as a part of an anatomy training?
Q6	Why?
-	Additional comments and explanations

TABLE IV
DATASETS GENERATED—FOR EACH SETUP—IN THE EXPERIMENTS FROM THE ANSWERS TO QUESTIONNAIRE 1

Id	*Description*
SF	Sensory factors
CF	Control factors
DF	Distraction factors
EF	Ergonomics factors
RF	Realism factors
OF	Other factors
3-D	3-D depth perception
SC	Rated score of the system
US	Perceived usefulness of the system

C. Tasks, Datasets, and Statistical Procedure

As previously mentioned, two tasks were used to compare the two setups for anatomy training. Task 1 was a low-complexity task where the main objective was to locate the heart of the virtual cadaver, grab it, and place it on a forensic tray. To accomplish this, the user had to first remove the three elements covering the heart (ribs, left lung, and sternum). Task 1 corresponds to the first task described in [1].

As for Task 2, its goal was to assembly the respiratory and the digestive system by selecting, grabbing, and placing their main components from a forensic tray. The difference in complexity between Task 1 and Task 2 is justified by the need to understand whether complex tasks are more detrimental to one type of paradigm/setup or another. Task 2 was updated with respect to the article presented in [1] in order to involve more elements (a total of ten) and more than one anatomic system.

These two tasks were chosen because they involve the understanding of the 3-D structure, position, and manipulation of organs. A more detailed explanation for the selection of these two tasks can be found in [1].

As the experiment was counterbalanced, both groups (A and B) needed to complete both tasks with both setups. Therefore, each participant had to complete four tasks (two with setup A and the same two tasks with setup B).

From the performance of the users completing these two tasks and from the subjective questionnaires, as shown in Tables II and III, a series of datasets—each of them containing 45 data elements—were collected, which are described next.

From Questionnaire 1, nine numeric datasets (six coming from the subjective factors in which the responses of the first 35 questions are grouped, and three coming from the last three elements of the questionnaire), shown in Table IV, were collected for each of the two setups. Therefore, there are 18 (nine for each setup) of these datasets, each of them containing 45 elements.

Authorized licensed use limited to: Universidad de Valencia. Downloaded on May 19,2023 at 08:07:25 UTC from IEEE Xplore. Restrictions apply.

TABLE V
OBJECTIVE DATASETS—FOR EACH SETUP—GENERATED IN THE EXPERIMENT

Id	*Description*
1.T	Total time needed to complete Task 1
1.C	Number of correctly selected elements in Task 1
1.M	Number of mistakenly selected elements in Task 1
2.T	Total time needed to complete Task 2
2.C	Number of correctly selected elements in Task 2
2.M	Number of mistakenly selected elements in Task 2

From Questionnaire 2, three Boolean datasets were collected from Q1, Q3, and Q5. In addition, three textual datasets were collected from Q2, Q4, and Q6. These six datasets contain 45 elements, showing the recommendation and preferences of the 45 users between the two setups.

Finally, from the objective data collected by the software application, three objective measures were recorded for each task: total time to complete the task (T), number of correctly selected elements (C), and number of mistakenly selected elements (M). These datasets are shown in Table V. There are 12 datasets (six datasets per setup) of this type, with 45 elements per dataset. These datasets address the second dimension of our hypothesis (objective performance).

As the order of testing matters, all these datasets were classified so that it is possible to know which system was tested first for each user. All these datasets were analyzed using IBM SPSS 26. First, we checked the normality hypothesis using the Kolmogórov–Smirnov test [66] and the Shapiro–Wilk test [67]. For the sake of brevity, we will not detail these normality tests, but all the numeric datasets passed these tests. Therefore, parametric tests can be applied to the numeric datasets collected during the experiments. As the experiment was counterbalanced, we applied both paired and unpaired t-tests—in order to compare mean values between the two groups—to the numeric datasets, as shown in Tables IV and V, as well as Cohen tests and a multifactorial ANOVA. We also applied a binomial test to the questions, as shown in Table III. All the analyses were two tailed and were conducted at the 0.05 significance level.

The parametric t-tests will help understand which one of the two setups provides better subjective and objective indicators, allowing us to test the research hypothesis. The binomial tests will tell us which system is considered more useful, preferred, and recommended by the participants, reinforcing our answer to the hypothesis, whereas the ANOVA will provide further useful insights into the data.

IV. RESULTS AND DISCUSSION

In this section, we present the results of the statistical analyses performed over the datasets collected during the experiments. First, we compare the subjective numeric datasets (those from Table IV) between the VR-based and the AR-based application (for the two tasks simultaneously since these datasets were collected once per setup). To avoid carryover effects, we compare the participants who used each of the systems first (between-subjects approach). Thus, this is an unpaired t-test comparing the average value of each of the factors when Setup A or Setup B are used first to complete both tasks. Table VI presents the results of this comparison. We do this because we aim to understand whether a VR table or an AR display produced the highest ratings when they are used for the first time. We can see that, in four parameters, the average value using the AR-based setup is significantly higher than the one obtained using the VR-based setup. The rest of the parameters do not show statistical significance but the Cohen's *d*- and the *t*-value show that the AR-based setup tends to provide higher values.

TABLE VI
STUDY OF STATISTICALLY SIGNIFICANT DIFFERENCES BETWEEN THE VR AND THE AR APPLICATION—FACTORS, 3-D PERCEPTION, SCORE, AND USEFULNESS

	VR (mean ± SD)	*AR (mean ± SD)*	*t*	*p*	*Cohen's d*
SF	6.239 ± 0.918	6.500 ± 0.551	−1.149	0.257	−0.355
CF	5.799 ± 1.069	6.205 ± 0.515	−1.609	0.115	−0.512
DF	6.087 ± 0.866	6.530 ± 0.551	−2.039	0.048	−0.626
EF	6.120 ± 0.879	6.648 ± 0.516	−2.445	0.019	−0.758
RF	5.696 ± 0.869	6.023 ± 0.675	−1.406	0.167	−0.424
OF	5.978 ± 0.938	6.205 ± 0.635	−0.943	0.351	−0.288
3D	5.348 ± 1.526	6.455 ± 0.739	−3.074	0.004	−0.978
SC	5.870 ± 1.100	6.318 ± 0.646	−1.658	0.105	−0.514
US	5.217 ± 1.678	6.273 ± 0.883	−2.623	0.012	−0.824

Groups sizes = 23 (A), 22 (B). DoF = 43.

TABLE VII
STUDY OF STATISTICALLY SIGNIFICANT DIFFERENCES BETWEEN THE VR AND THE AR APPLICATION—FACTORS, 3-D PERCEPTION, SCORE, AND USEFULNESS—GROUP A

	VR (mean ± SD)	*AR (mean ± SD)*	*t*	*p*	*Cohen's d*
SF	6.239 ± 0.918	6.587 ± 0.567	−1.566	0.132	−0.468
CF	5.799 ± 1.069	6.223 ± 0.800	−1.344	0.193	−0.454
DF	6.087 ± 0.866	6.449 ± 0.736	−1.559	0.133	−0.452
EF	6.120 ± 0.879	6.467 ± 0.688	−1.594	0.125	−0.444
RF	5.696 ± 0.869	6.362 ± 0.533	−3.649	0.001	−0.951
OF	5.978 ± 0.938	6.348 ± 0.749	−1.568	0.131	−0.438
3D	5.348 ± 1.526	6.391 ± 0.941	−2.547	0.018	−0.846
SC	5.870 ± 1.100	6.348 ± 0.885	−1.525	0.141	−0.482
US	5.217 ± 1.678	5.522 ± 1.473	−0.718	0.480	−0.193

Group size = 23. DoF = 22.

We also analyze the results using a within-subjects approach. Tables VII (paired t-test for participants in Group A) and VIII (paired t-test for participants in Group B) show the results of a repeated measures test, where we compare the subjective experience of those participants who tested one system first and then the other. We do this in order to identify possible carryover effects (the setup tested in second place might provide better results because the user had already some time to practice). However, this is not what happened and the analysis shows that Setup A does not provide any favorable results. On the contrary, Setup B provides two favorable significant results (RF and 3-D) in Table VII, and all but one (SF, which in any case has a *p*-value of 0.085) in Table VIII. In both cases, the effect size is high, as reflected by the value of Cohen's *d*. This is a clear sign that the

Authorized licensed use limited to: Universidad de Valencia. Downloaded on May 19,2023 at 08:07:25 UTC from IEEE Xplore. Restrictions apply.

TABLE VIII
STUDY OF STATISTICALLY SIGNIFICANT DIFFERENCES BETWEEN THE AR AND THE VR APPLICATION—FACTORS, 3-D PERCEPTION, SCORE, AND USEFULNESS—GROUP B

	AR (mean ± SD)	*VR (mean ± SD)*	*t*	*p*	*Cohen's d*
SF	6.500 ± 0.551	6.216 ± 0.683	1.808	0.085	0.461
CF	6.205 ± 0.525	5.438 ± 1.154	3.621	0.002	0.919
DF	6.530 ± 0.551	5.788 ± 1.171	3.787	0.001	0.863
EF	6.648 ± 0.516	5.761 ± 1.130	3.757	0.001	1.077
RF	6.023 ± 0.675	5.686 ± 0.851	2.573	0.018	0.442
OF	6.205 ± 0.635	5.500 ± 1.157	3.513	0.002	0.786
3D	6.455 ± 0.739	5.409 ± 1.333	3.511	0.002	1.009
SC	6.318 ± 0.646	5.364 ± 1.364	3.375	0.003	0.949
US	6.273 ± 0.883	5.364 ± 1.590	2.306	0.031	0.735

Group size = 22. DoF = 21.

AR-based setup provides better subjective experiences for this application.

Among these subjective factors, the one revealing the highest difference is the 3-D perception, which is significantly better in all three cases (see Tables VI–VIII) for setup B, with a difference of around one point out of seven. This is probably one of the leading causes of the superiority of this AR-based setup over the VR table.

In any case, both setups offer excellent results in terms of subjective measures, with both systems approaching the highest possible value (7). This is especially true in the AR-based setup, with all values above six points—in a seven-point scale—in Tables VI–VIII, with the exception of just one (the usefulness measure in Table VII). This means that the subjective experience is good in both cases and very good in the case of the AR-based setup.

We compare the results obtained in Table VI with the results shown in Table III of the article presented in [1] offers also some interesting insights. All nine values for the AR-based setup are higher than they were in our previous work and have a smaller dispersion. For instance, the overall score (SC) has increased from 5.905 (± 1.122) to 6.318 (± 0.646), and 3-D perception has changed from 5.619 (± 1.324) to 6.455 (± 0.739), a substantial improvement. Similar improvements can be observed in the within-subjects' analyses. We compare Tables VII and VIII versus Tables V and VI of the article presented in [1].

Next, we perform a similar analysis with the objective datasets. This time we can analyze the results for each task since we have datasets for both Task 1 and Task 2. Table IX presents the results of performing an unpaired t-test (between-subjects approach) comparing the average performance, for Task 1, of the participants who used Setup A first to the average performance of participants using Setup B first. No statistically significant differences are found between the two setups. Table X presents a similar analysis for Task 2, and now a small, but statistically significant difference, is found for dataset 2.C in favor of the AR-based setup. Fig. 6 shows the box plots of the time datasets (1.T and 2.T).

We can also analyze the results per group using a within-subjects approach. Tables XI (paired t-test for participants in

TABLE IX
STUDY OF STATISTICALLY SIGNIFICANT DIFFERENCES BETWEEN THE VR AND THE AR APPLICATION—OBJECTIVE DATASETS OF TASK 1

	1.T Time	*1.C Correct*	*1.M Mistaken*
VR (mean ± SD)	34.174 ± 25.473	3.000 ± 0.000	3.043 ± 3.022
AR (mean ± SD)	32.864 ± 19.219	3.000 ± 0.000	2.955 ± 2.554
t	0.194	-	0.106
p	0.328	-	0.495
Cohen's d	0.059	-	0.032

Groups sizes = 23 (A), 22 (B). DoF = 43.

TABLE X
STUDY OF STATISTICALLY SIGNIFICANT DIFFERENCES BETWEEN THE VR AND THE AR APPLICATION—OBJECTIVE DATASETS OF TASK 2

	2.T Time	*2.C Correct*	*2.M Mistaken*
VR (mean ± SD)	157.261 ± 81.421	9.174 ± 2.309	3.130 ± 3.584
AR (mean ± SD)	124.409 ± 93.943	9.682 ± 0.780	2.909 ± 3.571
t	1.255	–0.979	0.207
p	0.609	0.026	0.504
Cohen's d	0.375	–0.329	0.062

Groups sizes = 23 (A), 22 (B). DoF = 43.

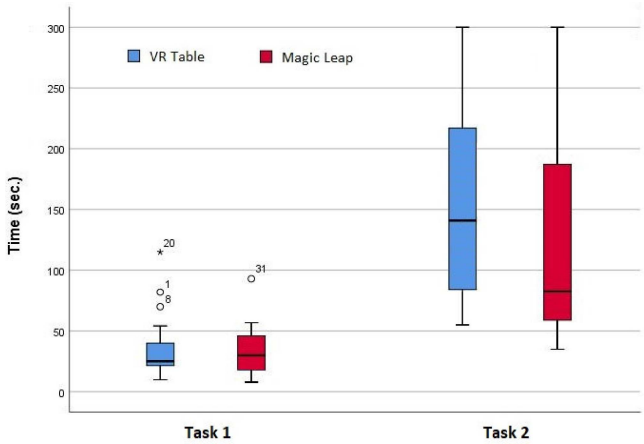

Fig. 6. Box plots (unpaired t-tests) for datasets 1.T (time, Task 1) and 2.T (time, Task 2).

TABLE XI
STUDY OF STATISTICALLY SIGNIFICANT DIFFERENCES BETWEEN THE VR AND THE AR APPLICATION—OBJECTIVE DATASETS OF TASK 1—GROUP A

	1.T Time	*1.C Correct*	*1.M Mistaken*
VR (mean ± SD)	34.174 ± 25.473	3.000 ± 0.000	3.040 ± 0.302
AR (mean ± SD)	30.739 ± 24.153	3.000 ± 0.000	0.170 ± 0.834
t	0.800	-	4.751
p	0.432	-	$<10^{-3}$
Cohen's d	0.138	-	5.052

Group size = 23. DoF = 22.

Authorized licensed use limited to: Universidad de Valencia. Downloaded on May 19,2023 at 08:07:25 UTC from IEEE Xplore. Restrictions apply.

TABLE XII
STUDY OF STATISTICALLY SIGNIFICANT DIFFERENCES BETWEEN THE VR AND THE AR APPLICATION—OBJECTIVE DATASETS OF TASK 2—GROUP A

	2.T Time	2.C Correct	2.M Mistaken
VR (mean ± SD)	157.260 ± 81.421	9.174 ± 2.309	3.130 ± 3.584
AR (mean ± SD)	88.565 ± 68.657	9.826 ± 0.834	1.043 ± 2.602
t	4.313	−1.789	3.372
p	$<10^{-3}$	0.087	0.003
Cohen's d	0.915	−0.415	0.675

Group size = 23. DoF = 22.

TABLE XIII
STUDY OF STATISTICALLY SIGNIFICANT DIFFERENCES BETWEEN THE AR AND THE VR APPLICATION—OBJECTIVE DATASETS OF TASK 1—GROUP B

	1.T Time	1.C Correct	1.M Mistaken
AR (mean ± SD)	32.864 ± 19.219	3.000 ± 0.000	2.955 ± 2.554
VR (mean ± SD)	31.682 ± 35.595	3.000 ± 0.000	2.909 ± 12.340
t	0.151	-	0.018
p	0.881	-	0.986
Cohen's d	0.043	-	0.006

Group size = 22. DoF = 21.

TABLE XIV
STUDY OF STATISTICALLY SIGNIFICANT DIFFERENCES BETWEEN THE AR AND THE VR APPLICATION—OBJECTIVE DATASETS OF TASK 2—GROUP B

	2.T Time	2.C Correct	2.M Mistaken
AR (mean ± SD)	124.409 ± 93.943	9.682 ± 0.780	2.909 ± 3.571
VR (mean ± SD)	119.045 ± 88.398	9.864 ± 0.468	4.727 ± 15.661
t	0.229	−0.890	−0.533
p	0.821	0.383	0.600
Cohen's d	0.059	−0.292	−0.189

Group size = 22. DoF = 21.

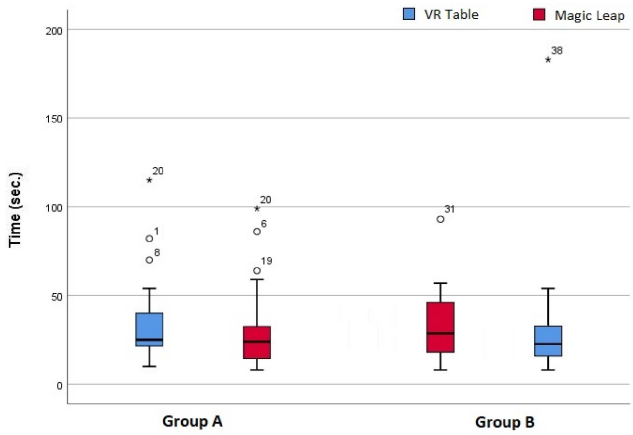

Fig. 7. Box plots (paired t-tests) for dataset 1.T (time, Task 1) and groups A (left) and B (right).

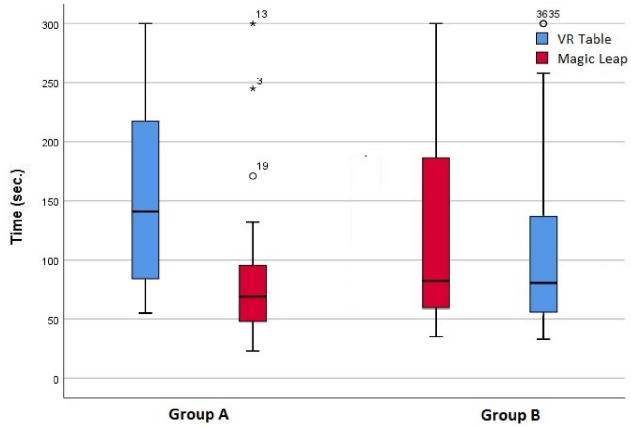

Fig. 8. Box plots (paired t-tests) for dataset 2.T (time, Task 2) and groups A (left) and B (right).

Group A completing Task 1), XII (paired t-test for Group A, Task 2), XIII (paired t-test for Group B, Task 1), and XIV (paired t-test for Group B, Task 2) present the results of these repeated measures tests.

This time, three significant results (for datasets 1.M, 2.T, and 2.M) occur for Group A, where the AR-based setup is shown to allow a better performance. In fact, the average time for Task 2 almost halves with respect to the VR table. This result makes sense since setup B setup provides better results in terms of time and errors when is tested second, whereas no statistical significance can be found for the same tasks when the order of the setups tested is reversed (see Tables XIII and XIV). This is most probably caused by a learning effect that cancels out the benefits of using setup B. The fact that the same type of interaction system is used for both setups contributes to this learning effect.

Figs. 7 and 8 show the box plots of datasets 1.T and 2.T, respectively, for groups A and B. The significant difference for Group A, dataset 2.T (time), in Task 2 is clearly visible in Fig. 8.

A retrospective comparison with the article presented in [1] can also be performed, although Task 2 has been changed and is more complex now. In any case, we can see that the results for both tasks have changed. We compare Tables IX and X with Table IV of the article presented in [1]. Now, the AR-based setup is not outperformed by the VR-based setup (in fact, the results lean in the opposite direction). Another interesting result is that the standard deviation for the time dataset in Task 2 has doubled, for both setups, with respect to the article presented in [1]. This probably means that, in this new experiment, some users still had important difficulties with this complex task but there were also several participants who were able to complete it very fast.

Next, we analyze the responses obtained from Questionnaire 2. First, we analyze the Boolean datasets coming from the two-choice questions (Q1, Q3, and Q5) in which users were prompted to decide between the two setups regarding usefulness, preference, and recommendation. As depicted in Table XV, the AR-based setup is clearly perceived as more useful (82.2% versus 17.8%), preferred (77.8% versus 22.2%), and recommended (73.3% versus 26.7%) over the VR-based setup. Within groups, the differences are similar, and sometimes overwhelming, as in Group B, Q1 (95.5% versus 4.5%). These results are confirmed by a binomial test, revealing that these differences are indeed statistically significant and not likely produced by chance. The

Authorized licensed use limited to: Universidad de Valencia. Downloaded on May 19,2023 at 08:07:25 UTC from IEEE Xplore. Restrictions apply.

TABLE XV
STUDY OF STATISTICALLY SIGNIFICANT DIFFERENCES IN USER RESPONSES FOR USEFULNESS, PREFERENCE, AND RECOMMENDATION

Question	*Group*	*VR*	*AR*	*Binomial p-value*
Q1 Usefulness	*A*	*7 (30.4 %)*	*16 (69.6 %)*	
	B	*1 (4.5 %)*	*21 (95.5 %)*	
	Total	*8 (17.8 %)*	*37 (82.2 %)*	$<10^{-3}$
Q3 Preference	*A*	*5 (21.7 %)*	*18 (78.3 %)*	
	B	*5 (22.7 %)*	*17 (77.3 %)*	
	Total	*10 (22.2 %)*	*35 (77.8 %)*	$<10^{-3}$
Q5 Recommendation	*A*	*7 (30.4 %)*	*16 (69.6 %)*	
	B	*5 (22.7 %)*	*17 (77.3 %)*	
	Total	*12 (26.7 %)*	*33 (73.3 %)*	*0.002*

TABLE XVI
SIGNIFICANT RESULTS OF THE WITHIN-SUBJECTS MULTIFACTORIAL ANOVA

Dependent variable	*Factor*	*F*	*p*	η^2
DF	*Age*	*2.932*	*0.031*	*0.352*
	*Age*Setup*	*3.091*	*0.032*	*0.314*
OF	*Age*	*2.959*	*0.030*	*0.354*
3D	*Setup*	*4.758*	*0.038*	*0.150*
SC	*Gender*Setup*	*4.204*	*0.050*	*0.135*
US	*Setup*	*5.345*	*0.029*	*0.165*

open-ended questions (Q2, Q4, and Q6) reveal that many users felt the AR-based setup "more engaging," "more realistic," "more natural," "more comfortable," "easier to use," and "more responsive." Nevertheless, some users wearing glasses disliked this setup for ergonomic reasons. This is an important lesson to bear in mind for the future exploitation of this technology.

Finally, we have performed a within-subjects multifactorial ANOVA, including all the subjective data (Group A and Group B), in order to analyze if there are significant interactions among the different features of the population and their responses to the questions, as shown in Table II. This allows us to know if certain groups of people rate these setups better than other groups. The following factors (independent variables) are considered: gender, age, and tested setup. The dependent variables are SF, CF, DF, EF, RF, OF, depth perception, usefulness, and score.

The analysis reveals that there are several significant effects, as shown in Table XVI. The effect of age on both the DF ($F[5, 27] = 2.932$, $p = 0.031$, and $\eta^2 = 0.352$) and the OF ($F[5, 27] = 2.959$, $p = 0.03$, and $\eta^2 = 0.354$) is not surprising since older people tend to underrate this kind of systems. The effect of the setup on both depth perception (3-D) ($F[1, 27] = 4.758$, $p = 0.038$, and $\eta^2 = 0.15$) and usefulness (US) ($F[1, 27] = 5.345$, $p = 0.029$, and $\eta^2 = 0.165$) simply reflects that the AR-based system is perceived as more useful and more realistic in terms of depth perception. The effect of the combination of gender and setup on the final score (SC) ($F[1, 27] = 4.204$, $p = 0.05$, and $\eta^2 = 0.135$) reflects that, for some reason, women using the Magic Leap provided higher score ratings. Finally, the distraction factor presents significant differences ($F[4, 27] = 3.091$, $p = 0.032$, and $\eta^2 = 0.314$) when analyzed by age groups and setup. This is explained because the DF using the VR-based setup is affected by age, and participants over 50 years felt more distracted with this setup than the rest.

In order to measure the learning effect, we have also performed a similar multifactorial ANOVA with the objective performance measures as dependent variables and the presentation order (the order in which the two setups were presented to the participants) as independent variable. The results show that the presentation order causes significant effects for the time dataset of Task 1 (1.T) in the VR setup ($F[1, 9] = 6.695$, $p = 0.029$, and $\eta^2 = 0.427$), for the errors dataset of Task 1 (1.M) in the AR setup ($F[1, 9] = 9.998$, $p = 0.012$, and $\eta^2 = 0.526$), and for the errors dataset of Task 2 (2.M) in the AR setup ($F[1, 9] = 521.241$, $p < 10^{-3}$, and $\eta^2 = 0.983$). The descriptors for the rest of datasets also suggest a learning effect caused by the presentation order, but due to high variances, statistical significance is not reached in these cases.

V. CONCLUSION

In a recently published paper [1], we preliminarily analyzed the suitability of applying VR and AR for anatomy training comparing an AR-based setup (using a Microsoft HoloLens device) with a semi-immersive VR-based setup (using a VR table), for anatomy training. The results of this previous research showed that the VR setup was clearly most suitable. In this article, we complete this research by comparing an improved version of the AR-based setup (using a Magic Leap One device) with the aforementioned VR-based system. Unlike in our previous research, the two setups now use the same interaction system, which has also been improved in both setups. Thus, the differences between the two setups lie on the visualization and conceptualization of the application for each of the paradigms (AR or VR). The objective of this new experiment is to confirm if the modifications made in the setups can change the previous photograph. Our hypothesis is that the improved AR-based setup will be more suitable, for the anatomy training application, than the VR-based setup, in terms of these three dimensions: subjective perception; objective performance; and explicit two-choice recommendation.

For this reason, we conducted an experimental research with 45 participants, comparing the use of an anatomy training software on which the users had to complete two anatomy-related tasks with the two setups. Objective and subjective data were collected. From the analysis performed to these data, we can conclude that the AR-based setup is now the preferred choice [dimension (iii)] of the participants and provides better results in terms of subjective perception-related measures [dimension (i)], confirming our initial hypothesis. One of the three objective performance-related datasets also shows a small but statistically significant advantage in favor of the AR-based setup. In fact, the AR-based setup also performs statistically better, in terms of total task time for one of the tasks, when tested second, whereas the VR-based setup does not show any performance advantage, even if it is tested second. This is an indication that the AR-based setup provides a slight performance advantage over the VR-based setup (dimension (ii)). This advantage is canceled

Authorized licensed use limited to: Universidad de Valencia. Downloaded on May 19,2023 at 08:07:25 UTC from IEEE Xplore. Restrictions apply.

out by the learning effect when the testing order is reversed. Regarding complexity, the fact that only a few differences are found between the results of Task 1 and Task 2 suggests that an increased complexity should not be a major factor in the choice of setup.

In addition, maximum average values for the subjective measures have increased with respect to our first research for most of the factors, which means that the improvements made have been successful. Objectives measures are not as comparable since the tasks are slightly more complex in this new experiment, but the AR-based setup has eliminated the performance gap with the VR-based setup, with respect to our first research. In any case, we can conclude that both setups—and this is probably the most important conclusion—offer excellent results in the subjective measures, with both systems approaching the highest possible values.

Regarding the limitations of the experiment, we still need to test the learning outcomes that the application could provide. No anatomy pre- and post-test experience/knowledge was performed, although the number of participants was sufficiently high, especially given the complexity of conducting such an experiment in the midst of a global pandemic. The context (anatomy, dissection, physiology, etc.) in which the training application is finally used may also influence the comparison. In addition, we are of course constrained by the technological limitations of the hardware being tested, so further changes in both setups could increase the usefulness of the application. This prevents us from stating categorically that the AR paradigm is better than the VR paradigm for this type of application. However, it does allow us to know how to obtain an appropriate setup for this problem.

In any case, we believe both setups are now ready to transfer these applications to teaching environments. This will be our next step in this long research journey so that we can finally test the effectiveness—in terms of learning outcomes—of these applications in real teaching environments. We will also continue researching about possible hardware and software improvements, especially on the VR-based setup. For instance, it would be interesting to compare the new semi-immersive AR-based setup with an improved version of the VR-based setup using an Oculus Quest 2 or an HTC Vive Cosmos HMD since the anatomy training application could also be implemented with a fully immersive VR setup.

REFERENCES

[1] R. S. Vergel, P. M. Tena, S. C. Yrurzum, and C. Cruz-Neira, "A comparative evaluation of a virtual reality table and a HoloLens-based augmented reality system for anatomy training," *IEEE Trans. Human-Mach. Syst.*, vol. 50, no. 4, pp. 337–348, Aug. 2020.

[2] H. Brun et al., "Mixed reality holograms for heart surgery planning: First user experience in congenital heart disease," *Eur. Heart J. - Cardiovasc. Imag.*, vol. 20, no. 8, pp. 883–888, Aug. 2019, doi: 10.1093/ehjci/jey184.

[3] J. K. Weeks et al., "Harnessing augmented reality and CT to teach first-year medical students head and neck anatomy," *Acad. Radiol.*, vol. 28, no. 6, pp. 871–876, 2021.

[4] L. Qian, J. Y. Wu, S. P. DiMaio, N. Navab, and P. Kazanzides, "A review of augmented reality in robotic-assisted surgery," *IEEE Trans. Med. Robot. Bionics*, vol. 2, no. 1, pp. 1–16, Feb. 2020.

[5] M. Ferro, D. Brunori, F. Magistri, L. Saiella, M. Selvaggio, and G. A. Fontanelli, "A portable da Vinci simulator in virtual reality," in *Proc. IEEE 3rd Int. Conf. Robot. Comput.*, 2019, pp. 447–448, doi: 10.1109/IRC.2019.00093.

[6] S. Casas, C. Portalés, and M. Fernández, "To move or not to move?: The challenge of including believable self-motion cues in virtual reality applications—Understanding motion cueing generation in virtual reality," in *Cases on Immersive Virtual Reality Techniques*. Hershey, PA, USA: IGI Global, 2019, pp. 124–144.

[7] S. Casas, C. Portalés, I. García-Pereira, and J. Gimeno, "Mixing different realities in a single shared space: Analysis of mixed-platform collaborative shared spaces," in *Harnessing the Internet of Everything (IoE) for Accelerated Innovation Opportunities*. Hershey, PA, USA: IGI Global, 2019, pp. 175–192, doi: 10.4018/978-1-5225-7332-6.ch008.

[8] H. F. Al Janabi et al., "Effectiveness of the HoloLens mixed-reality headset in minimally invasive surgery: A simulation-based feasibility study," *Surg. Endosc.*, vol. 34, no. 3, pp. 1143–1149, 2020.

[9] T. Jiang, D. Yu, Y. Wang, T. Zan, S. Wang, and Q. Li, "HoloLens-based vascular localization system: Precision evaluation study with a three-dimensional printed model," *J. Med. Internet Res.*, vol. 22, no. 4, 2020, Art. no. e16852.

[10] I. Kuhlemann, M. Kleemann, P. Jauer, A. Schweikard, and F. Ernst, "Towards X-ray free endovascular interventions—Using HoloLens for on-line holographic visualisation," *Healthcare Technol. Lett.*, vol. 4, no. 5, pp. 184–187, 2017.

[11] Y. Zuo et al., "A novel evaluation model for a mixed-reality surgical navigation system: Where Microsoft HoloLens meets the operating room," *Surg. Innov.*, vol. 27, pp. 193–202, 2020.

[12] E. Rae, A. Lasso, M. S. Holden, E. Morin, R. Levy, and G. Fichtinger, "Neurosurgical burr hole placement using the Microsoft HoloLens," in *Proc. Imag.-Guided Procedures, Robot. Interventions, Model.*, 2018, vol. 10576, pp. 190–197, doi: 10.1117/12.2293680.

[13] P. Pratt et al., "Through the HoloLens™ looking glass: Augmented reality for extremity reconstruction surgery using 3D vascular models with perforating vessels," *Eur. Radiol. Exp.*, vol. 2, no. 1, 2018, Art. no. 2.

[14] T. Romanus, S. Frish, M. Maksymenko, W. Frier, L. Corenthy, and O. Georgiou, "Mid-air haptic bio-holograms in mixed reality," in *IEEE Int. Symp. Mixed Augmented Reality Adjunct (ISMAR-Adjunct)*, pp. 348–352, 2019.

[15] C.-C. Teng et al., "Mixed reality patients monitoring application for critical care nurses," in *Proc. 3rd Int. Conf. Med. Health Inform.*, 2019, pp. 49–53, doi: 10.1145/3340037.3340050.

[16] H. B. Andersson, T. Børresen, E. Prasolova-Førland, S. McCallum, and J. G. Estrada, "Developing an AR application for neurosurgical training: Lessons learned for medical specialist education," in *Proc. IEEE Conf. Virtual Reality 3-D User Interfaces Abstr. Workshops*, 2020, pp. 407–412, doi: 10.1109/VRW50115.2020.00087.

[17] C. A. Neves et al., "Application of holographic augmented reality for external approaches to the frontal sinus," *Int. Forum Allergy Rhinol.*, vol. 10, no. 7, pp. 920–925, Jul. 2020, doi: 10.1002/alr.22546.

[18] M. C. Howard, "A meta-analysis and systematic literature review of virtual reality rehabilitation programs," *Comput. Hum. Behav.*, vol. 70, pp. 317–327, 2017.

[19] P. G. Wittkopf, D. M. Lloyd, O. Coe, S. Yacoobali, and J. Billington, "The effect of interactive virtual reality on pain perception: A systematic review of clinical studies," *Disabil. Rehabil.*, vol. 42, no. 26, pp. 3722–3733, 2020.

[20] B. Cieślik, J. Mazurek, S. Rutkowski, P. Kiper, A. Turolla, and J. Szczepańska-Gieracha, "Virtual reality in psychiatric disorders: A systematic review of reviews," *Complement. Therapies Med.*, vol. 52, 2020, Art. no. 102480.

[21] S. F. M. Alfalah, J. F. M. Falah, T. Alfalah, M. Elfalah, N. Muhaidat, and O. Falah, "A comparative study between a virtual reality heart anatomy system and traditional medical teaching modalities," *Virtual Reality*, vol. 23, pp. 229–234, 2019.

[22] C.-H. Chien, C.-H. Chen, and T.-S. Jeng, "An interactive augmented reality system for learning anatomy structure," in *Proc. Int. Multi Conf. Eng. Comput. Scientists*, 2010, pp. 370–375.

[23] C. V. Le, J. G. Tromp, and V. Puri, "Using 3D simulation in medical education: A comparative test of teaching anatomy using virtual reality: Virtual reality, augmented reality, artificial intelligence, Internet of things, robotics, industry 4.0," *Emerg. Technol. Health Med.*, vol. 12, pp. 21–33, 2018.

[24] A. M. Codd and B. Choudhury, "Virtual reality anatomy: Is it comparable with traditional methods in the teaching of human forearm musculoskeletal anatomy?," *Anat. Sci. Educ.*, vol. 4, no. 3, pp. 119–125, 2011.

Authorized licensed use limited to: Universidad de Valencia. Downloaded on May 19,2023 at 08:07:25 UTC from IEEE Xplore. Restrictions apply.

[25] J. Ferrer-Torregrosa, M. Á. Jiménez-Rodríguez, J. Torralba-Estelles, F. Garzón-Farinós, M. Pérez-Bermejo, and N. Fernández-Ehrling, "Distance learning ECTS and flipped classroom in the anatomy learning: Comparative study of the use of augmented reality, video and notes," *BMC Med. Educ.*, vol. 16, no. 1, 2016, Art. no. 230.
[26] J. Ferrer-Torregrosa, J. Torralba, M. A. Jimenez, S. García, and J. M. Barcia, "ARBOOK: Development and assessment of a tool based on augmented reality for anatomy," *J. Sci. Educ. Technol.*, vol. 24, no. 1, pp. 119–124, 2015.
[27] S. K. Ghosh, "Cadaveric dissection as an educational tool for anatomical sciences in the 21st century," *Anat. Sci. Educ.*, vol. 10, no. 3, pp. 286–299, 2017, doi: 10.1002/ase.1649.
[28] M. A. Bohl et al., "Evaluation of a novel surgical skills training course: Are cadavers still the gold standard for surgical skills training?," *World Neurosurg.*, vol. 127, pp. 63–71, 2019.
[29] World Health Org., "IARC monographs on the evaluation of carcinogenic risks to humans," Lyon, France, 2006. [Online]. Available: https://monographs.iarc.who.int/wp-content/uploads/2018/06/mono88.pdf/
[30] D. Chytas, M. Piagkou, M. Salmas, and E. O. Johnson, "Mixed and augmented reality: Distinct terms, different anatomy teaching potential," *Anat. Sci. Educ.*, vol. 14, pp. 519–520, 2021.
[31] S. Zafar and J. J. Zachar, "Evaluation of HoloHuman augmented reality application as a novel educational tool in dentistry," *Eur. J. Dent. Educ.*, vol. 24, no. 2, pp. 259–265, 2020.
[32] G. Riva, R. M. Baños, C. Botella, F. Mantovani, and A. Gaggioli, "Transforming experience: The potential of augmented reality and virtual reality for enhancing personal and clinical change," *Front. Psychiatry*, vol. 7, 2016, Art. no. 164.
[33] C. Trepkowski, D. Eibich, J. Maiero, A. Marquardt, E. Kruijff, and S. Feiner, "The effect of narrow field of view and information density on visual search performance in augmented reality," in *Proc. IEEE Conf. Virtual Reality 3D User Interfaces*, 2019, pp. 575–584.
[34] M. Frutos-Pascual, C. Creed, and I. Williams, "Head mounted display interaction evaluation: Manipulating virtual objects in augmented reality," in *Proc. 17th IFIP TC 13 Int. Conf. Human-Comput. Interaction*, 2019, pp. 287–308.
[35] A. Macaranas, A. N. Antle, and B. E. Riecke, "What is intuitive interaction? Balancing users' performance and satisfaction with natural user interfaces," *Interact. Comput.*, vol. 27, no. 3, pp. 357–370, May 2015.
[36] R. Serrano, P. Morillo, S. Casas, and C. Cruz-Neira, "An empirical evaluation of two natural hand interaction systems in augmented reality," *Multimed. Tools Appl.*, vol. 81, pp. 31657–31683, Apr. 2022, doi: 10.1007/s11042-022-12864-6.
[37] S. Park, S. Bokijonov, and Y. Choi, "Review of Microsoft HoloLens applications over the past five years," *Appl. Sci.*, vol. 11, no. 16, 2021, Art. no. 7259.
[38] W. Flynn, N. Kumar, R. Donovan, M. Jones, and P. Vickerton, "Delivering online alternatives to the anatomy laboratory: Early experience during the COVID-19 pandemic," *Clin. Anatomy*, vol. 34, pp. 757–765, 2021.
[39] J. Iwanaga, M. Loukas, A. S. Dumont, and R. S. Tubbs, "A review of anatomy education during and after the COVID-19 pandemic: Revisiting traditional and modern methods to achieve future innovation," *Clin. Anatomy*, vol. 34, no. 1, pp. 108–114, 2021.
[40] A. Sivananthan et al., "A feasibility trial of HoloLens 2™; using mixed reality headsets to deliver remote bedside teaching during COVID-19," *JMIR Formative Res.*, vol. 6, pp. 1–7, 2022.
[41] K. G. Byrnes, P. A. Kiely, C. P. Dunne, K. W. McDermott, and J. C. Coffey, "Communication, collaboration and contagion: 'Virtualisation' of anatomy during COVID-19," *Clin. Anatomy*, vol. 34, no. 1, pp. 82–89, 2021.
[42] D. Chytas et al., "The role of augmented reality in anatomical education: An overview," *Ann. Anatomy-Anatomischer Anzeiger*, vol. 229, 2020, Art. no. 151463.
[43] J. Zhao, X. Xu, H. Jiang, and Y. Ding, "The effectiveness of virtual reality-based technology on anatomy teaching: A meta-analysis of randomized controlled studies," *BMC Med. Educ.*, vol. 20, 2020, Art. no. 127.
[44] S. Dreimane and L. Daniela, "Educational potential of augmented reality mobile applications for learning the anatomy of the human body," *Technol., Knowl. Learn.*, vol. 26, pp. 763–788, 2021.
[45] M. L. Duarte, L. R. Santos, J. B. G. Júnior, and M. S. Peccin, "Learning anatomy by virtual reality and augmented reality: A scope review," *Morphologie*, vol. 104, pp. 254–266, 2020.
[46] M. Romand, D. Dugas, C. Gaudet-Blavignac, J. Rochat, and C. Lovis, "Mixed and augmented reality tools in the medical anatomy curriculum," *Stud. Health Technol. Inform.*, vol. 270, pp. 322–326, 2020.
[47] R. Wirza, S. Nazir, H. U. Khan, I. García-Magariño, and R. Amin, "Augmented reality interface for complex anatomy learning in the central nervous system: A systematic review," *J. Healthcare Eng.*, vol. 2020, 2020, Art. no. 8835544.
[48] U. Uruthiralingam and P. M. Rea, "Augmented and virtual reality in anatomical education–A systematic review," *Adv. Exp. Med. Biol.*, vol. 1235, pp. 89–101, 2020.
[49] C. P. R. Triepels et al., "Does three-dimensional anatomy improve student understanding?," *Clin. Anatomy*, vol. 33, no. 1, pp. 25–33, May 2020, doi: 10.1002/ca.23405.
[50] A. C. Stirling and C. Moro, "The use of virtual, augmented and mixed reality in anatomy education," in *Teaching Anatomy*. Berlin, Germany: Springer, 2020, pp. 359–366. [Online]. Available: https://doi.org/10.1007/978-3-030-43283-6_36
[51] N. Kumar, S. Pandey, and E. Rahman, "A novel three-dimensional interactive virtual face to facilitate facial anatomy teaching using Microsoft HoloLens," *Aesthetic Plast. Surg.*, vol. 45, pp. 1005–1011, 2021.
[52] S. G. Izard and J. A. J. Méndez, "App design and implementation for learning human anatomy through virtual and augmented reality," in *Information Technology Trends for a Global and Interdisciplinary Research Community*. Hershey, PA, USA: IGI Global, 2021, pp. 72–87.
[53] D. Schott et al., "A VR/AR environment for multi-user liver anatomy education," in *Proc. IEEE Virtual Reality 3D User Interfaces*, 2021, pp. 296–305.
[54] P. Maniam et al., "Exploration of temporal bone anatomy using mixed reality (HoloLens): Development of a mixed reality anatomy teaching resource prototype," *J. Vis. Commun. Med.*, vol. 43, no. 1, pp. 17–26, 2020.
[55] N. S. Birbara and N. Pather, "Real or not real: The impact of the physical fidelity of virtual learning resources on learning anatomy," *Anat. Sci. Educ.*, vol. 14, pp. 774–787, 2021.
[56] Y. P. Zinchenko et al., "Virtual reality is more efficient in learning human heart anatomy especially for subjects with low baseline knowledge," *New Ideas Psychol.*, vol. 59, 2020, Art. no. 100786.
[57] N. S. Birbara, C. Sammut, and N. Pather, "Virtual reality in anatomy: A pilot study evaluating different delivery modalities," *Anat. Sci. Educ.*, vol. 13, no. 4, pp. 445–457, 2020.
[58] E. A. Duncan-Vaidya and E. L. Stevenson, "The effectiveness of an augmented reality head-mounted display in learning skull anatomy at a community college," *Anat. Sci. Educ.*, vol. 14, pp. 221–231, 2021.
[59] K. Bogomolova et al., "The effect of stereoscopic augmented reality visualization on learning anatomy and the modifying effect of visual-spatial abilities: A double-center randomized controlled trial," *Anat. Sci. Educ.*, vol. 13, no. 5, pp. 558–567, 2020.
[60] R. Kurul, M. N. Ögün, A. N. Narin, Ş. Avci, and B. Yazgan, "An alternative method for anatomy training: Immersive virtual reality," *Anat. Sci. Educ.*, vol. 13, no. 5, pp. 648–656, 2020.
[61] H. Ro, J.-H. Byun, Y. J. Park, N. K. Lee, and T.-D. Han, "AR pointer: Advanced ray-casting interface using laser pointer metaphor for object manipulation in 3D augmented reality environment," *Appl. Sci.*, vol. 9, no. 15, 2019, Art. no. 3078.
[62] J.-J. Arino, M.-C. Juan, J.-A. Gil-Gómez, and R. Mollá, "A comparative study using an autostereoscopic display with augmented and virtual reality," *Behav. Inf. Technol.*, vol. 33, no. 6, pp. 646–655, 2014.
[63] M.-C. Juan, I. García-García, R. Mollá, and R. López, "Users' perceptions using low-end and high-end mobile-rendered HMDs: A comparative study," *Computers*, vol. 7, no. 1, 2018, Art. no. 15.
[64] D. Rodríguez-Andrés, M.-C. Juan, M. Méndez-López, E. Pérez-Hernández, and J. Lluch, "MnemoCity task: Assessment of childrens spatial memory using stereoscopy and virtual environments," *PLoS One*, vol. 11, no. 8, 2016, Art. no. e0161858.
[65] B. G. Witmer and M. J. Singer, "Measuring presence in virtual environments: A presence questionnaire," *Presence, Teleoperators Virtual Environ.*, vol. 7, no. 3, pp. 225–240, 1998.
[66] F. J. Massey Jr., "The Kolmogorov-Smirnov test for goodness of fit," *J. Amer. Statist. Assoc.*, vol. 46, no. 253, pp. 68–78, 1951.
[67] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, no. 3/4, pp. 591–611, 1965.

Authorized licensed use limited to: Universidad de Valencia. Downloaded on May 19,2023 at 08:07:25 UTC from IEEE Xplore. Restrictions apply.