5 CollabAR – Implementation Process

The previous chapter detailed the architectural design of CollabAR, that is, the main ideas and decisions taken to fulfil the goals proposed for the software development and the efforts done to align the technical aspects of the project with the educational goals present in the research. The chapter also identified key components of the development process to follow and proposed an overall plan for the implementation of the software.

This chapter will continue the technical analysis of the project, this time with a focus on the final implementation process and the important details that create opportunities for discussion. This analysis will highlight the research done to identify and propose a suitable interaction design that takes into consideration the context around CollabAR, the evolving tendencies in the implementation of AR experiences and all the in formation regarding good and bad practices that had been found in the literature.

It will also be an opportunity to compare the plans and ideas proposed in the architectural design against the actual implementation of the software, identify the reasoning behind such changes and how the context and circumstances of the development shaped the final product.

Finally, this chapter will detail the software development process used. The process proposed and implemented two iterations of development which emphasized testing procedures and early validations. There was also an important effort to integrate the development of the tool with the case study and evolved with the particularities of the course and the educational needs of the final users.

Chapter 3 detailed the reasoning behind the chosen development methodology, explaining how an iterative approach allowed for experimentation and flexibility, and that better accommodated the constraint imposed in time and resources to facilitate development following a “just-enough” approach to architecture, design and implementation.

As a result, CollabAR undertook a development process in two iteration cycles which derived in the testing and deployment of a prototype used for observation and data gathering. Both cycles were structured around a design phase, implementation and quick adjustments. In between iterations it was possible to allocate some time to data analysis and planification for the nest iteration cycle.

Iteration 1 focused on building and testing the functional AR technology behind CollabAR: the shared AR space and the fundamental AR interactions. The overall goal was to explore the main use cycle of creating a room, starting a session with other users and share visualizations of different digital objects.

The in between testing and analysis was used to better aligning CollabAR to the particularities of the Industry Project Course, as well as incorporating key observations provided by initial test users.

Iteration 2 focused on providing the needed functionalities to fulfil the uses cases identified in the architectural design process, those that provided most value to the users and the research.

A final analysis of the results was done to identify possible solutions to the issues and ideas generated from the data gathering, both at an implementation and design level. There is also a discussion of possible future steps for CollabAR, of how the software can evolve and the lessons learned for future or related scenarios.

The following sections will then describe this implementation process and provide a form of post-mortem of the activity to promote discussion and analysis of the results. Section 5.1 will first detail the research done as part of the interaction design of CollabAR, research that informed the approach that the software was going to take in relation to AR technologies and how to implement a set of interactions for the collaborative and multi-platform tasks present in the context of development. Section 5.2 will describe and show the results of the development process, organizing the analysis around the two iteration cycles described previously. Section 5.3 will act as a final analysis and conclusion of the results of the implementation process, the development of CollabAR and for the technical section of the research in general.

# 5.1 Interaction Design

An important effort in the construction of CollabAR was put in the design of the interactive elements at the core of the experience. The interactions that received the most attention during the design process can be classified in three:

* Interactions with digital objects in the AR spaced created for the room.
* Social interactions between participants in relation to group management and the objective activities of the course.
* The controls of the tool itself, the individual components and the flow of the work session.

These interactions were prioritised because they represent the tools that CollabAR uses to promote the behaviours and ideas around collaboration at the core of the objectives and design of the project. It is especially important in the design of social interactions and the structure that is being proposed for the work process but without interfering too much in a negative way or hindering the academical work of the students. In a similar light, the project also offers the opportunity to explore solutions to a variety of interaction problems associated with the use of AR in mobile devices such as:

* Simultaneously interact with digital objects that are shared with other users.
* Organize and select information.
* Use and understand social cues from other users.
* Plan and organize tasks for a common objective.
* Jump from using the AR application to any other external activity and vice versa.

The following section will explore all these aspects of the interaction design in two main approaches. The first one will focus on general AR design and will detail the research done to understand how to use mobile AR devices in social and collaborative activities, and the user-guided design followed to identify the best approach to the overall AR interactions needed for the implementation of CollabAR.

The second part of the section will detail the decisions taken for the top-down interactions implemented for CollabAR, that is, the final implementation and user experience for activities related to create a work session, interact with the digital objects in the room and structure the collaborative activities of the group using CollabAR.

## 5.1.1 Collaborative interactions in AR

The use of AR technology in the core of a software solution has moved beyond the speculative stage and is no longer confined to pure research. Successful implementations of AR in everyday usage and production, such as those reported by Kowalewski et al. (2017), demonstrate its practical applications. Furthermore, the availability of ready-to-use software like Manifest (Manifest AR Work Instruction Platform, 2022) and the accessibility offered by end-user products like the Meta Quest 3 (Meta, n.d.) highlight the need to consider any development within the broad spectrum of mixed realities as a product intended for the general consumer. This is why the development of CollabAR, although envisioned as a research prototype, aims for an implementation guided by a user-centred, human-centred design approach, as explained in chapter 3.

For a human-centred approach it is important to considers the requirements that are directly related to interactions of the end user with the core utility of the product and the

context in which it happens (Boy, 2017). This general design goal is highly relevant for CollabAR if seen as a product introducing AR technology to the collaborative education context. AR technology remains relatively new to most consumers and most of them have little or no contact with it in their daily routines (Chang, Kuo, & Du, 2023). Consequently, users have not yet developed a consistent digital literacy around these and similar technologies. Is equally uncommon, even for experience users, to have any form of expectations about the behaviour and allowed interactions of any particular development. Even when an application is deployed in more recognizable platforms, like a smartphone, the affordances and interaction methods that the users are accustomed to may not be the same or may be entirely absent.

Several efforts have been made to create a shared language for developers to use

in the interaction design of their applications, and to solve common problems with

consistent approaches. Research studies such as those conducted by Piumsomboon, Clark, Billinghurst, and Cockburn (2013) and Wobbrock, Morris, and Wilson (2009) have formulated and compiled an array of gestures tailored to various forms of Augmented Reality and

Mixed Reality (MR) interactions. These proposed sets of gestures encompass a broad

spectrum of general interactions, ranging from basic object selections and manipulation

to complex tasks like file browsing and content editing. However, these gestures are designed under the assumption of single-user interactions.

While some interactions may coincide with those found in single-user applications, others are unique to multi-user settings. Moreover, it is important to note that collaboration is not an inherent aspect of multi-user scenarios. Various forms of social interactions, facilitated by technology, can introduce their own unique challenges and characteristics. The objective of the research described in this section is to explore the particularities of collaborative interaction in AR. Using the same user-driven methodology proposed in the studies cited

previously, the aim is to identify common tasks executed in a collaborative setting using different forms of AR technology. Based on that, the end result is a proposed set of gestures that capture the most natural interactions that users perform and tend to converge to. The results were directly used as a design guideline for the interaction design of CollabAR, highlighting effective strategies to tackle particular tasks related to the use of AR in the collaborative scenario and the use of mobile devices as the main focus for deployment.

### 5.1.1.1 Research Methodology

The interaction design process was guided by a three-step procedure: 1.) Identification and selection of tasks, 2.) Data collection from participants proposing interactions for the identified tasks, and 3.) Analysis of the collected data to identify convergent approaches, recurrent issues, and common behaviours.

In total, 24 different participants were interviewed with a median age of 41 and standard

deviation of 11. Of the participants, 66% were women while 34% were men. In terms of demographics, 20% of the participants were university students with careers mostly in design and computer science. The remaining participants were professionals in different fields like IT, medicine, and administration.

Two different scenarios were designed for the test. In Scenario 1, participants were equipped with a smartphone, which they were required to hold in one hand, leaving

the other one free to interact during each task. Scenario 2 involved the use of a Microsoft HoloLens headset, which afforded the participants the freedom to use both hands. Figure 1 shows examples of each scenario.

Un hombre con gorra y lentes oscuros

El contenido generado por IA puede ser incorrecto.

Figure 1 Setups for Scenario 1 (using a smartphone) and Scenario 2 (using a HoloLens)

Both scenarios facilitated distinct modes of interaction, which were anticipated to influence the participants’ thought processes and instinctive responses for each task. It was hypothesized that:

* Scenario 1 would prompt individuals to interact with the screen rather than the free hand.
* Scenario 2 would prompt more air gestures.
* Each scenario would converge to its own set of gestures rather than both converging to the same general set.

It is also important to remark that at this point in the design of the software it was not clear if CollabAR was going to be implemented with full cross-platform compatibility in mind, and it was also important to identify how users proposed gestures for the same task in different scenarios, if they diverged enough that would be necessary to create to different input systems of if there was some form of convergent method that could be used or expanded upon.

As noted in previous sections, Piumsomboon et al. (2013) proposed a set of tasks related to the most common activities a person would need to perform in an AR application. These tasks are related to selection, manipulation, and transformation of digital objects. Nowadays these actions have become common and expected in any environment that requires 3D interactions, with the particularity that these tasks are all related to single-user experiences. Based on ideas present in previous works (Pinelle, Gutwin, & Greenberg, 2003; Scoular, Duckworth & Heard, 2020), a set of 14 tasks was identified, representing the most common activities done in a collaborative activity and that could be supported by an AR application. Three principles guided the final selection:

* All or most members of the team are present in the same place. Remote collaboration
* was considered but not enforced or assumed.
* All members of the team have access to the same AR application but not necessarily to the same platform nor do they have the same means of interaction.
* The context in which the collaboration is done can be varied (education, work, problem solving, etc.) with the objective of applying to any form of collaborative activity.

Table 1 shows the list of selected tasks divided into three distinct categories. The first category refers to asking another participant to pay attention to another element that is part of the experience, often relating to its position or some describing characteristic. This action has been generalized as pointing and has also been separated into three similar tasks with the purpose of analyzing whereas the type of element being pointed at would change the behavior of the participants. These tasks would ask to point at a digital construct created by the app, at a physical or real object in the real space used for the augmented experience and at another person participating in the experience with them, another human user.

The second category refers to the actions the team could be doing simultaneously while interacting with the digital constructs provided by the application. Most tasks identified for this category overlapped with common activities proposed in previous works, such as selecting, moving, and transforming objects. The collaborative context of this study would not change substantially the way participants would approach any of those actions. To gather more pertinent data, we only tasks related to sharing objects between team members were selected.

There is also a proposed distinction between a shared and a private in which each member of the team can interact with different digital constructs, a concept similar to the ideas discussed in works such as Reilly et al. (2014) or Lebeck, Ruth, Kohno, and Roesner (2018). The primary objective of this addition was to identify how the participants would visualize and interact with these spaces and to investigate whether the different scenarios would influence their ideas and responses.

The final category elements common to team management (Zigurs refers to & Munkvold, 2006), those that could benefit from AR augmentation, or at the very least, be compatible within an AR setting. This assumption was particularly relevant given the decision to maintain synchronous team interactions within the same physical space. This category also considers actions done by an individual managing an activity for several groups at the same time, like a teacher or a coach guiding the experience.



Table 1 Proposed collaboration tasks in AR

To facilitate the data gathering process, a simple application was built to allow the participants to visualize animations representing each task. This also allowed a more realistic sense of how the activities would be carried out using the assigned device. The participants were able to only see the animations but not interact with the digital constructs in any way. They could also repeat the animation any number of times and move between the animations of different tasks at will. The application worked in the exact same way in both scenarios. Figure 2 shows some examples of what the participants could see in the application.

Imagen de la pantalla de un video juego

El contenido generado por IA puede ser incorrecto.

Figure 2 Examples of the user view of the data gathering application

Each participant was assigned one of the two scenarios (smartphone or HoloLens), alternating between the options to ensure a balanced distribution. Each participant was given a brief introduction and an explanation of what was expected of them by the researcher. Additionally, they were given a brief tutorial of the basic usage of each setup. Special emphasis was placed on showing how they could put their hands in front of the camera to simulate an interaction, but they were still reminded that they could do whatever they thought would be most natural or straightforward to achieve the action presented to them.

All participants would then proceed through each task performing any number of interactions that would come to mind and trying to verbalize their thought processes and logic. Participants could propose any number of interactions. If none of the proposed interactions were air gestures, the researcher would remind them of that possibility. Tasks were always given to the participants in the same order to simplify the process of explaining the task context and managing the flow of the session.

The specific emphasis on air gestures is due to their strong association with AR technology, which often uses image recognition techniques to track the users’ hands and use them for interactions in a more natural and intuitive way (citation needed). It was also theorized that participants would not be necessarily familiar neither with the technology or those common interaction metaphors, so it was decided that an explicit mention of air gestures after the first round of proposals made by the users would help in the generation of ideas and would also help gather data concerning the hypotheses build for each scenario.

To capture the raw data, all sessions were recorded with a video camera for future analysis. The recordings took place over several weeks and in different spaces, sometimes with the participants seated and some other times with them on their feet, but always with the same equipment, the same test software and following the same procedure.

Another important decision was to gather data of each participant individually. Although the interactions were associated with group activities, the was no need nor the process gained anything by simulating the full teamwork experience. On the other hand, the process would require a more complex process, the coordination in time and space of several participants and for the mock-up software to also grew in complexity to accommodate several uses. It was deemed as an unnecessary development. The procedure clearly explained to the participants that the activity meant to simulate a teamwork scenario with other humans, and some measures were taken to add to the animations presented by the software simple representations of other users to give the participants a visual idea of the scenario presented to them.

Each of the recordings underwent analysis to create a transcript of all the participants’ comments and ideas, as well as to identify common thoughts, problems, and solutions. Video footage of each session was processed to compile a library of all the gestures performed for each task. Each individual performance of a gesture was given a score from 1 to 3, with the highest score given to the first gesture proposed by a participant. If a participant proposed two or more gestures or proposed a gesture after being prompted by the researcher, those gestures were given a lower score. This means that gestures with a higher score were performed the most and could be interpreted as the most natural or intuitive reaction for most of the participants. All gestures after the third proposal received a score of 1.

A total of 478 gestures were proposed for all tasks. Although each task clearly converged to a specific set of gestures, agreement in general was low. For each task, agreement was calculated using equation 1 (Wobbrock, Aung, Rothrock, & Myers, 2005).

In this equation, Pt is the total number of gestures performed for a task, and Ps is the number of times a distinct gesture was performed by different participants. This gives a value of agreement for each task up to 1, which would mean that every participant

proposed the same gesture for the task, while a value closer to 0 means that every participant proposed a different gesture. Figure 3 shows the level of agreement reached for each task and in each scenario in descending order.



Figure 3 Total agreement scores for each task, separated by scenario

Overall, tasks 1, 12 and 9 had the highest agreement, corresponding to pointing at an object, pausing your participation, and voting for an option on a menu. Although these tasks are the most prominent examples, all of them behave in the same manner, with most of the participants agreeing on one or two gestures, but the rest proposing a very granular or unique set, generally bringing down the agreement score significantly for all tasks. It is also important to note that the agreement score for both scenarios aggregated is a bit lower than when separated by scenario. Tasks 8, 9 and 13 had a better score in the mobile scenario, mainly because they were strongly associated with touch actions on the screen and most participants converged to such interactions. On the other hand, the HoloLens scenario had better scores for tasks 1 and 4 (point and hold and object), probably because they suggested obvious or natural hand movements to which most of the participants converged.

### 5.1.1.2 Proposed Interaction Gestures

Based on all the data provided by the participants, a library of 39 gestures was selected as the final user-driven selection. These gestures cover 56% of the total proposals and the set of each task represents between 50% and 70% of them. The following section will show the final agreement score for all the unique gestures identified in each task, as well a visual example of the gesture selected for the final set collection, the one that exemplifies the best the suggestion to which the users converged at. It is possible to check a library of the video captures for all the gestures identified in each task in Annex 1.

Tasks 1, 2 and 3 are related to pointing elements of your environment to the team

members. These tasks were the ones with the highest agreement scores in both scenarios.

Participants defaulted to pointing with one or two fingers towards the camera

in the general direction of the digital object or trying to cover it by superposing the

gesture over the object. Figure 4 shows the agreement score for the 11 unique gestures identified for the task.



Figure 4 Agreement scores for unique gestures in Task 1

The selected gestures involved pointing directly to the object with one or two fingers, with a stronger tendency for this gesture in the headset scenario, while mobile users defaulted more to tapping motions in the screen. The most important reasoning given by the users behind the tapping behaviour was that the motion was “natural” or “common”, especially for a touch screen. In scenario 1 some users also mention the naturality of the tapping motion in relationship to how they would select an option in a touching screen, highlighting the influence of previous knowledge when translating the actions to a scenario that is similar or somewhat familiar. The tapping motion also included stronger deviations of the gesture, like using the whole hand or exaggerating the pointing motion of the object to make clear the

intended selection and eliminate ambiguity in the selection, a concern mentioned by some users.

Another relevant gesture proposed was a variation on grabbing or holding the object to signify to others that “this object here that I am shaking is the one I am talking about”, to use the words of one user. The mechanics of holding the object varied a lot between users, sometimes using one hand or both, and with or without movement involved after the grabbing motion.

It is also important to mention that the task was described to the users as “pointing” to an object, which could have predisposed the users to a particular set of gestures. The task was described as such to convey the idea that the action was done for the benefit of another user, and that the main objective was to highlight the object to another person, not only to select the object. Nonetheless, a more neutral term could have been used, like “show”, “mention” or “indicate”. Figure 5a shows an example of the selected gestures for task 1, Figure 25a depicts T1G1: Pointing at the object with the index finger, while Figure 5b shows T1G3: Taping the object with the index finger.



Figure 5 Point at an object – Pointing Index (left) and Tapping Motion (right)

Task 2 yielded very similar results. Participants tended to use more descriptors or directions related to real-world elements. Phrases such as “the one next to the chair” or “the one next to your hand” were common, but these comments were always accompanied by the same pointing gesture. Figure 6 shows the agreement score for the 15 unique gestures identified for Task 2.

Gráfico, Gráfico de barras

El contenido generado por IA puede ser incorrecto.

Figure 6 Agreement scores for unique gestures in Task 2

A significant number of users defaulted to use the whole hand for the pointing gesture, but it was not possible to identify a reason for such a difference in comparison to Task 1. Figure 7 shows and example of this gesture. Another somewhat common proposal was to draw some figure to enclose the point in space being pointed at, like a circle, an X or some movement of the fingers.



Figure 7 Point at a place – Extended hand gesture

A similar situation presented in Task 3 when pointing to a person. In this scenario, the extended hand gesture was explicitly explained by the participants, who expressed reservations about pointing at someone else with a finger, considering it disrespectful. They preferred to point using the entire hand or by tapping on the screen. Otherwise, the same tendencies shown in the previous two tasks were also present in here. Figure 8 shows the agreement score for all the 13 unique gestures identified for Task 3.

Gráfico, Gráfico de barras

El contenido generado por IA puede ser incorrecto.

Figure 8 Agreement scores for unique gestures in Task 3

To wrap up the observation of the pointing category, it was observed that users defaulted to point to objects using an air gesture with either the index or the whole hand. Users also commonly referred to their previous experience with touch screens, founding natural the selection of objects using a tap in the air or the screen. There was not much difference between pointing digital or physical objects, but several participants had second thoughts and concerns when pointing people, considering the index gesture as rude or inappropriate.

Tasks 4, related to holding and manipulating an object, was designed to examine how the two scenarios would influence the participants’ interaction proposal, especially considering that Scenario 2 provided the freedom to use both hands. There were indeed instances of gestures only appearing in Scenario 1 and others only in scenario 2. The use of mobile phones led some participants to default to a familiar drag gesture over the screen to hold and move the digital object, which was only possible in that scenario.

On the other hand, the headset scenario strongly promoted the participants to use the perceived physicality of the digital object to try and grasp its contour using the index and thumb or even with the whole hand, as shown in Figure 9. The next most popular gestures used a simpler approach of grasping the object by completely closing the hand over it or pinching it with the index and the thumb. Nonetheless, participants strongly defaulted to the most complex yet more realistic approach when using a headset.

Figure 9 Hold an Object - Closed Fist (left), Contour (middle) and Pinch (right) gestures

Task 5 is the collaborative take of the object manipulation category and was meant mostly on gathering ideas from the users on how to organize the group while manipulating objects. The participants were specifically prompted to explain how they would share a held object or the act of sharing anything in general. In this case, both scenarios converged to the gesture of extending an open hand in the direction of the intended target, either to automatically trigger the sharing action (a very straightforward, abstract action) or as the first step of a process, expecting the other team member to extend their hand and take the object (a more realistic, physical action). Mobile users also converged to the motion of “throwing by swiping” the object

in the direction of the intended target. Examples of all these gestures can be seen in Figure 10.



Figure 10 Share an Object - Pass over (left) and Toss (right) gestures

For these gestures related to sharing and collaborative manipulating objects the participants proposed increasingly more complex gestures requiring full movements or patterns to convey the intended meaning (which are harder to show in static figures). Movement ranged from simple patters like swipes or tosses that only required a direction or a reference to the target team member, while others proposed complex series of movements normally involving the other user in the action, with specific protocols required to initiate the sharing of an object and its receipt from the second user. Figure 11 shows the agreement scores for all the gestures identified in Tasks 4 and 5.

Gráfico, Gráfico de barras

El contenido generado por IA puede ser incorrecto.

Gráfico, Gráfico de barras

El contenido generado por IA puede ser incorrecto.

Figure 11 Agreement scores for unique gestures in Task 4 (top) and Task 5 (bottom)

Tasks 6 and 8 corresponded to taking an object in and out of the private workspace of the participants. The shared and public space figure was an interesting concept that appeared in literature to emphasise the need to differentiate between public and private spaces in shared experiences. The shared versus private workspace create a boundary for each user to be able to keep work and experimentation as private work and be able to control what is shared with the team and what is not. In terms of interaction design it also offered the opportunity to understand how users would understand this concept, if at all, and if it was possible to converge into a set of gestures to represent possible actions needed for this scenario. It was decided to focus on the process of moving from and into the private space and how can this be signalled in the shared experience. In this case, both data gathering scenarios clearly converged to different gestures.

Both were similar in nature and consisted of moving the object with one hand to and from the area shown to participants as their private workspace. The difference appeared when participants in the mobile scenario saw a zone in the border of the phone’s screen and dragged the object there using a hold-and-drag gesture. Few people took the object with their hand, and if they did, they did not consistently use the same gesture proposed in Task 4.

On the other hand, headset users saw a zone in their physical space and moved the object there, sometimes holding it using the exact same gesture proposed in Task 4, but not consistently. Other times they would use a completely different gesture to hold and drop the object or would use a different one related more to push the object into the zone or drag it close to their own person. Figure 12 shows examples of the converged gestures for both tasks.

Figure 12 Put an Object in Private Space - Fist hold Inwards(left), Pinch Hold Inwards (middle) and Screen Swipe (right) gestures

This distinction between scenarios is very probable due to the way the private workspace was shown in each scenario: a 2D space on the screen for Scenario 1 and a 3D area near the user for Scenario 2. This could have prompted the participants to use the screen rather than an air gesture. Nonetheless, it is also probable that the screen interaction would have been one of the most prominent results even if both scenarios would have shown the same representation of the tasks, considering how for tasks 4 and 5 mobile users also defaulted to interactions using the screen despite the mostly 3D representation shown to them in the application.

The idea behind Task 7 was not necessarily collaborative in nature, mostly exploring a common action in any other software that could be expected by users in the functionality of CollabAR. Its implications in the coordination of action when manipulating objects could also be discussed. In the end, Task 7 represented one of the less agreed upon sets, and showed a variety of approaches to signal the process of undoing the previous action, but the most agreed upon comment among both scenarios was that such an action could better be implemented via a button or option in a menu, and not linked to a gesture, either in the air or the touch screen. This is important because this idea became more prevalent in the following tasks. Figure 13 shows some of the most common gestures proposed for the task, while figure 14 shows the distribution of the agreement score for all 22 gestures identified.

Mano de una persona con los brazos extendidos

El contenido generado por IA puede ser incorrecto.

Figure 13 Undo a Task - Push Sideways (left) and Wave (right) gestures

Gráfico

El contenido generado por IA puede ser incorrecto.

Figure 14 Agreement scores for unique gestures in Task 7

Tasks 9 and 10 asked the participants to vote in a debate or answer a multiple-choice question. The tasks were designed to identify differences in interaction when the participants had to interact with a clear user interface in comparison to a more physical selection involving objects in the digital space. Curiously, and due to the decision of keep the order of all the tasks the same for every participant, observations regarding this issue already appeared in tasks 6 and 8 when presenting headset users with a more physical representation of the private workspace in comparison to the UI representation shown to the mobile users. With these tasks, results were consistent among both scenarios and behaved as initially observed.

Task 9 showed a static menu on screen to represent the voting process, and as expected, it converged to simple selections over the user interface, normally a tap in the screen or a pointing gesture in the headset. Major distinctions consisted in changes in the pointing action, like using two fingers instead of one. The biggest variants used exaggerated version of a tap, like a stamp with the whole hand or a punch. Figure 15 shows examples of the selected gestures for this task.



Figure 15 Vote in a Static Menu - Air tap (left) and Screen tap (right) gestures

Task 10 promoted more variety. Pointing was still the most common gesture in both scenarios, probably because the 3-dimensionality of the objects promoted air gestures, even in Scenario 1. Holding the selected object was the other most common proposition. Like Tasks 6 and 8, the holding gesture was not consistent with the ones used in Tasks 4 or 5. Yet, both scenarios converged to the same holding gesture, particular for this task. It consisted of holding the contour of the object from the top. Because in Tasks 4 and 5, the holding gestures were accompanied by the expectation of having to manipulate the object in some fashion, participants probably preferred a more precise grasp that allowed easy rotation and inspection. Task 10, on the other hand, only required to signal the selection of the object, and merited a simpler hold. Figure 16 shows examples of the selected gestures for Task 10.



Figure 16 Vote with a Digital Object - Contour Hold (right) and Object Tap (left) gestures

We can also speculate that the way the objects were shown to the participants affected the decision-making process. The objects were anchored to the table or surface where the visualization was taking place, which probably promoted a natural reach-and-take motion. If the objects were floating in space or shown in a more abstract manner, reactions might have been different. Figure 17 shows the agreement score distribution for both version of the voting tasks and all the gestures identified.





Figure 17 Agreement scores for unique gestures in Task 9 (top) and Task 10 (bottom)

Tasks 11 through 14 presented activities common in managing a group such as organize a conversation, ask for attention, or signal a pause in your activities. These tasks presented the higher variety of proposed gestures, mainly due to their abstract nature and because participants considered they represented more a process than a simple action that could be triggered by a gesture.

Tasks 11, 12 and 14 converged to simple air gestures that commonly represent their respective actions in real life. Asking for attention, either for teammates (Task 11) or for every group (Task 14), was represented by raising a hand, often with the variation of waving the hand. Signalling a pause in participation (Task 12) was represented by showing the palm of one hand in a stop motion. Figure 18 shows examples of these gestures.

Figure 18 Raising hand gesture for Task 11 (left), Task 12 (middle) and Task 14 (right)

All these cases are good examples of proposals that would pose a conflict in an actual interaction design because they are practically identical, the main variation only being the height of the gesture or what side of the hand is shown to the camera. Among several other proposals for Task 4, Task 9 and Task 10, there is a conflict between natural, significant gestures, gestures detectable by the technology and complex gestures that make sense to a human but would add to much complexity to the process guided by a software tool.

Another interesting observation for these tasks was that a lot of participants also converged to a form of voice command to ask for attention or to communicate any status of the activity, arguing that it would be more natural to simply talk with your teammates if they are near, rather to rely on the software to control the whole action.

Task 13 represented an action meant to convey order to a group discussion, and a form of control of who was speaking in each time. Results behave similarly to the previous examples, with participants arguing that the tasks fitted more into a process than a single triggering gesture, effectively dividing the task in two distinct steps: a gesture for stopping someone from talking, then another for giving someone the turn to speak. Few participants tried to give a gesture that encompassed the whole process, and it was always a composite one, made of distinct gestures or movements for each part. The gestures for stopping people from speaking were similar to the gestures for Task 12 and pointing at people was the preferable method to signal the turn of speaking, like the gestures for Task 3. Figure 19 shows the agreement scores for all the gestures identified in Task 13, while Annex A contains video records of all gestures, especially for those that are hard to show in printed form.



Figure 19 Agreement scores for unique gestures in Task 13

### 5.1.1.3 Analysis and Conclusions

Participants also provided an extensive catalogue of ideas and commentary that accompanied the proposed gestures. These ideas and general conversation provided were compiled and analysed using transcripts or the recordings. The main ideas discussed were organized in major categories that contextualized each participant’s proposal or that promoted some form of discussion around the methodology, the setting, or the technology.

The first major element identified was related to all the things that influenced in some way the decision-making of the participants, and that prompted or facilitated the same types of gestures or the same type of ideas and issues to arise. For instance, participants assigned to Scenario 1, which used a smartphone, defaulted immediately to use the screen as their main proposed gesture of interaction. By far, the major intervention that the researcher had to do during the data gathering was to remind participants in Scenario 1 about air gestures, and when participants referred to some gesture to be “natural” or “logical” to solve a task, most of the time they were performing a touch gesture in the screen.

Of all the participants interviewed, 60% reported not being familiar with AR technologies, which can translate to most of them feeling more comfortable or seeing as more logical to use a known touch gesture in the screen or UI elements to perform any task. Even more, people that reported to be knowledgeable of AR also reported to understand how air gestures work and how they are recognized by the technology, as well as being aware of the limitations it possesses, which derived into preferring touch interactions or UI elements because they recognized air gestures as impractical or too difficult to be recognized.

In comparison, headset users in Scenario 2 had no other option than to propose air gestures due to the nature of the technology. We can hypothesise that given a more recognizable mean of interaction, like a controller with buttons, participants would have also defaulted to use it. An interesting aspect about the air gestures proposed by headset users is the realism of the actions. Hold gestures, for example, would consider the size and contour of an object and its position in space.

Finally, the way tasks were presented to the users through the application also affected the proposals. Some of the participants based their gesture mainly on the animation shown to them, like the movement of objects or some visual effects meant to signal or represent things, like a selection or someone performing an action. Similarly, any UI element predisposed the participants to think in UI interactions. Closely related to this, some participants assigned to Scenario 2 where completely fixated to the default select gesture of the HoloLens headset (a pinch with the index and thumb), which had to be taught to facilitate the interaction with the device.

Independent of the scenario, some other common situations were present when participants tried to explain or justify a gesture during the think-out-loud activity. The most common discussions emerged when the participants themselves found issues or flaws in the hypothetical scenarios created by them to explain or understand the gesture proposed. These were very valuable pieces of information because they showed how the user expected something to work and how they would use it in a real scenario. Some outstanding

examples were:

* People feel awkward mediating conversation by technology. The extra steps in the communication process were seen as a problem or in need of too much justification to feel natural.
* People expect the technology to detect a lot of social cues used in normal communications, such as sounds, gaze, and facial expressions. These cues also vary a lot between cultures.
* People who are not familiar with AR interactions feel more comfortable with gestures that mimic actions in real life or convey an understandable meaning, despite more abstract or simple gestures being easier to use. In contrast, the more familiarized the participants were with the technology, they expect simpler and more concrete gestures to be available (like touches, taps or drags).

This user-driven design exercise gathered 24 participants to identify natural air gestures

for interaction with an augmented reality application in a collaborative setting. The participants proposed a total of 478 gestures to cover 14 different tasks related to interaction with teammates and sharing a common goal. Of those gestures, 195 were unique propositions organized based on level of agreement to make the final selection of 39 gestures, which represent what the participants considered the most natural means of interaction with AR technologies either using a smartphone or a headset device. The resulting set of interactions is a useful tool to start thinking in the design of a collaborative AR application and the issues that must be solved in a concrete solution such as CollabAR.

This design approach proved to be a valuable technique, especially suited for early prototypes of an idea or to understand the impact of new technologies. The results obtained went beyond the proposed gestures, and a valuable collection of information was gathered related to how the users perceive the technology, how they see themselves using it and the issues, problems and opportunities that could arise in different contexts proposed by the users themselves, and that could have escaped the scope of a development team.

This research was proposed under the premise that collaboration is an important aspect of the value proposal that AR can convey. With the advent of new consumer products available on the market, it is necessary to understand how to use this technology in a social context, and how the process of communication itself can change when similar tools gain more popularity and mass adoption. Communication itself is a big aspect of any collaborative task, and we still need to learn more about how interactive technologies mediate those communication processes, how to mitigate the issues that will arise and capitalize on the benefits. Particularly, given the context in AR that this project holds, it can be interesting to research and analyse what it means to augment communication, collaboration and social interactions.

## 5.1.2 Implications for the implementation of CollabAR

The previously exposed research effort had a broader objective of understanding and identifying natural approaches that users proposed to interact with AR technologies and how they related to the means of interaction currently provided by available products. Through the research it was possible to identify the complex interactions that users proposed when offered a hands-free form of control and how they related to the slowly building common language being developed around these types of controls. The metaphors and approached currently being developed in the industry are not yet widespread enough in the market or among the final users, which prompt users to have more complex or elaborated expectation of what the technology is capable to do.

On the other hand, when users are presented with a mobile, touch-based device (a smartphone), the situation reverses, and the users start to rely in their knowledge and previous experiences with touchscreens, expecting or defaulting to gestures and interactions that are the norm in the mobile design environment. The conjunction of these two scenarios creates a situation in which users could be completely unfamiliar with the technology and would need a substantial learning period to properly use an app, or could be very familiar with the device, which facilitates learning but also has to take into account expectations and behaviours that could either inform or clash with the current implementation.

Analysing this information in the context of CollabAR, it is possible to draw some important conclusions that informed the future design and development process:

1. From the selected set of gestures, it was possible to observe that both the touch screen and headset scenarios converged to gestures that were similar or compatible between each other. There were few instances of radically different proposals for each scenario.

The selected gestures where at the top or near the top of the agreement ranking and consisted mostly of simple static gestures (a tap or a point selection) or common movements (swipes, waves of pinches). The commonality among these gestures is that the work the same or at least very similar in both testing scenarios and can also be represented in a full air-gesture representation with little or no change. These first responses or initial variations to the first responses reflected a tendency for users to default to known and tried approaches to interaction.

This behaviour is supported by the design implementation followed by modern (as the writing of this document) AR-able products such as the Meta Quest 3 and the Apple Vision, which opted for simple and recognizable gestures like taps and pinches, even when the hardware is capable of robust hand recognition. Figure 20 shows some examples.



Figure 20 Interaction Examples with the Meta Quest (top) and the Apple Vision (bottom)[[1]](#footnote-1)

This also echoes the data recorded from the ideas and justifications in the think-aloud sessions with the participants. Users indicated that the felt that the actions that could be performed in a smartphone for similar or parallel actions were more natural or correct and would even indicate that they would expect or wish for a functionality in the same venue for the application to be proper or correct.

For the development of CollabAR this indicates that the interaction design can focus on a single set of gestures for both mobile and headset deployment, opting for a “touch screen approach” that can be equally implemented on a screen or in a floating style UI. This would present familiar interactions for all users and would simplify the multi-platform approach.

2. The data shows that air gestures work better as abstractions, even when strong hand-recognition capabilities are present. Several users reported how unnecessary it felt to perform a complex gesture in a 3D space when a simple touch on a screen or in a UI option would suffice. This was especially present in the more process-driven actions like voting in a debate or pausing the current activity.

Given this observed behaviour it was considered as an opportunity to simplify the general interaction design of CollabAR to consider only standard gestures. Although air-gestures represent and interesting development and research topic, for the overall goals of CollabAR would divert time and resources that would not properly translate into data for the proposed research questions related to collaboration and TEL. A simple set of gestures supported by a clear UI promoted a simpler production and a more straightforward learning curve for the users.

3. It is important to focus more on user-driven process rather than software-driven ones. Besides providing a set of representative tasks for the data gathered, the set of activities proposed for the research was to prototype different ideas for the future functionalities of CollabAR. Tasks like the voting process and asking attention from other team members were ideas extracted from literature and that could offer value to the students using CollabAR.

Especial attention was given to the qualitative data extracted from the participants in task which such characteristics, and although every task converged to a distinct set of gestures, general comments and ideas mirrored confusion and frustration when trying to understand or justify the proposals. The main struggle related to ling what was viewed as a complex process to a single trigger action. Concerns on how to implement such gestures in a real application were also spoken. The data also showed that some participant found impractical or simply wrong to communicate to other people through the application when they were shared the same space and were at talking distance.

Paired with information obtained during the testing stage of the first implementation process, this prompted a strong change on the approach for the design of several of the functionalities of CollabAR. It was decided that a better approach was to hint or discuss ideal structures for the collaborative work of the users, but without enforcing it. This prompted flexibility in the usage of the tool and would lower the disruption that the data gathering process could create in the workflow of the students during the course sessions.

In summary, the research process on collaborative gestures informed several aspects of the interaction design of CollabAR in terms of understanding the expectations that users have with the technology, the best road to approach the most physical aspects of the interaction design and how to promote the ideas at the core of learning goals of the application. In general, this approach can be summarized as:

* A focus on simple gestures that prioritize touchscreen recognizable actions.
* Rely on process guided by the users and not the app.
* Promote structures rather than enforcing it

The following section will then show the final implementation process of CollabAR, the objectives achieved through the two development iterations executed, issues and problems encountered, and an analysis of the results obtained and the road ahead for future research and development.

# 5.2 Implementation Process

## 5.2.1 First Development Iteration

The main objective of the first iteration was to create a deployable test for the shared AR experience. Rather than approaching any of the use cases, the objective was to test different technologies and run trials on how to create the multi-user interaction. In more concrete terms, four objectives were established for a successful first iteration:

1. Create the “room” concept based on scanning the physical space and providing specific configuration parameters.
2. Create the “participant profile” by identifying the most important data needed for configuration and information display in the room.
3. Open the room for a session and allow participants to connect, acknowledging the interchange of information over the network and the visualization needed on AR.
4. Test the creation and interaction with digital objects during the work session.

At the beginning, the process focused more on testing different technologies and understanding the deployment of an AR solution over a network. Once the most relevant aspects of the infrastructure were understood, development move to the prototype of main usage loop, which can be specified as:

* Joining a configured room and visualizing other participants in the session
* Creating a workspace for the session.
* Adding and interacting with digital objects.
* Pointing at digital objects, other participants and other physical elements in the room.

Each of these objectives were considered as major milestones to achieve during the iteration and were used as the main markers for progress along the six months of development that the iteration approximately took. The development of each milestone was followed by a set of unstructured tests to gather quick feedback and incorporating changes and adjustments that could feasible be developed in parallel to the advancement of the other milestones. Important points of changes that were not possible to quickly incorporate or that required a major pivot in development were logged for the next iteration or even tagged as analysis points for future work.

The following sections will explore the results obtained for each milestone and the data gathered in each testing phase. It is important to remark that the process itself was not as structured or as sequential as explained. Although there were a clear plan and a tangible set of goals guiding the process, each milestone was work upon and tested in parallel and at different point of development, keeping a constant communication loop with testers and stakeholders but without following any formal sequence of steps or staged process. The following analyses are shown in isolation to facilitated communication and exposition but work itself was more fluid and less sequential.

### 5.2.1.1 The room and the workspace

The concept of the room was defined as the physical space used to hold the work session, represented in the app as the network session that coordinates all the participants and the digital objects. The room itself is constructed and coordinated by several elements of the architecture (see section 4.3.3), and is not centralized at all, neither in terms of processes nor software components. Nonetheless, this architecture is transparent to the user, who only needs to understand the room as a digital space that can be entered to initiate the session, and that holds all the digital object need to work.

Two major principles guided the implementation of the room. In first instance, the room hints at structure without enforcing it, a major design focus in the development of CollabAR. The room indicates that a work session is happening, and to enter the room means to start focusing on the tasks and goals planed for the session. If a user is not working, then that person shouldn’t be in the room, but without leaving the physical space shared with the other teammates.

Following a similar line of though, hopping between rooms acquires the meaning of getting in and out of different workflows, each one compartmentalizing goals, resources and results, something hard or even impossible to achieve in a physical sense.

The second principle is related to anchoring concepts and ideas to physical elements. It is a principle also linked to flexibility, to make the users think about allocating and space suitable to work and to gather all the needed resources, but allowing the use of any type of space, not only a literal, physical room. Any space could be augmented to fit the needs of the work session. The idea of collaborative augmentation is also at play here, so that the users can stablish a link between the selected space and all the digital information selected, adding a layer of information to the space that can signify whatever meaning the users want it to have (a wall for ideas, another for issues and another for opportunities i.e.). Figure 22 show the basic configuration of the room that is create ant the beginning of the work session.



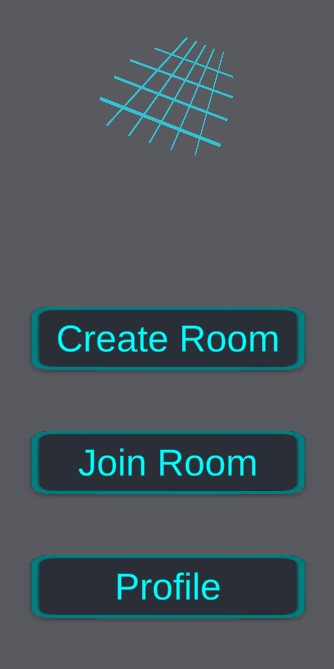
Figure 21 Basic Configuration of the Room

For the technical implementation it was also important to introduce the concept of the host. The implementation followed for Lightship required for a user to act as a host for the network session, since it was decided that the network architecture would follow a serverless approach to facilitate development.

There was the possibility to completely remove the terminology of a host from the side of the users since the concept was really only important for the technical implementation, but a couple of ideas made it worth it to use the term as part of the user experience:

* Users are familiar with the term in the context of video calls or video conferences (like zoom or Microsoft teams). The idea of “hosting” a meeting of that there is a person responsible for “making the call” is understandable to the users. Comparing the work session with a video conference could be beneficial for understanding the structure of the activity, although it also proved to bring some baggage with it as will be shown on chapter 6.
* It prompts the idea of designating roles to each participant and to allocate tasks and responsibilities.
* It made the process of creating and launching the room clear and easy to explain because a lot of known metaphor could be used (like the video call or being the host in a room or house).

As a matter of implementation, the host user only needs to select the room creation and provide a room name, which act as unique ID for that room in Lightship, as shown in Figure 22. On the other hand, a client user can search for all the available rooms created so far for CollabAR. In a more realistic scenario, in which several groups would be working at the same time, it would be necessary to implement more tools for users to filter and search with precision a room given its ID, as well as means to compartmentalize the creation of room inside bigger logical units like classroom or workshop, to create an easier and smother experience for several students at the same time using the application.

Interfaz de usuario gráfica, Aplicación

El contenido generado por IA puede ser incorrecto.Interfaz de usuario gráfica, Aplicación

El contenido generado por IA puede ser incorrecto.

Figure 22 Interface for creating a Room or joining an existing one

At this point the other responsibility of the host is to establish the goals for the session, although this step is not necessary for the activity to continue. It would be the responsibility of the role to manage the goals of the session, either by tracking their progress or by updating the goals as the session progresses. These activities, and in general the role of the host about the management of goals, are for the group to discuss, organize and implement. All the other responsibilities of the host, those related to the coordination of information through the network, are done at the backend of the software, and thus are completely transparent to the user.

With this process the room is created, and the users are now connected through the network, but the visualization has not started. Both host and clients need to scan with the camera the configured Target Image to anchor the room’s shared origin. Simple QR patterns are normally preferred for this step because they are easy to detect, but any image can work and helps to configure the room with any resource at hand for with an image that holds meaning for the group or the current activity.

Scanning the image will establish the origin point of the room for the local user and coordinate the current state of the room (other participants, play models and annotations). The origin point is also represented by a digital object that can be interacted with to get the status of the room. Figure 23 shows the emergent menu upon interaction with the room anchor. This is the view of the host, who can add and remove goals. It is also possible to give the option for the client to have the same ability to manage the goals of the session, once again giving the option to decide how much or little structure they want to implement in the group.

Interfaz de usuario gráfica, Texto, Aplicación, Chat o mensaje de texto

El contenido generado por IA puede ser incorrecto.

Figure 23 Room Information Menu

The final responsibility of the host is the configuration of the workspace. If the room anchor serves the purpose of indicating that the session in progress, then the workspace serves to delimit the interaction area of the room. The workspace is a simple plane that holds all the play models added to the room during the session. It helps the users to position of the objects in the space and focuses the attention of the group into a specific point, avoiding a typical problem in AR where is easy to “lose” objects or menus in the surrounding space when they are not directly on view (Brudy, 2013).

Other possible designs for the workspace were considered, specially from the technical implementation point of view:

* Using the detected planes in the room the way they were detected, relying on the output of the Plane Manager provided by Share AR.
* Using a default plane, or even a more complex representation like a table or a whiteboard. The model could originate from the target image similar to the Room Anchor.
* Letting the users position the workspace in the same way they would position an annotation, relaying on the AR Raycast Manager provided by AR foundation.

The final implementation used the plane manager but did not totally rely on the information detected from the physical space. In first instance, plane detection (as implemented by most AR frameworks at the moment of development) continuously improves the detail of the plane by iteratively merging the features detected in the space to get more complexity in the object. It was considered that the opportunity to have a quicker setup of the room con be beneficial in the context of the classroom.

Second, it was important to not rely completely on the physical layout of the room to encourage the group to set-up a meeting anywhere that was convenient or at hand. The only requirement that the environment needs to provide is for a plane to be able to be detected, which in the worst-case scenario can be the floor. Any detected plane can be then modified to the desire shape and be positioned in the space as a digital representation of the work area for the group, either matching a current table or just providing the space floating in the air. Other technical requirements must be taken into consideration when detecting a plane, but they are more circumstantial. For instance, good illumination and surfaces with a good contrast like colour, a pattern or an irregular surface are easier to detect than plane, monochromatic ones. In the end, a compromise was found between a quick setup of the workspace and giving the user the ability to adapt it to the needs of the session.

For the configuration process, the host can see the iterative recognition of any horizontal plane in the room and select one that suits the needs of the session. The host can also transform the selected plane by moving, rotating and scaling it until it fits into the desired shape and position, but that it is not a requirement for the process to continue. Figure 24 shows the host view for the plane detection stage and the fine-tuning options.

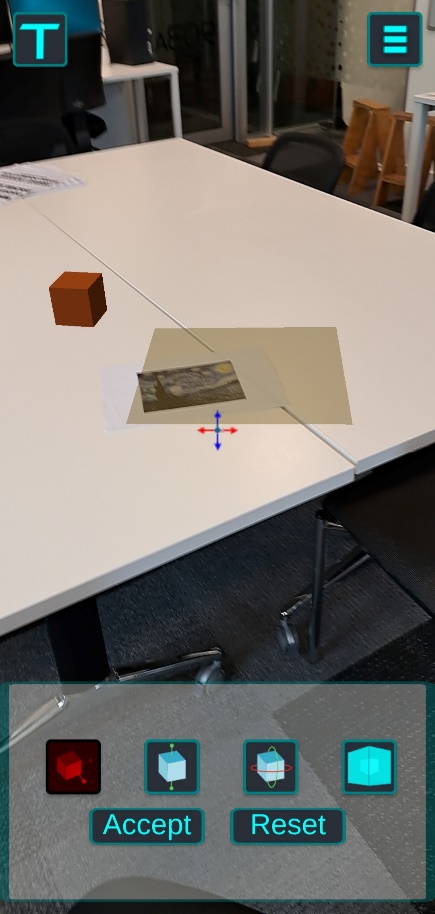


Figure 24 Iterative process for the plane detection (left) and workspace configuration (right)

With this the room configuration ends with the active responsibilities of the host, and the group can start the work session.

### 5.2.1.2 The User Anchor

With the main elements of the room represented it was also important to create that same element of visual representation in the digital space to each of the users in a session. For this there were two possible approaches:

* Recognizing the human face or figure and associate information of the participant to that detected feature.
* Add a digital avatar to the room that represents each participant.

There are important technological differences between these two options. For instance, face recognition is a common tool offered by most AR frameworks available but not at all categories of implementations in different devices (ARCore Supported Devices, n.d.), which could exclude potential hardware to be compatible with CollabAR. Incompatible hardware is not a surprising issue in this context, any technological product needs to consider the hardware ecosystem available in the market, but it was surprising to found how inconsistent was de compatibility of tools like AR core between different brands and different categories inside one brand. Elements like how new a device was, if it was a high-end or a low-end brand or the type of cameras available for the device had little relation to the presence of different AR features or not.

Another consideration regarding the face recognition option is data handling. There almost no issue for simple face recognition, that is, recognizing that a human face is present. Limitations and issues start to arise with more complex functionalities are needed. Recognizing and following a unique face, which would be the required functionality for this scenario, requires collecting and storing a significant amount of identifiable data for each test subject, which can quickly become impractical and an unnecessary ethical consideration.

This decision also became relevant thanks to the observed behaviour of test users during the AR gestures research. It was possible to notice that some users, those less familiar with AR technologies, had the expectation for the interface to behave in a more natural, complex way. The way the users are represented in the tool could also be part of those expectations, and for the user recognition to work as a seamless representation of each individual.

### 5.2.1.3 Pointing and interacting with digital objects

## 5.2.2 Analysis of the First Iteration

## 5.2.3 Second Development Iteration

## 5.2.4 Analysis of the Second Iteration

# 5.3 Conclusions

1. Images extracted from <https://www.youtube.com/watch?v=Exu7r2vZpcw> and <https://www.youtube.com/watch?v=dtp6b76pMak> [↑](#footnote-ref-1)