Enhancing user engagement through the user centric design of a mid-air gesture-based interface for the navigation of virtual-tours in cultural heritage expositions

Enhancing user engagement through the user centric design of a mid-air gesture-based interface for the navigation of virtual-tours in cultural heritage expositions

Vito M. Manghisi, Antonio E. Uva, Michele Fiorentino, Michele Gattullo, Antonio Boccaccio*, and Giuseppe Monno

Department of Mechanics, Mathematics and Management, Polytechnic University of Bari
Viale Japigia, 182, I-70126, Bari, Italy

*Corresponding author:

Antonio Boccaccio
E-mail address: a.boccaccio@poliba.it

Number of words: 6796 (References excluded)
8422 (References included)

Abstract

One of the most effective strategies that can be adopted to make successful cultural heritage expositions consists in attracting the visitors' attention and improving their enjoyment/engagement. A mid-air gesture-based Natural User Interface was designed, through the user-centric approach, for the navigation of virtual tours in cultural heritage exhibitions. In detail, the proposed interface was developed to "visit" Murgia, a karst zone lying within Puglia, very famous for its fortified farms, dolines, sinkholes, and caves. Including an "immersive" gesture-based interface was demonstrated to improve the user's experience thus giving her/him the sensation of "exploring" in a seamless manner the wonderful and rather adventurous sites of Murgia. User tests aimed at comparing the implemented interface with a conventional mouse-controlled one confirmed the capability of the proposed interface to enhance the user engagement/enjoyment and to make "more" natural/real, the virtual environment.

Keywords: Natural User Interface; Virtual Tour; User-centric Approach; Gesture Vocabulary Design.

1. Introduction

The Cultural heritage of a country represents a priceless patrimony not only for its inhabitants, to understand and explain the origin of customs and traditions but also for the touristic industry to attract visitors from abroad. Promotion of cultural, artistic, and environmental goods represents a crucial importance issue for a country as most of the touristic industry and hence, of the general economy, depends on how the financial resources are allocated for this scope. The use of new technologies, as well as novel interaction paradigms, were recognized to be one of the key-points to approach the mass audience and hence to adequately promote and present cultural heritage [1]. Among the technologies recently utilized to this purpose, Virtual Reality (VR) plays a role of predominant relevance [1–3]. By allowing the user to interact with the virtual world in an immersive environment, VR represents one of the most appealing and effective ways to improve the users' engagement and attract their attention on specific cultural heritage subjects.

It is a moral duty to share and spread the patrimony of a country among people and over time, but it is also a moral duty to preserve sites of interest and artworks although their large scale accessibility makes this task very difficult to accomplish. VR allows to replace the real visit with the exploration of reconstructed historical places and virtual museums [4], by means of virtual tours (i.e., immersive 360-degree panoramas [5]) and the manipulation of digital artifacts. Consequently, VR can provide both the preservation of these sites and the access of general public, and add all the potentialities related to the multiple virtual experiences, in terms of interaction and immersion [6].

Natural User Interfaces (NUI) that are at the basis of VR-based exhibitions, - such as virtual tours or virtual museums [7] -, are the object of the study of a large number of researchers throughout the world. For instance, the usability of gesture interfaces in virtual reality environments was investigated by Cabral et al. [8]. Navigation and selection techniques in virtual environments by means of gestures were studied by Dam et al. [9]. Fanini et al. [10], developed an engaging and shared gesture-based interface for visiting museums and exploring the treasures contained. Three principal

71 approaches are followed in the design of interfaces: (i) user-centric, (ii) iterative and (iii) centrist
72 analytical. The first approach utilizes an elicitation procedure where the user is asked to propose
73 possible gestures to execute specific commands/referents. Interface inputs are then designed on the
74 basis of the gesture proposals. Iterative and centrist analytical approaches design interfaces by basing
75 upon technological constraints and/or the experience of experts in the specific application field.
76 However, the user-centric approach represents the standard methodology in NUIs design [11] [12] as
77 it allows pursuing two important objectives: (i) lowering the cognitive load and (ii) improving the
78 user experience.

79 Touristic/cultural events are no longer simply places devoted to the exhibition of products,
80 customs, and flyers of a specific country but a privileged place to make culture accessible to the
81 largest mass audience. In such exhibitions, VR technologies can be conveniently utilized to 'immerse'
82 the user in virtual tours reproducing the typical environment of a country thus letting him feel as
83 physically present in 'that' place that can be thousands and thousands of miles away. In a previous
84 study [13], a multisensory system was proposed where a 20-foot container was utilized to host visitors
85 giving them visual, climatic and olfactory stimuli. In this study, we further developed this system by
86 designing and implementing a gesture-based interface for the navigation of virtual tours on a display
87 wall (the display was hypothesized to be fixed on one of four walls of the container) and investigating
88 the capability of such an interface to enhance the engagement/enjoyment of user's experience. In
89 detail, this interface was developed to "visit" Murgia, a karst topographic plateau of rectangular shape
90 lying within Puglia, very famous for its fortified farms and for being the seat of transhumance practice
91 in animal husbandry. A number of studies utilized gesture-based interactions with display walls
92 [3,14,15], but their application scenario is quite different from the navigation of virtual tours. Indeed,
93 scenes in virtual tours almost consist of spherical panoramas that users can explore by using a
94 zooming function, changing gaze direction or selecting active items. Other studies [16,17], also
95 reported in the literature, utilize VR technologies to carry out virtual tours but do not implement
96 gesture-based interfaces.

2. Research aim

From the review of the state of the art, it appears that no specific studies are reported in the literature focused on the design, the implementation and the validation of gesture-based interfaces for the navigation of virtual tours in cultural heritage exhibitions. In detail, no studies are available that adopt the user-centric gesture-elicitation procedure for the definition of the vocabulary of gestures capable of guaranteeing the users' engagement/enjoyment. Our hypothesis is that including an "immersive" gesture-based interface improves the user's experience thus giving her/him the sensation of "exploring" in a seamless manner the wonderful and rather adventurous sites of Murgia.

Therefore, the aims of this work are:

1. To describe the application of such an elicitation procedure to the design process of a mid-air gesture-based NUI;
2. To test the developed NUI and evaluate its effectiveness in improving the engagement and the enjoyment of users with respect to the classical mouse-controlled interface.

3. The gesture vocabulary design

The gesture vocabulary, i.e., the set of gestures utilized by the user to interact with the interface, has to be straightforward and intuitive in such a way that the resulting NUI satisfies the general requirements of the low cognitive load and the low physical fatigue [18]. The design process of the vocabulary was structured into three principal phases:

- The *Interface requirements definition* phase. In this phase, the set of commands that are the most suited to the specific scenario/application are defined. Three aspects must be considered in this phase: (i) the specific tasks that have to be performed, (ii) the environment/context where tasks must be accomplished, and (iii) the set of commands required to perform the tasks.

- The *Gesture elicitation* phase. Spontaneous gestures the users would intuitively use to trigger the interface commands are proposed and collected. The agreement analysis then follows.
- The *Vocabulary definition* phase. The gesture proposals, collected in the previous phase, are ranked according to their guessability. The Vocabulary of gestures will include, for each command, the most intuitive gesture among those proposed for that command.

Hereafter, the following definitions will be utilized in the present article:

- Referents: interface commands, in this study five referents were hypothesized to be necessary to conduct and control a virtual tour (see next Section 3.1);
- Gesture proposals: gestures proposed by the participants during the elicitation phase as interacting metaphors to "execute" a given referent;
- Vocabulary: a combination of different gestures devoted, each, to "execute" a given referent.

3.1 Interface requirements definition

The following five interface commands were hypothesized to be strictly necessary to properly conduct and control a virtual tour on a wall display:

- move the pointer on the screen;
- zoom-in;
- zoom-out;
- change gaze direction (i.e., the solid angle of the spherical pano visualized on the display);
- select items.

The context/environment where the virtual tour was hypothesized to take place is a container with an average space of 8 m^2 available for users. Containers are "portable" installations that can be easily transported to the location where cultural/touristic events are organized and properly equipped with all the devices required for a virtual tour. Following the hypothesis that the container can host up to 4 visitors simultaneously, the proposed virtual exhibition system was designed to switch the control

among users according to a defined policy. Users were hypothesized to be not familiar with gesture-based interfaces which were, therefore, designed as intuitive as possible.

3.2 Gestures elicitation procedure

In order to conduct the gesture elicitation process, a population of 29 participants (average age 21.3 years, SD=3.11) was first recruited, including 18 males and 11 females, all right-handed. All of them were students in Mechanical and Computer Engineering: five participants declared to use an Xbox Kinect for recreational purposes regularly; nine utilized the device a few times; fifteen never used it.

Before performing the elicitation, a preliminary investigation was carried out to evaluate the users' preferences on the two control modes that are available in the software suite *krpano* [19] utilized to implement the virtual tour. In detail, this suite includes two main components: (i) a set of tools to create and edit virtual-tours, and (ii) a viewer, embeddable into HTML pages, which allows the navigation of virtual tours using a web browser. In the default configuration, the tour is only sensitive to system events, such as those triggered by the keyboard and/or the mouse. Two control modes are available to change the gaze direction. The first one, which we will call as "drag&drop" mode, permits the user to grab the scene by keeping the left mouse-button pressed, and to change the gaze direction by moving it together with the mouse pointer. In the second control mode, which we will call as "move-to" mode, while the left mouse-button is kept pressed and the mouse is moved, a vector is defined, the direction and the amplitude of which allow controlling the changes of gaze direction and the speed of its movement, respectively. After explaining to all the participants the two control modes, they were asked to test both of them using the standard mouse-controlled interface. The test had a duration of about 4 minutes at the end of which participants were asked to express their preference. 27 participants (i.e., 93.1 % of the recruited population) preferred the drag&drop mode which led us to elicit gestures just for this control mode.

The elicitation procedure adopted a user-centric conscious bottom-up approach [11] where referents are first explained to the participants, then the gesture proposals for each referent are collected. The Wizard of Oz study-setup [20] was followed to carry out the elicitation procedure. The experimenter asked the participants to think of the possible mid-air hand-gesture she/he would use to trigger each of the five referents. Then, staying in front of the display wall (a professional UHD 85 inches Samsung QM85D monitor), - at the distance of 2 meters- showing a scene of the virtual tour, each participant was asked to execute the gesture previously thought for each referent. According to the Wizard of Oz study-setup, a hidden experimenter triggered the corresponding commands via the mouse thus giving the participant the impression of activating her(him)self the command via the executed gestures. All the gestures executed during this phase were recorded and analyzed. To this purpose, a desktop PC with a CPU Intel Core i7 6700 (3.4 GHz) 16GB RAM and a GeForce GTX 970 graphic adapter were utilized.

Fig. 1 shows and briefly describes all the gestures proposed for the execution of the five referents while a more detailed description of them is given in the Appendix A (see Supplementary Material). Acronyms were utilized to make more compact the notation. Fig. 2(a) shows the distribution of the gesture proposals, i.e., the number of times a given gesture was utilized to execute a specific referent. For each referent, the *Agreement Rate* (*AR*) was finally computed which is a measure of the cognitive load and, in detail, of the agreement that exists between the gestures proposed for each referent. *AR* ranges in the interval [0, 1]; for a given referent r_k, the value $AR(r_k) =$ 0 means that all the proposals collected for the referent r_k are different from each other, on the contrary, the value 1, indicates that all the proposals (gathered for r_k) are the same. Generally speaking, a low value of *AR* imply a high cognitive load and hence the necessity of re-designing the set of referents. The values of *AR* computed in this study were large enough, which led us to conclude that the five hypothesized referents do not require a high cognitive load. Further details on the computation of the Agreement Rate values are given in the Appendix B (see Supplementary Material).

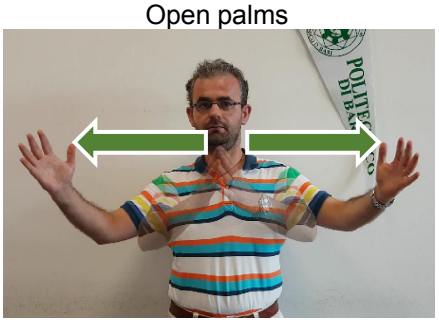
Move the pointer on the screen	**MHO**: Moving Hand with Open palm 	**MHIP**: Moving Hand with the Index finger Pointing 	
Zoom – in	**OHUP**: One Hand Unpinching 	**DTHC**: Distancing Two Hands with Clenched fists 	**DTHO**: Distancing Two Hands with Open palms 
Zoom – out	**OHP**: One Hand Pinching 	**BTTHC**: Bringing Together Two Hands with Clenched Fists 	**BTTHO**: Bringing Together Two Hands with Opened palms 
Change gaze direction	**OHPM**: One Hand Pointing and Moving 	**OHGM**: One Hand Grabbing and Moving 	
Select items	**OHPo**: One Hand Pointing 	**OHPu**: One Hand Pushing 	**OHC**: One Hand Clicking 

Fig. 1. Gesture proposals collected for each of the five hypothesized referents. The acronyms utilized to identify each gesture proposal are reported on the top of each picture.

It is worthy to note that 3 out of the 13 proposed gestures (Fig. 1) are almost the same: One Hand Pointing (OHPo), One Hand Pointing and Moving (OHPM) and Moving Hand with Index finger Pointing (MHIP).

3.3 Vocabularies definition

To select the best gesture for each referent, the *gesture guessability G* [21–23] was computed, which is a factor that 'measures' the intuitiveness of each gesture with respect to the corresponding referent. The *gesture guessability G* is given by:

$$G_i^k = \frac{P_i^k}{P_{TOT}^k} \; ; \; G \in]0,1] \qquad (1)$$

where P_i^k is the number of times the i^{th} gesture was proposed for the k^{th} referent and $P^k{}_{TOT}$ is the total number of the proposals collected for the k^{th} referent. It is worthy to note that while the Agreement Rate *AR* is a measure of the cognitive load required to execute a given gesture in response to a referent, the gesture guessability *G*, instead, is an indicator of the intuitiveness related to a gesture with respect to the corresponding referent. Ranking each gesture according to the guessability (Fig. 2(b)) led us to define the best vocabulary of gestures utilized to carry out the virtual tour (Fig. 3).

4. The NUI implementation

The hardware setup used for the NUI was the same as that described for the elicitation procedure with the addition of the Kinect v2 RGB-D camera that was utilized as a tracking sensor. This device was successfully employed in different domains of applicability that go from gaming [24] to ergonomics assessment [25] to the navigation of virtual environments [9] and complex 3D archaeological scenes [26].

The designed interface is based on a software developed using the C# language, the Windows Presentation Foundation libraries (.NET framework), and the Microsoft Kinect for Windows SDK 2.0. One of the essential and critical requirements of the proposed interface was to develop a gesture

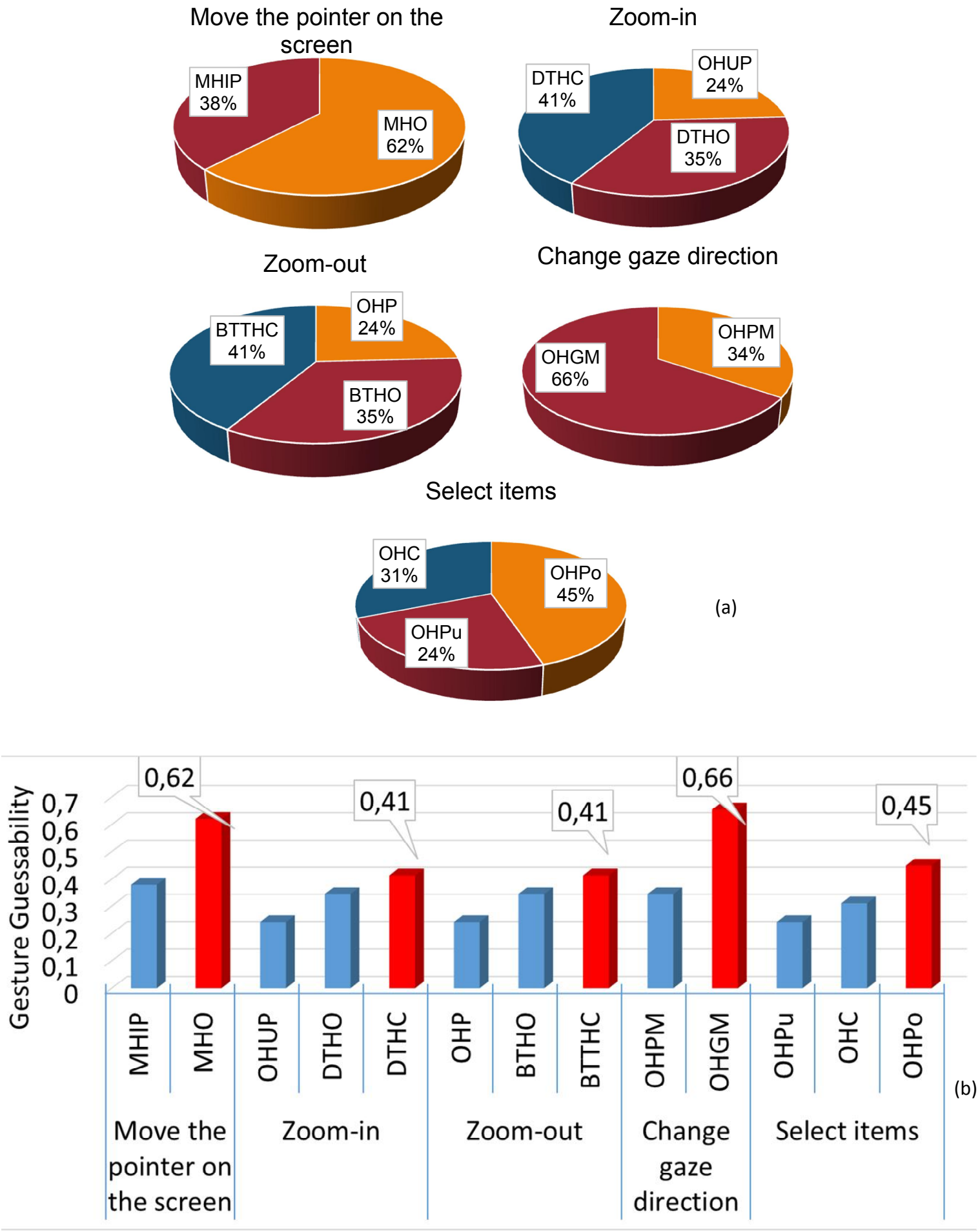


Fig. 2. (a) Percentage of gesture proposals collected for each referent. The acronyms utilized are specified in Figure 1. (b) Values of guessability computed for each gesture proposal. The histograms highlighted in red refer to the gestures with the highest guessability that were finally utilized in the best gesture vocabulary.

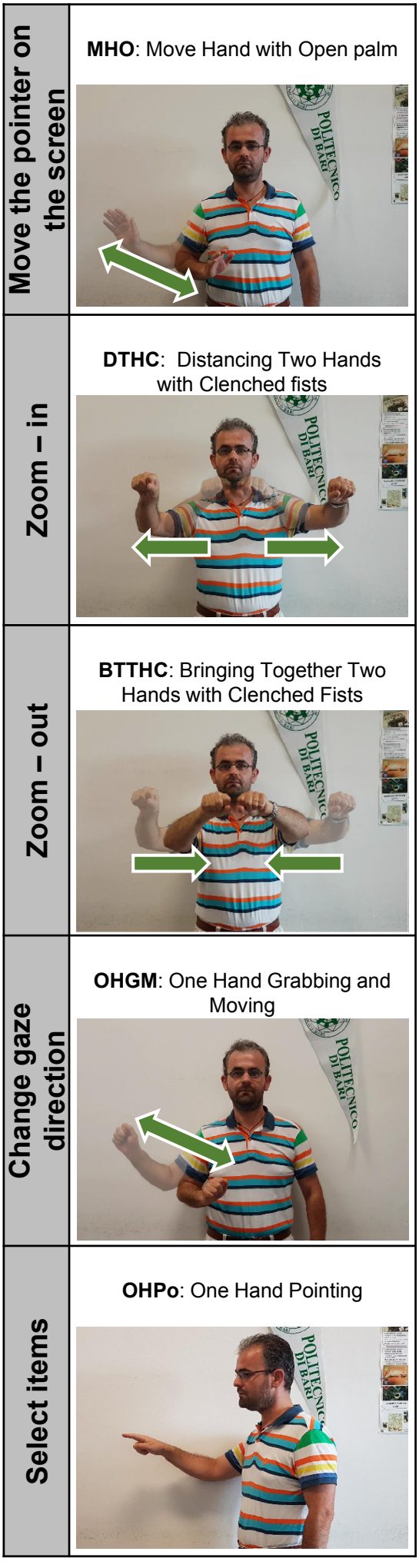


Fig. 3. Best vocabulary of gestures utilized to carry out the virtual tour.

recognition system reliable and capable of adapting to people differently shaped and sized (e.g., tall and short people, adults and children, left and right-handed users).

A control flow (Fig. 4) supervised the users' navigation, where the user's activity is considered as a state machine. The clock of this state machine is event-based and triggered by the arrival of a new frame from the depth sensor. The sequence of the main actions controlled for each new-frame is the following:

1. Definition of the user that is enabled to lead the navigation of the virtual tour, hereafter we will call her/him as the user-leader.
2. Definition of the user state. For each frame, the system has to detect if the user wants to execute a gesture, and if so, which action she/he intends to trigger on the interface.
3. Triggering of the detected actions.

4.1 User-leader definition.

Following the hypothesis that users will wander through the virtual tour in a standing position and based on the studies of the optimal position an user should occupy to maximize the accuracy and the reliability of the Skeleton Tracking algorithm for Kinect v2 [27], a "virtual area" was defined as the space located in front of the display wall, starting and terminating at 120 cm and 450 cm, respectively, from it (Fig.5). Amongst the visitors occupying the "virtual area," the user-leader is defined as the one that is the nearest to the sensor.

4.2 User's state definition.

The state machine that detects the state of the user-leader works through 5 states: unengaged, zoom, move the pointer on the screen, change gaze direction and select items (Fig. 6(a)).

These states are detected by evaluating, during the navigation, the state and the position of each hand of the user-leader. The Kinect v2 skeleton-tracking algorithm allows four states for each tracked hand to be identified: unknown, open, closed, and lasso (Fig. 6(b)). Based on the state of each hand and

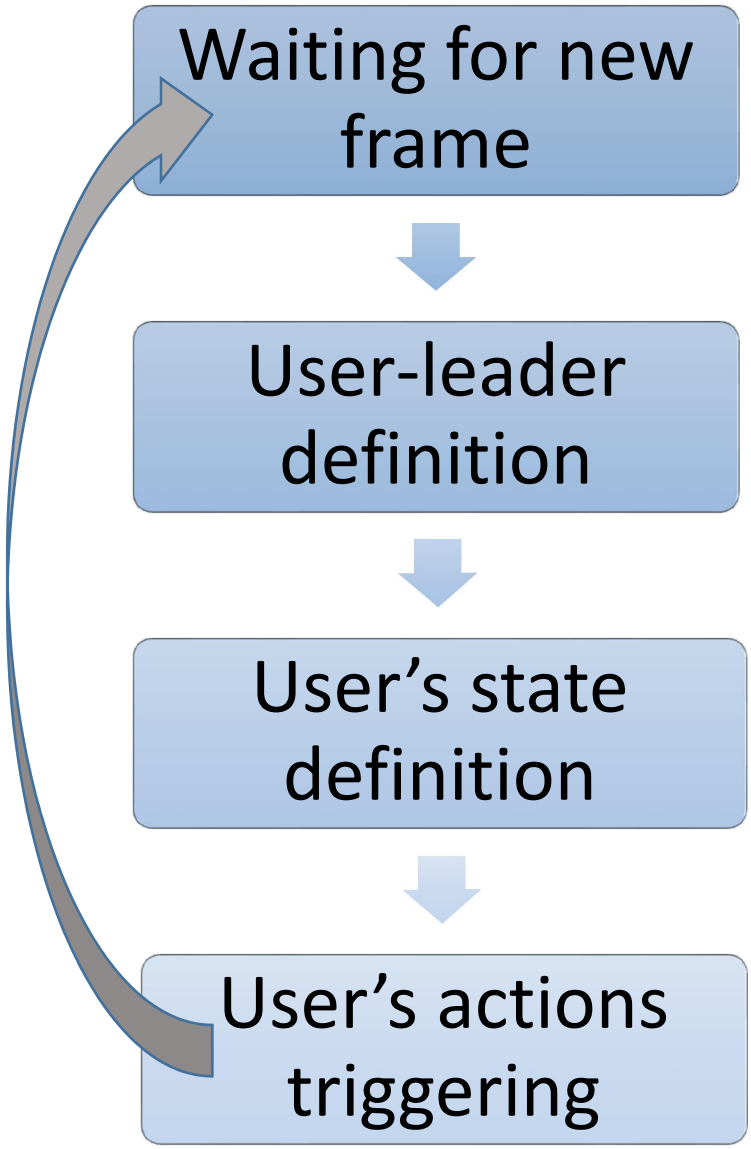


Fig. 4. Sequence of the main actions controlled for each new-frame detected by the sensor.

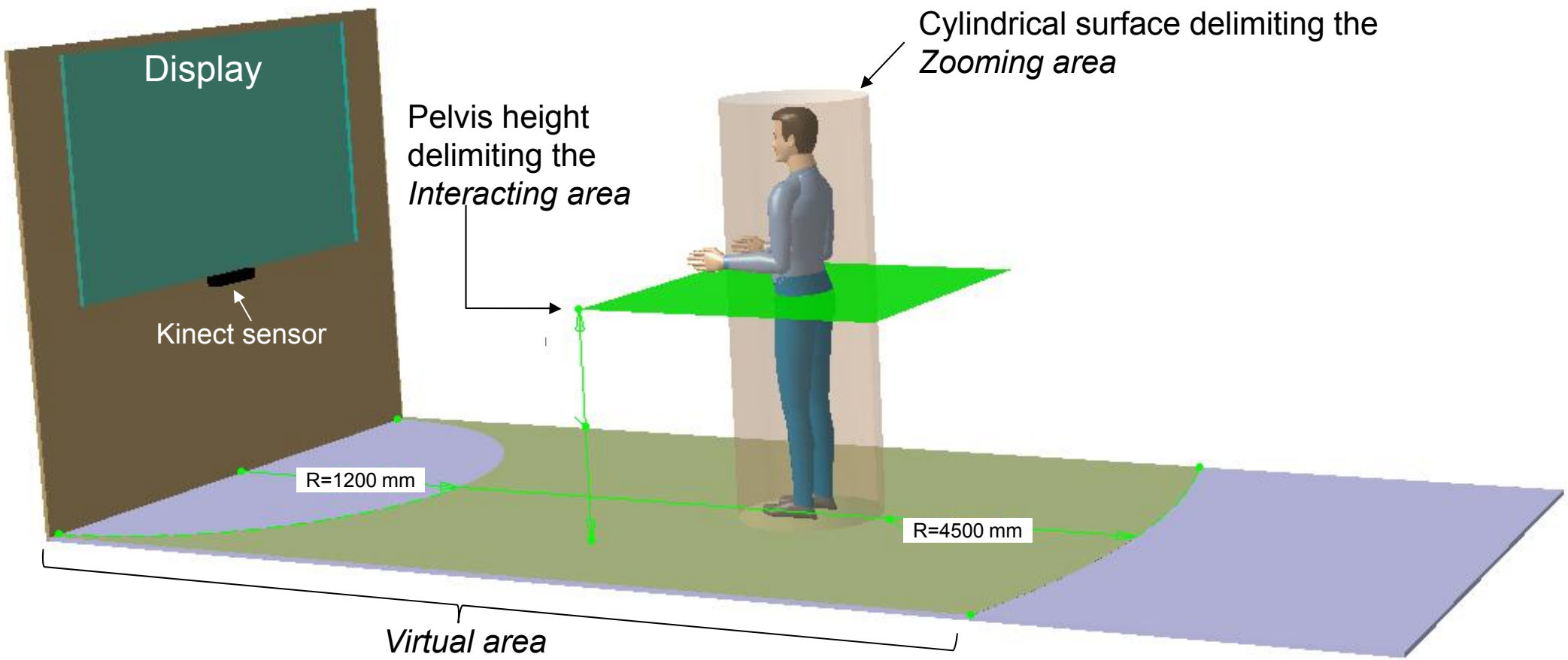
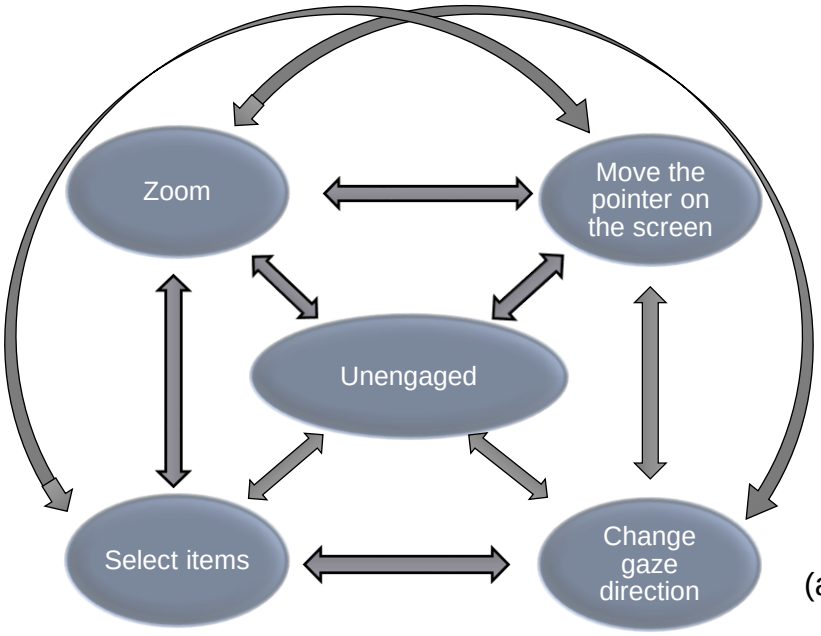
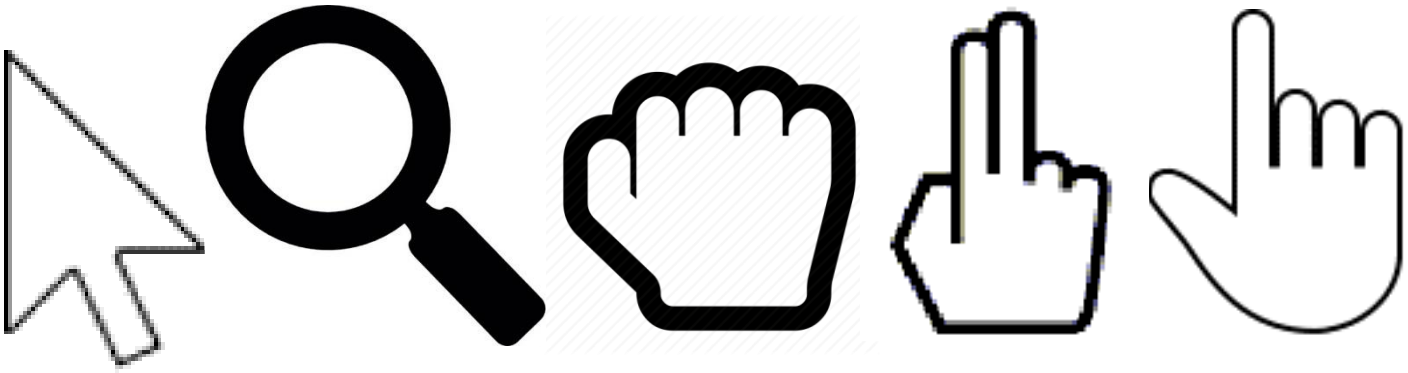


Fig. 5. Schematic of the working areas defined (with their limit dimensions) to control the user's interaction.



(a)

(1) (2) (3) (4) (b)



(1) (2) (3) (4) (5) (c)

Fig. 6. (a) The state machine that controls the user's behavior. (b) The hand states detected by the Kinect skeleton tracking algorithm for the tracked hands: b1) Unknown; b2) Open; b3) Closed; b4) Lasso. (c) Pointer icons shown on the screen and utilized as a feedback to make aware users about their current state: c1) Move the pointer on the screen; c2) Zoom; c3) Change gaze direction; c4) Select items; c5) Hotspots pointer on-over event.

241 the combination of the states of both hands, following a specific policy described below, the state
242 machine is identified.

243 To define the rules for the identification of the state machine, the following working areas were
244 defined.

245 *Interacting area*. Fixed an ideal horizontal plane (highlighted in green, Fig. 5) at the height of
246 the user's pelvis, the interacting area is defined as the space above the plane. If both the hands of the
247 user are below the plane, no interactions with the system are triggered. The contrary occurs once one
248 or both the hands are in the interacting area. The height of the plane delimiting the interacting area
249 was fixed at the level of the pelvis which is a good compromise between two opposite requirements.
250 A too "high" interacting area has the disadvantage to increase the fatigue related to the arm
251 movements, the so-called gorilla arm effect [28]. With a too "low" interacting area, instead, the risk
252 of unwanted interactions due to natural movements of the hands, can be run.

253 *Zooming area.* Fixed an ideal cylindrical surface around the user with a diameter equal to the
254 distance between the user's shoulders, the zooming area is defined as the space outside the cylinder
255 (Fig. 5). It is worthy to note that the designed interface allows the control from the right to the left
256 hand and vice-versa to be switched in real time. To this purpose, the control-hand was defined as the
257 hand that is kept at the largest distance with respect the floor. During the interaction, if the control-
258 hand is kept lower than the other for more than 1 second, the control switches to the other hand. With
259 this strategy, the user can change the control-hand in a natural manner and hence rest the tired arm,
260 but also the system is suitable for both, left- and right-handed users.

261 The state machine that controls the user's behavior starts from the Unengaged state (Fig. 6(a)).
262 In this state, the user-leader is not actively interacting with the interface. To start the interaction,
263 she/he has to raise one of her/his hands until reaching the interacting area (i.e., the space above the
264 ideal horizontal plane passing through the user's pelvis). When the user is engaged, her/his state
265 changes on the basis of the state of the control-hand and according to the following rules:

- If the state of the control-hand is open, the user's state turns to Move the pointer on the screen;
- If the state of the control-hand is closed and the state of other hand is open or it is outside the zooming area, the user's state turns to Change gaze direction;
- If the state of the control-hand is lasso, the user's state changes to Select items;
- If hands are inside the zooming area and their state is closed, the user's state changes to Zoom.

In order to improve the reliability of the system, thus avoiding any sudden and unwanted changes of the state (between the right and the incorrect one) of the hands, the strategic approach of the most voted policy was adopted. In detail, the state and the position of each hand in each frame were stored in four queue-like buffers (i.e., two queues for the state and two queues for the position). As soon as new frames are detected by the sensor, all the buffers are updated by removing the frame on the bottom and inserting the new one on the top. Then, the hand state is identified as the most frequent state in the buffer.

In order to make aware users about their interaction with the system, graphical cues were utilized as a feedback. The pointer icon (Fig. 6(c)) changed on the screen according to the actual state of the user. In addition to the pointers referring to the user's states: Move the pointer on the screen, Zoom, Change gaze direction, and Select items, a specific pointer for Hotspots on-over event (Fig. 6(c5)) was included.

5. User's actions triggering

The Move the pointer on the screen action allows the user to control the pointer position on the scene in front of her/him and can be triggered by executing the MHO (Moving Hand with Open palm) gesture (Fig. 1). To interact with the *krpano* viewer, a software simulation of the move-cursor event was utilized to trigger the actual action on the virtual tour. The pointer location is controlled by a ray

290 casting method, where the position of the head detected by the Kinect skeleton-tracking algorithm is
291 utilized as the projection center while the pointer position on the scene is determined by projecting
292 (on the screen) the ray that connects the head and the control-hand (Fig. 7).

293 Proper filtering methods were applied to minimize the jittering effects from which the Kinect
294 skeleton-tracking algorithm suffers. Regarding the location of the hands, a median filtering approach
295 was adopted, while regarding the head position, an accumulation method. This method stores, frame
296 by frame, in a queue-like buffer the position occupied by the head and assumes as the actual position
297 of the head the one computed as the mean value of all the positions stored in the buffer. As new
298 frames are acquired, if the computed mean values change more than a fixed threshold, then, the
299 position of the head is updated; if, conversely, the new computed mean value differ less than a fixed
300 threshold, then, the position of the head is hypothesized not to change. Although this approach allows
301 a perfect alignment between the cursor position on the scene and the target of the user's gaze, presents
302 two main limitations. The first is related to fatigue required to the user that, to control the pointer
303 position, has to keep her/his hands raised for a long time without rest. The second is related to the
304 interposition, between the user's eyes and the scene, of the user's control-hand. To overcome these
305 limitations (albeit in a partial way) the projection center was moved at a lower level (h=600 mm, see
306 Fig. 7) and towards the display wall (L=500 mm, Fig. 7). Adopting this solution, the user can control
307 the position of the pointer with smaller and at a lower level movements thus reducing the fatigue
308 required to execute them.

309 The Zoom actions allow the user to zoom-out and zoom-in on the scene and are triggered with
310 the execution of the *BTTHC* (Bringing Together Two Hands with Clenched fists, Fig. 1) and the
311 *DTHC* (Distancing Two Hands with Clenched fists, Fig. 1) gestures, respectively. The user acts as
312 she/he is at the center of a sphere on whose internal surface the scene is stuck. Therefore, to zoom-
313 out, she/he grabs the scene with her/his hands and reduces its dimensions by decreasing the distance
314 between hands. Conversely, to zoom-in, she/he grabs the scene with the hands and increases the
315 distance between them as if she/he is stretching the sphere. A software simulation of a mouse-wheel-

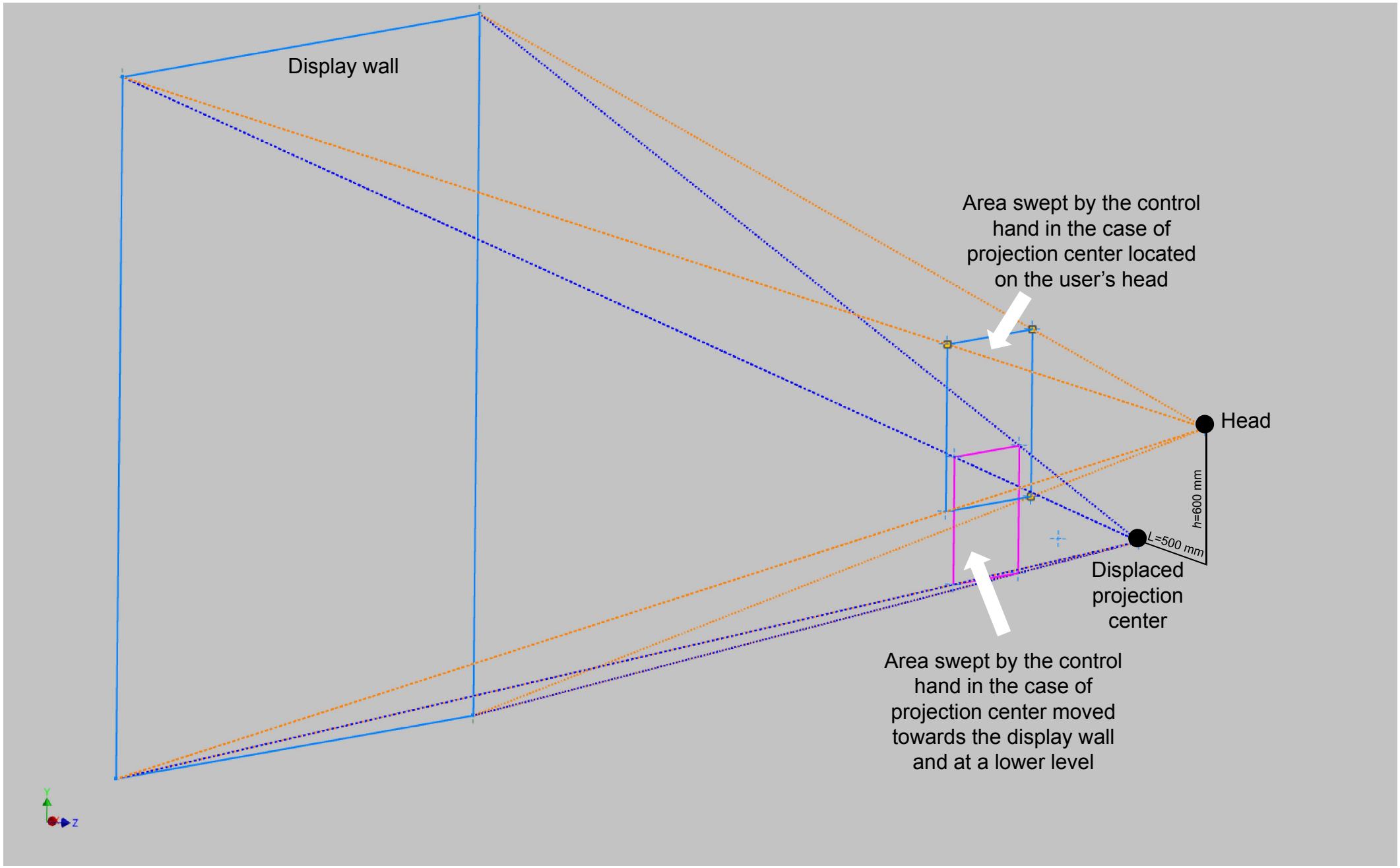


Fig. 7. Schematic of the ray casting method adopted to control the pointer location. The area (delimited by the magenta rectangle, in the case of projection center located in correspondence of the user's head) that the user's hand must sweep to roam the displayed scene becomes smaller (area delimited by the blue rectangle), when the projection center moves from the position of the head towards the display wall (by the quantity L=500 mm) and at a lower level (by the quantity h=600 mm).

316 scroll event triggers the actual action. The entity of the scrolling is proportional to the variation of the
317 relative distance between the two hands in two subsequent frames.

318 The Change gaze direction action that is triggered by executing the OHGM (One Hand
319 Grabbing and Moving, Fig. 1) gesture, allows the user to control the direction of her/his view. The
320 user acts as she/he is at the center of a sphere on whose internal surface the scene is stuck. Therefore,
321 to change the direction of view she/he only has to grab the sphere and drag it wherever she/he likes.
322 In this last action, the position of the pointer is controlled as in the Move the pointer on the screen
323 action. A software simulation of both a mouse left-button-pressed event and a move-cursor event
324 triggers the actual action on the virtual tour.

325 The Select items action allows the user to select/activate an item on the scene and is triggered
326 with the execution of the OHPo (One Hand Pointing, Fig. 1) gesture. To select an item (which is
327 indicated by the Hotspots pointer on-over event) the user has to move the pointer over the
328 corresponding hotspot and select it by pointing at it with her/his index finger. The Kinect skeleton-
329 tracking algorithm is tailored to detect the lasso state, where, both the index and the middle fingers
330 point. However, it was found that the algorithm detects the lasso state even in the case the user points
331 with the only index finger. A software simulation of a mouse left-button-click event triggers the actual
332 action on the virtual tour.

333 A short video-report available as Supplementary Material shows how the proposed interface
334 was utilized to navigate a virtual tour on Murgia.

335 6. NUI testing and evaluation

336 A user study was carried out to evaluate the effectiveness of the proposed gesture interface and,
337 in particular, to obtain a comparative evaluation in terms of perceived usability, user
338 engagement/enjoyment and overall users' preferences between gesture and mouse-controlled
339 interface. Given the strong legacy bias related to the use of the mouse-controlled interface, the
340 following three hypotheses were formulated:

H_1-The mouse-controlled interface will achieve a better or at least comparable score in terms of usability;

H_2- The gesture-based interface will achieve a better score in terms of enjoyment/engagement;

H_3- The overall users' preferences will be more oriented towards the gesture-based interface.

6.1 Experimental procedure

Two different experimental setups were adopted. In the first one, the user stood in front of the display wall at an average distance of 2 meters and navigated the virtual tour using the gesture interface. In the second experimental setup, the user kept the same position and wandered the tour by means of a wireless mouse controller using a desk as mouse support. Following Ayoub et al. [29], the desk upper surface was placed at 102 cm of height from the ground. Each participant tested both setups in two consecutive sessions. The order of execution was counterbalanced over users. In order to minimize any learning effect, a time of at least 15 minutes was awaited between the end of the first session and the beginning of the second one.

A total of 16 voluntary participants (13 males and 3 females, all right handed) were recruited among engineering, students, and researchers at Polytechnic University of Bari. The average age was 28.1 years (min 22 years, max 42 years, SD = 5.6 years). None of the participants ever interacted with a Kinect V2 sensor-based interface, 10 out of 16 had no previous experience with mid-air gesture interfaces while the other 6 had previous experience with the first version of the Kinect sensor.

Each session was supervised by an experimenter and consisted of a training- and a test-phase. In the training-phase the participant watched a video tutorial explaining the use of the interface under test and then had a free-training period of two minutes on a demonstrative tour (different with respect to the virtual tour of Murgia successively navigated). After a break of 3 minutes the test-phase started. In this phase, the user was asked to navigate freely the virtual tour of Murgia and to use each interacting metaphor at least twice. After a participant accomplished the minimal required tasks, she/he was free to interrupt the test phase. The experimenter recorded the time duration of this phase.

After each test-session, the participant was asked to estimate the time she/he spent to carry out the test and to fill in a usability satisfaction questionnaire [7] and a customized Intrinsic Motivation Inventory (IMI) questionnaire (the IMI questionnaire is freely available at http://selfdeterminationtheory.org/intrinsic-motivation-inventory, more details on the administered questionnaires are given in the next Section 6.2). Table 1 reports the independent and dependent variables of the experimental procedure.

After the participants filled in the questionnaires, the experimenter had an interview with them to gather their impressions of the navigation experience.

Table 1: Independent and dependent variables of the presented experiment.

INDEPENDENT VARIABLES		
Participants	16	13 males, 3 females
Interfaces	2	Gesture interface, Mouse-controlled interface
DEPENDENT VARIABLES		
Learnability		Mean of two answers on a 7 point likert scale
Interface efficacy		Mean of two answers on a 7 point likert scale
System efficacy		Mean of two answers on a 7 point likert scale
Time duration percent difference $t_{\%}$		$\frac{\text{Estimated time duration} - \text{Actual time duration}}{\text{Actual time duration}} * 100$
Interest/enjoyment		7 point likert scale
Perceived competence		7 point likert scale
Effort/importance		7 point likert scale
Value/Usefulness		7 point likert scale
Felt pressure and tension		7 point likert scale
Preferred interface		7 point likert scale
Easiest to use interface		7 point likert scale

6.2 Metrics: Administered Questionnaires

Following Barbieri et al. [7], a seven point Likert scale usability satisfaction questionnaire was administered including 6 items, which is suited to catch cognitive aspects related to user satisfaction.

The posed questions can be gathered in couples thus forming three sub-groups corresponding to the following three sub-scales: learnability, interface efficacy and system efficacy.

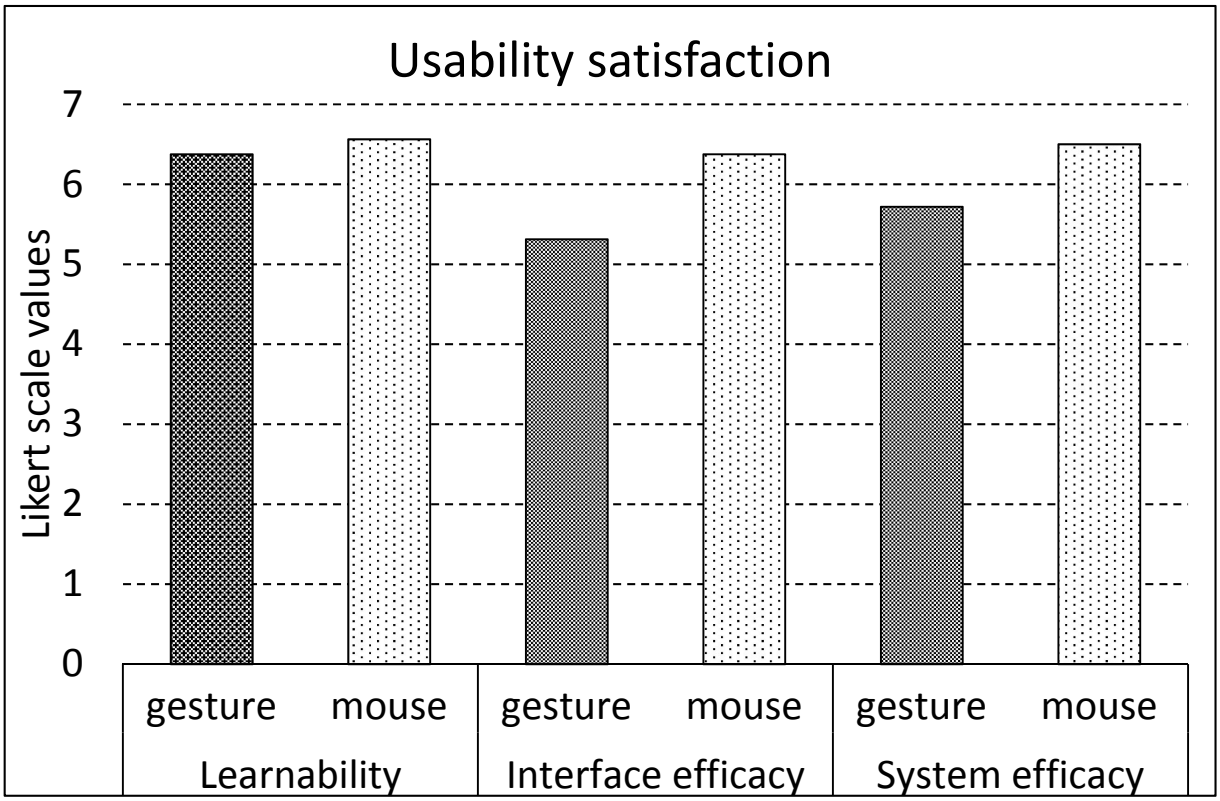
In order to compare the interfaces, the users' subjective experience was measured by administering a post-test Intrinsic Motivation Inventory (IMI) questionnaire which is a flexible assessment tool that determines the subjects' (i) interest/enjoyment, (ii) perceived competence, (iii) effort, (iv) value/usefulness, (v) felt pressure and tension, and (vi) perceived choice while performing a given activity, thus yielding six sub-scale scores [30] [31]. This test was successfully utilized in several experiments related to intrinsic motivation and self-regulation [32–35] as well as in studies regarding tasks execution in virtual environments [36–39]. Indeed, a customized version of the questionnaire was utilized that includes 22 questions and neglects the perceived choice sub-scale (i.e. the sub-scale (vi)).

The proposed questionnaires included two further questions: (i) the first one regarding the preferred and the easiest to use interface (gesture-based or mouse-controlled); (ii) the second one regarding the time the user thinks has spent to conduct the test. The comparison between this time and the actual time measured by the experimenter was expressed in terms of percent difference $t_{\%}$. In other words, if $t_{estimated}$ is the time estimated by the user to execute the test while t_{actual} is the actual time measured by the experimenter to carry out the test, $t_{\%}$ can be computed as:

$$t_{\%} = \frac{t_{estimated} - t_{actual}}{t_{actual}} \times 100$$

6.3 Results

The usability satisfaction questionnaire returned usability scores for the mouse interface higher than for the gesture interface (Fig. 8(a)).



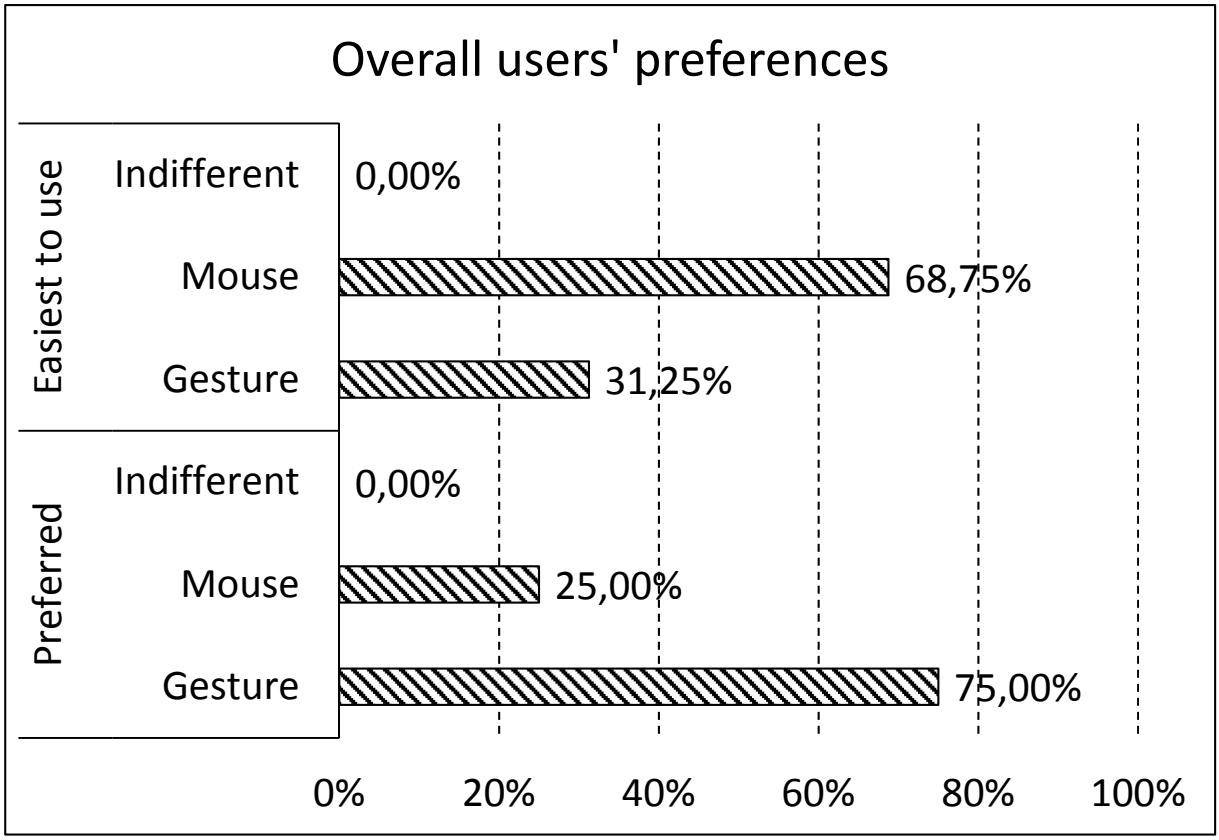


Fig. 8. (a) Average values computed over the 16 participants of the usability satisfaction questionnaire scores obtained for the different sub-scales: learnability, interface efficacy and system efficacy. (b) Overall users' preferences: Preferred and Easiest to use interface.

Paired samples t tests (Table 2) show that these differences are statistically significant for the Interface efficacy and the System efficacy sub-scales while no statistically significant differences can be seen for the Learnability sub-scale.

Table 2: Descriptive statistics and paired samples t test for the usability satisfaction questionnaire sub-scales.

Sub-scale	Interface	N	Mean	Std. D.	t(df=15)	p
Learnability	Gesture	16	6.38	0.62	-1.576	0.138
	Mouse	16	6.56	0.12		
Interface efficacy	Gesture	16	5.31	1.24	-2.573	0.021
	Mouse	16	6.38	0.81		
System efficacy	Gesture	16	5.72	1.00	-2.674	0.017
	Mouse	16	6.50	0.52		

In the IMI questionnaire, the gesture-based interface obtained, for the Interest/Enjoyment and the Value/Usefulness sub-scales, median scores significantly higher than the mouse-controlled interface counterpart (Table 3).

Table 3: Descriptive statistics and Wilcoxon Signed-ranks test for the IMI questionnaire sub-scales.

Sub-scale	Interface	N	Median	Min	Max	Z	p
Interest / Enjoyment	Gesture	16	30.50	26	35	-2.425	0.015
	Mouse	16	28.50	18	34		
Effort / Importance	Gesture	16	24.00	11	34	-1.337	0.181
	Mouse	16	21.00	11	32		
Perceived competence	Gesture	16	29.50	16	35	-0.427	0.669
	Mouse	16	29.00	22	35		
Pressure / Tension	Gesture	16	12.00	8	17	-1.355	0.176
	mouse	16	11.50	9	16		
Value / Usefulness	Gesture	16	18.00	14	21	-2.098	0.036
	Mouse	16	16.00	10	20		

The time spent to carry out the test was averagely underestimated by 12.08 % for the gesture-based interface and overestimated by 10.88 % for the mouse-controlled interface (Table 4). Statistically significant differences could be found between the two time estimations (Table 4).

Table 4: Descriptive statistics and paired samples t test for the values of $t_{\%}$ computed for the mouse-controlled and the gesture-based interfaces.

Variable	Interface	N	Mean $t_{\%}$	Std. D.	t(df=15)	p
Percent difference	Gesture	16	-12.08%	44.91%	-2.516	0.024
	Mouse	16	10.88%	47.42%		

Finally, users preferred the gesture-based interface (exact binomial test p=0.021) while judged the mouse-controlled interface as the easiest to use (Fig. 8(b)).

7. Discussion

A gesture-based interface was proposed to navigate a virtual tour on display walls. In particular, the proposed interface was developed to "visit" Murgia, a 4000 km^2 karst zone lying within Puglia, very famous for its fortified farms, dolines, sinkholes, and caves. The virtual tour was hypothesized to take place in a container, a "portable" installation that can be easily transported to the place where cultural/touristic events are organized and properly equipped with all the devices required for a virtual tour [40]. The goal pursued in the study is the development of an appealing interface capable of improving the users' engagement/enjoyment thus attracting their attention and interest towards the exploration of specific cultural heritage subjects.

After simulating a brief trip through the Murgia territory with its natural landscapes and its typical countryside's native vegetation, the tour shows Masseria Jesce, an ancient farm dating back to the 16^{th} century and located in the vicinity of the ancient Appian Way, in the countryside around Altamura (Bari, Italy). The farm includes amphitheater caves (Fig. 9(a)) such as a frescoed crypt dedicated to the Archangel St. Michael on the walls of which a deesis (Fig. 9(b)), (attributed to

Fig. 9. The tour shows Masseria Jesce, an ancient farm dating back to the 16th century. The farm includes amphitheater caves (a) such as a frescoed crypt dedicated to the Archangel St. Michael on the walls of which a deesis (b), (attributed to Giovanni da Taranto, 14th century) a Marian cycle (c) and a painting of St. Michael (d) (attributed to Didaco de Simone, 17th century) is represented.

Giovanni da Taranto, 14th century) a Marian cycle (Fig. 9(c)) and a painting of St. Michael (Fig. 9(d)) (attributed to Didaco de Simone, 17th century) is represented.

Given the novelty of the proposed gesture-based interface compared to the well-known mouse-controlled one, the hypothesis H_1 was formulated as: *users evaluate the mouse-controlled interface the most usable*. Indeed, the Interface efficacy and the System efficacy sub-scales (see Table 2) of usability satisfaction questionnaire confirmed our hypothesis H_1. Furthermore, coherently with this hypothesis, the overall users' preferences indicated the mouse as the easiest to use interface (Fig. 8(b)), which is consistent with the results of Cabral et al. [8] who found that the time to completion of simple pointing tasks performed with the use of gestures, is considerably slower when compared to that spent with a mouse-controlled interface. Regarding the sub-scale of Learnability (see Table 2), no statistically significant differences exist between mouse and gestures. This result leads us to conclude that the user-centric approach adopted in the design process, was capable of giving to the proposed interface an intuitiveness comparable with that of the mouse-controlled interface.

Regarding the Interest/Enjoyment sub-scale of the IMI questionnaire, the gesture-based interface reached a score significantly higher than the mouse-controlled interface (Table 3). Similarly, the score related to the sub-scale Value/Usefulness was significantly higher for the gesture-based interface, which can be interpreted as the argument that when interacting by gestures, users give more value to their navigation experience than when interacting by mouse. Furthermore, the computed values of $t_{\%}$ (Table 3) further confirm the fact that the interaction by gestures involves so much the user that she/he loses the sense of the time. These results, definitively, lead us to conclude that the hypothesis H_2*: the gesture-based interface will achieve a better score in terms of enjoyment/engagement*, holds true. This is consistent with the results of Cabral et al. [8] who found that gesture interfaces allow, compared to the mouse-controlled interface, to have a natural and intuitive access to computing resources that might be embedded in the environment. For the other sub-scales of the IMI questionnaire, no statistically significantly different scores have been obtained

462 (Table 3), which implies that the interfaces gesture-based and mouse-based are practically equivalent
463 from this point of view.
464 From the overall users' preferences (Fig. 8(b)) it appears that the preferred interface is the
465 gesture-based one. However, the mouse-controlled interface appears to be the easiest to use. These
466 results support the hypothesis *H_3: The overall users' preferences will be more oriented towards the*
467 *gesture-based interface.*
468 Interestingly, during the open interviews participants confirmed that the use of a new interface
469 modality was challenging and enjoyable at the same time. Compared to the gesture-based interface
470 the mouse-controlled one was boring even if most of the users judged it as the easiest to use interface.
471 Different participants claimed also to feel the gesture-based interface more "natural" and more
472 capable of making natural/real the virtual environment.
473 The proposed study presents some limitations. First, despite our efforts to reduce fatigue and
474 the choice of adaptive thresholds (utilized to define the areas for the interaction) based on user's
475 anthropometric data, some users still complain about the gorilla-arm effect [28]. The same limitation
476 was pointed out by Cabral et al., [8] who found that the use of gesture interfaces in virtual environment
477 causes fatigue. This suggests that the proposed gesture-based interface is attractive for novice users,
478 but needs to be further improved in the case of long-lasting user-interactions. Second, a user-centric
479 approach was adopted to define the gesture vocabulary. Generally speaking, sometime it is
480 cumbersome to develop a reliable gesture-recognition software capable of detecting a gesture
481 vocabulary defined via a user centric approach [41]. For instance, this issue was encountered to
482 recognize the gesture proposed for the referent "Select items," i.e., One Hand Pointing (OHPo, Fig.
483 1). The skeleton-tracking algorithm included in the Microsoft Kinect SDK was not designed to detect
484 such a specific gesture. A possible strategy that can be adopted to overcome this limitation consists
485 in identifying this hand state by using the joints information returned by the skeleton-tracking
486 algorithm and implementing some hand-tracking library. However, such a solution would
487 tremendously increase the computational cost thus slowing down the response time of the user-

interface and resulting in a poor user-experience. The issue of triggering the Select items referent was addressed by utilizing the detection of the lasso state. Preliminary observations confirmed, in fact, that even if the user executes the OHPo gesture, the gesture-recognition system recognizes it as a lasso gesture and consequently triggers the correct action on the interface. The evaluation/testing of the proposed gesture-based interface was carried out on participants 28±1 years old, who certainly possess a large experience with digital applications. The strong legacy bias related to the use of the mouse-controlled interface, led us to formulate the null hypotheses H_1, H_2 and H_3 as above described. Further studies should be carried out in the future to investigate how the proposed gesture-based interface is judged by adult and elderly people that may not have any experience in digital applications. A possible strategy that can be adopted to make more engaging the proposed gesture-based interface consists in including in the system further interaction units by means of which the user-leader can interact with other possible users or tour-guides present in the container. Such a strategy was successfully implemented by Fanini et al. [10] to explore the treasures of museums. Future studies should be oriented in this direction although we believe that the most important issue to implement this solution is represented by the limited space available in a container.

8. Conclusions

A gesture-based interface was proposed to navigate a virtual tour on display walls. In detail, this interface was developed to "visit" Murgia, i.e., a karst topographic plateau of rectangular shape lying within Puglia, very famous for its fortified farms and for being the seat of transhumance practice in animal husbandry. User tests aimed at comparing the implemented interface with a classical mouse-controlled one confirmed the capability of the proposed gesture-based interface to enhance the user engagement/enjoyment. The mouse-controlled interface appeared to be boring even if most of the users judged it as the easiest to use interface. Interestingly, all the users that participated in the tests underestimated the time spent to navigate the virtual tour via the gesture-based interface while

512 overestimated the time in which they utilized the mouse-controlled interface. This leads us to
513 conclude that the interaction by gestures involves so much the user and is so much interesting for
514 her/him that she/he loses the sense of the time.

515

9. Funding sources

This work was developed as part of the PAC02L2_00228 “VirtualMurgia - Smart-Multisense Ubiquitous System for Territorial Promotion of Apulia’s Culture and Traditions of Murgia” project, which was cofunded by the Apulia Region and the European Union under the Cohesion Action Plan.

10. References

[1] M. Carrozzino, M. Bergamasco, Beyond virtual museums: Experiencing immersive virtual reality in real museums, Journal of Cultural Heritage. 11 (2010) 452–458.

[2] M. Mortara, C.E. Catalano, F. Bellotti, G. Fiucci, M. Houry-Panchetti, P. Petridis, Learning cultural heritage by serious games, Journal of Cultural Heritage. 15 (2014) 318–325. doi:https://doi.org/10.1016/j.culher.2013.04.004.

[3] M. Pullambaku, T. Tsering, Multi-user and immersive experiencesin education: the Ename 1290 game, based on the use of MicrosoftKinect and Unity, (2016).

[4] S. Styliani, L. Fotis, K. Kostas, P. Petros, Virtual museums, a survey and some issues for consideration, Journal of cultural Heritage. 10 (2009) 520–528.

[5] S.E. Chen, Quicktime VR: An image-based approach to virtual environment navigation, in: Proceedings of the 22nd annual conference on Computer graphics and interactive techniques, ACM, 1995: pp. 29–38.

[6] J.P. Guerra, M.M. Pinto, C. Beato, Virtual reality-shows a new vision for tourism and heritage, European Scientific Journal, ESJ. 11 (2015).

[7] L. Barbieri, F. Bruno, M. Muzzupappa, Virtual museum system evaluation through user studies, Journal of Cultural Heritage. 26 (2017) 101–108.

[8] M.C. Cabral, C.H. Morimoto, M.K. Zuffo, On the usability of gesture interfaces in virtual reality environments, in: Proceedings of the 2005 Latin American conference on Human-computer interaction, ACM, 2005: pp. 100–108.

[9] P. Dam, P. Braz, A. Raposo, A study of navigation and selection techniques in virtual environments using microsoft kinect®, in: International Conference on Virtual, Augmented and Mixed Reality, Springer, 2013: pp. 139–148.

[10] B. Fanini, E. d Annibale, E. Demetrescu, D. Ferdani, A. Pagano, Engaging and shared gesture-based interaction for museums the case study of K2R international expo in Rome, in: Digital Heritage, 2015, IEEE, 2015: pp. 263–270.

[11] M. Nielsen, M. Störring, T. Moeslund, E. Granum, A Procedure for Developing Intuitive and Ergonomic Gesture Interfaces for HCI, in: A. Camurri, G. Volpe (Eds.), Gesture-Based Communication in Human-Computer Interaction, Springer Berlin Heidelberg, 2004: pp. 409–420. doi:10.1007/978-3-540-24598-8_38.

[12] M.R. Morris, J.O. Wobbrock, A.D. Wilson, Understanding users’ preferences for surface gestures, in: Proceedings of graphics interface 2010, Canadian Information Processing Society, 2010: pp. 261–268.

[13] V.M. Manghisi, M. Fiorentino, M. Gattullo, A. Boccaccio, V. Bevilacqua, G.L. Cascella, M. Dassisti, A.E. Uva, Experiencing the Sights, Smells, Sounds, and Climate of Southern Italy in VR, IEEE computer graphics and applications. 37 (2017) 19–25.

[14] G. Ren, C. Li, E. O'Neill, P. Willis, 3d freehand gestural navigation for interactive public displays, IEEE computer graphics and applications. 33 (2013) 47–55.
[15] R.-D. Vatavu, User-defined gestures for free-hand TV control, in: Proceedings of the 10th European conference on Interactive tv and video, ACM, 2012: pp. 45–48.
[16] M. Koehl, A. Schneider, E. Fritsch, F. Fritsch, A. Rachedi, S. Guillemin, Documentation of historical building via virtual tour: the complex building of baths in Strasbourg, in: Proceedings of the XXIV International CIPA Symposium on Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Strasbourg, France, 2013: pp. 2–6.
[17] K. Kwiatek, M. Woolner, Embedding interactive storytelling within still and video panoramas for cultural heritage sites, in: Virtual Systems and Multimedia, 2009. VSMM'09. 15th International Conference on, IEEE, 2009: pp. 197–202.
[18] J. Blake, Natural User Interfaces in .Net, Manning Publications, 2012.
[19] K. Reinfeld, krpano https://krpano.com/ last (accessed genuary 2017), (2016). https://krpano.com/ last accessed genuary 2017.
[20] N. Dahlbäck, A. Jönsson, L. Ahrenberg, Wizard of Oz studies—why and how, Knowledge-based systems. 6 (1993) 258–266.
[21] T. Piumsomboon, A. Clark, M. Billinghurst, A. Cockburn, User-Defined Gestures for Augmented Reality, in: P. Kotzé, G. Marsden, G. Lindgaard, J. Wesson, M. Winckler (Eds.), Human-Computer Interaction – INTERACT 2013, Springer Berlin Heidelberg, 2013: pp. 282–299. doi:10.1007/978-3-642-40480-1_18.
[22] J.O. Wobbrock, H.H. Aung, B. Rothrock, B.A. Myers, Maximizing the Guessability of Symbolic Input, in: CHI '05 Extended Abstracts on Human Factors in Computing Systems, ACM, Portland, OR, USA, 2005: pp. 1869–1872. doi:10.1145/1056808.1057043.
[23] J.O. Wobbrock, M.R. Morris, A.D. Wilson, User-defined Gestures for Surface Computing, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, Boston, MA, USA, 2009: pp. 1083–1092. doi:10.1145/1518701.1518866.
[24] B. Jones, R. Sodhi, M. Murdock, R. Mehra, H. Benko, A. Wilson, E. Ofek, B. MacIntyre, N. Raghuvanshi, L. Shapira, RoomAlive: Magical Experiences Enabled by Scalable, Adaptive Projector-camera Units, in: Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology, ACM, Honolulu, Hawaii, USA, 2014: pp. 637–644. doi:10.1145/2642918.2647383.
[25] V.M. Manghisi, A.E. Uva, M. Fiorentino, V. Bevilacqua, G.F. Trotta, G. Monno, Real time RULA assessment using Kinect v2 sensor, Applied Ergonomics. (2017).
[26] B.J. Fernández-Palacios, D. Morabito, F. Remondino, Access to complex reality-based 3D models using virtual reality solutions, Journal of Cultural Heritage. 23 (2017) 40–48. doi:https://doi.org/10.1016/j.culher.2016.09.003.
[27] Q. Wang, G. Kurillo, F. Ofli, R. Bajcsy, Evaluation of pose tracking accuracy in the first and second generations of microsoft kinect, in: Healthcare Informatics (ICHI), 2015 International Conference on, IEEE, 2015: pp. 380–389.
[28] J.D. Hincapié-Ramos, X. Guo, P. Moghadasian, P. Irani, Consumed Endurance: A Metric to Quantify Arm Fatigue of Mid-air Interactions, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, Toronto, Ontario, Canada, 2014: pp. 1063–1072. doi:10.1145/2556288.2557130.
[29] M. Ayoub, Work place design and posture, Human Factors: The Journal of the Human Factors and Ergonomics Society. 15 (1973) 265–268.
[30] R.M. Ryan, E.L. Deci, Intrinsic and extrinsic motivations: Classic definitions and new directions, Contemporary educational psychology. 25 (2000) 54–67.
[31] IMI, Intrinsic Motivation Inventory (2003), Available at www.psych.rochester.edu/SDT/measures/word/IMIfull.doc (Last accessed 2017/11/05), (n.d.). http://www.psych.rochester.edu/SDT/measures/word/IMIfull.doc.

[32] E.L. Deci, H. Eghrari, B.C. Patrick, D.R. Leone, Facilitating internalization: The self-determination theory perspective, Journal of personality. 62 (1994) 119–142.
[33] R.W. Plant, R.M. Ryan, Intrinsic motivation and the effects of self-consciousness, self-awareness, and ego-involvement: An investigation of internally controlling styles, Journal of personality. 53 (1985) 435–449.
[34] R.M. Ryan, Control and information in the intrapersonal sphere: An extension of cognitive evaluation theory., Journal of personality and social psychology. 43 (1982) 450.
[35] R.M. Ryan, R. Koestner, E.L. Deci, Ego-involved persistence: When free-choice behavior is not intrinsically motivated, Motivation and emotion. 15 (1991) 185–205.
[36] R. Colombo, F. Pisano, A. Mazzone, C. Delconte, S. Micera, M.C. Carrozza, P. Dario, G. Minuco, Design strategies to improve patient motivation during robot-aided rehabilitation, Journal of neuroengineering and rehabilitation. 4 (2007) 3.
[37] W.A. IJsselsteijn, Y. de Kort, J. Westerink, M. de Jager, R. Bonants, Virtual fitness: stimulating exercise behavior through media technology, Presence: Teleoperators and Virtual Environments. 15 (2006) 688–698.
[38] M. Mihelj, D. Novak, M. Milavec, J. Ziherl, A. Olenšek, M. Munih, Virtual rehabilitation environment using principles of intrinsic motivation and game design, Presence: Teleoperators and Virtual Environments. 21 (2012) 1–15.
[39] D. Novak, A. Nagle, U. Keller, R. Riener, Increasing motivation in robot-aided arm rehabilitation with competitive and cooperative gameplay, Journal of neuroengineering and rehabilitation. 11 (2014) 64.
[40] V.M. Manghisi, M. Fiorentino, M. Gattullo, A. Boccaccio, V Bevilacqua, G.L. Cascella, M. D'Assisti, A.E. Uva, Experiencing the Sights, Smells, Sounds, and Climate of Southern Italy in VR, IEEE Computer Graphics and Applications (in press November/December 2017). (2017).
[41] P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, J. Kautz, Online Detection and Classification of Dynamic Hand Gestures With Recurrent 3D Convolutional Neural Network, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

Figure legends

Fig. 1. Gesture proposals collected for each of the five hypothesized referents. The acronyms utilized to identify each gesture proposal are reported on the top of each picture.

Fig. 2. (a) Percentage of gesture proposals collected for each referent. The acronyms utilized are specified in Fig. 1. (b) Values of guessability computed for each gesture proposal. The histograms highlighted in red refer to the gestures with the highest guessability that were finally utilized in the best gesture vocabulary.

Fig. 3. Best vocabulary of gestures utilized to carry out the virtual tour.

Fig. 4. Sequence of the main actions controlled for each new-frame detected by the sensor.

Fig. 5. Schematic of the working areas defined (with their limit dimensions) to control the user's interaction.

Fig. 6. (a) The state machine that controls the user's behavior. (b) The hand states detected by the Kinect skeleton tracking algorithm for the tracked hands: b1) Unknown; b2) Open; b3) Closed; b4) Lasso. (c) Pointer icons shown on the screen and utilized as a feedback to make aware users about their current state: c1) Move the pointer on the screen; c2) Zoom; c3) Change gaze direction; c4) Select items; c5) Hotspots pointer on-over event.

Fig. 7. Schematic of the ray casting method adopted to control the pointer location. The area (delimited by the magenta rectangle, in the case of projection center located in correspondence of the user's head) that the user's hand must sweep to roam the displayed scene becomes smaller (area delimited by the blue rectangle), when the projection center moves from the position of the head towards the display wall (by the quantity L=500 mm) and at a lower level (by the quantity h=600 mm).

Fig. 8. (a) Average values computed over the 16 participants of the usability satisfaction questionnaire scores obtained for the different sub-scales: learnability, interface efficacy and system efficacy. (b) Overall users' preferences: Preferred and Easiest to use interface.

Fig. 9. The tour shows Masseria Jesce, an ancient farm dating back to the 16th century. The farm includes amphitheater caves (a) such as a frescoed crypt dedicated to the Archangel St. Michael on the walls of which a deesis (b), (attributed to Giovanni da Taranto, 14th century) a Marian cycle (c) and a painting of St. Michael (d) (attributed to Didaco de Simone, 17th century) is represented.

Supplementary Material

Enhancing user engagement through the user centric design of a mid-air gesture-based interface for the navigation of virtual-tours in cultural heritage expositions

Vito M. Manghisi, Antonio E. Uva, Michele Fiorentino, Michele Gattullo, Antonio Boccaccio*, and Giuseppe Monno

Department of Mechanics, Mathematics and Management, Polytechnic University of Bari, Italy

*Corresponding author:

Antonio Boccaccio
E-mail address: a.boccaccio@poliba.it

Appendix A

Gesture proposals description

For the Move the pointer on the screen referent, two different gestures were proposed (Fig. 1):

- Moving Hand with Open palm (MHO): Users keeping open the palm, moved the hand to control the pointer position on the screen;
- Moving Hand with Index Finger Pointing (MHIP): Users kept the index finger pointing at the pointer position on the screen while moved their hand to control the pointer position.

For the Zoom-in referent three different gestures were proposed (Fig. 1):

- One Hand Un-Pinching (OHUP): Two fingers of the dominant hand un-pinch, just as for touch interfaces;
- Distancing Two Hands with Open palms (DTHO): The user moves her/his hands in a plane approximately parallel to the frontal plane. Palms are kept open while the user distances them increasing their relative distance;
- Distancing Two Hands with Clenched fists (DTHC): The user moves her/his hands in a plane approximately parallel to the frontal plane. Fists are kept clenched while the user distances them increasing their relative distance.

For the Zoom-out referent there were three different gesture proposals (Fig. 1):

- One Hand Pinching (OHP): Two fingers of the dominant hand pinch just as for touch interfaces;
- Bringing Together Two Hands with Open palms (BTTHO): The user moves her/his hands in a plane approximately parallel to the frontal plane. Palms are kept open while the user brings together hands reducing their relative distance.
- Bringing Together Two Hands with Clenched fists (BTTHC): The user moves her/his hands in a plane approximately parallel to the frontal plane. Fists are kept clenched while the user brings together hands reducing their relative distance;

For the Change gaze direction referent two gestures were proposed (Fig. 1):

- One Hand Grabbing and Moving (OHGM): The user clenches her/his fist and changes the direction of view by dragging the scene with her/his hand;
- One Hand Pointing and Moving (OHPM): The user points at the scene with the index finger and changes the direction of view by moving the scene with her/his index finger.

For the Select items referent there were two different gesture proposals (Fig. 1):

- One Hand Pushing (OHPu): The user points at the target on the display by positioning the cursor on it, then she/he selects the target by pushing her/his hand toward it;
- One Hand Pointing (OHPo): The user points at the target on the display by positioning the cursor on it, then she/he selects the target by pointing at it with the index finger;
- One Hand Clicking (OHC): The user points at the target on the display by positioning the cursor on it, then she/he selects the target by "clicking" on it.

Appendix B

Agreement analysis

The *Agreement Rate AR* is defined as "the number of pairs of participants in agreement with each other divided by the total number of pairs of participants that could be in agreement".

In detail, the *AR* can be computed as:

$$AR(r_k) = \frac{|P|}{|P|-1}\sum_{P_i \subseteq P}\left(\frac{|P_i|}{|P|}\right)^2 - \frac{1}{|P|-1};\ AR \in [0, 1] \qquad \text{(B1)}$$

where, P is the set of all proposals for the referent r_k, $|P|$ the size of the set, and $|P_i|$ the size of subsets of similar proposals included in P. The Agreement rate AR ranges in the interval [0, 1]. For a given referent r_k, the value $AR(r_k) = 0$ refers to the case where all the proposals collected for that referent r_k are different from each other, the value 1 to the case where all the proposals (collected for r_k) are the same. The agreement rate AR for each of the five hypothesized referents (Table B1) was computed by implementing the AGATE tool (AGreement Analysis Toolkit, http://depts.washington.edu/aimgroup/proj/dollar/agate.html).

Table B1. Agreement rate computed for the five hypothesized referents

Referents	Agreement Rate *AR*
Move the pointer on the screen	0.512
Zoom-in	0.325
Zoom-out	0.325
Change gaze direction	0.532
Select items	0.320
average *(AR)*	0.403

It is worthy to note that values of the agreement rates too much low indicate a high cognitive load which requires to redesign the commands set. However, implementing the Variation between

agreement rates statistics (V_{rd} statistics), - which is a statistical significance test for comparing two or multiple agreement rates calculated from proposals elicited from the same participants (i.e., repeated measures design) [36] -, we found that the agreement rates related to all the hypothesized referents have statistically significant differences with respect to zero: (*AR*(*Move the pointer on the screen*) = 0.512, $V_{rd}(1) = 208.000$, $p = .001$; *AR*(*Zoom-out*) = 0.325, $V_{rd}(1) = 132.000$, $p = 0.001$; *AR*(*Zoom-in*) = 0.325, $V_{rd}(1) = 132.000$, $p = .001$; *AR*(*Select items*) = 0.320, $V_{rd}(1) = 130.000$, $p = .001$); *AR*(*Change gaze direction*) = 0.532, $V_{rd}(1) = 216.000$, $p = .001$)), which means, in other words, that the hypothesized referents do not require high cognitive load.