applied sciences

MDPI

Article

MuMIA: Multimodal Interactions to Better Understand Art Contexts

George E. Raptis *, Giannis Kavvetsos and Christina Katsini

Human Opsis, Patras, 26500 Western Greece, Greece; gkavvetsos@humanopsis.com (G.K.); ckatsini@humanopsis.com (C.K.)

* Correspondence: graptis@humanopsis.com

Abstract: Cultural heritage is a challenging domain of application for novel interactive technologies, where varying aspects in the way that cultural assets are delivered play a major role in enhancing the visitor experience, either onsite or online. Technology-supported natural human–computer interaction that is based on multimodalities is a key factor in enabling wider and enriched access to cultural heritage assets. In this paper, we present the design and evaluation of an interactive system that aims to support visitors towards a better understanding of art contexts through the use of a multimodal interface, based on visual and audio interactions. The results of the evaluation study shed light on the dimensions of evoking natural interactions within cultural heritage environments, using micro-narratives for self-exploration and understanding of cultural content, and the intersection between human–computer interaction and artificial intelligence within cultural heritage. We expect our findings to provide useful insights for practitioners and researchers of the broad human–computer interaction and cultural heritage communities on designing and evaluating multimodal interfaces to better support visitor experiences.

Keywords: human–computer interaction; multimodal interactions; eye tracking; voice; cultural heritage; museum; artificial intelligence

Citation: Raptis, G.E.; Kavvetsos, G.; Katsini, C. MuMIA: Multimodal Interactions to Better Understand Art Contexts. *Appl. Sci.* **2021**, *11*, 2695. https://doi.org/10.3390/app11062695

Academic Editor: Liliana Ardissono

Received: 25 February 2021
Accepted: 9 March 2021
Published: 17 March 2021

1. Introduction

During the last few decades, there has been an obvious and multifaceted evolution of the way that cultural heritage institutions (e.g., archaeological sites and museums) deliver collections, exhibits, and activities to their visitors. The establishment of a positive and enduring relationship with visitors remains the basic purpose of cultural heritage institutions, and thus, they drive the delivery of meaningful and sustainable visitor experiences. On the other hand, there is a constant technological evolution, which influences several aspects of the technology-related part of the visitor experience, such as the interactive nature between visitors and exhibits, which are augmented with virtual-form and physical-form cultural information. The increased technological capabilities allow for alternative methods of interaction and exhibition types. Therefore, cultural heritage institutions have the opportunity to convey their cultural and educational message more efficiently than before because the conventional passive display of an exhibit can be enriched with features, tools, and activities, which significantly improve the visitor experience and knowledge transfer to the visitors.

Various methods embed the latest technological advancements into a cultural heritage institution (e.g., archaeological site, museum). For example, interactive systems, which play the role of a virtual museum guide, allow the audience to remotely visit a museum using a computer system (e.g., desktop computer, touch-screen mobile device, and head-mounted display). Moreover, virtual-, augmented-, and mixed-reality applications provide high-end visualizations and expand the knowledge space of the exhibits [1], such as virtual monument preservation [2] and representation [3]. The use of a common smartphone for augmented-reality applications means that almost every modern museum visitor can access

such interactive systems [4]. Concerning mixed-reality applications, there are sophisticated new devices (e.g., mobile head-mounted displays) that can lift the visitor experience to a whole new level. Such methods are employed nowadays by cultural heritage institutions, and thus, they provide visitors with rich experiences and attract wider audiences than those who visit the institutions physically. Technological advances have led to enhanced ubiquity that helps people access cultural information that the physical conditions and limitations would not allow for. For example, many national authorities stopped the operation of cultural heritage institutions during the recent pandemic, and thus, people did not have access to art and cultural events for a long time.

Nowadays, there is an abundance of cultural information available, about almost every object we can think of. Handling and presenting such a huge amount of information is a challenging task for cultural heritage institutions, which could influence the way we access and enjoy cultural and artistic content in ubiquitous computing scenarios. Advanced and natural human–computer interactions are a key factor in enabling access to cultural heritage. Towards this direction, multimodal interactions with the support of artificial intelligence could be employed. Through multimodal interactions, visitors will have a more fruitful and natural visitor experience, taking advantage of the interaction with multiple senses. The enhancement of such multimodal interactions with artificial intelligence will provide a more personalized and adaptive visitor experience. Therefore, it would be interesting to investigate whether we can design a multimodal interface to support an enriched visitor experience and to support visitors in better understanding cultural contexts.

In this paper, we discuss the design and evaluation of such an interactive system, which aims to support visitors towards a better understanding of art contexts through the use of a multimodal interface, based on visual and audio interactions. We expect our findings to provide useful insights for practitioners and researchers of the broad human–computer interaction and cultural heritage communities on designing and evaluating multimodal interfaces to better support visitor experiences. The rest of the paper is structured as follows: in the next section, we discuss related works and present the motivation of this work. Next, we report the design and evaluation of a tool that supports multimodal interactions to help visitors better understand art contexts. Next, we discuss the lessons learned, the implications, and the limitations of our work; present future research steps; and conclude the paper.

2. Related Work and Motivation

In this section, we discuss works about multimodal interactions and multimodal technologies and interfaces in cultural heritage, focusing on visual- and voice-specific interactions. Next, we present the motivation of our work and state the research question that the paper aims to answer.

2.1. Multimodal Interactions

Human interaction with the world is inherently multimodal because we employ multiple senses, both sequentially and in parallel, to passively and actively explore our environment, to confirm expectations about the world, and to perceive new information [5]. In contrast to human experience with the natural world, human–computer interaction (HCI) has historically been focused on uni-modal communication [5]. This means that information has been communicated between people and computer systems primarily through a single mode, such as text on a screen with a keyboard for input. In recent years, multimodal interactions (i.e., providing the users with multiple modes of interacting with a system) have gained popularity in the HCI domain, as research indicates that multimodal interactions improve user experience. The use of multimodal technologies is growing in importance due to advances in hardware and software, the benefits that they provide to users, and the natural fit they have with the increasingly ubiquitous mobile computing environment [6].

Multimodal technologies aim to support the recognition of naturally occurring forms of human language and behavior through the use of recognition-based technologies [7]. The main purpose of such systems is to deliver natural and efficient interaction, but it turns out that there are several specific advantages of multimodality such as better flexibility and reliability [5]. They can also offer interaction alternatives to better meet the needs of diverse users with a range of usage patterns, and preferences [8,9]. Other advantages of multimodal interfaces include [10] improved efficiency, shorter and simpler speech utterances, better precision of spatial information, enhanced error avoidance, easier error resolution, and enhanced adaptability.

As discussed, multimodal technologies support user input and processing of more than one modalities (e.g., speech, pen, touch and multi-touch, gestures, gaze, and virtual keyboard). These modalities may coexist together on an interface; they can be used either simultaneously or alternately. To describe multimodal technologies, we have to consider the various natural human capabilities and means to communicate. Among others, *vision* and *hearing/speech* are the most used aspects based on human senses to support multimodalities in HCI [11]. Regarding vision, face location, gaze, facial expression, lipreading, and sign language are some of the input types that have been used [5]. Regarding hearing/speech, both speech and non-speech audio as input have been used [5].

Regarding the application of multimodal technologies, several interactive systems have been proposed. Based on the aspect of human–computer interaction, research has shown that, when applying multimodal interaction, error handling and context checking are improved [12], the understanding of visualization is improved [13], users' engagement is increased [14], behavioral engagement is detected [15], users' behavior is monitored [16], design practices are enhanced [17], users' performance is improved [18,19], etc. The multimodal technologies have been applied in a wide range of devices, such as mobile phones [12,20], tablet devices [13], desktop and laptop computers [14,15,17], glasses [16], tactile displays [21–24], and vehicle interfaces [18,19]. The reported multimodalities have been developed through varying communication channels, such as speech input/voice commands [12,13] gaze [18–20], touch [20], face location [14], facial expression [14,15], gestures [16,17], and touch-related actions (e.g., vibrations and pressures) [21–24]. Multimodal technologies are used in a wide range of application domains, such as education [14,15], security [20], data visualization [13], food and diet [16], interaction design [17], and the automotive industry [18,19]. The next section focuses on the cultural heritage domain.

2.2. Multimodal Technologies and Interfaces in Cultural Heritage

In the cultural heritage domain, multimodal technologies have been used to provide a more realistic visitor experience [25], to improve and provide a more immersive visitor experience [26], to safeguard and transmit intangible cultural heritage aiming for better education [27], to improve accessibility and to attract wider audiences [28], to improve collaborative meaning-making [29], to evoke more natural interactions [30], to provide a more efficient way for seeking and retrieving operations [31], and to better support cultural heritage professionals on the management of heritage assets [32]. For the implementation of multimodal interactions, a wide range of techniques have been used, based on haptics [25], visual inspection [25,28,29,32], speech and sound [25,27–30,32], gestures [27,28], recognition of bio-metric and physio-metric characteristics [27], etc. Next, we focus on interactions with sound/speech and visual/gaze interfaces within the cultural heritage domain.

2.2.1. Interactions with Sound/Speech Interfaces

Regarding interactions with sound/speech interfaces, in the cultural heritage domain, the application of sound has been mostly used to improve the sense of place and immersion for certain displays. This is typically met in audio guides through commentary that directs the visitor to a specific location or area on the cultural heritage asset. Besides the use of audio as output only, audio input is also considered a useful communication method for

multimodal interfaces, especially with the use of natural speech interaction. The primary benefit of such an input is that the users act naturally as they simply talk to the cultural heritage system. Algorithms that support text-to-speech and speech-to-text are typically implemented in such tools.

Besides typical audio-based interactive systems, which provide audio as output to the visitors (e.g., audio tour guides), sound/speech interfaces have also been used for voice interaction (e.g., visitors asking the system about cultural information). Research has shown that voice interaction within multimodal interfaces in cultural heritage improves visitors' immersion and learning process [33], that utilized spatial sound can be used to build more realistic visitor experiences [34], that acoustic speech and music sounds can be used to model and support voice interaction for safeguarding and transmitting intangible cultural heritage [27], that the combination of Internet-of-things and artificial intelligence can provide customized visits through a vocal environment [35], and that speech recognition drives natural interactions and enhances visitor experiences [36].

2.2.2. Interactions with Visual/Gaze Interfaces

Regarding visual/gaze interfaces, the data from specialized equipment that monitor and analyze eye movements (i.e., eye-tracking data) have been used for monitoring users' behavior and interactions in many domains such as drowsiness detection [37], diagnosis of clinical conditions [38], neuromarketing [39], and security [40]. In the cultural heritage domain, eye-tracking has been used to understand visitor experience [41] and learning experience [42]; to understand visitors' emotional states, appraisals, and visual exploration patterns [43]; to provide personalized and immersed visitor experiences [44]; to build visual attention models of visitors [45],; and to detect objects and areas of interest [46,47].

Raptis et al. [41] showed that, through eye-tracking analysis, different visitors' cognitive styles can be recognized as they lead to imbalances regarding understanding of the cultural content. Based on these differences, eye tracking can be used to model visitors' profiles and to deliver personalized and immersed experiences [44]. Pierdicca et al. [45] showed that, through eye tracking, it is feasible to build visual attention models of visitors to extract patterns and to provide art recommendations that would be interesting to them. These patterns and points of interest provide useful insights for optimizing existing augmented-reality applications for cultural heritage institutions. Garbutt et al. [48] demonstrated the value of mobile eye tracking for identifying areas of attention and for identifying eye-movement patterns within exhibition spaces. Mokatren et al. [46] used computer vision techniques for image-based positioning and for detecting objects and areas of interest in real-time. In the same vein, Toyama et al. [47] used mobile eye tracking to recognize cultural heritage objects and Cantoni et al. [49] to interact with artwork.

The aforementioned works present eye-tracking applications that are divided into two categories: *active* and *passive*. Active eye-tracking applications enable the users to use their eye movements as an input modality to control a device, a virtual world, a game, etc. Passive eye-tracking applications are used to observe and evaluate human attention objectively and non-intrusively, enabling the researchers and designers to better understand the users and to increase the impact of the visual designs and communication. For the scope of the paper, active eye-tracking plays an important role as it provides a natural and engaging interaction experience, considering that people naturally gaze at the world in conjunction with other activities [50]. When active eye tracking is combined with other input modalities (e.g., voice interaction and mouse/touch), the user experience can be enhanced. Table 1 presents a comparison between our proposed tool and other tools in cultural heritage that are primarily based on active eye tracking.

Table 1. Comparison between our proposed tool (*MuMIA*) and other tools in cultural heritage that are primarily based on active eye tracking.

Work	Sound/Speech Interface	Gaze Interface
Raptis et al. [44]	Not available	Eye-tracking was used to elicit visitors' cognitive profiles and to adjust cultural heritage applications (virtual tour and game) during visit time.
Pierdicca et al. [45]	Passive audio guide	Gaze data were used to build visitors' attention models to extract patterns and to provide art recommendations that would be interesting to the visitors during visit time.
Mokatren et al. [46]	The visitors were delivered with verbal information about the exhibit they were looking at. Gesture identification was also employed.	Mobile eye tracking was used to identify visitors' location and exhibits of interest, taking advantage of image-based object recognition techniques.
Toyama et al. [47]	After identifying an exhibit, the corresponding audio file was played and verbal information was delivered to the visitors.	Mobile eye tracking was used to detect gaze on exhibits, which were then identified through image-based object recognition techniques.
Garbutt et al. [48]	Not available	Mobile eye tracking was used to identify areas of attention on exhibits and identify eye-movement patterns within exhibition spaces during visit time.
Cantoni et al. [49]	Not available	Gaze interactions was used to enable visitors of an exhibition to select artworks, to perform image handling (e.g., scrolling and resizing), and to define areas of interest.
Our tool: *MuMIA*	Visitors ask for information about the areas of interest they look at. They can combine areas of interest and ask for information about them, including areas that they had looked at before. The system provides visitors with verbal information.	Gaze data are used to help visitors identify multiple areas of interest on an exhibit and ask the system for information about them. They can revisit the identified areas of interest.

2.3. Motivation

From the aforementioned discussion, we argue that voice and gaze interaction channels (through sound/speech and gaze interfaces) can be combined to build an interactive multimodal system that aims to improve the visitor experience. The motivation underlying our work is the design of a multimodal interface, based on eye tracking (gaze interface) and voice (sound/speech interface), for exploring virtual cultural heritage exhibitions. Through the evaluation of this multimodal interface, we aim to investigate whether we can provide the visitors with a better understanding of art contexts and can improve their experience in a cultural heritage institution (e.g., virtual museum). Therefore, the research question that this paper aims to answer is:

> " *Is the design of an interactive multimodal interface based on eye-tracking and voice communication within the cultural heritage domain feasible?* "

In case of a positive answer, we aim to answer the following question too:

> " *What is its impact on the visitor experience?* "

3. Design of the System

To answer the research question, we designed *MuMIA* (*Mu*lti-*M*odal *I*nteractions in *A*rt), which allows visitors to perform multimodal interactions with art exhibits to build a deeper knowledge space about the presented cultural information by creating strong and memorable connections with the varying types of information presented within the exhibits. The system consists of four main components (art exhibit, eye-tracking module, voice commands/sound interface, and repository), with the visitor being the main actor of the system. Figure 1 depicts the conceptual architecture of *MuMIA*.

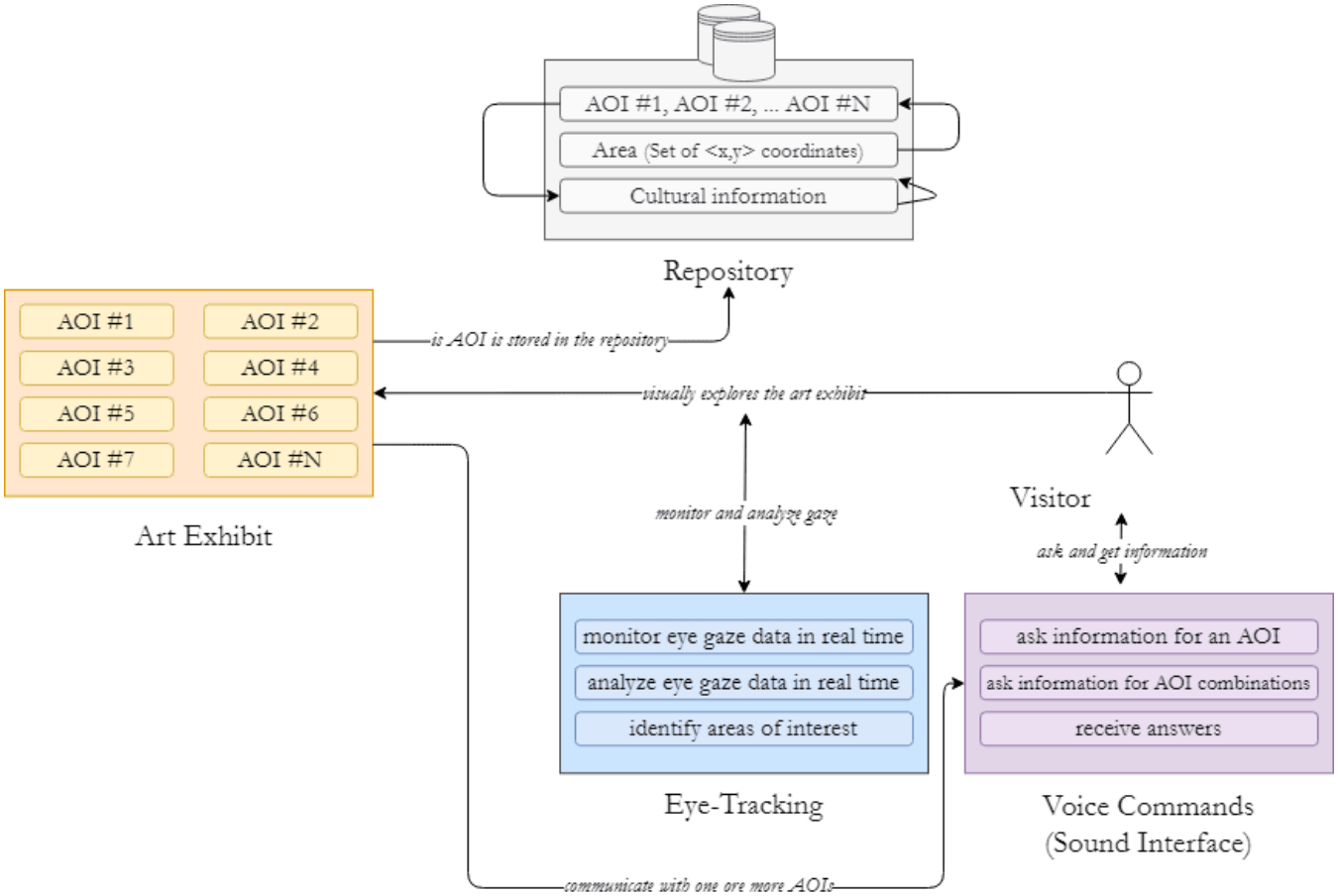

Figure 1. The conceptual architecture of *MuMIA* (Multimodal Interactions to Better Understand Art Contexts).

Each art exhibit consists of various areas that provide cultural information (i.e., areas of interest). Each area of interest (AOI) consists of cultural information that is related to the characteristics of the exhibits (e.g., theme, creator, era, and the information provided by the other areas of interest). Such information is stored in the *MuMIA* repository, and the stored elements develop interconnections with each other, considering that the information provided for each area of interest is often related to the information provided for some other area of interest.

When visitors stand in front of an art exhibit, they visually explore it and they receive brief and core information about it via audio (e.g., audio guide). Next, they discover special areas of interest by gazing at the exhibit, measured with the use of an eye-tracking apparatus (e.g., eye-tracking glasses). When an area of interest is identified, the visitor asks the system, through voice commands, to provide them with more information about the identified area of interest. The system then searches and retrieves the corresponding audio file from the repository.

After the corresponding file is retrieved, the system plays it to the visitor, and thus, they perceive the cultural experience through the audio channel. The visitors can gaze at

different areas of interest; can ask the system to provide them with more information or can build on previous knowledge, acquired through the multimodal interaction; and can pose new questions to the system, trying to connect different areas of interest and thus building a deeper understanding about the presented cultural information. The next subsections discuss in detail the (i) art exhibits, (ii) eye-tracking module, (iii) voice commands (sound interface), and (iv) repository. At the end of this section, we present a use-case scenario of *MuMIA* and implementation details for the prototype.

3.1. System Components

3.1.1. Art Exhibit

The art exhibit is every cultural artifact that can be presented in the cultural heritage institution, such as paintings and pediments with sculptures. Each art exhibit describes a specific context and theme, which are presented to the visitor. For example, *The School of Athens* (Italian: *Scuola di Atene*; Figure 2) is a fresco by Raphael (full name: Raffaello Sanzio da Urbino), who was an Italian Renaissance artist, and was painted between 1509 and 1511 as a part of Raphael's commission to decorate the rooms now known as the Stanze di Raffaello in the Apostolic Palace in the Vatican. represents philosophy and provides reference to Ancient Greek philosophers. Each art exhibit consists of several areas of interest, which contain critical cultural information regarding the context of the exhibit, the era, the painting technique, etc. In the previous example, a number of figures are depicted in *The School of Athens*. Besides well-known Ancient Greek philosophers, such as Plato and Aristotle, several other people, instruments, and relationships are depicted, that provide a deeper understanding of the underlying concepts if recognized and interpreted. For example, each philosopher represents different schools, such as the Platonist and the Peripatetic schools. Other people, other than Ancient Greek philosophers, can be identified, such as Leonardo da Vinci, whose work was one of the most striking influences in Raphael's work. Moreover, instruments, such as the compass used by Euclid, can be identified, and they provide a connection between an art-related or a science-related theme and the figure depicted (e.g., the intersection between mathematics and philosophy for Euclid). Such areas of interest provide meaningful concepts to visitors that are not visible during a simple and perfunctory scan of the exhibit. The areas of interest are defined by various stakeholders of the cultural heritage institution, such as historians and educators. Therefore, each art exhibit contains a collection of areas of interest (Equation (1)), and each area of interest is a collection of cultural information for the selected area of interest or for a combination with some other area of interest (Equation (2)).

$$Art\ Exhibit = \{AOI_1, AOI_2, \ldots, AOI_N\} \quad (1)$$

$$AOI = \{\langle cultural\ information, other\ AOIs\rangle, \langle cultural\ information, other\ AOIs\rangle, \ldots\} \quad (2)$$

3.1.2. Eye-Tracking Module

The eye-tracking module monitors and analyzes the visitors' gaze in real-time, and then, it matches it with the areas of interest within the art exhibit. When the area of interest is matched with the gaze pointer, the gaze pointer changes shape and format to inform the visitor that they have identified an area of interest. After an area of interest is identified, the visitor extracts cultural information about it through the use of voice commands (discussed in the next subsection). Therefore, the eye-tracking module performs three main functions:

- It monitors the eye gaze behavior of the visitors in real time: it captures saccades (i.e., the type of eye movement used to move the fovea rapidly from one point of interest to another) in real time in $<x, y>$ coordinates projected on the surface of the art exhibit. An area of interest is characterized by a collection of $<x, y>$ coordinates that build the total area (Equation (3)). Therefore, Equation (2) is enriched with that collection of

coordinates (Equation (4)). Moreover, a calibration procedure needs to be performed prior to the eye-tracking session to increase the validity of the captured eye gaze data.

- It analyzes the captured saccades in fixations (i.e., the periods of time when the eye is aligned with the target for a certain duration, allowing for the image details to be processed), and then, it extracts more complex metrics, such as fixation duration, fixation entropy, etc. To extract the fixations, we used a customized velocity threshold identification algorithm, with a minimum fixation duration set to 80 ms, as it is accepted to use fixations shorter than 100 ms when analyzing visual scene perception [51] as *MuMIA* does. Based on the position and the duration of the fixations, the system understands when and what the visitors gaze at, and thus, it activates (or not) the corresponding area of interest.
- It identifies the areas of interest, notifies the visitors (with the use of visual annotations), and then provides the visitors with a way to interact with them (through voice commands, as discussed in the next subsection).

$$Active\ Area_{area\ of\ interest} = \{\langle x_1, y_1\rangle, \langle x_2, y_2\rangle, \ldots, \langle x_N, y_N\rangle\} \quad (3)$$

$$AOI = \{\langle cultural\ information, other\ AOIs,\ Active\ Area\rangle, \ldots\} \quad (4)$$

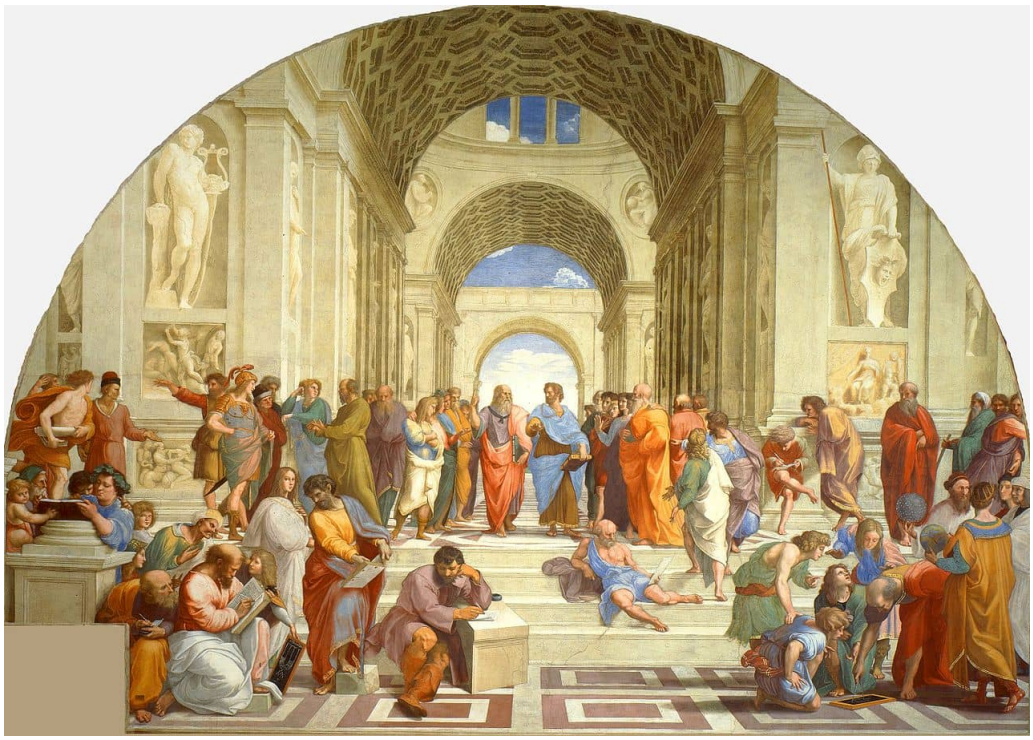

Figure 2. *The School of Athens* (Italian: *Scuola di Atene*) is a fresco by Raphael, which depicts the concept of philosophy. The most famous philosophers of ancient times move within an imposing Renaissance architecture, which was inspired by Bramante's project for the renewal of the early Christian basilica of St Peter. Some of these are easily recognizable. In the center, Plato points upwards with a finger and holds his book *Timeus* in his hand, flanked by Aristotle with *Ethics*; Pythagoras is shown in the foreground intending on explaining the diatessaron. Diogenes is lying on the stairs with a dish, while the pessimist philosopher, Heracleitus, a portrait of Michelangelo, is leaning against a block of marble, writing on a sheet of paper. Michelangelo was in those years executing the paintings in the nearby Sistine Chapel. On the right, we see Euclid, who is teaching geometry to his pupils; Zoroaster holding the heavenly sphere; and Ptolemy holding the earthly sphere. The personage on the extreme right with the black beret is a self-portrait of Raphael.

3.1.3. Voice Commands

The visitor uses voice commands to receive audio information about the art exhibit, an area of interest, or a combination of them. When an area of interest is identified (through the eye-tracking module discussed in the previous subsection), the visitor has the option to obtain information about it by asking the system "*Who is this?*", "*What is this?*", or "*Tell me more information about this*". For example, in *The School of Athens*, when the visitor gazes at the central figure and asks the system "*who is this?*", the system will tell the visitor that "*the central figure depicts Plato, who is represented by Leonardo da Vinci*" along with more details about Plato and Leonardo da Vinci in relation to the overall theme of the painting and the background of Raphael (e.g., the impact of Plato's work on philosophy and the influence of Leonardo da Vinci on Raphael's work). Each area of interest that is identified by the visitor is stored as an active artifact, and the visitor can select it again to hear its story and its connection with the art exhibit as many times as they want. Another feature of the system is that it provides the visitor with more complex information that connects different areas of interest. The visitor has access to such information by asking the system "*What is the connection between <area of interest 1> and <area of interest 2>?*", "*wWy is <area of interest> important for the exhibit?*", etc. For example, after both central figures of *The School of Athens* are identified, the visitor can ask the system "*What is the connection between Plato and Aristotle*", and the system will provide them with information about the relation between Plato and Aristotle (Aristotle was a student of Plato), about their notable ideas (e.g., Plato: theory of forms, Platonic idealism, and Platonic realism; Aristotle: the golden mean, reason, logic, biology, and passion) and main interests (e.g., Plato: rhetoric, art, literature,justice, virtue, politics, education, family, and militarism; Aristotle: politics, metaphysics, science, logic, and ethics) regarding philosophy, etc. Therefore, through the aforementioned techniques, we introduce an artificial intelligence aspect to the system, with each voice command triggering a function that retrieves specific types of cultural information and interconnections about the selected area(s) of interest and the art exhibit (Equation (5)).

$$f : Voice\ command(AOI) \rightarrow Cultural\ Information\ for\ AOI \quad (5)$$

3.1.4. Repository

The repository is a structure of data that contains the voice commands, the areas of interest, the location of the areas of interest at the art exhibit, and the cultural information that is related to the art exhibit. As discussed in the previous subsections, the areas of interest can be interrelated with each other, and thus, they can provide cultural information that is interconnected with the information provided by other areas of interest. The repository of the proposed system supports such types of relations. Various stakeholders can have access to the repository to manage and configure the cultural information and the areas of interest. For example, a historian can use the repository to define the areas of interest of an art exhibit and can provide information about them (e.g., their meaning, their importance for the creator, etc.), and an educator or a museum guide can use the repository to activate and deactivate areas of interest and information, so they can create a subset of cultural information that is related to the theme or the course of the visit in a given amount of time, etc.

3.2. Scenario

Our proposed system can be used in varying types of cultural heritage institutions and can support diverse types of art exhibits. A typical scenario of the use of the proposed system consists of the following steps:

1. The visitor enters a virtual gallery hall and starts exploring the various art exhibits. When an exhibit attracts their attention, they visit it (e.g., they stand in front of it).

2. The visitor visually explores the art exhibit, trying to discover areas that seem interesting. At the same time, the audio guide provides them with brief information about the exhibit and the gallery theme.
3. The visitor identifies an area of interest by gazing at it and asks the system to provide them with more information about what they gaze at, through the use of voice commands.
4. The system searches in the repository for information about the identified area of interest. When found, it retrieves the corresponding cultural information. It presents it to the visitor through the audio channel (e.g., sound interface).
5. After receiving the information, the visitor can gaze at other areas of interest to receive cultural information about them or can ask the system to provide them with more information about the gazed area of interest, in order to create deeper connections and to build a better understanding of the underlying contexts of the presented area of interest and its relation with other areas within the art exhibit.

3.3. Implementation

For the implementation of *MuMIA*, we used the Eye Tribe tracker (Table 2) to build the eye-tracking module. We used the Eye Tribe Java SDK, and thus Java programming language, to communicate with the Eye Tribe server and to interact with the eye-tracker in real-time. Each art exhibit was presented through a JAVA interface that could easily be deployed as a web application for a more universal access. To support the voice commands, we used speech recognition APIs. To implement the repository, we used data structures, an SQL database, and audio files.

Table 2. Technical characteristics of the Eye Tribe tracker.

Characteristic	Information
Sampling rate	60 Hz
Accuracy	0.5° (average)
Spatial resolution	0.1° (RMS)
Latency	<20 ms
Calibration	5, 9, 12 points
Operating range	45 cm–75 cm
Tracking area	40 cm × 30 cm at 65 cm distance
Screen sizes	Up to 24 inches

4. Evaluation Study

4.1. Method

The goal of this study was to evaluate the use of *MuMIA* by visitors of cultural heritage institutions, with an emphasis on the virtual exhibits. The visitors used *MuMIA* and a typical audio guide to explore *The School of Athens*. Eleven visitors participated in the study and shared their experiences with us, and we analyzed them following a thematic analysis approach based on the MUSETECH framework [52]. We discuss them in detail in the following paragraphs.

4.1.1. Procedure

The user study was divided into three stages: (i) preparation, (ii) main stage, and (iii) analysis of the collected data. They are discussed next:

S1 **Preparation**:Preparation started with the recruitment process, during which we contacted potential visitors who had experience in visiting both physical and virtual cultural heritage institutions and communicated the study motivation. Next, the people who were willing to take part in the study received more information about the study and we arranged a mutually agreed upon date and time to conduct the main stage of the study.

S2 **Main stage**: For the main stage of the user study, (i) the participants provided their consent; (ii) we presented them with the *MuMIA* system and the audio guide; (iii) the participants used *MuMIA* and the audio guide to explore the *School of Athens* exhibit; and (iv) we discussed with the participants their overall experience following a semi-structured interview approach in which we adopted the MUSETECH [52] model. During the aforementioned steps, we took notes and recorded the participants when necessary.

S3 **Analysis**:After all participants completed the main stage, we collected the data, transcribed the recordings, and performed a thematic analysis based on the dimensions described in the MUSETECH framework [52], which is discussed in more detail in the following paragraphs.

4.1.2. Participants

We recruited eleven people who had experience and interest in visiting both physical and virtual cultural heritage institutions. Seven participants were self-described as men and four participants were self-described as women. Their median age was 31 years; they had varying educational and professional backgrounds. We communicated the study, and we recruited the participants by inviting them to a virtual guide session, sending email invitations, directly contacting acquaintances of the authors, and posting call-out flyers on social media pages. All participants were informed about the study and provided their consent for data collection and analysis by the research team. The participation was voluntary, and the participants were free to withdraw at any time.

4.1.3. Visit Scenario

The visit scenario used by the study participants was the one described in Section 3.2 with *The School of Athens* as the art exhibit. The study participants could freely explore the exhibit by using *MuMIA* and a typical audio guide.

4.1.4. Apparatus

The study participants used the hardware and software provided by the Eye Tribe tracker (Section 3.3; Table 2) as the basis of the eye-tracking module. Moreover, they used a laptop computer, which was powerful enough to ensure good operation of the system (Intel Core i7, 8GB RAM), along with a 22-inch monitor. The apparatus was tested before the main study, and no problems (e.g., poor performance and glitches) were identified by the study participants or the researchers.

4.1.5. Thematic Analysis

To analyze the responses, we conducted an inductive thematic analysis as outlined by Braun [53]. We aimed to gain a rich understanding of the overall experience of the study participants when exploring a virtual exhibit through the use of multimodal interactions provided by *MuMIA*. The research team was involved in reviewing the transcribed data, generating the codes and the themes of the analysis, and interpreting the produced data through an iterative discussion process. The inductive thematic analysis resulted in codes that complied with the MUSETECH evaluation framework [52]. In their recent work, Damala et al. [52] presented the MUSETECH framework for evaluating cultural heritage technologies (e.g., software systems provided by a cultural heritage institution). It is based on three perspectives (professional, institution, and visitor) and four concepts (design, content, operation, and compliance). Each concept is divided into clusters and each cluster has various dimensions. Considering the goal of this study, we focused on the perspective of the *visitor*. A more comprehensive guide of the evaluation framework is provided by the MUSETECH companion [54].

4.2. Results

In examining the visitor experience when using *MuMIA*, we focused on the themes described in the MUSETECH framework. Note that the themes are not mutually exclusive. Each theme is discussed below with illustrative quotes, labeled by participant number (e.g., Participant #1); quotes that were not made in English were translated from the original language.

4.2.1. Experience Design and Narrative

The design of a cultural heritage tool is not only communicated in terms of message but also in terms of experience. Therefore, the cultural heritage tool should support the design of appropriate narratives and mediated experiences. The perspective of the visitor should be considered when trying to draw an artificial line between the design of the tool and the preoccupations of the cultural heritage institutions. The study participants found *MuMIA* more engaging than the typical audio guide, as it could guide them through micro-blocks of narratives that kept their attention during the visit. The study participants mentioned that they were absorbed and lost their sense of time when interacting with the multimodal interface.

> " *I totally lost myself in the visit experience! The feeling was great and I hope I had more time to explore more MuMIA exhibits and other aspects of the gallery.* " ∼ [Participant #4]

> " *The time slipped away so fast! I was fully concentrated and engaged with the exploration of the School of Athens; that categorization and the chunking of the narratives into micro-steps of acquiring knowledge definitely helped.* " ∼ [Participant #3]

That level of visitor engagement often results in increased learning, entertainment, and edutainment. The study participants found the micro-narratives as micro-blocks of knowledge that they could be more easily absorb rather than having a full detailed description of the art exhibit in a single interaction. Considering that such tasks are both instructive and entertaining and that pleasant and positive experiences are favored over unpleasant ones in our memory [55], the structure of *MuMIA* will help the edutainment goal of cultural heritage applications.

> " *I loved the features of MuMIA! The fact that I could visually explore the scene and select a figure or an instrument, that I knew nothing about, to receive information and know about it, was super helpful. For example, I didn't know what the central figure holds and I wouldn't be able to ask about it, if I couldn't gaze at it to show to the system somehow that I need more information about it.* " ∼ [Participant #10]

> " *MuMIA incorporates a great multi-modal interaction mechanism! The look-and-ask feature and the micro-blocks of narratives, I have the feeling that I could acquire quick knowledge and information about the art exhibit; which would normally take time when searching in the web and Wikipedia.* " ∼ [Participant #11]

Moreover, the multimodal interactions and the interconnected micro-narratives provided by *MuMIA* seem to have an affective impact and to improve cognitive and learning response when exploring art. The study participants could build connections between the areas of interest presented in the art exhibit in a more concrete and easy-to-follow way, which did not demand cognitive overload. We should mention that the volume of the presented cultural information is the same in *MuMIA* and a typical audio guide but that the micro-narratives and the self-exploration evoking helped visitors to perceive and absorb the cultural content in a more effective way.

> " *The system gave me the opportunity of self-exploration, meaning that I could explore, inspect, and process what I liked at a given time under a given motivation. So, it triggered*

my curiosity and it decreased my cognitive effort, helping me to better understand the concept of the art exhibit and hopefully remember it. " ∼ [Participant #2]

" *I am sure that a single pass through the whole information does not help you to remember, and thus learn, the presented cultural content, in contrast to a more goal-oriented approach, as the one supported in MuMIA. The visitors are implicitly guided to small chunks of information, which are more memorable and thus the visitors can recall the information in the future.* " ∼ [Participant #11]

The aforementioned results indicate the positive impact of experience design and narrative on the visitor experience. Considering that the visitors understood easily the design concept and what they could achieve with the specific technology, *MuMIA* can be used by cultural heritage professionals to integrate such multimodal technologies with art exhibitions and/or other information and communication technologies used in the cultural heritage institution. With the integration of multimodal technologies such as *MuMIA*, the cultural heritage institution would increase the level of innovation and business intelligence, leading to added value, recognizable brand name, uniqueness, and originality.

4.2.2. Interactions, Affordances, and Metaphors

The study participants had a positive perception of the utility and the usability of *MuMIA*. They mentioned that the interactions supported in the proposed system are meaningful, intuitive, and easy to use and follow. They could easily understand the concept of "look-and-ask" supported by the multimodal interaction, and they noticed no failure of the system.

" *While I had never used an eye-tracker before, its use was quite simple and intuitive. It was as simple as you look at an area that is interesting and you ask the system to provide more information about it.* " ∼ [Participant #11]

" *I noticed no failure of the system; it always provided information about the figure I was looking at. I enjoyed the fact that I could ask questions for combinations of figures, such as Michelangelo and Heraclitus, with the system providing me with more complex information.* " ∼ [Participant #10]

While *MuMIA*, similar to every software application and digital tool, includes an initial learning phase, it was found to be quite intuitive and the visitors did not face difficulties when learning to use it. The concept of "look-and-ask" was natural to them. Moreover, the familiar affordances and interaction metaphors contributed to the positive experience of the study participants, who mastered the interaction mechanisms quickly.

" *It was easy to learn how to use MuMIA; for people who face difficulties, maybe, it would be a good idea to include a tutorial phase to practice interactions before the actual visit phase.* " ∼ [Participant #6]

" *The concept was natural and friendly to me, so, I didn't need much time to understand how to use MuMIA. But, even if I had to spend more time to figure out how it works, I would had spent it, because the visitor experience was unique!* " ∼ [Participant #11]

While *MuMIA* was stable and credible, some study participants noticed a slight latency, which could affect the system responsiveness. Considering that the visitors' time is valuable and, often, visit time is limited, the responsiveness of a cultural heritage application is crucial. The study participants mentioned that the system did what was intended but that, sometimes, it took more time than they expected, and thus, it could have a negative impact on visitor experience.

> *" I didn't notice any failure during the operation, but a shortcoming that I noticed is that, at some times, there was a latency to follow my gaze. So, I caught myself trying to follow the [system] pointer rather than focusing on the content of the exhibit. "* ∼ [Participant #3]

Moreover, some study participants raised an issue about the ability to follow-up usage on other platforms or re-visits. *MuMIA* is based on eye-tracking, which may not be an established commercial technology, but it is growing fast and many techniques are deployed for accurate and fast eye tracking. In particular, recent advances in visual computing, gaze estimation algorithms, cameras, and processing power of computing devices have led to eye tracking being no longer constrained to desktop computers but also available on head-mounted displays [56–60], handheld mobile devices [61], and public displays [39]. Today, laptops such as Alienware 17 R4 and Acer Predator 21 X come with integrated eye trackers and smartphones such as the iPhone X and Huawei Mate 30 Pro are equipped with front-facing depth cameras capable of accurate gaze estimation.

> *" While I enjoyed MuMIA, I am not sure if it's a feasible solution for a museum context. I mean, what is the cost for a museum to have an eye-tracker for each exhibit or what is the cost for a visitor to have their own eye-tracker? In case that the cost is high, it might be a problem for adopting such solutions in real-life settings. "* ∼ [Participant #1]

Finally, regarding the interactions and affordances, the study participants found *MuMIA* to have a clear navigation system, which helped them not only navigate through the exhibit scene but also learn more about the presented art concept, as discussed previously. Moreover, the fact that the system supports interactions and, thus, visitor engagement using different senses helped the study participants use the system in a more efficient, effective, and fun way.

> *" Both navigating through the scene and navigating from eye-tracking to voice-commands and vice versa was super easy. It was more natural than using an audio guide, and I think, that it contributed into being more engaged with the cultural information. "* ∼ [Participant #5]

> *" Given that visitors can use MuMIA through multisensoriality, meaning both visual and auditory channels, definitely will help them navigate through the art concepts presented in each exhibit or even gallery, and thus, they will eventually have a more pleasant and informative visitor experience. "* ∼ [Participant #10]

The aforementioned results highlight that the interactions, affordances, and metaphors used in a system are crucial, and *MuMIA* achieved a high score, as its features felt natural to the study participants. This is important for a cultural heritage institution, as such novel but natural multimodal interactions create inspiring cultural experiences that differentiate the institution and make it unique. We should also mention that some participants raised issues and concerns regarding the ability to follow-up usage on other platforms and latency, which are important, but these can be overcome with the recent technological advances in the eye-tracking industry, and the information and communication technologies (e.g., increased computer power and faster network speeds).

4.2.3. Perceived Content Quality

The areas of interest provided a type of meaning-making mechanism for chunking of the cultural content, which was appraised by the study participants. Considering that the system allowed the study participants to access content that was interesting and relevant to them, they characterized the cultural content as having increased quality. Some of the study participants mentioned that they would like to have a personalization mechanism, especially for large art galleries, to have more focused attention on exhibits and areas that they are interested in.

> " *I was free to choose and filter what content I had access to, and this helped a lot at acquiring information and knowledge for art concepts that I was interested and not for concepts that were not relevant to my quest.* " ∼ [Participant #9]

4.2.4. Operation as Deployment and Setting Up

The study participants found that *MuMIA* can provide increased visitor experience quality and customer care in real-life settings, as it provided a visit that was flawless since they faced no problems during their interaction with the system. While this was expected for the audio-guide part, as it is a well-established technology in the cultural heritage domain, the multimodal interface introduced challenges that could lead to risks of failure during operation of the system. However, no such issues were identified by the study participants, since *MuMIA* operation and performance was trouble-free, and thus, *MuMIA* was a credible technological solution for supporting the multimodal interactions.

> " *The system can definitely be used in real-life settings, in a museum for example, as it is robust, credible, and performs flawlessly; with no doubt, it can increase the quality of the visitor experience when interacting with art exhibits.* " ∼ [Participant #7]

> " *To be entirely honest I was somewhat surprised that the system operation was that good! Given that it is based on a eye-tracking, I wouldn't be surprised if the prototype was not working well.* " ∼ [Participant #9]

A concern that was raised by the participants was whether *MuMIA* would be provided by the cultural heritage institutions or it would be supported by visitor-owned devices. During the last years, the proliferation of mobile devices has resulted in a "bring-your-own-device" culture. However, *MuMIA* is based on eye-tracking mechanisms, which while are credible are not mature enough to be integrated into each mobile device. However, some devices support such mechanisms, as discussed previously, and we argue that, in the near future, more and more mobile devices will support eye tracking, and thus, the "bring-your-own-device" culture will support the use of *MuMIA*.

> " *One thing that I am concerned about the tool is that it might not be integrable to the devices that the people bring with them when visiting a cultural heritage institution, such as their mobile devices; so, the set-up might be difficult in such conditions.* " ∼ [Participant #2]

Finally, the study participants found *MuMIA* to be robust, responsive, stable, and quick. The multimodal interactions worked well and provided an efficient and effective way for the visitor to communicate with the art exhibits.

> " *The responsiveness of the tool was impressive, given that it works with eye-tracking and voice commands. When gazing at some of the painting figures, it was giving me the opportunity to ask for information in zero time. I had also no problems with the voice commands, as it understood literally all the commands I gave.* " ∼ [Participant #5]

5. Discussion

Based on the results of the user study, the lessons learned and the implications of the works are directed towards three main themes: (i) evoking natural interactions, (ii) use of micro-narratives for supporting self-exploration, and (iii) implications beyond the cultural heritage domain We discuss these themes next.

5.1. Evoking Natural Interactions

The multimodal interface supported by *MuMIA* helped visitors to interact with the system naturally by following the "look-and-ask" approach. The participants visually explored the cultural scene (e.g., art exhibits) and identified areas (e.g., figures, shapes, and instruments) that attract their attention. After identifying such areas of interest,

they naturally asked the system to provide them with cultural information about them. Evoking natural interactions help people who communicate with software systems feel more confident; be more engaged [62]; have a decreased cognitive load [63]; have more fun [64]; and improve learnability [65], especially in augmented- and virtual-reality contexts, such as the one presented in the paper. Therefore, this is an important effect of the proposed system when considering the cultural heritage domain, where visitors typically use interaction systems for entertainment and learning purposes (i.e., edutainment goal).

Besides eye-tracking (i.e., visual interface), the voice communication channels (i.e., auditory interface) also help people be more engaged when interacting with systems, as they have been developed to support information exploration, to increase enjoyment with arts and entertainment, and to provide a complementary method of interaction with visual interfaces in the case study presented in the paper. The use of artificial intelligence for understanding voice commands and providing suitable verbal cultural information to the visitors contributed towards this direction. Artificial intelligence provides valuable application methods and techniques within the cultural heritage domain [66,67], as it introduces several advantages, such as the increased accessibility [68], the lower cost (e.g., use of voice modalities to build different versions of narratives compared to hiring a professional for recording voiceovers), the adaptability, and the transferability to new technologies [69].

5.2. Micro-Narratives for Supporting Self-Exploration through Multimodal Interaction

The contrast between the typical audio guide experience and the micro-blocks of narratives shows up in the participants' responses and is amplified by the multimodal interaction. Micro-narratives had a positive impact on perceiving the cultural content, which suggests that there is an obvious alternative to the conventional museum guide experience. This innovative way of transferring knowledge of a cultural heritage institution turns out to be not only entertaining but also quite effective. Each visitor is free to choose a unique path by setting questions to the system, fulfilling the purpose of self-exploration. Alongside micro-narratives, other types of structured knowledge acquisition blocks can also be considered, such as micro-augmentations [70], which have been shown to enhance the visitor experience as a whole, by surprising visitors, and by triggering their curiosity. Adopting such techniques that support visitors' self-exploration and knowledge acquisition could be used to enhance intra-group communication and perception, aiming to increase participation (e.g, between a family) and to build a more holistic visitor experience.

When designing micro-narratives, it is important to ensure that the content does not have a unique or static flow of reading, watching, or listening, but the visitors should be able to determine it dynamically [71]. Moreover, when creating a micro-narrative, we should ensure that, besides its obvious goal to meet the visitors' needs by providing them with a pleasant and inspiring experience, it should also be open for interacting and connecting various media elements. Chunking the cultural information in smaller units and delivering it to the visitors in a more natural and structured way is expected to evoke pleasant and positive experiences that would help towards a lower cognitive load and an increased edutainment. Therefore, it is crucial to engage various stakeholders (e.g., designers, institutions, curators, and educators) in a continual and interactive procedure aiming to cover various aspects [72], from basic ones that appear in the initial design stages (e.g., to identify the scope, the main idea, and the gist of a personal story in a creative, open-ended, and brainstorm-like process) to more complex ones that appear in the later production stages (e.g., authorship, copyright, and dissemination).

5.3. Implications Beyond the Cultural Heritage Domain

While the implementation of *MuMIA* presented in this paper aims to improve the user experience by supporting multimodal interactions, mainly of the visual and auditory systems, within cultural heritage contexts, *MuMIA* could be expanded to other domains, where there is a need to evoke self-exploration and edutainment when interacting with

software systems. Systems and applications such as *MuMIA* can have an impact on education, which is a domain rich in information that aims to help people acquire and build knowledge models. Considering that studies in this domain [14,15] provide evidence that multimodal interactions have a positive impact on the learning procedure with increased engagement, we argue that multimodal interactions through tools such as *MuMIA* would benefit the varying stakeholders in the education domain (e.g., both students and teachers).

Considering the findings from the evaluation study of *MuMIA*, we envisage the design of a multimodal framework based on gaze and voice interactions that could be adjusted to support diverse domains. Each domain has unique characteristics, as it aims to accomplish specific objectives (e.g., edutainment in cultural heritage and learning in education) for specific stakeholders (e.g., visitors and curators for cultural heritage, and students and teachers for education) through specific context dependencies (e.g., activities, time constraints, and space limitations). Studies, such as the reported one, could be conducted to identify the domain-specific characteristics in order for the framework to adjust on them and to better support the objectives of the stakeholders. Other modalities and combinations of them, such as tactile interaction, can be used as complementary to gaze and voice interactions—or even to replace them in case they provide more satisfactory and feasible solutions—to better serve human–computer interactions and to support users in achieving their goals effectively and efficiently.

6. Limitations, Future Work, and Ethical Considerations

6.1. Limitations and Future Work

The paper focused on the visitor, who is the main actor of such systems. However, more actors can be identified, considering that cultural heritage is a domain that attracts diverse types of stakeholders, such as cultural heritage institution managers, educators, designers, creators, and tourist guides. Each of them has unique characteristics and roles in varying stages of presenting, managing, and preserving the cultural content and information. While the paper did not discuss their role in depth, the architecture of the *MuMIA* system allows for the integration of mechanisms that support their roles (e.g., an educator could use an authoring tool to manage what information would be available for a tour with a specific theme and under a limited amount of time). As an immediate future work step, we aim to expand and evaluate the system to support other cultural heritage stakeholders in achieving their goals.

A limitation of the evaluation study presented in the paper is the limited sample size. Due to the COVID-19 pandemic and the nature of the study (i.e., physical presence of the participants in the lab room), only eleven people took part in the study. However, considering the characteristics of the study (e.g., experienced participants who were interested in the concepts presented in the study, qualitative approach, and in-person study), the sample size is similar to the sample sizes reported in works in the human–computer interaction community [73]. Therefore, we argue that the results could be generalizable regarding the visitors' experience on such cultural heritage schemes that are based on multimodal interactions and technologies. As a future step, we aim to increase the sample size, either by engaging more participants physically or by integrating remote-access eye-tracking modules, in order to validate the results in a larger population.

Moreover, regarding the participants' profiles, we should mention that we focused on people with specific characteristics: young people who are interested in arts. We made this decision to obtain an increased validity, considering the limited sample size. However, it would be interesting to investigate the perceptions, cognitive load, and visitor experience for people who are not interested in or are not familiar with art concepts. In the same vein, it would be interesting to evaluate the visitor experience through *MuMIA* from an age perspective, considering that age can influence perceived usefulness and satisfaction of technology in a cultural heritage setting [74]. Therefore, as a future step, we aim to recruit participants with diverse characteristics (e.g., different age groups, different cultural

backgrounds, and different levels of interest in arts) to study the generalizability of our proposed model.

Finally, we should stress that the use of eye trackers might influence visitor experience, as it is a technology that is not fully available yet. In physical visits, eye tracking could be applied through the use of eye-tracking glasses, while in virtual visits, eye tracking could be applied through integration into computer systems. The first entails the issue of familiarity, which could be amplified because of the use of other external devices, such as headphones. However, gaze and voice interactions can be integrated into head-mounted displays (e.g., Microsoft Hololens), which provide a rich spatial user interface. Research has shown that such systems deliver wider accessibility and enhanced user experience [75]. Moreover, visitors tend to have positive reactions when using eye-tracking glasses along with audio guides [46]; they also tend to feel more comfortable after a short time, as they tend to interact with the system as they would normally [44]. Regarding the integration of eye trackers into computer systems, it entails difficulties in terms of calibration. Recent research moves towards calibration-free eye tracking, which will increase the user experience as the interaction between the users and the eye trackers will be performed unobtrusively.

6.2. Ethical Considerations

Our research incorporated appropriate consideration of ethical issues into the design, the conduction, and the analysis of the user study. Our research involved interaction with human subjects, and it was performed following context-specific ethical guidelines. The researchers actively respected the human rights and dignities of all those involved in the studies and appropriately addressed questions of consent, capacity, power relations, deception, confidentiality, and privacy. All people participated voluntarily in our studies, and they all agreed to and signed a consent form for their interactions with *MuMIA* being recorded and analyzed anonymously as part of experimental user studies of the research group. The study participants could bring the experiment to an end at any time and for any reason. During the experiment, the researcher had the authority to bring it to an end if there was probable cause to believe that it could harm the participants in any way. All participants were informed about the experimental procedure and the rights they had as volunteers. However, no further details about the aim of the studies were provided to them to avoid bias effects.

7. Conclusions

Our work in this paper reveals the benefits of multimodal interaction solutions in cultural heritage. We presented the design and evaluation of an interactive system that implemented multimodality through eye tracking and voice commands that allowed the visitors to refer to any exhibit area that provides cultural information and to ask questions about it, aiming to build a better understanding of the presented art contexts. The results of the evaluation study revealed that *MuMIA* was more engaging than the typical audio guide, resulting in increased edutainment. The results provided evidence that the concept of the natural "look-and-ask" approach contributed to increased attention and enhanced visitor experience, through the use of micro-narratives as micro-blocks of knowledge transfer. The implications and the lessons learned from our work provide a step enlightening the dimensions evoking natural interactions within cultural heritage environments and adopting micro-narrative approaches for self-exploration and understanding of cultural content.

Author Contributions: Conceptualization, G.E.R.; methodology, G.E.R; validation, G.E.R. and G.K.; investigation, G.E.R. and G.K.; writing—original draft preparation, G.E.R. and G.K.; writing—review and editing, G.E.R. and G.K. and C.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Approval was obtained from the institutional ethics committee.

Informed Consent Statement: Study participants provided consent for data collection and analysis.

Data Availability Statement: Data availability is possible upon request.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AOI	Area of Interest
API	Application Programming Interface
MuMIA	Multi-Modal Interactions in Art
SQL	Structured Query Language

References

1. Bekele, M.K.; Pierdicca, R.; Frontoni, E.; Malinverni, E.S.; Gain, J. A Survey of Augmented, Virtual, and Mixed Reality for Cultural Heritage. *J. Comput. Cult. Herit.* **2018**, *11*, 1–36. [CrossRef]
2. Walmsley, A.P.; Kersten, T.P. The Imperial Cathedral in Königslutter (Germany) as an Immersive Experience in Virtual Reality with Integrated 360° Panoramic Photography. *Appl. Sci.* **2020**, *10*, 1517. [CrossRef]
3. Edler, D.; Keil, J.; Wiedenlübbert, T.; Sossna, M.; Kühne, O.; Dickmann, F. Immersive VR Experience of Redeveloped Post-industrial Sites: The Example of "Zeche Holland" in Bochum-Wattenscheid. *KN J. Cartogr. Geogr. Inf.* **2019**, *69*, 267–284. [CrossRef]
4. Raptis, G.E.; Katsini, C.; Chrysikos, T. CHISTA: Cultural Heritage Information Storage and reTrieval Application. In *In Digital Heritage: Progress in Cultural Heritage: Documentation, Preservation, and Protection*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 163–170. [CrossRef]
5. Turk, M. Multimodal interaction: A review. *Pattern Recognit. Lett.* **2014**, *36*, 189–195. [CrossRef]
6. Cutugno, F.; Leano, V.A.; Rinaldi, R.; Mignini, G. Multimodal Framework for Mobile Interaction. In Proceedings of the International Working Conference on Advanced Visual Interfaces, Capri Island Italy, 22–26 May 2012; pp. 197–203. [CrossRef]
7. Oviatt, S. Advances in robust multimodal interface design. *IEEE Ann. Hist. Comput.* **2003**, *23*, 62–68. [CrossRef]
8. Xiao, B.; Girand, C.; Oviatt, S. Multimodal integration patterns in children. In Proceedings of the Seventh International Conference on Spoken Language Processing, Denver, CO, USA, 16–20 September 2002.
9. Oviatt, S.; Lunsford, R.; Coulston, R. Individual differences in multimodal integration patterns: What are they and why do they exist? In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Portland, OR, USA, 2–7 April 2005; pp. 241–249.
10. Oviatt, S.; Cohen, P.; Wu, L.; Duncan, L.; Suhm, B.; Bers, J.; Holzman, T.; Winograd, T.; Landay, J.; Larson, J.; others. Designing the user interface for multimodal speech and pen-based gesture applications: State-of-the-art systems and future research directions. *Hum. Comput. Interact.* **2000**, *15*, 263–322. [CrossRef]
11. Jaimes, A.; Sebe, N. Multimodal human–computer interaction: A survey. *Comput. Vis. Image Underst.* **2007**, *108*, 116–134. [CrossRef]
12. Li, T.J.J.; Azaria, A.; Myers, B.A. SUGILITE: creating multimodal smartphone automation by demonstration. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, Denver, CO, USA, 6–11 May 2017; pp. 6038–6049.
13. Srinivasan, A.; Lee, B.; Henry Riche, N.; Drucker, S.M.; Hinckley, K. InChorus: Designing consistent multimodal interactions for data visualization on tablet devices. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 25–30 April 2020; pp. 1–13.
14. Aslan, S.; Alyuz, N.; Tanriover, C.; Mete, S.E.; Okur, E.; D'Mello, S.K.; Arslan Esme, A. Investigating the impact of a real-time, multimodal student engagement analytics technology in authentic classrooms. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Glasgow, UK, 4–9 May 2019; pp. 1–12.
15. Alyuz, N.; Okur, E.; Genc, U.; Aslan, S.; Tanriover, C.; Esme, A.A. An unobtrusive and multimodal approach for behavioral engagement detection of students. In Proceedings of the 1st ACM SIGCHI International Workshop on Multimodal Interaction for Education, Glasgow UK, 13 November 2017; pp. 26–32.
16. Bedri, A.; Li, D.; Khurana, R.; Bhuwalka, K.; Goel, M. Fitbyte: Automatic diet monitoring in unconstrained situations using multimodal sensing on eyeglasses. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 25–30 April 2020; pp. 1–12.
17. Speicher, M.; Nebeling, M. Gesturewiz: A human-powered gesture design environment for user interface prototypes. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Montréal, Canada, 21–26 April 2018; pp. 1–11.
18. Salminen, K.; Farooq, A.; Rantala, J.; Surakka, V.; Raisamo, R. Unimodal and multimodal signals to support control transitions in semiautonomous vehicles. In Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Utrecht, The Netherlands, 22–25 September 2019; pp. 308–318.

19. Politis, I.; Brewster, S.; Pollick, F. Language-based multimodal displays for the handover of control in autonomous cars. In Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Nottingham UK, 1–3 September 2015; pp. 3–10.
20. Khamis, M.; Alt, F.; Hassib, M.; von Zezschwitz, E.; Hasholzner, R.; Bulling, A. Gazetouchpass: Multimodal authentication using gaze and touch on mobile devices. In Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems, San Jose, CA, USA, 7–16 May 2016; pp. 2156–2164.
21. Lee, J.; Han, J.; Lee, G. Investigating the information transfer efficiency of a 3x3 watch-back tactile display. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, Seoul, Korea, 18–23 April 2015; pp. 1229–1232.
22. He, L.; Xu, C.; Xu, D.; Brill, R. PneuHaptic: delivering haptic cues with a pneumatic armband. In Proceedings of the 2015 ACM International Symposium on Wearable Computers, Osaka, Japan, 7–11 September 2015; pp. 47–48.
23. Lee, J.; Lee, G. Designing a non-contact wearable tactile display using airflows. In Proceedings of the 29th Annual Symposium on User Interface Software and Technology, Tokyo, Japan, 16–19 October 2016; pp. 183–194.
24. Shim, Y.A.; Lee, J.; Lee, G. Exploring multimodal watch-back tactile display using wind and vibration. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Montréal, QC, Canada, 21–26 April 2018; pp. 1–12.
25. Christou, C.; Angus, C.; Loscos, C.; Dettori, A.; Roussou, M. A versatile large-scale multimodal VR system for cultural heritage visualization. In Proceedings of the ACM Symposium on Virtual Reality Software and Technology, Limassol, Cyprus, 1–3 November 2006; pp. 133–140.
26. Liarokapis, F.; Petridis, P.; Andrews, D.; de Freitas, S. Multimodal serious games technologies for cultural heritage. In *Mixed Reality and Gamification for Cultural Heritage*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 371–392.
27. Dimitropoulos, K.; Manitsaris, S.; Tsalakanidou, F.; Denby, B.; Crevier-Buchman, L.; Dupont, S.; Nikolopoulos, S.; Kompatsiaris, I.; Charisis, V.; Hadjileontiadis, L.; et al. A Multimodal Approach for the Safeguarding and Transmission of Intangible Cultural Heritage: The Case of i-Treasures. *IEEE Intell. Syst.* **2018**, 1–1. [CrossRef]
28. Santoro, C.; Paterno, F.; Ricci, G.; Leporini, B. A multimodal mobile museum guide for all. In Proceedings of the Mobile Interaction with the Real World Workshop, Singapore, 9 September 2007; pp. 21–25.
29. Ho, C.M.; Nelson, M.E.; Müeller-Wittig, W. Design and implementation of a student-generated virtual museum in a language curriculum to enhance collaborative multimodal meaning-making. *Comput. Educ.* **2011**, *57*, 1083–1097. [CrossRef]
30. Santangelo, A.; Augello, A.; Gentile, A.; Pilato, G.; Gaglio, S. A Chat-Bot Based Multimodal Virtual Guide for Cultural Heritage Tours; In Proceedings of the 2006 International Conference on Pervasive Systems & Computing, Las Vegas, NV, USA, 26–29 June 2006; pp. 114–120.
31. Carmichael, J.; Larson, M.; Marlow, J.; Newman, E.; Clough, P.; Oomen, J.; Sav, S. Multimodal indexing of digital audio-visual documents: A case study for cultural heritage data. In Proceedings of the 2008 International Workshop on Content-Based Multimedia Indexing, London, UK, 18–20 June 2008; pp. 93–100.
32. Cutugno, F.; Dell'Orletta, F.; Poggi, I.; Savy, R.; Sorgente, A. The CHROME Manifesto: Integrating Multimodal Data into Cultural Heritage Resources. In Proceedings of the Fifth Italian Conference on Computational Linguistics, Torino, Italy, 10–12 December 2018.
33. Neto, J.N.; Silva, R.; Neto, J.P.; Pereira, J.M.; Fernandes, J. Solis' Curse-A Cultural Heritage game using voice interaction with a Virtual Agent. In Proceedings of the 2011 Third International Conference on Games and Virtual Worlds for Serious Applications, Athens, Greece, 4–6 May 2011; pp. 164–167.
34. D'Auria, D.; Di Mauro, D.; Calandra, D.M.; Cutugno, F. A 3D audio augmented reality system for a cultural heritage management and fruition. *J. Digit. Inf. Manag.* **2015**, *13*. [CrossRef]
35. Sernani, P.; Vagni, S.; Falcionelli, N.; Mekuria, D.N.; Tomassini, S.; Dragoni, A.F. Voice interaction with artworks via indoor localization: A vocal museum. In Proceedings of the International Conference on Augmented Reality, Virtual Reality and Computer Graphics. Springer: Cham, Switzerland. Lecce, Italy, 8–11 June 2020; pp. 66–78.
36. Ferracani, A.; Faustino, M.; Giannini, G.X.; Landucci, L.; Del Bimbo, A. Natural experiences in museums through virtual reality and voice commands. In Proceedings of the 25th ACM international conference on Multimedia, Mountain View, CA, USA, 23–27 October 2017; pp. 1233–1234.
37. Picot, A.; Charbonnier, S.; Caplier, A. Drowsiness detection based on visual signs: blinking analysis based on high frame rate video. In Proceedings of the 2010 IEEE Instrumentation & Measurement Technology Conference, Austin, TX, USA, 3–6 May 2010; pp. 801–804.
38. Xu, G.; Zhang, Z.; Ma, Y. Improving the performance of iris recogniton system using eyelids and eyelashes detection and iris image enhancement. In Proceedings of the 2006 5th IEEE International Conference on Cognitive Informatics, Beijing, China, 17–19 July 2006; Volume 2, pp. 871–876.
39. Zhang, Y.; Chong, M.K.; Müller, J.; Bulling, A.; Gellersen, H. Eye Tracking for Public Displays in the Wild. *Pers. Ubiquitous Comput.* **2015**, *19*, 967–981. [CrossRef]
40. Katsini, C.; Abdrabou, Y.; Raptis, G.E.; Khamis, M.; Alt, F. The Role of Eye Gaze in Security and Privacy Applications: Survey and Future HCI Research Directions. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 25–30 April 2020; pp. 1–21. [CrossRef]

41. Raptis, G.E.; Fidas, C.; Avouris, N. Do Game Designers' Decisions Related to Visual Activities Affect Knowledge Acquisition in Cultural Heritage Games? An Evaluation From a Human Cognitive Processing Perspective. *J. Comput. Cult. Herit.* **2019**, *12*, 4:1–4:25. [CrossRef]
42. Rainoldi, M.; Neuhofer, B.; Jooss, M. Mobile eyetracking of museum learning experiences. In *Information and Communication Technologies in Tourism 2018*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 473–485.
43. Pelowski, M.; Leder, H.; Mitschke, V.; Specker, E.; Gerger, G.; Tinio, P.P.; Vaporova, E.; Bieg, T.; Husslein-Arco, A. Capturing aesthetic experiences with installation art: An empirical assessment of emotion, evaluations, and mobile eye tracking in Olafur Eliasson's "Baroque, Baroque!". *Front. Psychol.* **2018**, *9*, 1255. [CrossRef] [PubMed]
44. Raptis, G.E.; Fidas, C.; Katsini, C.; Avouris, N. A Cognition-Centered Personalization Framework for Cultural-Heritage Content. *User Model. User Adapt. Interact.* **2019**, *29*, 9–65. [CrossRef]
45. Pierdicca, R.; Paolanti, M.; Quattrini, R.; Mameli, M.; Frontoni, E. A Visual Attentive Model for Discovering Patterns in Eye-Tracking Data—A Proposal in Cultural Heritage. *Sensors* **2020**, *20*, 2101. [CrossRef] [PubMed]
46. Mokatren, M.; Kuflik, T.; Shimshoni, I. Exploring the potential of a mobile eye tracker as an intuitive indoor pointing device: A case study in cultural heritage. *Future Gener. Comput. Syst.* **2018**, *81*, 528–541. [CrossRef]
47. Toyama, T.; Kieninger, T.; Shafait, F.; Dengel, A. Museum Guide 2.0–an eye-tracking based personal assistant for museums and exhibits. In Proceedings of the International Conference "Re-Thinking Technology in Museums", Limerick, Ireland, 26–27 May 2011; pp. 103–110.
48. Garbutt, M.; East, S.; Spehar, B.; Estrada-Gonzalez, V.; Carson-Ewart, B.; Touma, J. The embodied gaze: Exploring applications for mobile eye tracking in the art museum. *Visit. Stud.* **2020**, *23*, 82–100. [CrossRef]
49. Cantoni, V.; Merlano, L.; Nugrahaningsih, N.; Porta, M. Eye Tracking for Cultural Heritage: A Gaze-Controlled System for Handless Interaction with Artworks. In Proceedings of the 17th International Conference on Computer Systems and Technologies 2016, Palermo, Italy, 23–24 June 2016; pp. 307–314. [CrossRef]
50. Sibert, L.E.; Jacob, R.J.K. Evaluation of Eye Gaze Interaction. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Hague, The Netherlands, 1–6 April 2000; pp. 281–288. [CrossRef]
51. Helmert, J.R.; Joos, M.; Pannasch, S.; Velichkovsky, B.M. Two visual systems and their eye movements: Evidence from static and dynamic scene perception. In Proceedings of the Annual Meeting of the Cognitive Science Society, Stresa, Italy, 21–23 July 2005; Volume 27.
52. Damala, A.; Ruthven, I.; Hornecker, E. The MUSETECH Model: A Comprehensive Evaluation Framework for Museum Technology. *ACM J. Comput. Cult. Herit.* **2019**, *12*, 7:1–7:22. [CrossRef]
53. Braun, Virginiaand Clarke, V. Using thematic analysis in psychology. *Qual. Res. Psychol.* **2006**, *3*, 77–101. doi:10.1191/1478088706qp063oa. [CrossRef]
54. Damala, A.; Ruthven, I.; Hornecker, E. The MUSETECH Companion: Navigating the Matrix. In *Guide or Manual*; University of Strathclyde: Glasgow, UK, 2019.
55. Falk, J.H.; Dierking, L.D. *The Museum Experience Revisited*; Routledge: Oxfordshire, UK, 2016.
56. Ahuja, K.; Islam, R.; Parashar, V.; Dey, K.; Harrison, C.; Goel, M. EyeSpyVR: Interactive Eye Sensing Using Off-the-Shelf, Smartphone-Based VR Headsets. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2018**, *2*, 57:1–57:10. [CrossRef]
57. Fuhl, W.; Tonsen, M.; Bulling, A.; Kasneci, E. Pupil Detection for Head-mounted Eye Tracking in the Wild: An Evaluation of the State of the Art. *Mach. Vis. Appl.* **2016**, *27*, 1275–1288. [CrossRef]
58. George, C.; Khamis, M.; Buschek, D.; Hussmann, H. Investigating the Third Dimension for Authentication in Immersive Virtual Reality and in the Real World. In Proceedings of the 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), Osaka, Japan, 23–27 May 2019; pp. 277–285. [CrossRef]
59. Hirzle, T.; Gugenheimer, J.; Geiselhart, F.; Bulling, A.; Rukzio, E. A Design Space for Gaze Interaction on Head-mounted Displays. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Glasgow, UK, 4–9 May 2019; pp. 625:1–625:12. [CrossRef]
60. Liu, S.; Wilson, J.; Xia, Y. Eye Gazing Passcode Generation Crossing Augmented Reality (AR) and Virtual Reality (VR) Devices. US Patent 9824206B1, 21 November 2017.
61. Khamis, M.; Alt, F.; Bulling, A. The Past, Present, and Future of Gaze-enabled Handheld Mobile Devices: Survey and Lessons Learned. In Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services, Barcelona, Spain, 3–6 September 2018. [CrossRef]
62. Brondi, R.; Avveduto, G.; Carrozzino, M.; Tecchia, F.; Alem, L.; Bergamasco, M. Immersive Technologies and Natural Interaction to Improve Serious Games Engagement. In *Lecture Notes in Computer Science*; Springer: Berlin/Heidelberg, Germany: 2016; pp. 121–130. [CrossRef]
63. Galais, T.; Delmas, A.; Alonso, R. Natural interaction in virtual reality. In Proceedings of the 31st Conference on l'Interaction Homme-Machine Adjunct, Grenoble, France, 10–13 December 2019. [CrossRef]
64. McMahan, R.P.; Alon, A.J.D.; Lazem, S.; Beaton, R.J.; Machaj, D.; Schaefer, M.; Silva, M.G.; Leal, A.; Hagan, R.; Bowman, D.A. Evaluating natural interaction techniques in video games. In Proceedings of the 2010 IEEE Symposium on 3D User Interfaces (3DUI). IEEE, Waltham, MA, USA, 20–21 March 2010. [CrossRef]
65. Cantoni, V.; Cellario, M.; Porta, M. Perspectives and challenges in e-learning: towards natural interaction paradigms. *J. Vis. Lang. Comput.* **2004**, *15*, 333–345. [CrossRef]

66. Pisoni, G.; Díaz-Rodríguez, N.; Gijlers, H.; Tonolli, L. Human-Centered Artificial Intelligence for Designing Accessible Cultural Heritage. *Appl. Sci.* **2021**, *11*, 870. [CrossRef]
67. Bordoni, L.; Ardissono, L.; Barcelo, J.; Chella, A.; de Gemmis, M.; Gena, C.; Iaquinta, L.; Lops, P.; Mele, F.; Musto, C.; et al. The contribution of AI to enhance understanding of Cultural Heritage. *Intell. D* **2013**, *7*, 101–112. [CrossRef]
68. Díaz-Rodríguez, N.; Pisoni, G. Accessible cultural heritage through explainable artificial intelligence. In Proceedings of the Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization, Genoa, Italy, 12–18 June 2020; pp. 317–324.
69. Caggianese, G.; De Pietro, G.; Esposito, M.; Gallo, L.; Minutolo, A.; Neroni, P. Discovering Leonardo with artificial intelligence and holograms: A user study. *Pattern Recognit. Lett.* **2020**, *131*, 361–367. [CrossRef]
70. Antoniou, A.; O'Brien, J.; Bardon, T.; Barnes, A.; Virk, D. Micro-Augmentations: Situated Calibration of a Novel Non-Tactile, Peripheral Museum Technology. In Proceedings of the 19th Panhellenic Conference on Informatics, Athens, Greece, 1–3 October 2015, pp. 229—234. [CrossRef]
71. Rizvic, S.; Djapo, N.; Alispahic, F.; Hadzihalilovic, B.; Cengic, F.F.; Imamovic, A.; Okanovic, V.; Boskovic, D. Guidelines for interactive digital storytelling presentations of cultural heritage. In Proceedings of the 2017 9th International Conference on Virtual Worlds and Games for Serious Applications (VS-Games), Athens, Greece, 6–8 September 2017. [CrossRef]
72. Sylaiou, S.; Dafiotis, P. Storytelling in Virtual Museums: Engaging A Multitude of Voices. In *Visual Computing for Cultural Heritage*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 369–388.[CrossRef]
73. Caine, K. Local Standards for Sample Size at CHI. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, CA, USA, 7–16 May 2016; p. 981–992. [CrossRef]
74. Kang, J.; Jang, J.; Jeong, C. Understanding museum visitor satisfaction and revisit intentions through mobile guide system: moderating role of age in museum mobile guide adoption. *Asia Pac. J. Tour. Res.* **2018**, *23*, 95–108. [CrossRef]
75. Hammady, R.; Ma, M.; Strathearn, C. User experience design for mixed reality: a case study of HoloLens in museum. *Int. J. Technol. Mark.* **2019**, *13*, 354–375. [CrossRef]