A Review of Interaction Techniques for Immersive Environments

Becky Spittle, Maite Frutos-Pascual, Chris Creed and Ian Williams

Abstract—The recent proliferation of immersive technology has led to the rapid adoption of consumer-ready hardware for Augmented Reality (AR) and Virtual Reality (VR). While this increase has resulted in a variety of platforms that can offer a richer interactive experience, the advances in technology bring more variability in display types, interaction sensors and use cases. This provides a spectrum of device-specific interaction possibilities, with each offering a tailor-made solution for delivering immersive experiences to users, but often with an inherent lack of standardisation across devices and applications. To address this, a systematic review and an evaluation of explicit, task-based interaction methods in immersive environments are presented in this paper. A corpus of papers published between 2013 and 2020 is reviewed to thoroughly explore state-of-the-art user studies, which investigate input methods and their implementation for immersive interaction tasks (pointing, selection, translation, rotation, scale, viewport, menu-based and abstract). Focus is given to how input methods have been applied within the spectrum of immersive technology (AR, VR, XR). This is achieved by categorising findings based on display type, input method, study type, use case and task. Results illustrate key trends surrounding the benefits and limitations of each interaction technique and highlight the gaps in current research. The review provides a foundation for understanding the current and future directions for interaction studies in immersive environments, which, at this pivotal point in XR technology adoption, provides routes forward for achieving more valuable, intuitive and natural interactive experiences.

Index Terms—Augmented Reality, Virtual Reality, HCI, Interaction, Input, Tasks, Usability, Multimodal, Immersive

1 Introduction

Immersive technologies encompass the spectrum of Virtual Reality (VR), Augmented Reality (AR) and Mixed Reality (MR) environments, which collectively are referred to as Extended Reality (XR). Over recent years, the technical advances in immersive technology have prompted an unprecedented growth in commercial hardware and software capabilities, which have taken XR from concept through to a near-natural, fully commercial possibility [6, 84].

Commonly, immersive technologies have been developed as an expansion of methods, theories and interaction approaches provisioned by 2D displays [4], to introduce novel ways of interfacing with computer-generated information [2]. Immersive technologies allow tasks to be performed directly in a real or virtual 3D spatial context [2], and go beyond the sedentary nature of 2D environments, to provide more enriched and engaging 3D experiences [4].

Interaction is essential in 3D immersive environments, yet is arguably more complicated to deliver effectively than in other fields of human-computer interaction [2]. As XR interfaces require novel configurations of interface components, namely devices, techniques and metaphors, a broader range of input and output modalities for interaction are provided, resulting in a myriad of opportunities to design new interaction approaches [46].

The range of approaches used for XR interaction are often more closely aligned with how we apply human-to-human interaction (namely speech, gaze, hand gesture and touch [56]) than traditional desktop environments [2]. This naturally creates a range of interaction possibilities, which can be tailored to our senses and communication methods and mapped to different use cases (i.e. based on environment, context, activity and application).

Interactions include aural cues (i.e. speech and para-linguistics), visual cues (i.e. gaze and gesture) and environmental information (i.e. object manipulation, writing and drawing) [6]. By exercising logic, considering context and building on an extensive body of interaction research, designers and developers of immersive technology are empowered to create the most relevant interpretations of human-to-human interaction and apply this understanding to deliver more natural interactions for XR environments [40].

Although many developments have been made, that transfer a natural level of interaction, XR researchers are unable to directly apply full comprehension of human-human communication for interaction with all virtual content. This is primarily because immersive technologies provide additional opportunities that exceed what is capable with human-human interactions [6], namely by offering advanced or beyond human interaction possibilities (i.e. speech, head and gestures for object control and manipulation) [56].

Furthermore, as immersive technologies combine different levels of reality and virtuality (i.e. real and virtual objects coexisting in the same immersive environment), the interaction paradigms employed are highly dependent on the nature of the content the user is interacting with, and interactions will differ between real and virtual objects. For example, interactions in AR with a virtual object can be applied more flexibly (i.e. the user able to execute object transformations at a distance [97, 101]). However, if the content is real, the same extended interaction possibility is not viable.

These inconsistencies across XR technologies present an interaction paradox for users. This also creates a spectrum of challenges and design choices to provide the most realistic, usable and valuable immersive experiences.

1.1 Transferable Interactions

As we move towards ubiquitous applications of immersive technologies [32] and to avoid the ad-hoc development of bespoke XR solutions, it is essential to understand how inputs can be best mapped to different tasks for XR environments.

Interaction in immersive environments can be divided into explicit and implicit inputs. Explicit interactions are defined as any intentional input provided to execute distinct tasks and manipulate the scene, notably to interact with virtual content within the 3D environment [79]. Implicit interactions are a combination of inherent motion and location awareness within the interactive space, which triggers an inherent interaction (i.e. walking around a spatially registered object).

Explicit interactions can be based on either a single stream of data, i.e. solely hand, head or speech information (unimodal input), or more than one input can be used to manipulate the scene, i.e. to separate functions for different tasks, or to add an extra source of data to improve system reliability (multimodal input) [56]. However, there is a current lack of clarity around what context, situation and application the advances between unimodal and multimodal interaction are best suited.

Additionally, as applications employ various modes of input and output, they often imply separate system requirements (i.e. the hard-

• *DMT Lab, School of Computing and Digital Technology, Birmingham City University, United Kingdom. E-mail: becky.spittle@bcu.ac.uk.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

ware and software/logic that is required [79]). Therefore, the range of existing and emerging immersive devices and technologies offer diverse interaction mappings and system architectures. This results in a lack of standards, a staggered workflow for content producers and a less seamless, immersive experience for the end-user [32, 84].

To address these points, the review specifically considers interaction techniques to perform explicit tasks (i.e. selection, translation, rotation etc.) and user evaluation/testing (i.e. captured objective and subjective study measures) in regards to input methods. This is to propose an evaluation of recent work, highlight the advantages and disadvantages of different unimodal and multimodal interaction techniques (based on freehand, head-based, speech-based and hardware-based inputs) and recommend the most valuable research directions.

By exploring how content producers can fully reap the benefits of XR interaction capabilities, we can work towards making interaction more standardised and transferable across the range of XR tasks, devices and use cases.

The paper is structured as follows. Section 2 details the methods employed to capture the data for review. This includes the inclusion/exclusion criteria, the categorisation and factors for analysis. Section 3 details the analysis of the literature, presenting quantitative values and insights for the factors under review. These are presented under the primary categories of XR devices, namely Handheld, Headworn and Multiple-Displays. Section 4 discusses the key findings for each of the factors in the review, with section 5 providing recommendations, conclusions and directions for future work.

2 METHOD

The focus of this review surrounds immersive technologies and provides a true representation of the input techniques explored for XR interaction. Searches were not restricted to specific publishing venues, meaning a range of papers were considered. This included full and short papers sourced from journals and conference proceedings.

Paper quality was assessed based on how thoroughly the research addressed the factors defined as key areas for exploration (in table 2). Although affiliations were taken into account and several highly cited papers were included in the review, publication impact was not a primary concern. Population size was also noted, to help determine the impact of surveyed works, yet the number of participants did not influence whether a paper was included.

The methods that were applied to filter, collect and prepare the data for analysis are further defined in the following subsections.

2.1 Data Collection

A sample of 182 eligible papers was collated from ACM Digital Library, IEEE Explore and other databases prevalent in the fields of HCI and computer science, such as Springer, Elsevier, IFIP and Oxford Press.

To define the corpus of papers for consideration, information was required for factors surrounding the type of study, display used, testing conditions, experimental set-up/design and the data collected.

A variation of search terms were applied to the advanced search engines of the chosen databases, as categorised in table 1. Search terms concerning 'Study Type' or 'Technology' were referenced in the Title. The remaining search terms were searched within the Abstract, apart from those classified as 'General', where they were applied to the body of text.

2.1.1 Inclusion/Exclusion Criteria

To inform inclusion/exclusion criteria, the review conducted by Bai et al. [7] was considered. This work provides a reference point for current evaluation techniques, trends and challenges, which are provided to benefit XR researchers intending to design, conduct and interpret usability evaluations. Consequently, their considerations were deemed transferable for this review.

To ensure papers were relevant and comparable, they had to meet the following criteria:

- **Display**: They should consider a) Headworn (HMDs or smart glasses), b) Handheld (wireless smart devices), or c) Static (monitor) displays. These display types were targeted as they are ubiquitous, consumer-level devices that are also widely employed for XR research. However, where output was delivered to the participant via a monitor, studies were also required to consider either a headworn or handheld display.

  Even though the display conditions included are heterogeneous (papers reporting on multiple combinations of hardware setups), those only considering less accessible displays, such as CAVE, smart mirrors and projection environments were also excluded. This is because these display types are more restricted to specific domains (i.e. applicable for ad-hoc, research and business applications, as opposed to more generalisable consumer interactions).

- **Input**: Studies had to concern one or more of the following inputs a) Speech b) Head c) Freehand d) Hardware-Based interaction with handheld smart devices (i.e. touchscreen or 6-DoF motion gestures). These inputs were defined as they are the most widespread and applicable to interaction with the targeted display devices. These inputs are also generally straightforward to implement using the built-in components of XR devices.

  Although some studies used hardware switches/controllers, eye gaze and marker-based interaction, they were only included for review if they considered at least one of the targeted inputs (as defined in table 2). For example, if a study used head input for pointing but used a physical button/switch to initiate a selection, or if an external input type was included in comparison to a target input, then the paper was deemed to provide value to the review.

  If the paper only examined external inputs across all conditions (i.e. a dedicated controller for pointing and selecting), then it was not included. Studies that were deemed to predominantly consider effects of output, as opposed to input, were also discounted.

- **Study type**: Studies were required to explore interaction for AR/VR applications. They also had to consider user accomplishments of application tasks or interactions, based on the defined input methods, or low-level tasks which assess human perception or cognition. However, this had to be strongly related to input approaches, implementing at least one form of explicit interaction.

  Papers that were found to consider interaction outside of XR technologies were classed as false positives. Studies that focused on novel hardware technologies were also excluded, as well as those primarily considering implicit interaction and output effects (i.e. to guide users to the correct interaction).

Table 1. Search Terms: Query applied to the IEEE and ACM databases, where each row of the table represents 'AND' and each comma between search terms represents 'OR'.

Topic	Search Terms	Location
Study Type/ Technology	elicit*, compar*, virtual, augmented, mixed, VR, AR, MR, immersive	Title
Display/Input	mobile, HMD, HWD, head mounted, head worn, tablet, smart phone, interact*, Input, technique*	Abstract
Interaction	method, intuitive, natural, modality, multimodal, ambigu*	Abstract
Modality	speech, voice, head, hand, gesture*	Abstract
Tasks	point*, select*, manipulat*, mov*, translat*, position*, rotat*, scal*, menu	Abstract
Use case	environment*, context*, scenario*, condition*, adapt*, hands free, eyes free	Abstract
General	participant*, subject*, user*, study	Full text

Table 2. Data categorisation approach: The factors assessed for the data analysis and their definitions.

Factor	Categorisation	Definition
Display Type	Headworn Display	Head Mounted/Headworn Displays (HMDs/HWDs)/smart-glasses
	Handheld Display	Smartphones/tablets
	Multiple Displays	A combination of Headworn with Handheld, or one of these displays alongside a static display (i.e. desktop monitors/ TV screens)
Input	Freehand	Using predefined gestures or unconstrained hand input with no wearable devices
	Speech-based	Using specific voice commands or natural language
	Head-based	Gaze interaction, orientations, rotations and head gestures
	Hardware-Based	Where a handheld display or external controller is employed; such as a touchscreen/touch-pad, button/switch, or 6-DoF manipulation of a handheld device
Type of Study	Elicitation	Where the users were asked to define their own interaction methods
	Assessment	Where users were asked to use a specific input/task combination and researchers assessed usability and feasibility for a given application/parameter
	Comparison	Where parameters (i.e. interaction methods or input/task combinations) were evaluated against a baseline or each other
Use case	Testing environment	-
	Lab	Constrained research setting
	Wild	Realistic use setting
	Scenario	-
	Static	Where interactions are conducted from a single position
	In motion	Where participants are free to move, or where interactions are performed whilst in motion
Tasks	Pointing	Searching for interactive elements i.e. via a cursor or ray casting
	Selection	Initiating/confirming an interaction
	Translation	Moving or relocating an interactive element
	Rotation	Changing the orientation of an interactive element
	Scaling	Reducing or enlarging the size of an interactive element
	Viewport control	Zooming and panning within an environment via a specific function (as opposed to implicitly moving around a scene)
	Menu-Based	Displaying a structured set of tabs, commands and/or utilities for the user to interact with
	Abstract	Non-spatial interactions such as editing (delete, undo, redo, insert, group; among others as in [64]), as well as interactions that could not be directly categorised as any other task

- **Participants**: Papers should clearly state the number of participants, the purpose of the study and its contribution.

- **Publication date**: Studies should have been published between 2013 and 2020. 2013 was defined as the cut-off date due to the impactful work presented by Piumsomboon et al. [64]. For their research, the surface taxonomy provided by Wobbrock et al. [92] was adapted to be better suited to AR gesture design. This resulted in the first user-defined taxonomy for intuitive hand interaction with holograms.

Of the 182 papers initially deemed to fulfil inclusion/exclusion criteria, 35 papers were selected for full review from ACM DL and 22 from IEEE Xplore. These publications were complemented by 11 papers from Springer, Elsevier, IFIP or Oxford Press.

This resulted in a corpus of 68 papers, which represents roughly a third of the eligible publications. More recent and relevant studies were prioritised, to provide an in-depth, state-of-the-art representation of current technologies and input capabilities.

2.2 Data Analysis

When conducting the review, there were five predominant areas of interest that embodied the factors considered. Table 2 provides the categorisations and definitions that were applied to analyse the sample of papers.

The first primary research area is *Display Type*, which is defined as the hardware employed for visualising virtual content. *Input* concerned the interaction methods observed as part of the user studies, which were used to interface with the display. *Type of Study* refers to the type of user evaluation conducted, with *Use Case* exploring the conditions that studies are conducted under. This involved reporting on the testing environment and users' scenario, particularly their pose (i.e. whether they were instructed to remain seated or if they were free to move), and highlighting to what extent interaction approaches were pre-defined and restricted for the research.

The final consideration was *Tasks*, which defined the interactions that the research reported on. The task categorisations were informed by the work of Piumsomboon et al. [64] and represent distinct functions, which are often combined to complete more complex activities in immersive environments.

These five factors are primary considerations for interaction and are commonly explored in reviews. For example, Hertel et al. [38] extract prevalent characteristics of interaction techniques based on input method and task and develop a taxonomy that sorts and groups them accordingly. Dey et al. [24] also discuss these factors in their review to identify primary application areas. They describe the methods and environments that are used for user studies, to propose guidelines and future research opportunities. Furthermore, the factors represent themes considered by LaViola et al. [46], where theoretical foundations, devices, techniques and design guidelines are explored in detail.

To clearly dissect information and highlight patterns and trends, data was extracted from each paper and coded within a matrix (based on the factors in table 2). There were three matrices, separated by display type (Headworn, Handheld and Multiple displays). The range of categories defined were not strictly binary, with papers being codified into more than one category where applicable (i.e. a significant number of papers examined more than one input method in comparison; or combination when multimodal approaches were explored).

3 Analysis

This section provides a summary of the data captured and highlights identified trends. Initially, a top-level analysis is conducted to encapsulate the data, reporting on the factors that were defined as key areas for exploration in section 2.2.

Following this, the data was analysed by display type. This was to provide a breakdown of the inputs employed, testing conditions implemented and tasks observed for different immersive platforms. The data is then further evaluated, regarding the current and projected state of XR interaction, in section 4.

3.1 Top-Level Review

This subsection summarises the data captured from the 68 papers included for review [1]. Of these papers, 54 were sourced from conferences and 14 from journals. The data discussed is presented for handheld, headworn and multiple displays in figures 1, 2 and 3, respectively.

3.1.1 Display Type

Roughly two-thirds of studies employed only headworn displays. There were an equal number of papers that implemented either solely handheld displays or multiple displays. Overall, 55 papers were found to target AR technologies and 13 were classified as VR. 3 papers reported to provide insight into both AR and VR.

3.1.2 Input

Most papers investigated either hardware-based input (22 of which considered interaction with external hardware controllers [10, 20]) or freehand gesture. Head was explored slightly less, closely followed by speech. A total of 36 papers were found to include multimodal input techniques.

3.1.3 Type of Study

All studies were identified as assessments, the majority of which also included a comparison. There were considerably fewer papers reporting on elicitation studies. Although it was not a focus of the review, information was also captured surrounding the factors that were assessed and/or compared.

As input is strongly related to how users respond to output, papers notably included visual parameters as variables (such as distance and scale) to test input approaches. Comparison studies generally analysed more than one input technique or display/interaction device, either under AR/VR conditions, or sometimes considering an immersive application against a standard, non-immersive baseline [14].

Relating to study type, an overview of participant sample and study protocol conditions is also provided, based on the parameters listed below:

- **Participant Sample:** The average number of participants was 22.48 (SD = 11.22), with the largest sample being 73 [12] and the smallest 12 [14, 36, 60, 66, 88, 97, 100].
- **Participant age:** 8 out of 68 studies did not report on average sample age. 7 papers provided vague demographics, either stating their participants were above 18 [99], or briefly referring to the ages of participants without explicitly stating their range [14]. For the remaining 53 studies, the average age was 27.72 (SD = 5.23).
- **Participant experience:** 55 studies reported on participants' relevant background experience (i.e. with the technologies, devices and interaction paradigms involved). 12 of these studies involved participants with previous basic or intermediate experience using relevant technologies, while 7 involved a sample with no previous experience. 36 studies included participants with different levels of experience, with 2 papers also reporting to include experts in their recruitment.
- **Study duration**: 37 papers reported on studies that were conducted during a single iteration, 35 of which stated average completion times per participant. These times ranged from 20 and 90 minutes for each user. 24 papers did not report on the duration of studies or testing sessions. 7 papers reported on longitudinal studies, capturing data on different occasions from the same participants, thus evaluating further learnability of the systems involved.

Another aspect addressed as part of study type was the kind of contribution. 50 papers were proposed to address or understand fundamental problems associated with explicit interaction in immersive environments. These studies included results on a more general scale and were not conducted to address practical issues. 18 papers were considered only relevant to a specific implementation, whereas 12 explored fundamental findings and went on to apply them to a final application.

Notable areas of contribution surround selection [12, 26, 30], object manipulation [20, 63, 90], text entry [48, 95], game interaction [14, 82], character control and animation [3, 96], human-human [86] and human-robot collaboration [31, 43], map exploration [76], UI (user interface) and menu-based interaction [8, 67], Medical/Healthcare [68, 74], interactive learning [9, 57] and AR assistants [51, 99]. Some studies could be classified into more than one of these categories, such as the work of Sadri et al. [74], which focuses on anatomic model manipulation for medical applications.

3.1.4 Use Case

The majority of studies were conducted under constrained, predetermined conditions in a laboratory environment. Only a small number of studies were delivered outside of the research lab (in the wild).

Even though the majority of studies used mobile technologies (untethered headworn and handheld devices), most papers reported on studies conducted from a single position in the testing space. Few studies focused on employing the freedom of movement offered by such devices.

3.1.5 Tasks

65 papers discussed a combination of tasks for their evaluations. Selection tasks were by far the most prevalent, followed by pointing and translation. Although reported slightly less than translation, transformation tasks were also broadly included (rotation slightly more than scale), as well as UI/menu-based interaction. Viewport control, such as zooming and panning, was explored considerably less.

Studies often assessed more complex interactions by adopting different combinations of explicit tasks. The majority of combinations included 3 tasks, which were noted by 23 papers, followed by 2 tasks (included in 18 papers). In 13 papers, 5 or more tasks were considered, and 4 tasks were featured in 11 papers.

Data was captured from participants based on a range of objective and subjective factors. 67 papers reported on quantitative metrics and 63 presented qualitative feedback. In 62 of the papers, both quantitative and qualitative measures were considered. This is likely the case as a mixed-methods approach is held as the most valid and reliable [77]. Only 5 papers include solely quantitative data and 1 paper qualitative data.

Data captured was namely error/accuracy and completion times (as objective metrics for assessments/comparisons). Subjective responses were usually collected via custom or industry-standard questionnaires (such as NASA-TLX [39], System Usability Scale (SUS) [16] and User Experience Questionnaire (UEQ) [78]). These were generally quantified for analysis alongside objective measures.

Many studies also captured more in-depth subjective feedback in the form of interviews, recorded observations and think-aloud protocols. Elicitation studies primarily quantified subjective agreement rates to define a consensus of user-defined gestures.

3.2 Handheld Display

Data captured for studies that considered solely handheld display devices, namely smartphones and tablets, is detailed in the following subsections. An overview of the data can be found in figure 1.

3.2.1 Study Type

All studies employing solely a handheld device addressed a specific parameter as a factor for assessment, to explore the influence of output or approach on interaction performance. This included how a pointer or cursor is indicated or behaves [60, 97], where the user performs the gesture (front or back of the device) [42], the impact of task on interaction [5, 42, 55, 75, 80, 96, 97], or the size/distance of an interactive element [51, 70, 97]. 8 of the papers also explored the benefits of a novel technique or interface.

[1] List of the 68 papers included in the review (Last accessed 8th September 2021) - https://1drv.ms/w/s!Ago1DH6X9D1OyXRsGiPbefkdsPyd?e=Fs2Yhj

All comparison studies used the input method as a variable, however, Tanikawa et al. [81] also considered the effect of display devices, by comparing a smartphone with a tablet. Furthermore, Kim and Lee [42] explored to what extent a wide-angle lens improved usability (enhancing FOV). Although tasks were performed under AR conditions using handheld for all of the comparisons, in some instances [5, 51, 59, 60, 81], touchscreen input was also used as a baseline to observe the effectiveness of other inputs (such as freehand gesture or multimodal approaches).

The elicitation study employed motion gestures in 6-DoF, where participants were asked to define motions to control an augmented character (first by manipulating a human-like doll and then a mobile device [96]). The gestures were later implemented within a novel interface for assessment, using the smartphone display.

6 papers offered fundamental contributions, whereas 5 papers considered their contribution on a general scale, as well as applying it to a specific application. There were 2 instances where the research was exclusively application based [31, 96].

As highlighted in figure 1, studies were predominantly conducted in a controlled environment, under laboratory conditions. However, Mayer et al. [51] employed a less controlled, outdoor environment for part of their experiment. Participants were also asked to remain static for the majority of studies. Where free motion was permitted during testing, 3 studies examined device motion and trajectories. No papers were found to report on human-motion data.

3.2.2 Input Methods

In line with the review by Goh et al. [35], the majority of studies employed hardware-based input via the handheld device itself. The touchscreen display was used in all studies for at least one condition (i.e. for interaction with GUI elements and for intuitive object manipulation via touchscreen legacy gestures [31]).

6 papers also considered manipulation of handheld displays for explicit interactions, with almost half of the studies implementing freehand interaction. As illustrated in figure 1, speech and head-based inputs were explored least.

Some studies reported on novel interaction approaches that discussed at least two types of input. For example, touch and hand were compared [5, 42, 59] and combined [42, 59, 80] in several papers. Hand, touch and device manipulation were also evaluated in the work of Su et al. [80]. Furthermore, Qian and Tether [70] compared hand gesture with dwell-based selection via device manipulation, whilst Mayer et al. [51] considered head-based interaction with speech and implicit hardware-based input. A single study also noted the impacts of different combinations of multimodal interaction (touch, hand, speech), with alternative output conditions [59]. In total, 8 papers investigated multimodal methods, however, 6 employed solely hardware-based input (a mixture of touchscreen interaction and device movement).

3.2.3 Tasks

As presented in figure 1, the task most often observed with handheld displays was selection, closely followed by translation. Approximately half of the studies explored rotation and pointing tasks. Abstract and scaling tasks were considered by close to a third of studies, whilst menu and viewport manipulation via a specific function (manipulating displayed content based on users' POV [75]), were examined least.

In terms of the input methods used to complete the different tasks, 10 papers implemented touch interaction for selection. Physical movement of the device with 6-DoF was generally employed for explicit pointing and manipulation tasks, using some kind of visual indicator (i.e. a rod, cursor or raycast [75, 97]), however, Tanikawa et al. [81] only considered movements with up to 3-DoF. Gestures with the physical device were also compared with standard touch gestures for object manipulation, through techniques such as multi-touch interaction [42].

Object manipulation tasks were achieved by combining touch with physical device movement in 5 papers (where touch triggered the interaction and movement defined the translation/rotation/scaling axis and behaviour). Mayer et al. [51] went beyond hand and hardware-based interaction by implementing speech for abstract commands and head gaze for pointing. As well as this, Nazri and Rambli [59] assessed how users freely employ different forms and combinations of input (speech and hand) alongside standard touch interaction, to complete a gamified task.

The tasks were delivered differently depending on the study design. Assessments predominantly investigated predefined tasks and interaction methods, which were most often taught to participants through a training stage. Comparisons primarily observed the impact of different interaction methods on task execution, whereas the elicitation study explored user gestures based on a defined list of actions, to understand user approaches to different types of tasks.

3.3 Headworn Display

The following subsections elaborate on the data captured for studies considering headworn display devices in standalone. An overview of the data for headworn displays is provided in figure 2.

3.3.1 Study Type

Of the papers represented in figure 2, 22 assessed how interaction is affected by different tasks and 21 papers measured the impacts of output. Changes in output were notably related to the size or distance of virtual content, which was explored in 13 of the publications. 19 assessments also concerned novel applications or techniques. 6 papers reported on the number of fingers/hands employed for mid-air interactions.

Few papers recognised factors surrounding environmental conditions. Only 2 papers were found to report on the influence of lighting when interacting indoors and outdoors [15, 49], one of which also discussed the impact of ambient noise levels [49] when employing speech input. 3 papers were found to report on longitudinal studies, to assess learning curves [48, 67].

Where factors were also compared, 31 studies discussed different interaction methods or techniques. 2 of these studies examined the device type, where different interaction form factors were explored [32, 82]. Alallah et al. [1] also compared the affects of input and from which point of view (performer vs observer).

Elicitations were again considered least. These studies were related to small target selection [12], multimodal interaction (speech and gesture) [90, 91] and gesture interaction [62, 64], for manipulation tasks, animation in VR [3], or more general input selection; when employing smart glasses for game interactions in public spaces [82].

The majority of studies were conducted under controlled laboratory conditions, with few papers reporting on results gathered in more realistic environments. 2 papers addressed both controlled and uncontrolled conditions. Studies conducted 'in the wild' were primarily related to specific use cases (i.e. a cultural heritage site [15], care home [68] or in an industrial environment [69, 86]), with Alallah et al. [1] exploring fundamental interaction in public spaces. Participants were again asked to remain static for most studies. 4 studies were found to consider both static and mobile conditions.

3.3.2 Input Methods

As shown in figure 2, the input method explored most with headworn displays was freehand interaction. This was followed by head-based input, which was included in more than half of the papers. Hardware-based and speech interaction were considered least, but still occurred relatively frequently.

Multiple input types were explored in most of the studies, with only 9 papers reporting on a single input modality. The publications were mostly observing 2 input methods (in 17 papers), or 3 input methods (in 9 papers). These input methods represented different permutations of hand, head-based, speech and hardware-based inputs. 23 papers applied at least one combination of multimodal input (i.e. to decouple inputs to complete distinct tasks [57] or to couple inputs to improve the accuracy of interactions [37]), whereas 8 papers used multiple inputs solely in comparison as individual techniques.

The most frequent multimodal input combination was head with a hardware controller, which was included in 9 papers. This was followed by hand with speech and hand with head, both of which were used in 8 papers. Head input with speech was also explored in 4 papers. Some

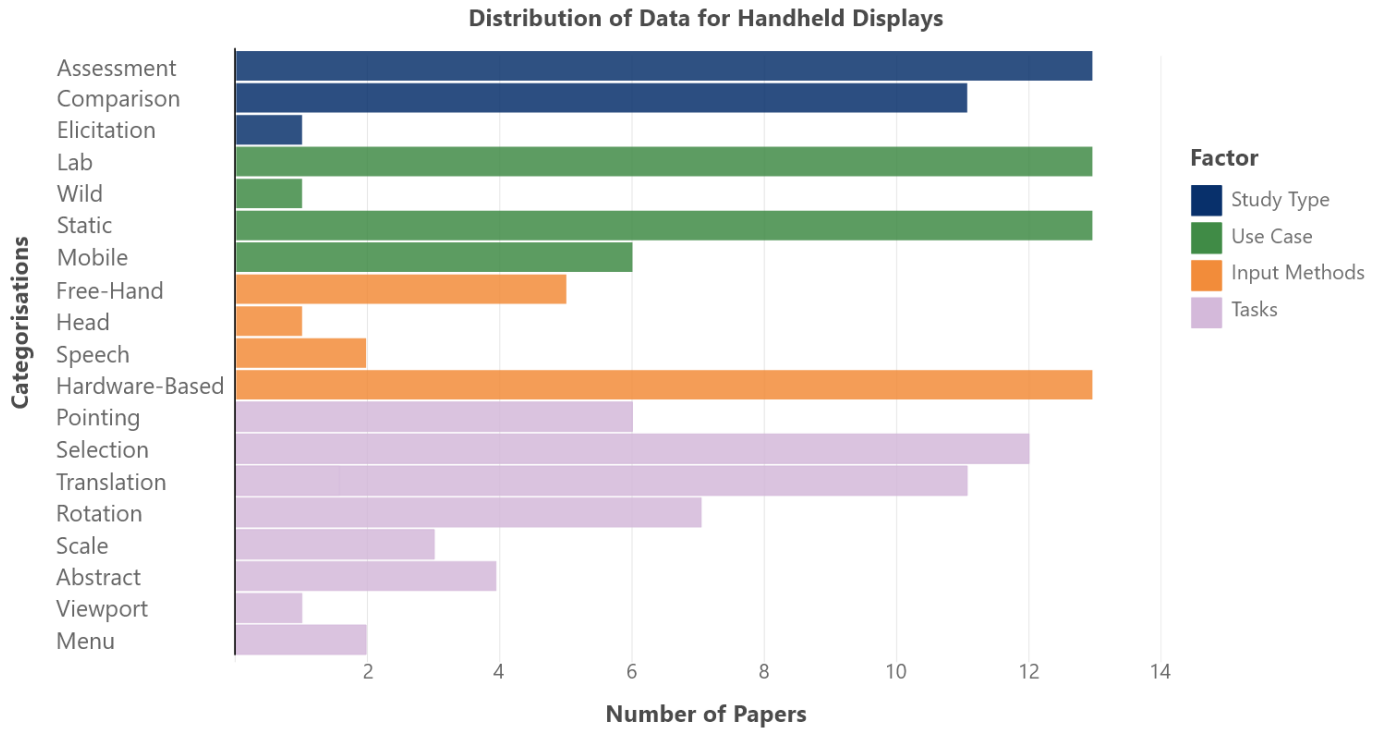


Fig. 1. Distribution of data for the 13 papers considering solely handheld displays (focusing on study type, use case, input methods and tasks).

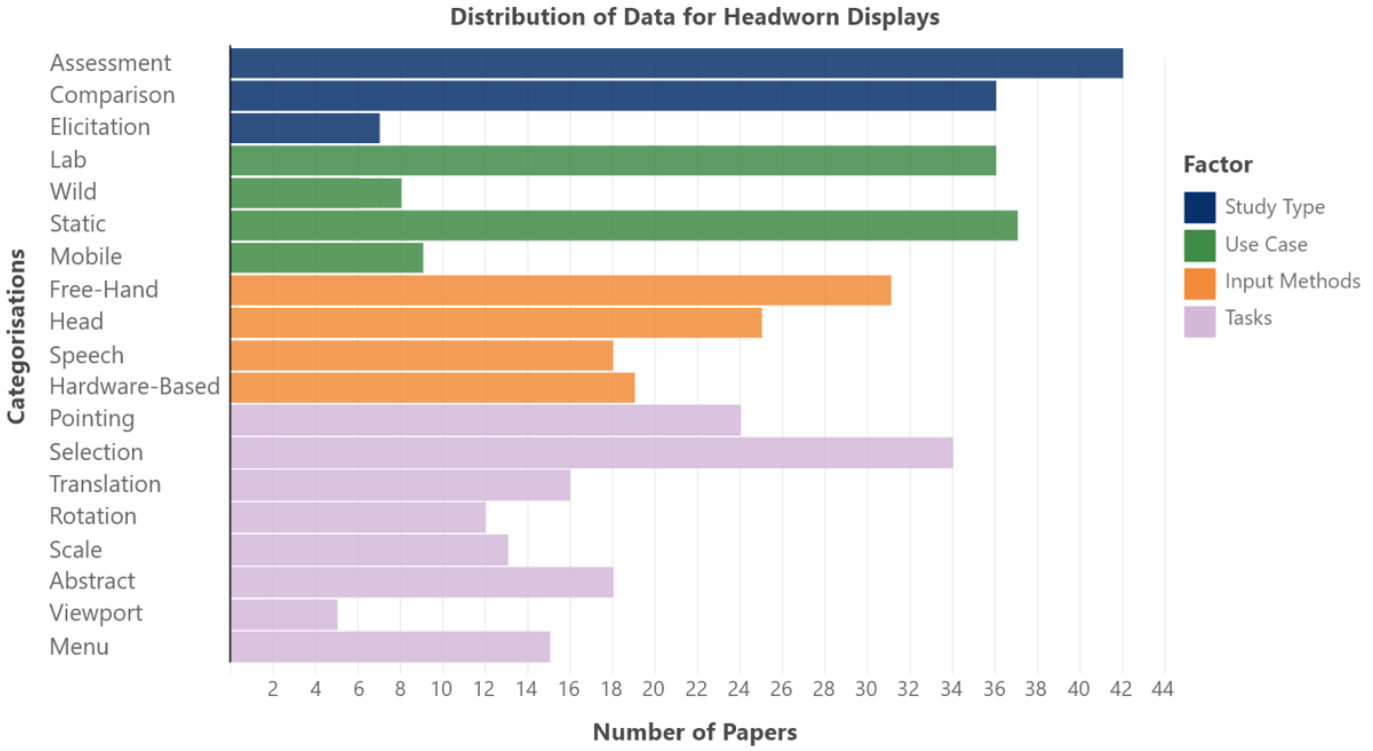


Fig. 2. Distribution of data for the 42 papers considering solely headworn displays (focusing on study type, use case, input methods and tasks).

studies considered multiple combinations of hand, head, speech and hardware-based inputs. For example, Tung et al. [82] investigated how users naturally choose to apply these inputs in public spaces.

Furthermore, 10 papers concerning head or speech input discussed how systems could adapt for hands-free interaction approaches. This predominantly included applications for healthcare or maintenance [8, 45, 68, 74, 86], where users are generally required to operate their hands to complete real-world tasks, and for text entry, where it may be inconvenient to use an external controller, or look down to type on a smartphone [48, 95].

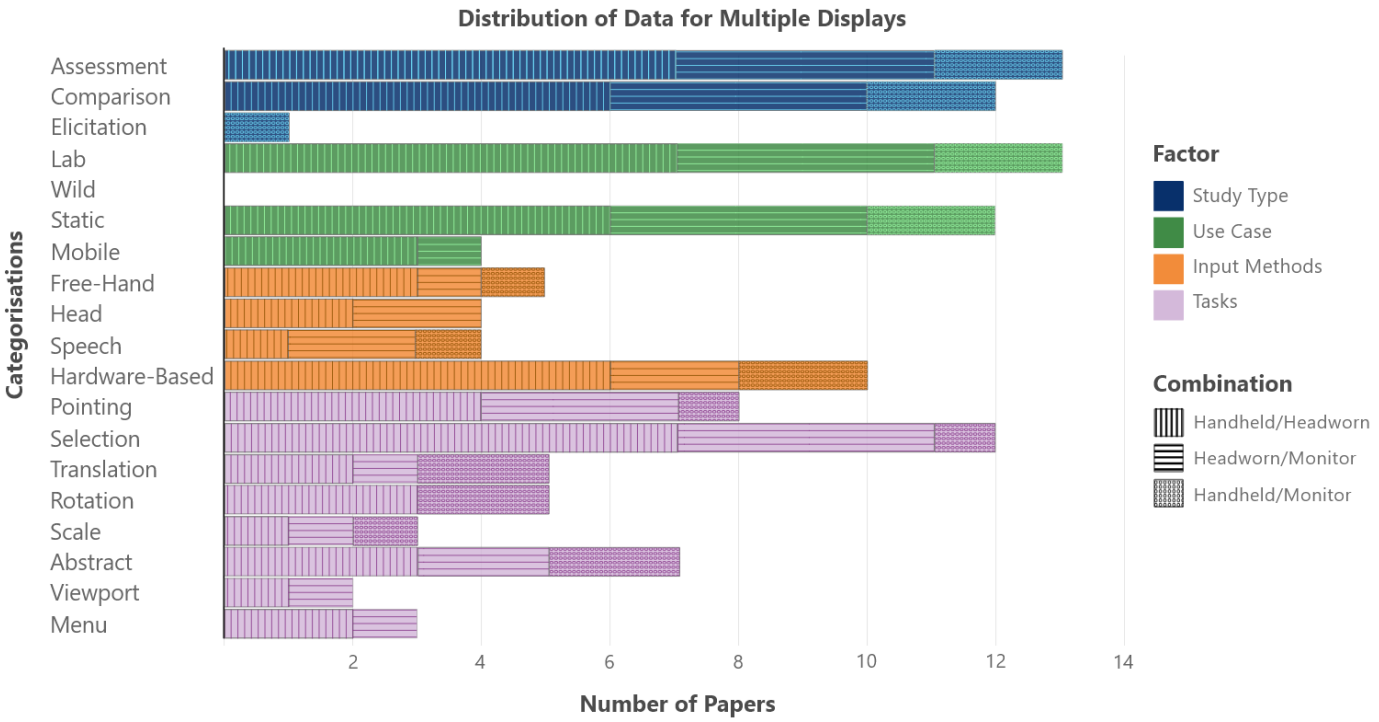


Fig. 3. Distribution of data for the 13 papers considering multiple displays, broken down by handheld/headworn (7 papers), headworn/monitor (4 papers) and handheld/monitor (2 papers) - focusing on study type, use case, input methods and tasks.

3.3.3 Tasks

As depicted in figure 2, selection tasks were again by far the most widely reported. However, for headworn displays, this was followed by pointing, which was explored by more than half of the papers. Abstract tasks, translation, menu-based interactions, scaling and rotation were addressed by a similar number of papers, whereas viewport control was explored considerably less.

15 studies investigated a combination of 3 different tasks and 10 explored 2 tasks. 4 studies were found to employ 4 tasks, with those considering more than this predominantly being elicitation studies. Only 2 studies reported on a single task, evaluating more abstract, indirect interactions; either assessing multimodal input, namely voice and hand gestures for interacting with a virtual character [21], or evaluating speech and conversational interfaces for indoor wayfinding and navigation in AR [99].

In terms of multimodal input methods, where head gaze is combined with a controller, it was generally to separate functions for pointing and selection mechanisms, i.e. using head to identify an object of interest, through a form of visual output (such as a raycast), and an external binary form of input (such as hand-gesture or controller [101]) to confirm the interaction. As previously highlighted, speech was generally employed to accompany freehand or head-based interactions. Papers often explored the affects of unimodal and multimodal combinations on task performance and usability.

Freehand gesture was considered for selection in 23 papers and for direct controls for canonical tasks, such as translation, rotation and/or scaling, in 16 papers. In 2 papers, freehand interaction was also used for indirect gesture controls to provide instructions to virtual avatars [21, 82].

Speech was predominantly used for abstract tasks (applied in 10 papers to trigger discrete interactions). Head input was employed in 7 papers for menu-based interactions, with 5 papers applying dwell for selections in at least one condition. Head input was also used for abstract interactions in 5 papers and object manipulation in 4 papers. In some cases, head gestures such as nods, shakes and tilts were utilised to manipulate an interface or virtual object [61, 68].

Where viewport control via a specific function was employed, 2 papers were related to VR interaction [12, 98] and 3 to AR interaction. Applications of viewport control for AR covered map exploration [76] and game input [82]. It was also employed for an elicitation study, where users chose to manipulate the scene to interact with distant objects (i.e. metaphorically zooming/pulling objects towards them) as opposed to physically approaching interactive content [62].

3.4 Multiple Display Types

Finally, we highlight the data captured from studies that considered multiple display devices (which is presented in figure 3). These are classified as handheld/headworn, handheld/monitor and headworn/monitor.

3.4.1 Study Type

Where multiple display types were included in studies, they were explored in combination in 8 instances and comparison in 5.

All headworn/monitor studies were assessments and comparisons. Qian and Teather [71] discussed the impact of target distance and input methods on performance, and Rao et al. [72] compared interaction behaviours, with either speech or vision-based context anchoring. Both studies employed a desktop set-up in combination with smart glasses, to complete the assigned task.

Wang et al. [88] also combined both types of display, to compare unimodal and multimodal interaction. This was the only study to explore the impacts of using 3 inputs simultaneously (eye gaze, gesture and speech). The work of Bothen et al. [14] compared a standard desktop-based gaming experience to a VR game, which employed head gaze for interaction.

One of the papers considering monitor/handheld reported on an elicitation study, where a TV monitor was used to display referents and the handheld display to design interactions [25]. The other paper compared an immersive desk, monitor-based set-up, to interaction on a tablet or smartphone [9], for interacting with an educational AR magazine. Both studies assessed how the task affects interaction methods and subjective preferences.

The multi-platform combination that was most frequently employed was handheld/headworn. 3 studies combined devices in tandem for a seamless multi-platform experience [36, 87, 100]. The remaining papers

compared headworn and handheld interactions [50]. Again, assessments primarily concerned the influence of task, which was addressed in 5 papers, and output, which was explored in 3 papers. However, 1 study investigated in-pocket text input by the thigh [36] and another considered how walking different path types affected interaction in locomotion [34].

3.4.2 Input Methods

The general trends for input devices used with combined displays are shown in Figure 3. Where a monitor was used with a headworn device, 1 study combined speech (which was captured by a headworn display) and hardware-based input (via a standard desktop setup) [72]. Another considered head input, in comparison and conjunction with eye gaze [71].

Wang et al. [88] explored how the number of inputs affects performance metrics, comparing different combinations of eye gaze, hand gesture and speech. Finally, Bothen et al. [14] examined head gaze and standard interaction with an Xbox controller, observing how the task and gaming experience of participants impacted objective and subjective results.

For handheld/monitor, Dong et al. [25] employed both touch-based surface and hardware-based (6-DoF) gestures, whereas Bazzaza et al. [9] considered multimodal interaction, as a combination of hand, speech and hardware-based input.

Where handheld/headworn was explored, hardware-based interaction was used most frequently. This was followed by hand, head and speech, respectively. 2 papers explored multimodal interaction [87,100], both of which implemented hand and hardware-based input. Waldow et al. [87] also investigated using head-based input alongside gesture for constrained object manipulation.

3.4.3 Tasks

As shown in figure 3, all headworn/monitor conditions considered selection, most of which also involved pointing. The studies concerning speech input both included abstract interactions, with Wang et al. [88] reporting the only instance of a scaling task in this category. As well as pointing and selection, Bothen et al. [14] explored translation, menu-based and viewport tasks, representing interactions such as aiming, shooting and walking within a VR game. None of the headworn/monitor studies were found to address rotation tasks.

For handheld/monitor, both studies considered translation, rotation and abstract interactions, however, Dong et al. [25] also examined scaling, and Bazzaza et al. [9] pointing and selection.

In handheld/headworn conditions, selection was again reported in all papers. This was followed by pointing in over half of the papers. Rotation and abstract tasks were explored slightly less, ahead of translation and menu-based interaction. Viewport and scale tasks were considered least.

The 2 papers that investigated multimodal interaction, with handheld/headworn devices in combination, incorporated the most number of tasks. Zhu et al. [100] employed all tasks except scale, and Waldow et al. [87] considered 4 tasks (pointing, selection, rotation and scale).

In the only instance that head-only input was recorded [27], pointing, selection and menu were considered. Similarly, the only paper involving speech input concerned abstract tasks (for text editing), which is in line with most speech-based, headworn conditions.

4 Discussion

To consider how interactions are currently assessed and employed for a range of immersive applications, 68 papers that included at least one usability study were collated and reviewed. A common set of attributes were defined as outlined in section 2.2 and data was extracted from each paper. The results, key trends and findings are discussed in the following sections.

4.1 Display Type

As highlighted in table 2, display type was classified into 3 categories; Headworn, Handheld and Multiple displays. The following subsections provide an overview of the types of technologies that were used to conduct user studies. The findings are also evaluated in more detail, to understand how interaction relates to the type of display employed and to highlight future implications.

4.1.1 General Findings

The sample of papers included for review predominantly employed headworn displays. This was primarily smart glasses (such as those manufactured by Epson, Daqri or Google) and industry-standard HWDs (such as Magic Leap and Hololens v1). VR headsets were also considered, namely Oculus technologies. As well as this, some studies employed custom made or adapted headsets, i.e. implementing a Leap Motion IMU onto a glasses frame [49].

Handheld devices were generally standard consumer-level platforms, such as Apple and Android smartphones or tablets. However, as seen with headworn displays, there was one instance where hardware was added to a handheld tablet display, by attaching a Leap motion IMU [42]. A single study also considered interaction on a Microsoft tablet-PC [75]. Many studies utilised tools such as Google's speech API and development frameworks such as ARKit, to develop the applications being assessed.

In terms of studies that used multiple displays, those including a monitor were all restricted to a defined interaction zone. Even so, they provide insights into how we may use untethered immersive technologies alongside ubiquitous displays such as laptops, desktop monitors and TV screens.

Few studies included in the review assessed how combining handheld and headworn displays affects interaction and usability. However, where studies did use both devices, they seemed to gain positive results. For example, Zhu et al. [100] propose that usability is improved when employing a familiar device (smartphone), alongside a less familiar form of interaction (HWD).

Despite the potential of portable technologies, there are issues surrounding interaction for these devices; notably concerning ergonomics and technological constraints such as tracking and recognition. Field of View was also revealed to be a major factor affecting usability in both headworn and handheld conditions, as well as depth perception and occlusion.

In an attempt to mitigate technological constraints, Kim et al. [42] incorporated a wide-angle lens on a handheld tablet, which was shown to improve the quality of freehand input techniques and provide more useful and natural interactions. As well as this, they used a leap motion to extend freehand interaction capabilities. Adjustments such as offsets to visual output were also found to make freehand input techniques more appropriate for handheld displays [42, 97].

4.1.2 Evaluation

Studies often implemented a range of hardware to remove technical constraints of current systems. However, as the testing set-up and apparatus employed for studies will indirectly affect the results that are reported, and provide an unrealistic outlook on the technology, it is not desirable for users to require additional equipment, such as tracking arrays and fiducial and colour markers, to interact. This is because, in most realistic future use cases, this equipment would be removed [83]. Some studies mitigated technical issues and constraints by employing a wizard of oz study [90, 91], or applied semantics to prompt speech interaction, however, this introduced other issues such as latency [45].

Despite the limitations of such approaches, it is necessary to consider how unrestricted and unconstrained input in realistic use cases influences interaction approaches, by focusing on factors external to those affected by the technology [85]. This is due to the quality of hardware and software being in constant flux.

The participants employed to conduct user studies are also key to uncovering the most appropriate inputs with different display types. For example, Munsinger et al. [57] highlight the importance of considering different audiences, as even if technologies employed for user testing work well for one group of users (i.e. as shown with average adults with the Microsoft HoloLens), this level of performance will not necessarily translate to all users, such as children. Therefore, it is important to consider how display devices used for immersive technologies can

be developed for different groups of users, so they are ubiquitously accessible.

Finally, although headworn and handheld displays are effective when used in standalone, they can also complement each other to utilise the benefits of both input provisions. For example, as hardware-based input via a controller has been found efficient for selection with headworn displays [30,43], a mobile phone could become a universal control method to use alongside different types of headworn display, i.e. as a straightforward way to separate pointing (head) and selection (hardware-based touch input) mechanisms.

As well as this, applying the headworn device as the display and handheld device as the controller removes the need to apply adaptations to output (which improves rotation tasks when manipulating objects on solely handheld devices), as the visual perspective is no longer bound by the orientation of the handheld display. As the handheld display is not required to be within FOV of the headworn display, this input technique could also reduce fatigue when compared to freehand input.

4.2 Input Methods

The input methods explored were categorised as freehand, hardware-based, speech-based and head-based. These were observed across both handheld and headworn display types, with the research being focused on input techniques that were achievable using the display devices themselves. As well as considering these input methods in standalone, we also evaluate how they could be combined as multimodal interaction techniques.

The following sections provide an overview of how different input methods were employed. Future research directions are also introduced, based on the advantages and disadvantages of these input techniques for the tasks defined in table 2.

4.2.1 General Findings

Although freehand gesture input was the most considered form of interaction, this is representative of HWDs, where it is not necessary to hold a device. Hand gesture was found to be less appropriate for interaction with handheld displays. However, freehand input could be useful in some instances; especially stationary applications when manipulating virtual objects or models [42].

Freehand gesture techniques are generally more suitable when employed for fun and at the users' leisure, specifically when time completion is not a factor [30, 87], as the input method is more intuitive [82] and maximises enjoyment [87]. This was especially found to be the case when not employed for extended periods to induce fatigue [11, 26, 82], and if not restricted by technical constraints surrounding tracking [21, 49].

Much of the research surrounding freehand input justifies the chosen interaction paradigms based on previous highly cited elicitation studies, primarily the work presented by Piumsomboon et al. [64]. As elicitation studies are based on instinctive, user-defined approaches, freehand input design is often based on legacy gestures (as detailed in section 4.3), where users employ inputs that simulate interactions with existing technologies (desktop/touchscreen). For example, air taps that mimic mouse clicks for selection were generally proposed [64, 90], and metaphoric scaling following the laws of touch screen pinch gestures were found to be preferred over isomorphic gesture paradigms, which simulates how we may stretch/scale objects for real-world interaction [32].

Hardware-based gestures with handheld display devices (i.e. manually manipulating the device with 6-DoF) were again generally more beneficial for applications that do not require high precision. This notably includes tasks surrounding character control or game interaction, where gestures could be performed indirectly, such as to make an avatar jump [96] or for throwing tasks [25]. When used in conjunction with a HWD, touchscreen interaction was also found to be more suitable for object transformation tasks than mid-air hand gestures, as well as having better overall ease of use [87].

For handheld devices, the most beneficial form of input for translation tasks was found to be a combination of the built-in components of the device (touch screen and physical manipulation). Interaction was generally found to improve when manipulation functions such as rotation and translation were separated by DoF [80]. Rotation tasks were more difficult to achieve through manipulation of the handheld display, due to the range of comfortable movement and issues with perception. Despite this, rotation could be more viable when used in conjunction with a headworn display [87], or with technical adjustments such as rendering the display to match users' perspective [75].

Touch-based interaction (as employed for interacting with touchscreen displays) is employed as standard for ubiquitous technologies like tablets and smartphones. However, where unimodal touchscreen interaction was used for direct selection and manipulation under immersive conditions, it was often found to be the most error-prone and least preferred form of interaction [59, 81]. The exception to this was when touchscreen input was compared to device motion gestures with 6-DoF for straightforward interactions (that can be employed as single/double taps) [25].

Although interactions with external handheld controllers were generally more efficient and socially accepted with headworn devices [30,95], users most often preferred the concept of interaction techniques that did not involve additional hardware-based input devices [34]. Inputs that did not require additional hardware were also sometimes found to be more intuitive and usable than standard interactions (such as game controllers) for new users [14]. This is because interactions such as character control can be initially difficult to achieve with input methods like analogue sticks, which often require accurately balancing movements in the X, Y and Z dimensions.

Speech is an ideal form of input, as natural language can be employed to easily define and represent concepts in the real and virtual world [51, 93]. However, speech interaction is limited by the quality of formal logic and recognition [45], and users have concerns surrounding privacy and social acceptance [34, 67, 82].

Speech is arguably the most error prone input method, comparison studies identifying speech to be the least robust form of interaction, due to systems struggling to adapt to the range of inconsistencies presented by natural language (which includes accents, dialects and ambiguities [40]). Despite this, research suggests that machine learning is advancing [18, 23] and in some instances, speech was the most robust input method [21, 49].

Speech has also been found to be less natural for applications with a single user [67]. However, as speech is inherently employed for human-to-human communication, voice-based input is arguably more appropriate for collaborative environments, such as for remote assistance applications, as verbal communication can be applied more intuitively [86]. Employing speech interaction has also been found to increase memory retention and learning for educational applications [18].

Where speech commands are abstract but relate to visible objects in the environment, head or hand input can also be employed alongside speech to improve system understanding, by correcting any ambiguities presented by natural language i.e. by providing context for "that" when indicating an object of interest [51, 93]. Although speech interfaces can benefit from natural language understanding, Zhao et al. [99] revealed that natural language algorithms are primarily beneficial for new users, and are less useful once a user is accustomed to using an application.

As speech is naturally employed to communicate concepts, it is also more difficult to apply for spatial interactions such as object manipulation, as it is difficult to precisely communicate intentions [91]. Instead, speech is especially beneficial for more abstract interactions, such as "delete" and "create" tasks, as it is more difficult to define gestures for non-direct, conceptual interactions [63, 90].

Head-based input (notably based on gaze and orientation information), similar to speech, was found to be useful for short, discrete tasks. Therefore, head input could be beneficial for abstract interactions such as switching or menu-based controls [20]. Head was also often used for pointing, to define an area or object of interest [51, 93], or as a cursor to select interactive elements (as employed for typing interfaces). This was generally achieved through dwell [22], or in combination with an external selection mechanism to confirm interactions, i.e. via a controller [8] or finger tap gesture [95].

As well as this, Yu et al. [98] exemplifies how head input can be used

to navigate content in the depth dimension, which could be beneficial for users with impairments, or where users are required to employ their hands for external tasks [48, 74], as well as when interacting in public contexts [30].

Although head-based interaction has been found accurate for both handheld and headworn conditions, especially where interactive content is large and in close proximity [51], head input was still found to be less accurate and natural with handheld than for headworn. This was especially the case as distance increased [51]. Therefore, like freehand interaction, head input is predominantly more appropriate for headworn displays. This is likely due to ergonomic factors, as users are required to hold the device in a less natural position under handheld interaction.

Instead, head gaze information is generally more suitable when used alongside another form of input, such as speech interaction, to correct borderline ambiguities and provide a system with context [93]. Although deictic hand gestures can also be used to indicate an object or area of interest [51], this form of input is less discreet and less ideal for repetitive interactions; due to fatigue [19].

Even though multimodal interactions were considered regularly, the review suggests that usability studies tended to combine just two modalities simultaneously. This was notably gesture and speech [21,91], head and controller [34] or touchscreen alongside physical movement of the handheld device [80]. Despite this, there is research to suggest that additional modalities could provide enhanced usability when applied to distinct tasks, through methods such as physical decoupling [53]. For example, head could be employed for pointing to indicate selections and interact with menus, and gesture alongside speech for object manipulation [88]. However, even though multimodal input can enhance interaction capabilities, as users tend to employ simultaneous multimodal input sparsely [91], both unimodal and multimodal interaction capabilities should always be permitted [83].

4.2.2 Evaluation

Results suggest that inputs can be mapped for enhanced interaction across both headworn and handheld devices, based on their effectiveness for fulfilling different tasks in immersive environments. Consequently, we suggest that research could focus on exploring combinations of inputs, based on the tasks they are most suited to, as opposed to a single input method to complete more complex interactions. This will help to understand to what extent interaction approaches can be balanced between input types, as well as to what degree they are appropriate and accepted by users, in different use cases and scenarios.

As a whole, for headworn displays, head was found to be beneficial for pointing tasks [27], hand for object manipulation [74], and speech for abstract tasks and commands [91]. For handheld devices, a plausible mapping for head pointing interactions on headworn displays is raycasting [97], or rod techniques [81]. Again, hand interaction could be used for more intuitive and enjoyable interactions with handheld displays [65,87]. 6-DoF gestures were also found to be beneficial when used in conjunction with headworn devices, such as for applications in gaming [25] or for object manipulation [100], which could prove to be more usable and precise than touchscreen gestures for interaction [25].

Although different input methods have been found most suited to certain tasks, in the past, studies surrounding immersive technologies have most frequently considered unimodal interaction techniques. These methods permit the user to manipulate content via a single input, for example, through solely gesture, speech, or a hardware controller [56, 83]. This means that the majority of applications restrict users, and are not fully reaping the benefits of immersive systems, as the combination of more than one modality can improve system understanding (i.e. to resolve issues surrounding unimodal input techniques) and enhance user experience [83].

Multimodal interaction capabilities are therefore beneficial, as they can provide the user with an adaptive interface, which makes interaction more intuitive and straightforward to employ. Multiple inputs can account for issues such as situational impairments, environmental conditions, and issues surrounding spatial awareness and 'fat finger' with freehand interaction (in mid-air and on touchscreen devices). Multimodal input can also aid with selection and manipulation tasks, such as translation or rotation, and help to correct speech ambiguities, when delivering commands via natural language [35, 40, 56].

The high number of multimodal input approaches that appear within this review (explored in 36 papers) confirms that there is an increasing amount of research considering how multiple inputs can be combined, to improve interaction and usability. However, currently, there is a lack of grounding to define the most appropriate input methods for the distinct tasks employed for immersive environments, when interacting with different devices and in various use-cases.

Multimodal communication capabilities provide opportunities to convey maximised transferability and interaction suitability, across immersive interfaces and devices. Interaction would benefit from the complementary nature of more than one input modality [47], which would also introduce a means to correlate proxies for natural interaction (when applied to different display types, tasks and use cases [2, 51]).

For example, although hand gesture is arguably the most intuitive form of input, mid-air hand interaction is unsuitable when the user is required to interact for prolonged periods of time, due to fatigue (which relates to "Gorilla Arm" [19]). Instead, it was found that hand gesture would be better implemented for specific tasks and interactions, namely object relocation [63], and used alongside additional modes of input, such as speech, to make interactions like scaling less cumbersome [63, 90]. This will increase enjoyment and engagement, and provide more usage scenarios [21].

Although multimodal input is highly beneficial for XR applications, it must be ensured that input methods are carefully designed. Systems should also apply unimodal input where appropriate, to limit physical and mental workload [91]. Understanding how inputs could be mapped for different use cases, alongside different output modalities, will be important for the future of immersive technologies, as applications become more widespread [24].

As well as the inputs that are used, the paradigms and mappings employed to communicate the different inputs are important for interaction design. Although studies that apply legacy gestures arguably provide more intuitive gesture designs, such approaches are likely to limit the potential of XR technologies. This is because legacy gestures are defined based on user instinct, which is strongly informed by their past experiences with ubiquitous devices.

Consequently, researchers should consider how to limit the affects of legacy bias, to avoid simply replicating standard interaction with 2D displays. This will ensure interaction approaches are fully reaping the benefits of input capabilities provided by immersive technologies.

Finally, the nature of the immersive environment (i.e. AR/MR or VR) and the capabilities of the technologies employed will influence the appropriateness of different input methods. XR interaction techniques are notably affected by technological embodiment (to what extent the technology becomes an extension of the human body), perceptual presence (psychological perception which ranges from feeling part of the real-world location to feeling transported elsewhere) and behavioural interactivity (the capacity to directly and/or indirectly modify and control the system, by responding to feedback in real time) [29].

Whereas AR allows the user to dictate the real environment as well as virtual content, VR applications completely substitute the real-world surroundings and generally aim to provide the user with the sense of being transported elsewhere. To effectively interact in VR, users are required to interpret the state of the virtual environment and respond accordingly. In VR, the real environment and user's body is hidden or virtually represented. This means inputs are required to be accurately mapped and clearly indicated, to maximise the level of embodiment/presence and provide effective interactivity.

Furthermore, as the user is not aware of the real-world surroundings in VR, input techniques are more limited by the size and nature of the interaction space than in AR. In AR, the user can arguably interact and navigate the environment more confidently, as they can appropriately adapt inputs to the real interaction space. For example, the user can more easily adjust their pose/input method, or pause their interaction, if an obstacle becomes apparent.

However, as AR/MR merges digital content with the real world, further issues are introduced. This includes layer interference and

problems with light/colour blending, which affects immersive content in terms of visibility, depth ordering, object segmentation and scene distortion. Surrounding people and objects also introduce noise, which can hinder the accomplishment of different tasks. Issues such as limited FOV, world tracking and context matching in AR (based on the real environment) can also impede interaction for users and make it difficult to effectively adapt and respond to content in real-time [95].

Additionally, AR interaction could be affected by social acceptance more so than VR. In VR, the user has a lower awareness of bystanders, meaning users could feel less conscious of observers in the interaction space. The tolerance to external devices, such as hardware controllers, may also differ. For example, hardware controllers can be represented more easily by virtual objects (i.e. a tool in VR), to match the context of the application and maximise embodiment. VR applications are also primarily restricted to a predefined interaction zone, whereas AR is more likely to be employed for sporadic interactions (i.e. when on the go), meaning external hardware would presumably be more cumbersome to use.

Although the impact of the type of XR technology on interaction is considered in this paper, it is not explored in depth. We intend to revisit this review in the future to learn more about the relationship between AR/VR input techniques and approaches, as well as the distinctions that may influence users' interaction preferences.

4.3 Type of Study

A primary consideration when conducting the review was the study type (Assessment/Comparison/Elicitation). This element relates to the factors and variables that were considered and introduced to observe user interaction approaches, which are explored further in the following sections.

4.3.1 General Findings

When considering the studies that were solely assessments, they concerned an application-specific development. This is where researchers were interested in refining a novel interaction approach [15], or validating an application [28, 69, 100].

As highlighted in section 3.1, studies also generally focused on measuring performance and general usability. Although a mixed-methods approach offers a more in-depth analysis, and it is promising to see the number of studies now adopting such approaches, many studies were measuring the same factors. Even though this increases comparability, few studies considered more abstract measures such as social acceptance and learnability. Few papers also reported on long-term studies (the longest being 14 days [67]) or environmental factors such as noise or lighting conditions [49].

Comparison studies often considered the influence of input modality, with hand being used both in comparison and conjunction with speech [90, 91], or touchscreen-based gestures being compared to 6-DoF gestures [25]. The differences between two types of smart glasses/HWDs on interaction were also compared [32, 82]. This was a trend across all comparison studies, however, input method was compared more so than the device type.

Another factor that was compared by Bothen et al. [14] was the type of users (how their level of experience impacted results). They revealed that this factor significantly affected which input methods were the most appropriate for interaction. Alallah et al. [1] also compared the suitability of different inputs based on perspective (performer vs observer).

All elicitation studies considered how various tasks affected interaction approaches, yet Pham et al. [62] also focused on the impact of scale on interaction, and Tung et al. [82] considered social acceptability (how users approached interaction when in a public context).

The extent to which participants were restricted for elicitation studies was also a notable factor. For example, most studies only permit interaction through hand gestures [17, 62, 64], whereas Tung et al. [82] allowed participants to interact via multiple modalities (head, eye, speech, handheld input device) and Williams et al. [90, 91] through speech and/or hand gesture.

Previous research has also shown that elicitations have been designed so that users are seated, and responding to referents on a 2D monitor [25]. However, some studies considered delivering referents via the display itself [62, 64, 91], with Pham et al. [62] permitting users to physically move around the space and utilise the portability of HWDs, to assess how distance and scale of interactive content affects interaction.

Some elicitation studies focus on the influence of input modality, with hand being used both in comparison and conjunction with speech [90, 91], or touchscreen-based gestures being compared to 6-DoF gestures [25]. Although different combinations of input techniques were explored for elicitations, the inputs produced are often limited by the study design, by introducing bias from the referents used. This includes text prompting for speech interaction [90], or animations that encourage users to interact in a specific way [64]. Users are also often tempted to resort to interaction metaphors from their previous experience with technologies [32, 64].

Although 1 elicitation study permitted participants to freely interact via all of the inputs considered [82], the majority opted to implement hand gesture. This could also be due to past experience within the real world and with ubiquitous technologies, where generally interfaces and objects are operated manually or bi-manually.

Whereas many elicitation studies highlight patterns of reusable (i.e. a single gesture used for more than one function) and reversible gestures (i.e. the same gesture performed in opposing directions to complete different functions) [64], Pham et al. [62] state that designers need to account for scale, and not simply reuse gestures across different hologram sizes. They also highlighted the benefit of capturing the trajectories of inputs, as well as the gestures used, to account for variations in proposals (i.e. a clap or a pinch both representing a squashing motion [62]). This finding corresponds to the notion that gestures performed via different input methods can be mapped (i.e. hardware-based inputs can somewhat correspond to freehand gesture inputs [3]).

4.3.2 Evaluation

Assessments most frequently focused on capturing performance metrics and data surrounding general usability, by measuring factors such as time and error, and utilising a narrow set of questionnaires/Likert scales, as detailed in section 3.1. However, failing to explore factors outside of time, error and general usability when assessing interaction techniques is arguably detrimental.

We argue that a more diverse range of measures should be included for user studies, as considerations such as novelty and social acceptance are important when developing for realistic, long-term applications [73, 82]. Measures surrounding how interaction is impacted by environmental conditions are also important for understanding factors such as system robustness [83]. Therefore, it would be beneficial for a wider range of influences encompassing usability, such as novelty, social acceptance and robustness under diverse conditions, to be included as measures for assessments more frequently.

As discussed in section 3, comparison studies generally considered distance and scale of virtual content as variables. Research suggests that the nature of output significantly affects user approaches to interaction [62], therefore it is important to consider. However, studies could go beyond the size and distance of content to measure the impacts of a range of properties, such as colour, shape (i.e. uniform and non-uniform objects [32]) and the realism of interactive content. This includes factors surrounding visual semantic information that prompt psychological responses, such as different materials and temperatures [13].

Another factor often compared was the type of input, which is useful to uncover the most appropriate interaction techniques for different tasks. However, research should more frequently consider how the device type affects the results of input methods, as different types of display (i.e. optical/video, see-through/pass-through) will likely produce mixed findings [52]. As well as this, research should also consider devices with diverse topological structures (i.e. smart glasses vs headworn displays) and different types of handheld displays (i.e. tablets and smartphones), with distinct physical interfaces and screen

resolutions, which affect the suitability of interactions [82].

Another factor that was often disregarded was comparing the type of users. Although many papers capture participants' previous experience with technology, this is not often a primary consideration. However, Bothen et al. [14] highlights the importance of understanding participants past experience with interaction methods and technologies, to appropriately contextualise results. Consequently, we suggest that a more diverse range of participants would help to gain a better understanding of how to apply input techniques more universally. We argue that this diversity should go beyond experience to also include factors such as age, gender and culture, which are equally likely to affect interaction preferences and approaches.

Interestingly, 1 study was also found to compare the appropriateness of input methods based on 2 different perspectives; performer and observer [1]. Results surrounding the impact on both the user and those in their surroundings will become more significant, as immersive technologies become more widespread and are more often used in public environments. It will be necessary to also consider how interaction techniques affect bystanders, by exploring factors such as comfort, privacy and cultural/social acceptance in different environments and from a range of perspectives.

A primary limitation of elicitation studies is legacy bias (as introduced in section 4.2), however, methods have been outlined to tackle this affect [54,85]. These include *production* (requiring users to produce multiple interaction proposals for each referent), *priming* (encouraging users to consider capabilities of a new form factor or sensing technology) and *partners* (inviting users to participate in elicitation studies in groups, rather than individually) [54]. Despite this knowledge, few elicitation studies were found to employ these techniques [85,90]. Even though these methods can introduce further bias and complications [54], we concur that it would be highly beneficial to explore these methods for elicitations further.

Similar to our review, Villarreal-narvaez et al. [85] also reveal how elicitations primarily focus on hand-based input design, without considering multimodal input possibilities. Where participants were permitted to use any type of input [82], they generally opted to use freehand interaction. This is likely because it is not standard to interact via head and speech-based inputs outside of human-to-human communication, meaning participants are less likely to propose these types of interactions. However, as highlighted by this review, this does not mean that they are not more suited for specific tasks, or easily learned and understood by users [48, 74].

Further elicitation studies would therefore be needed, to understand how natural, multimodal approaches are applied under various real-world scenarios. This requires carefully preparing studies to allow for unrestricted approaches, that minimise sources of bias, notably through applying methods such as production, priming or pairing [54], and designing referents that do not prompt participants (i.e. by avoiding text labels, animations or task instructions [64, 90]).

Studies that allow users to create interactions for their own imagined applications of XR could provide more valuable insights, especially when considering ubiquitous applications of portable technologies. This implication is in line with a recent review of 216 elicitation studies [85], which highlights the possible sources of bias surrounding these restrictions, which may be negatively influencing user-defined approaches. Descriptions and designs of elicitation studies are often stripped from the context of use and the conditions in which the experiment took place, which limits the applicability of results.

Elicitation studies also tend to produce similar findings, which allude to reversible/reusable gestures, impacts of experience with previous technologies, as well as difficulties providing hand gestures for abstract tasks. This is likely the case as elicitation studies are primarily designed following the same methods (notably based on the research of Piumsomboon et al. [64]). Consequently, reconsidering approaches to elicitation studies will ensure that XR interfaces go beyond replicating interaction with standard platforms, to fully reap the benefits of immersive technologies [54, 85].

4.4 Use Case

The final factor discussed is use case, which is concerned with how differences surrounding users' situation, activities and environment impact interaction. Considerations regarding the use case are detailed, as well as the implications of failing to consider a diverse range of variables for user studies.

4.4.1 General Findings

The high percentage of studies conducted in lab environments represents the lack of experimentation in real-world conditions, which is in line with the results presented by Dey et al. [24]. This highlights no change in trends from 2005-2014. Ideally, studies would be conducted in (or simulate) real use cases, to maximise the value of the results generated. However, this is still not the case, with only 9 studies considering interaction in a realistic scenario.

As detailed in section 3 studies were predominately delivered in lab-based environments. Researchers sometimes attempted to simulate realistic conditions in a lab setting [44, 61], however, the majority of reviewed papers were highly controlled and restricted to a single condition.

Research suggests that factors sparsely explored, such as the pose and location of the user, impacts the appropriateness of input techniques. For example, when comparing 2 studies that explored interaction in public settings, where users were seated, hand gesture was by far the most employed input over any other type of modality and was preferred [82]. However, where participants were standing in open space, hand gesture was regarded as the least preferable input method [1].

Another factor relating to users' situation is the level of encumberment. Where users are required to employ their hands to operate the real environment, hands-free interaction is desirable. In such cases, head and/or speech input could be used as an alternative input method [74].

Another key area that requires further research is how interaction is affected by locomotion. Despite portability being a primary benefit of untethered display types, there were only 19 studies that allowed for locomotion when testing, and even fewer directly observed how movement affects interaction [34]. However, portable technologies are capable of going beyond what is plausible with static displays. They provide opportunities to effectively use immersive technologies for a broader range of applications and scenarios, as when the user is multitasking or on the go [83]. In circumstances where users are in locomotion, inputs could be adapted (i.e. walking path could be referenced via head directionality, for more subtle interaction in public settings [58]).

When considering testing differences in studies for AR or VR head-worn displays, VR is more likely to require viewport control, whereas this is employed less frequently overall with AR. However, several studies reported that in AR room-scale environments, participants preferred to interact from a distance as opposed to walking towards content [62, 89]. Manipulation of the scene could therefore provide further agency, or 'superpowers' to users, for situations where it is not desirable to physically approach interactive elements, such as when interacting in public places or under collaborative conditions.

An area that requires further attention also relates to the length of studies. In one case where a 5-day study was conducted [98], user performance tended to reach its peak after 3 days of practice, with users producing a steady performance from that point on. This suggests that where short studies are conducted, appropriate inputs could be dismissed simply because they have short learning curves.

4.4.2 Evaluation

Although conducting studies under highly controlled conditions will reveal usability when interacting in an ideal environment and scenario, the key to practical applications of immersive technologies is understanding how they can maintain usability and robustness under a range of diverse conditions, as is the case in real scenarios [83]. Therefore testing should focus more on external conditions that may affect performance and usability on a broader scale.

Factors surrounding the use case include users' location (i.e. whether interacting indoors or outdoors, the nature of their environment and the

ambient levels of light/noise), the crowdedness of an interaction space; in terms of the size of the environment and the density of surrounding people and objects (which can be measured subjectively or objectively), as well as the current state/activity of the user. This final category relates to considerations such as the level and type of encumberment (i.e. number of hands occupied and the types of objects being held, or if the user is in locomotion), and the task scenario (whether interaction is associated with fun or serious applications).

The results of highly controlled lab-based studies are arguably less applicable to standard interaction applications and environments. This is potentially a factor preventing widespread implementation of immersive technologies for practical use cases, that move beyond commercial applications.

The conditions a study is conducted under strongly relates to the concept of use case (the interaction scenario and environment). As a prominent finding is that use case has a strong influence on the most appropriate interaction methods [49, 82], user studies should aim to consider a more diverse range of variables and simulate realistic interaction conditions more closely. Because of the lack of diversity in study conditions, many results could be misleading, as users may even prefer different inputs in different use cases, and have better performances, after learning how to employ them [41].

When considering the growing range of application types for XR technologies, testing needs to explore the factors which affect interaction approaches and over a longer period, as opposed to only the objective measures of input techniques under ideal interaction conditions in a single instance. This will ensure that research can move away from observing usability for ad-hoc implementations, towards a more universal understanding of interaction with XR technologies, as they become more ubiquitous.

5 Conclusions and Recommendations

As we move towards consumer-level immersive applications, AR and VR technologies will become broader and more intertwined. Input designers will need to consider in what contexts applications are employed, and provide input techniques that are capable of adapting to users' situations, activities and surroundings; within both real and virtual environments. However, the interaction methods currently employed to develop applications are arguably not sustainable for the increasing emergence and diverse use cases of immersive technologies.

To address this, we have explored how different inputs have been applied and received, for a range of XR applications in different domains. This has led to the identification of trends and the primary advantages and disadvantages of input techniques, which are employed for consumer-level handheld and headworn devices.

Overall, results highlight the present absence of a single uniform solution to interaction. Furthermore, due to the range of users/use cases and devices, we highlight the current challenge for researchers and developers in applying robust logic, to seamlessly adapt inputs to tasks and scenarios. Despite this, the patterns highlighted in this review do confirm the appropriateness of certain input modalities for XR tasks (see tables 3 and 4). Findings also suggest that the most appropriate interaction approaches can be predicted, based on valuable trends attributed to the device, task and use case.

Based on the 68 papers reviewed, the following recommendations are also provided to prompt future research directions:

Test with a wider variety of user groups.
As highlighted in section 4.3, although participant demographics and past experience is often noted, user group is not generally a primary consideration. However, different users may have contrasting preferences surrounding input techniques. By conducting user studies with a more diverse range of user groups, patterns may be presented surrounding preferences for inputs, which could make it more straightforward to adapt interaction to each user.

Different user groups can be defined by considering a combination of factors, such as age, gender, cultural background, ability/disability and technology usage. Through creating mappings of how these considerations affect interaction approaches and user preferences (i.e. through tree data structures), we can work towards making immersive technologies more personalised for individual users, and more representative of a true population.

Pay closer attention to task scenarios.
As well as considering user demographics, we should pay close attention to the scenarios that users will be applying immersive technologies. As highlighted in sections 4.2 and 4.4, the suitability of different interaction techniques depends on the context that applications are being employed (i.e. for fun/at leisure, or for more serious tasks where time and error considerations are of high importance).

By considering the context of different immersive applications, and how AR/VR technologies will be used for a range of consumer use cases, we can better understand the advantages and disadvantages of input methods. The design of applications can then be tailored, to ensure they are transferable for the range of scenarios that immersive technologies will be used.

Consider how users' activity/situation will impact interaction.
Building on the task scenario, we should also consider under what activities and situations a user will interact. Key factors associated with users' activity and situation are highlighted in section 4.4.

Impairments, whether permanent or due to a users situation/activity, will directly impact the most appropriate interaction techniques. Consequently, it is important to understand how users adapt behaviours and interactions, depending on their circumstances, so designers can adapt input techniques accordingly. Because immersive technologies offer a broad range of use cases, the influence of activity/situation will be important to consider, and account for, when designing interaction techniques.

Further explore environmental and social constraints.
As well as understanding the impacts of users' activity/situation, we must also explore how the environment (and how the social acceptance associated with this environment) will impact interaction preferences. Usability studies should focus on testing in, or simulating, real-world scenarios, under diverse conditions. This will help to maximise social acceptance of immersive technologies and system usability/robustness.

As discussed in sections 4.3 and 4.4, research should be exploring how input approaches are affected by different social and environmental factors. It will also be important to consider how these factors can be measured and, depending on these variables, how different input modalities can harmonise the nature and flow of interaction.

Although testing in real conditions is not generally practical for scientific research, it is important to deliver more theoretical studies that focus on the future of interaction with these technologies. By understanding how different variables related to society and environment impact interaction, we can design input techniques that are more appropriate for realistic use cases/conditions.

Consider the provisions of emerging and future technologies.
Although it is important to research what is currently achievable, we should also be considering what we expect to be possible with XR technologies in the future (keeping this suggestion in mind will also help to address all of the recommendations provided). As detailed in section 4.1, this could be achieved by designing studies that eliminate the issues surrounding current technologies, or systems could be adapted/enhanced by modifying existing equipment. Adopting such techniques will ensure researchers are more in line with what is achievable when novel technologies are released. As opposed to recycling input approaches, we can focus on constantly making them better, as the technologies used for AR/VR are continuously improving.

Investigate how inputs/devices could be employed simultaneously.
As detailed in section 4.2, few studies have been designed to consider different combinations of input (primarily only 2 modalities), and how they can be used simultaneously, to improve usability. The findings of this review suggest that multimodal input can improve interaction by decreasing fatigue, improving system understanding and providing more interaction capabilities. We also note the benefits of using multiple displays simultaneously, which can provide multimodal inputs across two platforms (i.e. a smartphone coupled with a headworn display).

Table 3. Mapping the most appropriate inputs to distinct tasks on handheld displays: advantages and disadvantages.

Input Method	Advantages	Disadvantages
Hand	+ Can be combined with hardware-based techniques to provide enhanced performance for object **manipulation tasks** (translation/rotation/scale) [42] + Intuitive to employ [5, 42, 70, 80] + Able to be performed either at front or back of device [42] + More enjoyable and immersive for close range interaction [70, 80]	- **Direct manipulation** affected by hand-occlusion [42, 97] - Significantly slower than screen dwell techniques for **selection** [70] - Not always practical to employ as users generally require at least one hand to hold the device/prone to induce fatigue [5, 42]
Head	+ Effective for **pointing/**identifying objects and regions of interest [51] + Can be referenced to decrease completion time for **Abstract** speech commands as interaction requires shorter and less precise utterances [51]	- Affected by distance/location of targets (too close or too far) [51] - Requires holding phone in unnatural position to capture head directionality information [51] - Requires experiencing a learning curve [51]
Speech	+ Effective for **Abstract/menu-based interactions** and has a lower workload than hand/hardware-based input [51] + Can be used to improve interaction experience/ provide more interaction capabilities [59]	- Requires longer, more precise utterances when used in standalone [51]
Hardware-based	+ Raycasting techniques are fast and effective for **pointing/selecting** large, visible content [55, 97] + Hardware-based gestures (with 6-dof) provides an easy, natural and intuitive method for object/character control **(translation/rotation/scaling)** and can produce higher agreement rates than hand gestures (based on motion/ direction as opposed to hand gesture design) [96] + Touch and motion inputs can be separated into independent mechanisms (i.e. for pointing/selecting or translation/rotation) to improve usability [71, 80] + Touchscreen legacy gestures are generally easy and comfortable to employ for simple object manipulation tasks [31]	- Multitouch/ motion gesture interaction is often found more cumbersome for selection/ object manipulations (translation/ rotation/scale) and is prone to error, namely due to finger occlusions/sensor tracking [42, 81, 96] - raycasting techniques are less effective for pointing/selecting if targets are occluded or small [60, 97] - Touchscreen/ motion gestures have higher task-load than voice/Gaze [51] - Precision of hardware-based techniques for selection/ object manipulation is highly dependent on type of interactive content and the design of output (i.e. rod/cursor length and appearance [60, 81, 97]) - Motion inputs often require system adaptations such as user perspective rendering [75] to provide usable interactions for **rotation tasks** and target expansion for pointing/selecting and menu-based interactions [60] - Touchscreen-based interaction does not mimic object manipulations in the real world [5] and when used alone limits interaction capabilities [42, 59]

Therefore, we recommend considering how more intuitive forms of interaction, such as hand gesture and hardware-based input, can be best used alongside inputs like speech and head/gaze, for different tasks in immersive environments.

Investigate how inputs/devices could be employed interchangeably. Although we recommend that multimodal inputs should become more widely explored, to utilise all forms of input inherent to consumer devices more frequently, multimodal input is not always required/useful for all types of interaction. Therefore, it is also important to understand how to balance the use of unimodal and multimodal inputs, to maximise the effectiveness, usability and flow of interactions.

Even though different inputs are more suited to certain tasks (as defined in tables 3 and 4, and discussed in section 4.2), it is important to consider how to best employ techniques interchangeably, to minimise negative consequences such as fatigue, frustration and cognitive load. This also applies to different devices (i.e. some tasks are more suited to handheld displays and others better employed with headworn displays).

Further explore similarities/differences between AR and VR interaction. Exploring to what extent AR interaction is transferable to VR (and vice versa) is another important research direction. Although there are differences between AR and VR which affect interaction, they also require the consideration of very similar factors, especially regarding input methods and tasks. Therefore, it will be interesting to highlight and explore the factors that impact the appropriateness of different interaction techniques in AR and VR (such as those introduced in section 4.2). Researchers can then better establish to what extent a common set of interaction guidelines could be mapped and adopted for the spectrum of XR technology.

Revisit approaches to Elicitation studies.
To effectively understand how different input modalities can be employed simultaneously and interchangeably, for different XR environ-

Table 4. Mapping the most appropriate inputs to distinct tasks on headworn displays: advantages and disadvantages

Input Method	Advantages	Disadvantages
Hand	+ Most Intuitive [32, 41, 74, 89] + Useful for object manipulation tasks **(translation/rotation/scale)** [32, 63] + Effective when used occasionally/in moderation [11] + Accurate for **selection when content is in arms reach** [89, 101] + Gesture metaphors can be employed directly (i.e. pulling content closer [12]) for **viewport control**, or indirectly (i.e. employing a control metaphor based on joysticks for viewport control [76], or for tasks that require lower precision, to reduce fatigue [82])	- Prone to induce fatigue [11, 26, 30, 43, 95, 101] - Difficult to use gestures for more **abstract** interactions [63, 90] - Difficult to interact with smaller/ distant/ more dense content [63, 101] - **Scaling was sometimes found to be less practical/intuitive** [63], hand gesture being more appropriate for scaling when adopting metaphoric legacy gestures [32] - Generally not the most appropriate input for applications where time/error is a concern [26, 30, 43] (i.e. effected by engagement/disengagement times [74] and boundaries of interaction zone due to limited FOV [94]) - Lacks tangible support [66] - Effected by social acceptance [1, 82]
Head	+ Effective **pointing/selection** mechanism [22, 26, 30, 43] + Less physically demanding that hand input [26, 43] + Most effective primary input for **hands-free applications** [11, 26, 43] + Discreet head movements such as nods or tilts are effective for **menu-based/abstract interactions** such as switching [68, 98] and can be employed as opposed to dwell for **selection**, to provide more control over the pace of interaction [48] + Provides an effective additional source of input to improve accuracy/prediction models [93] and account for ambiguities [37] + Shown to be faster than hand input for **translation/scale** tasks [74]	- Dwell interaction is slower and more demanding that employing an external controller (i.e. clicker/touch pad for selection [22, 26, 30, 43]) - Less intuitive than hand input and has a short learning curve [41, 48] - Effected by social acceptance [1, 48] - **Rotation tasks** are difficult to achieve [74]
Speech	+ Most appropriate for **abstract interactions** [21, 90, 91] + Effective **hands-free selection/menu-based mechanism** [67, 93] + **Can aid with scaling/rotation tasks** when used alongside hand input [90], especially as size of content decreases and the number of objects increases [63] + Allows user to focus on the task as opposed to the means of interaction [93] + Not effected by distance of interactive elements [89]	- Difficult imagining **rotation/translation tasks** via speech [63, 89] - Low preference and social acceptance [89] - Often experiences high error rates (especially with shorter utterances) [49]
Hardware-based	+ Allows for less noticeable interactions as input is not dependent on computer vision technologies (indirect control) [30] + Offers tangible support [65] + Shown to provide better performance/user experience than other techniques for **pointing/selecting tasks** [30, 33, 95] + Often deemed the least tiring technique for selection [30,43]	- Requires additional hardware (less practical/cost efficient) [95] - Not as accurate as head or speech input for **selecting distant content** [89]

ments, we must reconsider how interactions are designed and delivered. As highlighted in sections 4.3 and 4.2, this can be achieved by employing carefully designed elicitation studies, that go beyond providing standard referents, to place minimal restrictions on users. By exploring how a range of users adapt their input choices and behaviours (in different representative scenarios, environments and conditions), we can begin to understand how to adapt system behaviours accordingly.

5.1 Limitations

Although reviewing a corpus of papers has provided an overview of the trends surrounding explicit interaction, this research (as with other reviews) is limited by the search criteria, the databases employed and the publication dates included.

Furthermore, the review does not consider the citation count for particular papers and therefore the potential significance of each paper discussed. If this were considered, papers deemed most influential could be prioritised and potential richer insights found. However, as citations accumulate over time, it is most likely that this approach would exclude, or negatively bias, more recent papers (which could prove influential in future XR development [24]). Sample size for each study was also considered but was not used as part of the inclusion/exclusion criteria. Again, this may have impacted the potential significance of the results, however, we believe that this leads to a more representative review of publications.

Another possible limitation is that both AR and VR technologies were considered for the review. Although these technologies share many similarities, especially surrounding input techniques, their differences will impact users' preferences and approaches (due to factors such as the provided level of embodiment/awareness and variations in interaction approaches with real and virtual content).

Finally, owing to the proliferation of some input paradigms (notably hand/manual input) the review has a higher number of studies using specific devices/inputs. While this may inherently skew/bias some of the findings, it is representative of published data. However, we still recommend further exploration of alternative modes for inputs (i.e head, gaze, speech) for future research in immersive technology.

Despite these limitations, this review helps to contextualise the use of input modalities for different commonplace tasks in immersive environments. Future research directions are highlighted, as well as some notable advantages and shortcomings of interaction approaches.

REFERENCES

[1] F. Alallah, A. Neshati, Y. Sakamoto, K. Hasan, E. Lank, A. Bunt, and P. Irani. Performer vs. observer. *Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology*, 11 2018.

[2] J. Aliprantis, M. Konstantakis, R. Nikopoulou, P. Mylonas, and G. Caridakis. Natural interaction in augmented reality context. In *VIPERC@IRCDL*, 2019.

[3] R. Arora, R. H. Kazi, D. M. Kaufman, W. Li, and K. Singh. Magicalhands: Mid-air hand gestures for animating in vr. *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*, 10 2019.

[4] B. Bach, R. Sicat, J. Beyer, M. Cordeil, and H. Pfister. The hologram in my hand: How effective is interactive exploration of 3d visualizations in immersive tangible augmented reality? *IEEE Transactions on Visualization and Computer Graphics*, 24:457–467, 01 2018.

[5] H. Bai, G. A. Lee, M. Ramakrishnan, and M. Billinghurst. 3d gesture interaction for handheld augmented reality. *SIGGRAPH Asia 2014 Mobile Graphics and Interactive Applications on - SA '14*, pages 1–6, 11 2014.

[6] H. Bai, P. Sasikumar, J. Yang, and M. Billinghurst. A user study on mixed reality remote collaboration with eye gaze and hand gesture sharing. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI 2020, April 25–30, 2020, Honolulu, HI, USA, 04 2020.

[7] Z. Bai and A. F. Blackwell. Analytic review of usability evaluation in ismar. *Interacting with Computers*, 24:450–460, 11 2012.

[8] C. Bailly, F. Leitner, and L. Nigay. Head-controlled menu in mixed reality with a hmd. *Human-Computer Interaction – INTERACT 2019*, pages 395–415, 2019.

[9] M. W. Bazzaza, B. Al Delail, M. J. Zemerly, and J. W. Ng. iarbook: An immersive augmented reality system for education. *2014 IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE)*, pages 495–498, 12 2014.

[10] V. Becker, F. Rauchenstein, and G. Sörös. Investigating universal appliance control through wearable augmented reality. *Proceedings of the 10th Augmented Human International Conference 2019*, pages 1–9, 03 2019.

[11] I. Belkacem, I. Pecci, and B. Martin. Pointing task on smart glasses: Comparison of four interaction techniques. *arXiv:1905.05810 [cs]*, 05 2019.

[12] S. Bhowmick, P. Kalita, and K. Sorathia. A gesture elicitation study for selection of nail size objects in a dense and occluded dense hmd-vr. *IndiaHCI '20: Proceedings of the 11th Indian Conference on Human-Computer Interaction*, pages 12–23, 11 2020.

[13] A. D. Blaga, M. Frutos-Pascual, C. Creed, and I. Williams. Too hot to handle: An evaluation of the effect of thermal visual representation on user grasping interaction in virtual reality. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 04 2020.

[14] S. Bothén, J. Font, and P. Nilsson. An analysis and comparative user study on interactions in mobile virtual reality games. *Proceedings of the 13th International Conference on the Foundations of Digital Games*, pages 1–8, 08 2018.

[15] N. Brancati, G. Caggianese, M. Frucci, L. Gallo, and P. Neroni. Experiencing touchless interaction with augmented content on wearable head-mounted displays in cultural heritage applications. *Personal and Ubiquitous Computing*, 21:203–217, 11 2016.

[16] J. Brooke. Sus: A quick and dirty usability scale. *Usability Eval. Ind.*, 189, 11 1995.

[17] E. Chan, T. Seyed, W. Stuerzlinger, X.-D. Yang, and F. Maurer. User elicitation on single-hand microgestures. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 05 2016.

[18] C. S. Che Dalim, M. S. Sunar, A. Dey, and M. Billinghurst. Using augmented reality with speech input for non-native children's language learning. *International Journal of Human-Computer Studies*, 134:44–64, 02 2020.

[19] N. Cheema, L. A. Frey-Law, K. Naderi, J. Lehtinen, P. Slusallek, and P. Hämäläinen. Predicting mid-air interaction movements and fatigue using deep reinforcement learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–13, New York, NY, USA, 2020. Association for Computing Machinery.

[20] D. L. Chen, R. Balakrishnan, and T. Grossman. Disambiguation techniques for freehand object manipulations in virtual reality. *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 285–292, 03 2020.

[21] Z. Chen, J. Li, Y. Hua, R. Shen, and A. Basu. Multimodal interaction in augmented reality. *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Banff, AB, Canada,5-8 Oct. 2017:206–209, 10 2017.

[22] L. Chittaro and R. Sioni. Selecting menu items in mobile head-mounted displays: Effects of selection technique and active area. *International Journal of Human–Computer Interaction*, 35:1501–1516, 11 2018.

[23] C. S. C. Dalim, A. Dey, T. Piumsomboon, M. Billinghurst, and S. Sunar. Teachar: An interactive augmented reality tool for teaching basic english to non-native children. *2016 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct)*, pages 82–86, 09 2016.

[24] A. Dey, M. Billinghurst, R. W. Lindeman, and J. E. Swan II. A systematic review of usability studies in augmented reality between 2005 and 2014. *2016 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct)*, pages 49–50, 09 2016.

[25] Z. Dong, T. Piumsomboon, J. Zhang, A. Clark, H. Bai, and R. Lindeman. A comparison of surface and motion user-defined gestures for mobile augmented reality. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 04 2020.

[26] A. Esteves, Y. Shin, and I. Oakley. Comparing selection mechanisms for gaze input techniques in head-mounted displays. *International Journal of Human-Computer Studies*, 139:102414, 07 2020.

[27] A. Esteves, D. Verweij, L. Suraiya, R. Islam, Y. Lee, and I. Oakley. Smoothmoves: Smooth pursuits head movements for augmented reality. *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*, 10 2017.

[28] F. E. Fadzli and A. W. Ismail. Voxar: 3d modelling editor using real hands gesture for augmented reality. *2019 IEEE 7th Conference on*

Systems, Process and Control (ICSPC), pages 242–247, 12 2019.

[29] C. Flavián, S. Ibáñez-Sánchez, and C. Orús. The impact of virtual, augmented and mixed reality technologies on the customer experience. *Journal of Business Research*, 100:547–560, 11 2018.

[30] J. Franco and D. Cabral. Augmented object selection through smart glasses. *Proceedings of the 18th International Conference on Mobile and Ubiquitous Multimedia*, 11 2019.

[31] J. A. Frank, M. Moorhead, and V. Kapila. Realizing mixed-reality environments with tablets for intuitive human-robot collaboration for object manipulation tasks, 08 2016.

[32] M. Frutos-Pascual, C. Creed, and I. Williams. Head mounted display interaction evaluation: Manipulating virtual objects in augmented reality. *Human-Computer Interaction – INTERACT 2019*, 11749:287–308, 2019.

[33] P. Ganapathi and K. Sorathia. Investigating controller less input methods for smartphone based virtual reality platforms. *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*, 09 2018.

[34] D. Ghosh, P. S. Foong, S. Zhao, C. Liu, N. Janaka, and V. Erusu. Eyeditor: Towards on-the-go heads-up text editing using voice and manual input. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 04 2020.

[35] E. S. Goh, M. S. Sunar, and A. W. Ismail. 3d object manipulation techniques in handheld mobile augmented reality interface: A review. *IEEE Access*, 7:40581–40601, 2019.

[36] J. Henderson, J. Ceha, and E. Lank. Stat: Subtle typing around the thigh for head-mounted displays. *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services*, pages 1–11, 10 2020.

[37] R. Henrikson, T. Grossman, S. Trowbridge, D. Wigdor, and H. Benko. Head-coupled kinematic template matching: A prediction model for ray pointing in vr. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1,14, 04 2020.

[38] J. Hertel, S. Karaosmanoglu, S. Schmidt, J. Braker, M. Semmann, and F. Steinicke. A taxonomy of interaction techniques for immersive augmented reality based on an iterative literature review. *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 10 2021.

[39] T. N. T. T. T. L. Index. Tlx @ nasa ames - home, 12 2020.

[40] P. Jackson. Understanding understanding and ambiguity in natural language. *Procedia Computer Science*, 169:209–225, 2020.

[41] H. J. Kang, J.-h. Shin, and K. Ponto. A comparative analysis of 3d user interaction: How to move virtual objects in mixed reality, 03 2020.

[42] M. Kim and J. Y. Lee. Touch and hand gesture-based interactions for directly manipulating 3d virtual objects in mobile augmented reality. *Multimedia Tools and Applications*, 75:16529–16550, 02 2016.

[43] D. Krupke, F. Steinicke, P. Lubos, Y. Jonetzko, M. Gorner, and J. Zhang. Comparison of multimodal heading and pointing gestures for co-located mixed reality human-robot interaction. *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 10 2018.

[44] W. S. Lages and D. A. Bowman. Walking with adaptive augmented reality workspaces. *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 356–366, 03 2019.

[45] F. Lamberti, F. Manuri, G. Paravati, G. Piumatti, and A. Sanna. Using semantics to automatically generate speech interfaces for wearable virtual and augmented reality applications. *IEEE Transactions on Human-Machine Systems*, 47:152–164, 02 2017.

[46] J. J. Laviola, E. Kruijff, R. P. Mcmahan, D. A. Bowman, and I. Poupyrev. *3D user interfaces : theory and practice*. Addison-Wesley, 2017.

[47] M. Lee, M. Billinghurst, W. Baek, R. Green, and W. Woo. A usability study of multimodal input in an augmented reality environment. *Virtual Reality*, 17:293–305, 09 2013.

[48] X. Lu, D. Yu, H.-N. Liang, X. Feng, and W. Xu. Depthtext: Leveraging head movements towards the depth dimension for hands-free text entry in mobile virtual reality systems. *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 1060–1061, 03 2019.

[49] F. Manuri and G. Piumatti. A preliminary study of a hybrid user interface for augmented reality applications. *Proceedings of the 7th International Conference on Intelligent Technologies for Interactive Entertainment*, pages 37–41, 2015.

[50] B. Marques, J. Alves, M. Neves, I. Justo, A. Santos, R. Rainho, R. Maio, D. Costa, C. Ferreira, P. Dias, and B. S. Santos. Interaction with virtual content using augmented reality. *Proceedings of the ACM on Human-Computer Interaction*, 4:1–17, 11 2020.

[51] S. Mayer, G. Laput, and C. Harrison. Enhancing mobile voice assistants with worldgaze. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 04 2020.

[52] D. Medeiros, M. Sousa, D. Mendes, A. Raposo, and J. Jorge. Perceiving depth: Optical versus video see-through. *Proceedings of the 22nd ACM Conference on Virtual Reality Software and Technology*, 11 2016.

[53] P. Mohan, W. Boon Goh, C.-W. Fu, and S.-K. Yeung. Head-fingers-arms: Physically-coupled and decoupled multimodal interaction designs in mobile vr. *The 17th International Conference on Virtual-Reality Continuum and its Applications in Industry*, pages 1–9, 11 2019.

[54] M. R. Morris, A. Danielescu, S. Drucker, D. Fisher, B. Lee, c. schraefel, and J. O. Wobbrock. Reducing legacy bias in gesture elicitation studies. *interactions*, 21:40–45, 05 2014.

[55] A. Mossel, B. Venditti, and H. Kaufmann. 3dtouch and homer-s. *Proceedings of the Virtual Reality International Conference: Laval Virtual*, pages 1–10, 03 2013.

[56] S. S. Muhammad Nizam, R. Zainal Abidin, N. Che Hashim, M. C. Lam, H. Arshad, and N. A. Abd Majid. A review of multimodal interaction technique in augmented reality environment. *International Journal on Advanced Science, Engineering and Information Technology*, 8:1460, 09 2018.

[57] B. Munsinger, G. White, and J. Quarles. The usability of the microsoft hololens for an augmented reality game to teach elementary school children, 09 2019.

[58] F. Müller, M. Schmitz, D. Schmitt, S. Günther, M. Funk, and M. Mühlhäuser. Walk the line: Leveraging lateral shifts of the walking path as an input modality for head-mounted displays. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 04 2020.

[59] N. I. A. M. Nazri and D. R. A. Rambli. The roles of input and output modalities on user interaction in mobile augmented reality application. *Proceedings of the Asia Pacific HCI and UX Design Symposium*, pages 46–49, 12 2015.

[60] P. Perea, D. Morand, and L. Nigay. Target expansion in context: the case of menu in handheld augmented reality. *Proceedings of the International Conference on Advanced Visual Interfaces*, pages 1–9, 09 2020.

[61] A. Pereira, E. J. Carter, I. Leite, J. Mars, and J. F. Lehman. Augmented reality dialog interface for multimodal teleoperation. *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 08 2017.

[62] T. Pham, J. Vermeulen, A. Tang, and L. MacDonald Vermeulen. Scale impacts elicited gestures for manipulating holograms. *Proceedings of the 2018 on Designing Interactive Systems Conference 2018 - DIS '18*, 2018.

[63] T. Piumsomboon, D. Altimira, H. Kim, A. Clark, G. Lee, and M. Billinghurst. Grasp-shell vs gesture-speech: A comparison of direct and indirect natural interaction techniques in augmented reality. *2014 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 09 2014.

[64] T. Piumsomboon, A. Clark, M. Billinghurst, and A. Cockburn. User-defined gestures for augmented reality. *Human-Computer Interaction – INTERACT 2013*, 8118:282–299, 2013.

[65] C. Plasson, D. Cunin, Y. Laurillau, and L. Nigay. Tabletop ar with hmd and tablet. *Proceedings of the 2019 ACM International Conference on Interactive Surfaces and Spaces*, pages 409–414, 11 2019.

[66] C. Plasson, D. Cunin, Y. Laurillau, and L. Nigay. 3d tabletop ar. *Proceedings of the International Conference on Advanced Visual Interfaces*, 09 2020.

[67] M. Pourmemar and C. Poullis. Visualizing and interacting with hierarchical menus in immersive augmented reality. *The 17th International Conference on Virtual-Reality Continuum and its Applications in Industry*, pages 1–9, 11 2019.

[68] M. Prilla, M. Janßen, and T. Kunzendorff. How to interact with augmented reality head mounted devices in care work? a study comparing handheld touch (hands-on) and gesture (hands-free) interaction. *AIS Transactions on Human-Computer Interaction*, 11:157–178, 09 2019.

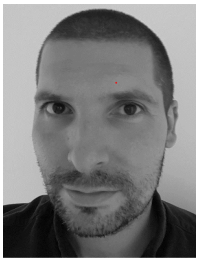
[69] A. Pringle, S. Hutka, J. Mom, R. van Esch, N. Heffernan, and P. Chen. Ethnographic study of a commercially available augmented reality hmd app for industry work instruction. *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, pages 389–397, 06 2019.

[70] J. Qian, D. A. Shamma, D. Avrahami, and J. Biehl. Modality and depth in touchless smartphone augmented reality interactions. *ACM International*

Conference on Interactive Media Experiences, pages 74–81, 06 2020.
[71] Y. Y. Qian and R. J. Teather. The eyes don't have it: An empirical comparison of head-based and eye-based selection in virtual reality. *Proceedings of the 5th Symposium on Spatial User Interaction*, 10 2017.
[72] N. Rao, L. Zhang, S. L. Chu, K. Jurczyk, C. Candelora, S. Su, and C. Kozlin. Investigating the necessity of meaningful context anchoring in ar smart glasses interaction for everyday learning, 03 2020.
[73] I. Rutten and D. Geerts. Better because it's new: The impact of perceived novelty on the added value of mid-air haptic feedback. CHI '20, page 1–13, New York, NY, USA, 2020. Association for Computing Machinery.
[74] S. Sadri, S. A. Kohen, C. Elvezio, S. H. Sun, A. Grinshpoon, G. J. Loeb, N. Basu, and S. K. Feiner. Manipulating 3d anatomic models in augmented reality: Comparing a hands-free approach and a manual approach. *2019 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 93 – 102, 10 2019.
[75] A. Samini and K. L. Palmerius. A study on improving close and distant device movement pose manipulation for hand-held augmented reality. *Proceedings of the 22nd ACM Conference on Virtual Reality Software and Technology*, pages 121–128, 11 2016.
[76] K. A. Satriadi, B. Ens, M. Cordeil, B. Jenny, T. Czauderna, and W. Willett. Augmented reality map navigation with freehand gestures. *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 593–603, 03 2019.
[77] J. Schoonenboom and R. B. Johnson. How to construct a mixed methods research design. *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 69:107–131, 07 2017.
[78] M. Schrepp. User experience questionnaire handbook, 09 2015.
[79] M. Speicher, B. D. Hall, and M. Nebeling. What is mixed reality? *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 05 2019.
[80] G. E. Su, M. S. Sunar, and A. W. Ismail. Device-based manipulation technique with separated control structures for 3d object translation and rotation in handheld mobile ar. *International Journal of Human-Computer Studies*, 141:102433, 09 2020.
[81] T. Tanikawa, H. Uzuka, T. Narumi, and M. Hirose. Integrated view-input ar interaction for virtual object manipulation using tablets and smartphones. *Proceedings of the 12th International Conference on Advances in Computer Entertainment Technology*, pages 1–8, 11 2015.
[82] Y.-C. Tung, C.-Y. Hsu, H.-Y. Wang, S. Chyou, J.-W. Lin, P.-J. Wu, A. Valstar, and M. Y. Chen. User-defined game input for smart glasses in public space. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*, pages 3327–3336, 2015.
[83] M. Turk. Multimodal interaction: A review. *Pattern Recognition Letters*, 36:189–195, 2014.
[84] A. E. Uva, M. Fiorentino, V. M. Manghisi, A. Boccaccio, S. Debernardis, M. Gattullo, and G. Monno. A user-centered framework for designing midair gesture interfaces. *IEEE Transactions on Human-Machine Systems*, 49:421–429, 10 2019.
[85] S. Villarreal-Narvaez, J. Vanderdonckt, R.-D. Vatavu, and J. O. Wobbrock. A systematic review of gesture elicitation studies. *Proceedings of the 2020 ACM Designing Interactive Systems Conference*, 07 2020.
[86] J. Väyrynen, M. Suoheimo, A. Colley, and J. Häkkilä. Exploring head mounted display based augmented reality for factory workers. *Proceedings of the 17th International Conference on Mobile and Ubiquitous Multimedia*, pages 499–505, 11 2018.
[87] K. Waldow, M. Misiak, U. Derichs, O. Clausen, and A. Fuhrmann. An evaluation of smartphone-based interaction in ar for constrained object manipulation. *Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology*, pages 1–2, 11 2018.
[88] Z. Wang, H. Yu, H. Wang, Z. Wang, and F. Lu. Comparing single-modal and multimodal interaction in an augmented reality system. *2020 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, 1:165–166, 11 2020.
[89] M. Whitlock, E. Harnner, J. R. Brubaker, S. Kane, and D. A. Szafir. Interacting with distant objects in augmented reality. *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 42–48, 03 2018.
[90] A. S. Williams, J. Garcia, and F. Ortega. Understanding multimodal user gesture and speech behavior for object manipulation in augmented reality using elicitation. *IEEE Transactions on Visualization and Computer Graphics*, 26:3479–3489, 12 2020.
[91] A. S. Williams and F. R. Ortega. Understanding gesture and speech multimodal interactions for manipulation tasks in augmented reality using unconstrained elicitation. *Proceedings of the ACM on Human-Computer Interaction*, 4:1–21, 11 2020.
[92] J. O. Wobbrock, M. R. Morris, and A. D. Wilson. User-defined gestures for surface computing. *Proceedings of the 27th international conference on Human factors in computing systems - CHI 09*, page 1083–1092, 2009.
[93] E. Wolf, S. Klüber, C. Zimmerer, J.-L. Lugrin, and M. E. Latoschik. "paint that object yellow": Multimodal interaction to enhance creativity during design tasks in vr. *2019 International Conference on Multimodal Interaction*, pages 195–204, 10 2019.
[94] W. Xu, H.-N. Liang, Y. Chen, X. Li, and K. Yu. Exploring visual techniques for boundary awareness during interaction in augmented reality head-mounted displays, 03 2020.
[95] W. Xu, H.-N. Liang, A. He, and Z. Wang. Pointing and selection methods for text entry in augmented reality head mounted displays. *2019 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 279 – 288, 10 2019.
[96] H. Ye, K. C. Kwan, W. Su, and H. Fu. Aranimator: In-situ character animation in mobile ar with user-defined motion gestures. *ACM Transactions on Graphics*, 39, 07 2020.
[97] J. Yin, C. Fu, X. Zhang, and T. Liu. Precise target selection techniques in handheld augmented reality interfaces. *IEEE Access*, 7:17663–17674, 2019.
[98] D. Yu, H.-N. Liang, X. Lu, T. Zhang, and W. Xu. Depthmove: Leveraging head motions in the depth dimension to interact with virtual reality head-worn displays. *2019 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 103 – 114, 10 2019.
[99] J. Zhao, C. J. Parry, R. dos Anjos, C. Anslow, and T. Rhee. Voice interaction for augmented reality navigation interfaces with natural language understanding. *2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ)*, pages 1–6, 11 2020.
[100] F. Zhu and T. Grossman. Bishare: Exploring bidirectional interactions between smartphones and head-mounted augmented reality. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 04 2020.
[101] K. Özacar, J. D. Hincapié-Ramos, K. Takashima, and Y. Kitamura. 3d selection techniques for mobile augmented reality head-mounted displays. *Interacting with Computers*, 12 2016.

Becky Spittle is a PhD student within the Digital Media Technology Lab (DMT Lab) at Birmingham City University. Her research interests are centred around Human-Computer Interaction (HCI), User Experience (UX) Design, Immersive technologies (Augmented Reality/Mixed Reality/Virtual Reality, AR/MR/VR) and Multimodal Interaction. Her PhD research explores the Transferability of Interaction Techniques for Immersive Technologies. She is keen to apply her knowledge of UX design and user-centred research practices, to provide further meaningful contributions to HCI and AR/VR fields.

Dr Maite Frutos-Pascual is a senior lecturer and active researcher at the Digital Media Technology Lab in Birmingham City University, UK. She specialises in Human Computer Interaction (HCI), immersive technologies (Augmented Reality and Virtual Reality AR/VR), usability, user analysis, interactive systems and sensor data analysis and integration. Her special interest is on virtual object manipulation, supervising PhD students in this area and collaborating with industry partners in bringing immersive systems outside laboratory environments. She has an extensive list of research outputs in key HCI and AR/VR venues.



Dr Chris Creed is an Associate Professor and head of the Human Computer Interaction group in the Digital Media Technology Lab (DMT Lab) at Birmingham City University. His core research interest is in the design and development of assistive technology for disabled people across a range of impairments and has extensive experience in leading collaborative technical projects exploring the use of innovative technologies.

Dr Ian Williams received his PhD from Manchester Metropolitan University in 2008 in low level feature analysis and Artificial Intelligence for multiple scale edge detection in biomedical images. He is an Associate Professor and head of the Digital Media Technology Lab (DMT Lab) at Birmingham city University. His work spans many concepts of visual and interactive computing with a key emphasis on creating novel methods for improving the Quality of Experience for users interacting and using Augmented Reality (AR) and Virtual Reality (VR) systems.