Frontiers in Virtual Reality

TYPE Review
PUBLISHED 31 July 2023
DOI 10.3389/frvir.2023.1171230

OPEN ACCESS

EDITED BY
Andrea Sanna,
Polytechnic University of Turin, Italy

REVIEWED BY
Mark Billinghurst,
University of South Australia, Australia
Mingze Xi,
Commonwealth Scientific and Industrial Research Organisation (CSIRO), Australia

*CORRESPONDENCE
Richard Nguyen,
richard.nguyen@usherbrooke.ca

RECEIVED 21 February 2023
ACCEPTED 18 July 2023
PUBLISHED 31 July 2023

CITATION
Nguyen R, Gouin-Vallerand C and Amiri M (2023), Hand interaction designs in mixed and augmented reality head mounted display: a scoping review and classification.
Front. Virtual Real. 4:1171230.
doi: 10.3389/frvir.2023.1171230

COPYRIGHT
© 2023 Nguyen, Gouin-Vallerand and Amiri. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Hand interaction designs in mixed and augmented reality head mounted display: a scoping review and classification

Richard Nguyen[1]*, Charles Gouin-Vallerand[2] and Maryam Amiri[3]

[1]DOMUS Laboratory, Department of Computer Science, Université de Sherbrooke, Sherbrooke, QC, Canada, [2]DOMUS Laboratory, Business School, Université de Sherbrooke, Sherbrooke, QC, Canada, [3]VMware Canada, Ottawa, ON, Canada

Mixed reality has made its first step towards democratization in 2017 with the launch of a first generation of commercial devices. As a new medium, one of the challenges is to develop interactions using its endowed spatial awareness and body tracking. More specifically, at the crossroad between artificial intelligence and human-computer interaction, the goal is to go beyond the Window, Icon, Menu, Pointer (WIMP) paradigm humans are mainly using on desktop computer. Hand interactions either as a standalone modality or as a component of a multimodal modality are one of the most popular and supported techniques across mixed reality prototypes and commercial devices. In this context, this paper presents scoping literature review of hand interactions in mixed reality. The goal of this review is to identify the recent findings on hand interactions about their design and the place of artificial intelligence in their development and behavior. This review resulted in the highlight of the main interaction techniques and their technical requirements between 2017 and 2022 as well as the design of the Metaphor-behavior taxonomy to classify those interactions.

KEYWORDS

augmented reality, mixed reality, hand interaction, POST WIMP, hand grasp, gestures, machine learning, scoping review

1 Introduction

Mixed Reality embodies experiences that involve both the physical world and virtual contents. Following the popular Virtuality-Reality Continuum from (Milgram et al., 1995), it encompasses Augmented Reality, which consists in adding virtual content on the real world and Augmented Virtuality, which consists in representing physical objects in virtual environments.

For this new medium, hand interactions are one of the most popular modalities to manipulate virtual content in mixed reality. Indeed, as we are used to grab and manipulate physical objects to explore the real world, this modality is intuitive and perceived as natural. Recent headsets such as HoloLens 2, Meta 2 or Magic Leap One support natively hand interactions. Besides, external sensors such as Leap Motion (Kim et al., 2019) and Myo Armband (Bautista et al., 2020) are also used to improve or enable hand interactions with older headsets. This modality has been made possible with the progress of computer vision and to some extent the progress of machine learning. Mixed reality headsets are endowed with sensors and computer vision algorithms that allow the machine to understand the context around the user, including but not limited to, tracking the hand movements in real-time. As such, it is possible to design more intelligent user interfaces where virtual content and physical objects are manipulated seamlessly.

According to Billinghurst et al. (2015)'s new interface medium adoption steps, mixed and augmented reality are still at the second step of adoption where they still use the desktop computer interface metaphors called the Window, Icon, Menu, pointer (WIMP) paradigm. The goal is to go beyond this paradigm and reach the third step of adoption by designing new interface metaphors suited to mixed and augmented reality. By definition mixed reality encompasses interactions with both virtual content and the real physical world. As such, the new interface and its metaphors will aim at closing the gap between what is real and what is computer generated. This can be done, for example, by enabling the user to manipulate virtual content as he would do with their physical counterpart or allowing him to seamlessly switch from interacting with the real and the virtual using diverse modalities and effectors.

In this paper, we are proposing a scoping literature review on hand interactions in mixed reality over the period 2017–2022. The year 2017 was chosen as the starting point of our review because researchers and developers have been able to start prototyping with commercial devices with the form under which they are popularized nowadays. Indeed HoloLens 1 and Meta 2[1] were released in 2016 and were mostly shipped globally early 2017. Magic Leap One a direct competitor was also announced at the end of 2017 for a release in 2018. We aim at providing in this paper an outline of the hand interaction techniques designed and their technical requirements to give students and researchers an overview of the recent trends in the context where accessible headsets for prototyping has been made available. We also propose a Metaphor-Behavior taxonomy to describe hand interactions which extends the model of Frutos-Pascual et al. (2019) and Macaranas et al. (2015), and was formulated while making this review.

2 Related works

2.1 What is an interaction in mixed reality?

In the work of Hand (1997), a user in virtual reality is acting to achieve four fundamental tasks.

1. Navigation;
2. Selection;
3. Manipulation;
4. Application Control.

The model proposed by Hand (1997) can be applied to mixed reality as the user is also evolving in a 3D environment. The difference is that in this context, the environment is composed of both virtual and physical content as targets of the interactions. The navigation task refers to exploring the environment by changing the point of view on the scene. In virtual reality, it consists in changing the point of view of the camera that represents the head of the user. To extend this model to augmented and mixed reality, the first task will also refer to moving in the physical world that is spatially registered by the system. Indeed, in augmented and mixed reality, there is a need to synchronize the position of the user in the virtual environment reference frame and the physical world environment reference frame. This is mostly done by the computing of a common origin through calibrations to match the two coordinate systems. More advanced systems also map the real environment topography to a 3D mesh that allows virtual content to interact with physical surfaces.The selection task refers to designating an object as the focus of the interaction with the user. The object can be both virtual and real in mixed reality.The manipulation task refers to changing the properties of a designated object such as the position or its shape. The application control task refers to triggering functionalities and communicating information to the application. To achieve those goals, the user has several modalities to convey his intention to the application and interact with the environment. The most popular modalities in commercial devices are: hand control, controllers, voice commands and head/eyes control.

2.2 Hand interaction

Based on the previous definition, hand interactions can be separated into two categories in most commercial mixed reality applications.

- gesture based interactions;
- trajectories and simulated physic-based interactions.

A gesture according to Bouchard et al. (2014) is an expressive and meaningful body motion (hand, face, arms, etc.) that convey a message or more generally, embed important information of spatio-temporal nature. In the work of Vuletic et al. (2019), a systematic review on hand gesture, hand gestures can be classified by three ways: temporal classification, contextual classification and instruction based classification. Temporal classification separates gestures in two classes: static or dynamic, the first one being hand poses while the second one is composed of several hand moves. Contextual classification describe what the gestures are used for and how do they convey the information in comparison to speech language. In this classification model, gestures can be separated in two classes: communicative and manipulative. In the communicative class the sub-categories are.

- Symbolic gestures, representing a symbolic object or concept with a cultural semantic without having a direct morphological relation to the designated object;
- Semaphoric gestures, formalizing a dictionary of gesture-to-action without any semantic background;
- Pantomimic gestures, imitating a sequence of actions to describe a narrative.
- Iconic gestures, illustrating the content of the speech by describing a shape, a spatial relation or an action of an object;
- Metaphoric gestures, illustrating the content of the speech by describing an abstract concept;
- Modalizing symbolic gestures or Beat gestures, following the rhythm of the speech while emphasizing on part of it;
- Cohesive gestures, highlighting the continuation of a topic that has been interrupted by using a recurrent gesture.

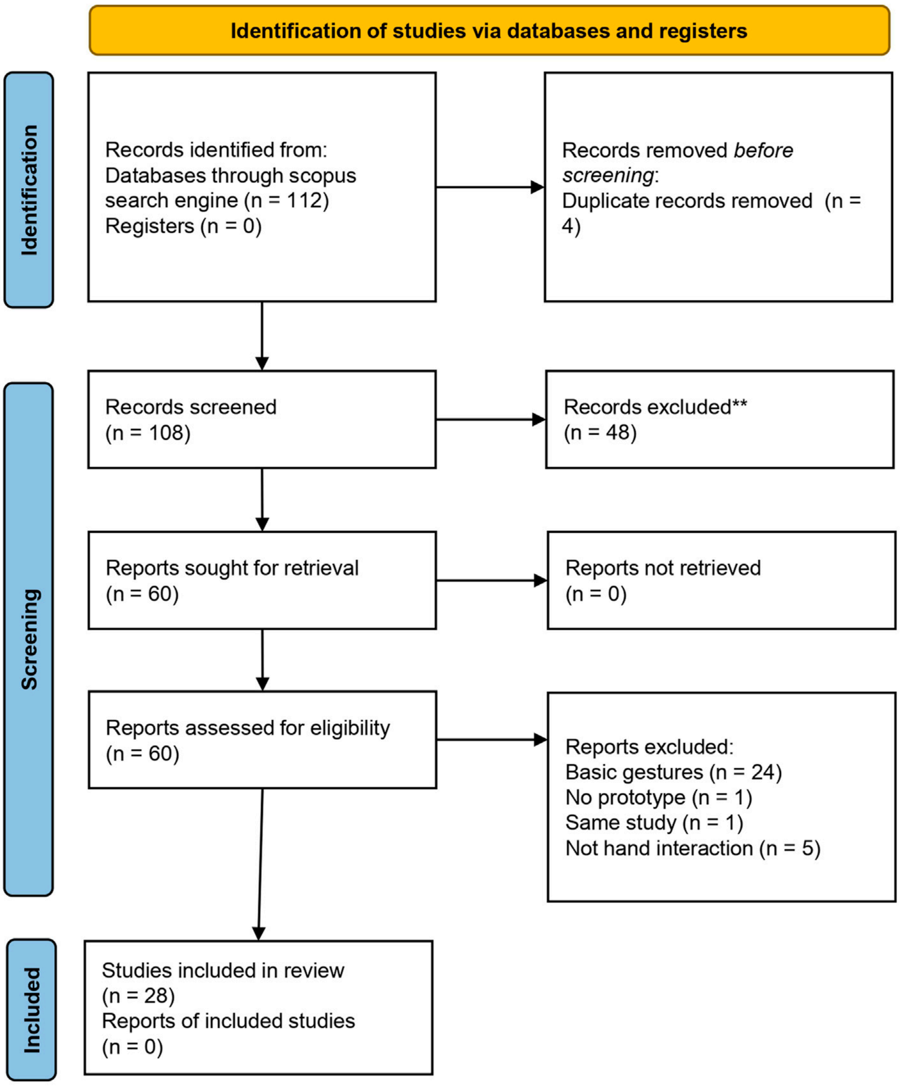
Identification of studies via databases and registers
Identification
Screening
Included
Records identified from:
Databases through scopus search engine (n = 112)
Registers (n = 0)
Records removed *before screening*:
Duplicate records removed (n = 4)
Records screened
(n = 108)
Records excluded**
(n = 48)
Reports sought for retrieval
(n = 60)
Reports not retrieved
(n = 0)
Reports assessed for eligibility
(n = 60)
Reports excluded:
Basic gestures (n = 24)
No prototype (n = 1)
Same study (n = 1)
Not hand interaction (n = 5)
Studies included in review
(n = 28)
Reports of included studies
(n = 0)

FIGURE 1
Screening process.

- Adaptors, releasing body tension through unconscious gesture.

The first three are independent from speech and can communicate on their own while the last five are complementing speech language. For the second class, any gesture that affects a spatial component of an object is considered as manipulative. Deictic gestures that are pointing gestures can be considered as both communicative and manipulative as they communicate the object of the focus while also manipulating the direction which is a spatial component. Finally, instruction based classification separates gestures into two categories: prescribed or free-form. The former designates predefined dictionary of gestures that need to be learnt while the latter is non-restrictive. The hand gestures reviewed by Vuletic et al. (2019) were classified in one of those groups or a group formed by a combination of those gesture categories. As mentioned by the authors, contextual classification has the flaw that is tightly tied to the context of speech. According to them, a potential research on a classification that decouples hand gestures from speech could be interesting for both ergotic (gestures for manipulation) and epistemic (gestures for learning from tactile exploration).

Koutsabasis and Vogiatzidakis. (2019) established a review on mid-air interactions with empirical studies such as gesture elicitations and user studies. Koutsabasis and Vogiatzidakis. (2019) note that there are no standard for gestures design and that the design and implementation are targeted for selected users and for a specific context. In terms of classification, Koutsabasis and Vogiatzidakis. (2019) grouped the reviewed interactions in types that describes what the interaction actually do: Targeting, Navigate, Pan, Typing, Rotate, Select, Point, 3D model shaping, Grabbing 3D object, Travel, Zoom, Other. The first three are 2D interactions while the next seven are 3D interactions. There are both 2D and 3D interactions for Zoom and Other. This classification can be considered as a more granular classification of the model of Hand. (1997). Indeed, we can group the subcategories this way.

1. Navigation: Navigate, Zoom, Pan, Travel;
2. Selection: Select, Point, Grabbing 3D object;
3. Manipulation: 3D model shaping;
4. Application control: Typing.

3 Methodology

3.1 Systematic review

For this review, we did a scoping review following the PRISMA filtering method (Page et al., 2021) as described in this section and as shown in Figure 1. The research questions we will answer in this review are.

- RQ1: What hand interactions have been designed and how to classify them?
- RQ2: What are the apparatus and algorithms for the implementation of those interactions?
- RQ3: What impact did the availability of commercial mixed reality headset have on hand interaction design and prototyping?

To gather the articles, we did one request on the database scopus with the filters.

- ("augmented reality" OR "mixed reality") AND ("HMD" OR "head-mounted display" OR "head mounted display" OR "helmet mounted display" OR "helmet-mounted display" OR "HoloLens" OR "egocentric" OR "glass" OR "headset"), contextualizing the request on the subject of mixed and augmented reality;
- AND ("hand gesture" OR "hand interaction" OR "hand manipulation" OR "hand grasp" OR "mid-air interactions"), restricting the articles to hand interactions.

In the first part of the filters, we decided to include augmented reality as the border between mixed reality (MR) and augmented reality (AR) is mobile and Milgram Virtuality-Reality Continuum (Milgram et al., 1995) describes AR being part of MR. In this context, we consider experiences in which the user is mainly in the real world and where virtual contents are added and interacts with the user and the physical world. The second part of the filters is targeting hand interactions. We also included egocentric point of view research because work on hand interactions, that are using sensors in an egocentric point of view, can be adapted for commercial mixed reality headsets as they are endowed with similar sensors. The scope of the research has been limited to 2017 to 2021 because the goal of this review is to explore the recent research that bloomed with the release of commercial products in 2017. The request was last updated on the 23 October 2022 and resulted in 112 articles of which we removed four duplicates. The Figure 2 illustrates the year distribution of the articles.

3.1.1 First screening

To filter the articles, we started with the screening on titles and abstracts using three inclusion criteria of which at least one must be met.

- the article describes an interaction technique:
- the article describes a technical solution to support interactions;
- the article is comparing different interaction techniques.

Furthermore, we also defined three exclusion criteria which invalidate all articles that meet one or several of them.

- the articles describe a technical solution for low-cost AR/MR such as Google cardboard;
- the article describes human-robot interactions;
- the article describes multi-user collaboration applications.

The first exclusion criteria is more precisely aiming at filtering articles that focuses more on algorithms that tackle the computing and streaming cost issues on cheaper HMD device rather than the interactions themselves which are the focus of this review. 50% of the articles describe an interaction technique, 24% of them describe a technical solution to implement interactions and 12% of them compare different techniques.

This first screening resulted 60 articles that were kept for this scoping review.

3.1.2 Second screening

For the second screening made on the content of the articles, the exclusion criteria are.

- the article describe a use case for natively supported interactions on commercial devices;
- the article does not describe any prototype.

The first criterion allows us to keep articles focused on interaction techniques and implementations instead of use cases. This criterion filtered 24 articles which is the biggest part of the articles that were rejected in this second screening. We then removed one article that described a concept of architecture without prototype. We also removed five articles that did not describe any hand interaction. Finally, we kept the most recent version of the work of among Bautista et al. (2018) and Bautista et al. (2020) which resulted in 28 articles in total or 26% of the total articles from the initial request.

4 Analysis

4.1 Extending frutos and macaranas taxonomy

In this litterature review, we came across two articles from Frutos-Pascual et al. (2019) and Serrano et al. (2022) that uses a taxonomy introduced by Macaranas in order to compare two popular hand interaction modality in commercial mixed reality HMD.

4.1.1 Macaranas taxonomy

Macaranas et al. (2015) introduced this taxonomy to classify strategies to make intuitive interactions using as criteria the type of mental scheme used to learn them. The three classes are.

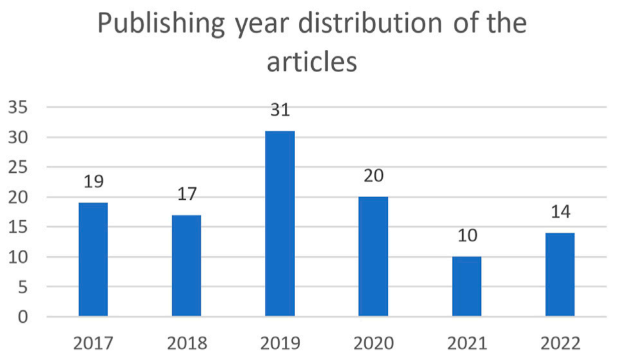

FIGURE 2
Year distribution of the articles reviewed.

- Metaphoric mapping, which are based on the images that link repeated outcomes from everyday iteractions to conceptual metaphors. An example given by the author is the fact that the height of a pile can be associated to the concept of high quantity. As such, interactions that are based on going up and down to increase or decrease an output are using this image and are hence metaphoric mappings;
- Isomorphic mapping, which are based on the one-to-one spatial relation between the user input and the system output. The main focus of this mapping is the correlation between the spatial movement of the input and the effect produced. The output can be physical or abstract. The authors give the example of a User Interface (UI) element made of empty ticks horizontally lined up that is mapped to the sound volume. The spatial movement on the horizontal line fills the ticks which are isomorphically mapped to the sound volume as each tick represent an quantity of volume;
- Conventional mappings, which are based on the interactions adapted from previous interfaces the user has used. The authors underline that they exclude in this mapping the interactions grounded on image schema-based metaphors and one-to-one mappings so as to differentiate from the two formers classes. The authors illustrate in their third figure using the rotation direction convention learned through the reading of a clock and the usage of a screw to learn how to use a control knob to increase the sound volume.

From our understanding of Macaranas et al. (2015)'s model, all three strategies to make an interface intuitive are not mutually exclusive. Indeed, the rotation of the knob is mapped to a volume quantity and thus can also be considered as an isomorphic mapping. On top of that, the rotation in the correct direction defined across those three objects is associated to the concept of an increase (the increase of time for the clock, and the increase in the progression to achieve the task of closing the jar or screwing for the jar cap and the screw respectively) can also be considered as a metaphoric mapping.

4.1.2 Application of macaranas taxonomy on hand interactions for commercial HMD

Frutos-Pascual et al. (2019) and Serrano et al. (2022) are applying Macaranas et al. (2015) model for strategies to make interactive system intuitive to compare HoloLens 1 and respectively Meta 2 and Magic Leap One interactions. Both authors designate HoloLens 1 interactions as Metaphoric Mappings as opposed to Meta 2 and Magic Leap One, being Isomorphic Mappings. Meta 2 and Magic Leap One interactions are considered as isomorphic mappings because when an object is grabbed using Meta 2 headset, the object position becomes linked to the hand position by means of a virtual representation of the hand. The hand tracking interaction creates a one-to-one spatial mapping between the hand joints movement and the feedback from the simulated physics in the virtual environment. On the opposite side, HoloLens 1 is considered by both authors as using a metaphoric mapping because it reminds of the mouse clicks on desktop computer. However, in our opinion, it can be argued that the whole interaction of HoloLens 1 relies on not the image of a mouse but the cursor-based interaction from the WIMP paradigm and thus is a conventional mapping. Besides, the mouse cursor image is not really associated to any abstract concept as described in the metaphorical association image scheme in Macaranas et al. (2015) paper. The research of Frutos-Pascual et al. (2019) and Serrano et al. (2022) highlight the need for a classification model for hand gestures unbound by speech as suggested by the review of Vuletic et al. (2019) because they are using a model designed to classify learning strategies and not hand interactions. Moreover, we have not found other work using this Macaranas et al. (2015) to classify hand interactions which imply that it is not standardized yet.

4.2 Our extension for the specific case of hand interactions

The model we propose to classify interactions only describes the way the user is performing an interaction and the way users are learning how to use them. If we need to describe the use case of the interactions in our review as defined in the classification of Koutsabasis and Vogiatzidakis. (2019), we keep the model of Hand. (1997) which is more general. As the model classifies the interactions using two criterias, we propose two axis.

- The Interaction Behaviour axis defined by Isomorphic and Gestural in the extremes;
- The Metaphor axis defined by Conventional and Realistic in the extremes.

In the first axis, Isomorphic refers to the same definition proposed by Macaranas et al. (2015) as a one-to-one spatial mapping between the user hand movement and the effect output in mixed reality but adding the idea that it is unprescribed. Gestural as opposed to that is encompassing the association of a motion to a punctual predefined effect and the extreme being pose based interaction where the user only needs to shape his hand in a specific shape to trigger an action. For example, of HoloLens 1 head pointing and air tap selection based interaction is in the gestural side of the spectrum because it is mainly relying on the air tap gesture as a predefined trigger of the action. As opposed to that, Meta 2 and Magic Leap One hand tracking manipulation are more isomorphic as they rely on the direct contact between the hand virtual representation and the mixed reality content. The better is the

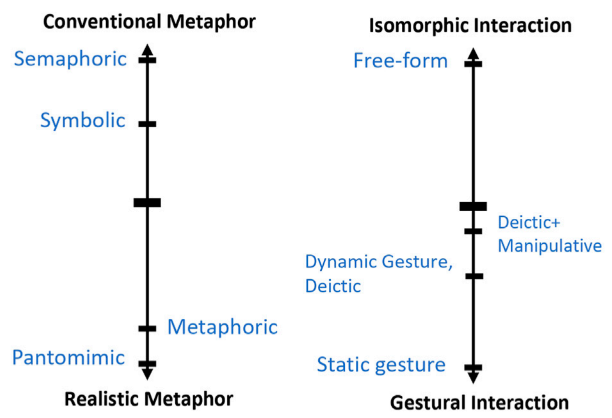

FIGURE 3
Independent communicative gesture categories placed in the proposed taxonomy.

isomorphy between the real hand and the virtual hand serving as an effector, the better the system react to the user's input. It can be noted though that categorizing the headsets globally is not totally accurate as it depends on the interactions used. For example, on Meta 2 and Magic Leap One, the user can grab an object. The grab which triggers the selection of the object is a gestural interaction but the manipulation of the object is isomorphic. To compare with Vuletic et al. (2019)'s gesture classifications described in 2.2, the extreme isomorphy would correspond to free-form gestures while the extreme gestural would be static gestures. 1) Deictic and Dynamic Gestures would be in the gestural side of the continuum but closer to the limit as they can map a motion of the user to a continuous output.

In the second axis, metaphor refers to the mental image that helps the user learn to execute the interaction. We use as a definition of Conventional Metaphor a similar definition to Macaranas et al. (2015)'s Conventional Mapping but broaden to include all mental models defined for a specific context and that are learned through repetitions. As opposed to that, a Realistic Metaphor is based on the user's everyday experience and repeated patterns in the real world. For example, if the user is mimicking the shape of binoculars to zoom on a content, the metaphor is realistic. Oppositely, if he uses a two-fingers pinch, referring to the touchscreen interaction counterpart, the metaphor is conventional. We defined this classification as a continuum because a gesture can be more or less conventional compare to others. To explain this aspect, we will use the comparison to Vuletic et al. (2019) model. In Figure 3 We have placed the gestures subcategories that can be used independently from speech language in the Metaphor axis. We also placed metaphoric gestures even if Vuletic et al. (2019) originally included them in speech related gestures because the definition is close to what Macaranas et al. (2015) designate as Metaphoric Mapping. The most realistic gestures would be Pantomimic gestures because they represent exactly what a user would be doing in reality. The most conventional gestures would be Semaphoric gestures because they have no semantic background and are learned solely by repetitions from zero for the context of usage. In between, we have Symbolic gestures that have an acknowledge meaning in the conventional side. In the realistic side, we have Metaphoric gestures that are tied to a similarity between a pattern in the real world and an abstract concept. Furthermore, a Symbolic gesture is actually a Semaphoric gesture that has been culturally ingrained in the gesture language by the repetition usage and of which the meaning has been commonly accepted. As such, in our model, a conventional interaction will shift toward the realistic ends of the continuum as it is becoming a standard in the society. To put it more simply, as an interaction designed for a specific system is democratized at the scale of the society, it becomes a daily interaction that people can use as a reference mental model when learning new interactions. As a result, the placement of an interaction on the Metaphor continuum can evolve over time. This the reason why we would position HoloLens 1 air-tap based interaction on the conventional side of the metaphor spectrum but close to the frontier with realistic metaphors.

Our taxonomy is also illustrated in Figure 4 with the example of interactions aiming at navigating in a virtual text content for each quadrants. In the top left corner of the graph, the Isomorphic Interaction based on Conventional Metaphor is represented by the usage of a scroll bar widget. The user is able to grab the handle of the widget which creates the one-to-one spatial relation between the widget and his hand. The user is also familiar with similar widgets that can be found on computers and mobile phones. In top right corner, the widget can be replaced by a mid-air scrolling gesture which consists in repeating a vertical translation down or up. One translation is registered as a gesture and translated to the movement of the content. The metaphor is conventional as it is a recurrent gesture for touchscreen interfaces. In the bottom left corner, the user is manipulating a virtual book. The pages of the book have a direct spatial relation with the finger representations of the user, and the metaphor is correlated to the manipulation of a real book. Finally, in the bottom right corner, the user is doing the mimic of turning the pages of a book. The gesture of turning one page is recognized and translated to the browsing of the content. The metaphor is the same as the precedent interaction but the behavior is gestural because only the punctual movement is registered and interpreted by the system.

This model we propose do not aim at being exhaustive in the classification of the interactions but to give a more global idea of what type of interaction have been designed with a simple description of the learning cue and the relation between the hand movement and the output.

4.3 Analysis of the reviewed paper using the metaphor-behavior taxonomy

The Figure 5 together with Tables 1, 2 summarize the classification of the selected articles. In the screened articles, the distribution of the articles in the taxonomy is illustrated in Figure 6. We can observe that in terms of metaphor used, the distribution is almost even. When it comes to the behaviour of the interaction, there are more gestural interactions than isomorphic ones. For both axis, there are papers that designs and evaluates interactions in both side of the spectrum. For example, gesture authoring tools like Mo et al. (2021) paper can be used to design interactions based on both types of metaphors. We decided to not classify the article of Mueller et al. (2017) as they propose a hand tracking technical solution for

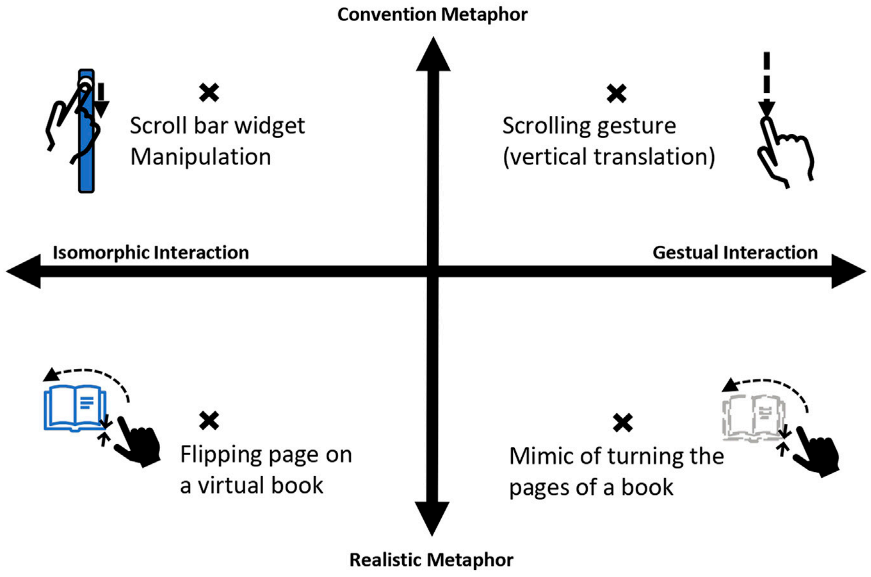
Convention Metaphor
Scroll bar widget Manipulation
Scrolling gesture (vertical translation)
Isomorphic Interaction
Gestual Interaction
Flipping page on a virtual book
Mimic of turning the pages of a book
Realistic Metaphor

FIGURE 4
Illustration of the Taxonomy of interactions by behaviour and metaphor (The position relative to the extremity of the axis is not relevant in this figure).

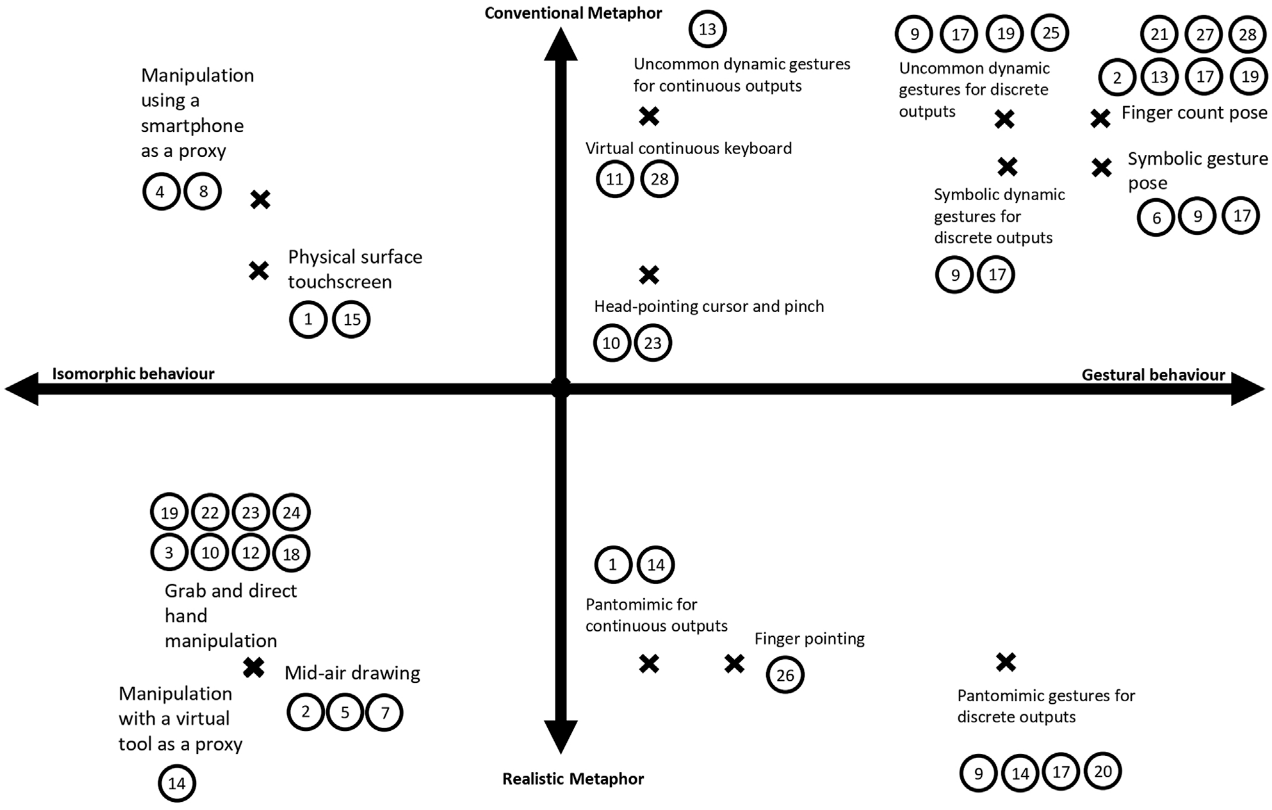
Conventional Metaphor
Manipulation using a smartphone as a proxy
4 8
Physical surface touchscreen
1 15
13
Uncommon dynamic gestures for continuous outputs
Virtual continuous keyboard
11 28
Head-pointing cursor and pinch
10 23
9 17 19 25
Uncommon dynamic gestures for discrete outputs
Symbolic dynamic gestures for discrete outputs
9 17
21 27 28
2 13 17 19
Finger count pose
Symbolic gesture pose
6 9 17
Isomorphic behaviour
Gestural behaviour
19 22 23 24
3 10 12 18
Grab and direct hand manipulation
Mid-air drawing
2 5 7
Manipulation with a virtual tool as a proxy
14
1 14
Pantomimic for continuous outputs
Finger pointing
26
Pantomimic gestures for discrete outputs
9 14 17 20
Realistic Metaphor

FIGURE 5
Classification of the screened articles. Articles numbered using the tables 1 and 2.

TABLE 1 Table of screened articles (1/2).

Number	Authors	Title
1	Plasson et al. (2020)	3D Tabletop AR: A comparison of mid-air, touch and Touch + Mid-Air interaction
2	Zhang et al. (2020)	ARSketch: Sketch-Based User Interface for Augmented Reality Glasses
3	Ababsa et al. (2020)	Combining hololens and leap-motion for free hand-based 3d interaction in mr environments
4	Lee and Chu. (2018)	Dual-MR: Interaction with mixed reality using smartphones
5	Chang et al. (2017)	Evaluating gesture-based augmented reality annotation
6	Lin and Yamaguchi. (2021)	Evaluation of Operability by Different Gesture Input Patterns for Crack Inspection Work Support System
7	Lu et al. (2019)	FMHash: Deep Hashing of In-Air-Handwriting for User Identification
8	Yu et al. (2017)	Geometry-aware interactive AR authoring using a smartphone in a wearable AR environment
9	Mo et al. (2021)	Gesture Knitter: A Hand Gesture Design Tool for Head-Mounted Mixed Reality Applications
10	Frutos-Pascual et al. (2019)	Head Mounted Display Interaction Evaluation: Manipulating Virtual Objects in Augmented Reality
11	Lee et al. (2019)	HIBEY: Hide the keyboard in augmented reality
12	Jailungka and Charoenseang. (2018)	Intuitive 3D model prototyping with leap motion and microsoft hololens
13	Sun et al. (2019)	MagicHand: Interact with iot devices in augmented reality environment
14	Jang et al. (2017)	Metaphoric Hand Gestures for Orientation-Aware VR Object Manipulation With an Egocentric Viewpoint

TABLE 2 Table of screened articles (2/2).

Number	Authors	Title
15	Xiao et al. (2018)	MRTouch: Adding touch input to head-mounted mixed reality
16	Mueller et al. (2017)	Real-Time Hand Tracking under Occlusion from an Egocentric RGB-D Sensor
17	Caputo et al. (2021)	SHREC 2021: Skeleton-based hand gesture recognition in the wild
18	Kim et al. (2018)	SWAG Demo: Smart Watch Assisted Gesture Interaction for Mixed Reality Head-Mounted Displays
19	Bautista et al. (2020)	Usability test with medical personnel of a hand-gesture control techniques for surgical environment
20	Min et al. (2019)	VPModel: High-Fidelity Product Simulation in a Virtual-Physical Environment
21	Choudhary et al. (2021)	Real-Time Magnification in Augmented Reality
22	Schäfer et al. (2022)	The Gesture Authoring Space: Authoring Customised Hand Gestures for Grasping Virtual Objects in Immersive Virtual Environments
23	Serrano et al. (2022)	An empirical evaluation of two natural hand interaction systems in augmented reality
24	Su et al. (2022)	A Natural Bare-Hand Interaction Method With Augmented Reality for Constraint-Based Virtual Assembly
25	Hu et al. (2018)	3D separable convolutional neural network for dynamic hand gesture recognition
26	Su et al. (2021)	Smart training: Mask R-CNN oriented approach
27	Shrestha et al. (2018)	Computer-Vision based Bare-Hand Augmented Reality Interface for Controlling an AR Object
28	Lee et al. (2022)	Virtual Keyboards with Real-Time and Robust Deep Learning-Based Gesture Recognition

the issue of hand-object occlusion, which could be used to design any type of interaction from the classification.

In this section, we will be answering to RQ1 using the proposed taxonomy.

4.3.1 Comparison between isomorphic interactions and gestural interactions

The article of Frutos-Pascual et al. (2019), on which this taxonomy is built on, is among the articles screened in this review. As mentioned above, Frutos-Pascual et al. (2019) compare the interactions from the headsets HoloLens 1 and Meta 2. Similarly, Serrano et al. (2022) compare HoloLens 1 and Magic Leap One. The authors designate the interactions on HoloLens as Metaphoric Mappings and the interactions on Meta 2 and on Magic Leap as Isomorphic Mappings. However, following our model, HoloLens 1 cursor head-pointing and pinch gesture selection-based interactions are Gestural Interactions using a Conventional Metaphor. Meta 2 and Magic Leap One trajectory

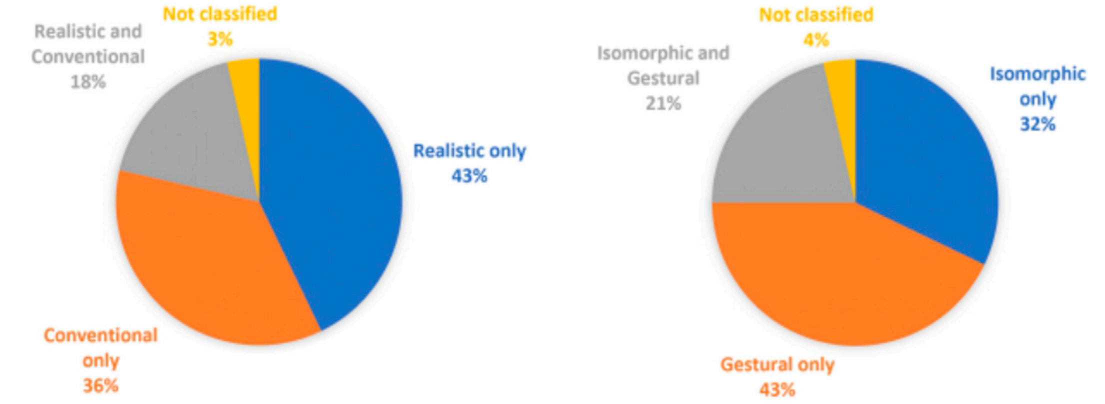

FIGURE 6
Distribution of the classification in the reviewed articles.

and simulated physic-based interactions are Isomorphic Interactions based on Realistic Metaphors. The user studies from Frutos-Pascual et al. (2019) show that Meta 2 interactions are preferred, qualified as more natural and useable, and require less cognitive charge according to the participants. Several other articles from the screening (Jailungka and Charoenseang, 2018; Kim et al., 2018; Ababsa et al., 2020; Bautista et al., 2020) also highlight the good usability of grasping objects. HoloLens 1 interactions only outperform Meta 2 interactions in terms of precision for the scaling which is a manipulation task. On the opposite side, Serrano et al. (2022) show that although Magic Leap One is preferred in the subjective questionnaires, there are no statistically significant differences between HoloLens 1 interactions and Magic Leap One interactions either for the objective data (accuracy, number of mistakes, time completion) and the subjective criteria (usefulness, preference and recommendation).

In our opinion, the studies, instead of comparing the behavior of the interactions, are reflecting the importance of the metaphors used by both interactions. Indeed, Serrano et al. (2022) explicitly use the realism of the interactions to oppose HoloLens 1 and Magic Leap One. In both articles (Frutos-Pascual et al., 2019; Serrano et al., 2022), Meta 2 and Magic Leap One interactions are favored because the realistic metaphor of grasping objects to explore the world is well ingrained in the knowledge of the users. Besides, even if HoloLens 1 interactions are using the conventional metaphor of the WIMP paradigm, the translation of the interface from 2D screen interactions to 3D mid-air interactions requires the user to adjust the spatial perception from the 2D mental image. According to Frutos-Pascual et al. (2019), the high performance of HoloLens 1 interactions for the scaling might be due to the similarity with the interactions learned on desktop computers and the lack of realistic metaphors for this kind of manipulation task. To go back to the behavior of the interaction, the cursor pointing is gestural because of the selection trigger gesture but also has a part of isomorphic behavior since the user is interacting with 3D handles spatially constrained to the head cursor when selected. The reason why HoloLens 1 interactions are better than Meta 2 for this specific task can be related to the fact that a common practice in human computer interaction research to increase the precision of an interaction is to change the mapping of the movement. Meta 2 has, by nature of the default interaction, a one-to-one mapping while the HoloLens 1 mapping can be changed for a different scaling to meet the precision required for the task. Serrano et al. (2022) also support the fact that nonrealistic interactions can work better as, according to them a realistic interaction that is not realistic enough can have the similar problem of the uncanny valley (McMahan et al., 2016) found in the field of robotics.

To finish with the comparison between isomorphic and gestural interactions, the article of Bautista et al. (2020) compares Meta 1 interactions and the gesture interface from the Myo armband. The authors use a similar differentiation as our taxonomy by qualifying Meta interactions as Manipulation Control, which corresponds to our isomorphic interactions and Myo armband interface as Gesture Control, which corresponds to our gestural interaction. The comparison of the two interaction techniques in user studies is done through an evaluation of the task of the control of the application. The study shows that compared to Meta 1, Myo armband is more comfortable, has less errors, requires less help from the researcher to the participants in order to complete the task and has a better completion time. However, the article does not detail the steps executed by the participants which makes it hard to understand what kind of gestures were used and to analyze the reason why they are better performing. Furthermore, Myo armband might be preferred because of the uptime and accuracy of the recognition of the interactions. Meta 2 is limited to sensors in an egocentric point of view which has a very limited volume of tracking in front of the user as opposed to Myo armband which detects gestures using electromyography data which works as long as the user is wearing the armband. This factor can significantly impact the number of errors, the time of completion and the perceived comfort.

4.3.2 Interactions mainly isomorphic

On the isomorphic side of the graph in Figure 5, there are five groups.

1. Grab and direct hand manipulation (Jailungka and Charoenseang, 2018; Kim et al., 2018; Frutos-Pascual et al., 2019; Ababsa et al., 2020; Bautista et al., 2020; Schäfer et al., 2022; Serrano et al., 2022; Su et al., 2022);
2. Manipulation with a virtual tool as a proxy (Jang et al., 2017);
3. Mid-air drawing (Chang et al., 2017; Zhang et al., 2020);
4. Physical surface touchscreen (Xiao et al., 2018; Plasson et al., 2020);
5. Manipulation using a smartphone as a proxy (Yu et al., 2017; Lee and Chu, 2018);

The first three groups are using Realistic Metaphors with a one-to-one spatial relation between the movement of the hand and the feedback of the system. More specifically, the first and second groups allow the user to manipulate virtual content directly using the hands or a virtual tool in contact with the content and are intuitive because of the counterpart interactions in the real world. In the first group, Su et al. (2022) and Schäfer et al. (2022) go further in the implementation of the interaction with a more realistic design of the grab action to select an object. The former considers realistic collisions and physical constraints for the grab and for the assembly of virtual components on virtual and physical objects while the latter proposes an authoring tool that records custom hand grab poses for each virtual object. In Schäfer et al. (2022)'s paper, their user studies show that the custom grab is perceived as more useable than the pinch and the standard closing hand grab. Making the behavior of virtual content realistic through physics models improves the experience as the user is less prone to mistakes and understands more easily how to manipulate the content thanks to his prior experiences. The third group is the reproduction of drawing and writing in the virtual environment. The articles from Zhang et al. (2020) and Chang et al. (2017) enable respectively 2D mid-air sketching and 3D annotations on the physical world. The former emphasizes on the usage of convolutional neural networks (CNN) for gesture recognition and sketch auto-completion. The latter shows the benefit of the common practice of limiting the degree of freedom in interactions. Indeed, in their user studies, 2D mapped drawing were preferred and more precise than simple 3D mid-air drawing. Furthermore, according to Chang et al. (2017), cleaning the visual feedback for the annotations by showing a beautified version makes the drawing process faster. The works of Zhang et al. (2020) and Chang et al. (2017) show that it is possible to go a step further in the design of a more intelligent interaction by analyzing the trajectories of the hand and improving the effect returned by the system. To end with mid-air drawing, Lu et al. (2019) proposes a user identification using hashcode of the mid-air writing signature in mixed reality. As the system uses a CNN, the main problem is the requirement to retrain the network for each additional signature.

Groups 4 and 5 are using conventional Metaphor from the usage of smartphones and tablets. Indeed, in the fourth group, Plasson et al. (2020) and Xiao et al. (2018) are using the spatial understanding of mixed reality to implement an intelligent adaptive user interface by converting physical surfaces into touchscreens. The haptic feedbacks make the interaction more natural. The works of the researchers (Xiao et al., 2018; Plasson et al., 2020) contribute to reduce the gap between the virtual environment and the real environment. Finally, the fifth group takes advantage of the familiarity the users have with using smartphones to interact with digital content as well as the extra sensors of the smartphones to improve the spatial understanding of mixed reality. Lee and Chu. (2018) propose to capture virtual object on the screen of a smartphone. The user is then able to manipulate the content by moving the smartphone or using the touchscreen interactions. Yu et al. (2017) use the smartphone as a pointer to select virtual object as if the user is manipulating a laser. The strong point of their works is the ubiquity of smartphones which makes the interfacing between the user and the mixed reality environment more flexible.

4.3.3 Interactions mainly gestural

On the gestural side of the graph in Figure 5, we identified ten groups.

1. Finger count pose (Shrestha et al., 2018; Sun et al., 2019; Bautista et al., 2020; Zhang et al., 2020; Caputo et al., 2021; Choudhary et al., 2021);
2. Symbolic gesture pose (Caputo et al., 2021; Lin and Yamaguchi, 2021; Mo et al., 2021);
3. Uncommon dynamic gestures for discrete outputs (Hu et al., 2018; Bautista et al., 2020; Caputo et al., 2021; Mo et al., 2021);
4. Symbolic dynamic gestures for discrete outputs (Caputo et al., 2021; Mo et al., 2021)
5. Pantomimic gestures for discrete outputs (Jang et al., 2017; Min et al., 2019; Caputo et al., 2021; Mo et al., 2021);
6. Finger-pointing (Su et al., 2021);
7. Head-pointing cursor and pinch (Frutos-Pascual et al., 2019; Serrano et al., 2022);
8. Virtual continuous keyboard (Lee et al., 2019; Lee et al., 2022);
9. Uncommon dynamic gestures for continuous outputs (Sun et al., 2019);
10. Pantomimic for continuous outputs (Jang et al., 2017; Plasson et al., 2020);

The first two groups described gestural-only interactions and differ slightly in the metaphor used to learn them even if they can be categorized as conventional. More specifically, in the both groups, users are learning to do a specific pose, which is static. However, the first group is what Vuletic et al. (2019) call as semaphoric gestures since the gestures have no semantic background, which is the extreme of the conventional side of the spectrum. Indeed the pose are based on finger extension. Depending on the number of fingers extended, a different functionality is mapped to the pose. The main advantage of this technique is the ease of detection by computer vision algorithms as the different poses are distinguishable. Oppositely, the second group uses symbolic metaphors which means that the poses have cultural semantic values. For example, the "ok" gesture found in Lin and Yamaguchi (2021) and Caputo et al. (2021) works, which consists in making a circle using the index and the thumb, is recognized in both articles and has commonly the cultural meaning of an agreement in English-speaking countries. This type of metaphor is a double-edged sword as it can make the interaction more intuitive for a set of population but can also be confusing or error inducing for other populations where the meaning is different. Indeed, the "ok" gesture is actually offensive in Brazil and got Richard Nixon booed by the crowd at Rio de

Janeiro in the 1950s, while in Japan, it represents money (Reuters, 1996). It can be noted that Mo et al. (2021) and Caputo et al. (2021)'s paper are classified in all the group described above as they are respectively a modular architecture for gesture design and a hand gesture recognition competition result summary. As such, they support gestures in all of the mentioned categories.

Groups 3, 4 and 5 which cover gestural interactions that are dynamic, are slightly more isomorphic than the first two groups as illustrated in our comparison with Vuletic et al. (2019)'s model in Figure 3. The difference between the three groups is again the learning metaphor. In terms of metaphor, the groups 3 and 4 have a similar relationship between each other as the one between groups 1 and 2 as they are both using conventional metaphor but are more specically using semaphoric and symbolic gestures respectively. We named the group 3 as "uncommon gestures" because the articles describe a variety of unrelated gestures such as doing a circle motion (Shrestha et al., 2018; Mo et al., 2021). The group 5 however is on the opposite of the spectrum in terms of metaphor as it is using pantomimic gestures which are at the extreme of the realistic side of the metaphor spectrum. In this group, Jang et al. (2017) proposes to invoke virtual objects by reproducing hand grasp poses as if the real object was in hand, while Min et al. (2019) augment an inert 3D printed camera prototype with visual feedbacks to the mimick of the usage of a real camera. With the definition given by Vuletic et al. (2019), a pantomimic is by nature dynamic as it is a sequence of actions.

Going farther to the left on the isomorphic-gestural spectrum, we have the group 6 that is composed of a single article from Su et al. (2021). The authors designed a deictic gesture to select a physical object in mixed reality. This gesture is as realistic as pantomimic gestures because it is using the exact motion that we use in reality to point at objects. It is a more realistic way to select an object compare to the group 7 which designate head-pointing cursor based interactions found in HoloLens 1. For the latter, a real-time spatial mapping between the user's head movement and a virtual content can be established after the user selects an item by doing the selection gesture "pinch" which consists in touching the index with the thumb to simulate a mouse "click". As mentioned in 4.2, HoloLens 1 interactions while being a conventional metaphor is placed close to the frontier with realistic metaphor because it is based on the WIMP paradigm that have become part of our daily lives across multiple devices from desktop computers to mobile devices. The reason why group 6 is more gestural than group 7 is that the finger point described in Su et al. (2021)'s work has a discrete output as its task is only the selection.

The last three groups have the same position as group 7 in the behaviour spectrum. The group 8 and 9 in particular have the exact same position in the metaphor spectrum as well because they are using semaphoric gestures. Even though the group 8 design a character input interaction based on keyboards, the trigger gestures do not have any semantic values. Lee et al. (2019) design a symbol typing interaction where the position of the hand during the movement allows the user to select a letter, to type in a letter, to delete a letter and to select a word from the auto-completion system. The gesture is totally different from the writing or the typing activity in the physical world. Similarly, Lee et al. (2022) use a specific gesture that uses the number of extended fingers as well as a thumb motion to trigger the typing. In the group 9, the only article is the work from Sun et al. (2019) and it describes a dynamic gesture which is triggered by a finger extension based pose and then map movement of the hand to a continuous value on a smart connected object. An example given in the article is the control of the sound volume with a horizontal translation with three fingers extended. Finally, the group 10 encompasses dynamic pantomimic gestures that are used to control continuous output. Plasson et al. (2020) use the metaphor of pulling the string of a floating helium balloon. The distance between the hand and the base of a virtual quantified stack regulates the value of the output on the stack as if it was the height of a floating balloon. Jang et al. (2017) article is also in this group because the user can use the invoked tools to manipulate virtual objects using the properties and functionalities of the tool invoked.

4.4 Algorithms and devices

In this section, we will answer to RQ2 and RQ3.

4.4.1 Devices

In terms of devices, Figure 7 summarizes the different devices used in the reviewed articles. Thirteen articles implemented their prototypes on HoloLens 1. 50% of those articles use external devices and/or sensors with a HMD to improve capture of the interactions. This trend is due to the limitations of the first version of HoloLens which only allowed limited head pointer cursor-based interactions and voice commands. As mentioned previously, the article from Lee and Chu. (2018) and Yu et al. (2017) use smartphones and their embedded sensors to complement mixed reality interactions. The other articles improve the hand interactions with a better hand tracking using external sensors such as depth cameras, with LeapMotion being the most popular, and Motion capture systems, such as OptiTrack (Figure 7). The articles based on the Rokid Glass and the Oculus Rift headsets also use depth cameras to support hand interactions. On the opposite, the article on Meta, HoloLens 2 and Magic Leap One do not use any external sensors because of the native support of real-time hand tracking. There are six articles that are presenting research that are aimed at MR HMD but have yet to test on actual devices.

4.4.2 Algorithms

In this review, ten articles are describing the algorithms used to implement their interaction techniques. The remaining articles mostly implement interactions build on natively supported interactions and hand tracking from the headset and sensors after calibration. Following the trend of CNN, four articles are based on this solution. In this context CNNs are used to.

- track the hand joints (Mueller et al., 2017; Zhang et al., 2020);
- detect gestures (Hu et al., 2018; Sun et al., 2019; Zhang et al., 2020; Caputo et al., 2021; Su et al., 2021; Lee et al., 2022);
- process trajectories (Lu et al., 2019; Zhang et al., 2020).

More specifically, Mueller et al. (2017) propose two CNNs that uses both RGB images and depth data to compute hand joint

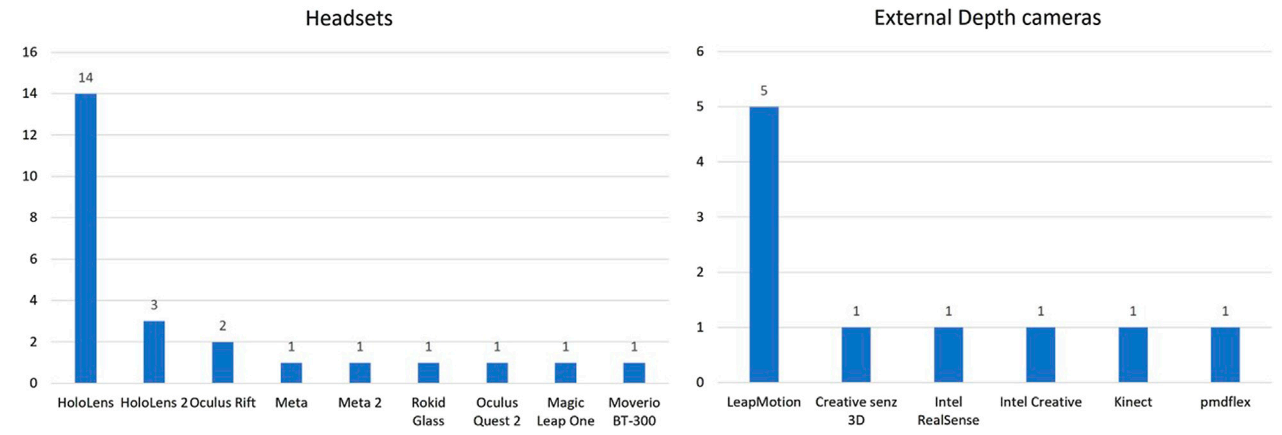

FIGURE 7
Distribution of devices in the reviewed articles.

trajectories in hand-object occlusion situations by localizing the hand center and regressing the joints relative to this center.

Concerning gestures, Mo et al. (2021), Jang et al. (2017), Caputo et al. (2021), Min et al. (2019), Schäfer et al. (2022) and Shrestha et al. (2018) present alternative solutions to CNN. Firstly, Mo et al. (2021) represent gestures using Hidden Markov Model (HMM). In their solution Gesture Knitter, each gesture is split into two components. The first one called the gross gesture component is a primitive gesture that describes the movement of the hand palm. It is represented using a 12 layers HMM. The second component is the fine gesture which describes the movement of the fingers in the palm center referential and is represented using a 8 layers HMM. To infer the gesture, the authors use the Viterbi algorithm to determine the most likely sequence of hidden layers. Secondly, Jang et al. (2017) use an innovative method which consists in describing hand pose with a voxel grid where each voxel is activated if enough 3D cloud points from the hand are in that cell. The encoding of a voxel plan and the change of the encoding over time represent a pattern that is used to discriminate the gestures in random decision trees. Finally, the article of Caputo et al. (2021) summarizes the result of a hand gesture recognition competition in 2021, based on hand joints data from LeapMotion. Several solutions presented are based on popular algorithms such as Transformers or Recurrent Neural Network (RNN). As opposed to traditional CNN, those solutions are adapted to spatial-temporal data which are mandatory to represent a gesture. Caputo et al. (2021) highlight that Long Short Term Memory algorithms, a popular variant of RNN, have not been explored in the competition even if they have significant good results in the literature. The winning solution of the competition is a modified version of the CNN called the spatial-temporal CNN. The hand movement is represented using a spatial-temporal graph where the joints of the hands are nodes. In the graph, the nodes are.

- linked according to the structure of the hand skeletons, which encapsulates the spatial information;
- linked if they represent the same nodes in consecutives images of the movement, which encapsulates the temporal information.

Each graph representing the movement on a time window is sent to a spatial-temporal CNN to localize and detect a gesture. Then, for the classification, each movement is represented with a gradient histogram which is compared using cosine similarity to a set of histograms from known gestures. The article from Min et al. (2019) also uses gradient histograms but classifies with a Support Vector Machine (SVM) algorithm. Schäfer et al. (2022) use a more basic solution to recognize the custom grab gesture recorded on their virtual object. If the hand of the user is close to a virtual object, a frame of the hand joints is compared to the existing gesture data using euclidian distance. Similarly, Shrestha et al. (2018) also use a very simple gesture classifier by finding calculating the number of finger tip outside of a circle centered around the palm to establish how many fingers are extended.

The last article from Xiao et al. (2018) uses short-range depth and infrared data to create touch surfaces in the physical world. Their in-house algorithm DIRECT presented in their previous work (Xiao et al., 2016) detects physical planes and the contacts between the fingers and those plans.

Finally, in terms of computing solutions, most of the papers use CPU and GPU on personal computers except for the works of Zhang et al. (2020), Xiao et al. (2018) and Min et al. (2019) where the CPU embedded on the headset (HoloLens 1) or glass (Rokid) is used. Network solutions to distribute the computing power have not been proposed in the reviewed articles.

5 Summary and discussion

The result of this review gave insight on the behavior of the interaction designed in the academic since commercial MR headsets have been made available for research and development and on their technical implementations. The Tables 3, 4 gives an overview of the articles analyzed in this paper.

TABLE 3 Summary of the article reviewed (1/2).

Article	Interaction behavior	Interaction metaphor	Headset(s)	External sensor(s)	Key Algorithm(s)
Plasson et al. (2020)	Isomorphic	Realistic	HoloLens 1	OptiTrack	
Zhang et al. (2020)	Isomorphic, Metaphoric	Realistic	Rokid Glass	pmd flexx	CNN
Ababsa et al. (2020)	Isomorphic	Realistic	HoloLens 1	LeapMotion	
Lee and Chu. (2018)	Isomorphic	Conventional	HoloLens 1	iPhone	
Chang et al. (2017)	Isomorphic	Realistic	HoloLens 1		
Lin and Yamaguchi. (2021)	Gestural	Conventional	HoloLens 2		
Lu et al. (2019)	Isomorphic	Realistic		LeapMotion	CNN
Yu et al. (2017)	Isomorphic	Conventional	HoloLens 1	Smartphone, Smartwatch	
Mo et al. (2021)	Gestural	Realistic, Conventional	HoloLens 2		multivariable HMM, Virterbi algorithm
Frutos-Pascual et al. (2019)	Isomorphic, Gestural	Realistic, Conventional	HoloLens 1, Meta 2		
Lee et al. (2019)	Gestural	Conventional	HoloLens 1		
Jailungka and Charoenseang. (2018)	Isomorphic	Realistic	HoloLens 1	LeapMotion	
Sun et al. (2019)	Gestural	Conventional	HoloLens 1	Depth camera	CNN
Jang et al. (2017)	Isomorphic, Gestural	Realistic	Oculus Rift	Intel Realsense SR3000	Voxel encoding, Random forest
Xiao et al. (2018)	Isomorphic	Conventional	HoloLens 1	Kinect	DIRECT
Mueller et al. (2017)				Intel RealSense SR300	CNN
Caputo et al. (2021)	Gestural	Realistic, Conventional		LeapMotion	Transformers, CNN, GRU

TABLE 4 Summary of the article reviewed (2/2), (* only used HoloLens 1 for making the dataset).

Article	Interaction behavior	Interaction metaphor	Headset(s)	External sensor(s)	Key Algorithm(s)
Kim et al. (2018)	Isomorphic	Realistic	Oculus Rift Development Kit 2	Ovrvision, Creative Senz3D, Samsung Gear Live	
Bautista et al. (2020)	Isomorphic, Gestural	Realistic, Conventional	Meta 1	Myo armband	
Min et al. (2019)	Gestural	Realistic	HoloLens	LeapMotion	SVM
Choudhary et al. (2021)	Gestural	Conventional	HoloLens 2	Logitech 4K Brio HDR	
Schäfer et al. (2022)	Gestural	Realistic	Meta (Oculus) Quest 2		
Serrano et al. (2022)	Isomorphic, Gestural	Realistic, Conventional	HoloLens 1, Magic Leap One		
Su et al. (2022)	Isomorphic	Realistic	HoloLens	LeapMotion	
Hu et al. (2018)	Gestural	Conventional	(HoloLens 1)*		Frame Diff, CNN
Su et al. (2021)	Gestural	Realistic	Moverio BT-300		CNN
Lee et al. (2022)	Gestural	Conventional		Webcam	CNN
Shrestha et al. (2018)	Gestural	Conventional		Webcam	YCB segmentation, k-curvature angle treshold

5.1 Interaction behaviors

When it comes to comparing the behaviors of the interactions, isomorphic interactions are perceived as more natural and intuitive because the one-to-one spatial relation between the manipulated content and the body movement is similar to the daily interactions the user has with objects in the real-world. The other advantage of isomorphic interaction is the diversity of the possible designs. Indeed, a variety of virtual proxy objects, inspired by the behavior of real world objects, can be created to complement the interactions. However, this type of interaction is more prone to close contact interactions and still requires the use of a 3D version of the WIMP paradigm to trigger intangible basic application control functionalities. On the opposite side, gestural interactions have the advantage of empowering the user with interactions beyond what the physical world can offer with mid-air distance interactions. Besides, gestural interactions do not require visual cues to function which reduces the cognitive load of the user. Nevertheless, by design, this type of interaction requires a limited vocabulary that translates gestures into output in the system which makes the system inflexible.

5.2 Learning metaphors

For both types of interactions, the learning curve and the perception of naturalness depend on the metaphor used for the mechanism of the interaction. In general, realistic metaphors make the interaction more intuitive as it is more relatable to the user. However, Serrano et al. (2022) warn that insufficiently realistic metaphors can break the naturalness similar to the uncanny valley phenomenon in the field of robotics. Conventional metaphors are often used because of the ease of recognition by algorithms and for some specific contexts to make the interaction more efficient for the aimed task. As shown by Frutos-Pascual et al. (2019), some specific tasks, such as the scaling, require more precision which can come at the cost of naturalness. This opposition between usability and efficiency reminds us of the notion of Flexibility and efficiency of use from Nielsen Usability Heuristics (Nielsen, 2022) that were initially devised for web design. This notion promotes the design of advanced shortcuts hidden from novice users that can speed up the interaction for experts. In our situation, the interaction itself is not necessarily a shortcut but a more efficient way to manipulate an object that is harder to learn for a new adopter of mixed and augmented reality. Thus, we believe that the design of hand interactions should factor the targeted users and the context of use to weight usability against efficiency. The steep learning curve when the user first encounter the conventional metaphor implicates that the interaction must justify a significant efficiency and usefulness to use this kind of metaphors. Factoring the learning need is even more important for gestural interactions with conventional metaphors because each gesture must be learned individually. In practice, a mixed reality application might need a set of natural interactions to facilitate the adoption of the medium and a more advanced set of interactions that are akin to shortcuts on desktop and mobile computers. In addition, long-term usages also need to be evaluated as muscle fatigue and health impacts are important factors in the ergonomy an user interface. These factors have not been explored in this review.

5.3 Apparatus

Before the review, in relation to our RQ3, we expected that the availability of commercial headsets would make prototyping easier with the native support of hand tracking and the progress of sensors equipped on newly released product. However, as highlighted, a for older headset like HoloLens 1, only simple gestures like air tap were supported which made researchers use external sensors to be able to design more complexe gestures. Furthermore, we also expected more use of RGB-D data in research using more modern HMD that has such sensors like Magic Leap One, Meta 2 and HoloLens 2. For example, the work of Mueller et al. (2017) could be applied to support hand-object occlusion providing that RGB-D pairs could be streamed. However, the articles that uses those HMD are limited to the usage of the hand joints data from the native hand tracking. This can be due to the fact that it is difficult to have access to raw sensor data for commercial devices. HoloLens 2 for example, has the API HoloLens2ForCV Ungureanu et al. (2020) to have access to all sensor data and more recently a publication provided a streaming architecture for HoloLens 2 sensor data with code sample Dibene and Dunn. (2022) to ease the process. However, a known issue is the difficulty to retrieve a matching pair of the RGB camera and the real time depth camera for hand tracking which is a requirement for RGB-D computer vision algorithms. This observation highlights the need to have more research and development friendly MR headsets that has a simple access and streaming to common data used in most of state of the art algorithms.

5.4 Trends

During our screening, we have observed that a significant amount of papers were proposing solutions to add hand tracking to low cost AR devices. This trend shows that there is an interest to make AR more affordable as indeed the MR HMD used in the reviewed articles cost more than a thousand dollars. A practical use case for a low computing cost RGB only hand tracking and hand gesture recognition is the support of hand interactions for Google Cardboard, for example,.

In the industry, commercial devices are trending toward supporting isomorphic interactions. Many examples can be not only found in mixed and augmented reality such as Magic Leap One, Meta 2 or HoloLens 2, but also in virtual reality such as Meta Quest (formerly Oculus Quest). We believe that the interest in isomorphic interactions comes from the need of appealing to new adopters. Besides, isomorphic interactions are also a good technical showcase of the hand-tracking technology prowess. The interest in gestural interactions has however not dwindled as across the review articles, about half of them propose this type of hand interaction. In our opinion, both types of interactions will be used jointly in the future for hybrid interactions or complementary interactions for different usages. The progress of hand tracking will also benefit on the development of gestural interactions as shown in the competition SHREC (Caputo et al., 2021) where the algorithms solely used data from hand joints.

When it comes to the technical implementations, researchers are using diverse data to track hand skeletons and/or detect hand gestures. Even if there are various algorithms including those that have been found in prior research as noted by Vuletic et al. (2019) review, such as HMM and SVM, in coherence with the popularity of neural networks, a lot of articles use solutions based on CNNs and RNNs. The issue with those solutions is the lack of flexibility because of the significant amount of data required to train the network and the necessity to retrain the network, or at least part of it through transfer learning, before adding new gestures. As such the work of Mo et al. (2021) and Schäfer et al. (2022) on gesture authoring tools, that are modular and that break down hand gestures into small primitives, are a promising perspective for prototyping interactions as well as making adaptive user interactions. With the same idea of being able to recognize gesture that were not originally in the training dataset, a recent machine learning field called Zero Shot Learning also addresses this concern. The idea of this technique is to train a model to distinguish a set of seen class using a semantic description. After training, the model would then be able to create new classifier for an unseen class if given a semantic description. The work of Madapana and Wachs. (2019) or Wu et al. (2021) for example, is an application of this technique for hand gesture recognition. The accuracy of unseen class still needs improvement but might be a promising tool for prototyping gestures in mixed reality without retraining models in the future. On top of the training time issue, CNN and RNN also require a significant computing power for real-time inference. As such, most of the prototypes are using client-server based solutions where the computing is done remotely from the headset. This solution is tackling two issues related to hardware limitations: network and computation latency, network bandwith and battery power.

5.5 Challenges

As mentioned above, gesture recognition and hand-tracking have tight computation power needs. As such, in the design of hand interactions, it is needed to decide what part of the computation is embedded in the mixed and augmented reality headset and what part is distributed. The different solutions for distributed computing are also a big consideration from the computing architecture such as edge computing or cloud computing to the communication protocols between the headsets and the remote computing machines. The distribution of the computing power also implies that the interaction recognition algorithms might need mechanisms to support the processing of the calculations on several devices and graphic cards. Currently, hand tracking specifically are embedded on headsets however those lightweight algorithms are subjected to stability issues especially when it comes to occlusion. It can be noted that hand-object occlusion has been little explored in the reviewed articles as only Mueller et al. (2017) propose an algorithm that tackles hand-object occlusion while, Myo armband, used by Bautista et al. (2020), work in those constrained situations due to the nature of the sensors. Both solutions require external sensor or computing machine that communicates with the headset. In terms of external sensor, another alternative to vision-based tracking solutions are data glove which works in hand-object occlusion situations. The disadvantages often raised for those type of devices are the price and the discomfort especially in the context where users are also using their hands to manipulate physical objects on top of the virtual content. Glauser et al. (2019) propose a hand tracking glove called Stretch Sensing Soft Glove that can be fabricated for cheap which can be a solution to prototype gesture recognition with object in hand. Another challenge that Myo armband tackles, is the tracking volume. Indeed, as mixed and augmented reality headsets mainly use front cameras to spatialize the user and track his hands, the volume of interaction is very limited. The immersion of the user is broken everytime his hands are going out of the hand-tracking coverage. One way to circumvent this issue with the current limited hardware is to give awareness of the tracking volume to the user through visual cues. Solutions similar to Myo armband might become the next replacement for physical controllers to complement hand tracking. Meta (formerly Facebook) for example, is working on their EMG armband for AR/VR input[1]. It should be noted that Myo armband was used in the reviewed article only for gesture detection and the aspect of hand tracking was not explored. The work of Liu et al. (2021) is an example of the future use of EMG armband towards a complete hand tracking of the user's skeleton.

5.6 Litterature gap

During our screening we have observed that only a small portion of the articles are comparing different interactions (12% of the articles). In our opinion, this fact is related to the lack of a standard method to evaluate an interaction in the literature. Indeed, in the reviewed articles, the designed interactions were often evaluated using qualitative (mainly usability and learnability questions using Likert-Scale questionnaires) and quantitative metrics (such completion time, precision and errors). However, each article uses their own metrics and questionnaires as popular standard usability test like the System Usability Scale are not appropriate for the specific case of hand interactions. In Koutsabasis and Vogiatzidakis. (2019)'s review, similarly a variety of metrics across the reviewed papers can be found which support the fact that there is a lack of standard evaluation method. As such, for a comparison to be made, the authors need to be able to implement multiple different interactions as a benchmark, which is time consuming. An interesting research topic could be the establishment of a formal evaluation method for hand interactions using the basis of the gesture evaluation metrics for both usability and performance grouped in Koutsabasis and Vogiatzidakis. (2019)'s review. We believe that the formal evaluation should also contextualise the interaction. The task targeted by the interaction should be factored in the evaluation using for example, the model of Hand. (1997) or a more granular classification such as what Koutsabasis and Vogiatzidakis. (2019) propose. External factors such as the hardware limitations (volume of tracking, occlusion handling, . . .) should also be considered in the impact of the rating of the interactions. Indeed, when Bautista et al. (2020) compare Meta 2 headset interactions with Myo armband, we mentioned that the limited range of hand tracking on the headset might have had an impact on the usability questionnaire results. Thus, the results of the evaluations are strongly dependant time

[1] Augmented reality headset from the company Meta which is no longer operating. (different from the current Meta, formerly Facebook) https://tech.facebook.com/reality-labs/2021/3/inside-facebook-reality-labs-wrist-based-interaction-for-the-next-computing-platform/

when the research was conducted. The standard specification of hardware limitations and contextual use case would allow an easier comparison across the papers and highlight the necessity of a further testing when the hardware has been improved.

As mentioned, in the section Challenges, hand-object occlusion is still a technical problem for mixed reality interactions. As such, there are currently little research publications considering object in hand while doing hand interactions. Furthermore outside of simply factoring hand-object occlusion in the recognition process, a step further could be to step towards multimodal interactions by mixing hand interactions with tangible interactions. An example would be the works of Zhou et al. (2020) which use the shape of the hand grip to create a virtual interactable surface. A breakthrough in hand tracking or the usage of affordable data glove would allow more research of this type to flourish as it would require less heavy apparatus.

Finally, in this review, bimanual interactions have been little explored except for Mo et al. (2021)'s work that proposes the support of two-handed gestures in their authoring tool. As in the real world, we are used to manipulate objects with both hands, it can be a promising lead to designing more natural interactions.

Author contributions

RN is a PhD student and the principal author of the article. He worked on the whole process of the scoping review, proposed the classification model and wrote the article. CG-V is the supervisor of RN. He guided RN in the review methodology, suggested relevant information to extract from the review, discussed the classification model and reviewed the article. MA is an industrial collaborator from VMware, working in the project. She reviewed the article. All authors contributed to the article and approved the submitted version.

Funding

This project is funded by the Canadian funding organization MITACS (project IT27213) and the company VMware Canada.

Acknowledgments

The authors would like to acknowledge Guillaume Spalla and François Racicot-Lanoue who provided helpful insights and documents to write this review article.

Conflict of interest

MA was employed by VMware Canada.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The authors declare that this study received the funding from the company Vmware Canada. The funders had the following involvement in the study: the reviewing of the manuscript and the decision to publish.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Ababsa, F., He, J., and Chardonnet, J. R. (2020). "Combining hololens and leap-motion for free hand-based 3d interaction in mr environments 12242 LNCS," in *Augmented reality, virtual reality, and computer graphics*, Lece, Italy (New York, NY: Springer International Publishing), 315–327. doi:10.1007/978-3-030-58465-8_24

Bautista, L., Maradei, F., and Pedraza, G. (2018). "Augmented reality user interaction to computer assisted orthopedic surgery system," in *ACM international conference proceeding series*, Merida, Mexico (New York, NY: Association for Computing Machinery). doi:10.1145/3293578.3293590

Bautista, L., Maradei, F., and Pedraza, G. (2020). "Usability test with medical personnel of a hand-gesture control techniques for surgical environment," in *International journal on interactive design and manufacturing (IJIDeM)* (Paris, France: Springer-Verlag France), 1031–1040. doi:10.1007/s12008-020-00690-9

Billinghurst, M., Clark, A., and Lee, G. (2015). A survey of augmented reality. *A Surv. augmented Real.* 8, 73–272. doi:10.1561/1100000049

Bouchard, K., Bouzouane, A., and Bouchard, B. (2014). "Gesture recognition in smart home using passive RFID technology," in *Proceedings of the 7th international conference on PErvasive technologies related to assistive environments*, Rhodes, Greece (New York, NY: Association for Computing Machinery), 1–8. doi:10.1145/2674396.2674405

Caputo, A., Giachetti, A., Soso, S., Pintani, D., D'Eusanio, A., Pini, S., et al. (2021). Shrec 2021: Skeleton-based hand gesture recognition in the wild. *Comput. Graph.* 99, 201–211. doi:10.1016/j.cag.2021.07.007

Chang, Y., Nuernberger, B., Luan, B., and Höllerer, T. (2017). "Evaluating gesture-based augmented reality annotation," in 2017 IEEE Symposium on 3D User Interfaces, Los Angeles, CA, USA, 18-19 March 2017 (IEEE). doi:10.1109/3DUI.2017.7893337

Choudhary, Z., Ugarte, J., Bruder, G., and Welch, G. (2021). "Real-time magnification in augmented reality," in *Symposium on spatial user interaction* (New York, NY: Association for Computing Machinery), 1–2. doi:10.1145/3485279.3488286

Dibene, J. C., and Dunn, E. (2022). HoloLens 2 sensor streaming. arXiv. doi:10.48550/arXiv.2211.02648

Frutos-Pascual, M., Creed, C., and Williams, I. (2019). "Head mounted display interaction evaluation: Manipulating virtual objects in augmented reality 11749," in *IFIP conference on human-computer interaction*, Paphos, Cyprus (Heidelberg: Springer-VerlagBerlin). doi:10.1007/978-3-030-29390-1_16

Glauser, O., Wu, S., Panozzo, D., Hilliges, O., and Sorkine-Hornung, O. (2019). Interactive hand pose estimation using a stretch-sensing soft glove. *ACM Trans. Graph.* 38, 41–15. doi:10.1145/3306346.3322957

Hand, C. (1997). A survey of 3D interaction techniques. *A Surv. 3d Interact. Tech.* 16, 269–281. doi:10.1111/1467-8659.00194

Hu, Z., Hu, Y., Liu, J., Wu, B., Han, D., and Kurfess, T. (2018). 3d separable convolutional neural network for dynamic hand gesture recognition. *Neurocomputing* 318, 151–161. doi:10.1016/j.neucom.2018.08.042

Jailungka, P., and Charoenseang, S. (2018). "Intuitive 3d model prototyping with leap motion and microsoft hololens 10903 LNCS," in *Human-computer interaction. Interaction technologies*, Las Vegas, NV (Heidelberg: Springer-VerlagBerlin), 269–284. doi:10.1007/978-3-319-91250-9_21

Jang, Y., Jeon, I., Kim, T. K., Woo, W., Hong, J. H., Heo, Y. M., et al. (2017). Five new records of soil-derived *trichoderma* in korea: *T. albolutescens*, *T. asperelloides*, *T. orientale*, *T. spirale*, and *T. tomentosum*. *egocentric Viewp.* 47, 1–8. doi:10.5941/MYCO.2017.45.1.1

Kim, H. I., Lee, J., Yeo, H. S., Quigley, A., and Woo, W. (2018). "SWAG demo: Smart watch assisted gesture interaction for mixed reality head-mounted displays," in *Adjunct proceedings - 2018 IEEE international symposium on mixed and augmented reality*, Munich, Germany (Piscataway, NJ: Institute of Electrical and Electronics Engineers Inc.), 428–429. doi:10.1109/ISMAR-Adjunct.2018.00130

Kim, M., Choi, S. H., Park, K. B., and Lee, J. Y. (2019). User interactions for augmented reality smart glasses: A comparative evaluation of visual contexts and interaction gestures. *A Comp. Eval. Vis. contexts Interact. gestures* 9, 3171. doi:10.3390/app9153171

Koutsabasis, P., and Vogiatzidakis, P. (2019). Empirical research in mid-air interaction: A systematic review. *A Syst. Rev.* 35, 1747–1768. doi:10.1080/10447318.2019.1572352

Lee, C. J., and Chu, H. K. (2018). "Dual-MR: Interaction with mixed reality using smartphones," in *Proceedings of the ACM symposium on virtual reality software and technology*, Tokyo, Japan (New York, NY: Association for Computing Machinery). doi:10.1145/3281505.3281618

Lee, L., Yung Lam, K., Yau, Y., Braud, T., and Hui, P. (2019). "Hibey: Hide the keyboard in augmented reality," in 2019 IEEE International Conference on Pervasive Computing and Communications, Kyoto, Japan, 11-15 March 2019 (IEEE). doi:10.1109/PERCOM.2019.8767420

Lee, T. H., Kim, S., Kim, T., Kim, J. S., and Lee, H. J. (2022). Virtual keyboards with real-time and robust deep learning-based gesture recognition. *Conf. Name IEEE Trans. Human-Machine Syst.* 52, 725–735. doi:10.1109/THMS.2022.3165165

Lin, Y. C., and Yamaguchi, T. (2021). "Evaluation of operability by different gesture input patterns for crack inspection work support system," in *2021 60th annual conference of the society of instrument and control engineers of Japan*, Tokyo, Japan (Piscataway, NJ: Institute of Electrical and Electronics Engineers Inc.), 1405–1410.

Liu, Y., Lin, C., and Li, Z. (2021). WR-hand: Wearable armband can track user's hand. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 1–27. doi:10.1145/3478112

Lu, D., Huang, D., and Rai, A. (2019). "FMHash: Deep hashing of in-air-handwriting for user identification," in IEEE International Conference on Communications, Shanghai, China, 20-24 May 2019 (IEEE). doi:10.1109/ICC.2019.8761508

Macaranas, A., Antle, A. N., and Riecke, B. E. (2015). What is intuitive interaction? Balancing users' performance and satisfaction with natural user interfaces. *Interact. Comput.* 27, 357–370. doi:10.1093/iwc/iwv003

Madapana, N., and Wachs, J. (2019). "Database of gesture attributes: Zero shot learning for gesture recognition," in 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), Lille, France, 14-18 May 2019 (IEEE), 1–8. doi:10.1109/FG.2019.8756548

McMahan, R. P., Lai, C., and Pal, S. K. (2016). "Interaction fidelity: The uncanny valley of virtual reality interactions," in *Virtual, augmented and mixed reality*, Toronto, Canada. Editors S. Lackey and R. Shumaker (New York, NY: Springer International Publishing), 59–70. doi:10.1007/978-3-319-39907-2_6

Milgram, P., Takemura, H., Utsumi, A., and Kishino, F. (1995). "Augmented reality: A class of displays on the reality-virtuality continuum," in *Proceedings volume 2351, telemanipulator and telepresence technologies* (Boston, MA, United States: SPIE), 282–292. doi:10.1117/12.197321

Min, X., Zhang, W., Sun, S., Zhao, N., Tang, S., and Zhuang, Y. (2019). VPModel: High-fidelity product simulation in a virtual-physical environment. *IEEE Trans. Vis. Comput. Graph.* 25, 3083–3093. doi:10.1109/TVCG.2019.2932276

Mo, G., Dudley, J., and Kristensson, P. (2021). "Gesture knitter: A hand gesture design tool for head-mounted mixed reality applications," in *Conference on human factors in computing systems - proceedings*, Yokohama, Japan (New York, NY: Association for Computing Machinery). doi:10.1145/3411764.3445766

Mueller, F., Mehta, D., Sotnychenko, O., Sridhar, S., Casas, D., and Theobalt, C. (2017). "Real-time hand tracking under occlusion from an egocentric RGB-d sensor," in *Proceedings of the IEEE international conference on computer vision*, Venice, Italy (Piscataway, NJ: Institute of Electrical and Electronics Engineers Inc.), 1163–1172. doi:10.1109/ICCV.2017.131

Nielsen (2022). 10 usability heuristics for user interface design. arXiv.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *Br. Med. J. Publ. Group Sect. Res. Methods & Report.* 372, n71. doi:10.1136/bmj.n71

Plasson, C., Cunin, D., Laurillau, Y., and Nigay, L. (2020). "3d tabletop AR: A comparison of mid-air, touch and touch+mid-air interaction," in *PervasiveHealth: Pervasive computing technologies for healthcare*, Salerno, Italy (New York, NY: Association for Computing Machinery). doi:10.1145/3399715.3399836

Reuters (1996). *What's a-o.k.* USA: lewd and worthless beyond.

Schäfer, A., Reis, G., and Stricker, D. (2022). The gesture authoring space: Authoring customised hand gestures for grasping virtual objects in immersive virtual environments. *Mensch Comput.* 2022, 85–95. doi:10.1145/3543758.3543766

Serrano, R., Morillo, P., Casas, S., and Cruz-Neira, C. (2022). An empirical evaluation of two natural hand interaction systems in augmented reality. *Multimedia Tools Appl.* 81, 31657–31683. doi:10.1007/s11042-022-12864-6

Shrestha, D., Chun, J., and Lee, H. (2018). Computer-vision based bare-hand augmented reality interface for controlling an AR object. *Ar. object* 10, 1. doi:10.1504/IJCAET.2018.10006394

Su, K., Du, G., Yuan, H., Wang, X., Teng, S., Li, D., et al. (2022). "A natural bare-hand interaction method with augmented reality for constraint-based virtual assembly," in Conference Name: IEEE Transactions on Instrumentation and Measurement (IEEE), 1–14. doi:10.1109/TIM.2022.3196121

Su, M. C., Chen, J. H., Trisandini Azzizi, V., Chang, H. L., and Wei, H. H. (2021). Smart training: Mask R-CNN oriented approach. *Smart Train. Mask. r-CNN oriented approach* 185, 115595. doi:10.1016/j.eswa.2021.115595

Sun, Y., Armengol-Urpi, A., Reddy Kantareddy, S., Siegel, J., and Sarma, S. (2019). "MagicHand: Interact with iot devices in augmented reality environment," in 26th IEEE Conference on Virtual Reality and 3D User Interfaces, VR, Osaka, Japan, 23-27 March 2019 (IEEE). doi:10.1109/VR.2019.8798053

Ungureanu, D., Bogo, F., Galliani, S., Sama, P., Duan, X., Meekhof, C., et al. (2020). HoloLens 2 research mode as a tool for computer vision research. arXiv.

Vuletic, T., Duffy, A., Hay, L., McTeague, C., Campbell, G., and Grealy, M. (2019). Systematic literature review of hand gestures used in human computer interaction interfaces. *Int. J. Human-Computer Stud.* 129, 74–94. doi:10.1016/j.ijhcs.2019.03.011

Wu, J., Zhang, Y., and Zhao, X. (2021). "A prototype-based generalized zero-shot learning framework for hand gesture recognition," in *2020 25th international conference on pattern recognition*, Milan, Italy (Piscataway, NJ: Institute of Electrical and Electronics Engineers Inc.), 3435–3442. doi:10.1109/ICPR48806.2021

Xiao, R., Hudson, S., and Harrison, C. (2016). "Direct: Making touch tracking on ordinary surfaces practical with hybrid depth-infrared sensing," in *Proceedings of the 2016 ACM international conference on interactive surfaces and spaces*, Niagara Falls, Ontario (New York, NY: Association for Computing Machinery), 85–94. doi:10.1145/2992154.2992173

Xiao, R., Schwarz, J., Throm, N., Wilson, A., and Benko, H. (2018). MRTouch: Adding touch input to head-mounted mixed reality. *MRTouch Adding touch input head-mounted Mix. Real.* 24, 1653–1660. doi:10.1109/TVCG.2018.2794222

Yu, J., Jeon, J., Park, J., Park, G., Kim, H. I., and Woo, W. (2017). Geometry-aware interactive AR authoring using a smartphone in a wearable AR environment. *Lect. Notes Comput. Sci.* 2017, 416–424. doi:10.1007/978-3-319-58697-7_31

Zhang, Z., Zhu, H., and Zhang, Q. (2020). "ARSketch: Sketch-based user interface for augmented reality glasses," in *MM 2020 - proceedings of the 28th ACM international conference on multimedia* (New York, NY: Association for Computing Machinery), 825–833. doi:10.1145/3394171.3413633

Zhou, Q., Sykes, S., Fels, S., and Kin, K. (2020). "Gripmarks: Using hand grips to transform in-hand objects into mixed reality input," in *Conference on human factors in computing systems - proceedings*, Honolulu, Hawaii (New York, NY: Association for Computing Machinery). doi:10.1145/3313831.3376313