State of the Art in Extended Reality - Multimodal Interaction

Ismo Rakkolainen, Ahmed Farooq, Jari Kangas, Jaakko Hakulinen,
Jussi Rantala, Markku Turunen, and Roope Raisamo

Technical report

HUMOR project, January 2021

Tampere University, Finland

Contents

1. Introduction

The Business Finland-funded HUMOR (Human Optimized XR) project investigates the perceptual issues of extended reality (XR). As a part of the project, each partner prepares a state-of-the-art (SoA) analysis in their fields of expertise. These SoAs find out the exact top of the line knowledge on the perception of virtual reality (VR) and extended reality (XR).

This state-of-the-art report on **multimodal interaction** forms a basis for the HUMOR project deliverables and helps to align the research efforts with other partners and with the best laboratories around the world.

The topic of virtual or extended reality is extensive and it has a long history. XR augments or replaces the user's view with synthetic objects. XR can be used as an umbrella term encompassing virtual reality (VR), augmented reality (AR) and mixed reality (MR). MR has also been used as an umbrella term for reality-AR-VR continuum (Milgram, Kishino 1994), and the XR terminology is not thoroughly strictly defined. In the recent years XR has become capable of generating increasingly realistic experiences and the human senses can be stimulated in more advanced ways. However, there are still many caveats, and many aspects can still be improved in XR technology.

There is a vast number of books and academic papers written about VR and XR, and we do not try to give a comprehensive, balanced view of all of it. For general overview of XR, we refer to the many books and surveys on it, e.g., Jerald (2015), Schmalstieg & Höllerer (2016), Billinghurst et al. (2015), Van Krevelen & Poelman (2010), and Rash et al. (2009). We thus assume that the reader knows the basic things about VR and related technologies. Even for this interaction section, Google Scholar gives 968,000 hits for "Human Computer Interaction" and 39,300 hits for "multimodal interaction" (with quotation marks on).

To make this SoA report more succinct, we will limit this section of multimodal VR to topics, trends, cutting-edge research results and specific hardware, which are relevant for the HUMOR project. We focus on recent trends of HMD-based XR, which are relevant to perceptual issues of XR. Furthermore, in this report we cannot describe all the possible multimodal interaction methods in depth but try to emphasize and focus on the most useful and frequently used ones.

2. Multimodal Interaction

In the field on human-computer interaction (HCI), **multimodal interaction** makes use of several input and output modalities in interacting with technology. Application fields with a slightly different emphasis include also, e.g., human-technology interaction (HTI), human-information interaction (HII), and human-robot interaction (HRI). Examples of the interaction modalities are speech, gestures (embodied interaction), gaze and touch. Technology can "read" or "listen to" these modalities, or actions of the body, as input. At the same time, the human receives input via their senses, e.g., sight, touch, or audio, that is generated as output by the technology. Sometimes modalities are also further split to sub-modalities, e.g., touch can be divided to sense of touch, sense of temperature etc.

The term **Modality** can be defined from different perspectives. In human-technology interaction, the human perspective relates both to input and output of human modalities.

- *human perceptual modalities:* visual, auditory, olfactory, and tactual (tactile, kinesthetic, haptic).
- *human output modalities*: motor system, speech, breath, bio-electric (EMG, EEG, GSR, ECG).

"*Modality is a way of representing information in some physical medium. Thus, a modality is defined by its physical medium and its particular ways of representation.*" Niels Ole Bernsen, 2008

There are many ways to categorize multimodal interaction. In Chapter 3 we will primarily present interaction with the human senses which are frequently used with XR. In Chapter 4 we present some modalities and methods which may become more important for XR in the future. Multimodal interaction can transform how people communicate remotely, train for tasks, and how people visualize, interact, and make decisions based on information.

2.1 Human senses

Humans have a variety of "built-in" senses (sight, hearing, touch, smell, taste, balance, etc.) and highly developed abilities to perceive, integrate, and interpret visual, auditory, haptic, and other sensory information. The visual and auditory modalities generally outperform the other sensory channels in terms of bandwidth and spatial and temporal resolution.

A baby looks, observes, touches, and recognizes things before she can talk. The physical world is observed with numerous simultaneous and coherent sensations of several senses and it is three-dimensional. The human observer actively processes the multisensory information using perceptual and cognitive mechanisms. The multitude of senses and their mutual interactions are meticulously and marvelously built, and human perception and cognition are very complex sensory fusion systems. A large amount of information is processed and only a small fraction gets through into our consciousness. The humans have a natural ability to operate, remember places, and perceive things in three dimensions.

The user wants to focus on the task, not on the interface between the user and the task. A good user interface should be so intuitive and easy-to-use that it would not require any substantial training. It should also facilitate expert users for fast interaction. Bad usability, on the other hand, breaks the naturalness and immersion.

Historically humans had to adapt to the computer world through punch cards, line printers, command lines and machine language programming. Engelbart's online system (1968), Sutherland's "the Sword of Damocles" (1968) and Bolt's "Put That There" (1980) were very visionary demonstrations of multimodal interfaces at their times. Rekimoto & Nagao (1995) and Feiner et al. (1997) presented computer augmented interaction with real world environments.

User interfaces change along with the changing use contexts and emerging display, sensor, actuator and user tracking hardware innovations in computing. The success of Nintendo's Wii and Apple's iPhone show the paramount role of user interfaces. Smart phones, desktop computers and XR each need different kinds of UIs. Discovering better human-technology interfaces is a modern gold rush, and XR UIs are not satisfactorily finalized yet.

The current computer user interfaces (UI) do not often take advance of all the human sensory capabilities but are mostly limited to visual domain and cognitively to 2D metaphors such as windows, icons, menus, and pointers (WIMP) and rely on desktop metaphor. The UIs for mobile computing utilize a few more sensors, but they have also many limitations such as small screens. However, XR can present information in a more natural way. The ability to recreate real sensations or create entirely new ones can enrich communication between computer and human and improve the way we interact with them. Fig. 1 shows a comparison of some human-computer interaction styles.

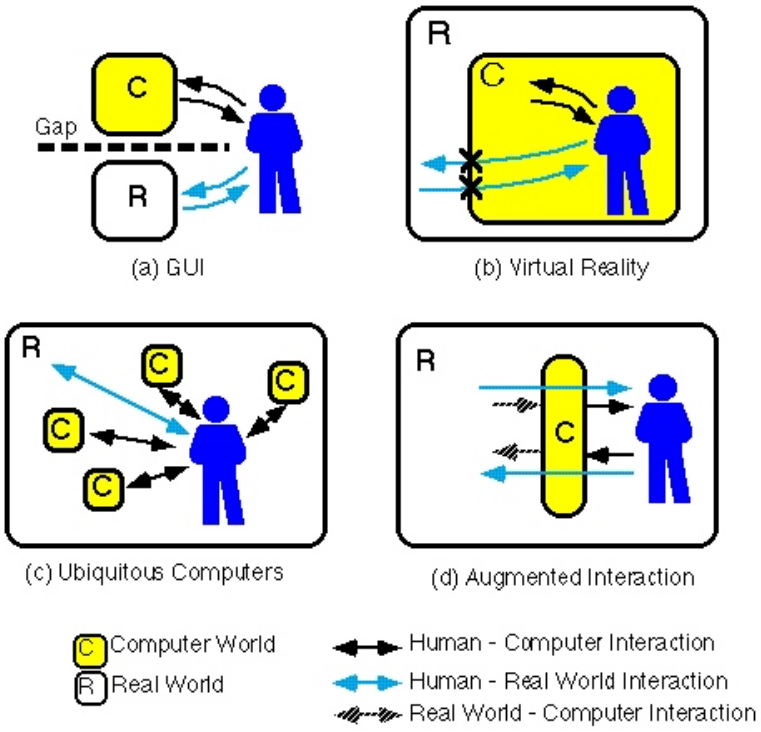

Figure 1. A comparison of HCI styles (Rekimoto & Nagao, 1995).

For a long time, the user interface research has tried to find better and more intuitive UIs (van Dam, 1997). General HCI results and guidelines cannot always be directly applied to XR, as users operate in 3D environments, input and output devices may be different, and so on. Immersion in VR is one major distinction from most other HCI styles or from traditional 3D computer graphics. VR encloses the user into a synthetically generated world and enables the user to enter "into the image".

Interactivity is one of the key features of VR. The PC's desktop metaphor is not well suited to interacting with XR. Interactions between humans and virtual environments rely on timely and consistent sensory feedback and spatial information. Effective feedback helps users to get information, notifications, and warnings. For example, haptic interaction and the feedback of events triggered by gaze enable a user to touch virtual and distant objects as though they were real and within reach.

Recently computers have started to adapt to humans through cameras, microphones, and other sensors with the help of artificial intelligence (AI), and they can recognize human activities and intentions. Human-computer interaction has become much more natural and multisensory, even though keyboard and mouse are still the prevalent forms of HCI in many contexts.

2.2. Introduction to multimodal interaction

Multimodal (multisensory) interaction employs several human senses and means of communication to provide efficient and flexible means for user input and output and user experience. They seek to leverage natural sensing (input) and rendering (output). Various multimodal interaction methods can be alternative (using one at a time) or complementary (using several simultaneously). The interface should choose or understand the right modality for the current task.

When multiple modalities are used in interaction, they can be combined in different ways. Laurence and Nigay (1993) defined a design space categorizing multimodal input to categories. In *exclusive* multimodality, modalities are used in independent and sequential manner. In this case, independent modalities are used at different times to complete unrelated tasks, and they are not combined in any way.

Examples of this are:

- selecting objects with gestural interaction, and then entering unrelated information with dictation
- recording speech and then closing all unnecessary windows with a mouse
- looking at visual content on a screen, then listening to some music

From the system viewpoint, the interpretation of the different modalities in different situations is the main support for multimodality in this case.

In *concurrent* multimodality, modalities are used in independent and parallel manner. The use of modalities is parallel, but they are unrelated, i.e., for different tasks.

Examples include:

- browsing pictures with hand gestures while performing a voice search
- driving a car while speaking on a hands-free mobile phone

- hearing a weather report while feeling the low power haptic indicator on the phone or car

Simultaneous use of modalities supports multitasking. From the system point of view this kind of multimodality can be similar to the unimodal interaction.

Alternate multimodality has combined and sequential use of modalities. The use of modalities is sequential, one after another, but they are linked together to complete the same task.

Examples include:

- performing a gesture indicating the start of speech input by raising a controller close to the mouth, then giving a speech command
- pointing an object with a controller and giving a voice command to copy it
- seeing a photo attachment from an email and then listening to the email's content

Synergistic multimodality features combined and parallel use of modalities. The use of modalities is parallel, and they are linked together to complete the same task.

Examples of this kind of multimodality include:

- selecting graphical objects with a touch screen gesture and changing their color with a spoken command at the same time
- pointing and giving a related speech command at the same time
- feeling (haptics) and listening to a dataset

Oviatt (2017) grouped theoretical foundation of benefits of Multimodal Interaction to three groups. Gestalt theory-based writings and results emphasize the point that multimodal information is more than just the sum of its parts. A classic example related to this is so called McGurk effect where visually seeing a person saying "Ga", but at the same time hearing the person say "Ba" results in you actually thinking the person says "Da". This illustrates how we human combine multiple input to find the best explanation for the entire stimulus. Another group of theories Oviatt discusses are related to working memory and related mental resources. These raise awareness to the fact that have, to an extent, their own resources for different modalities (e.g., separate visual and auditory memories) and using multiple modalities can use human resources more efficiently than unimodal interaction. The third group in Oviatt's overview is embodied cognition and related theories. These theories and related research have shown that doing is an important part of understanding and learning. For example, mirror neurons are key part of this. They are neurons in our brains which activate both when we perceive some action (e.g., somebody speaking) and when we do it ourselves (speak).

Technically, most of human-computer interfaces are multimodal but, in practice, interfaces are considered multimodal only when they have two or more primary modes of interaction. Typing with a traditional keyboard does provide haptic feedback as the user feels the keys and audio feedback when the keys click. Still only the visual feedback of text appearing on screen is considered primary and one input modality of hitting the keys is used. Therefore, the interaction is considered unimodal. Historical developments also play a role here as "traditional" interaction methods are not usually called multimodal, only "novel" technologies like speech recognition and gesture tracking are usually accepted as "multimodal".

Multimodal interaction covers a wide array of technologies, which enable, e.g., haptics, hand or body gestures, facial expression, gaze interaction, speech, scents, and brain-computer interfaces. Multimodality is one tool to create UIs with better user experience and more efficient interaction. Fig. 2 depicts some typical interaction approaches for smart glasses that are an example of potential technology for enabling multimodal interaction in the XR context.

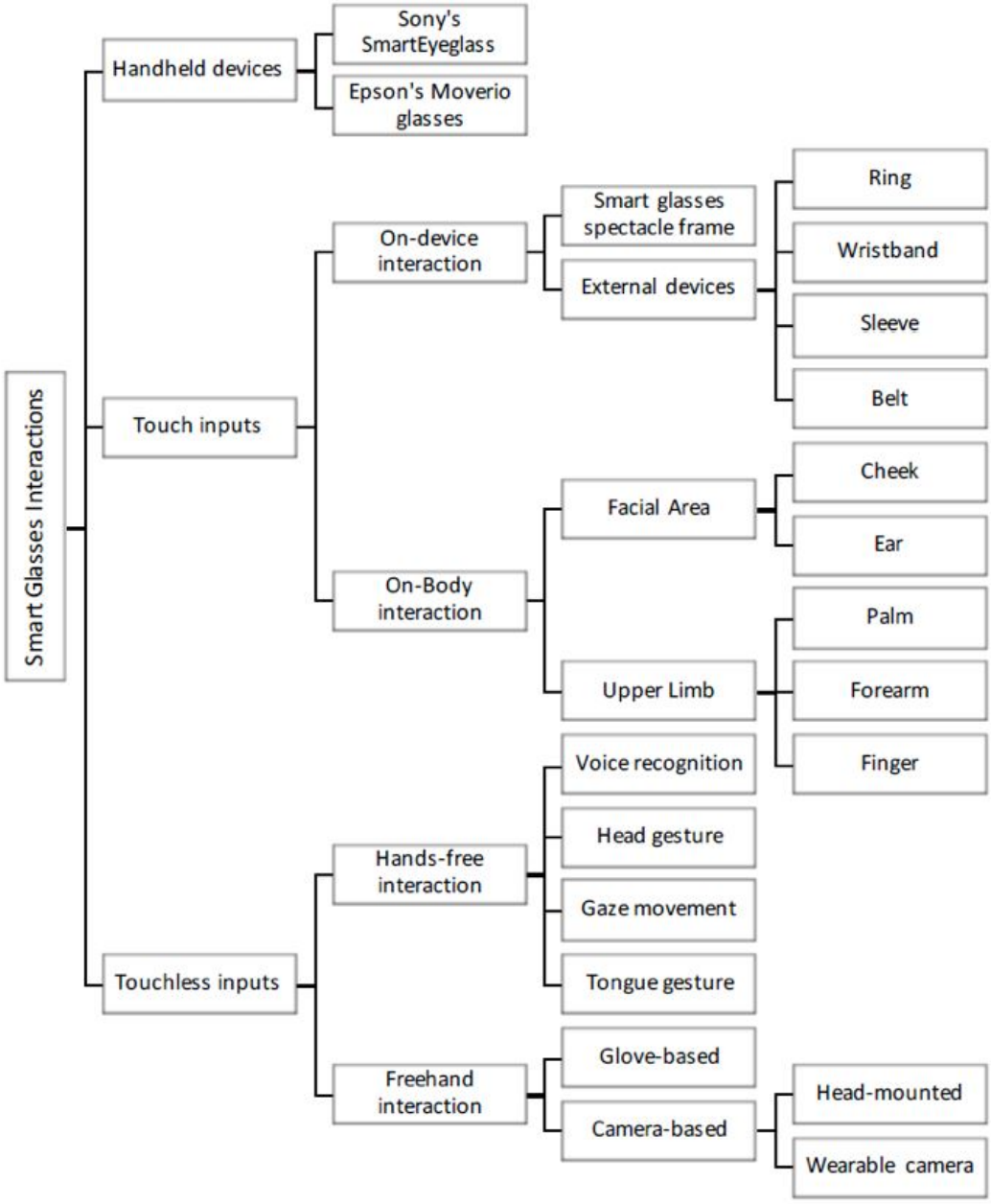

Figure 2. Classification of interaction approaches for smart glasses (Lee et al. 2018).

Multimodal UIs process simultaneously several modalities, such as speech, movements, gestures, postures, facial expression and articulation, handwritten input, emotion and mood recognition, user behavior, etc. Physiological measurements, such as EEG, heart rate, and respiratory effort, are also useful in understanding the user. Using several modalities for the same information provides an increased bandwidth of information transfer and their weaknesses offset by the strengths of others. Visual or auditory feedback is often used in multimedia and XR systems, but other multimodal techniques are very helpful on feedback and they might

improve XR to produce an immersive, believable experience. Achieving truly multisensory experiences is the Holy Grail of human-technology interaction (Flavián et al. 2020).

User interfaces should feel natural, seamless, and human-friendly. Perceptual UIs (Turk 2014) emphasize the multitude of human modalities and their sensing and expression power. In addition to human senses, they also combine an understanding of human communication, motor, and cognitive skills. Various kinds of 3D UIs (LaViola et al. 2017) take advantage of user's spatial memory, position, and orientation.

Skin and various body parts can be used as input or output surfaces (Bergström & Hornbæk 2019), and they form an always available, naturally portable, and on-body I/O system. For example, Skinput (Harrington et al. 2010) appropriates the human body for acoustic transmission, allowing the skin to be used as an input surface.

Deformable interfaces and input is one form of recent HCI, which can use various kinds of soft and malleable materials (e.g., rubber), or the physical input can be deformed, or they allow users to input in ways that are unlikely with rigid interfaces (e.g., bend, stretch) (Boem & Troiano 2019).

Mid-air or touchless interaction is one style of HCI (Koutsabasis & Vogiatzidakis 2019; Mewes et al. 2017), in which users can interact with digital content and UI controls through body movements, hand gestures, speech or other touchless means. It is useful in exergames or when touching a device would be inconvenient, e.g., while baking with messy hands, or for sterility in surgery. Many emerging technologies (e.g., depth cameras and other sensors, ultrasound haptics) or wearable devices (e.g., sensor wristbands or Microsoft HoloLens) are well suited to implement mid-air UIs. Fig. 3 depicts some typical touchless interaction tracking methods.

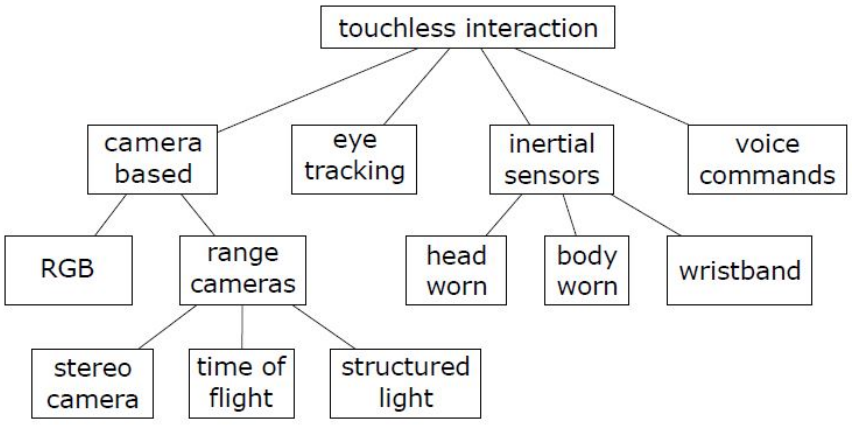

Figure 3. Overview of typical touchless interaction methods and devices (Mewes et al. 2017).

The XR scenes are often viewed with head-mounted displays (HMD), which are now feasible for consumers. A major problem of current XR is limited technical properties of displays and related tracking of users' head movements. Another major problem is inadequate interaction and UIs. The user e.g., can't directly see any hand-held controllers through the immersive VR or 360° video.

HMDs usually employ visual and auditory senses, but no other human modalities. Many emerging technologies are changing this, and most multimodal technologies have been tried with HMDs but usually not with the various combinations of them. Multimodality and integration of a wide range of sensors such as hand,

facial expression and gaze tracking is starting to take place also on commercial HMDs, e.g., DecaGear (2020) and HP Reverb G2 (2020). Multimodal means employ several human senses to provide efficient and natural means for user input and output and they can enhance the HMD viewing experience.

There are also many other ways to view XR scenes, such as various kinds of 3D, light field and holographic displays (Benzie et al. 2007), CAVE virtual rooms, fogscreens, or spatial augmented reality (Bimber & Raskar 2005), which uses projectors. IBM's director of automation research, Charles DeCarlo, presented a stunning vision of immersive projection screens for telepresence already in 1968 (DeCarlo 1968). It described a home in Phoenix, where a photorealistic live ocean scene was projected onto curved screens and realistic audio completed the immersion. The "immersive home theatre" was used for telepresence and teleconferencing. The author already foresaw VR replacing reality.

A Cave automatic virtual environment (CAVE) is an immersive VR environment where projectors create synchronized imagery to several adjoining walls. The ceiling and floor can also be used as projection surfaces. Often the projections are stereoscopic, to be viewed with synchronized shutter glasses. The CAVE can provide a very immersive environment for one or a couple of tracked viewers. There are many variations of this basic setup. Large, partially immersive projection screens are manufactured in several forms and configurations. They can be, e.g., domes or curved panoramic screens, or large, flat screen areas with several seamlessly edge-blended images, producing ultrawide field of view, high-resolution images. They fill most of the human field of view and can produce good immersion.

The FogScreen (Rakkolainen & Palovuori 2002; Rakkolainen et al. 2015) is a permeable mid-air projection display. The user can reach or walk through the 2D screen consisting of thin, planar mist in mid-air, which feels dry to the touch. It is essentially touchless and enables images, XR objects and UIs to float in thin air. It could be used e.g., as a hygiene touchscreen or a mid-air 3D display.

Spatial Augmented Reality (SAR) employs projectors and spatially aligned optical elements, such as mirror beam splitters, transparent screens, or holograms, in order to blend real and synthetic objects together. SAR displays can overcome some technological and ergonomic limitations of conventional AR systems, e.g., they do not require any head-mounted or wearable gear.

Some of the multimodal technologies could become mainstream features and applications of XR in a few years' time. They have a vast number of applications e.g., in industry, health care, entertainment, design, architecture and beyond. Well-considered multimodal experiences and improved interaction may be the key sweet spot for next-generation XR!

2.3 Three-dimensional user interfaces

3D user interfaces try to take advantage of user's spatial perception, memory, position and orientation. Multimodal 3D interaction in XR is frequently used for many tasks such as navigation, selection, manipulation, system control, and communication (LaViola et al. 2017). 3D User interfaces can feel very natural for the user, but on the other hand they do not need to follow the laws of physics, thus enabling many intriguing UIs. If

only real-life like interaction is used, it has a severe limitation as the range where the users can reach with their hands is limited.

Presumably the first experiment on 3D UIs was “the Sword of Damocles” by Sutherland (1968), and it was also the first augmented reality display. In an early work, Leftwich (1993) presented Infospace, where information and virtual objects surround the user on a desktop setting (see Fig. 4 left). As another early example, Billinghurst et al. (1998) envisioned several kinds of 3D UIs for mobile smart glasses users (see Fig. 4 right).

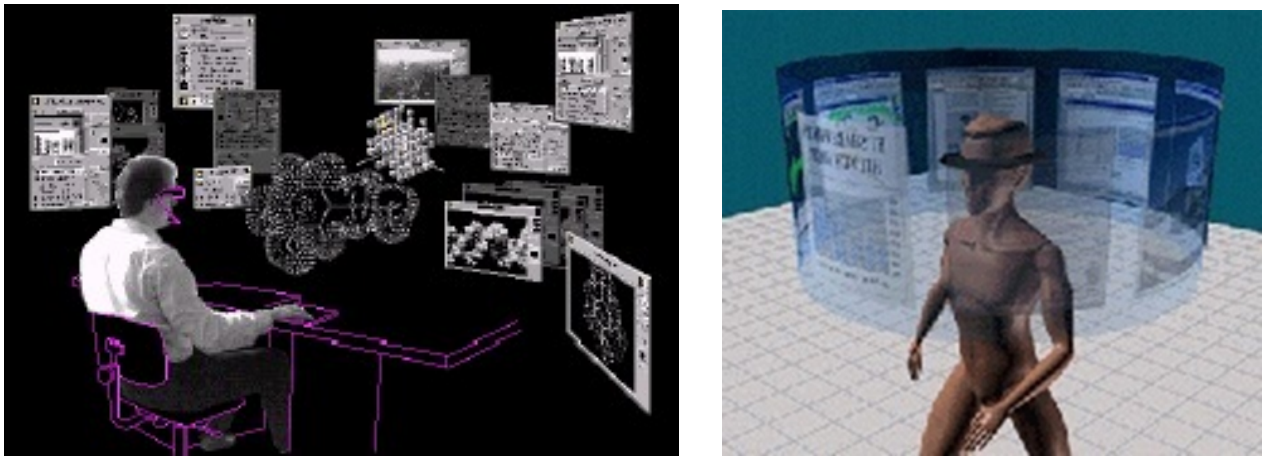

Figure 4. The data and virtual objects can surround the user (Leftwich 1993; Billinghurst et al. 1998). They still use 3D windows for information, but current systems often use AR and synthetic objects merged with real objects and environments.

3D interaction in XR can encompass many kinds of purposes and forms. General categories for 3D UI interaction (LaViola et al. 2017) include navigation (travel, wayfinding), selection (pointing, choosing objects), manipulation (grabbing, changing object position, orientation, scale, shape, etc.), system control (changing the system state or mode of interaction), and communication (with other users or agents). In practice it can mean clicking buttons, control of speed, or continuous adjustment of various things (e.g., lighting, virtual environment scaling). Through interaction the user could also move avatars, characters, or objects, bring down walls, keep balloons in air, etc.

Various interaction solutions have been created to overcome the limitations of the real world and thus to add more flexibility and possibilities to 3D interaction (e.g., Go-Go by Poupyrev et al. (1996) or other methods of scaled movement). The game Portal enables to create “wormholes” for portal jumps, which would be impossible in real world. Spicer et al. (2017) discuss some challenges and opportunities of 3D UIs for mixed reality. However, the range and other 3D UI solutions are still being studied and deviced (see e.g., Esmaeili et al. 2020).

As the desktop metaphor is not well suited to interacting with XR, several new metaphors have been formed for 3D UIs. Hand gestures are used in many proposed 3D UIs. As with gestures and deaf sign languages in general, there is no universal gestural vocabulary or a 3D UI standard. Simple gestures such as pointing are self-evident, but usually each system has its own set of gestures and convention for the meaning of specific

gestures. Due to human memory and other limitations, there can’t be too many gestures to learn and to memorize. Suitable feedback confirms the gestural action and helps the user with the system.

3. Core Multimodal Interaction Techniques

In this section we will focus on the most common modes of interaction in XR, excluding standard visual and audio properties, perception, and interaction, which are covered in SoA reports from other HUMOR partners. We will cover interaction modalities, such as auditory interfaces, gaze interfaces, gestural interfaces, and haptic interfaces. In section 4 we will also briefly discuss some other, rarely used or currently experimental modalities such as BCI, taste and scent, which have potential to be used more widely in XR in the future.

One important recent development in multimodal interaction is the role of artificial intelligence (AI), which can perform many tasks especially in visual domain. AI-based systems can e.g., render realistic synthetic scenes and humans, or recognize scenes, people, text, products, or emotions. For example, Microsoft has developed a Seeing AI (https://www.microsoft.com/en-us/ai/seeing-ai), which can describe the world around the user. It is helpful e.g., for blind and visually impaired users. In a similar vein, Google Project Guideline has helped a blind man to run solo (https://blog.google/outreach-initiatives/accessibility/project-guideline/). AI approaches are becoming commonplace also for other senses, and this may have seminal implications on the development of multimodal user interfaces.

3.1 Auditory interfaces

Auditory interfaces can be utilized in XR, and they come in many forms (Freeman et al. 2017). The audio can be ambient, directional, musical, speech or sonic. They do not match to every context or usage situation due to e.g., annoyance, privacy, or noise. There are also bone conduction headsets, which leave ears free.

Virtual reality simulates real or fictional environments. Audio in VR can therefore be split to two, audio which simulates sounds in the virtual space and audio which acts as an element of interaction. When interaction is part of the virtual world, sound can be an instance of both categories. In AR, similar use of sound is possible, but the presence of real-world sounds must be considered in all design. In addition, augmentation can be done in form of sounds. This is not common approach but there is significant potential.

Audio as a medium is public type communication, i.e., everybody in the shared space can hear the sounds in there. However, most XR solutions utilize headphones, so this aspect is not critical. Audio is a modality, which can capture users' attention efficiently, even if their visual attention is somewhere else. Because of this, audio is often used for warnings. As a modality, audio is temporal; once it is played out, the user cannot go back to it unless an explicit interaction solution is created. This is different from visual information in many cases.

Sound has a spatial aspect. Human beings can detect the direction of sound using different elements. Two ears allow to use stereo hearing to position sound to left or right. Sound arrives to the two ears with a slightly varying timing and intensity, and the human brain can estimate the direction of the sound source from them. Depending on the direction, the typical accuracy can be a few degrees (in front of the listener) or dozens of degrees (on the back). Rotating the head can help this process, and blind people typically develop a better accuracy. Bats even navigate using (ultra)sound echolocation.

In XR, users' heads are usually tracked, so supporting this kind of directional hearing is possible when volume and timing of audio sent to left and right ears is adjusted. However, there are also more complex elements we utilize to understand where the audio is coming from. The individual form of the ears shapes the frequency balance of the audio we hear depending on where the audio source is. This enables us to tell whether sound comes from above or below. Echoes and reverberation can tell us about the size, shape and materials of an environment. A stone cathedral sounds very different from a room full of pillows. All this makes creating a realistic spatial audio challenging. This is further complicated by the fact that head-related transfer function is slightly different in different people, but recent work has studied machine learning to do this personalization (Miccini & Spagnol, 2020) and various approaches providing results good enough for most uses (Wolf et al. 2020). This allows universal solutions for most applications, but extreme realism requires per user parameters.

Audio can convey information about the environment in different ways. Spatial information can tell about the shape and materials of the environment and by detecting recognizable sounds we can understand a lot from our environment. Human voices, especially speech can provide a lot of information (see Speech section for more), including emotional information.

Human beings are quite limited in how well we can interpret spatial audio. Especially in complex environment with many surfaces where audio can echo from, a simplified model may be easier and even more natural that fully realistic spatial audio. Audio is an important part of multimodal interaction, audio and visual modalities can support each other, visuals helping the interpretation of audio and audio can support awareness of surrounding when visual focus is directed to a specific direction.

Sonification uses nonspeech sound to render complex data. A simple example is to play EEG data with auditory signals in hospitals. Some parameters can be e.g., frequency, pitch, loudness, timbre, temporal structure, or spatial location. Continuous audio can be used to support awareness of users as people can spot quite subtle changes is repeating sounds. Audio can also help in fine control of systems, for example while driving a car people without much conscious effort monitor changes in speed via sound.

Auditory icons and earcons are sounds or audio elements used in interfaces. Icons are sounds which are similar to real world sounds and their meaning can be understood through this mapping. Earcons are similarly short sounds but their meaning must be, to an extent, learnt as the meaning is encoded into parameters like rhythm and pitch.

Music can be utilized in XR in a similar way as in many other media. It can create emotion but communicating information via music is also possible.

Audio created by user's actions is a significant part of multimodal interaction in some cases. By hearing things like button clicks interaction with devices can be more efficient than without.

3.2 Speech

Speech input and output, often referred as voice as well, is one form of auditory information. Apple's Siri, Google Assistant, Microsoft's Cortana, Amazon's Alexa, and other similar applications can interpret many

spoken languages, and AI can make it difficult to discern if a remote discussion partner is a computer or a real person. Nonverbal aspects of speech can also be used. These include emotions, psychophysiological states, intonations, pronunciation and accents, parameters of a speaker's voice, etc.

Most VR and AR headsets include audio input, and all have audio output support. Speech technologies include speech recognition (speech to text), speech synthesis (text to speech), speaker recognition and identification and emotion recognition from speech. All of these have developed greatly in the last decades and several companies including Google and Nuance provide speech and language technologies for numerous languages. The quality of speech recognition especially has evolved to a level where it can be applied to various domains. Deep neural networks play a significant role on it (Alam et al. 2020). In overall, speech technologies have reached such maturity that speech-based interaction is now possible in many, even if not most, domains and situations.

There are several main motives to use speech in XR environments. First, voice input provides both hands and eyes free usage. This is particularly important in professional settings where users are typically focused on the task at hand, and the benefits of XR are most obvious in tasks where hands-on activities are performed, and user's hands – and usually eyes as well - are thus occupied. Typical examples include industrial installation and maintenance tasks (Burova et al. 2020). Second, voice input is efficient and expressive. For example, speech is an efficient way to communicate selections from large sets of possible values which can be categorized, and people have names for the things they need to talk about. The power of natural language is the main factor for efficiency, and this applies both for speech input and speech output. Natural languages can efficiently communicate abstract concepts and relations. In contrast, natural language is weak when one needs to communicate about direction, distances, and spatial relations. Combination of speech and gestures can, at best, combine the two modalities to efficient multimodal communication.

As an output, speech can be relatively slow if large amounts of content are played out to the user. Most people can read faster than talk or listen. However, both listening and reading in XR environments differs greatly from what we have used to do in desktop and mobile environments. First, rendering of text in XR environments has its challenges, and in general textual modality is not as efficient in XR environments as in other environments. Second, typical XR usage scenarios, on the other hand, favor spoken output, since they are often private since people are wearing headphones, and their eyes are typically focused on other tasks.

Finally, in addition to efficiency and hands and eyes free usage, speech is a very natural way for people to communicate and it is often preferred over other modalities. Especially if virtual characters are utilized in virtual worlds, it is expected that communication with them takes place in spoken, natural manner. This requires, however, not only robust speech recognition, but also sophisticated dialogue modelling techniques. Luckily, there has been work with (spoken) natural-language dialogue systems for decades, and this knowledge can be applied for virtual environments. This is also the approach used in the voice assistants, such as Siri. The resulting human-like conversational embodied characters can be efficient guides in professional applications, for example.

Speech and language technologies require data-intensive and substantial development efforts, and therefore the technology works better with widely spoken languages. Languages with few speakers have varying level of support and languages with small market potential tend to be lacking support. Finnish, as an example, is a middle-tier language. While the number of speakers is small, domestic research, overall positive attitude towards technology and higher than average GDP have made the Finnish language market valuable enough for, e.g., Google to implement Finnish language support to their solutions. In contrast, variants of Sami language and even Finnish variant of Swedish are lacking support currently.

In some XR applications, the use environment is challenging for speech. Noisy environments can make both speech input and output challenging. For example, AR solutions for industrial use need to consider this aspect. Still, quite often efficient noise-cancelling close-talk microphones are used in this context, which can reduce the effects of noisy environment even completely, so this cannot be considered as a major challenge in the long-run.

In speech-based interaction, error management is critical. The user must be kept aware of how the system has recognized their speech and there must be ways to correct the situation. Error management may take significant time, reducing the efficiency of voice as an interaction modality. In this respect multimodal interaction, such as combining voice input with gestures can be efficient. However, since one of the main motives to use voice input in XR and AR environments is hands-free interaction, this should be done with caution. In this sense, combining e.g., voice and gaze input might be more viable solution than the typical combination of voice and gestures. Still, it should be remembered that eyes are usually often needed for hands-free interaction, so completely auditory interaction might be a viable solution for error correction at least as an alternative.

Speech can also be used to create and describe 3D models and virtual environments. As said, using speech to specify positions and dimensions is not natural for people so most likely benefits come from specifying more conceptual and abstract parameters like copying values from other objects and specifying relative values.

For navigation, symbolic navigation, e.g., using place and object names, can be efficient via speech.

Speech has the most potential in VR when used as an element of multimodal communication. Combination of speech and gestures is a natural match when gestures are used for pointing and communicating dimensions and direction while speech communicates more abstract information like operations and relations (e.g., Bolt's "Put That There" system). This can be efficient in those interfaces where user's hands and eyes can be used for interaction with the virtual environment, e.g., they are not occupied with the main tasks (such as manipulating the physical world). This is particularly true in AR environments in which the primary task is often occupying user's hands and eyes. However, in many virtual environments where tasks resemble real-world tasks voice can be efficient for other than the main task related interaction. Typical examples include educational usage (e.g., learning to perform real-world tasks) and remote operation situations.

3.3 Gesture recognition technologies

Human activity recognition is widely used e.g., in HCI, XR, security, video surveillance and home monitoring. Tracking physical objects is also important, especially if the user is wearing objects such as handheld controllers, HMDs, smart glasses, etc. In fact, user and object tracking is essential for XR. External objects at the environment can also be tracked.

Gesture recognition is a well-known method for HCI and XR. A tracking system or motion sensing input device is needed to recognize the moving gestures or static postures and the position and orientation of the HMD. Modern HMDs embed many sensors for position, orientation, motion, and gesture tracking, and they can track the user's gestures and/or hands.

Several tracking technologies can be used for gestural interaction tracking for XR (Cardoso 2019). Kinect depth camera and Ultraleap tracker popularized gestures for games, HMDs, and for many other applications. In the recent years, the tracking software and hardware has improved tremendously, and environmental tracking and hand tracking are possible for a stand-alone HMD. Often computer vision (CV) methods are used for tracking arms, hands, or fingers. It is convenient, as it often requires no user-mounted artifacts.

Hand or finger tracking is important for most human gesture recognition applications, as it is an essential part of the natural human communication. Hand gesture recognition and hand pose estimation is very challenging due to the complex structure and dexterous movement of the human hand, which has 27 degrees of freedom (DOF). It can also make very fast and delicate movements. Deep-learning-based methods are very promising in hand pose estimation.

There are many kinds of position and orientation trackers. Sensor fusion combines several of these tracking methods. One widely used method is a tiny, built-in inertial measurement unit (IMU) for orientation tracking which often contains accelerometers, gyroscopes, and magnetometers. Optical trackers use light (often IR light) for tracking. Magnetic (e.g., Polhemus, Razor Hydra) tracking systems use magnetic fields for tracking and are thus not limited to line-of-sight to any device. Acoustic tracking can also be used to locate an object's position. Typically, it uses three ultrasonic sensors and three ultrasonic transmitters on devices. For hand tracking there are also e.g., various kinds of bend sensors and stretch sensors.

Outside-in tracking uses external devices in fixed locations (e.g., HTC Vive base stations) to optically track the position (and often also the orientation) of an HMD in real-time. It has a good accuracy and low latency but has a limited range. Inside-out tracking is built into an HMD, and cameras facing outward from it keep track of the position, orientation and movement (e.g., Hololens 2, Magic Leap 1, Oculus Quest 2, DecaGear). It is less cumbersome and spatially limited but may not be very accurate. Typical outside-in base stations can track only hand-held controllers for gestures, and inside-out tracking does not have a very good accuracy for hand or body gesture tracking.

A simple and low-cost method to track gestures is to use a single webcam to detect body or hand movements and gestures, but this method is limited in many ways (e.g., accuracy, occlusions, lighting conditions, etc.).

RGBD or depth cameras is one category of gesture trackers. They usually contain an RGB camera and a depth sensor and their outputs are a color image and a depth map.

The depth sensor can be based on many technologies. One way is to project IR light patterns onto the environment (Kinect 1.0) and calculate the depth based on the distorted patterns. Some cameras emit light and measure the time it arrives back to the camera (time of flight cameras such as Kinect 2.0, Intel RealSense D400, etc., or the cameras built into the Microsoft HoloLens 2 HMD). An RGB stereo camera and CV algorithms can also find out the depth (e.g., Stereolabs ZED 2). Ultraleap Controller and Ultraleap Stereo IR 170 use a stereo IR-camera pair and IR illumination for accurate and latency-free finger, hand and gesture tracking. It can be used as a built-in or accessory sensor on an HMD. Solid state LiDAR cameras use MEMS mirror scanning and a laser for high-resolution scanning and they can nowadays be very small (e.g., Intel RealSense L515 with 61 mm diameter and 100 g of weight). An interesting emerging tracking technology is the small-size Google Soli and KaiKuTek Inc.'s 3D gesture sensor, which both use a 60 GHz frequency-modulated radar signal.

Hand-held controllers, data gloves or full-body VR suits can track the user's movements and possibly also give some tactile feedback (see the Haptics section). Data gloves or full-body VR suits (see Section on Haptics) can be more precise than cameras, and often they do not require a line-of-sight to cameras, but a user must put them on, wear and possibly also calibrate before use. They may have hygiene problems for multiple users, especially in times of pandemics. They may also be tethered and have more limited operational range than e.g., with camera-based methods.

There are some advanced hand-held controllers such as Valve Knuckles, which have a large set of various built-in sensors, including grip force sensor and finger tracking. This enables many advanced features with the controllers. DecaGear V1 hand-held controllers (DecaGear 2020) can also track these advanced features.

Some scenarios and upcoming products propose to move the 3D rendering to the cloud employing 5G networks. This would free the local processor to focus much more on tracking and thus reduce the tracking latency.

There is a massive literature on gesture tracking methods employing CV-based and other methods. Rautaray & Agrawal (2015) made a survey on CV-based hand gesture recognition for HCI. Cheng et al. (2015) made a survey on hand gesture recognition using 3D depth sensors, 3D hand gesture recognition approaches, the related applications, and typical systems. They also discuss deep-learning-based methods. Vuletic et al. (2019) made a review of hand gestures used in HCI. Chen et al. (2020) made a comprehensive and timely review of real-time sensing and modeling of the human hands with wearable sensors or CV-based methods. Alam et al. (2020) provides a comprehensive survey on intelligent speech and vision applications using deep neural networks. They also summarize on running deep neural networks on hardware-restricted platforms, i.e., within limited memory, battery life, and processing capabilities. Beddiar et al. (2020) review and summarize the progress of human activity recognition systems from the computer vision perspective. HCI and CV have traditionally been two distinct research communities.

Most of the CV-based hand gesture recognition approaches are either 3D model-based methods or appearance-based methods. Furthermore, most of the CV-based systems have three phases: detection, tracking and recognition (Rautaray & Agrawal 2015). In special cases, e.g., custom-colored or reflective gloves or other props may improve the CV tracking accuracy.

Motion capture

Motion capture (Mo-cap) trackers are designed to record the position and orientation of human bodies or objects, usually in real time. Motion capture is mainly used in medical or sports applications, film making, TV studios, etc., and it is not widely used for common HMD users (even though beacons or cameras may track e.g., the HMDs or hand-held controllers). Mo-cap devices usually track acoustic, inertial, LED, magnetic or reflective markers, or a combination of these. Some systems use active markers that emit light, some use passive (optical or magnetic) markers, and CV-based markerless systems are also being developed. The most common Mo-cap types are optical (e.g., Vicon) or magnetic (e.g., Polhemus) tracking systems. Acoustic tracking or full-body suits can also be used. While motion capture solutions can provide high-resolution data (millimeter level or better) with good data rate (many systems provide about 200Hz or more), the systems usually require calibration and wearing markers. This means motion capture can be used in prototyping but rarely in end user applications.

3.4 Gestural interaction in XR

On a general level, manual gestures can be split to mid-air gestures, gestures done with hand-held devices and touch-based gestures. In XR the mid-air gestures and gestures done with hand-held controllers are in many cases the same and they play often significant role but use of buttons often makes interaction more explicit than fully natural mid-air gestures. Some of controllers and HMDs have touch pads which allow touch-based gestures. However, rarely are complex gestures with these used. However, already simple swipes increase the interaction vocabulary of such devices. Current hand-held controllers utilize mostly wrist rotations, but pen-type devices can potentially support more precise and faster movements (Li et al. 2020). In addition to manual gestures, facial gestures are important especially in human-human interaction. Detecting facial gestures is done in various prototypes to support remote collaboration in VR. Gaze is an important element of human-human communication, as dyadic and triadic gaze provide insight to attention of other people. See dedicated chapter for details about gaze. Facial gestures are not, at the moment, widely used as explicit input to the system.

Tracking the user's movements and postures (see Section 3.3) form the basis for the gesture-based interaction. Gestural interaction can be used in many settings, e.g., in homes, offices and surgical operation rooms. Smartphones, tablets, laptop and desktop computers can also interpret user gestures with their sensors. The short sci-fi movie World Builder (https://www.youtube.com/watch?v=VzFpg271sm8) shows an illustrative example of how gestures could be used to build a life-like virtual environment.

Hand, body, or head gestures can be used for human-to-human or HCI interaction and are very helpful or even essential in many contexts. Deaf sign languages (of which there are many) are one specialized example.

Pointing with hands or index finger is a method learned already in early childhood. The finger is an intuitive, convenient, and universal pointing tool, which is always on and available. It is used across all cultures and does not require any literacy skills. The meaning of some gestures varies across cultures, e.g., waving goodbye in Europe means “come here” in India. The gestures may also have emotional, social, cultural, and other meanings and levels.

Bolt’s “Put-That-There” (1980) was an early multimodal interface combining gestural interface with speech input. It enabled pointing at an object on a screen with a hand and giving commands with speech, for example “Put that there”. This is an example of deictic gestures, i.e., pointing indicating objects and locations. Deictic gestures are natural for people and they are actively studied in VR context (see e.g., Mayer et al. 2020; Henrikson et al. 2020). In VR, there are two common ways to implement points, cursor-based relative pointing where hand movements move cursor and ray casting, where a ray is drawn through two points in user’s body, e.g., from shoulder to wrist. While relative pointing works on displays, in immersive environments ray pointing is usually utilized and often it is also visibly shown to the user. Another early work were Krueger’s installations since the 1970’s (Krueger et al. 1985) which enabled to interact with visual art using viewer’s body movements.

Gestures and postures have been used in VR solutions for decades as hand-held controller with tracking technology have been utilized in many systems and are standard in most systems today. In AR systems, gestures are also common but rarely utilize controllers. In an early work Mann (1997) presented a finger mouse for a head-mounted camera and display. It enabled control of a cursor with a finger, allowing the user to draw outlines of objects. Finger pointing was used also in an interface for wearable computing (Starner et al. 1997). It allowed the user to replace a mouse with his/her finger. Finger pointing was used e.g., to control the local wearable computer’s pulldown menus or draw an image.

Gestures are commonly used in VR interaction since they match the immersive nature of VR experiences. Direct touching and pointing are natural ways to interact in immersive virtual reality, especially when the user is standing, and locomotion is by walking. However, interaction does not need to be a replica from reality, but it can use more powerful and flexible methods (see section on 3D UIs).

Gestural interaction is often used with XR. In fact, it is usually even more important for XR than HCI. One early work on a gestural UI with computer vision-based hand tracking for HMD was Kölsch et al. (2006). Li et al. (2019) made a review on gesture interaction in virtual reality. Chen et al. (2017) compared gestures and speech as input modalities for AR. They found that speech is more accurate, but gestures can be faster.

Superwide field-of-view (FOV) on HMDs conveys peripheral information, improves immersion, situational awareness, and performance (Ren et al. 2016) in some tasks and is generally preferred by audiences. It has also an impact on user interfaces and interaction. There are some recent superwide FOV HMDs, both academic (e.g., Rakkolainen et al. 2017) and commercial superwide FOV HMDs (e.g., Pimax 8K, StarVR One). Gestures with them can be much wider than usually, but this is a relatively little researched field. Ultraleap Stereo IR 170 is capable of tracking 170x170° FOV and it has also a longer tracking range (10cm to 75cm).

3.5 Locomotion interfaces

Locomotion interfaces enable users to move around in a virtual space and make them feel as if they indeed are moving from one place to another (Olivier et al. 2017). The user can e.g., fly on a flight simulator or walk or bike in a virtual environment, when in reality she is just moving on a pneumatic platform, treadmill or on a similar VR motion platform. It is also possible to trick a walking user visually in VR to feel that she is walking straight when she is actually walking in circles.

In the last few years, locomotion interfaces and tracking have progressed substantially. Applications that require users to traverse virtual space often relay of existing tools such as teleportation, astral body projection, blinking, tunneling etc., however each technique may have its issues especially with respect to motion sickness. This is because in a typical implementation the system is using head or hand-based locomotion.

Using the controllers or headset to navigate can be awkward and may confuse the user. Recently, wearable devices and sensors coupled with HMDs provide hip-based tracking (e.g., DecaMove sensor (DecaGear 2020) to clip on pants or belt). Hip tracker can not only make inverse kinematics and body tracking easier but also reduce motion sickness by using hip-based locomotion over hand / head-based locomotion.

Locomotion by walking has been found to support user's spatial understanding and help way finding. An approach which has received significant amount of interest is redirected walking where limited space to walk is used so that the virtual environment is distorted to make the user walk within the physical limitations. This way the user feel she is walking the virtual environment and receives most of the benefits (immersion and spatial understanding) of real walking while reasonable physical space is used. For example, IEEE VR conference had a dedicated session on the topic ("3DUI - Navigation - Redirected Walking"). Similar enhanced experience has been made also for jumping in VR (Havlík et al. 2019).

Other ways to locomotion by walking are treadmills, where the user can walk naturally while the floor below her is moving – even to any direction. Yet other ways are curved, slippery walking platforms where the user walks kind of naturally, but actually stays in the same spot, the user stepping in place, or the user sitting on a chair and wearing slippers, which can sense movement, thus simulating walking (e.g., Cybershoes for Oculus Quest, https://www.kickstarter.com/projects/cybershoes/cybershoes-for-oculus-quest). Another approach is a "VR hamster ball" VirtuSphere, which fully surrounds the user, enables to walk to any direction. Large robotic arms carrying the seated user is one way to create wild sensations of motion. Various locomotion platform products include Kat walk (https://www.kat-vr.com/pages/kat_walk_c), Infinadeck (https://infinadeck.com/), Omnideck (https://www.omnifinity.se/), Stewart platform (https://en.wikipedia.org/wiki/Stewart_platform), and 3dRudder (https://www.3drudder.com/).

Audi Holoride (2019) is a gaming/VR platform for backseat passengers. Instead of causing motion sickness on top of carsickness, it takes advantage of the stops, accelerations, bends, and other movements of the car, and transports them to the virtual environment. The motions of the virtual world and the car are in sync. Hence the car becomes a locomotion platform. It even seemed to reduce any symptoms of motion sickness and nausea.

Boletsis (2017) has done a recent review of different locomotion techniques. Fig. 5 shows the current locomotion techniques found for the review.

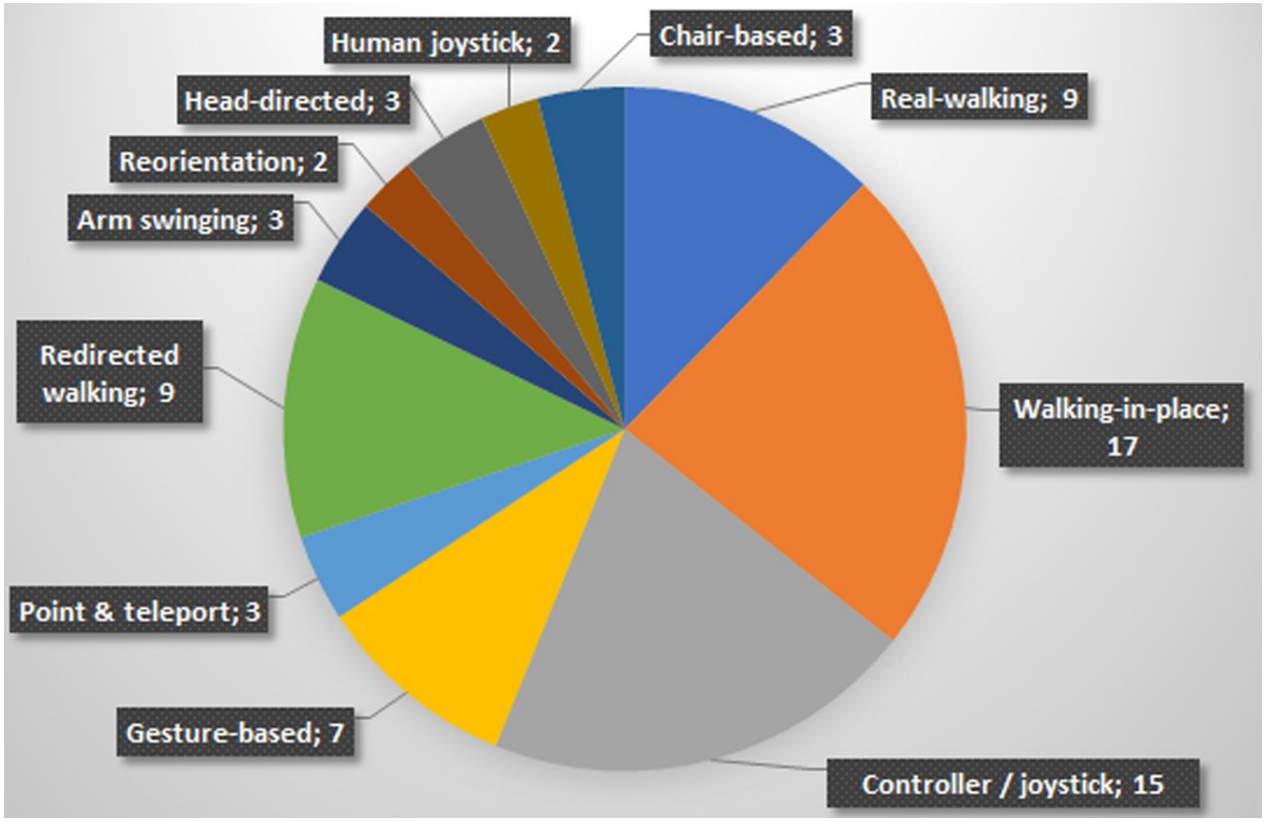

Figure 5. A survey of current locomotion techniques (Boletsis 2017).

3.6 Gaze

Humans use gaze to study their environments, to look at objects and (sometimes) to communicate with others. Eye tracking means using (various) sensor technologies to identify the locations and objects that the gazer is looking at, or a sequence of such gaze locations. Knowing that, the eyes can then be used as a control tool in HCI or as an indicator for understanding the person's mental state or intentions.

Recent advances in eye tracking technology have lowered the prices and made the technology more generally available. For example, Tobii is offering a gaming-oriented gaze-tracking product with a price of around 200€ (Tobii 2020). A gaze aware game will know what the gamer is paying attention to and may adapt the game content and interaction possibilities based on gaze behavior.

Gaze can be utilized in HCI in various ways. Gaze can be used to infer the user's interests based on gaze patterns (as above), or gaze can be used to provide specific commands. Users' interests depend on the individuals, context, tasks, culture, etc. Fig. 6 depicts the heatmaps (visualizations of gaze concentrations) of Korean and Finnish viewers of the Korean scene.

Figure 6. A Korean scene on the left have different fixation point distributions in the Korean (middle) and Finnish (right) viewers. The Korean viewers seem to concentrate on the people around the front table, while the Finnish viewers are also looking at the Kimchi preparation, which they probably are unfamiliar with (Isokoski et al. 2018).

As eye is primarily used for sensing and observing the environment, using gaze as an input method can be problematic since the same modality is then used for both perception and control. The tracking system needs to be able to distinguish casual viewing of an object from intentional selection act, in order to prevent the "Midas touch" problem where all viewed items are selected. A common method to prevent erroneous activations is to introduce a brief delay, "dwell time", to differentiate viewing and gaze-control (Majaranta et al. 2009). The user needs to stare the object, at least, for the duration of the dwell time to activate it. Blink and wink detection can also be used as control tools in gaze interaction (Kowalczyk & Sawicki 2019). In multimodal settings also other control methods, such as body gestures (Schweigert et al. 2019), audio commands (Parisay et al. 2020) or a separate physical trigger (Nukarinen et al. 2018b), can be used to activate a gaze-based selection.

There are three commonly used methods to utilize gaze as an explicit command: dwell-select (described above), gaze gestures (Hyrskykari et al. 2012), and smooth-pursuit-based interactions. Dwell-select is based on long fixations on a target in order to select and activate it. Gaze gestures are based on system recognizing a sequence of rapid eye movements (saccades) that follow a specified pattern to activate a command (Drewes & Schmidt, 2007; Istance et al. 2010). The sequence needs to be such that it doesn't often occur naturally while casually viewing the environment. Finally, the smooth pursuit interaction is based on recognizing a continuous movement of gaze while tracking a specific moving target (Vidal et al. 2013; Esteves et al. 2015). The system deduces that if the gaze follows a similar trajectory to that target, there is a match, and a selection or a (target specific) command is triggered. Recent example of pursuit-based selection in VR is a study by Sidenmark et al. (2020).

Several technical solutions have been developed for tracking eye movements (Duchowski 2007) and defining gaze position, or gaze vector (Hansen & Ji 2010). The most common method is analyzing a video image of the eye, using video-oculography (VOG). For each video frame captured by a camera located somewhere close to the user's eye, the tracking software detects several visual features, such as pupil size, pupil center, and so on. VOG-based trackers typically require a calibration before the gaze point can be estimated. The VOG system fits naturally to HMD as the cameras can easily be installed close to the display elements facing the user's eye. In addition to VOG, eye movements can also be detected by electro-oculography (EOG), based on the cornea-

retinal potential difference (Majaranta & Bulling 2014). EOG systems require the sensors to touch the skin close to eyes, which is possible to arrange in HMDs. EOG is most useful in detecting relative eye movements (e.g., gaze gesture recognition) when the exact point-of-gaze is not needed.

Eye tracking has been studied for a long time, and gaze control has a long history as an input method for human-technology interaction (e.g., Hutchinson et al. 1989). At first the gaze interaction method was mostly used for special purposes, like typing tools for disabled (Majaranta & Räihä, 2002) who couldn't use other methods, but through an active research and affordable new trackers gaze-based interfaces can now be included in many new devices. Through research it has been demonstrated how eye tracking could enhance the interaction with mobile phones (Rozado et al. 2015), tablets (Holland & Komogortsev 2012), smart (watches (Akkil et al. 2015), smart glasses (Zhang et al. 2014) and public displays (Zhang et al. 2013).

Recent development of the gaze tracking in VR have made it easy to study gaze behavior in a virtual environment (Clay et al. 2019), There are already several commercial HMDs with integrated eye trackers (e.g., HTC Vive Pro Eye, Fove, Magic Leap 1, Varjo, HP Reverb G2, Pico Neo2 Eye), and eye tracking is expected to become a standard feature. For mixed reality use the gaze-based input has been used in studies with HMDs (e.g., Piumsomboon et al. 2017; Nukarinen et al. 2018; Nukarinen et al. 2018b). Also, Meissner et al. (2019) have used gaze tracking in VR to study shopper behavior. Using virtual reality-based system makes it extremely easy and fast to modify the shopping environment for straightforward comparison trials. Tobii (Tobii VR, 2020) and Varjo (Varjo Eye Tracking in VR, 2020) are providing integration tools for gaze data collection and analysis for VR-based use cases. Burova et al. (2020) utilized gaze tracking in development of AR solutions using VR technology. By analyzing gaze, it is possible to understand how AR content is perceived and check also real-world related aspects like safety of AR use in risky environments. Such safety issues can be found early with a VR prototype. Also, Gardony et al. (2020) discuss how the gaze tracking can be used to evaluate the cognitive capacities and intentions of users to tune the UI for improved experience in mixed reality environments.

While the focus of gaze interaction studies has usually been in intentional control of HCI using gaze, the research interest for other use cases is steadily growing. As described in the game example and in the shopping example above the computational systems can utilize the gaze data also to "know" of the user's attention or interests, and optionally to adapt to that. Other examples could be that the system might notice user confusion by following gaze behavior (Sims & Conati 2020; DeLucia et al. 2014) and offer help, recognize cognitive state of a person (Marshall 2007), make early diagnoses of some neurological conditions (Boraston & Blakemore 2007) or analyze the efficiency of advertisement (Dreze & Hussherr 2003). Human gaze tracking constitutes an important part of the vision of Augmented Human (Raisamo et al. 2019). In a way, humans also analyze (often unconciously) other people's interest or state-of-mind by their gaze behaviour.

Research on gaze UIs is often divided into research on gaze-only UIs where only gaze information is used for input, gaze-based UIs where the gaze information is the main input modality, and gaze-added UIs where gaze data is used to add functionality on an otherwise functional UI. For example, in gaze-controlled assistive sys-

tems, gaze-based interaction is the main (sometimes the only) method of communication and control (Majaranta & Räihä 2002). Alternatively, information from the user's natural gaze behavior can be exploited subtly in the background as an input channel in attentive interfaces in a wide variety of application areas (Hyrskykari et al. 2005; Hansen et al. 2005). The research on HMD integrated gaze tracking naturally falls into the latter category as all the other input modalities are also available.

One example of how gaze data can be used for human behavior analysis in industrial tasks is presented in Burova et al. (2020). As seen in the figure below, gaze data can be used to identify where industrial personnel are looking at when they are performing operations. This can be used not only to analyze what users have seen and in which order, but what they have missed. These can be crucial information in safety critical environments, since it would be possible to detect e.g., hazardous situations and unsafe working conditions even when other measures (e.g., task-completion and error rates) show that the tasks have been completed successfully.

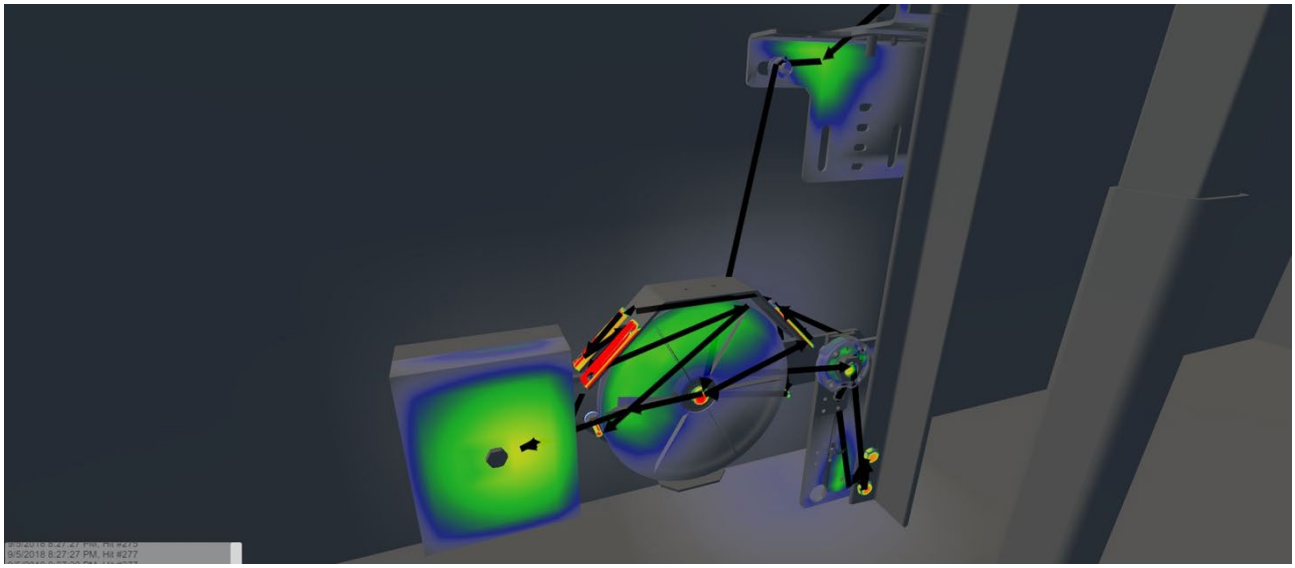

Figure 7. Gaze-data analysis applied for an industrial operation.

3.7 Haptics

Haptics is an integral part of our lives and activities. Humans use the sense of touch to grasp, explore, walk, and manipulate in the real world. The sense of touch is delicately and marvelously built, it is a very complex system, and it pervades the whole body. It comprises of cutaneous inputs from the various types of mechanoreceptors in the skin and kinesthetic inputs from the muscles, tendons, and joints. It provides updated information, e.g., on 3D shape and texture of objects, the position of the limbs, balance, and the muscle stretch (Biswas & Visell 2019). Mechanoreceptors have various densities in various body parts. The sense of touch associates a certain tactile stimulation with pressure, vibration, pain, temperature, or pleasure.

Haptic devices are an interface for communication between human and computer. The sense of touch must be artificially recreated, e.g., in interactive computing, virtual worlds and robot teleoperation. Mechanoreceptors

in the human body are artificially stimulated to produce expedient sensations of touch. This can enhance realism and human performance. Tactile feedback is usually provided in direct contact to skin, which seems intuitive for the sensation of touch. Some haptic devices can also provide kinesthetic feedback which can stimulate muscles, joints and tendons as well as the skin receptors. Many technologies can be used, e.g., tactile gloves, exoskeletons or proxy devices. There are several surveys on haptics in general (e.g., Culbertson et al. 2018; Berjemo & Hui 2017; Choi & Kuchenbecker 2013; Biswas & Visell 2019), and recent surveys on haptics for VR (Wang et al. 2019).

The fidelity of current tactile display technologies is very rudimentary compared to audiovisual displays or to the capabilities and complexity of human tactile sensing (Biswas & Visell 2019). The shortcomings of tactile display technologies amount to several orders of magnitude (Hamza-Lup et al. 2019). Many shortcuts and approximations must be used in order to mass-produce haptic displays for general use such as device degrees of freedom, response time, workspace, input / output position & resolution, continuous / maximum force and stiffness, system latency, dexterity and isotropy. As haptics is a personalized method of interaction, the mass-produced approach can often create inconsistent outputs. Moreover, interaction devices being developed today are only able to generate artificially encoded signals to the skin which may also contribute towards lower information transfer rate and higher cognitive load as compared to visual and auditory modalities. Having said this, during interaction where auditory or visual modalities are restricted or in conditions where the two modalities are insufficient at creating an immersive interaction experience, even low-resolution haptic feedback can improve user experience substantially.

Moreover, most haptic systems are part of a multimodal interaction platform consisting of audio and visual components. Similar to these components, haptic feedback also needs to be encoded, generated and delivered to convey a specific spectrum of information. Like other communication channels, the haptic channel can either contain unique or redundant information, in any case end-to-end communication needs to happen effectively and with minimum latency to ensure that the multimodal experience is natural and immersive across all the available modalities. This can be difficult to achieve as devices and components providing the information may use very different technologies and may relay the information to the user at different rates.

Unlike the auditory and visual information within a multimodal system, haptic feedback consists of artificially encoded signals which can take longer to encode, generate, and relay to the user (Fig. 8). If this issue is not accounted for while designing system interaction, user experience can suffer considerably. For that reason, it is important to understand how haptic feedback is created and relayed to the user in comparison with other information channels of the system.

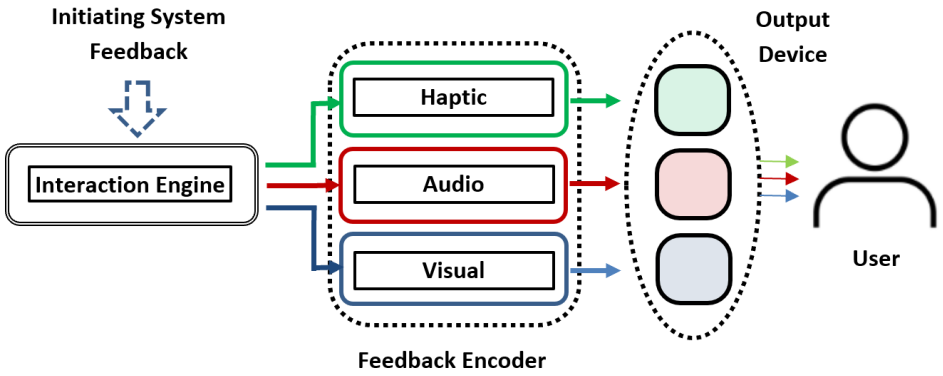

Figure 8. Illustration of the need for a bottom-up approach to encoding, generating, and delivering multimodal information to the user.

XR interaction is a prime example of how haptic feedback can greatly enhance the immersion, performance, and quality of the interaction experience (Wang et al. 2019; Biswas & Visell 2019), but the technology is still very limited. The lack of realistic and natural haptic feedback prevents deep immersion during object contact and manipulation. It is very disappointing and confusing to reach toward a visually accurate virtual object and then feel rudimentary tactile signals (or no tactile signals at all) when picking it up.

Hand-held controllers provide only global vibrotactile feedback, even though they are easy to use, unlike many specialized and cumbersome haptic devices, gloves, and full-body suits such as Teslasuit. If a haptic device requires laborious mounting, adjustments, and calibration, it is not very user-friendly and may discourage their use, but on the other hand, once on, they improve the feeling of presence. Skin-integrated adhesive bandages or patches (Yu et al. 2019) may provide a less obtrusive approach. Vibrotactile feedback has also been employed directly on HMDs (e.g., Oliveira et al. 2016; Kaul & Rohs 2017; Nukarinen et al. 2018).

In the following subsections we take a look as some of the promising technologies which can improve tactile feedback for VR / XR interaction in four key areas: Non-contact; Surface-based or hand-held; Wearable; Multi-device and full-body interaction.

Non-contact interaction

Current VR / XR interaction uses a mixture of mid-air gestures and various controllers for tracking and delivering tactile feedback. As the technology improves, para-personal and mid-air gestures may partly replace fixed hand-held controller and should create a natural and seamless interaction space. However noncontact-based interaction has a significant drawback: the lack of unobtrusive, tactile feedback. Interaction without it can feel unnatural and can lead to uncertainty. The ability to 'feel' content in mid-air can address fundamental usability challenges with gestural interfaces (Freeman et al. 2014, 2017).

Effective touchless tactile feedback techniques (Iwamoto et al. 2008; Carter et al. 2013; Long et al. 2014; Rakkolainen et al. 2020) have the potential to help users interact with virtual environments by improving non-contact gesture performance and enhancing tactility through touchless haptic space. Using calibrated ultrasound (Rakkolainen et al. 2020) or pneumatic (Farooq et al. 2014) transducer arrays, researcher have been successful at bringing complex 3D virtual objects to the physical space. Mid-air tactile feedback is unobtrusive

and maintains the freedom of movement. It can improve user engagement and create a more immersive interaction experience.

Ultrasound haptics creates acoustic radiation force, which produces small skin deformations and thus elicit the sensation of touch. It has been combined with some 3D displays (e.g., Hoshi et al. 2009) and VR systems (Martinez et al. 2018; Furumoto et al. 2019). The array can be placed, e.g., on a table in front of the user (Kervegant et al. 2017; Martinez et al. 2018), or directly on the HMD (Sand et al. 2015b), as depicted on Fig. 9. The user can see objects through an HMD and feel them naturally in mid-air.

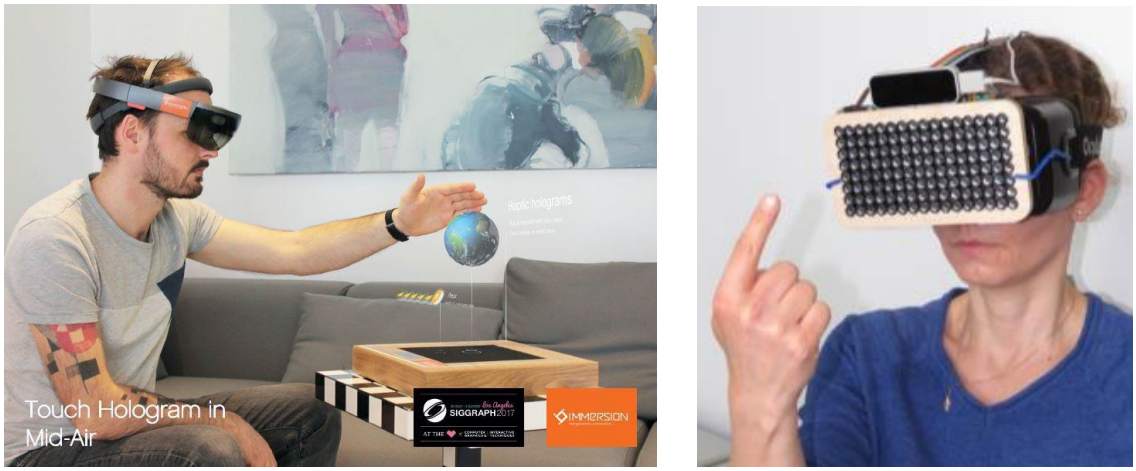

Figure 9. Left: Mixed reality ultrasound haptics with the array on a fixed position (Kervegant et al. 2017). Right: Mid-air ultrasound haptics with the array in front of a VR HMD (Sand et al. 2015).

Research results on perception of mid-air ultrasound haptics feedback suggests that the technique can provide similar properties as vibrotactile feedback on the perceptibility of frequencies. Early research focused on the detection of one (Hoshi et al. 2010) or multiple points of feedback (Carter et al. 2013). The sense of touch in fingers and palms is the most sensitive to vibration of 150 – 250 Hz (Ryu et al. 2010). A good form of mid-air tactile feedback for a button click is a single 0.2 s burst of 200 Hz modulated ultrasound (Palovuori et al. 2014).

The minimum distance on the skin that is required to recognize the position difference between a projected visual point image and a tactile stimulation is about 10 – 13 mm regardless of the stimulation patterns (Yoshino et al. 2012). The average localization error of a static point is 8.5 mm (Wilson et al. 2014). Stimulation of multiple points along the trajectory, longer durations (50–200 ms) and longer traveling distances (>3 cm) all improve movement perception. Raza et al. (2019) presented a perceptually correct haptic rendering algorithm which is independent of the hardware.

Several aspects of the use and interaction of mid-air ultrasound haptics have been researched. Carter et al. (2013) provided multi-point haptic feedback above an interactive screen. The transducers were under a front projection screen, which allowed ultrasound to pass through it. A similar visioacoustic screen (Yoshino & Shinoda 2013) allowed ultrasonic pressure to pass through it while back-projecting light on screen. They also presented two interaction layers: a guide layer for hand guidance, and an operational layer near the screen for button presses etc. Mid-air haptic shapes are not easily identified even after a learning phase (Rutten et al.

2019). Linear shapes are easier to recognize than circular shapes, and increasing age decreases tactile sensitivity.

An interactive system (Matsubayashi et al. 2019) enables a user to manipulate 3D object images with multiple bare fingers receiving haptic feedback. It improves the recognition of the object surface angle and position and it enables the user to hold and move the object easily even if it is not visible. In the future, flexible printed circuit technology could enable very thin and transparent ultrasonic emitters (Van Neer et al. 2018), which could possibly even be pasted onto a visual display. The printing technology can also bring down the cost significantly.

Surface-based or hand-held interaction

Although VR / XR interaction space may not require direct physical contact, most haptic systems rely on surface-based interaction to track and deliver vibrotactile feedback. Devices such as touchscreens and smart surfaces augment virtual objects and environments into physical space to create mixed reality interaction. System such as the Haptic Stylus (Farooq et al. 2016) use discrete point on the touchscreen to locate and tract interaction and subsequently deliver tactile and kinesthetic feedback to the user in combination with visual and auditory modalities. Similar to inertial-based interaction through standalone 3D devices (Phantom, Falcon etc.) used in conventional VR / XR system, the Haptic Stylus approach regulates a physical manipulandum linked to the virtual environment.

Smartphones have also been utilized as either interaction tools (Desai et al. 2017) or the core platform (Chuah & Lok, 2012) for VR / XR interaction. Initiatives such as Google's Cardboard VR, Samsung's Gear VR, and Apple AR kit extend the onboard resources of a conventional smartphone to become a platform for consuming VR / XR content. The onboard visual, auditory, and haptic interaction environments, although limited (Pokemon Go), can serve as a rudimentary window into creating experience not possible before, without dedicated hardware. This type of interaction has been researched to extend VR / XR to multiple devices and use cases. The hybrid approach taken by Brown University researchers (Portal-ble system) utilizes visual interface of a mobile device but instead of using the touchscreen as the point of interaction, the system uses rear-mounted sensors to track the user's hands in real time. Essentially, the mobile device becomes a window into the virtual environment and the users can reach out and interact with the virtual environment (Qian et al. 2019). Although this implementation may improve visual interaction, it takes away the highly informative haptic feedback needed to interpolate virtual objects. Similarly, in other implementations where smart surfaces or touchscreens are used to create AR / XR experiences, haptic interaction either takes a back seat or becomes so impractical an interaction mode that it is substituted with visual and auditory modalities, thus drastically reducing the immersive experience. This means that dedicated hardware needs to be developed to communicate with the main device (smartphone, tablet, etc.) and generate the necessary haptic signals required for meaningful user interaction in VR / XR environments.

Other techniques such as "inForce" at MIT lab use projection (Nakagaki et al. 2019) displays to overlay the user's physical workspace by augmenting virtual objects and environments whereas, tangible objects carefully

selected for the interaction experience provide tactility needed to complete the immersion process. Similar techniques have been used in various mixed interaction environments from creating augmented eating experiences (Allman-Farinelli et al. 2019, Daniel et al. 2019, Stelick et al. 2018) to complex training procedures (Kaluschke et al. 2018, Brazil et al. 2018, Karafotias et al. 2017). However, the techniques either require custom designed interaction surfaces or novel hand-held devices which need to be mapped to the virtual environment in real time to provide meaningful tactile or kinesthetic experiences in combination with visual, auditory, and even olfactory feedback. One example of this is the PlayStation 5 dual sense controller (Doucet et al. 2020) which allows the user to experience various kinesthetic (through adapted trigger buttons) and tactile signals (through individually actuated wings) overlayed on to the game mechanics and environment. However, this implementation needs a dedicated layer of encoded haptics on top of the visual and auditory feedback layer, which requires additional information from the developers. For that reason, encoding the relevant information across the different multimodalities needs to be done in such a way that no one output overpowers the rest. Moreover, immersive interaction experiences can only be created if the interaction engine and feedback devices (Fig. 8) need to encode and deliver the calibrated signals to the user for a natural multimodal experience.

Wearable device interaction

Due to the challenges in providing direct mid-air actuation for VR / XR interaction, wearable devices have been used to indirectly relay tactile and kinesthetic information to the user. Conventional wearables such as gloves, rings, wristband, and watches can serve as an always-on interface between events and triggers within the virtual environment and the physical space. Using Bluetooth LE connections and integrated actuation components and drivers these devices are an ideal platform to relay basic tactile feedback. In some cases (fitness tracker, and rings), the onboard sensor can also provide tracking or movement information which can be relayed to the VR / XR central device to improve the overall experience. However, as these devices are not designed to provide rich complex actuation signals to the users, in most cases their haptic output is limited. Moreover, due to their size and smaller onboard batteries, these devices cannot reliably generate sensible feedback signals for extended sessions. And lastly, most devices utilize wireless connections that prioritizes efficiency rather than low latency reliable actuation feedback, which essentially translates to unreliable haptic output compared to auditory and visual experience (Fig. 8).

On the other hand, gloves that track the user's movements in virtual environments and deliver tactile / kinesthetic feedback to one of the most sensible part of the body in real-time are an ideal tool to support VR / XR interaction. However, existing haptic gloves either restrict the natural motion and maximum output force of the hand or are too bulky and heavy to be worn for extended interaction session (Ma & Ben-Tzvi 2015). To develop reliable experiences using a haptic glove, the design and implementation needs to keep the heft of the glove down and yet maximize the workspace and force output range. Until a few years ago the only reliable force-feedback gloves were the CyberGrasp™ of Immersion Corp (now CyberGlove Systems) (Turner et al. 1998) or the Master II of Rutgers University (Bouzit et al. 2002). These and other similar prototype devices used DC motors, artificial muscles, shape memory alloys or dielectric elastomers to create finger movements

and tactile simulation on the hand. However, as VR / AR interaction has become more mainstream, private companies have started developing their own adaptations of the haptic glove. Companies like HaptX, VRgluv and Tesla have developed complex exoskeleton-based force feedback devices that are reliable and manageable for extended interaction session.

These devices boost five finger interaction and wireless solutions with enhanced degrees of freedom (5-9 DoF) and can concurrently provide actuation and tracking. The exoskeleton approach ensures that sufficient force feedback can be generated for a natural and immersive experience in most VR / XR environments. However, as with most cutting-edge technologies, these devices have some limitations. Firstly, most of these devices are work-in-progress and lack reliable driving software to integrate conventional VR / XR environments. Custom haptic encoding and tactile layers need to be added to virtual environments to control each device. Moreover, most of the devices lack the ability to sense environmental and user specific forces being applied during the interaction, which can make them susceptible to overdriving the motor mechanism that provides the force feedback component. Furthermore, even though some of these devices use lightweight alloys (magnesium, fiberglass etc.) the exoskeleton structure means that users still need to wear and manipulate gloves weighing 500g (each) or more which can become difficult during longer interaction sessions. And lastly, as most of these commercial products have not been extensively tested, results on user perception and long-term user experience are still limited. However, the available technical specifications suggest that the current development of haptic gloves have the potential to create more immersive and natural feedback experiences for VR / XR interaction than has been previously available. Table 1 is a collection of current and upcoming devices and their technical specifications to illustrate the development in the area of haptic gloves.

Table 1. A list of currently available haptic gloves and their prototype specification for VR / XR interaction (adapted from Jérôme & Emmanuel, 2018).

Device	Type	Nr fingers	Wireless	Actuator	Force-feedback	Tactile feedback	Hand tracking	Active DoFs	Weight (g)	Published price
Gloveone	Glove	5	yes	Electromagnetic	no	yes	yes	10	na	499 €
AvatarVR	Glove	5	yes	Electromagnetic	no	yes	yes	10	na	1,100 €
Senso Glove	Glove	5	yes	Electromagnetic	no	yes	yes	5	na	$ 599
Cynteract	Glove	5	yes	Electromagnetic	yes	no	yes	5	na	na
Maestro	Glove	5	yes	Electromagnetic	yes	yes	yes	5	590	na
GoTouchVR	Thimble	1	yes	Electromagnetic	no	yes	no	1	20	na
Tactai Touch	Thimble	1	yes	na	no	yes	no	1	29	na
CyberGrasp	Exosk.	5	no	Electromagnetic	yes	no	no	5	450	$ 50,000
Dexmo	Exosk.	5	yes	Electromagnetic	yes	no	no	5	320	$ 12,000
HaptX	Exosk.	5	no	Pneumatic	yes	yes	yes	na	na	na
VRgluv	Exosk.	5	yes	Electromagnetic	yes	no	no	5	na	$ 579
Sense Glove DK1	Exosk.	5	yes	Electromagnetic	no	no	no	5	300	999 €
HGlove	Exosk.	3	no	Electromagnetic	yes	no	no	9	750	30,000 €
Noitom Hi5	Glove	5	yes	Electromagnetic	no	yes	yes	9	105	$ 999
Sensoryx VR Free	Glove	5+5	yes	Electromagnetic	no	yes	yes	10	na	600 €
Dextarobotics Dexmo	Exo-Gloves	5+5	yes	Electromagnetic	yes	yes	yes	11	300	na
Tesla Glove	Exosk.	5	yes	Electromagnetic 3x3 display/finger	yes	yes	yes	9	300	$ 5,000

Multi-device and full-body interaction

Another method of providing tactile and force feedback is to use multiple wearable devices that either interact with each other or communicate with the Interaction Engine (Fig. 8) to provide a comprehensive haptic experience. Some systems utilize this type of feedback (Lylykangas et al. 2015) to create mimicking movements between users while other adaptations (Ramírez-Fernández et al. 2015) are designed for limb motor therapy and to improve performance and generate a lower mental workload. In most cases the systems are untethered to each other and serve as a haptic interface between the *Interaction Engine* and the wearer. Some adaptations of this technology can create large-area stimulation through smart clothing (Lindeman et al. 2004) or through

small puck-like devices known as AHDs (Autonomous Haptic Devices) that can be attached to any part of the body (Farooq et al. 2020). However, these techniques do not provide full-body tracking and interaction which can be very useful in complex VR / XR environments that are designed for complete immersion.

Full-body motion reconstruction and haptic output for VR applications can enable natural interaction and a much higher level of immersion (Kasahara et al. 2017). VR applications which aim to increase the feeling of presence in VR (sense of being) need to be able to track full-body movements of the user and provide real time feedback throughout the body (Jiang et al. 2016). Research by Slater and Wilbur (2016) illustrate that VR / AR immersion requires the entire virtual body whereas presence requires the user to also identify with that virtual body (virtual self-image). In other words, for a high sense of presence, the users must recognize the movements of their virtual body as their movements and be able to sense the interaction in real-time to achieve virtual immersion (Caserman et al. 2020). Recent research is using motion capture technologies to create realistic user models for VR applications however, the tactile and kinesthetic feedback associated with these models is very limited. If the purpose of creating authentic virtual representations of the user is to increase immersion, the lack of haptic feedback can severely hamper this goal.

To solve this issue research into full-body haptic stimulation is being carried out focusing on wearable clothing (Table 2). Although there are a number of wearable tracking solutions (PrioVR, Perception Neuron2.0, Smart-suit Pro, Xsens, etc.), most of them only focus on tracking and sensing full body or joint movements. However, some startups and research labs are stepping in, to develop and promote full-body vibrotactile and kinesthetic feedback solution using electromagnetic and microfluid technologies. Some of the solutions are still work in progress whereas many are extensions of other wearable devices (i.e., gloves). Next is a list of some of these systems that are currently available or publicly shared.

Table 2. A list of currently available full-body haptic suits and their specification for VR / XR interaction.

Device	Coverage	Tracking	Feedback type	Haptic library	Dev. API	HMD platform	Price ($)
Nullspace VR	32 Independent zone (chest, abdomen, shoulders, arms & hands)	Yes	Vibrotactile (via Bluetooth & wired)	Fixed 117 effects with customization effects editor	Unity 3D	Multiplatform with 3rd party tracking	299 (vest only)
Tesla Suit	Fullbody suit, 80 embedded electrostatic channels	Yes, with Biometrics with 10 MC sensors	EM Vibrotactile (via Bluetooth)	Customizable from 1-300Hz, 1-260ms & 0.15amp per channel	Unity 5 & Unreal Engine 4	Multiplatform	13,000 (suit & gloves)
Axon VR / HaptX	Full body suit, gloves, and pads	Yes, magnetic motion tracking	Vibrotactile micro-fluidic technology with force feedback exoskeleton	Customizable temperature, vibration, motion, shape, and texture rendering	Unity 5, Unreal Engine, Steam VR HaptX SDK	Multiplatform	N/A enter-prise solution
Tactsuit (bHaptics)	Full-body suit, gloves, pads, and feet guard upto 70 indivi-dual zones (x16 or x40)	Yes	Vibrotactile actuators (via Bluetooth LE) (comes in 2 versions: x16 & x40)	Customizable	Unity 5, Unreal Engine	Multiplatform	x16=299 X40=499 (pre-order price)
Rapture VR	HMD and Vest by uploadVR	Yes	Vibrotactile	N/A	Unity 5, Unreal Engine	Only for VOID Experience	N/A
Synesthesia Suit	Vest, gloves, & pads with 26 active zones	Yes	Vibrotactile	Customizable (Triggered alongside audio feedback)	PS VR, Unity 5, Unreal Engine	Multiplatform including PlayStation	Under develop ment
Haptika	Gloves, vest, pads	No	Vibrotactile	Customizable	No	Multiplatform	N/A
HoloSuit	Glove, jacket & pants with 40 sensors and 9 actuation elements	Yes	Vibrotactile	Customizable	Unity, Unreal Engine 4 & Motion-Builder	Multiplatform	Under develop ment
Woojer	Vest (with 6 cus-tom actuators 2 on sides, back and front for stereo haptics) and Waist straps (1 actuator)	No	Vibrotactile con-trolled through audio-based signals (1-200Hz range) over TI wireless controller	Customizable (Triggered along-side audio feedback	Used over any audio-based interface	Multiplatform / open	Vest 349 Strap 129
NeoSensory exoskin VR suit	Haptic Jacket, vest and wrist device with 32 actuation motors	No	Vibrotactile feedback with adjustable frequency signals	Customizable	Custom SDK and API for Unity and UE	Multiplatform	400 SDK and develop er package
Shockwave *(Under dev-elopment, Kickstarter)*	Vest, and wear-able straps (on the legs) with 8 zones	Full body tracking (using 8 wireless IMUs)	64-point vibrotactile feedback (HD haptics)	Customizable	Unity, Unreal Engine 4	Compatible with most VR headset (requires dev. support)	300 for kick-starter

Dynamic physical environments

Researcher have also been exploring dynamic physical environments that are tethered to virtual experiences to enhance immersion. Elements of the physical space are either directly connected to virtual experiences or act as an extension to the interaction carried out in the virtual environment by relaying physical forces to virtual actions. Various adaptations of the Haptic Floor (Visell et al. 2009; Bouillot & Seta, 2019) are prime examples of relaying physical forces in connection to virtual environments. Individual segments of the floor are set of a platform that are pivot or vibrate in connection to the virtual environment or user's action to support visual and auditory feedback in virtual environments. Each segment of the floor acts as an individual pixel within the interaction scheme can provide meaningful tactile and kinesthetic information enhancing the overall experience.

Other dynamic environments track the physical and virtual movements of the user and supplement auxiliary support or cues by introducing artificial forces. One adaptation of this is the ZoomWalls (Yixian et al. 2020) which create dynamically adjustable walls that simulate a haptic infrastructure for room-scale VR. Multiple wall segments are mounted on a mobile platform that track and follow the user within the physical space, and if needed orient the users to similar artificial surrounding by simulating walls, doors and walkways corresponding to their virtual interface.

Another such type of environmental haptics feedback is the CoVR system (Bouzbib et al. 2020), which utilizes a robotic interface to provide strong kinesthetic feedback (100 N) in a room-scale VR arena. It consists of a physical column mounted on a 2D Cartesian ceiling robot (XY displacements) with the capacity of resisting to body-scaled users' actions such as pushing or leaning; and acting on the users by pulling or transporting them. The system is also able to carry multiple potentially heavy objects (up to 80 kg) which users can freely manipulate within a joint interaction environment. However, in both cases, virtual and physical tracking plays a crucial role in such applications. There is also a need to have various elements of the movable environment follow the user in real time while the user interacts with the VR environment. This can be an issue where the user may be using an HMD with limited or no visual passthrough capabilities. In such scenarios, users may unintentionally bump into these objects, which can cause injuries. Moreover, larger environments with more than one user interacting with the dynamic space would require far more resources, limiting scalability of such setup.

However, dynamically adjustable environmental interaction is a new research area and novel solution may be adopted in the future that can enhance the usability and efficiency of similar approaches. In fact, researchers from Microsoft have adapted a similar approach to create wearable devices that can create the forces related to interacting with spherical objects. The PIVOT system (Kovacs et al. 2020) is a wrist-worn haptic device that renders virtual objects into the user's hand on demand. Similar to the design premise of dynamic environments utilized by Bouzbib et al. and Yixian et al., PIVOT uses actuated joints that pivots a haptic handle into

and out of the user's hand, rendering the haptic sensations of grasping, catching, or throwing an object anywhere in space. Unlike existing hand-held haptic devices and haptic gloves, PIVOT leaves the user's palm free when not in use, allowing users to make unencumbered use of their hand. PIVOT also enables rendering forces acting on the held virtual objects, such as gravity, inertia, or air-drag, by actively driving its motor while the user is firmly holding the handle. Authors suggest that wearing a PIVOT device on both hands can add haptic feedback to bimanual interaction, such as lifting larger objects for user.

4. Emerging and Future Multimodal UI Prospects

There are many experimental and emerging technologies which are intriguing for interaction purposes, but which are not yet widely used or deployable for XR UIs. In this chapter we present some examples of prospects for the future multimodal XR interfaces. Some of them may become widely used in the future, others maybe not. The time will tell.

4.1 Facial expression interfaces

Facial expressions, eye tracking, emotion recognitions, etc. are elements of human-human communication, albeit usually they happen unconsciously. For example, we may focus on eyes mainly if the eye behaviour is unusual, or if we try to find out if another person is serious or speaking the truth. However, most of the HCI does not take much of these nuances into account. In XR, multimodal gaze and gesture has been utilized in collaboration, for example by Bai et al. (2020).

Facial expression and emotion recognition interfaces are a relatively new concept in VR / XR interaction. HMDs partly block user's face, hence there are specific challenges in the XR context. However, some sensors can be placed inside the HMD, and on the other hand the HMDs are becoming smaller and may ultimately become ultralight smart glasses. In the last few years, researchers as well as the industry have integrated facial recognitional technology into HMDs to track and relay user expression to virtual avatars. Apart from the social applications, this type of technology can be useful in creating immersive reaction between users as well as NPCs without creating complex facial models.

Devices such as Decagear (Decagear 2020) utilize facial tracking and mapping in real-time. Similarly, Tobii has been working with other companies to implement eye tracking seamlessly into VR / XR headsets to improve foveated rendering and create precursors for locomotion, thereby reducing motion sickness.

4.2 Scent

The sense of smell is known as a chemical sense because it relies on chemical transduction. It is more difficult to digitize the sense of smell and scents in HCI context compared to sounds and light (Obrist et al. 2016). Scents have been underrepresented in VR (LaViola et al. 2017). However, technology for enabling scents in XR is advancing rapidly. An increasing body of research shows that scents affect the user in numerous ways. For example, scents can enrich the user experience (LaViola et al. 2017; Munyan et al. 2016), increase immersion (Hopf et al. 2020), sense of reality (Baus et al. 2019) and presence (Ranasinghe et al. 2018), affect emotion, learning, memory and task performance (Tortell et al. 2007), and enhance a training experience in applications such as shopping, entertainment and simulators (Murray et al. 2016; Obrist et al. 2016; Cheok & Karunanayaka 2018).

Different technologies exist for delivering scents to XR. The easiest way to deliver digitized odors to a user is by using ambient scents (Spence et al. 2017). Ambient scent is present in the environment instead of emanating from a specific object (Spangenberg et al. 1996). Ambient scents can be created with various scent-emitting devices placed in a room. However, it is difficult to rapidly change from one scent to another or change the intensity of scent unless the scented space is relatively small similarly to the sensory reality pods by Sensiks Inc. Active directing of the scented air with a fan or air cannon provides a little more control.

Fig. 10 illustrates more precise approaches to deliver scents and Fig. 11 shows related examples of recent scent display prototypes. Scented air can be directed from a remote scent display to an HMD with tubes (Salminen et al. 2018). Alternatively, it is possible to produce more compact scent displays that are attached to a VR controller (Niedenthal et al. 2019), worn on the user's body (Amores et al. 2018; Wang et al. 2020) or connected directly to the HMD (Brooks et al. 2020; Kato & Nakamoto 2019; Narumi et al. 2011; Ranasinghe et al. 2018). The advantages of wearable scent display typically include better spatial, temporal and quantitative control because the scents can be delivered near or in the nostrils.

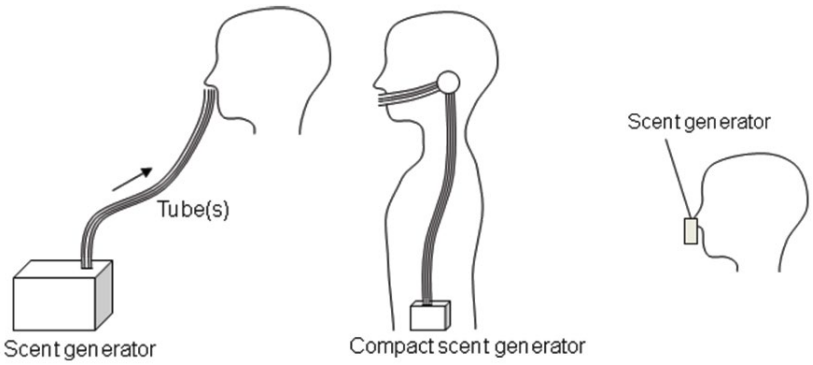

Figure 10. Alternative scent delivery methods suitable for XR (Yanagida, 2012).

Figure 11. Examples of scent displays for XR. Olfactory display attached to an HTC Vive controller (left, Niedenthal et al. 2019), on-face olfactory interfaces (middle, Wang et al. 2020), and an addon for VR mask stimulating the trigeminal nerve in the user's nose with different scents (right, Brooks et al. 2020).

Before scents can be delivered to a user, they must be vaporized from stocked form of odor material. The most typical solutions are natural vaporization, accelerated vaporization by air flow, heating, and atomization (Yanagida 2012). Feelreal, a Kickstarter-funded addon for VR masks, used atomization for releasing up to nine different scents. A limiting factor in all scent displays is the number of possible scents that can be created. The

displays use stocked odor material that typically can present 1-15 scents. In addition, it is often difficult to blend multiple odors. However, recent research indicates that it could be possible to synthesize scents on demand by creating a mixture of odorants that humans perceive similarly to the original scent (Ravia et al. 2020). This is a major step towards technology that digitizes and reproduces scents similarly to what is already possible by recording sounds and taking photographs.

4.3 Taste

Taste (gustation) is also a chemical sense and even less often used than scents, especially in XR (Cheok & Karunanayaka 2018). Taste perception is often a multimodal sensation composing chemical substance, sound, smell, and haptic sensations (Iwata 2008). In fact, taste perception largely originates from the sense of smell (Auvray & Spence 2008) because scents travel through orthonasal (sniff) and retronasal (mouth) airways while eating. Many XR applications such as those aimed for augmenting flavor perception have therefore used scents (Aisala et al. 2020; Narumi et al. 2008) instead of attempting to stimulate the sense of taste directly. It is also possible to develop technology for stimulating specifically the sense of taste, targeting to stimulate one or more of the five basic taste sensations that taste buds can sense: salty, sweet, bitter, sour, and umami.

The three main approaches to creating taste sensations are ingesting chemicals, sensing electrical stimulation of the tongue, and using thermal stimulation (Kerruish 2019). TasteScreen (Maynes-Aminzade 2005) used a questionable method of requiring users to lick a computer screen with a thin layer of flavoring chemical. Vocktail (Ranasinghe et al. 2017) built a cocktail glass with embedded electronics for creating electrical stimulation at the tip of the tongue. Typically, there is significant interpersonal variation in the robustness of taste perception resulting from electrical stimulation. Some users perceive a vague taste, while others report only a metallic sensation (Nakamura & Miyashita 2012; Ranasinghe et al. 2013). Therefore, in many concepts taste stimulation is supported by simultaneous stimulation of other senses to create a more convincing multisensory experience. The last approach, thermal stimulation, was used in Affecting Tumbler (Suzuki et al. 2014), which was a cup designed for changing the flavor perception of a drink by heating the skin around the user's nose. Fig. 12 illustrates the use of thermal and electrical stimulation.

Figure 12. Examples of taste interfaces. Changing flavor perception with thermal stimulation (left, Suzuki et al. 2014) and using electrical stimulation at the tip of the tongue to alter the existing flavors of a beverage (right, Ranasinghe et al. 2017).

Even though initial empirical findings have suggested that the prototypes can alter taste perceptions (e.g., Suzuki et al. 2014), more research is needed. Taste stimulation typically requires other supporting modalities to create applications that are meaningful, function robustly, and are pleasant to use. Compared to other modalities, we are still in the early stages of development for taste (Obrist et al. 2016). However, HMDs and other wearable devices for XR offer a good technological platform for further development.

4.4 Exhalation interfaces

Exhalation interface is also possible, albeit rarely used. It can provide a limitedly controlled hands-free interaction. It is almost always available, even to persons with quadriplegia. Hands-free blowing is useful and quick when the user's hands are preoccupied with another task. Blowing can also be very discreet, as very low sound levels are produced. It can be limitedly controlled in both magnitude and direction.

Blowing has been used e.g., as a communication method for people with disabilities, for health applications, or for speech therapy. Blowing as an interaction method has been used also for VR art, play and entertainment (e.g., Sra et al. 2018; Kusabuka & Indo, 2020). Furthermore, it has been proposed for computer, mobile phone or smartwatch user interfaces (e.g., Chen et al. 2019).

Breathing or blowing air as an interaction method have been proposed also for VR. Anemometers or microphones can be fitted onto a VR headset, thus always remaining optimally located. It can deepen the immersion in games and simulations. As exhalation and inhalation are interaction methods for some real-world interfaces, such as musical instruments, they could be used in simulated training of related skills.

Sra et al. (2018) proposed four breathing actions as a directly controlled input interaction for VR games. As breathing sensors, they explored microphone, temperature sensor, and a BioHarness on chest, which could also detect inhalation actions. Their user study showed that breathing UI was found to give a higher sense of presence and be more fun. They also proposed several design strategies for blowing with games.

Chen et al. (2019) used a headset microphone as a blowing sensor and classified the input into categories to improve the measuring accuracy. The number of interaction types is limited, as people cannot skillfully control many forms of blowing. Their user tests with various applications indicated that blowing improves users' interest and experience, specifically in VR applications.

4.5 Tongue interfaces

Tongue can be used for hands-free control, and it matches well with XR systems. Tongue is a fast and precise organ, comparable to head, hands, or eyes for the purposes of user interaction. Tongue-based interaction is hands-free and very discreet and personal. A thin mouthguard with suitable sensors would make tongue an intriguing alternative interaction approach.

There are many kinds of proposed and experimental tongue-controlled interaction systems. A simple method is to stick the tongue out, point to various directions, and use a camera to sense the movements. A camera

could also observe the cheek skin near mouth. Several approaches and sensor types have been experimented with (e.g., an array of textile pressure sensors attached to the user's cheek, a wearable around the ear which reads brain signals and tongue muscle signals, EMG signals detected at the underside of the jaw, or intraoral electrode array or capacitive touch sensors attached to a mouthguard). Fig. 13 illustrates some of them.

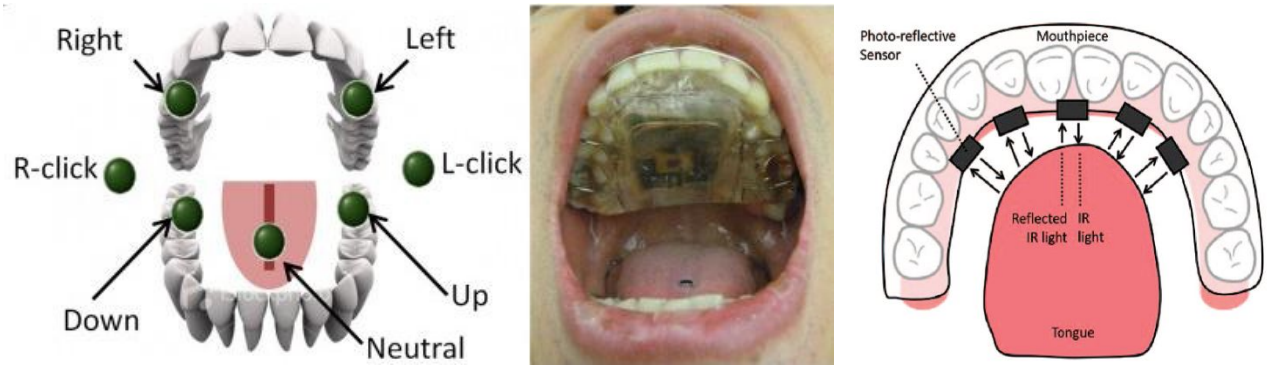

Figure 13. Examples of tongue interfaces. Left/middle: mouth-mounted buttons for using a mouse. Right: TongueInput (Hashimoto et al. 2018) measures tongue motion using photo-reflective sensors.

Tongue UI is a promising general interaction method, but currently it is used mostly as an assistive technology for patients suffering from speech impairments and full-body paralysis.

4.6 Brain-computer interface

Many of potential future user interfaces are based on bioelectrical signals in tissues, organs or the nervous system that can be measured, or feedback given through bioelectrical signals. These include electroencephalogram (EEG), electrocardiogram (ECG), electromyogram (EMG), electrooculography (EOG), mechanomyogram (MMG), magnetoencephalogram (MEG) and galvanic skin response (GSR).

The ultimate interface would be a “mind-reading” direct link between user’s thoughts and a computer. Brain-computer interface (BCI) is two-way (input and output) communication between brain and a device, unlike one-way neuromodulation. BCI is not a sense in itself, but it bypasses all human sensors and nerves and stimulates directly and non-invasively the brain with various signals in order to create synthetic sensations. Feeding visual, auditory, haptic, taste, smell, or other sensations directly to the brain could open up entirely new avenues for XR, but this is presumably still far in the future. To illustrate the potential, there is an intriguing sci-fi movie Brainstorm (1983), https://www.youtube.com/watch?v=cOGAEAJ4xJE&t=1385 describing the neuromodulation and BCI sensory feeding.

BCI input can be based on surgically implanted prostheses or on external, non-invasive devices such as EEG sensors (Abiri et al. 2019). There are several non-invasive neuroimaging methods, such as electroencephalography (EEG), functional magnetic resonance imaging (fMRI), and functional near-infrared spectroscopy (fNIRS). EEG is the most widely used for VR currently. Several non-invasive commercial devices can read

human brain activities (e.g., Emotiv, MindWave, Neuroware, Open BCI, Brain Co, Neurosity, iDun, Paradromics, Looxidlabs, or NeuroSky). They use it as input to perform actions with computers or other devices. At least Looxid Labs already sells EEG headsets that can be retrofitted to HMDs.

The feedback (output) can be given through brain stimulation using various methods. Transcranial magnetic stimulation (TMS) and transcranial focused ultrasound stimulation (tFUS, TUS) are some possible non-invasive input methods. TMS has been used e.g., for helping a blindfolded user to navigate a 2D computer game only with direct brain stimulation (Losey et al. 2016). tFUS has superior spatial resolution and the ability to reach deep brain areas. TMS, tFUS (Lee et al. 2016) and other brain stimulation methods can also elicit synthetic visual perception (phosphenes) when given onto visual cortex, even though it is very coarse with current methods.

HMDs can have various physiological sensors close to the skin, eyes, and skull. The PhysioHMD system (see Fig. 14) merges several types of biometric sensors to an HMD and collects sEMG, EEG, EDA, ECG and eye-tracking data (Bernal et al. 2018). EEglass (Vourvopoulos et al. 2019) is a prototype of an HMD employing EEG for BCI. Luong et al. (2020) estimate the mental workload of VR applications in real-time with the aid of physiological sensors embedded in the HMD. Barde et al. (2020) have made a review on recording and employing the user's neural activity in virtual environments.

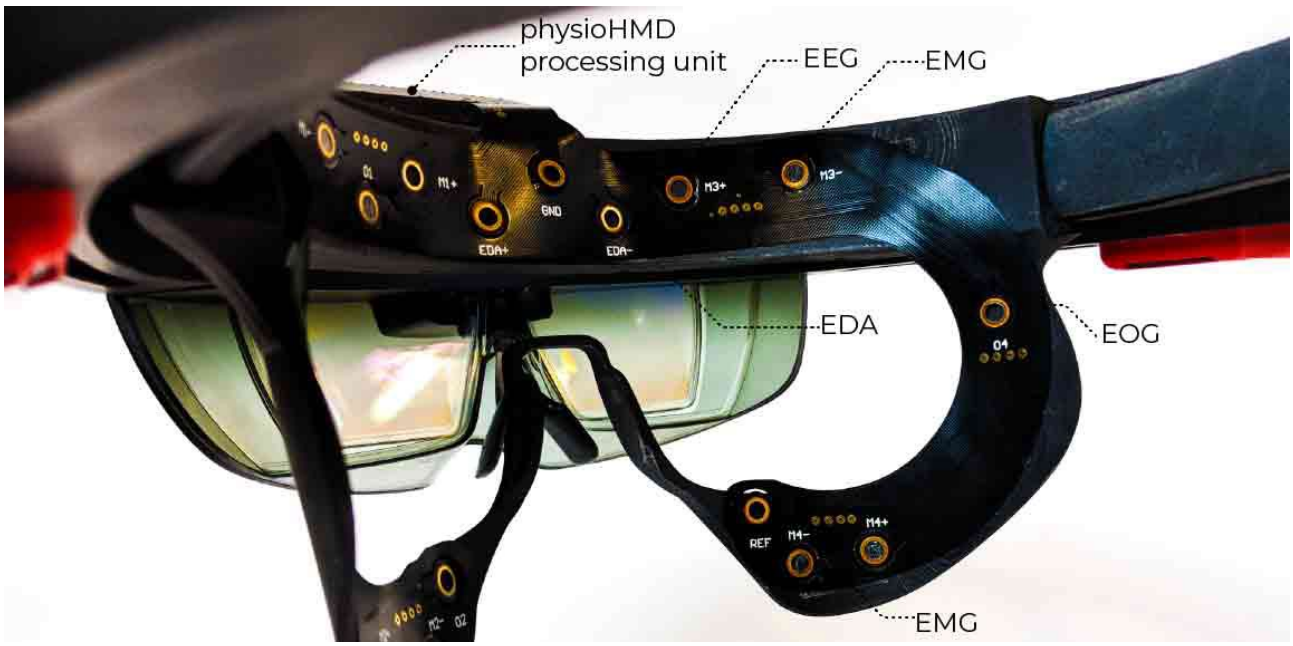

Figure 14. PhysioHMD records physiological data such as EMG, EEG, EDA, ECG through the contact with the skin (Bernal et al. 2018).

Elon Musk's Neuralink Inc. (http://www.neuralink.com/) demonstrated recently Gertrude, a pig with a coin-sized computer chip implant. Human experiments are due soon, and they intend to do a variety of things, e.g., to solve ailments such as memory loss, hearing loss, depression and insomnia or restoring some movement to people with quadriplegia. Ultimately, they hope to fuse humankind with artificial intelligence.

BCI-XR research is high risk, high reward work. Potentially it is a very disruptive technology, and closely related to Augmented Human. However, the few conducted experiments on creating realistic synthetic sensations have very lame results so far. On the other hand, input through EEG or EMG has gained better results.

BCI is still very limited in its capabilities, and it is used mostly for special purposes such as implanted aid for paralyzed people or to prevent tremors caused by Parkinson's disease. One practical line of research is to create synthetic vision for blind people. Recently machine learning has helped to e.g., classify mental or emotional states.

4.7 EMG and other biometric signal interfaces

Electromyography (EMG) reads the electrical activity from muscles and EOG reads specifically the electrical activity of the muscles near the eyes. This can be used for input in XR systems. Some EMG sensors can be attached to user's muscles wearables, or datasuits (see Fig. 15).

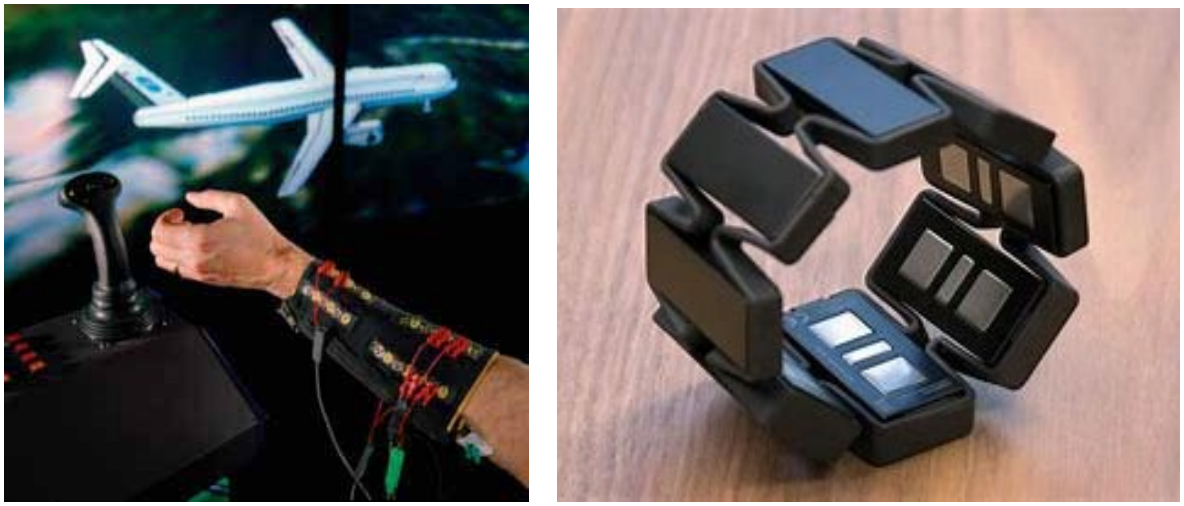

Figure 15. Examples of EMG interfaces. Left: Reading EMG from arm muscles to steer an airplane simulation. Right: Thalmic labs' Myo armband.

Some HMDs have embedded EMG and EOG physiological sensors. The beforementioned PhysioHMD system (Bernal et al. 2018) merges several types of biometric sensors to an HMD, including EMG. Barde et al. (2020) have made a review on recording the neural activity for their use in virtual environments.

Bioelectrical signals can be used also for feedback. For example, Sra et al. (2019) have added proprioceptive feedback to VR experiences using galvanic vestibular stimulation.

4.8 Other potentially relevant technologies for HMDs

There are many emerging or disruptive technologies and sciences, which are not directly interaction methods per se, but which may have immense implications for XR interaction and HMD technology. This in part will feed back to these sciences and other fields of applications, including industry, business, training, education, etc.

Improving displays, foveated rendering, and tracking will naturally affect XR interaction. Furthermore, some of the clearly impactful technologies for XR are improving GPUs and/or practically infinite processing power through supercomputers or quantum computers together with very fast networking and 5G mobile networks, as rendering can then happen on the cloud. This could enable for example immersive, fully photorealistic remote work or teleconferencing. They could enable being together with other people or visiting remote places while the experience would be almost indistinguishable from reality. Another benefit of low-latency networks and cloud rendering is that the HMDs can be relatively simple, low-cost, and very lightweight, or they might even become contact lens displays (e.g., Mojo Lens, https://www.mojo.vision/). Contact lens displays would even enable to view information or watch movies with eyes shut. Qualcomm Snapdragon XR2 5G Platform enables cloud rendering, numerous sensors, fast XR hardware and software development, etc.

Other potentially relevant technologies for XR interaction and HMDs include AI, XR chatbots (agents), battery technology, nanotechnology, miniaturization of sensors and actuators, IoT, robots, new materials, flexible sensor and/or actuator patches on skin or on earbuds, smart contact lenses, distributed ledgers, social media and social gaming in virtual environments (c.f., Facebook, Second Life), phosphenes with neuromodulation, etc. Some of these challenges and opportunities are discussed in more length e.g., by Spicer et al. (2017).

XR can make paintings come alive, e.g., the Dreams of Dali (https://www.youtube.com/watch?v=zQ2-oJOkTKc). AI can restore historical images and videos (e.g., https://www.youtube.com/watch?v=2s_hIs8s_N4). AI could also transfer them to virtual environments, and even become a time machine, by reconstructing historical scenes and people to become alive again. 3D scanning is developing fast, and it is easy to scan physical environments (e.g., Leica BLK360, Matterport), but using only 2D historical (coarse) image materials are a challenging source of information. To the best of our knowledge, historical images or videos have not yet been transformed to virtual environments in any meaningful way.

Biometric methods are not multimodal interaction methods in the strict sense, but e.g., iris scanning would be easy to embed to an HMD and fingerprint reading for data gloves, and thus personalize and authorize selected content. On the other hand, biometric and other sensing technologies have also ethical and privacy concerns. For example, unscrupulous companies, or criminals such as authoritarian governments or mafia organizations could spy on people and exploit them in multiple ways. And if they can, they will. As Mark Pesce (2021) puts it: "*The concern here is obvious: When it [Facebook Aria] comes to market in a few years, these glasses will transform their users into data-gathering minions for Facebook. Tens, then hundreds of millions of these AR spectacles will be mapping the contours of the world, along with all of its people, pets, possessions, and peccadilloes. The prospect of such intensive surveillance at planetary scale poses some tough questions about who will be doing all this watching and why*".

In addition to technologies, also other issues will have an impact on future XR technology, usage and applications – directly or indirectly. Social trends, cultural issues, economy, business, politics, geopolitics, demographics, pandemics, etc. will alter sentiment, prosperity, innovation, and many other things, and those things will have indirect impact on the development and usage of XR.

What else? Think outside the box! What are the implications of new technologies applied to XR interaction?

4.9 Augmented human

Augmented human (Farooq & Grudin 2016; Raisamo et al. 2019; Mueller et al. 2020) is a new paradigm to extend human abilities and senses fluently through technology. In a way, the Augmented human paradigm is like multimodal interaction on steroids, not using just a handful of modalities but a vast number of advanced sensor and actuator technologies which generate a large volume of data, but which is presented in concise and coherent manner. Its processing requires e.g., machine learning, signal processing, computer vision, and statistics. It is no more just interfacing with a device, but integrating with it (e.g., Hainich 2009).

Augmented human technologies provide new, smart, and stunning experiences in unobtrusive ways. Light-weight, comfortable and yet efficient augmentation technology can be very useful and have a significant impact on various human activities. For a person requiring assistance in living due to deteriorated vision, smart glasses can enhance visual information and turn it into speech. The glasses can also augment cognition and support memory by fetching answers to spoken questions. Special clothes can provide augmented skin that senses the touches and movements assisted by a therapist and integrates them with a training program stored in the smart glasses. Embedded sensors in the clothes can also notice imbalance in movements and save information of physical reactions so that the therapy instructions can be adapted accordingly. Physical augmentation is also possible, e.g., with lightweight exoskeletons or robotic prostheses, which amplify the user's physical strength or endurance.

5. Conclusions

Multimodal interaction could revolutionize the use of computers and phones, and applications such as CAD, data visualization, digital signage, tele-presence, home automation, and entertainment. Multimodal interaction for XR is essential for the usability of the virtual and augmented environments. Interaction methods from the PC desktop context are not usually effective. 3D user interfaces have been developed for many purposes, and they match the multidimensional and multisensory VR and AR worlds better. However, new concepts, paradigms and metaphors are needed for advanced interaction and for novel and emerging hardware.

Most of the multimodal interaction technologies are still immature for universal use. All the current approaches for multimodal interaction have their strengths and weaknesses. Also, no single approach is likely to dominate. The applied technology will largely depend on the context and application. Research is going on in many universities and companies.

Perfectly seamless and fluid interaction in HCI is not always required, or even possible. Good enough is good enough. Even though the human perceptual system is delicately and meticulously designed, it has some perceptual shortcomings which can be taken advantage of. Many tricks and approximations can be used to create satisfactory multimodal interaction for general use. This is also one of the tenets of the HUMOR project. Even shortcut technologies can make the audiences believe they are seeing magic. On the other hand, the technologies must fit for human perception and cognition. As stated by Gardony et al. (2020) "*When done poorly, MR experiences frustrate users both mentally and physically leading to cognitive load, divided attention, ocular fatigue, and even nausea. It is easy to blame poor or immature technology but often a general lack of understanding of human perception, human-computer interaction, design principles, and the needs of real users underlie poor MR experiences. ..., if the perceptual and cognitive capacities of the human user are not fully considered then technological advancement alone will not ensure MR technologies proliferate across society.*".

Any product must be reasonably priced for the purpose, and it must meet a demand. Even if something is cool, few want to pay a lot of money for it. For any technology to penetrate the markets, there are many non-technical issues to consider, such as revenue, marketing, early adaptors, consumer needs, demand and acceptance, IPR, backward compatibility, price, timing, manufacturability, luck, etc. These will have a paramount impact on the emergence of XR systems with multimodal interaction.

Multimodal XR interaction in 2040?

How will a typical XR system look like in 10 or 20 years from now? Or does it even make any sense to talk about XR systems, in the same vein as multimedia PC is an obsolete term nowadays, as all the PCs are that? Will we be wearing our 6G Flex-Communicator in our pocket, on hand or near the eyes, depending on the

context? How can it bring added value to our lives and augment and assist us in our daily routines and special moments?

Can XR systems in 2040 immerse all of the user's senses? Near-perfect visuals and audio are easier to implement, but there will be grand challenges and probably insurmountable obstacles to produce e.g., effective haptic, locomotion or gustation systems, which do not encumber the user. Only if neuromodulation or BCI will take giant steps in the future, it might be possible. Yet again, this is task-, context- and cost-dependent. Is unencumbered interaction even necessary? People use daily all kinds of instruments, devices and tools for various tasks in real life, so why not for XR? In any case, some kind of a display is needed in XR. Furthermore, the camera-based Kinect gesture sensor was unencumbered, but it never became a permanent success story. For some purposes, a sufficiently immersive XR system may be possible and useful. For some AR systems, immersion may not even be desirable. And XR has also many social, societal, and other issues to be solved (for example, see Fig. 16).

Figure 16. Will XR become a new addiction and a form of escapism, among other things?

6. References

Abiri, R., Borhani, S., Sellers, E., Jiang, Y., & Zhao, X. (2019). A comprehensive review of EEG-based brain–computer interface paradigms. Journal of neural engineering, 16(1), 011001.

Aisala, H., Rantala, J., Vanhatalo, S., Nikinmaa, M., Pennanen, K., Raisamo, R., & Sözer, N. (2020). Augmentation of Perceived Sweetness in Sugar Reduced Cakes by Local Odor Display. In Companion Publication of the 2020 International Conference on Multimodal Interaction (ICMI '20 Companion), 6 pages.

Akkil, D., Kangas, J., Rantala, J., Isokoski, P., Špakov, O., & Raisamo, R. (2015). Glance awareness and gaze interaction in smartwatches. In Proc. CHI 2015 Conf. on Human Factors in Computer Syst. (pp. 1271-1276).

Alam, M., Samad, M., Vidyaratne, L., Glandon, A., & Iftekharuddin, K. (2020). Survey on deep neural networks in speech and vision systems. Neurocomputing, 417, 302-321.

Allman-Farinelli, M., Ijaz, K., Tran, H., Pallotta, H., Ramos, S., Liu, J., Wellard-Cole, L., & Calvo, R. (2019). A Virtual Reality Food Court to Study Meal Choices in Youth: Design and Assessment of Usability. JMIR Form Res 2019;3(1):e12456. DOI: 10.2196/12456

Amores, J., Hernandez, J., Dementyev, A., Wang, X., & Maes, P. (2018). BioEssence: A Wearable Olfactory Display that Monitors Cardio-respiratory Information to Support Mental Wellbeing. In 2018 40th Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society (EMBC) (Vol. 2018-July, pp. 5131–5134).

Auvray, M., & Spence, C. (2008). The multisensory perception of flavor. Consciousness and Cognition, 17(3), 1016–1031.

Bai, H., Sasikumar, P., Yang, J., & Billinghurst, M. (2020). A User Study on Mixed Reality Remote Collaboration with Eye Gaze and Hand Gesture Sharing. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (pp. 1-13).

Barde, A., Gumilar, I., Hayati, A., Dey, A., Lee, G., & Billinghurst, M. (2020). A Review of Hyperscanning and Its Use in Virtual Environments. In Informatics 7(4), 55. Multidisciplinary Digital Publishing Institute.

Baus, O., Bouchard, S., & Nolet, K. (2019). Exposure to a pleasant odour may increase the sense of reality, but not the sense of presence or realism. Behaviour & Information Technology, 38(12), 1369–1378.

Beddiar, D., Nini, B., Sabokrou, M., & Hadid, A. (2020). Vision-based human activity recognition: a survey. Multimedia Tools and Applications, 79(41), 30509-30555.

Benzie, P., Watson, J., Surman, P., Rakkolainen, I., Hopf, K., Urey, H., ... & Von Kopylow, C. (2007). A survey of 3DTV displays: techniques and technologies. *IEEE transactions on circuits and systems for video technology*, *17*(11), 1647-1658.

Bergström, J., & Hornbæk, K. (2019). Human-Computer Interaction on the Skin. ACM Computing Surveys (CSUR), 52(4), 1-14.

Bermejo, C., & Hui, P. (2017). A survey on haptic technologies for mobile augmented reality. arXiv preprint arXiv:1709.00698.

Bernal, G., Yang, T., Jain, A., & Maes, P. (2018). PhysioHMD: a conformable, modular toolkit for collecting physiological data from head-mounted displays. In Proceedings of the 2018 ACM International Symposium on Wearable Computers (pp. 160-167).

Bernsen N. (2008) Multimodality Theory. In: Tzovaras D. (eds) Multimodal User Interfaces. Signals and Commmunication Technologies. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-78345-9_2

Billinghurst, M., Bowskill, J., Dyer, N., Morphett, J. (1998). Spatial Information Displays on a Wearable Computer. IEEE Computer Graphics and Applications, 18(6), 24–31.

Billinghurst, M., Clark, A., & Lee, G. (2015). A Survey of Augmented Reality. Foundations and Trends® in Human–Computer Interaction, 8(2-3), 73-272.

Bimber, O., & Raskar, R. (2005). Spatial Augmented Reality: merging real and virtual worlds. Peters.

Biswas, S., & Visell, Y. (2019). Emerging material technologies for haptics. Advanced Materials Technologies, 4(4), 1900042.

Boem, A., & Troiano, G. M. (2019). Non-Rigid HCI: A Review of Deformable Interfaces and Input. In Proceedings of the 2019 on Designing Interactive Systems Conference (pp. 885-906).

Boletsis, C. (2017). The new era of virtual reality locomotion: A systematic literature review of techniques and a proposed typology. *Multimodal Technologies and Interaction*, *1*(4), 24.

Bolt, R. (1980). "Put-that-there": voice and gesture at the graphics interface. ACM Comput. Graphic. 14 (3), 262–270.

Boraston, Z., & Blakemore, S. J. (2007). The application of eye-tracking technology in the study of autism. The Journal of physiology, 581(3), 893-898.

Bouillot, N., & Seta, M. (2019). A Scalable Haptic Floor Dedicated to Large Immersive Spaces. In Proceedings of the 17th Linux Audio Conference (LAC-19) At: CCRMA, Stanford University (USA).

Bouzbib, E., Bailly, G., Haliyo, S., & Frey, P. (2020). CoVR: A Large-Scale Force-Feedback Robotic Interface for Non-Deterministic Scenarios in VR. In Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology (UIST'20) (pp. 209–222). DOI:https://doi.org/10.1145/3379337.3415891

Bouzit, M., Popescu, G., Burdea, G., & Boian, R. (2002). The Rutgers Master II-ND Force Feedback Glove, Proceedings of IEEE VR 2002 Haptics Symposium.

Brazil, A., Conci, A., Clua, E., Bittencourt, L., Baruque, L., & da Silva, N. (2018). Haptic Forces and Gamification on Epidural Anesthesia Skill Gain. Entertainment Computing, 25, pp. 1-13.

Brooks, J., Nagels, S., & Lopes, P. (2020). Trigeminal-based Temperature Illusions. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (pp. 1–12).

Burova, A., Mäkelä, J., Hakulinen, J., Keskinen, T., Heinonen, H., Siltanen, S., & Turunen, M. (2020). Utilizing VR and Gaze Tracking to Develop AR Solutions for Industrial Maintenance. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (pp. 1-13).

Cardoso, J. (2019). A Review of Technologies for Gestural Interaction in Virtual Reality. Recent perspectives on gesture and multimodality. Cambridge Scholars Publishing.

Carter, T., Seah, S. A., Long, B., Drinkwater, B., & Subramanian, S. (2013). UltraHaptics: multi-point mid-air haptic feedback for touch surfaces. In Proceedings of the 26th annual ACM symposium on User interface software and technology (pp. 505-514).

Caserman, P., Garcia-Agundez, A., & Göbel S. (2020). A Survey of Full-Body Motion Reconstruction in Immersive Virtual Reality Applications, IEEE Transactions on Visualization and Computer Graphics, 26(10), 3089-3108, doi: 10.1109/TVCG.2019.2912607.

Chen, Y., Bian, Y., Yang, C., Bao, X., Wang, Y., De Melo, G., Liu J., Gai W., Wang, L. & Meng, X. (2019). Leveraging Blowing as a Directly Controlled Interface. In 2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (pp. 419-424). IEEE.

Chen, Z., Li, J., Hua, Y., Shen, R., & Basu, A. (2017). Multimodal interaction in augmented reality. In 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC) (pp. 206-209). IEEE.

Chen, W., Yu, C., Tu, C., Lyu, Z., Tang, J., Ou, S., Fu, Y., & Xue, Z. (2020). A Survey on Hand Pose Estimation with Wearable Sensors and Computer-Vision-Based Methods. Sensors, 20, 1074.

Cheng, H., Yang, L., & Liu, Z. (2015). Survey on 3D hand gesture recognition. IEEE transactions on circuits and systems for video technology, 26(9), 1659-1673.

Cheok, A., & Karunanayaka, K. (2018). Virtual taste and smell technologies for multisensory internet and virtual reality. Cham: Springer.

Choi, S., & Kuchenbecker, K. J. (2012). Vibrotactile display: Perception, technology, and applications. Proceedings of the IEEE, 101(9), 2093-2104.

Chuah, J., & Lok, B. (2012). Experiences in Using a Smartphone as a Virtual Reality Interaction Device. International Journal of Virtual Reality, 11(3), 25-31. https://doi.org/10.20870/IJVR.2012.11.3.2848

Clay, V., König, P., & Koenig, S. (2019). Eye Tracking in Virtual Reality. Journal of Eye Movement Research. 12(1).

Cronin, S., & Doherty, G. (2019). Touchless computer interfaces in hospitals: A review. Health informatics journal, 25(4), 1325-1342.

Culbertson, H., Schorr, S., & Okamura, A. (2018). Haptics: The present and future of artificial touch sensation. Annual Review of Control, Robotics, and Autonomous Systems, 1, 385-409.

DecaGear. (2020). Megadodo simulation games pte. Ltd. https://www.deca.net/decagear/ (accessed 15.12.2020).

DeCarlo, C. (1968). Computers. Toward the Year 2018, Foreign Policy Association (ed.), Cowles Educational Corporation, New York 1968.

DeLucia, P., Preddy, D., Derby, P., Tharanathan, A., & Putrevu, S. (2014). Eye movement behavior during confusion toward a method, In Proc. Human Factors and Ergonomics Society 58th Ann. Meet., 58, (pp. 1300-1304).

Desai, A., Peña-Castillo, L., & Meruvia-Pastor, O. (2017). A Window to Your Smartphone: Exploring Interaction and Communication in Immersive VR with Augmented Virtuality. In Proceedings of 14th Conference on Computer and Robot Vision (CRV), (pp. 217-224), doi: 10.1109/CRV.2017.16.

Drewes, H., & Schmidt, A. (2007). Interacting with the computer using gaze gestures. In Proceedings of the INTERACT 2007 Conference on Human-Computer Interaction. Berlin, Germany: Springer.

Drèze, X. & Hussherr, F. (2003). Internet advertising: Is anybody watching? Journal of interactive marketing, 17(4), pp.8-23.

Duchowski, A. (2007). Eye tracking methodology: Theory and practice. London, UK: SpringerVerlag.

Engelbart, D. (1968). A demonstration at AFIPS Fall Joint Computer Conference, San Francisco, CA, Dec. 9, 1968.

Esmaeili, S., Benda, B., & Ragan, E. D. (2020). Detection of Scaled Hand Interactions in Virtual Reality: The Effects of Motion Direction and Task Complexity. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)* (pp. 453-462). IEEE.

Esteves, A., Velloso, E., Bulling, A., & Gellersen, H. (2015). Orbits: Gaze interaction for smart watches using smooth pursuit eye movements. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology* (pp. 457-466).

Farooq, A., Weitz, P., Evreinov, G., Raisamo, R., Takahata, D. (2016). Touchscreen Overlay Augmented with the Stick-Slip Phenomenon to Generate Kinetic Energy. In Adjunct Proceedings of ACM User Interface Software Technology (UIST '16), pp. 179-180, http://dx.doi.org/10.1145/2984751.2984758.

Farooq, A., Coe, P., Evreinov, G., & Raisamo, R. (2020). Using Dynamic Real-Time Haptic Mediation in VR and AR Environments. In: Ahram, T., Taiar, R., Colson, S., Choplin, A. (eds) Human Interaction and Emerging Technologies. IHIET 2019. Advances in Intelligent Systems and Computing, vol. 1018. Springer, Cham.

Farooq, U., & Grudin, J. (2016). Human-computer integration. Interactions, 23(6), 26-32.

Feiner, S., MacIntyre, B., Höllerer, T., & Webster, A. (1997). A touring machine: Prototyping 3D mobile augmented reality systems for exploring the urban environment. Personal Technologies, 1(4), 208-217.

Flavián, C., Ibáñez-Sánchez, S., & Orús, C. (2021). The influence of scent on virtual reality experiences: The role of aroma-content congruence. Journal of Business Research, 123(October 2020), 289–301.

Freeman, E., Wilson, G., Vo, D., Ng, A., Politis, I., & Brewster, S. (2017). Multimodal feedback in HCI: haptics, non-speech audio, and their applications. In The Handbook of Multimodal-Multisensor Interfaces: Foundations, User Modeling, and Common Modality Combinations-Volume 1 (pp. 277-317).

Furumoto, T., Fujiwara, M., Makino, Y., & Shinoda, H. (2019). BaLuna: Floating Balloon Screen Manipulated Using Ultrasound. In 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), (pp. 937-938).

Gardony, A., Lindeman, R., & Brunyé, T. (2020). Eye-tracking for human-centered mixed reality: promises and challenges. In Optical Architectures for Displays and Sensing in Augmented, Virtual, and Mixed Reality (AR, VR, MR) (Vol. 11310, p. 113100T). International Society for Optics and Photonics.

Hainich, R. (2009). The End of hardware: augmented reality and beyond. BookSurge.

Hamza-Lup, F., Bergeron, K., & Newton D. (2019). Haptic Systems in User Interfaces: State of the Art Survey. In Proceedings of the 2019 ACM Southeast Conference (ACM SE '19), 141–148.

Hansen, J., Hansen, D., Johansen, A., & Elvesjö, J. (2005). Mainstreaming gaze interaction towards a mass market for the benefit of all. In C. Stephanidis (Ed.) Universal Access in HCI: Exploring New Interaction Environments, Vol. 7 (Proc. HCIII 2005). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Hansen, D., & Ji, Q. (2010). In the eye of the beholder: A survey of models for eyes and gaze. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(3), 478–500.

Harley, D., Verni, A., Willis, M., Ng, A., Bozzo, L., & Mazalek, A. (2018). Sensory VR: Smelling, touching, and eating virtual reality. In *Proceedings of the Twelfth International Conference on Tangible, Embedded, and Embodied Interaction* (pp. 386-397).

Harrison, C., Tan, D., & Morris, D. (2010). Skinput: appropriating the body as an input surface. In Proc. of the SIGCHI conference on human factors in computing systems (pp. 453-462).

Havlík T., Hayashi D., Fujita K., Takashima K., Lindeman RW. & Kitamura Y. (2019) JumpinVR: Enhancing jump experience in a limited physical space. In SIGGRAPH Asia 2019 XR, 19-20.

Heilig, M. (1962). U.S. Patent No. 3,050,870. Washington, DC: U.S. Patent and Trademark Office.

Henrikson, R., Grossman, T., Trowbridge, S., Wigdor, D., & Benko, H. (2020). Head-Coupled Kinematic Template Matching: A Prediction Model for Ray Pointing in VR. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1-14).

Holland, C., & Komogortsev, O. (2012). Eye tracking on unmodified common tablets: Challenges and solutions. In Proceedings of the ETRA 2012 Symposium on Eye Tracking Research and Applications (277-280).

Holoride GmbH. Holoride: Virtual Reality meets the real world (2019). https://www.audi.com/en/experience-audi/mobility-and-trends/digitalization/holoride-virtual-reality-meets-the-real-world.html. (accessed 15.12.2020).

Hopf, J., Scholl, M., Neuhofer, B., & Egger, R. (2020). Exploring the Impact of Multisensory VR on Travel Recommendation: A Presence Perspective. In *Information and Communication Technologies in Tourism 2020* (pp. 169-180). Springer, Cham.

Hoshi, T., Abe, D., & Shinoda, H. (2009). Adding tactile reaction to hologram. In RO-MAN 2009-The 18th IEEE International Symposium on Robot and Human Interactive Communication (pp. 7-11). IEEE.

Hoshi, T., Takahashi, M., Iwamoto, T., & Shinoda, H. (2010). Noncontact tactile display based on radiation pressure of airborne ultrasound. IEEE Transactions on Haptics, 3(3), 155-165.

HP. (2020). HP Reverb G2, https://www8.hp.com/us/en/vr/reverb-g2-vr-headset.html. (accessed 15.12.2020).

Hutchinson, T., White, K., Martin, W., Reichert, K., & Frey, L. (1989). Human-computer interaction using eye-gaze input. IEEE Transactions on Systems, Man, and Cybernetics, 19(6), 1527-1534.

Hyrskykari, A., Majaranta, P. & Räihä, K-J. (2005) From Gaze Control to Attentive Interfaces. In Proceedings of HCII 2005.

Hyrskykari, A., Istance, H., & Vickers, S. (2012). Gaze gestures or dwell-based interaction? In Proceedings of the Symposium on Eye Tracking Research and Applications (ETRA '12). (pp. 229–232).

Isokoski, P., Kangas, J., & Majaranta, P. (2018). Useful approaches to exploratory analysis of gaze data: enhanced heatmaps, cluster maps, and transition maps. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications* (pp. 1-9).

Istance, H., Hyrskykari, A., Immonen, L., Mansikkamaa, S., & Vickers, S. (2010). Designing gaze gestures for gaming: an investigation of performance. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications* (pp. 323-330).

Iwamoto, T., Tatezono, M., & Shinoda, H. (2008). Non-contact method for producing tactile sensation using airborne ultrasound. In International Conference on Human Haptic Sensing and Touch Enabled Computer Applications, 504-513. Springer, Berlin, Heidelberg.

Iwata, H. (2008). Taste interfaces. HCI Beyond the GUI: Design for Haptic, Speech, Olfactory, and Other Nontraditional Interfaces", Edited by Kortum, P., Elsevier Inc., USA.

Jerald, J. (2015). The VR book: Human-centered design for virtual reality. Morgan & Claypool.

Jérôme, P., & Emmanuel, V. (2018). Touching Virtual Reality: a Review of Haptic Gloves. ACM SIGGRAPH 2019 Emerging Technologies.

Jiang, F., Yang, X., & Feng, L. (2016). Real-time full-body motion reconstruction and recognition for off-the-shelf VR devices. In Proceedings of 15th ACM SIGGRAPH Conf. Virtual-Reality Continuum Appl. Ind. - Volume 1, 2016, pp. 309–318.

Kaluschke, M., Weller, R., Zachmann, G., Pelliccia, L., Lorenz, M., Klimant, P., Knopp, S., Atze, J., & Mockel, F. (2018). A Virtual Hip Replacement Surgery Simulator with Realistic Haptic Feedback. 2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), pp. 759-760.

Karafotias, G., Korres, G., Sefo, D., Boomer, P., & Eid, M. (2017). Towards a Realistic Haptic-based Dental Simulation, 2017 IEEE Int. Symp. on Haptic, Audio and Visual Environments and Games (HAVE).

Kasahara, S., Konno, K., Owaki, R., Nishi, T., Takeshita, A., Ito, T., Kasuga, S., & Ushiba, J. (2017). Malleable embodiment: Changing sense of embodiment by spatial-temporal deformation of virtual human body. In Proceedings of CHI Conf. Human Factors Comput. Syst., 2017, pp. 6438–6448.

Kato, S., & Nakamoto, T. (2019). Wearable Olfactory Display with Less Residual Odor. In 2019 IEEE International Symposium on Olfaction and Electronic Nose (ISOEN) (pp. 1–3). IEEE.

Kaul, O., & Rohs, M. (2017). Haptichead: A spherical vibrotactile grid around the head for 3D guidance in virtual and augmented reality. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (pp. 3729-3740).

Kervegant, C., Raymond, F., Graeff, D., & Castet, J. (2017). Touch hologram in mid-air. In ACM SIGGRAPH 2017 Emerging Technologies (pp. 1-2).

Kerruish, E. (2019). Arranging sensations: smell and taste in augmented and virtual reality. The Senses and Society, 14(1), 31–45.

Koutsabasis P., & Vogiatzidakis P. (2019). Empirical Research in Mid-Air Interaction: A Systematic Review, Int. J. Human–Computer Interaction, 35:18, 1747-1768.

Kovacs, R., Ofek, E., Gonzalez Franco, M., Siu, A., Marwecki, S., Holz, C., & Sinclair, M. (2020). Haptic PIVOT: On-Demand Handhelds in VR. In Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology (pp. 1046-1059).

Kowalczyk, P., & Sawicki, D. (2019). Blink and wink detection as a control tool in multimodal interaction. Multimedia Tools and Applications, 78(10), 13749-13765.

Krueger, M., Gionfriddo, T., & Hinrichsen, K. (1985). VIDEOPLACE—an artificial reality. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 35-40).

Kusabuka, T., & Indo, T. (2020). IBUKI: Gesture Input Method Based on Breathing. In *Adjunct Publication of the 33rd Annual ACM Symposium on User Interface Software and Technology* (pp. 102-104).

Kölsch, M., Bane, R., Hollerer, T., & Turk, M. (2006). Multimodal interaction with a wearable augmented reality system. IEEE Computer Graphics and Applications, 26(3), 62-71.

LaViola Jr, J., Kruijff, E., McMahan, R., Bowman, D., & Poupyrev, I. (2017). 3D user interfaces: theory and practice. Addison-Wesley Professional.

Lee, W., Kim, H., Jung, Y., Chung, Y., Song, I., Lee, J., & Yoo, S. (2016). Transcranial focused ultrasound stimulation of human primary visual cortex. Scientific reports, 6(1), 1-12.

Lee, L., & Hui, P. (2018). Interaction methods for smart glasses: A survey. IEEE access, 6, 28712-28732.

Leftwich, J. (1993). InfoSpace: A Conceptual Method of Interacting with Information in a Three-Dimensional Virtual Environment. In Proceedings of the third International Conference on Cyberspace.

Li, Y., Huang, J., Tian, F., Wang, H., & Dai, G. (2019). Gesture interaction in virtual reality. Virtual Reality & Intelligent Hardware, 1(1), 84-112.

Li, G., Mcgill, M., Brewster, S., & Pollick, F. (2020). A Review of Electrostimulation-based Cybersickness Mitigations. IEEE the 3rd International Conference on Artificial Intelligence & Virtual Reality.

Li, N., Han, T., Tian, F., Huang, J., Sun, M., Irani, P., & Alexander, J. (2020). Get a Grip: Evaluating Grip Gestures for VR Input using a Lightweight Pen. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1-13).

Lindeman, R., Page, R., Yanagida, Y., & Sibert, J. (2004). Towards full-body haptic feedback: The design and deployment of a spatialized vibrotactile feedback system. In Proc. ACM Symposium on Virtual Reality Software and Technology (VRST). 146-149. 10.1145/1077534.1077562.

Long, B., Seah, S. A., Carter, T., & Subramanian, S. (2014). Rendering volumetric haptic shapes in mid-air using ultrasound. ACM Transactions on Graphics (TOG), 33(6), 1-10.

Losey, D., Stocco, A., Abernethy, J., & Rao, R. (2016). Navigating a 2D virtual world using direct brain stimulation. Frontiers in Robotics and AI, 3, 72.

Luong, T., Martin, N., Raison, A., Argelaguet, F., Diverrez, J., & Lécuyer, A. (2020). Towards Real-Time Recognition of Users' Mental Workload Using Integrated Physiological Sensors Into a VR HMD. In IEEE International Symposium on Mixed and Augmented Reality (ISMAR) (13p).

Lylykangas, J., Heikkinen, J., Surakka, V., Raisamo, R., Myllymaa, K., & Laitinen, A. (2015). Vibrotactile stimulation as an instructor for mimicry-based physical exercise. *Advances in Human-Computer Interaction*, *2015*.

Ma, Z., & Ben-Tzvi, P. (2015). Design and optimization of a five-finger haptic glove mechanism. *Journal of Mechanisms and Robotics*, *7*(4).

Majaranta, P. & Räihä, K. J. (2002). Twenty years of eye typing: systems and design issues. In *Proceedings of the 2002 symposium on Eye tracking research & applications* (pp. 15-22).

Majaranta, P., Ahola, U. K., & Špakov, O. (2009). Fast gaze typing with an adjustable dwell time. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 357-360).

Majaranta, P., & Bulling, A. (2014). Eye tracking and eye-based human-computer interaction. In S. Fairclough & K. Gilleade (Eds.), Advances in physiological computing (pp. 39–65). Berlin, Germany: Springer.

Mann, S. (1997). Wearable computing: a first step toward personal imaging. IEEE Computer, 30(2), 25–32.

Marshall, S. (2007). Identifying cognitive state from eye metrics, Aviat. Space Envir. MD., 78(5), pp. B165-B175

Martinez, J., Griffiths, D., Biscione, V., Georgiou, O., & Carter, T. (2018). Touchless haptic feedback for supernatural vr experiences. In 2018 IEEE Conf. on Virtual Reality and 3D User Interfaces (VR), 629-630.

Matsubayashi, A., Makino, Y., & Shinoda, H. (2019). Direct finger manipulation of 3d object image with ultrasound haptic feedback. In Proc. 2019 CHI Conf. on Human Factors in Computing Systems (pp. 1-11).

Mayer, S., Reinhardt, J., Schweigert, R., Jelke, B., Schwind, V., Wolf, K., & Henze, N. (2020). Improving Humans' Ability to Interpret Deictic Gestures in Virtual Reality. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20). (pp. 1–14).

Maynes-Aminzade, D. (2005). Edible Bits: Seamless Interfaces between People, Data and Food. In CHI 2005 Extended Abstracts, 2207–2210.

Meißner, M., Pfeiffer, J., Pfeiffer, T. & Oppewal, H. (2019). Combining virtual reality and mobile eye tracking to provide a naturalistic experimental environment for shopper research. Journal of Business Research, 100, 445-458.

Mewes, A., Hensen, B., Wacker, F., & Hansen, C. (2017). Touchless interaction with software in interventional radiology and surgery: a systematic literature review. *International journal of computer assisted radiology and surgery*, *12*(2), 291-305.

Miccini, R., & Spagnol, S. (2020). HRTF individualization using deep learning. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)* (pp. 390-395). IEEE.

Milgram, P., & Kishino, F. (1994). A taxonomy of mixed reality visual displays. IEICE Transactions on Information and Systems, 77(12), 1321-1329.

Mueller, F., Lopes, P., Strohmeier, P., Ju, W., Seim, C., Weigel, M., ... & Nishida, J. (2020). Next Steps for Human-Computer Integration. In Proc. 2020 CHI Conf. on Human Factors in Computing Systems (pp. 1-15).

Munyan, B. G., Neer, S. M., Beidel, D. C., & Jentsch, F. (2016). Olfactory Stimuli Increase Presence in Virtual Environments. PLOS ONE, 11(6), e0157568.

Murray, N., Lee, B., Qiao, Y., & Muntean, G. (2016). Olfaction-enhanced multimedia: A survey of application domains, displays, and research challenges. ACM Computing Surveys (CSUR), 48(4), 1-34.

Nakagaki, K., Fitzgerald, D., Ma, Z., Vink, L., Levine, D., & Ishii, H. (2019). inFORCE: Bi-directional `Force' Shape Display For Haptic Interaction. In Proc. 13th International Conf. on Tangible, Embedded, and Embodied Interaction (TEI '19) (pp. 615-623).

Nakamura, H., & Miyashita, H. (2012). Development and evaluation of interactive system for synchronizing electric taste and visual content. In Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12 (p. 517).

Narumi, T., Nishizaka, S., Kajinami, T., Tanikawa, T., & Hirose, M. (2011, May). Augmented reality flavors: gustatory display based on edible marker and cross-modal interaction. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 93-102).

Niedenthal, S., Lunden, P., Ehrndal, M., & Olofsson, J. (2019). A Handheld Olfactory Display For Smell-Enabled VR Games. In 2019 IEEE Int. Symp. on Olfaction and Electronic Nose (ISOEN) (pp. 1–4). IEEE.

Nigay, L., & Coutaz, J. (1993). A design space for multimodal systems: concurrent processing and data fusion. In Proc. INTERACT'93 and CHI'93 conference on Human factors in computing systems (pp. 172-178).

Nukarinen, T., Kangas, J., Rantala, J., Pakkanen, T., & Raisamo, R. (2018). Hands-free vibrotactile feedback for object selection tasks in virtual reality. In Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology (pp. 1-2).

Nukarinen, T., Kangas, J., Rantala, J., Koskinen, O., & Raisamo, R. (2018). Evaluating ray casting and two gaze-based pointing techniques for object selection in virtual reality. In *Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology* (pp. 1-2).

Obrist, M., Velasco, C., Vi, C., Ranasinghe, N., Israr, A., Cheok, A., ... & Gopalakrishnakone, P. (2016). Sensing the future of HCI: touch, taste, and smell user interfaces. Interactions, 23(5), 40-49.

Oliveira, V., Nedel, L., Maciel, A., & Brayda, L. (2016). Spatial discrimination of vibrotactile stimuli around the head. In 2016 IEEE Haptics Symposium (HAPTICS) (pp. 1-6). IEEE.

Olivier, A., Bruneau, J., Kulpa, R., & Pettré, J. (2017). Walking with virtual people: Evaluation of locomotion interfaces in dynamic environments. IEEE transactions on visualization and computer graphics, 24(7), 2251-2263.

Palovuori, K., Rakkolainen, I., & Sand, A. (2014). Bidirectional touch interaction for immaterial displays. In Proceedings of the 18th International Academic MindTrek Conference: Media Business, Management, Content & Services (pp. 74-76).

Parisay, M., Poullis, C. & Kersten, M. (2020). EyeTAP: A Novel Technique using Voice Inputs to Address the Midas Touch Problem for Gaze-based Interactions. arXiv preprint arXiv:2002.08455.

Pesce, M. 2021. AR's Prying Eyes. IEEE Spectrum.

Piumsomboon, T., Lee, G., Lindeman, R., & Billinghurst, M. (2017). Exploring natural eye-gaze-based interaction for immersive virtual reality. In 2017 IEEE Symp. on 3D User Interfaces (3DUI), 36-39. IEEE.

Poupyrev, I., Billinghurst, M., Weghorst, S., & Ichikawa, T. (1996). The go-go interaction technique: non-linear mapping for direct manipulation in VR. In *Proceedings of the 9th annual ACM symposium on User interface software and technology* (pp. 79-80).

Qian, J., Ma, J., Li, X., Attal, B., Lai, H., Tompkin, J., ... & Huang, J. (2019). Portal-ble: Intuitive free-hand manipulation in unbounded smartphone-based augmented reality. In Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology (pp. 133-145).

Raisamo, R., Rakkolainen, I., Majaranta, P., Salminen, K., Rantala, J., & Farooq, A. (2019). Human augmentation: Past, present and future. International Journal of Human-Computer Studies, 131, 131-143.

Rakkolainen, I., & Palovuori, K. (2002). Walk-thru screen. In Projection Displays VIII (Vol. 4657, 17-22). International Society for Optics and Photonics.

Rakkolainen, I., Sand, A., & Palovuori, K. (2015). Midair User Interfaces Employing Particle Screens. IEEE computer graphics and applications, 35(2), 96-102.

Rakkolainen, I., Raisamo, R., Turk, M., Höllerer, T., & Palovuori, K. (2017). Extreme field-of-view for head-mounted displays. In 2017 3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON) (pp. 1-4). IEEE.

Rakkolainen, I., Freeman, E., Sand, A., Raisamo, R., & Brewster, S. (2020). A Survey of Mid-Air Ultrasound Haptics and Its Applications. IEEE Transactions on Haptics.

Ramírez-Fernández, C., Morán, A., & García-Canseco, E. (2015). Haptic feedback in motor hand virtual therapy increases precision and generates less mental workload. In Proceedings of 9th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth), (pp. 280-286).

Ranasinghe, N., Cheok, A., Nakatsu, R., & Do, E. (2013). Simulating the sensation of taste for immersive experiences. In Proceedings of the 2013 ACM international workshop on Immersive media experiences - ImmersiveMe '13 (pp. 29–34).

Ranasinghe, N., Jain, P., Thi Ngoc Tram, N., Koh, K. C. R., Tolley, D., Karwita, S., ... & Yen, C. C. (2018). Season traveller: Multisensory narration for enhancing the virtual reality experience. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1-13).

Ranasinghe, N., Nguyen, T., Liangkun, Y., Lin, L.-Y., Tolley, D., & Do, E. (2017). Vocktail: A Virtual Cocktail for Pairing Digital Taste, Smell, and Color Sensations. In Proceedings of the 2017 ACM on Multimedia Conference (MM'17) (pp. 1139–1147).

Rantala, J., Majaranta, P., Kangas, J., Isokoski, P., Akkil, D., Špakov, O., & Raisamo, R. (2020). Gaze interaction with vibrotactile feedback: Review and design guidelines. Human–Computer Interaction, 35(1), 1-39.

Rash, C., Russo, M., Letowski, T., & Schmeisser, T. (2009). *Helmet-mounted displays: Sensation, perception and cognition issues*. U.S. Army Aeromedical Research Laboratory, Fort Rucker, AL, USA. ISBN 978-0-615-28375-3. Available at https://apps.dtic.mil/docs/citations/ADA522022.

Rautaray, S., & Agrawal, A. (2015). Vision based hand gesture recognition for human computer interaction: a survey. Artificial intelligence review, 43(1), 1-54.

Ravia, A., Snitz, K., Honigstein, D., Finkel, M., Zirler, R., Perl, O., … & Sobel, N. (2020). A measure of smell enables the creation of olfactory metamers. Nature, (December 2018).

Raza, A., Hassan, W., Ogay, T., Hwang, I., & Jeon, S. (2019). Perceptually Correct Haptic Rendering in Mid-Air using Ultrasound Phased Array. IEEE Transactions on Industrial Electronics, 67(1), 736-745.

Rekimoto, J., & Nagao, K. (1995). The world through the computer: Computer augmented interaction with real world environments. In Proc. 8th ACM Symp. on User interface and software technology (pp. 29-36).

Ren, D., Goldschwendt, T., Chang, Y., & Höllerer, T. (2016). Evaluating wide-field-of-view augmented reality with mixed reality simulation. In 2016 IEEE Virtual Reality (VR) (pp. 93-102). IEEE.

Rozado, D., Moreno, T., San Agustin, J., Rodriguez, F. B., & Varona, P. (2015). Controlling a smartphone using gaze gestures as the input mechanism. Human–Computer Interaction, 30, 34–63.

Rutten, I., Frier, W., Van den Bogaert, L., & Geerts, D. (2019). Invisible touch: How identifiable are mid-air haptic shapes? In Ext. Abstracts of the 2019 CHI Conf. on Human Factors in Computing Systems, pp. 1-6.

Ryu, J., Jung, J., Park, G., & Choi, S. (2010). Psychophysical model for vibrotactile rendering in mobile devices. Presence: Teleoperators and Virtual Environments, 19(4), 364-387.

Salminen, K., Rantala, J., Isokoski, P., Lehtonen, M., Müller, P., Karjalainen, M., ... & Telembeci, A. (2018). Olfactory display prototype for presenting and sensing authentic and synthetic odors. In Proceedings of the 20th ACM International Conference on Multimodal Interaction (pp. 73-77).

Sand, A., Rakkolainen, I., Isokoski, P., Kangas, J., Raisamo, R., & Palovuori, K. (2015). Head-mounted display with mid-air tactile feedback. In Proceedings of the 21st ACM Symposium on Virtual Reality Software and Technology (VRST) (pp. 51-58).

Schmalstieg, D., & Hollerer, T. (2016). Augmented reality: principles and practice. Addison-Wesley Professional.

Schweigert. S., Schwind, V., & Mayer, S. (2019). EyePointing: A Gaze-Based Selection Technique. In Proc. Mensch und Computer 2019 (MuC'19) (pp. 719–723).

Sidenmark, L., Clarke, C., Zhang, X., Phu, J., & Gellersen, H. (2020). Outline Pursuits: Gaze-assisted Selection of Occluded Objects in Virtual Reality. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1-13).

Sims, S,D. & Conati, C. (2020). A Neural Architecture for Detecting User Confusion in Eye-tracking Data. In Proc. 2020 International Conference on Multimodal Interaction (ICMI '20). pp. 15–23.

Slater, M., & Wilbur S. (1997). A Framework for Immersive Virtual Environments (FIVE): Speculations on the role of presence in virtual environments, Presence: Teleoperators Virtual Environments, vol. 6, no. 6, pp. 603–616, 1997.

Spangenberg, E., Crowley, A., & Henderson, P. (1996). Improving the Store Environment: Do Olfactory Cues Affect Evaluations and Behaviors? Journal of Marketing, 60(2), 67.

Spence, C., Obrist, M., Velasco, C., & Ranasinghe, N. (2017). Digitizing the chemical senses: Possibilities & pitfalls. International Journal of Human Computer Studies, 107(June), 62–74.

Spicer, R., Russell, S., & Rosenberg, E.S. (2017). The mixed reality of things: emerging challenges for human-information interaction. In Next-Generation Analyst V (Vol. 10207, p. 102070A). International Society for Optics and Photonics.

Sra, M., Xu, X., & Maes, P. (2018). Breathvr: Leveraging breathing as a directly controlled interface for virtual reality games. In Proc. 2018 CHI Conference on Human Factors in Computing Systems (pp. 1-12).

Sra, M., Jain, A., & Maes, P. (2019). Adding Proprioceptive Feedback to Virtual Reality Experiences Using Galvanic Vestibular Stimulation. In Proc. 2019 CHI Conf. Human Factors in Computing Systems (pp. 1-14).

Starner, T., Mann, S., Rhodes, B., Levine, J., Healey, J., Kirsch, D., Picard, R., & Pentland, A. (1997). Augmented Reality Through Wearable Computing. Presence, 6(4), 386-398.

Stelick, A., Penano, A., Riak, A., & Dando, R. (2018), Dynamic Context Sensory Testing–A Proof of Concept Study Bringing Virtual Reality to the Sensory Booth. Journal of Food Science, 83: 2047-2051. doi:10.1111/1750-3841.14275

Sutherland, I. (1968). A Head-mounted Three-dimensional Display. Proceedings of Fall Joint Computer Conference 1968, AFIPS Conf. Proceedings, Thompson Books, Washington, D.C., Vol. 3, pp. 757-764.

Suzuki, C., Narumi, T., Tanikawa, T., & Hirose, M. (2014). Affecting tumbler. In Proceedings of the 11th Conference on Advances in Computer Entertainment Technology - ACE '14 (pp. 1–10).

Tobii gaming (2020). https://gaming.tobii.com/ (accessed 15.12.2020).

Tobii VR (2020). https://vr.tobii.com/ (accessed 15.12.2020).

Tortell, R., Luigi, D., Dozois, A., Bouchard, S., Morie, J., & Ilan, D. (2007). The effects of scent and game play experience on memory of a virtual environment. Virtual Reality, 11(1), 61-68.

Turk, M. (2014). Multimodal interaction: A review. Pattern Recognition Letters, 36, 189-195.

Turner, M., Gomez, D., Tremblay, M., & Cutkosky, M. (1998). Preliminary tests of an arm-grounded haptic feedback device in telemanipulation. In Proc. ASME Dyn. Syst. Control Div., 1998, vol. DSC-64, 145–149.

Van Dam, A. (1997). Post-wimp user interfaces. Communications of the ACM, 40(2):63-67.

Van Krevelen, D., & Poelman, R. (2010). A survey of augmented reality technologies, applications and limitations. International journal of virtual reality, 9(2), 1-20.

Van Neer, P., Volker, A., Berkhoff, A., Schrama, T., Akkerman, H., van Breemen, A., ... & Gelinck, G. (2019). Development of a flexible large-area array based on printed polymer transducers for mid-air haptic feedback. In Proc. Meetings on Acoustics ICU (Vol. 38, No. 1, p. 045008). Acoustical Society of America.

Varjo Eye Tracking in VR (2020). https://varjo.com/blog/how-to-do-eye-tracking-studies-in-virtual-reality/ (accessed 15.12.2020).

Vidal, M., Bulling, A., & Gellersen, H. (2013). Pursuits: Spontaneous interaction with displays based on smooth pursuit eye movement and moving targets. In Proceedings of the UbiComp 2013 Conference on Pervasive and Ubiquitous Computing (pp. 439-448).

Visell, Y., Law, A., & Cooperstock, J. (2009). Touch Is Everywhere: Floor Surfaces as Ambient Haptic Interfaces. In IEEE Transactions on Haptics, 2(3), 148-159, 2009, doi: 10.1109/TOH.2009.31.

Vourvopoulos, A., Niforatos, E., & Giannakos, M. (2019). EEGlass: An EEG-eyeware prototype for ubiquitous brain-computer interaction. In Adjunct Proc. 2019 ACM International Joint Conf. on Pervasive and Ubiquitous Computing and Proc. 2019 ACM International Symp. on Wearable Computers (pp. 647-652).

Vuletic, T., Duffy, A., Hay, L., McTeague, C., Campbell, G., & Grealy, M. (2019). Systematic literature review of hand gestures used in human computer interaction interfaces. International Journal of Human-Computer Studies, 129, 74-94.

Wang, Y., Amores, J., & Maes, P. (2020). On-Face Olfactory Interfaces. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1-9).

Wang, D., Yuan, G., Shiyi, L. I. U., Zhang, Y., Weiliang, X., & Jing, X. (2019). Haptic display for virtual reality: progress and challenges. Virtual Reality & Intelligent Hardware, 1(2), 136-162.

Wang, D., Ohnishi, K., & Xu, W. (2019). Multimodal haptic display for virtual reality: A survey. IEEE Transactions on Industrial Electronics, 67(1), 610-623.

Weiser, M. (1993). Some computer science issues in ubiquitous computing. Comm. ACM, 36(7), 75-84.

Wilson, G., Carter, T., Subramanian, S., & Brewster, S. (2014). Perception of ultrasonic haptic feedback on the hand: localisation and apparent motion. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 1133-1142).

Wolf, M., Trentsios, P., Kubatzki, N., Urbanietz C., & Enzner, G. (2020). Implementing Continuous-Azimuth Binaural Sound in Unity 3D. In IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW), (pp. 384-389), doi: 10.1109/VRW50115.2020.00083.

Yanagida, Y. (2012). A survey of olfactory displays: Making and delivering scents. In *SENSORS, 2012 IEEE* (pp. 1-4). IEEE.

Yixian, Y., Takashima, K., Tang, A., Tanno, T., Fujita, K., & Kitamura, Y. (2020). ZoomWalls: Dynamic Walls that Simulate Haptic Infrastructure for Room-scale VR World. In Proc. 33rd ACM Symposium on User Interface Software and Technology (UIST'20) (pp. 223–235).

Yoshino, K., Hasegawa, K., & Shinoda, H. (2012). Measuring visio-tactile threshold for visio-tactile projector. In 2012 Proceedings of SICE Annual Conference (SICE) (pp. 1996-2000). IEEE.

Yoshino, K., & Shinoda, H. (2013). Visio-acoustic screen for contactless touch interface with tactile sensation. In 2013 World Haptics Conference (WHC) (pp. 419-423). IEEE.

Yu, X., Xie, Z., Yu, Y., Lee, J., Vazquez-Guardado, A., Luan, H., ... & Ji, B. (2019). Skin-integrated wireless haptic interfaces for virtual and augmented reality. Nature, 575(7783), 473-479.

Zhang, L., Li, X.-Y., Huang, W., Liu, K., Zong, S., Jian, X., … Liu, Y. (2014). It starts with iGaze: Visual attention driven networking with smart glasses. In Proceedings of the MobiCom 2014 Conference on Mobile Computing and Networking (pp. 91-102).

Zhang, Y., Bulling, A., & Gellersen, H. (2013). Sideways: A gaze interface for spontaneous interaction with situated displays. In Proceedings of the CHI 2013 Conference on Human Factors in Computer Systems.