*"University of Turku Technical Reports, No.9 — August 2015"*

# NOVEL INTERACTION TECHNIQUES FOR MOBILE AUGMENTED REALITY APPLICATIONS

## A Systematic Literature Review

Lauri Härkänen | Seppo Helle | Lauri Järvenpää | Teijo Lehtonen

Lauri Härkänen
University of Turku, Technology Research Center, 20014 Turun yliopisto, Finland
lauri.harkanen@utu.fi

Seppo Helle
University of Turku, Technology Research Center, 20014 Turun yliopisto, Finland
seppo.helle@utu.fi

Lauri Järvenpää
University of Turku, Technology Research Center, 20014 Turun yliopisto, Finland
lauri.jarvenpaa@utu.fi

Teijo Lehtonen
University of Turku, Technology Research Center, 20014 Turun yliopisto, Finland
teijo.lehtonen@utu.fi

www.trc.utu.fi

ISSN 2341-8028 | ISBN 978-951-29-6214-3

Technology Research Center

Turun yliopisto
University of Turku

**Abstract**

This study reviews the research on interaction techniques and methods that could be applied in mobile augmented reality scenarios. The review is focused on the most recent advances and considers especially the use of head-mounted displays. In the review process, we have followed a systematic approach, which makes the review transparent, repeatable, and less prone to human errors than if it was conducted in a more traditional manner. The main research subjects covered in the review are head orientation and gaze-tracking, gestures and body part-tracking, and multimodality – as far as the subjects are related to human-computer interaction. Besides these, also a number of other areas of interest will be discussed.

**Keywords**

human-computer interaction, review, mobile, wearable, head-mounted, see-through, near-eye, HMD, NED, gestures, multimodality, tracking, body part-tracking, eye-tracking, gaze-tracking, augmented reality, mixed reality, virtual reality, AR, VR, MR

# Contents

# 1 Introduction

In the recent years, advances in ICT technology have resulted in a quick increase in the types and amounts of mobile devices. It is more than likely that both the variety and the availability of wearable devices are going to grow, following a similar steep curve in the coming years. Smart watches are an early example of this trend. In addition, we now have functional head-mounted (HMD) or near-eye (NED) displays that can be used for both virtual and augmented reality applications – granted that none of these devices are yet even near the goal of having high enough resolution and the field of view to match the human eye. Regarding commercial products, probably the most advanced non-see-through HMD is Oculus Rift DK2, which is still under development (www.oculus.com). There are also various see-through HMDs, of which it is difficult to name the technical leader. See for example www.epson.com and www.vuzix.com. Recently, Microsoft revealed that they as well are developing a see-through HMD, called "hololens" (www.microsoft.com/microsoft-hololens).

These new wearable devices require that we rethink the way how the user interacts with the computer. The traditional mouse and keyboard interface cannot be considered at all with mobile devices. The touch screen is a potential solution to many interaction tasks. However, it would be beneficial to find interaction methods that do not require any handheld tools, but rather where the users could interact with the computer directly with their body.

In this review, we analyze the recent research on potential techniques for interacting with augmented reality applications, with a specific focus on the methods that could be used together with head-mounted displays. The structure of this article is three-fold. First, we will discuss both the methodology and the conduction of the review process. The next chapter holds the analysis of the various articles that we chose to discuss in detail. In this chapter, we will first take a look at the articles that consider head orientation and gaze. After that, we shall discuss various techniques of tracking body parts and using human gestures for interacting with computers in AR contexts. Finally, we will review the more marginal interaction techniques. In the last chapter of this review, we will briefly discuss a topic that is common to many articles discussed previously, and which some articles even hold as their main focus, i.e. multimodality of interaction methods.

# 2 Systematic Literature Review

This article follows the main principles of a systematic literature review, proposed by Kitchenham & Charters[2]. The results of the searches have been filtered, categorized, and analyzed with the selected criteria so that each phase can be back-tracked and – at least in principle – repeated by other researchers. The filtering, evaluation, and reading of the articles was committed by two persons. The optimal case would be that at least two persons would read the article in question, and if their opinions differ, a third person would give his or her sentence. However, due to the large amount of the articles, we were unable to fully follow the guideline for a systematic literature review, according to which each paper should be evaluated by multiple persons. It is acknowledged by the authors that this increases the chance of human misjudgments regarding the selection of relevant articles.

The purpose of this review is to bring together and to analyze the most relevant articles regarding the interaction techniques for HMD-based AR applications. We keep our focus on the very recent publications while also taking into account some significant older studies. During the filtering phase, we decided to delimit the review to the last decade. In the end, the oldest papers that we read through are dated to 2006.

The research questions of this review were: 1. What new control/interaction methods have been proposed for AR/VR/MR systems? 2. What promising interaction methods have emerged to be used with head-mounted displays? 3. Are there any novel control devices that could be used with HMDs and if there are, then how could they improve the user experience?

These questions were kepth in mind both when forming the queries and during the whole review process. The final phrasing of the query was following:

*("augmented reality" OR "mixed reality" OR "virtual reality") AND ("head-mounted" OR "head-worn" OR "glasses" OR "see-through") AND (controller OR control OR input OR interaction OR experience OR interface)*

All the literature queries were conducted in November 2014 using the following search engines: IEEExplore, ACM, Web of Science, Science Direct, Wiley Online Library, CiteSeerX, and Springer Link. We tried to delimit the queries to title-abstract-keyword fields only. Unfortunately, in the case of Springer Link, we were unable to do this. In order to keep the amount of results reasonable, we had to slightly modify the phrasing. After some testing, we found out that omitting the phrases "augmented reality" and "virtual reality", and keeping only "mixed reality" in the query, gave results that we managed to handle. Due to this, it is possible that some relevant articles were not found from Springer Link.

The exact numbers of the query results were: ACM 2998, IEEE 492, Web of Science 650, Wiley 35, Citeseer 265, Springer 898, and Science Direct 73. The total count of the results was thus 5411. All the results were imported into a spread sheet – mostly manually but partly with an aid of small parser scripts. Next, the results were manually evaluated by one person, chiefly based on the title but in some cases also on the abstract. With this method, the amount of articles was reduced to 1145. The remaining articles were sorted by the publication year and then evaluated by another person with the same criteria. In this phase, we decided to delimit the research only to the last ten years, which approximately halved the amount of the articles. After the second evaluation, 196 articles remained for more profound assessments. All the remaining articles were then analyzed based on abstracts and conclusions. For the full reading phase, we chose 91 articles, of which majority was found via ACM Search. More than half of the articles were later filtered out in the reading phase, because in the closer scrutiny their content was not deemed relevant enough. The final amount of the articles discussed in this study is therefore 41. The full list of the articles can be found at the end of this paper.

After the articles were chosen for reading, they were roughly categorized. The biggest category (n=26) we labeled as "mobile, general or mixed", of which only five articles were directly related to HMDs. After this, the most prominent category was "optical tracking of body parts" (n=19). Other two significant categories were "head orientation or gaze" (n=10) and "wearable, gesture-related"(n=8). Besides these four themes, the topics ranged from auditory interaction and feedback to projection-based systems. Some of the papers even discussed more exotic subjects like haptics, volumetric displays, and jamming (mouldable) interfaces. In addition, many of the articles in question consider the topic of multimodality, which means combining multiple user interaction techniques together so that they can be used together for the same end.

## 2.1 Previous Reviews

One of the most notable reviews of AR-related research was written by Zhou et al. [P41]. The authors covered 276 papers that were presented in ISMAR proceedings between the years 1998 and 2008. Besides analyzing the content of the papers, they also took into account the amount of citations of each paper. Interaction techniques was one of the major subjects found in the review. As a topic, only the tracking was more popular than the interaction. Considering the citation counts, the interaction related articles shared the second place together with the calibration and the augmented reality applications.

Most of the interaction studies discussed in Zhou et al. [P41] considered multimodality and gestural interaction. However, only one of the these described a system that comprised an HMD of any kind (See Sandor[3]). It is also worth of noticing that eye-tracking and gaze interaction were not discussed at all, whereas in our review they form one of the major topics. These both findings are directly related to the fact that after 2008 we have seen substantial development in both display technology and mobile devices.

# 3 Interaction Techniques

## 3.1 Head Orientation and Gaze

The question of head orientation as an interaction technique is directly related to pose tracking.[1] In order to augment any virtual content so that it aligns with the physical objects as planned, the camera pose must be known. This is not a trivial task in mobile AR usage, where the position of the camera cannot be known beforehand. In coarse positioning, most common methods utilize GPS (outdoor), RFID, or beacons (indoor). The fine tracking usually relies on computer vision. Assuming that the challenges in pose tracking can be solved, the head orientation information becomes an intriguing possibility for tracking the user's attention.

The head pose alone is not enough, however. In order to provide a complete human-computer interface, we have to combine it with other techniques. Zhang B. et al. [P39] developed a prototype system called "HOBS", where the selectable objects were first coarsely scanned with the head pose information and then the list of selectable objects was refined with different methods. The prototype used infrared (IR) emitters mounted on a Google Glass and receivers attached to the physical objects that were tracked. The refinement techniques were a) naive alphabetical listing, which was handled with the standard touch pad of the Glass, b) sorting the list by signal strength, and c) sensor together with IR-based head motion tracking. In all the techniques, the confirmation was handled with the touch pad. According to their study, the users preferred c) over b) and b) over a). In other words, the users clearly inclined toward more natural methods over the combination of artificial techniques.

Besides moving the head, humans also tend to move their eyes. It is therefore natural that the next step in the quest for tracking the human attention is to track the actual eye movement and to calculate the actual target of the gaze. Eye-tracking as a research subject is already quite old, having roots in the 1940's. For a concise overview on the topic, see Baldauf et al. [P4]. Some commercial products that are able to track the gaze quite accurately are already found on the markets. (See for example www.asleyetracking.com and www.eyetracking-glasses.com).

Not many researchers have attempted to incorporate the eye-tracker into an HMD and apply it in AR context. The "iGaze", described in Zhang L. et al. [P40] is one of the few exceptions. It is a custom-built HMD with an eye-tracker that communicates through Visual Attention-Driven Networking (VAN) with other HMDs and objects that have a VAN device attached to them. According to Zhang L. et al. the VAN protocol can run on top of the existing networking protocols like Wi-Fi. Moreover, it enables the users to communicate directly with each other after establishing the communication link via gaze.

The main weakness of both "iGaze" [P40] and "HOBS" [P39] is the necessity of having physical receivers attached to the objects that are tracked. This might not be a hindrance if we only consider the communication between users, but it is definitely a disadvantage when it comes to the interaction between the user and multiple physical objects. An alternative method is to rely on pose tracking to provide accurate information about the surroundings and to cast rays from a point that represents the center of the vision (which can be either one eye or the middle point between the eyes) to the virtual world augmented over the physical world. This approach has been taken by Schuchert et al. [P29], who developed an HMD prototype for museum applications. Like "HOBS" [P39], it as well had IR sensors mounted inside the HMD.

Another example of a wearable gaze-tracking system is "KIBITZER" [P4]. Unlike the other setups discussed here, it did not embody a display at all but provided only audio feedback via ear phones. The positioning was based on the GPS and the accelerometer of a mobile phone, both mounted on a helmet. The processing was handled by a laptop that was carried in a back bag. The eye-tracking, which was handled with iView X HED apparently worked quite well with the coarse positioning approach, and the researchers reported that the accuracy of the system was "sufficient for determining the POIs (Point of Interest) in the users' gaze". For the current commercial products of the manufacturer, see www.smivision.com/en/gaze-and-eye-tracking-systems/products/overview.htm.

Presuming that the user's view pose can be mapped to the virtual world with a sufficient accuracy, it seems evident that the eye-tracking could already be applied in an AR context with or without HMDs. Tracking the gaze direction, however, might not be enough in order to draw conclusions about the user's attention. In Ajanki et al. [P1], a prototype is described, where the AR system displays relevant information about the physical world by analyzing the user's gaze patterns and adapts the system to the inferred preferences.

It is also possible to combine the tracking data to the information available on the human perception. In Vidal et al. [P36], some examples are given of how to use the data gathered from the gaze tracking to create smarter interfaces. One option is to use the known patterns of eye movement. For example, vestibulo-ocular reflex (VOR) is a subconscious correction motion of the eye. It occurs when the head moves while the gaze is simultaneously kept focused at a target. The VOR does not manifest itself if one looks at a heads-up display (HUD), hence a combination of head movement an the VOR indicates that the user is looking through the display, not at it. Therefore, if the VOR is detected, all the HUD elements could be hidden so that they are not blocking the sight. Similar deductions about the user's attention can be drawn by detecting the vergence of the eyes or the easily detectable patterns that occur when one reads a text. Another aspect is to design the user interface in the way that it adapts to the gaze direction. For example, there is a substantial difference between the foveal and the peripheral visions: unlike the first, the latter is bad at detecting colors but excels in detecting movement.

The gaze is fairly good indicator of the user's attention, but how can we actually interact with the objects we gaze at? The phenomenon of not being able to precisely control what objects the user selects with his or her gaze is called the Midas touch problem. If the user is unable to control what to select, the gaze can hardly be used as a method of interacting with the computer. To overcome the Midas touch issue, one could implement a short delay for indicating a selection. This can be unnatural and slow because pausing is required each time the user wishes to make a selection.

It can also be problematic because humans tend to sometimes gaze at things that they are actually not giving any attention to. Hence, many other solutions have been proposed. With the "KIBITZER" system, Baldauf et al. [P4] employed a prolonged blink of the eye as the selection mechanism, whereas in in Elepfandt & Grund [P8] the gaze was combined to voice commands in order to provide the user the ability to select the object they gaze at.

Yet, what would be more natural than pointing with your eye and clicking with your mind? Schmalstieg et al. [P28] explored this idea by tracking the activation patterns of the brain EEG. They were able to develop a system where the user could access a test system simply by gazing its display. Alas, the brain-computer interface is currently too experimental for any practical AR purposes. Nevertheless, other modalities that aim for the same goal of achieving more natural human-computer interfaces are worth taking a look.

## 3.2 Body Part-Tracking and Gestures

### 3.2.1 Wearable Sensors

Wearable sensors are an extensively studied area in AR/VR interaction research. With wearable sensors and actuators the assumption – or goal – is to avoid causing excessive restrictions to the user's movements and actions. In the case of research prototypes, this goal is often not fulfilled. Yet, it seems possible that this technology could eventually grow into commercial products, with which the sharp edges typical for research prototypes could be honed to fulfill the requirements of unrestricted interaction.

Sensors recognizing hand, arm, and finger movements form a major subcategory of wearable sensors. Various principles can be used for that though. Magnetic sensors together with permanent magnets embedded in gloves is one option, of which the system described in Lakatos et al. [P17] is a fine example. In it, the glove sensors recognize pinching actions between the thumb and fingers. Each finger is mapped to a specific function: index finger selects, middle finger creates a new object, and ring finger deletes. The system also contains a handheld screen, and the relative locations of the hand and screen have different meanings. Above the screen surface, the user's hand controls the spatial 3D parameters. The 3D objects can be modified behind the surface, whereas on the surface, a touch interface can be used for animation, annotations, and scene properties. For tracking, tags in the tablet and 19 fixed cameras in the surroundings were used. This makes it essentially a non-portable system.

Finger movement can also be recognized with computer vision algorithms. Kim et al. [P15] describe a system where an IR camera is worn on the wrist together with IR light sources: one laser and four LEDs. The system is attached to the palm side of the arm, and the field of view covers the hand (palm and fingers) unless the hand is over-arched outwards. Crossed fingers and hand-held objects are also problematic for the sensor. Assuming that the camera and the IR sources are positioned a few centimeters away from the wrist, this kind of device would not be very convenient to wear constantly, even if other technical problems could be solved. The authors provide a fairly extensive list of earlier work about the use of cameras in various places of the body or the head area. Those could offer more convenient solutions from the wearer's point of view, but the risk of occlusions, for example, is even

higher.

Nanayakkara et al. [P23] describe a very unusual, ring-like device called "EyeRing". The device includes a camera is used to recognize objects that are pointed with the finger. The system also uses speech recognition for launcing different applications and functions. The application demonstrations described in the article include a shopping assistant for visually impaired persons and the ability to read aloud banknotes and price tags. Third application demonstration that was brought forth by the authors consider children that could not yet read. With the aid of the application, they were able to interpret a text before actually being able to read it.

The "ShoeSoleSense" described in Matthies et al. [P20] is yet another approach to wearable devices. The device is integrated into the insoles of the shoes, with the intention of leaving the user's hands completely free. The input functions are based on pressure sensors in the soles, recognizing pressure from the big toe, sides of the feet, and the heel. The device recognizes jumping and can give feedback to the user via haptic and temperature signals. This kind of device allows some control without any hand interaction, which may be desirable in many cases. The authors suggest that the concept could be used as a supplementary device, not completely replacing other means of scene manipulation in VR.

Computer vision algorithms require a fairly high amount of processing and data traffic, which means the power requirements cannot be negligible. However, vision-based approaches seem to be an important research and development path, whether the devices will be worn in a finger, on a wrist, or at some other part of the body.

### 3.2.2 Optical Tracking

Optical methods to track body parts are closely related to computer vision, which is employed to understand what the camera perceives. On one hand, many issues in that field are problems also when it comes to tracking body parts. On the other hand, any advances in computer vision can also contribute to the questions of tracking and understanding the body movement.

There are two main approaches in trying to interpret the body movement for the computer. One option is to track the whole body part and reconstruct a 3D representation of it. This approach requires special cameras or sensors, and will be discussed shortly. Another approach is to use a standard RGB camera and to deduce from the 2D image only what is necessary. A fine example of this approach is the system described in Song et al. [P31], where machine learning algorithms were applied for teaching a computer to understand signs based on the training data it had been given. The authors claim that the data is fully customizable without rewriting the code. Moreover, the algorithm should be fast enough to run in a smart watch. The disadvantage of this approach is that it can only be used like a language, which burdens the user with a cognitive overhead for using yet another method that must be learned and remembered.

Tracking an object in 3D provides more information on the tracked object than a 2D tracking method, which allows more complicated gestures to be used in the user interface. In addition, tracking the 3D pose of an object can be used as a starting point for providing free interaction with virtual objects in 3D space, which means that we do not have to interpret gestures for the computer. There are some attempts to achieve this with RGB cameras. Terajima et al. [P34] describe a prototype with a high frame rate camera that is used to track the fingers of the user while he or she types in the air. The authors postulated the pressing gestures in real time from the

2D image, which requires deducing the 3D position of the fingers. Jung et al. [P14] were even able to approximate the 3D pose of multiple body parts from a single 2D image. Their system requires pre-registration of the background, which is used to calculate the depth information of the body parts. This means that their solution will most likely lack any mobile applications.

Currently most solutions for tracking the 3D pose of a body part involve additional sensors or multiple cameras. In Colaço et al. [P7] a Time of Flight (ToF) sensor was used together with a standard RGB camera, both mounted on an HMD. The ToF sensor, built using off-the-shelf hardware, comprised three unfocused, baseline-separated photodiodes and an omnidirectional, pulsed Light-Emitting Diode (LED). With this setup, the researchers were able to achieve robust tracking, which could detect even in-air drawing and writing. The setup is also low-power, light-weight, and relatively unobtrusive. The disadvantage of the system, as noted by the authors, is that the functional area of the ToF sensor is very prone to interference by other objects resembling a hand. Thus, the tracking works properly only with one hand and in the conditions where enough empty space is available.

Another approach for obtaining the 3D pose of an object is to mimic the human perception, which is based on stereographic vision. This method is followed both in Song et al. [P32] and Manders et al. [P19], where stereo cameras are used to enable the user interaction with the virtual content – the former with a finger and the latter with both hands. On one hand, the approach described in Song et al. [P32] suffers from the same issue as the approach in Colaço [P7] because it implements only shape detection. On the other hand, the approach described in Manders et al. [P19] considers only very coarse movements. Although both of these prototypes could be used as a good starting point for building a mobile AR solution, neither of them is directly applicable as such.

Regarding the use of stereo cameras, an encouraging example is described in Akman et al. [P2]. By combining the depth information with color data, the researchers were able to track the full 3D pose of both hands with robustness similar to marker tracking. The prototype was tested with a laptop, which was able to operate at 25–35 fps.

Yet another option is to use a structured light projection-based RGB-D camera to acquire the depth information of the tracked objects. Like ToF but unlike a standard stereo camera, RGB-D is an active device that emits light onto its surroundings. Due to this, it has disadvantages similar to other active methods, being susceptible to interference by other lights and objects, not being able to handle reflective surfaces, and having a limited operation range. The benefit of the approach is a robust tracking performance in close distances. One of the recent applications of this technique is described in Bai et al. [P3], where an RGB-D camera was combined with a Google Glass. The focus of the study was to explore the possibilities of combining gestures to the native touch pad of the Glass. However, their pilot test (n=5) suggested that the combination was deemed as mentally stressful. The test subjects clearly preferred using solely gestures above the combination.

Thus far, there remains the unsolved challenge of accurately tracking the 3D pose of a body part. The positive side is that we merely lack the proper equipment. There are already multiple tracking solutions available. Considering hand-held devices, the main problem is to successfully incorporate either a stereo camera or a depth sensor into the device so that the device is not too heavy to carry and so that it can run the system with decent frame rate and without draining its batteries too soon. As the matter of fact, this is what Google is currently attempting with its Project Tango

(www.google.com/atap/projecttango).

## 3.3 Other Techniques

While doing this review, we encountered many interesting interaction related publications, which may not be at the core of our interest, yet which comprise ideas or practical solutions that could be of use with HMDs in an AR context. Next, we will briefly discuss some of these.

The efforts for providing solutions that leave the user's hands free for interaction with the physical world are well grounded and sound. However, the absence of tangible control devices leads to the issue of not having any haptic feedback. This shortcoming could be overcome with ultrasonic waves, as suggested in Monnai et al. [P21], but we have not yet found any mobile applications of this technique. The haptics can, nevertheless, be embodied into an eyewear and be used together with eye-tracking to provide feedback about the objects the user looks at, as shown in Rantala et al. [P25].

In some situations, employing an interaction device might be inevitable for providing reliable and fluent interaction. It should be noted that the interaction device does not have to be active. For example, in Chakraborty et al. [P6] a simple physical cube was used as a tool to manipulate virtual objects. The cube was optically tracked and it was entirely passive, due to which it did not use any power or have any wires attached to it.

If we use tools in the interaction, we could also mould them according to our wishes. At least this is what Follmer et al. [P9] seem to have thought when they took a technique called particle jamming from the robotics and explored its possible applications in human-computer interfaces. The jamming technique is based on altering the particle stiffness, which makes the object either mouldable or stiffens it in form. With this technique, it would be possible to have one tangible tool that could function differently depending on the context. To give us some examples, the researchers made four prototypes: 1) mouldable clay with adjustable stiffness, 2) a transparent haptic lens that could be used to feel the surface beneath the lens, 3) a shape-deforming haptic interface behind a tablet that leaves the display unobstructed, and 4) a deformable device that could turn into a phone, a remote control, a watch, or a game controller, depending on the shape it is given.

Another appealing idea is based on the notion of substitutive displays, where any surface could suffice as a temporary display. This result can be achieved with mobile projectors (See Sand & Rakkolainen [P27], Song et al. [P30], and Winkler et al. [P38]), which makes the content visible also for others than the actual user of the system. An interesting projection-based implementation is described in Ridel et al. [P26], where the user could point with his or her hand at a predefined area, where the projector would then display the augmented content. Ridel et al. [P26] built the abovementioned system for the purposes of museum exhibition, and it seems that this is a notable option in situations where no pointing device is desired and where the system can be stationary. However, it is also possible to display the image on an HMD so that it appears as it was actually projected on the physical environment. In Nakanishi & Horikoshi [P22] this approach was used together with a ring type microphone in the user's finger (as an input device), with which the user could decide where to position the substitutive display. The more broadly applicable idea here is that it is possible to use the physical environment as an interface to the

virtual content even if the image is actually displayed on an HMD.

Currently the AR solutions on mobile devices are mostly implemented following the "magic lens" metaphor, where the camera is used as an eye and the augmented content is combined to the video stream and then displayed for the user on the screen. The advantage of this approach is that it provides the best possible resolution, saturation, and brightness levels. Hence, the video-see-through (or non-see-through) HMDs provide much better image quality than see-through displays, presuming that the camera is good enough. Especially in HMD use, video-see-through requires that the perspective of the camera viewport is corrected so that it matches with the user's perspective. This can be done, but it requires processing. In Unuma et al. [P35] and Pucichar et al. [P24] a perspective correction with hand-held mobile devices is described. This is not a simple task, since the alignment and the distance between the device and the user's eyes continuously alter when the device is held in hands. With both of these methods, the researchers were able to correct the perspective, so that the video stream displayed on the hand-held device appears correctly aligned with the physical context, as the user perceives it.

In addition to mounting transparent displays on the user's head, it is likewise possible to build tablet computers with transparent displays, as described in two articles by Hincapié et al. [P11][P12]. Considering mobile devices, there are a few advantages in using transparent displays. First, this simplifies the challenges of object tracking, binocular parallax, the registration, and the rendering processes, as noted in Hincapié et al. [P11]. Moreover, it enables forms of use that are difficult or impossible to achieve with opaque displays, like using the tablet for overlaying virtual content directly on top of physical objects when the device is in contact with the object. In Hilliges et al. [P10] a static transparent display was combined with a perspective correction, so that the users were able to interact with virtual objects appearing behind the display naturally with their hands. Even though their system was immobile, there might be some principles here that could be applicable to in mobile use as well.

# 4 Multimodality

Relying on a single input modality can cause problems like the Midas touch phenomenon that was discussed earlier. The user may accidentally do something that will be interpreted as a command, and this can happen more easily if only one modality is used as a trigger. The possibility of unintentional triggers decreases when multiple simultaneous actions are required. However, this requires that all the modalities work seamlessly together. A single modality like hand gestures or speech commands may also be problematic in certain environments and tasks that interfere with the specific modality.

Combining voice to either a gaze or body part-tracking system, was attempted in many articles in this review. For example, in Billinghurst [P5] hand gestures, which were tracked with a stereo camera, were combined to speech input, whereas in Elepfandt & Grund [P8] voice commands were used to overcome the Midas touch issue, as we noted earlier. In Tamaki et al. [P33] audio was used as the main feedback method in a system where an ear-mounted camera was used for fingertip-tracking, since the system did not comprise a display at all. As demonstrated in Tamaki et al. [P33] and the abovementioned "KIBITZER" system [P4], a display is not a required component in all use cases. In AR context, this decision has the obvious advantage of leaving the eyes free to perceive the world.

Besides sound, also haptics can be used to give the user better feedback regarding the information about the virtual environment. In addition to the articles already discussed above, like Monnai et al. [P21] that raised the possibility of using ultrasonic waves for producing haptic feedback, one additional study should be mentioned here, in the context of multimodality. In Lv et al. [P18] a standard mobile phone was used to track the user's foot while he or she plays a football game. In order to achieve haptic feedback from dribbling and kicking, the researchers attached a vibrator to the user's foot. The graphics of the game were displayed on the mobile device, where the user could also use fingers to intercept shots of the other players. The setup might sound awkward, but in the preliminary user studies the multimodal version of the game was deemed substantially more interesting and fun to play than the standard version without being overly difficult to learn.

One of the most exotic systems that we encountered is described in Wang et al. [P37], where multimodal interaction was studied in the context of immersive virtual reality. The user had to manipulate a virtual environment while sitting in a chair. Interaction tools available for the user were a modified Wii controller that was employed like a magic wand and an arm-mounted tablet that gave the user a god-like view of the virtual world. The primary view mode was displayed on a non-occlusive HMD (i.e. non-see-through yet allowing the user to look by it), which provided a first person perspective to the world. The users were able to complete

predefined manipulation tasks better with using both tablet and the wand together with the HMD than by using only a single technique. The only caveat was that both modalities had to be kept synchronized or else the system was considered confusing.

The synchronization of the different modalities was focal also in Irawati et al. [P13], which is one of the oldest papers in our review and which was also discussed in Zhou et al. [P41]. The main innovation of Irawati et al. [P13] was to augment an earlier paddle gesture-based interaction system with a speech interface. Both modalities were interpreted under the same system in order to keep the user commands consistent.

The question of which interaction modalities or combinations of multiple modalities users prefer, was touched in many articles. In many cases, the user tests were rudimentary. Nevertheless, some papers were more dedicated to the user studies than the others. For example in Kollee et al. [P16] the researchers compared gestures with voice and touch interfaces in an office environment, where the users had to move images from a display to another. The interaction techniques in question were: 1) head gestures, where the selection was done by nodding; 2) hand gestures, where the users used a pushing movement to select; 3) hand gestures with grasping and opening the fist; 4) simple voice commands; and 5) touchpad on the Google Glass where a list of selectable objects were also displayed. In an elicitation study that was conducted prior to the actual user tests, the test subjects (n=16) were first presented video material about the interaction methods and then asked which they would prefer. They regarded nodding (1) as the worst technique, and thus it was not taken into account in the testing phase. In the actual user tests (n=12), touchpad (5) outperformed the gestural techniques (2 and 3), but only with a small margin, whereas the speech input (4) performed worst.

No far-reaching conclusions can be drawn from the user studies of this kind, as the implementations of the techniques vary greatly. Hence, in many cases the conclusions are in conflict with other studies. For example, in Billinghurst [P5] the findings of the user study were that a multimodal interface, in which speech and gestures were combined, was faster than gestures-only but merely as fast as speech-only. Regarding the speech interface, this is quite the opposite to the findings of Kollee et al. [P16]. More research is therefore required before the questions of usability can be answered.

It should be noted that in some cases a multimodal interface has been considered as stressful (Bai et al. [P3]) and that a singlemodal interface might have been preferred over a compound technique (Zhang B. et al. [P39]). This all comes down to design: not just any technique can be thrown together with another. Ideally, the user could interact with the computer by using the techniques that are familiar from everyday life or the same methods they interact with the physical world. If this is not possible, the learning curve that is required to handle the interaction tools should be moderate. In the best case, combining multiple modalities can empower the user to interact effectively and without giving a thought to the interface system, which leaves the brain capacity for the actual task in hand.

# 5 Conclusion

This review began with the question of novel interaction methods and techniques regarding mobile AR usage. Fortunately we did not remain empty-handed on this quest, since various techniques and methods were found on the articles and discussed on this paper. However, many of these can become common only with further advances in tracking techniques and head-mounted display technology.

Regarding applications where head-mounted displays can be used, tracking the human body is apparently the most explored area in the current interaction research. In this field of study, hand movements are naturally of primary interest. The range of applied technologies for tracking the body motion is wide. These technologies include optical sensing and computer vision in both visible spectrum and infra-red band, magnetic sensors, muscle activity sensors, and many other approaches. With HMDs, it would also be beneficial to be able to track the eye-movement, because it is the best indication of where the user's attention is focused at.

Multimodality – the combined use of different control methods – appears to be a prerequisite for providing robust and reliable usability. Therefore, it deserves a fair amount of serious research. It remains to be seen, however, which methods will actually become commonly associated with particular tasks.

For mobile use, unrestricting and unobtrusive wearable components are required but not yet widely available. This seems to be a true challenge for many solutions proposed in the reviewed articles. Small and unnoticeable parts are necessary especially when one attempts to bring a product into the markets, where the outlook of the product is tightly linked to the social acceptance that it requires in order to become popular.

Some high-tech marginal solutions, like direct brain links, are actively researched, but it seems that such solutions are still a long way from having practical applications. It is also possible that some currently unknown concept based on more conventional technology will arrive and change the whole field of study. Although many individual challenges have been overcome, there still remains plenty of work to be done before mobile augmented reality becomes a true commodity.

# Appendix A Glossary

**HMD**
Head-Mounted Display

**NED**
Near-Eye Display

**AR**
Augmented Reality

**MR**
Mixed Reality

**VR**
Virtual Reality

**GPS**
Global Positioning System

**RFID**
Radio-Frequency Identification

**IR**
Infra-Red

**VAN**
Visual Attention-Driven Networking

**Wi-Fi**
The most common WLAN standard

**WLAN**
Wireless Local Area Network

**POI**
Point of Interest

**ROI**
Region of Interest

**VOR**
Vestibulo-Ocular Reflex

**HUD**
Heads-Up Display

**EEG**
Electroencephalography

**LED**
Light-Emitting Diode

**RGB**
Red-Blue-Green

**RGB-D**
Red-Blue-Green-Depth

**ToF**
Time of Flight

# Appendix B Search strings

- **ACM (2998 hits), IEEE (492 hits), Web of Science (650 hits), Wiley (35 hits), Citeseer (265 hits), Science Direct (73 hits):**
  ("augmented reality" OR "mixed reality" OR "virtual reality") AND ("head-mounted" OR "head-worn" OR "glasses" OR "see-through") AND (controller OR control OR input OR interaction OR experience OR interface)
- **Springer (898 hits):**
  "mixed reality" AND ("head-mounted" OR "head-worn" OR "glasses" OR "see-through") AND (controller OR control OR input OR interaction OR experience OR interface)

**Acknowledgements**

The research has been carried out during MARIN2 project (Mobile Mixed Reality Applications for Professional Use) funded by Tekes (The Finnish Funding Agency for Innovation) in collaboration with following partners: Defour, Destia, Granlund, Infrakit, Integration House, Lloyd's Register, Nextfour Group, Meyer Turku, BuildingSMART Finland, Machine Technology Center Turku and Turku Science Park. The authors are from Technology Research Center, University of Turku, Finland.

# Bibliography

## Selected papers

[P1] Antti Ajanki et al. "An Augmented Reality Interface to Contextual Information". In: *Virtual Reality* 2011.15 (2010), pp. 161–173.

[P2] Oytun Akman et al. "Multi-cue hand detection and tracking for a Head-mounted augmented reality system". In: *Machine Vision and Application* 24 (2013), pp. 931–946.

[P3] Huidong Bai, Gun A. Lee, and Mark Billinghurst. "Using 3D Hand Gestures and Touch Input for Wearable AR Interaction". In: CHI'14. Toronto, Canada: ACM, 2014, pp. 1321–1326.

[P4] Matthias Baldauf, Peter Fröhlich, and Siegfried Hutter. "KIBITZER: A Wearable System for Eye-gaze-based Mobile Urban Exploration". In: Augmented Human Conference. Megéve, France: ACM, 2010.

[P5] Mark Billinghurst. "Hands and Speech in Space: Multimodal Interaction with Augmented Reality interfaces". In: ICMI'13. Sydney, Australia: ACM, 2013, pp. 379–380.

[P6] Arpan Chakraborty et al. "CAPTIVE: A Cube with Augmented Physical Tools". In: CHI 2014. Toronto, ON, Canada: ACM, 2014, pp. 1315–1320.

[P7] Andrea Colaço et al. "Mime: Compact, Low-power 3D Gesture Sensing for Interaction with Head-mounted Displays". In: UIST'13. St. Andrews, UK: ACM, 2013, pp. 227–236.

[P8] Monika Elepfandt and Martin Grund. "Move It There, Or Not? The Design of Voice Commands for Gaze with Speech". In: Gaze-in'12. Santa Monica, California, USA: ACM, 2012.

[P9] Sean Follmer et al. "Jamming User Interfaces: Programmable Particle Stiffness and Sensing for Malleable and Shape-changing Devices". In: UIST'12. Cambridge, Massachussetts, USA: ACM, 2012, pp. 519–528.

[P10] Otmar Hilliges et al. "HoloDesk: Direct 3D Interactions with a Situated See-Through Display". In: CHI'12. Austin, Texas, USA: ACM, 2012, pp. 2421–2430.

[P11] Juan David Hincapié-Ramos et al. "cAR: Contact Augmented Reality with Transparent-Display Mobile Devices". In: PerDis'14. Copenhagen, Denmark: ACM, 2014, pp. 80–85.

[P12] Juan David Hincapié-Ramos et al. "tPad: Designing Transparent-Display Mobile Interactions". In: DIS'14. Vancouver, BC, Canada: ACM, 2014, pp. 161–170.

[P13] Sylvia Irawati et al. ""Move the Couch Where?" : Developing an Augmented Reality Multimodal Interface". In: ISMAR'06. IEEE, 2006, pp. 183–186.

[P14] Jinki Jung, Kyusung Cho, and Huyn S. Yang. "Real-time Robust Body Part Tracking for Augmented Reality Interface". In: VRCAI 2009. Yokohama, Japan: ACM, 2009, pp. 203–208.

[P15] David Kim et al. "Digits: Freehand 3D Interactions Anywhere Using a Wrist-Worn Gloveless Sensor". In: UIST'12. Cambridge, Massachusetts, USA, 2012, pp. 167–176.

[P16] Barry Kollee, Sven Kratz, and Tony Dunnigan. "Exploring Gestural Interaction in Smart Spaces using Head Mounted Devices with Ego-centric Sensing". In: SUI'14. Honolulu, HI, USA: ACM, 2014, pp. 40–49.

[P17] Dávid Lakatos et al. "T(ether): Spatially-Aware Handhelds, Gestures and Proprioception for Multi-User 3D Modeling and Animation". In: SUI'14. Honolulu, HI, USA: ACM, 2014, pp. 90–93.

[P18] Zhihan Lv et al. "Multimodal Hand and Foot Gesture Interaction for Handheld Devices". In: *Trans. Multimedia Comput. Commun. Appl.* 11.1s (2014), 10:1–10:19.

[P19] Corey Manders et al. "A gesture control system for intuitive 3D interaction with virtual objects". In: *Comp. Anim. Virtual Worlds* 2010.21 (2009), pp. 117–129.

[P20] Denys J. C. Matthies et al. "ShoeSoleSense: Proof of Concept for a Wearable Foot Interface for Virtual and Real Environments". In: VRST. Singapore: ACM, 2013, pp. 93–96.

[P21] Yasuaki Monnai et al. "HaptoMime: Mid-air Haptic Interaction with a Floating Virtual Screen". In: UIST'14. Honolulu, HI, USA: ACM, 2014, pp. 663–667.

[P22] Mikiko Nakanishi and Tsutomu Horikoshi. "Intuitive substitute interface". In: *Pers Ubiquit Comput* 17 (2013), pp. 1797–1805.

[P23] Suranga Nanayakkara et al. "EyeRing: A Finger-Worn Input Device for Seamless Interactions with our Surroundings". In: AH'13. Stuttgart, Germany: ACM, 2013, pp. 13–20.

[P24] Klen Čopič Pucihar, Paul Coulton, and Jason Alexander. "The Use of Surrounding Visual Context in Handheld AR: Device vs. User Perspective Rendering". In: CHI 2014. Toronto, ON, Canada: ACM, 2014, pp. 197–206.

[P25] Jussi Rantala et al. "Glasses with Haptic Feedback of Gaze Gestures". In: CHI 2014. Toronto, ON, Canada: ACM, 2014, pp. 1597–1602.

[P26] Brett Ridel et al. "The Revealing Flashlight: Interactive Spatial Augmented Reality for Detail Exploration of Cultural Heritage Artifacts". In: *Journal on Computing and Cultural Heritage* 7.2 (2014), 6:1–6:18.

[P27] Antti Sand and Ismo Rakkolainen. "Mixed Reality with Multimodal Head-mounted Pico Projector". In: Laval Virtual VRIC'13. Laval, France: ACM, 2013.

[P28] Dieter Schmalstieg et al. "Gaze-directed Ubiquitous Interaction Using a Brain-computer Interface". In: Augmented Human Conference. Megève, France: ACM, 2010.

[P29] Tobias Schuchert, Sascha Voth, and Judith Baumgarten. "Sensing Visual Attention Using an Interactive Bidirectional HMD". In: Gaze-in'12. Santa Monica, California, USA: ACM, 2012.

[P30] Hyunyoung Song et al. "MouseLight: Bimanual Interactions on Digital Paper Using a Pen and a Spatially-Aware Mobile Projector". In: CHI 2010. Atlanta, Georgia, USA: ACM, 2010, pp. 2451–2460.

[P31] Jie Song et al. "In-air Gestures Around Unmodified Mobile Devices". In: UIST'14. Honolulu, HI, USA: ACM, 2014, pp. 319–329.

[P32] Peng Song, Hang Yu, and Stefan Winkler. "Vision-based 3D Finger Interactions for Mixed Reality Games with Physics Simulation". In: VRCAI 2008. Singapore: ACM, 2008.

[P33] Emi Tamaki, Takashi Miyaki, and Jun Rekimoto. "Brainy Hand: An Ear-worn Hand Gesture Interaction Device". In: CHI 2009. Boston, MA, USA: ACM, 2009, pp. 4255–4260.

[P34] Kazuhiro Terajima, Takashi Komuro, and Masatoshi Ishikawa. "Fast Finger Tracking System for In-air Typing Interface". In: CHI 2009. Boston, MA, USA: ACM, 2009, pp. 3739–3744.

[P35] Yuko Unuma, Takehiro Niikura, and Takashi Komuro. "See-through Mobile AR System for Natural 3D Interaction". In: IUI'14. Haifa, Israel: ACM, 2014, pp. 17–20.

[P36] Mélodie Vidal, David H. Nguyen, and Kent Lyons. "Looking At or Through? Using Eye Tracking to Infer Attention Location for Wearable Transparent Displays". In: ISWC'14. Seattle, WA, USA: ACM, 2014, pp. 87–90.

[P37] Jia Wang and Robert Lindeman. "Coordinated 3D Interaction in Tablet- and HMD-Based Hybrid Virtual Environments". In: UIST'14. Honolulu, HI, USA: ACM, 2014, pp. 70–79.

[P38] Christian Winkler et al. "Pervasive Information through Constant Personal Projection: The Ambient Mobile Pervasive Display (AMP-D)". In: CHI 2014. Toronto, ON, Canada: ACM, 2014, pp. 4117–4126.

[P39] Ben Zhang et al. "HOBS: Head Orientation-based Selection in Physical Spaces". In: SUI'14. Honolulu, HI, USA: ACM, 2014, pp. 17–25.

[P40] Lan Zhang et al. "It Starts with iGaze: Visual Attention Driven Networking with Smart Glasses". In: MobiCom'14. Maui, Hawaii, USA: ACM, 2014, pp. 91–102.

[P41] Feng Zhou, Henry Been-Lirn Duh, and Mark Billinghurst. "Trends in Augmented Reality Tracking, Interaction and Display: A Review of Ten Years of ISMAR". In: IEEE International Symposium on Mixed and Augmented Reality. Cambridge, UK: IEEE, 2008, pp. 193–202.

## Other references

[1] Antti Euranto et al. "Model-Based Tracking Initialization in Ship Building Environment". In: *University of Turku Technical Reports* 2 (2014).

[2] B. A. Kitchenham and S. Charters. "Guidelines for Performing Systematic Literature Reviews in Software Engineering. Version 2.3." In: *EBSE Technical Report* (2007).

[3] C. Sandor et al. "Immersive mixed-reality configuration of hybrid user interfaces". In: ISMAR '05. 2005, pp. 110–113.