When Bigger is Simply Better After all:
Natural and Multi-Modal Interaction with Large Displays Using a Smartwatch

Franca Rupprecht & Carol Naranjo
Computergraphics & HCI
Technische Universität Kaiserslautern
Kaiserslautern, Germany
Email: rupprecht@cs.uni-kl.de
Email: valero@cs.uni-kl.de

Achim Ebert & Joseph Olakumni
Computergraphics & HCI
Technische Universität Kaiserslautern
Kaiserslautern, Germany
Email: ebert@cs.uni-kl.de
Email: josepholakunmi@gmail.com

Bernd Hamann
Department of Computer Science
University of California Davis
Davis, US
Email: hamann@cs.ucdavis.edu

Abstract—**Smartwatches as latest technology of smart devices offer great opportunities for intuitive and natural interaction techniques. The inbuilt sensors of the smartwatches enable consistent user interaction and hands-free, heads-up operations. The utilization and usability of wrist gestures and in-air non touch gestures has been demonstrated in several studies. In combination with Large Display Devices (LDD), smartwatches can be used as control devices in a natural and intuitive way. The most common way to interact with large display devices has been through the keyboard/mouse interaction model, which gives the user wide interaction capabilities but limits the user in his physical space. However, providing more physical space for the user in order to walk around and explore the application projected limits the number of interaction modality. Often the only interaction modality performed with LDDs that do not limit the user to a steady device are pointing gestures. Using smartwatches as control interfaces for LDDs unfetters users from a steady control technology, as already demonstrated with, e.g., the Microsoft®PowerPoint smartwatch enhancement. Users are able to start presentations and switch to the next or previous slide by a simple button touch. But smartwatches offer a much higher potential as control devices of LDDs. More recently, there has been an increasing adoption of natural communication means, such as speech, touch or non-touch gestures (in-air gestures) for interacting with digital devices. Thus, in this paper we describe the design and utilization of a multi-modal interaction interface based on a smartwatch combining the input modalities: (1) touch gestures, (2) non-touch gestures, and (3) speech. In a user study, we employed the interface to different fields of application and discuss critically the suitability of the technology. It can be said, bigger is simply better after all.**

Keywords–smartwatch; human-computer interaction; multi-modal interaction; large display device; speech input.

I. Introduction and Related Work

By the mid of 2017, we have access to a wide variety of inter-connected devices that communicate with their surroundings and expand interaction possibilities. For example, smartwatches have embedded sensors and decent processing units, and they have been considerably improved and become broadly available. Despite the increase of power from these ubiquitous devices, the amount of information they can display and the input capabilities via touch gestures are defined by their display sizes and are therefore limited. In spite of the small and usually poor displays on smartwatches, big-screen Televisions (TVs) and display monitors are becoming cheaper and more prevalent. As a result, display technologies are becoming less expensive as well, and there has been a steady increase in the use of the large screen displays. Using large screen displays for just viewing or public display purposes has never be a problem with researches suggesting they are very beneficial for data exploration, collaboration, and data organization [1]. Interacting with them (efficiently) has been a key problem, with several researches looking into different ways to improve interaction with large screens. Currently, one of the most prevalent way of interacting with large display is still through touch-screen interaction [2]. Touch-screen interaction in large displays suffers from a lot of shortcomings, some of which are: difficulty to reach extreme corners of the screen; privacy of input; arm fatigue due to distances between buttons, occlusion from the finger performing the touch input [3]. Additionally, the utilization of large screen metaphors in head-mounted-displays as presented in [4] requires eyes-free, natural, and intuitive interaction techniques as it is simply not possible to touch this virtual screen.

Due to the problems encountered with touch-screen interaction with large displays, several researches explored the possibilities of touch-less / remote interaction devices. Ardito et al. [2] reported 34% of surveyed paper were focused on interaction with large displays using a remote device. These devices interact with large displays mostly through wireless protocols, such as Bluetooth, Wireless Fidelity (Wi-Fi), infrared, Short Message Service (SMS), etc. Lots of research efforts have been put into this technique of interaction, employing different devices, such as smart-phones, Wiimote, etc.; however, the smartwatches are often overlooked and underutilized, despite being easily accessible.

Over the past few years, smartwatches embedded with sensors and decent processing units have been undergoing improvements in technology and sales (adoption). According to Statista [5], about 75 million smartwatches where sold in 2017. Therefore, it is noticeable that these wearable devices are becoming widely available with an adoption rate and predicted to grow even further in the coming years.

Despite the increase in processing power and capabilities of these portable devices, the amount of information that can be displayed is highly limited by the screen sizes. However, smartwatches are fitted with processing power, sensors, and input possibilities, such as touch screens, microphones, heart rate monitors, etc. These sensors and input devices can be exploited to interact with these large display devices using natural modalities of interactions, such as tilt, touch, non-touch gestures and speech. When combining these modalities appropriately with the right user interfaces, they can create novel interaction modalities other than touch-screen displays or desktop interaction models.

Technologies for large-screen displays and smartwatches have limitations, and the lack of capabilities of one device can be compensated by the capabilities of another (display/screen-estate vs. sensors). The possibilities for natural interaction to be achieved with these devices has been explored, see [6] for example. The sensors and input capabilities of the smartwatch can be exploited to support the interaction with large-display devices, using natural interactions, such as touch, gesture or speech. Speech enhances overall interaction capabilities enormously. The use of speech is often the easiest, most natural way to interact with other humans but also computers [7]. In many of today's systems, using speech and gestures is supported in the user interface, creating a concerted and more natural user interaction [8]. Such natural interaction enables innovative interaction possibilities, going far beyond those offered by a remote control or desktop interaction model. Previous work covered the aspects of body movement gestures (non-touch gestures) [9], which will be enhanced in this work with the most natural way of interaction: speech.

In this paper, an approach for fusing multiple interaction modalities, such as speech, touch-gesture, and non-touch gestures by using a smartwatch for user interaction with large display devices is introduced in Section 2. In Section 3, we investigate in depth the concepts of multi-modal interaction and speech recognition within different usage contexts.

We first present concepts of multi-modal interaction and speech recognition in a general manner. Subsequently, we demonstrate the adaptation and utilization of these concepts in a first prototype system for three different scenarios. Therefore, we explain the system implementation in Section 4. The system is evaluated within a user study, in Section 5, in order to document the benefits or shortfalls offered by the combination of the various input modalities. Afterwards, we will critically discuss the suitability of such interfaces in different fields of application and make suggestions of suitable setups for these different kind of scenarios. In Section 6, we will discuss the conclusion and give suggestions for further work.

II. Multi-Modal Interaction

The term multi-modal interaction refers to the combination of several (multi) natural methods of communication for the purpose of interacting with a system. Natural modalities of communication are, amongst others, gesture, gaze, speech, and touch [10]; thereby making it more intuitive to untrained users. This interaction interface allows a user to employ their skilled and coordinated communicative behavior to control systems in a more natural way. Hence, multi-modal systems incorporate different modalities.

Modality refers to the type of communication used to convey or acquire information. It is the way an idea is expressed or the manner in which an action is performed [11], and it defines the type of data exchange. The state that determines the way information is interpreted in order to extract or convey meaning is referred to as *mode*. For example, gesture modality can provide data that can be interpreted into different modes of communication, such as *tilt* or *shake*. When multiple modalities are in use, it is paramount to fuse them in a way that is most suitable and natural.

Central to this concept is the ability to combine data perceived by the user, *fusion*. While on the output end, multiple channels (mostly independent of one another) can also be used to convey information, which is called *fission*.

In multi-modal systems, the decision to fuse or not to fuse the data from different modalities depends on the suitability of the intended usage of the data. The absence of multi-modal fusion is called *independent* multi-modal interaction whereas the presence is termed *combined* [11]. Combination of audio from two microphones or a microphone-array for a stereo effect can be said to be fusion. Fission on the other hand, is the splitting and dissemination of information through several channels [12], used for outputting information in more immersive ways. This could be the transmission of text, speech, vibration feedback, and audio cues concurrently, to allow a more accurate interpretation.

III. Concept of a Multi-Modal Interaction Interface

In order to provide a proof-of-concept of multi-modal interaction system using a smartwatch and speech, we had to fully understand the general concept and explore its feasibility. Research papers, such as [13] for example, provide insight into multi-modal interaction, guiding us to determine viability of some of our envisioned approaches.

A. Modes of interaction.

As multi-modal systems become more prevalent, new and novel ways of interacting with systems are continuously being discovered and improved, techniques, such as gaze, smile, gesture, speech, and touch, amongst others, are not uncommon in modern studies in Human Computer Interaction (HCI). The modes of interaction for our implementation are chosen in order to get the most potential out of the smartwatch as main interaction device, keeping in mind the common capabilities embedded on it as well as the restrictions of size and limited processing power.

1) Speech Input: The use of speech as an interaction modality is not a new technique in HCI. Actually, it has gone through numerous evolutions to attain the level of stability it presently supports today, with some systems almost enabling free form communication. Several speech based interaction systems exist today ranging from software based speech input systems (e.g., Siri) to dedicated standalone devices (e.g., Xperia Ear). Although speech has proven very useful for hands-free interaction, it can, however, be hindered by problems, such as ambient noise, privacy, and limited support for accents. Numerous *Software Development Kits* (SDKs) have been developed from research projects aiming to improve the process of speech recognition and analysis. They can be classified into two main categories: online and offline analysis. The online based analysis engines leverage powerful cloud architecture for speech recognition thereby offloading processing from the device, which serves as input interface. Some examples of popular *Application Programming Interfaces* (APIs) are: Google Speech API [14] or Microsoft's Bing Speech API [15]. Offline analysis engines allow analysis from within the system/application without the need of a network connection,

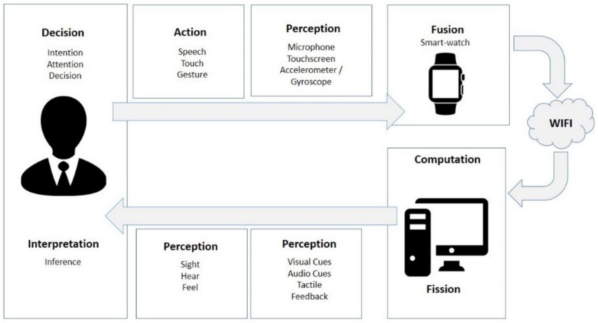


Figure 1. Multi-modal interaction system. The main components are a smartwatch and a large display enabling speech, touch, and gesture modalities perceived by the system through microphone, touchscreen, accelerometer, and gyroscope. Visual and audio cues, as well as tactile feedback are perceived by the user due sight, hear, and feel.

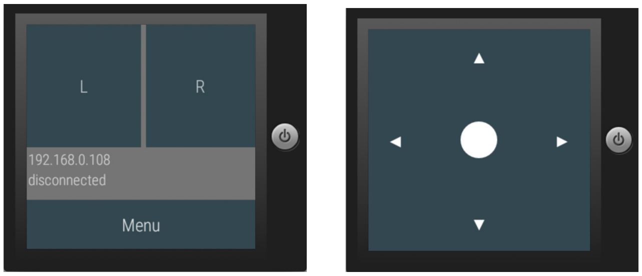
an example is CMU Sphinx open source speech recognition [16].

2) Gesture Input: Gestures used as interaction modes, adds more meaning to interaction with systems or interfaces, by empowering users to achieve goals in more natural ways that current modalities, such as mouse and keyboard does not allow [17]. Gesture input allows more natural communication modes, such as pointing, body movement, tilt, shake, etc. to interact with systems. A popular example is movement tracking with Kinect camera [18], which uses a RGB camera, depth sensor and a microphone array for acquiring user's complete body interaction and voice commands. Another popular gesture based input interface is the Wii remote control [19], which enables numerous ways of interacting with systems using gesture and movement [20]. Most mobile phones and smartwatches of recent age, come equipped with sensors that can be used to easily detect gestures of various forms which can range from a mere shake, down to imitation of steering wheel tilt for racing mobile games.

3) Touch Input: Touch is the most common input method for smart devices [21]. However, smartwatches compared to common smart devices have an even smaller form factor and are worn on a wrist, which demands a reconsideration of common smart device input techniques. Touch is more difficult on smartwatches, which typically have small screens, exacerbating the fat finger problem [22] and no multi-touch capabilities. Therefore, touch input on smartwatches should be designed in a way that even inaccurate touches are successful, i.e., very precise touch points (e.g., too small buttons) should be avoided, but at the same time, a good distribution of User Interface (UI) elements on the smartwatch's display can accelerate the task completion.

B. Interaction model.

An overview of the interaction model is shown in Figure 1, based on the multi-modal interaction concept, showing the main components needed to achieve the desired level of interaction capabilities. The figure emphasizes which modalities are used to support a user in the decision making process. We designed and implemented a smartwatch application that enables users to interact with large display devices, allowing users to interact with a Personal Computer (PC) using the provided combination of interaction modalities, i.e., speech, tilt gesture, and touch. Since the smartwatch is completely separated from the display or PC, one wants to communicate with, a system architecture is needed that enables transfer of user interaction data and interpretation of this data on the receiving large-display. In order to capture speech, gesture, and touch input, the smartwatch must have a microphone, gyroscope, accelerometer, and touch screen. Further, the smartwatch must be capable of communicating with the PC using a wireless network, requiring the smartwatch to have its own board and a Wi-Fi communication chip.



(a) Watch UI for mouse/keyboard mode (b) Watch UI for game controller mode

Figure 2. Watch User Interface for different modes

C. Application scenarios.

Multi-modal interaction can be used in many contexts, ranging from navigating a map on a white-board for controlling a robot, or navigating on a smartwatch menu. However, a "near-perfect" interaction paradigm used in one context could be inappropriate in another context; different contexts have different requirements in terms of precision and responsiveness [23]. We propose and implement two concepts to suit better different case scenarios: Free form and Tailored.

Concept A: Free Form The first category of prototypes were geared towards providing total freedom of usage to users, i.e., the input device will appear as a non customized input device, thereby appearing as an alternative to both keyboard and mouse; and also game controller. For this category the watch exists as a prototype considered suitable for "Windows, Icons, Menus, Pointing" (WIMP) and video games thereby existing as a direct substitution. The UI for the mouse/keyboard mode and the game controller-mode are shown in Figure 2.

Concept B: Tailored for Productivity According to the *productivity* category in the survey provided by Ardito et al. [2], this category refers to interfaces and interaction techniques customized for specific applications. In this mode, the UI will be created and interaction capabilities are adapted to specific use cases. We explored two use cases in our work: a data-exploration application (a common use for large displays) and a navigation game.
To better suit users' preferences, the device will provide options for customizations based on preference, such as sensitivity, inverted axis, or orientation. To support our targeted scenarios well, the application provides two options: toggling orientation to match wrist-use mode and air mouse hand-held mode in which the watch is simply hold in the hand in analogy to the usual smartdevice usage.

D. Gesture implementation.

Two classes of gestures are handled by our prototype, (1) tilt and (2) face-down. Users are alerted to calibrate the app to detect their watch's central position, used as reference point for interpreting sensor data. In order to support the tilting functionality, the data from the IMU sensors are combined and processed to obtain the updated rotation in the world coordinate system. The procedures `getRotationMatrix` and `getOrientation` of Android SDK's SensorManager are used. The procedure `getRotationMatrix` returns the rotation matrices in an orthonormal basis system[24].
The rotation matrix resulting from this process is passed on to the `getOrientation` of the `SensorManager`, returning angular vector data (in radian) as an array of `Azimuth`, `Pitch`, and `Yaw` values. This data is converted to degrees and normalized before sent to the server.
The tilt implementation was changed by using the **rotation vector**, which greatly simplifies the process, as we merely need to subscribe one sensor. The rotation vector represents the orientation as a combination of angle and axis in which the device is rotating, described as an angle θ around axis (x,y,z) [25]. The coordinate system, however, remains the same. Also, we implemented the hand flick gesture into the final system in order to perform **next**, **previous** or **toggling** through a list.
The **face-down gesture** relies on the use of gravity data. The desired effect is achieved when gravity readings are mainly in z-axis direction, with values being $v \leq -9.0m/s^2$ and readings close to 0 in the other two axis-directions. The face-down gesture is only used in this first prototype to enable listening for speech interaction. Before issuing a speech command, a user must twist the watch, forcing it to face down towards the floor, invoking the listening module and causing the watch to vibrate indicating readiness for speech input. We considered different methods to determine a very good solution for speech recognition. *Shake-to-Speak* is an operational mode where the smartwatch is quickly shaken in any direction to activate the speech listener. Another solution for the speech listening module is to continuously listen to all spoken words. Both approaches lead to many false positives and require high computation times. Using a dedicated gesture to wake up the listening module is less resource-intensive than listening continuously, and a dedicated gesture-based approach also produces less false positives.

E. Touch-Track-pad.

Furthermore, we enhanced the touch input modalities with touch-track input in order simulate mouse movements and pan activities. This could be applied by tracking finger touch movements and its velocity across the screen.

F. Speech implementation.

Several speech recognition engines were considered. The CMU pocket sphinx was adopted in this prototype, mainly due to its light-weight form and portability. Speech processing is done off-line. Two types of speech recognition were implemented, *keyword-targeted translation* and *free-form speech* for typing. Free-form speech recognition is made possible through dedicated keywords. Text synthesized from speech is transmitted as normal text in a JSON format to the server.

IV. System implementation

A. Architecture

For the implementation of our concept, we adopted a *component-based architecture*. The overall system was divided into two separate components. The system uses a *smartwatch app* component, included in the smartwatch module, and a *server* component, part of the large display module that is executed on the PC end. Both components only transmit to one another but do not rely on each other for processing capabilities, as computations and data transformations are done locally in both components.

1) Smartwatch application: For the development of the smartwatch application, the Sony smartwatch 3 [26], called SWR50 in the following, was used. The SWR50 is equipped with a microphone, Wi-Fi, and an Inertial Measurement Unit (IMU) with accelerometer, gyroscope and magnetometer, making it suitable for our purposes. It also runs the Android Wear O.S 1.5, which enables direct socket communication to help reducing latency. As shown in Figure 3, the app depends on data provided by the O.S' Sensor Manager.

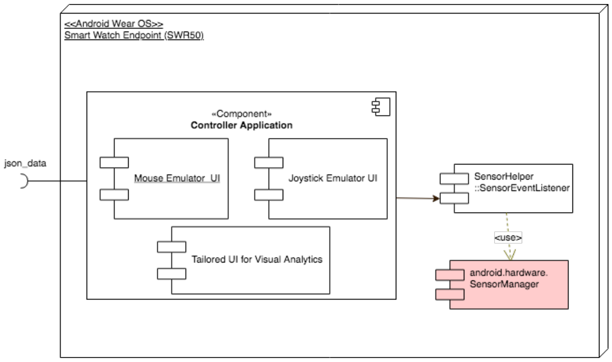

<<Android Wear OS>>
Smart Watch Endpoint (SWR50)
«Component»
Controller Application
json_data
Mouse Emulator UI
Joystick Emulator UI
Tailored UI for Visual Analytics
SensorHelper ::SensorEventListener
<use>
android.hardware. SensorManager

Figure 3. Smartwatch app and explicitly accessed system APIs (red).

2) Large Display Device: The large display is controlled from a Windows machine, where a *server application* is deployed to receive and interpret the data sent from the smartwatch in order to execute the intended action(s). On the same machine,it is deployed the corresponding dedicated application for the case of the tailored scenario.

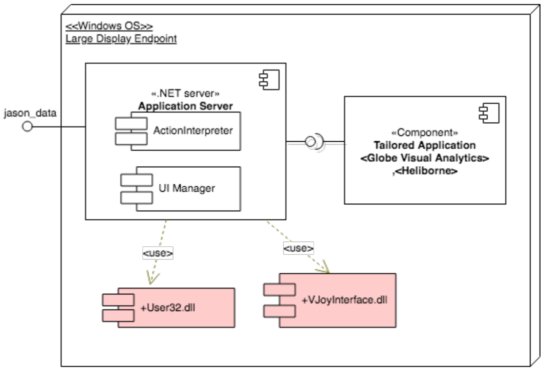
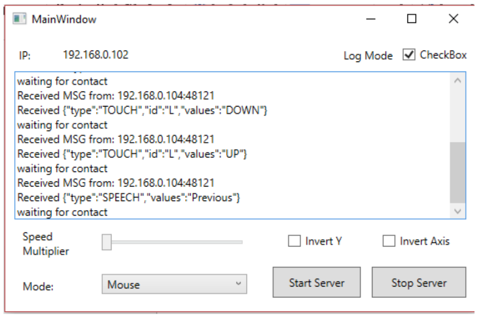


Figure 4. Large display components

a) Server application: The server application is implemented using the Microsoft .NET framework. It interprets the data from the smartwatch and decides what actions to execute. The server enacts the user's intention through the appropriate dependencies for the Windows Operative system, see in Figure 4. The server UI see in Figure 5 was designed in a straightforward manner to provide useful capabilities, including axis inversion, speed-of-mouse control and a drop-down box that allows a user to switch between three contexts to support a specific scenario via an appropriate mode. The three modes are: *mouse mode, key navigation mode*, and *game controller mode*.



IP: 192.168.0.102

Log Mode ☑ CheckBox

```
waiting for contact
Received MSG from: 192.168.0.104:48121
Received {"type":"TOUCH","id":"L","values":"DOWN"}
waiting for contact
Received MSG from: 192.168.0.104:48121
Received {"type":"TOUCH","id":"L","values":"UP"}
waiting for contact
Received MSG from: 192.168.0.104:48121
Received {"type":"SPEECH","values":"Previous"}
waiting for contact
```

□ Invert Y □ Invert Axis

Mode: Mouse

Figure 5. User Interface of the server application. The received data packets from the smartwatch are displayed for debugging purposes.

In *mouse mode*, mouse movement is simulated by tilting the smartwatch in the corresponding direction. Angular tilt data of smartwatch motion is mapped to mouse velocity. Hence, a steep tilt causes the mouse to move at high speed. In *keyboard mode*, speech is used as text input and the keyboard's cursor keys are simulated by mapping tilt angle and direction of smartwatch motion. In *controller mode*, the server acts as a feeder to the Vjoy controller, interpolating the angular values from tilt data to match an analog stick axis. This mode imitates a virtual joystick's movement, mimicking a controller analog stick with the smartwatch's tilting motions.

b) Communication Protocol: Communication between the smartwatch application and the server is enabled by a User Datagram Protocol (UDP) socket connection. Although UDP lacks reliability and congestion control, it is energy-efficient [23]. Furthermore, due to the absence of reliability logic and status packets (ACK/NACK), a high throughput and low latency can be ensured. Regarding our system, packet loss can be tolerated for our gesture data-based approach, but a lag would be detrimental for a smooth user experience. The data sent to the server is serialized as JSON since it is a light protocol that also helps in reducing the load in communication.

V. User Study

We presented two different applications covering examples for data exploration and immersive navigation, which are adequate applications to demonstrate large display interaction. Participants performed 18 tasks in total, whereby we measured their success rates in order to determine the systems effectiveness. Afterwards, surveys and interviews about usability and user satisfaction were carried out.

A. Study setting

7 university students participated in the study (undergraduate and graduate students, 21-29 years old; 5 were male and 2 female; 3 participants had experience with wearable devices and 3 participants are using speech commands frequently). In order to measure the usability and adaptability of the setup we followed the taxonomy of tasks for large display interaction, according to Foley [27]. Thus, the following task types are realized in both applications: (1) Position, (2) Orient, (3) Select, (4) Path, (5) Quantify, (6) Text. If we can demonstrate that all tasks types according to the task taxonomy are applicable in an adequate way, it can be stated that the system is usable and adaptable for large display interaction.

B. Case 1: Visual analytics - Data exploration around the globe over the years

The visual analytic application is based on the Unity3D-Globe provided by Aldandarawy [28]. A 3D globe showing the worlds population is centered on the screen, as shown in Figure 6. Area's population values are discrete data sets shown as color-coded bars attached to the country/area. The height of the bar and the color denotes the amount of population.

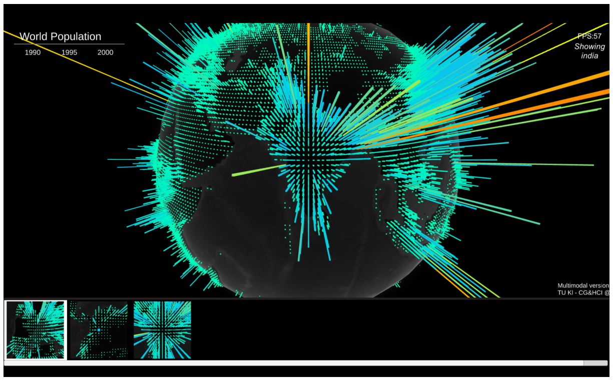


Figure 6. Unity3D-Globe application controlled with multi-model interaction techniques.

TABLE I. CASE 1: ACTION MAPPING.

No.	Action	Interaction	Task type
1	Rotate globe	Touch	Position, Orientation
2	Change data per year	Touch & Gesture	Select
3	Zoom in/out	Speech	Quantify
4	Get specific data	Speech	Text
5	Take a screenshot	Speech	Select
6	Change the mode	Gesture	Path

TABLE II. SPEECH COMMANDS FOR DATA EXPLORATION APPLICATION.

Command	Action
"Go to <Country name>"	Locate the globe to the desired location
"Zoom in" / "Zoom out"	Zoom in or zoom out in the current location
"Capture"	Take a screenshot of current location
"Remove"	Remove selected screenshot

The application is initially created for a mouse/keyboard setting but could be easily enhanced for improved interaction technology. The stated task types are mapped to the following actions inside the application as shown in Table I. The integrated speech commands are listed in Table II.

C. Tasks for Case 1

Users are given a labeled world map and a list of country names. At the beginning, users were asked to perform simple navigation tasks in order to explore the control capabilities. In the next step, users were asked to use the learned interaction techniques in order to explore the data. The following tasks had to be performed:

Control exploration phase:

1) Rotate the globe in all directions (watch control - touch & tilt).
2) Show the data for the year 1995 (watch flickering in year mode).
3) Zoom in and out (voice control).
4) Locate and view each of the countries (alternate voice and/or watch control), capture the view in few locations (voice control).
5) Remove a selected capture (voice control).

Data observations:

1) Observe the population growth in Hong Kong from year 1990 to 2000.
2) Capture the view of the current location.
3) Compare the population between Europe and Asia in the year 2000 using captures.
4) Remove existing captures (voice control).
5) Compare the population between France, Colombia, and India in the year 2000 using captures.

D. Case 2: Immersive navigation - Heliborne

The application for immersive navigation (see in Figure 7), called Heliborne [29], is also initially created for mouse/keyboard setup but could be enhanced for improved interaction technology. The application is a simple helicopter simulator controlled with multi-model interaction techniques provided by the smartwatch. Heliborne is a helicopter combat game that simulates combats and terrains, helicopter and gunships from 1950 to modern day machines. It is not a real helicopter flight simulator game, but a flight game with flight physics toned down to a control scheme make flying and playing simple and fun. Although complex maneuvers may still require some degree of expertise, the basics can be easily picked up and enjoyed.

Figure 7. Heliborne – a helicopter simulator controlled with multi-model interaction techniques.

The stated task types are mapped to the following actions inside the application as shown in Table III and linked with the speech commands listed in Table IV.

E. Tasks for Case 2

Users have a print out copy of the map in the application, highlighting specific locations. Analog to the first application, introductory users were ask to perform simple navigation tasks in order to explore the control capabilities. In the next step, users were asked to use the learned interaction techniques in order to explore the simulation world and to perform combined tasks. The following tasks had to be performed:

TABLE III. CASE 2: ACTION MAPPING.

No.	Action	Interaction	Task type
1	Raise Altitude	Speech	Quantify
2	Reduce Altitude	Speech	Quantify
3	Control flight direction	Touch & Tilt	Position
4	Move Camera	Touch & Gesture	Orientation
5	Fire/Stop Fire	Speech & Touch	Select
6	Roll left/right	Touch & Speech	Position, Quantify
7	Switch Weapon	Flick wrist	Selection
8	Select gun	Speech	Selection

TABLE IV. SPEECH COMMANDS FOR IMMERSIVE NAVIGATION APPLICATION.

Command	Action
“Go up” / “Go down”	Raise/reduce altitude of the helicopter
“Enough”	Clears previous command
“Bank left” / “Bank right”	Role the helicopter left/right
“Open fire”	Starts fire
“Give me guns”	Selects gun as weapon
“Give me rockets”	Selects rockets as weapon

Control exploration phase:

1) Raise/Reduce altitude of the helicopter (voice control).
2) Control helicopter in all directions (watch control).
3) Move the camera in left/right direction.
4) Roll left/ right (touch and voice control).
5) Select a gun and fire (voice control & flickering).

Simulation world observation:

1) Visit the camp located around longitude 6.8 and latitude 35; count the number of Silos in that settlement. (From the starting point behind you).
2) Visit the other camp located around longitude 6 and latitude 45; count the number of silos situated there.
3) Travel to the rendezvous point at longitude 4.8 and latitude 65: locate the orange signal, destroy as much of the surrounding structure around the location as you can, before landing.

F. Procedure

Conducting the whole experiment took about 45 minutes per participant. We determined 5 minutes to introduce the setups and basic interfaces. Then the participants carried out the tasks described underneath. Before each task, the concept and input modalities have been introduced and exemplary demonstrated. The participants were asked to get familiar with the corresponding device before the actual tasks have been conducted (10 minutes per application).
After the task execution session, we conducted a survey and interview. The survey included 5 aspects listed in Figure 9. During the interview, we asked for the reasons for their ratings. We also asked about general usability issues and solicited detailed feedback about the system and the experience of multi-modal interaction with the smartwatch.

G. Results

1) Effectiveness: In order to measure the effectiveness of the system, we measured the users’ success rate of each task performance. The averaged success rates defines the total accuracy value of the executed tasks. The effectiveness value is calculated by multiplying the success rate with the normalized task difficulty. Table V summarizes the accuracy and effectiveness results for both demonstrated applications.

TABLE V. ACCURACY AND EFFECTIVENESS VALUES OF BOTH APPLICATIONS.

Application	Tasks	Accuracy	Effectiveness
Visual analytics	10	96.25	95.17
Immersive navigation	8	82.5	76.73

2) User Acceptance: For the user acceptance, the descriptive statistic values mean, median, and standard deviation based on 5 point likert scale are calculated. In total, we asked 21 questions, covering the usability aspects Suitability, Learnability, Controllability, Error Tolerance, and Attitude toward using the technology, as described by Venkatesh et al. [30]. The Boxplot in Figure 8 shows the distribution of the collected user acceptance measures sorted per asked question.

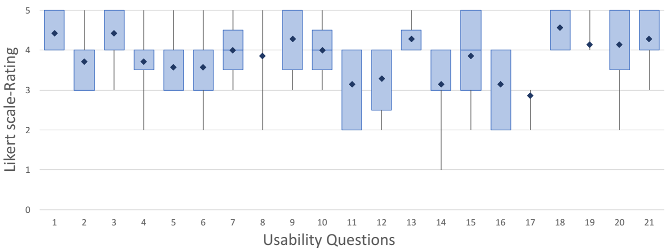


Figure 8. Boxplot displays the data distribution of the usability measures.

The average acceptability over all questions is 3.81, showing a quite good result. From direct feedback with the user, an overall satisfaction was stated, the subjects felt that learning the system was easy, using the system was fun, and the system would make their work more interesting. Users mentioned not having the feeling of complete control over the scene but also stated that it would be easy to become skillful in using the system. The attitude toward using the technology has an average value of 4.28, higher than the suitability value with 4.07, and the controllability with 3.71 in average. Table VI summarizes the acceptability measures per usability category, which are visualized in Figure 9.

TABLE VI. AVERAGE RATING OF USABILITY CATEGORIES.

Usability categories	Average rating
Suitability	4,07
Learnability	3,71
Controllability	3,71
Error Tolerance	3,28
Usage attitude	4,28

3) Quantitative Assessment: As described in [31], the qualitative results of the assessment provide a performance quantification basis that results in a scalar usability value U.

The aim of this evaluation was to proof that the system is suitable for those kinds of applications and large display interaction. Thus, the usability categories suitability and users attitude towards using the system was more important implicating the association of a higher weight in the usability score calculation. As we focused less on evaluating the quality of the implemented tasks, as well as the interaction techniques themselves, the categories error tolerance and learnability are less weighted. Table VII summarizes the weights and scoring of each usability category, leading to a satisfactory overall usability score of U = 0.8905.

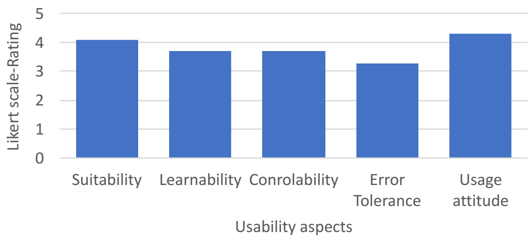


Figure 9. Average user ratings of usability categories from questionnaire.

TABLE VII. WEIGHTS AND SCORING OF USABILITY CATEGORIES TO CALCULATE USABILITY SCORE U OF THE SYSTEM.

Category	(1)	(2)	(3)	(4)	(5)
$w(s)$	0,3	0,1	0,2	0,1	0,3
$v(s)$	1	0,66	0,78	0,66	1
$w(s) \cdot v(s)$	0,3	0,06	0,15	0,06	0,3
$\sum$	0.8905				

H. Discussion

We could demonstrate that all interaction task types according to the task taxonomy are applicable in an adequate way. It can be stated that the system is usable and adaptable for large display interaction. The visual analytics application, compared to the immersive navigation application, incorporates less degree of freedom, making the control easier. As expected, the accuracy value of the visual analytics application (96.25 %) is higher than the form of the immersive navigation application (82.5 %). Analog observations are found for the effectiveness value (95.17 % vs. 76.73 %). Thus, the first application shows very good results; implying suitability of the system for this kind of task. The immersive navigation application, however, was in total more difficult. It could be observed that the control techniques felt less cumbersome towards the end of the evaluation. After executing the interviews, it is to expect that the effectiveness of the system with this kind of applications will increase after a longer training phase as users felt they could become easily skillful at using the system.

We could prove that multi modal interaction realized with the use of a single smartwatch is usable for exploration tasks and adaptable as large display interaction. With overall satisfactory user feedback and an usability score of 0.8905 over 1, the presented system demonstrates a more natural and novel way of interaction.

As mentioned in the title of this paper, we will critically discuss the suitability of such interfaces in different fields of application. Although, it could be shown that the utilization of a smartwatch as control interface is usable for these kind of tasks, it is still to discuss if the smartwatch is really suitable in these or other scenarios.

In the case of controlling a Virtual Reality-scene (VR-scene), using the device motion sensors of smartphones is a quite common approach. The smartphone is simply tilted in the direction of movement and the tilting angle is transferred to the virtual control object. Hereby, the smartphone is held in the hand. The smartwatch, however, is attached to the wrist. Tilting gestures in all 3 Degrees Of Freedom (DOF's) is limited due to the physical constraints of the human forearm and thus, higly unnatural. Accordingly, handling of tilting with smartphones feel natural and intuitive, but not with the smartwatch. Although, usable in general and commonly applied as flick gesture, are this kind of small non-touch gesture not suitable for controlling VR-scenes.

Bigger scalled non-touch gestures, like circling the arm or swiping using the complete arm, fit to controll a VR-scene (see [9]). These kind of gestures, however, need more time than small and quick movements like tilting. But, on the contrary, bigger or longer gestures lead to longer reaction time and thus are not suitable in competing applications and VR games.

Horak et al. [32] presented enhanced exploration capabilities by the combination of a smartwatch and an interactive large display within information visualization applications. This kind of interaction does not require quick reaction time or even different DOF's. As such, the watch acts as filtering technique, while the large screen is giving the overview of the investigated dataset. Getting additional information on the smartwatch is again higly limited while larger scaled smartdives as smartphones and tablet pc's are commonly used in that kind of scenario that bring a higher degree of information presentation and interaction capabilities.

Smartwatches, however, can enhance the efficiency of collaborations arising in design, simulation or data analysis, including visualization, as presented in [33]. Additional to smartdevices, the smartwatch is used to give at-the-glance information without distracting from the actual tasks and collaborations. Using the smartwatch as an alternative to the used smartdevices would decrease the quantity and quality of information presentation and interaction capabilities.

Following, although smartwatches constitute an hand free alternative for controlling applications on large displays, the limitations of the technology is omnipresent and bigger scaled devices seem to be more suitable and powerful. Concluding, either bigger gestures (mid-air gestures) with the utilization of the smartwatch, bigger time frames for computation (no competing scenarios) or even bigger scaled hand-held devices are more promising for further trends. Leading to the teasing statement in the title that bigger is simply better after all.

VI. Conclusion and Future Work

The combination of touch gestures, non-touch gestures, and speech leads to a more natural and novel ways of interaction. Speech interaction as the most natural way of interaction enhances the range of common interaction techniques significantly. Together with touch- and non-touch gesture a wide range of natural and intuitive interaction capabilities are provided. The lightweight and portability of a smartwatch makes it very convenient to handle and fuse all the modalities into one single system. Based on first prototype combining touch, non-touch gestures, and speech as interaction techniques performed with a smartwatch we could improve the system for better performance and usability. The results are described and incorporated. The performed user study of the final system provided some useful ways of combining speech, gesture, haptic, and touch interaction modes with a smartwatch, showing an effectiveness value of 95.71 % and 76.73 %. As such, the system is suitable and adaptable as efficient interaction techniques for controlling large displays. We could gather overall satisfactory user feedback resulting in an usability score of 0.8905. Following, the presented system demonstrates more natural and novel way of interaction for large displays and in general. In further work, we will work on a system that allows to easily link application-functionality with input modalities. Therefore, we will provide a gesture library for smartwatches basing on the device motion data. Further one, we will enhance existing systems like IN^{2}CO [34] with these kind of input modalities and will investigate the usability of such an enhanced system.

Acknowledgment

This research was funded by the German Research Foundation (DFG) as part of the IRTG 2057 ”Physical Modeling for Virtual Manufacturing Systems and Processes.”

References

[1] L. Lischke and et al., “Using smart phones for large-display interaction,” in Proceedings of the 2015 International Conference on Interactive Tabletops & Surfaces, pp. 501–504, 2015.

[2] C. Ardito, P. Buono, M. F. Costabile, and G. Desolda, “Interaction with large displays: A survey,” ACM Computing Surveys (CSUR), vol. 47, no. 3, p. 46, 2015.

[3] B. Badillo, D. A. Bowman, W. McConnel, T. Ni, and M. Silva, “Literature Survey on Interaction Techniques for Large Displays,” Oct. 2006. [Online]. Available: http://eprints.cs.vt.edu/archive/00000925/

[4] A. Schneider, D. Cernea, and A. Ebert, “Hmd-enabled virtual screens as alternatives to large physical displays,” in Information Visualisation (IV), 2016 20th International Conference, pp. 390–394. IEEE, 2016.

[5] Statistia. Smartwatch unit sales worldwide from 2014 to 2018 (in millions). [Online]. Available: https://www.statista.com/statistics/538237/global-smartwatch-unit-sales/

[6] J. Rekimoto, “Tilting operations for small screen interfaces,” in Proceedings of the 9th annual ACM symposium on User interface software and technology, pp. 167–168. ACM, 1996.

[7] C. I. Nass and S. Brave, Wired for speech: How voice activates and advances the human-computer relationship. MIT press Cambridge, MA, 2005.

[8] R. Bolt, “Put-that-there: Voice and gesture at the graphics interface,” in ACM Computer Graphics 14, 3, pp. 262–270, 1980.

[9] F. A. Rupprecht, A. Ebert, A. Schneider, and B. Hamann, “Virtual reality meets smartwatch: Intuitive, natural, and multi-modal interaction,” in Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems, pp. 2884–2890. ACM, 2017.

[10] A. Jaimes and N. Sebe, “Multimodal human–computer interaction: A survey,” Computer vision and image understanding, vol. 108, no. 1, pp. 116–134, 2007.

[11] L. Nigay and J. Coutaz, “A design space for multimodal systems: Concurrent processing and data fusion,” in CHI ’93 Proceedings of the INTERACT ’93 and CHI ’93 Conference on Human Factors in Computing Systems, pp. 172–178, 1993.

[12] A. E. Ali, “Multimodal interaction,” Mar. 2011. [Online]. Available: https://www.slideshare.net/Abd0/multimodal-interaction-an-introduction

[13] M. Turk, “Multimodal interaction: A review,” Pattern recognition letters, vol. 36, pp. 189–195, 2014.

[14] Google Developers, “Google cloud speech api documentation,” Jan. 2018, (Accessed on 01/30/2018). [Online]. Available: https://cloud.google.com/speech/docs/?hl=en

[15] Microsoft, “Microsoft speech service,” Jan. 2018, (Accessed on 01/30/2018). [Online]. Available: https://docs.microsoft.com/en-us/azure/cognitive-services/speech/home

[16] CMUSphinx Group, “Cmusphinx open source speech recognition,” Jan. 2018, (Accessed on 01/30/2018). [Online]. Available: https://cmusphinx.github.io/

[17] S. Bhandari and Y.-K. Lim, “Exploring gestural mode of interaction with mobile phones,” in CHI’08 Extended Abstracts on Human Factors in Computing Systems, pp. 2979–2984. ACM, 2008.

[18] Microsoft, “Kinect for windows sensor components and specifications,” 2018, (Accessed on 01/24/2018). [Online]. Available: https://msdn.microsoft.com/en-us/library/jj131033.aspx

[19] ERP – Enterprise Resource Planning, “Nintendo wii specifications,” 2018, (Accessed on 01/24/2018). [Online]. Available: http://www.tech-faq.com/nintendo-wii-specifications.html

[20] M. Schreiber, M. von Wilamowitz-Moellendorff, and R. Bruder, “New interaction concepts by using the wii remote,” in 13th International Conference on Human-Computer Interaction, pp. 261 – 270. Springer-Verlag Berlin, Heidelberg, 2009.

[21] X. Chen, T. Grossman, D. J. Wigdor, and G. Fitzmaurice, “Duet: exploring joint interactions on a smart phone and a smart watch,” in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 159–168. ACM, 2014.

[22] D. Schmidt, J. Seifert, E. Rukzio, and H. Gellersen, “A cross-device interaction style for mobiles and surfaces,” in Proceedings of the Designing Interactive Systems Conference, pp. 318–327. ACM, 2012.

[23] S. Carrino, A. Péclat, E. Mugellini, and O. A. Khaled, “Humans and smart environments: A novel multimodal interaction approach,” in ICMI ’11 Proceedings of the 13th international conference on multimodal interfaces, pp. 105–112, 2011.

[24] Google Inc., “Sensor manager,” Jun. 2018. [Online]. Available: https://developer.android.com/reference/android/hardware /SensorManager.html

[25] Google. Motion sensors. Portions of this page are modifications based on work created and shared by the Android Open Source Project and used according to terms described in the Creative Commons 2.5 Attribution License. [Online]. Available: https://developer.android.com/reference/android/hardware /SensorManager.html

[26] Sony Mobile Communications Inc., “Smartwatch 3 swr50 specifications,” Jan. 2018, (Accessed on 01/30/2018). [Online]. Available: https://www.sonymobile.com/global-en/products/smart-products/smartwatch-3-swr50/specifications/

[27] J. D. Foley, V. L. Wallace, and P. Chan, “The human factors of computer graphics interaction techniques,” IEEE computer Graphics and Applications, vol. 4, no. 11, pp. 13–48, 1984.

[28] A. Aldandarawy, “Unity3d-globe,” Dec.2017, (Accessed on 12/31/2017). [Online]. Available: https://github.com/Dandarawy/Unity3D-Globe/blob/master/README.md

[29] Jet Cat Games Interactive Entertainment, UAB, “Heliborne – helicopter combat game,” Jan. 2018, (Accessed on 01/18/2018). [Online]. Available: http://www.jetcatgames.com/

[30] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, “User acceptance of information technology: Toward a unified view,” MIS quarterly, pp. 425–478, 2003.

[31] P. Mayring, “Qualitative social research,” A guide to qualitative thinking, vol. 5, 2002.

[32] T. Horak, S. K. Badam, N. Elmqvist, and R. Dachselt, “When David Meets Goliath,” Proc. 2018 CHI Conf. Hum. Factors Comput. Syst. - CHI ’18, pp. 1–13. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3173574.3173593 Apr. 2018.

[33] F. A. Rupprecht, G. Kasakow, J. C. Aurich, B. Hamann, and A. Ebert, “Improving collaboration efficiency via diverse networked mobile devices,” Journal on Multimodal User Interfaces, vol. 12, no. 2, pp. 91–108, Jun 2018.

[34] F.-A. Rupprecht, B. Hamann, C. Weidig, J. C. Aurich, and A. Ebert, “IN2CO - A Visualization Framework for Intuitive Collaboration,” in EuroVis 2016 - Short Papers, E. Bertini, N. Elmqvist, and T. Wischgoll, Eds. The Eurographics Association, 2016.