A Multimodal Interface for Virtual Information Environments

Jeffrey T. Hansberger[1(✉)], Chao Peng[2], Victoria Blakely[2], Sarah Meacham[2], Lizhou Cao[2], and Nicholas Diliberti[2]

[1] Army Research Laboratory, Huntsville, AL 35816, USA
jeffrey.t.hansberger.civ@mail.mil
[2] University of Alabama in Huntsville, Huntsville, AL 35816, USA

Abstract. Continuing advances in multimodal technology, machine learning, and virtual reality are providing the means to explore and develop multimodal interfaces that are faster, more accurate, and more meaningful in the interactions they support. This paper describes an ongoing effort to develop an interface using input from voice, hand gestures, and eye gaze to interact with information in a virtual environment. A definition for a virtual environment tailored for the presentation and manipulation of information is introduced along with a new metaphor for multimodal interactions within a virtual environment.

Keywords: Multimodal interface · Gesture recognition · Virtual environment

1 Introduction

The concept of multimodal interfaces has captured the imagination of science fiction audiences and shown significant benefits among HCI researchers [1]. The mouse and keyboard, however, remain the primary method of interacting with this digital information. Continuing advances in multimodal technology, machine learning, and virtual reality are providing the means to explore and develop multimodal interfaces that are faster, more accurate, and more meaningful in the interactions they support. This paper will describe our ongoing effort to develop an interface using input from voice, hand gestures, and eye gaze to interact with information in a virtual environment.

The mouse has been a ubiquitous input device because it presents the metaphor of pointing that is known, efficient, and meaningful to the user. In conjunction with the WIMP (windows, icons, menus, and pointer) interface, the mouse provides an effective way of interacting with information. The mouse and its accompanying WIMP interface, however, afford indirect interactions with the information and goals of the user. Using a mouse, the user does not directly manipulate an object. They use a mouse on a two-dimensional horizontal surface whose movement is then translated to a two-dimensional vertical screen to manipulate elements of the WIMP interface. These steps and resulting task distance between the user and their goal has been defined as the gulf of execution by Norman [2]. A smaller gulf of execution will enable faster and more efficient task accomplishment with a smaller chance of error. Part of the potential of multimodal interactions is that it can afford a much smaller gulf of execution through the use of multiple and more direct input options.

J. Y. C. Chen and G. Fragomeni (Eds.): HCII 2019, LNCS 11574, pp. 59–70, 2019.
https://doi.org/10.1007/978-3-030-21607-8_5

Research with some of these alternative input modalities such as voice, eye gaze, and gestures has demonstrated the benefits of reducing the gulf of execution by providing faster and more efficient interactions. The use of eye gaze in place of a mouse for pointing at objects has proven to be a significantly faster technique [3]. There is also evidence that using voice commands is more efficient than activating the same option with a mouse and menu system [4].

The development and availability of multimodal systems that include two modalities has been rare and interfaces that use more than two modalities are even more scarce. Technological advancements in these multimodal domains of eye tracking, voice and gesture recognition however, has improved the accuracy, speed, and accessibility of the technologies monitoring and interpreting these modalities. We believe the technology in these areas is mature enough to develop a working prototype of a multimodal interface that has been designed from the ground up to integrate input from these three modalities: (1) eye gaze, (2) voice, and (3) hand gestures.

2 Related Work

2.1 Eye Gaze Input

One of the primary ways people direct their attention is by moving their eyes to visually explore and inspect the environment. Eye fixations have been shown to indicate what a person is currently working on or attending to and requires little cognitive effort [5]. Tracking a person's eye movement can be dated back to the late 19th century when Louis Emile Javal examined eye saccades while reading [6]. Eye tracking efforts are often used to understand what people are attending to or analyzing their scanning pattern to improve the design and effectiveness of a product [7].

Researchers have also explored the use of eye tracking as an input modality for interaction. Research has shown that eye gaze can be faster for selection than a mouse and can be particularly beneficial for hands free tasks and larger screen workspaces [3, 8]. Bolt used eye movements in user-computer dialogues [9, 10] while Glenn used them to actively track moving targets [11]. Researchers have also identified disadvantages and challenges with the use of eye tracking as an input modality.

Eye trackers have traditionally been limited in everyday use as they can be intrusive for the user, too sensitive to head movements, accuracy issues, and difficult to administer [12]. Another challenge using eye tracking as an input modality is called the Midas touch problem [13]. This problem occurs when interface elements are activated unintentionally by the user due to the fast and unintentional movement of the eyes. Potential solutions have been proposed such as limiting the use of eye gaze to selection and not activation and setting timing thresholds for dwell times before an item is activated [14].

2.2 Voice Command Input

Speech is widely regarded as the most natural method of communication and as such has been considered an important area of development for enhancing input capabilities.

With the development of larger vocabulary data sets and new algorithms, speech recognition technology has made extensive progress in the past few decades in achieving near instantaneous responses [15]. Early attention in this domain centered on human-machine performance comparisons centering on acoustic-phonetic modeling, language modeling, and error rates both in prime environments free of noise as well as degraded environments filled with noise pollution [16]. What started with simple machines recognizing only a few sets of sounds progressed to automatic speech recognition systems which use statistical models of speech derived from Hidden Markov Models [17, 18]. This technology recently has led to the development of spoken dialog systems allowing for multimodal inputs and the use of machine learning, resulting in high quality speech recognition.

Utilizing only a limited set of spoken command words can improve the accuracy and speed of a speech recognition system. Past research has shown that such spoken command word recognition systems can be faster than a keyboard and mouse interface [4, 19]. It has also been shown in some domains that even if the task takes longer to do with speech, the users prefer the speech input method over mouse interactions [19]. Current speech recognition systems require little training because they leverage commonly used vocabulary commands (i.e. using the natural command 'Stop' rather than less intuitive or longer phrases) [20]. This mode of input can both reduce cognitive load and increase system usability overall [21].

2.3 Hand Gesture Input

The use of hand gestures to communicate information is a large and diverse field. For brevity, we will reference the taxonomy work of Karam and Schraefel [22] to identify 5 types of gestures relevant to human-computer interaction: deictic, manipulative, semaphoric, gesticulation, and language gestures [23].

Deictic gestures consist primarily of a pointing gesture to spatially identify an object in the environment. Bolt's "Put-That-There" study in 1980 [24] defined and used hand gestures in this way for a graphical user interface (GUI). Manipulation with gestures controls objects by closely coupling the actions of the gesture with that object. Examples of this would be to move, relocate, or physically alter an object with a gesture [25, 26]. Semaphoric gestures are defined as a set of static and dynamic gestures that communicate a standard meaning when performed. An example of a static Semaphoric is a halt/stop gesture [27–29]. Gesticulation is one of the most natural uses of hand gestures and it consists of the gestures that accompany conversational speech [30]. The last form of gestures is language gestures, which represent the hand motions for sign language that have grammatical and lexical meaning associated with them [31].

A number of technological approaches are available to track and identify hand gestures. Optical solutions with external cameras that track the user's motion can include two basic types, a marker based system and markerless motion capture. The marker based system uses input from multiple cameras to triangulate the 3D position of the user wearing special markers while the markerless motion capture uses one or more cameras and computer vision algorithms to identify the user's 3D position. For issues of practicality, the markerless motion capture represents the optical motion capture of choice for general use. The Microsoft Kinect and the Leap Motion sensor are examples

of markerless motion capture systems that are both affordable and accessible to consumers, researchers, and developers. However, these types of optical sensors must have an unobscured view of the user's hands, which can force the user's arms and hands into a high fatigue posture [27]. In addition, these sensors have shown to be limited in their gesture recognition accuracy and reliability [32, 33].

Another approach that does not use any optical devices is an inertial measurement unit (IMU) system. The IMU approach consists of several sensors placed on the user or in the clothing the user wears. Each IMU consists of a gyroscope, magnetometer, and accelerometer to wirelessly transmit the motion data of the user to a computer, where it is translated to a biomechanical model of the user. IMU gloves and suits have typically been used by the movie and special effects industry but recent crowdsourcing efforts like the Perception Neuron IMU suit have provided more affordable IMU based motion capture solutions. IMU solutions do require the user to wear the sensors in the form of gloves or straps but unlike the optical solutions, it does not provide constraints on where the user's hands must be to perform the gestures. As long as the sensors are within Wi-Fi range of the router, there are no constraints on the position, orientation, or worry of obscuring the hands from an external camera source.

2.4 Multimodal Systems

Multimodal systems involve two or more of the input modalities mentioned above and beyond. One of the primary goals of multimodal systems is to leverage naturally occurring behaviors and use them to interact with digital information. Essentially it allows the user to interact with digital information in many of the same ways they interact with everyday physical objects. Thoughtful implementation of these modalities can reduce the gulf of execution mentioned earlier to improve task efficiency.

Bolt's "Put-That-There" study was one of the earliest implementations of a multimodal system integrating speech and pointing gestures [24]. Other studies have shown that there is a strong user preference to interact multimodally when given the chance [1, 19, 34, 35]. Performance is likewise improved for many tasks that include verbal tasks [1], manipulation of 3D objects [35], and drawing tasks [36]. The flexibility of multiple modalities also allows for easier error recovery [37] and allows the user to select the modality they are most comfortable using, which provides a more customized user experience. These are all important benefits to consider when designing multimodal systems for future technology and virtual environments [38].

3 Virtual Information Environment

3.1 Virtual Information Environment (VIE) Attributes

We define a virtual information environment (VIE) as a virtual environment whose primary purpose is to facilitate information foraging and processing activities. A VIE should allow the user to (1) view information, (2) control how it is organized, and (3) allow interaction with the desired information elements. The navigation requirements are reversed for a VIE compared to typical virtual environments. In most virtual

environments, the user can navigate through the environment to view and experience different aspects of the environment. With a VIE, the user is stationary and the information is moved and interacted with relative to the user's stationary position. This avoids the challenging issue of navigation within a virtual environment that many VR experiences struggle with.

The multimodal prototype developed was a digital photo management application. Fig. 1 shows the basic console with a view of the VIE from the perspective of the user. The interface allows zooming within the image collection to capture the elements of Shneiderman's visual information seeking mantra of providing an overview, while allowing the user to zoom and filter in order to obtain details on demand [39]. In order to reduce the potential for motion sickness during zooming actions with the information, the image collection is contained within the curved console. Nonmoving anchors or frames in the virtual environment help mitigate motion sickness [40]. While the images inside the console may be zooming in and out based on user input, the rest of the environment provides a nonmoving anchor.

Fig. 1. Over-the-shoulder view of the VIE and a user viewing a photo collection based on time. The timeline graph shown can be zoomed in to see that each part of the graph is composed of the images taken during that part of the timeline. The images are framed on the top and bottom by the non-moving VIE console.

Another attribute of the VIE is that most of the information visualizations, graphs, and analytics is presented primarily in a 2D fashion. Past research has found that 2D graphs are generally more accurate in presenting the intended information relative to

3D graphs [41]. Most of the graphs and information visualizations in the VIE are presented in a 2D fashion where we reserve using the third dimension for special cases where it can add new information for the user. In summary, one of the primary purposes of the VIE is to support the user in searching, manipulating, and understanding information. It does this by presenting information to the user in a virtual environment where they are stationary and most clearly presents the data, which is primarily with 2D visualizations. The second purpose of the VIE is to break the glass that separates the digital information from the user.

3.2 Bridging Digital Information with User Input Modalities

One of the most important attributes of the VIE is that it creates an environment where both digital information and multiple input modalities from the user can be directly represented. This is in contrast to WIMP interfaces where digital information is presented behind a glass monitor and interacted indirectly with a mouse and keyboard, creating a wide gulf of execution. When multiple modalities are monitored, recognized, and translated in real-time to the VIE, users have the ability to interact directly with the digital information in the same ways they interact with a physical object. This is when the full capabilities and potential of multimodal input can be realized.

3.3 A Metaphor for Multimodal Input

The use of a metaphor can help both the design and use of an interface. For design purposes, a metaphor can help identify the issues and maintain consistency. For use purposes, a metaphor can provide a schema that informs the user's current and future actions with the interface. Ware [42] and Hinckley [43] identified four control metaphors for 3D interactions:

- Eyeball-in-hand metaphor (camera metaphor): The view and perspective is controlled by the user's hand movement.
- Scene-in-hand metaphor: This is a first-person perspective view of an object where objects can be manipulated directly with a hand motion.
- Flying vehicle control (flying metaphor): This is a locomotion metaphor that covers ways to navigate through a virtual environment that includes flying, walking, jumping or riding.
- Ray casting metaphor: Object selection and navigation can occur by casting a ray at a target object or location.

We add a fifth metaphor:

- Conversation metaphor: This metaphor establishes that information elements in the virtual environment will respond to multiple input modalities of the user as if it is an active participant in a conversation. This metaphor leverages the ray casting metaphor and applies it specifically to eye gaze driven ray casting for object selection. It also expands that metaphor to include other modality inputs such as speech and hand gesture input. Each information element can respond across these different modalities, sometimes in different ways, sometimes in the same way.

The information responses are loosely based on social and conversation conventions between two humans, particularly the intent behind the action of the sender and the expected response of the receiver. The sender is the human user while the receiver is the information element in the VIE. The information element responds in a limited but similar manner as another person would. Eye gaze indicates where someone's attention is being allocated to in the environment while speech and hand gestures indicates the user's intent. The way each of these modalities supports the conversation metaphor is explored in the next section.

4 Multimodal Interactions

4.1 Eye Gaze

Technology. The technology used to monitor eye gaze in real-time is the Tobi Pro Glasses 2 installed inside an Oculus head mounted display (HMD). The eye gaze data is fed into Unity and the digital photo management application (i.e., VIE) to aid in the selection of targets and objects.

Approach. Eye gaze indicates where a person's attention is currently focused within the environment. This is typically a reliable indicator what the user is interested in or what they are currently working on. The use of eye gaze to select objects can be much faster than other selection strategies [3]. We, therefore, use eye gaze in a limited but focused manner. The eye gaze data is used purely as a selection function based on dwell time on an object. In the case of our current application, most of the objects are images, but there are other objects in the console including filters, bins the images can be sorted to, and other control devices to support the visual information seeking mantra [39]. We do not represent a cursor icon of any sort within the environment but instead, highlight the object that is currently selected based on eye gaze data.

4.2 Speech

Technology. Speech is monitored and analyzed in real-time by an open source speech recognition algorithm called Snowboy (https://snowboy.kitt.ai). Snowboy is a key word speech recognition capability that runs on raspberry pi hardware. The system requires each key word to be trained by the individual user. Training consists of repeating the key word or phrase 3 times through an online interface. The user is required to do that for each keyword to create an individual speech model that can then be loaded onto the raspberry pi hardware. Once this individual model is created and loaded, no additional changes are necessary unless new key words are added to the vocabulary list. The current key word vocabulary is around 30 words that consist of commands like "center", "home", and "activate". In order to account for terminology preferences among users, some actions are activated by more than one term such as "zoom in", "enhance", and "magnify".

Approach. The choice of using a key word approach instead of natural language processing was due to a combination of available technology, speed, and accuracy.

Based on initial testing, accuracy levels of the key word system are above 95%. In addition to the inherent speech recognition capabilities, the vocabulary list can be customized to further improve accuracy levels by selecting key words that are phonetically different from one another.

The key commands are primarily used to manipulate the view and organization of the photos in the VIE. They replace some of the functions found in the menu system of a typical WIMP interface. For example, a user can state "Filter vehicles" to apply a filter that shows only vehicles in the photo collection. Key commands can be applied generally or they can be specific to a particular photo in the VIE. Using the eye gaze input data allows the system to know which photo is being attended to and can use that location information within the VIE to zoom into when the user says a command such as, "enhance".

4.3 Gestures

Technology. Several commercial over-the-shelf motion capture systems were tested to provide real-time tracking of the user's hands and fingers. Issues arose when testing these systems regarding their accuracy, reliability, and programming flexibility. These issues motivated us to create a custom set of motion capture gloves with IMU technology. The use of IMU technology was critical in order to adopt the supported gestures described in Hansberger *et al.* [27] and avoid significant user fatigue. These gestures have been tested in both gaming environments [44] and with a digital photo management application [45]. A convolutional neural network was trained to recognize a set of 22 gestures. The training dataset was composed of 3D rotation data of finger joints recorded from the glove's IMU sensors. In order to meet the goal of real-time gesture recognition, we reduced the network's complexity by reducing the amount of feature layers and the number of weight parameters in the training phase of the network, and made the network find archetypal features of each gesture. As a result, the classification model produced by the network maintained a high recognition accuracy, and was able to classify new data samples by scanning a real-time stream of joint rotations during the use of the multimodal interface.

Approach. The use of hand gestures during speech is so natural and ubiquitous that people gesticulate as much whether the person they are talking to can see them or not [46]. The position of their arms and hands when they gesticulate is typically with their elbows bent at a 90-degree angle with their hands near their waist area [47]. In crafting our gesture vocabulary, we leveraged semaphoric type gestures used in the arm position that most gesticulation occurs [20, 27]. The gestures selected are commonly used semaphoric gestures that also have applicability to manipulate actions within a VIE. This results in short, familiar, and meaningful gestures that can be executed while the user is seated with their arms in a supported posture by a set of armrests (Fig. 2). Future gestures that allow for direct manipulation of VIE objects include actions such as pinching and pulling two ends of a photo to enlarge it.

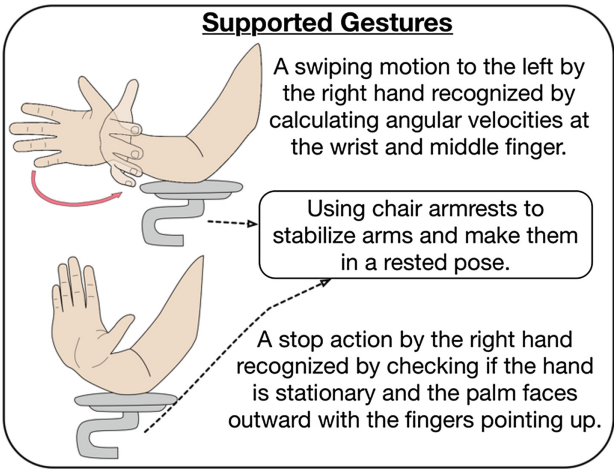

Fig. 2. Illustration of the supported gestures using an armrest of a chair as support. Two example semaphoric gestures are shown, a swipe and a stop gesture.

4.4 Multimodal Discussion

Each of these modalities offer potentially faster and more natural methods that can help reduce the gulf of execution between the user and their information related task. It is when they are integrated and designed as a single input system when the potential of a multimodal system is evident.

These modalities complement one another because we are not asking any single modality to do too much or to perform functions that they are not well suited for. Eye gaze performs the basic selection function that can then be used with either speech or gesture manipulation. With every function or task in the VIE, we have tried to provide at least two means to complete a task. For example, to zoom in on a photo, the user can look at an image and either say "zoom in" or perform a "come here" gesture. This flexibility aids in error recovery by providing alternatives for the user if one method is not effective but it also allows the user to customize their pattern of interactions within the VIE based on their individual preferences. For example, if a person, based on individual differences, prefers to interact verbally, they have the option to utilize that modality to a greater extent. This leads to greater flexibility and increased user satisfaction overall.

The application of the conversation metaphor has helped guide the multimodal system discussed here. It has motivated us to think more broadly about information and how it can be more naturally and directly manipulated in a virtual environment. More importantly, it has addressed the challenge of designing actions in the VIE that respond to multiple modalities that will help explore multimodal research questions in the future.

5 Future Directions

Future efforts in this area include a series of experiments that will examine the performance, engagement, and user experience levels that the multimodal system provides within the VIE. In addition to the VIE digital photo management application being developed, we have also developed a 2D touchscreen version that mirrors all the same functionalities. Future experiments will be able to examine the differences between unimodal and trimodal interfaces in order to better understand the advantages and disadvantages of both.

References

1. Oviatt, S.: Multimodal interactive maps: designing for human performance. Hum.-Comput. Interact. **12**, 93–129 (1997)
2. Norman, D.: Design of Everyday Things. Basic Books, New York (2013)
3. Sibert, L., Jacob, R.: Evaluation of eye gaze interaction. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (2000)
4. Karl, L., Pettey, M., Shneiderman, B.: Speech versus mouse commands for word processing: an empirical evaluation. Int. J. Man-Mach. Stud. **39**(4), 667–687 (1993)
5. Just, M., Carpenter, P.: Eye fixations and cognitive processes. Cogn. Psychol. **8**, 441–480 (1976)
6. Huey, E.: The psychology and pedagogy of reading. MIT Press, Cambridge (1968)
7. Donegan, M., Morris, J., Corno, F., Signorile, I., Chio, A.: Understanding users and their needs. Univers. Access Inf. Soc. **8**, 259–275 (2009)
8. Ware, C., Mikaelian, H.: An evaluation of an eye tracker as a device for computer input. In: Proceeding of ACM CHI+GI 1987 Human Factors in Computing systems Conference (1987)
9. Bolt, R.: Gaze-orchestrated dynamic windows. Comput. Graph. **15**(3), 109–119 (1981)
10. Bolt, R.: Eyes at the interface. In: Proceeding of ACM Human Factors in Computer Systems Conference (1982)
11. Glenn, F.: Eye-voice-controlled interface. In: Proceeding of 30th Annual Meeting of the Human Factors Society, Santa Monica (1986)
12. Morimoto, C., Mimica, M.: Eye gaze tracking techniques for interactive applications. Comput. Vis. Image Underst. **98**, 4–24 (2005)
13. Jacob, J.: Eye tracking in advanced interface design. In: Virtual Environments and Advanced Interface Design, pp. 258–288, June (1995)
14. Bednarik, R., Gowases, T., Tukiainen, M.: Gaze interaction enhances problem solving: effects of dwell-time based, gaze-augmented, and mouse interaction on problem-solving strategies and user experience. J. Eye Mov. Res. **3**(1), 1–10 (2009)
15. Stedmon, A., Patel, H., Sharples, S., Wilson, J.: Developing speech input for virtual reality applications: a reality based interaction approach. Int. J. Hum.-Comput. Stud. **69**(1–2), 3–8 (2011)
16. Lippmann, R.: Speech recognition by machines and humans. Speech Commun. **22**(1), 1–16 (1997)
17. Davis, K., Biddulph, R., Balashek, S.: Automatic recognition of spoken digits. J. Acoust. Soc. Am. **24**(6), 627–642 (1952)

18. Baum, L.: An inequality and associated maximization technique in statistical estimation for probabilistic functions of markov processes. Inequalities **3**, 1–8 (1972)
19. Cohen, P., Oviatt, S.: The role of voice input for human-machinge communication. In: Proceedings of the National Academy of Sciences (1995)
20. Barfield, W., Baird, K., Bjorneseth, O.: Presence in virtual environments as a function of type of input device and display update reate. Displays **19**, 91–98 (1998)
21. Hone, K., Baber, C.: Designing habitable dialogues for speech based interaction with computers. Int. J. Hum Comput Stud. **54**(4), 637–662 (2001)
22. Karam, M., Schraefel, M.C.: A taxonomy of gestures in human computer interactions. Technical report, University of Southampton (2005)
23. Quek, F., et al.: Multimodal human discourse: gesture and speech. ACM Trans. Comput. Hum. Interact. **9**(3), 171–193 (2002)
24. Bolt, R.: "Put-that-there": voice and gesture at the graphics interface. In: Proceedings of the 7th Annual Conference on Computer Graphics and Interactive Techniques (1980)
25. Rekimoto, J.: Pick-and-drop: a direct manipulation technique for multiple computer environments. In: Proceedings of the 10th annual ACM Symposium on User Interface Software and Technology (1997)
26. Rubine, D.: Combining gestures and direct manipulation. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (1992)
27. Hansberger, J.H., et al.: Dispelling the gorilla arm syndrome: the viability of prolonged gesture interactions. In: International Conference on Virtual, Augmented and Mixed Reality (2017)
28. Baudel, T., Beaudouin-Lafon, M.: Charade: remote control of objects using free-hand gestures. Commun. ACM **36**(7), 28–35 (1993)
29. Cao, X., Balakrishnana, R.: Visonwand: interaction techniques for large displays using a passive wand tracked in 3D. In: Proceedings of the 16th Annual ACM Symposium on User Interface Software and Technology (2003)
30. Wexelblat, A.: Natural gesture in virtual environments. In: Proceedings of the Conference on Virtual Reality Software and Technology (1994)
31. Bowden, R., Zisserman, A., Kadir, T., Brady, M.: Vision based interpretation of natural sign languages. In: Exhibition at ICVS03: The 3rd International Conference on Computer Vision Systems (2003)
32. Brown, M., Stuerzlinger, W., Filho, E.: The performance of un-instrumented in-air pointing. In: Proceedings of Graphics Interface Conference (2014)
33. Guna, J., Jakus, G., Pogacnik, M., Tomazic, S., Sodnik, J.: An analysis of the precision and reliability of the leap motion sensor and its suitability for static and dynamic tracking. Sensors **14**(2), 3702–3720 (2014)
34. Hauptmann, A.: Speech and gestures for graphic image manipulation. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (1989)
35. Kefi, M., Hoang, T., Richard, P., Verhulst, E.: An evaluation of multimodal interaction techniques for 3D layout constraint solver in a desktop-based virtual environment. Virtual Real. **22**(4), 339–351 (2018)
36. Leatherby, J., Pausch, R.: Voice input as a replacement for keyboard accelerators in a mouse-based graphical editor: an empirical study. J. Am. Voice Input/Output Soc. **11**(2) (2002)
37. Suhm, B.: Multimodal interactive error recovery for non-conversational speech user interfaces. Ph.D. thesis, Fredericiana University (1998)
38. Nizam, S., Abidin, R., Hashim, N., Lam, M., Arshad, H., Majid, N.: A review of multimodal interaction technique in augmented reality environment. Int. J. Adv. Sci. Eng. Inf. Technol. **8**(4–2), 1460–1469 (2018)

39. Shneiderman, B.: The eyes have it: a task by data type taxonomy for information visualizations. In: Proceedings 1996 IEEE Symposium on Visual Languages (1996)
40. Hettinger, L., Riccio, G.: Visually induced motion sickness in virtual environments. Presence Teleoperators Virtual Environ. **1**(3), 306–310 (1992)
41. Hughes, B.: Just noticeable differences in 2D and 3D bar charts: a psychophysical analysis of chart readability. Percept. Mot. Skills **92**(2), 495–503 (2001)
42. Ware, C., Osborne, S.: Exploration and virtual camera control in virtual three dimensional environments. Comput. Graph. **24**(2), 175–183 (1990)
43. Hinckley, K., Pausch, R., Goble, J., Kassell, N.: A survey of design issues in spatial input. In: Proceedings of the 7th Annual ACM Symposium on User Interface Software and Technology (1994)
44. Peng, C., Hansberger, J., Shanthakumar, V., Meacham, S., Blakely, V., Cao, L.: A case study of user experience on hand-gesture video games. In: 2018 IEEE Games, Entertainment, Media Conference (GEM) (2018)
45. Peng, C., Hansberger, J.T., Cao, L., Shanthakumar, V.: Hand gesture controls for image categorization in immersive virtual environments. In: 2017 IEEE Virtual Reality (VR) (2017)
46. Cadoz, C.: Les Realites Virtuelles. Flammarion, Dominos (1994)
47. Kendon, A.: Gesture: Visible Action as Utterance. Cambridge University Press, Cambridge (2004)