# A REVIEW OF TECHNOLOGIES FOR GESTURAL INTERACTION IN VIRTUAL REALITY

Jorge C. S. Cardoso

*CISUC, Department of Informatics Engineering, University of Coimbra, P-3004 516 Coimbra, Portugal*

*CITAR, School of Arts, Portuguese Catholic University, Porto, Portugal*

*jorgecardoso@ieee.org*

# INTRODUCTION

Interest in virtual reality (VR) technology has recently resurfaced due to the emergence of low-cost consumer devices with considerable graphics quality. The following is a common understanding of “virtual reality” that signals one of the characteristics that differentiates VR from other types of interactive systems – the need for constantly sensing participants’ actions.

“a medium composed of interactive computer simulations that sense the participant's **position and actions** [emphasis added] and replace or augment the feed-back to one or more senses, giving the feeling of being mentally immersed. . .” (Sherman and Craig 2003)

Even if the VR world does not include a complete representation of the user’s body, users still expect to, e.g., “touch” a virtual object with their hands and see the object move. This requires a computer system that is constantly sensing the position of the user’s body or body part – gestures performed by users are fundamental for the interaction with the VR world. (Kurtenbach and Hulteen 1990)’s definition of gesture is adopted in this work:

“A gesture is a motion of the body that contains information. Waving goodbye is a gesture. Pressing a key on a keyboard is not a gesture because the motion of a finger on its way to hitting a key is neither observed nor significant.”

Any body part is considered relevant if it is possible for a computer system to detect its movement.

# Motivation

Different VR experiences have different requirements for the sensing detail and accuracy of users’ actions and for the body parts that need to be sensed. For a designer of a VR experience, it is important to know the available technical options. However, given the breadth of technical possibilities, designers of VR experiences struggle to know them all and understand their characteristics and limitations. From entire rooms equipped with wide area tracking systems that track the position and orientation of a user’s body through markers placed in a motion-capture suit, to systems focused only on users’ hands that employ computer vision techniques to identify specific hand poses, the possibilities for automatic gesture recognition are immense. Having a catalogue of examples of application of different gestural technologies would be an asset for learning about the possibilities and for comparing the different approaches when deciding the details of a VR project.

# Objective

The objective of this work was to produce a database of possible technologies for implementing gesture-based interaction for VR environments, with concrete examples of applications of that technology describing the details about the devices used, the setting in which the experience takes place, the tracked body part, and the gestures that were detected. This objective was achieved through a review of scholarly publications and analysis of their descriptions of VR systems.

The specific contributions of this work are:

1. review of literature about technologies for gestural interaction in the context of virtual reality,
2. an analysis of the main characteristics and limitations of the reviewed technologies,
3. an online, searchable database with the reviewed examples categorized by type of hardware device needed, tracked body part, and usage setting. The resulting Web-based database can be found in (Cardoso 2017).

# Delimitations

A literature review is potentially a never-ending task, and any review must necessarily impose limits on the extension of the reviewed works and the depth of analysis. In this work, head-mounted displays (HMDs) were excluded from the analysis. HMDs usually have rotation sensors that are used to detect the perspective over the 3D environment. These sensors can be used to detect head movements for purposes other than changing the 3D perspective. However, the type of device is very well known and usage of the sensors beyond changing the 3D perspective is small. Touch input in 2D screens was also left out, although touch input can also be considered a gesture. Technologies specific to locomotion in VR environments were also not included. Locomotion is a very large area of research in VR which would require its own review of technologies.

# Related workd

Literature reviews related to gestural technologies have been done before by various researchers. These works try to provide a high-level view of a research body by aggregating and categorizing individual examples, providing a top-down view over the field. Reviews may help identifying gaps in the research which represent opportunities for further research. In general, literature review works also allow newcomers to the field to rapidly understand and construct a mental map of the main approaches for their fields. Due to their nature, review works tend to be focused on a subset of the field and have a specific audience. For example, (Zabulis, Baltzakis, and Argyros 2009), and later (Rautaray and Agrawal 2015), both wrote reviews of vision-based hand gesture recognition for human-computer interaction. Their focus was on a specific detection technology – computer vision – and a very specific body part – hands. Their review analyzed and was structured around the algorithms for the three layers of computer-vision systems: detection, tracking, and recognition. Detection is “responsible for defining and extracting visual features that can be attributed to the presence of hands in the field of view of the camera(s)” (Zabulis, Baltzakis, and Argyros 2009). Tracking is responsible for associating the detected objects in consecutive frames. Recognition is responsible for assigning labels to the data extracted from the previous layers that correspond to gestures. The current review of gestural technologies for VR includes various examples where computer vision techniques have been employed, but the details of the algorithms used are not presented. The work by (Zabulis, Baltzakis, and Argyros 2009) and (Rautaray and Agrawal 2015) can be seen as a complement for the work presented in this paper, and a reference for algorithm details for the specific case of vision-based detection technologies. (Cheng, Yang, and Liu 2015) wrote a survey on 3D hand gesture recognition, motivated by the emergence of depth sensors such as the Kinect. They structured the analysis in three stages: acquisition, recognition, and applications. In the acquisition stage, they describe various 3D sensors that can be used for hand gesture, but also existing datasets that represent various catalogued gestures. They include sensors such as Kinect 1.0, Kinect 2.0 (time of flight camera), Leap Motion, 2D cameras, multi-camera systems, and range cameras. In the recognition stage, they describe hand modelling and gesture recognition algorithms, discriminating between static and continuous gesture recognition. In the application stage, their analysis focuses on gaming, sign language, virtual manipulation, daily assistance, palm verification, and human-robot interaction. In my review, the focus is primarily on the technologies, which correspond roughly to the acquisition stage, but I do not limit the analysis to depth sensors. Also, instead of describing detection algorithms and applications, my concern is in providing a high-level view of the technology by presenting examples of gesture-based interactions.

This work aims not only to provide an overview of the various technologies that have been used for gestural interaction in the context of VR, but also to provide a searchable database of concrete examples.

# METHODOLOGY

This survey followed a systematic approach comprising the following steps.

An initial metadata search on IEEE Xplore Digital Library was performed, using two keywords: “virtual reality” AND “gesture”. The metadata of over 630 search results was downloaded and converted to a spreadsheet format.

A first scan over the search results was performed to 1) eliminate publications that were inaccessible (i.e., behind paywalls), 2) discard publications that did not mention details of the use of any gestural technology. This first scan was performed by reading the abstract. If any doubts still existed after reading the abstract, the implementation (or equivalent) section was rapidly skimmed to confirm the use of any gestural technology. The intent of this initial scan was just to reduce the total number of publications to be analyzed. About 91 publications were removed from the list due to paywalls, and about 316 were removed due to not describing the use of any gestural technology with sufficient detail.

The remaining publications were read in more detail to extract the characteristics of the VR system and associated gestural technology. A few additional publications were removed after closer inspection for not describing the use of gestural technology or interactions with sufficient detail. The final list contains 220 publications. Reading focused on the implementation, results, and discussion sections of the publications and aimed at providing answers to the following questions:

* Devices: what physical devices where employed to sense user’s actions?
* Body part: what body part was the system targeting?
* Spatial setting: what was the spatial setting in which the user interacted with the system (e.g. using an HMD, in a CAVE, a standard desktop, a large frontal projection)?
* Gestures: What gestures was the system attempting to recognize?
* Issues: Did the researchers explicitly report and issues they found while using the gestural technology?

The categorization of the devices in this paper and in the web-based database of gestural technologies is obviously subjective. Categories were chosen to reflect the most obvious differences in how the devices are used or worn but also the differences in the types of gesture recognition they support.

# Limitations

The size of the sample of publications and the fact that they originate from a single database (IEEE Xplore Digital Library) may be pointed out as limitations of the methodology for this work. However, although from a single database, the list of publications in the sample originated from about 130 venues with a date span from 1993 to 2016. Thus, the publication sample appears to have enough variability to cover most of gestural technologies used in the context of VR.

# FINDINGS

# Statistics

Two hundred and twenty publications were analyzed in this work. The publication year spanned from 1993 to 2016. About 50% of the publications analyzed were published in the last 5 years (Figure 1).

These publications originate in 139 different venues as determined by the analysis of the first part of the DOI suffix code.



Figure 1. Distribution of publication year.

Detection of hands constitutes an overwhelmingly majority of use cases for gestural technologies (Figure 2). Almost 70% of the examples found target the hands. This is consistent with other surveys (Badi et al. 2016) and is obviously expected since hands produce most of the gestures in human communication.

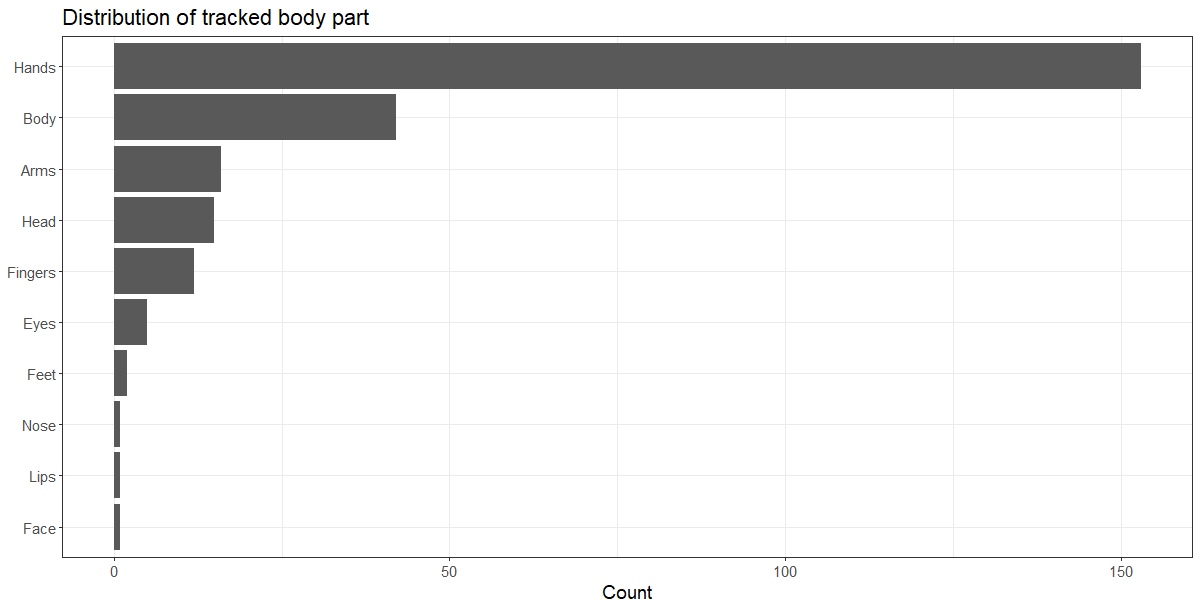


Figure 2. Distribution of tracked body part.

Interestingly, in terms of detection technology, the sample of publications shows more examples that use depth cameras (RGBD) than traditional RGB cameras (Figure 3). Of those, most examples use the Microsoft Kinect device. This shows the success of the device for research on gestural interaction given that it was only introduced in 2010.

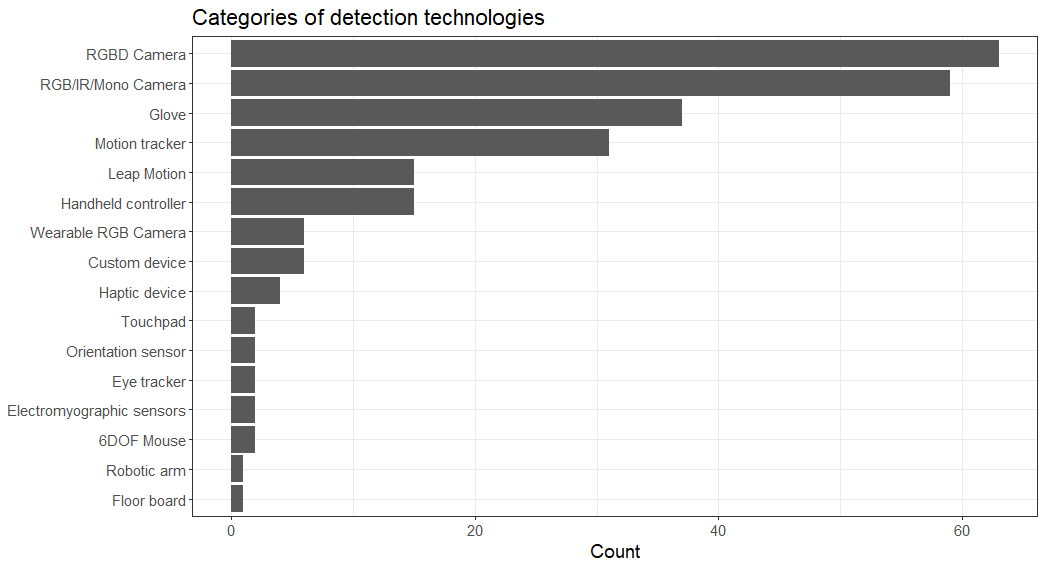


Figure 3. Distribution of detection devices.

The following sections describe the first five categories of detection devices presenting examples and the main characteristics and limitations.

### RGBD Camera

This category includes depth cameras, i.e., cameras capable of calculating the distance of each individual picture element to the camera. These cameras usually contain a color (RGB) camera and a depth (D) sensor, although some cameras contain only the depth sensor.

The technology for depth cameras varies. Some are based on the projection of a pattern of light – usually infrared (IR) over the scene. The objects on which that pattern lays cause distortions that depend on the distance. The resulting pattern is then read by an IR camera allowing the device to calculate the distance of each picture element.

Other cameras can emit a light source and measure the time it takes for it to be reflected to the camera, allowing them to calculate the distance of the picture element. Other types of camera work like humans, using a pair of color cameras and computer vision algorithms to infer the distance of the various objects on a scene. Regardless of the type of sensing technology, the information that results from depth cameras is a depth map and a color image.

Depth cameras make it easier or at least more robust to detect objects in front of them. Because depth cameras usually operate with IR light, they are usually less sensitive to lighting conditions that regular RGB cameras.

These devices are usually programmed by using a software development kit (SDK) that analyses the depth data and extracts high-level information about objects or people in front of the camera. For example, a common functionality provided by these SDKs is the detection of human bodies and reporting a simplified skeleton model for each detected body.

RGBD cameras have been used to track and recognize gestures made with various body parts (hands, arms, legs, overall body). (Zhou Ren, Jingjing Meng, and Junsong Yuan 2011), for example, used a Kinect device to detect various hand postures corresponding to the numbers 1 to 9 to play a Sudoku game (see Figure 4).



Figure 4. Hand gestures for playing Sudoku. Adapted from (Zhou Ren, Jingjing Meng, and Junsong Yuan 2011)

In a virtual reality installation – Livia's Villa Reloaded – dedicated to the archaeological site of Villa Ad Gallinas Albas, (Pietroni, Forlani, and Rufa 2015) used a Kinect sensor to allow visitors to navigate through the 3D environment. Among other gestures and body movements, the installation allowed users to use their left arm to indicate the viewing direction. The arm’s relative position was detected by a Kinect sensor positioned in front of the user.

Accuracy of depth cameras vary, but depth data usually has an error on the order of at least a few centimeters. Additionally, researchers report several issues to be aware of when using this kind of technology:

* They are sensitive to infrared light so usage in places with sunlight or other sources of IR lighting is limited
* They are sensitive to reflections and shiny surfaces
* Care must be taken to position them correctly so that they are able, e.g., to detect a human body
* Skeletal tracking may take a few seconds after detection of the body
* Depth accuracy may vary according to lighting conditions
* Depth cameras usually operate on lower framerates, so they may not be able to detect very fast movements.

### RGB/Monochrome/IR Camera

This category includes traditional web cameras as well as other color (RGB) cameras, but also more specialized monochrome or IR cameras. These cameras differ on the type of light they can sense and on the number of output color channels, but otherwise they all output a stream of bi-dimensional color data. We also include in this category projects that require multiple synchronized or unsynchronized cameras.

RGB cameras usually require a very controlled lighting environment and a static background for the detection algorithms to work robustly. When detecting body parts, particularly hands and faces, the algorithms may be sensitive to skin color. Some systems require users to wear special clothing or gloves or physical props of a specific color to be detected, but in general they provide a device-free interaction with minimal setup for new users. Despite these limitations, these cameras are usually very cheap, small, and easy to install and adapt to various settings, making them an attractive alternative for gestural interaction. Both depth cameras and traditional color cameras are interesting because they allow an unencumbered interaction, without cables or additional devices. However, they are not very good at detecting small movements and body parts. For example, if detecting the relative position of fingers with detail is important, cameras may not be the best choice.

There is a vast knowledge body about computer vision algorithms for detecting various body parts in various settings using RGB cameras. (Hyosun Kim and Fellner 2003) used whitepaper on fingertips illuminated with a “black light” source to provide several kinds of finger gestures for manipulating 3D objects in a back-projection screen. Cameras were installed at the top of the projection screen facing the user and overhead black lights illuminated the scene and made the fingertips stand out from the background.

RGB Cameras are highly flexible in their use and can be applied to the detection of very localized body parts. (Czyzewski et al. 2014) in the LipMouse application for example, used a web camera to detect lip gestures (see Figure 5). The overall mouth position is translated to cursor movements on the screen, but specific lip gestures such as opening the mouth or forming a kiss gesture can trigger specific mouse actions such as double click.



Figure 5. The main application window for LipMouse

The issues found be researchers when using RGB cameras as gesture recognition devices vary greatly depending on the type of recognition (body part and algorithm) and camera installation setting. However, in general, the main issues are

* Lighting. Detection algorithms typically require controlled lighting conditions, e.g., making sure that there is enough contrast between the object of interest and its surroundings. Algorithms also usually require very stable lighting conditions. A change in illumination may require recalibrating the system.
* Skin color. Some algorithms are very sensitive to skin color requiring specific calibrations for each user.
* Occlusion. As with RGBD cameras, an RGB camera can only detect objects in its line of sight.

### Glove

The “Glove” category includes both passive (e.g., simple colored gloves using in conjunction with a vision system) and active gloves (e.g., gloves with sensors that report finger bending or 3D hand position and orientation). However, in this description we focus on active gloves. There are many different types of active gloves, with different sensing capabilities. Some sense only two states for each finger - bent or not bent, others provide a continuous bend value, some require cables to connect to a computer, some are wireless. Some gloves can sense the hand orientation and position relative to a base station, others need to be fitted with additional sensors for this.

Gloves obviously target at a specific body part and, as such, can give more precise readings of the relative positioning and orientation of the fingers than RGBD or RGB cameras. Gloves can be a good alternative for systems that only need to track users’ hands. Gloves don’t have the line-of-sight problem that cameras have. The downside is that they always represent an additional prop that users must wear and thus require setting up. Also, when the system is meant to be used by various users over time, gloves can be subject to the usual wear-and-tear and sanitary problems. In this survey, no example of usage of gloves in public settings was found.

(Khambaty et al. 2008) developed a custom tethered glove system comprising bend sensors to sense the position and movement of the fingers and wrist, and two sensors to measure the pitch and roll of the wrist. They applied this glove to a sign language recognition (see Figure 6) that speaks out the letters and words.

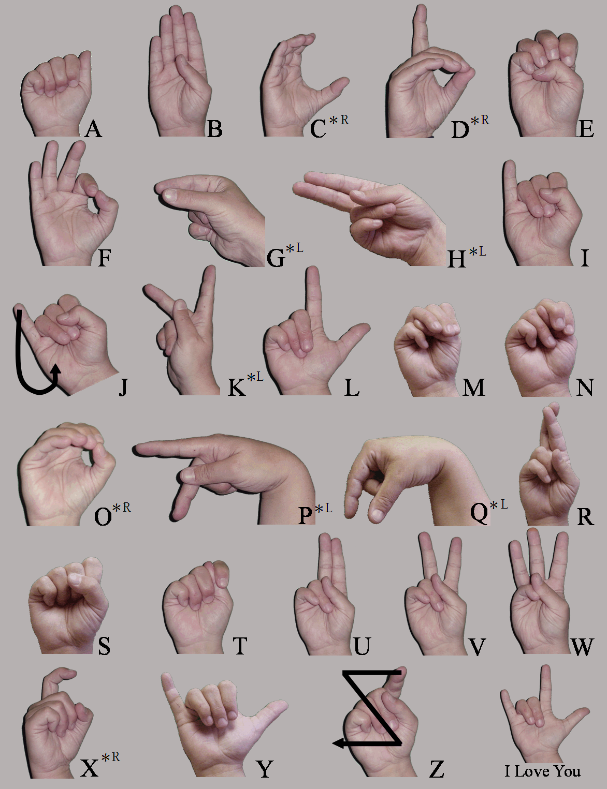


Figure 6. Sign language alphabet. From (Commons 2015a).

Using gloves for gestural interaction faces a few issues that are very dependent on the specific glove technology used:

* Some gloves have positional trackers to determine the position and orientation of the hands, but these trackers were made for desktop interaction and have a limited range.
* Some gloves require wires connected to a computer which may interfere with the natural hand gestures.
* The cost of a glove system can be relatively high.

### Motion tracker

Motion trackers are generic devices that can be used to track the position and orientation of various kinds of objects, including human bodies. Motion trackers can be implemented with different technologies, but the most common are optical motion trackers and magnetic motion trackers. Optical systems typically use several infrared cameras on a room’s ceiling and reflective markers (there are also systems that use active markers that emit light) placed over the user’s body (on a special suit) or on objects to be tracked. Optical systems are subject to occlusion issues since markers may be temporarily occluded by other objects. Magnetic systems emit a magnetic field that the sensors placed on the objects to be tracked can sense to determine their location and orientation. Because these systems use a magnetic field, they are not susceptible to occlusion problems. However, these systems are susceptible to magnetic interference from metals and electronic equipment. Magnetic trackers can usually only track objects in a smaller volume than optical systems.

Regardless of the technology, these systems are usually very accurate and provide the position and orientation of several points over the object of interest. With the locations of the individual points, one can determine the position and orientation of the overall object(s).

(Portillo-Rodriguez et al. 2008) used an optical motion capture system to detect upper body movements and recognize Indian dance gestures. In Figure 7 it is possible to observe a user wearing a suit with reflective white markers and various motion capture cameras mounted in the top frame.



Figure 7. User being tracked by an optical motion tracking system. From (Commons 2015b).

Motion tracker usage should consider:

* Their cost. Motion trackers are very expensive and can be found only in a few institutions. They are not home devices.
* Their setup time. If the user’s body is to be tracked, time must be taken for the user to put on a suit or for placing the markers in the correct location. It is not a walk-up-and use system.

# CONCLUSIONS

We have presented a review of technologies for gestural interaction in virtual reality environments. This review was based on a literature analysis based on publications from the IEEE Xplore database. Examples of use of gestural technologies from these publications were analyzed and compiled in an online searchable database at (Cardoso 2017).

The results of this work can be used by designers of VR experiences to better understand what technologies have already been used to provide which kind of gestural detection. Although VR has recently gained much attention, it has focused on a narrow set of consumer devices. It is important that interaction designers know what else is out there to design novel ways of exploring virtual worlds, without being too constrained by the possibilities of a small number of commercial devices.

# REFERENCES

Badi, Haitham, Mohammed Fadhel, Sana Sabry, and Mohamed Jasem. 2016. “A Survey on Human–computer Interaction Technologies and Techniques.” *International Journal of Data Science and Analytics* 2 (1–2). Springer International Publishing: 1–11. doi:10.1007/s41060-016-0018-x.

Cardoso, Jorge C. S. 2017. “VR Gesture Survey Online Database,” February. doi:10.6084/m9.figshare.4616332.v1.

Cheng, H, L Yang, and Z Liu. 2015. “A Survey on 3D Hand Gesture Recognition.” *IEEE Transactions on Circuits and Systems for Video Technology* PP (99): 1. doi:10.1109/TCSVT.2015.2469551.

Commons, Wikimedia. 2015a. “File:ABC Pict.png --- Wikimedia Commons{,} the Free Media Repository.” https://commons.wikimedia.org/w/index.php?title=File:ABC\_pict.png&oldid=181073053.

———. 2015b. “File:MotionCapture.jpg --- Wikimedia Commons{,} the Free Media Repository.” https://commons.wikimedia.org/w/index.php?title=File:MotionCapture.jpg&oldid=146289622.

Czyzewski, A., P. Dalka, Kosikowski, B. Kunka, and P. Odya. 2014. “Multimodal Human-Computer Interfaces Based on Advanced Video and Audio Analysis.” *Advances in Intelligent Systems and Computing* 300: 87–102. doi:10.1007/978-3-319-08491-6\_8.

Hyosun Kim, and D.W. Fellner. 2003. “Interaction with Hand Gesture for a Back-Projection Wall.” In *Proceedings Computer Graphics International, 2004.*, 395–402. IEEE. doi:10.1109/CGI.2004.1309240.

Khambaty, Yousuf, Rolando Quintana, Mehdi Shadaram, Sana Nehal, Muneeb Ali Virk, Waqar Ahmed, and Ghayas Ahmedani. 2008. “Cost Effective Portable System for Sign Language Gesture Recognition.” In *2008 IEEE International Conference on System of Systems Engineering*, 1–6. IEEE. doi:10.1109/SYSOSE.2008.4724149.

Kurtenbach, G., and E. Hulteen. 1990. “Gestures in Human-Computer Communications.” In *The Art of Human Computer Interface Design*, edited by Brenda Laurel, 309–17. Addison-Wesley.

Pietroni, Eva, Massimiliano Forlani, and Claudio Rufa. 2015. “Livia’s Villa Reloaded: An Example of Re-Use and Update of a Pre-Existing Virtual Museum, Following a Novel Approach in Storytelling inside Virtual Reality Environments.” In *2015 Digital Heritage*, 511–18. IEEE. doi:10.1109/DigitalHeritage.2015.7419567.

Portillo-Rodriguez, Otniel, Oscar O. Sandoval-Gonzalez, Carlo A. Avizzano, Emanuele Ruffaldi, Davide Vercelli, and Massimo Bergamasco. 2008. “Development of a 3D Real Time Gesture Recognition Methodology for Virtual Environment Control.” In *RO-MAN 2008 - The 17th IEEE International Symposium on Robot and Human Interactive Communication*, 279–84. IEEE. doi:10.1109/ROMAN.2008.4600679.

Rautaray, Siddharth S., and Anupam Agrawal. 2015. “Vision Based Hand Gesture Recognition for Human Computer Interaction: A Survey.” *Artificial Intelligence Review* 43 (1). Kluwer Academic Publishers: 1–54. doi:10.1007/s10462-012-9356-9.

Sherman, William R, and Alan B Craig. 2003. *Understanding Virtual Reality: Interface, Application, and Design*. *The Morgan Kaufmann Series in Computer Graphics and Geometric Modeling*. Vol. 12. http://www.loc.gov/catdir/description/els031/2002107270.html.

Zabulis, X, H Baltzakis, and A Argyros. 2009. “Vision-Based Hand Gesture Recognition for Human-Computer Interaction.” In *The Universal Access Handbook*, 34.1-- 34.30. Lawrence Erlbaum Associates, Inc. (LEA). http://users.ics.forth.gr/~argyros/mypapers/2009\_06\_book\_hci\_gestures.pdf.

Zhou Ren, Jingjing Meng, and Junsong Yuan. 2011. “Depth Camera Based Hand Gesture Recognition and Its Applications in Human-Computer-Interaction.” In *2011 8th International Conference on Information, Communications & Signal Processing*, 1–5. IEEE. doi:10.1109/ICICS.2011.6173545.