

CHAPTER

10

MULTIMODAL INTERACTION IN MIXED REALITY

*Ajune Wanis Ismail, Mohamad Yahya Fekri Aladin, and
Muhammad Anwar Ahmad*

10.1 INTRODUCTION

Mixed reality (MR) is a technology where the digital world coexists with our physical world. Digital information can understand our physical world, making a virtual three-dimensional (3D) robot that can hide itself behind a physical couch and play hide-and-seek with us. This technology differs from augmented reality (AR), which only allows virtual content to be overlaid onto the physical world through a video stream. MR has grown to be so much more than just for display, covering environmental understanding and interaction metaphor. The advancement of graphics processors, central processors, computer vision, and input systems has made MR, no longer a fiction or computer-generated imaginary, possible. To accompany MR technology, natural and real interaction needs to be implemented to enhance the user experience when interacting with virtual content. There are many natural input systems based on human form, such as gaze, facial expressions, gestures, and speech. These natural inputs are essential for enabling users to interact with virtual content in the most intuitive way. This chapter explains the implementation of multimodal

interaction for MR using speech and gesture for the MR interface. AR has produced a lot of user interaction, but the MR environment has seen very little of it. With the advancement of techniques, multimodal interaction has been recommended by many researchers because complementary inputs enable users to interact using more than one input method. The remaining issue is user interaction, the user interaction needs to be as natural as possible, and the 3D object manipulation methods need to be as intuitive as possible (Ismail, 2011). This research must solve the occlusion issue to produce an intuitive multimodal interaction technique in an MR environment. Depth mapping errors occur during the capturing process, and the errors need to be minimized to correspond with the real object, such as a user's hand.

10.2 ARCHITECTURE FRAMEWORK

This section discusses the stages of developing the MR environment to achieve the research objective. The following subsections detail the methodology for developing a multimodal MR interface. Figure 10.1 discusses how to set up an MR environment. MR interface design involves gesture inputs and speech inputs; it is a basic modality to enable multimodal interaction. Hand gestures are detected using a Leap Motion Controller, while speech input is captured through a basic microphone connected to the computer and integrated with the speech recognition system in the MR space. Two primary input types are considered when constructing the MR environment: virtual elements and real-world elements. The virtual components consist of 3D objects such as cubes, cones, and spheres, whereas the real elements are real-time image projections within the MR view. The user's hand is represented as a 3D skeletal model, and user interface components are presented using two-dimensional (2D) image overlays.

In typical MR setups, head-mounted displays (HMDs) like the Microsoft HoloLens can recognize both hands and track their motion data, including position, orientation, and coordinates (Xiao et al., 2018). However, in this research, the Leap Motion Controller is

employed for hand gesture input. The device is mounted on an HMD to provide precise gesture recognition within the MR workspace, as shown in Figure 10.1. Additionally, a ZED Mini red-green-blue-depth (RGB-D) (depth-sensing) camera is attached to the HMD. Unlike infrared-based systems, the ZED Mini utilizes two red-green-blue (RGB) cameras and applies a 3D reconstruction algorithm to capture depth information for the real-world scene. Figure 10.1 illustrates the whole setup for capturing user input in the MR environment.

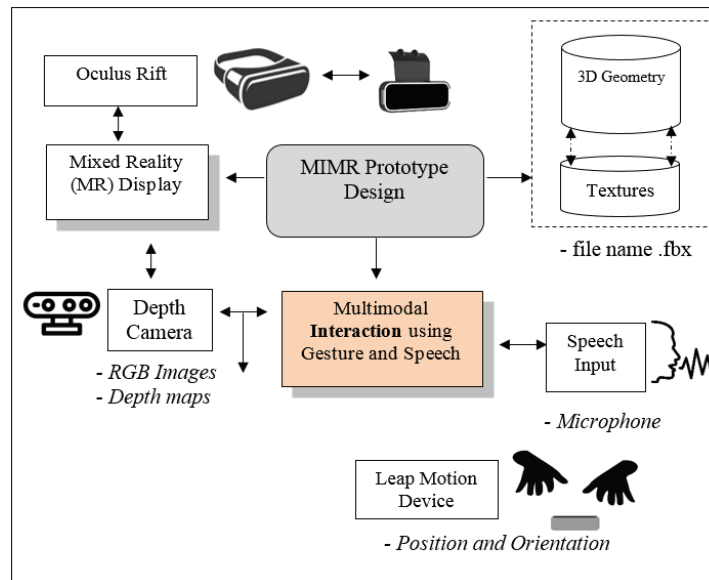


Figure 10.1 Setting up the MR environment

10.3 INNOVATION APPROACH

This study adopts the Leap Motion Device to take advantage of the powerful hand tracking. The device works by having two cameras and three infrared LEDs that track infrared light that is outside the visible light spectrum. The sensor reads the data and later streams it via Universal Serial Bus (USB) to Leap Motion Software. The software then calculates the finger positions and orientation. The prototype will then utilize this information to facilitate gesture interactions. The