Object selection and scaling using multimodal interaction in mixed reality

M Y F Aladin [1,2] *, A W Ismail [1,2], N A Ismail [1,2] and M S M Rahim [1,2]

[1] Mixed and Virtual Environment Research Lab (mivielab), ViCubeLab, Universiti Teknologi Malaysia, 81310 Johore, Malaysia
[2] School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia, 81310 Johor, Malaysia

* yahyafekri@gmail.com

Abstract. Mixed Reality (MR) is the next evolution of human interacting with the computer as MR has the ability to combine the physical environment and digital environment and making them coexist with each other. Interaction is still a huge research area in Augmented Reality (AR) but very less in MR, this is due to current advanced MR display techniques still not robust and intuitive enough to let the user to naturally interact with 3D content. New techniques on user interaction have been widely studied, the advanced technique in interaction when the system able to invoke more than one input modalities. Multimodal interaction undertakes to deliver intuitive multiple objects manipulation with gestures. This paper discusses the multimodal interaction technique using gesture and speech which the proposed experimental setup to implement multimodal in the MR interface. The real hand gesture is combined with speech inputs in MR to perform spatial object manipulations. The paper explains the implementation stage that involves interaction using gesture and speech inputs to enhance user experience in MR workspace. After acquiring gesture input and speech commands, spatial manipulation for selection and scaling using multimodal interaction has been invoked, and this paper ends with a discussion.

1. Introduction

Mixed Reality (MR) is a technology where digital world coexists with our physical world. Digital information has the ability to understand our physical world making a virtual 3D robot can hide behind a physical couch and play hide and seek with us. Differs from Augmented Reality (AR) where virtual content can only be overlaid onto the physical world through a video stream. MR term originates from [1]. Since then, MR has grown so much more than just for display but covering environment understanding and interaction metaphor. MR, no longer fiction or computer-generated imaginary is made possible by the advancement of graphic processer, central processor, computer vision and input systems. To make sense the MR technology, a natural and real interaction needs to be implemented to enhance the user experience when interacting with the virtual content [2]. There are many natural input systems based on the human form, such as gaze, facial expressions, gesture, and speech. These natural inputs are crucial to let the user interact with virtual contents in the most natural manner. This paper explains about implementing multimodal interaction for MR using speech and gesture for MR interface. User interaction has been produced in AR, very less in the MR environment. Due to the advanced technique, multimodal inter-action has been recommended by most of the researchers due to the complementary inputs, and more than one input more user to interact. The remaining issues in user interaction, the user interaction needs to as natural as possible as agreed by [3], and the 3D object

manipulation methods need to be as intuitive as possible [4]. An intuitive multimodal interaction technique in the MR environment can be achieved when the occlusion issue can be improved in this research. There are depth mapping errors during the capturing process, and the errors need to be minimized to correspond with the real object such as a user's hand.

2. Related Works
MRTouch [5] is a multitouch input solution for MR, and in their system, the user uses their physical environment as if they were touchscreens as shown in Figure 1. User is able to use their finger to interact with the virtual inter-faces affixed to surfaces. He incorporates Microsoft HoloLens as dis-play device and takes advantage of HoloLens gesture recognition mod-ule to produce their multitouch interaction.

Figure 1. MyTouch application interfaces [5].

Another research by Meegahapola [6] produced a system that enhanced in-store shopping experience through a smart phone or in short SPMRA – phone-based MR application. This application, instead of using a costly device such as HoloLens, incorporate a mobile phone as the headset by using a Samsung Gear VR or Google Cardboard. This application claims to integrate multimodal interaction using gaze and physical button on the side of the Samsung Gear VR. CETA [7], MR system for school-aged children to enhance mathematical learning using Tangible User Interaction (TUI). The system uses a low-cost android tablet, a mirror, a holder to hold the tablet, and a set of rectangle wooden blocks that acts as the tangible interface. While in architecture design, [8] has designed a prototype for multimodal to generate displays in 3D environments so users can use a variety of devices to interact in the same networked environment. However, it was fully immersive VR environment not in MR. Same as [9], they proposed a multimodal interaction to enhance creativity during designing tasks with a pulsing microphone icon to give participants feedback and indicate when they can interact in a VR environment.

In the AR environment, Ismail et al. [10] have produced a system called Multimodal Interaction in Augmented Reality or MIAR for short. The input modalities used in MIAR is natural hand gesture and speech. A leap motion device is used to track and recognize the gestures performed by the user and a microphone is used to capture the user's speech command. Piumsomboon et al. [11] adopted the use of gesture and speech input modalities in a system called Gesture-Speech Interface for Augmented Reality or G – SIAR for short and the project setup is as illustrated in Figure 2. The free-hand gesture interaction techniques or Grasp-Shell and multimodal interaction technique Gesture-Speech is constructed to be compared and evaluated.

Figure 2. G-SIAR setup.

Table 1 shows the related works in multimodal that has been listing by year in 2015 until 2020. The five years of works indicate that the projector-based started with mobile-based and headed to see-

through headset and projectors. Table 2 shows the gesture interaction that commonly invoked in MR/AR. There are two categories, gesture input using real hand and fingertip using marker-based. Marker-based is when the user's hands were attached with the black and white physical marker and interactions such as pointing, grabbing, push and pressing can be defined.

Table 1. Previous works on multimodal interaction in AR and MR from 2015-2020.

Year	Research	Interaction technique	Display
2015	Vision-Based Technique and Issues for Multimodal Interaction in AR [12]	Face, Gesture, Speech	See-through based, Mobile-based
	MIAR: Gesture and Speech Input in AR Environment [10]	Speech, Gesture	PC-based
	ARZombie [13]	Face, Gesture	Mobile-based
2016	In Situ CAD Capture [14]	Speech, Gesture	Mobile-based
	Semantics-based Software [15]	Speech, Gesture	Mobile-based
2017	Adaptive Multimodal Input [16]	Speech, Motion, Gesture, Adaptive	Mobile-based
	Multimodal Interaction in Augmented Reality [17]	Speech, Gesture	Mobile-based
	Mobile AR with the Hololens [18]	Speech, Gesture, Tangible	See-through based
2018	HandsInTouch [19]	Gesture, Sketch	See-through based, Pc-based
	Space Tentacles [20]	Gesture, Speech	Pc-based
	A Multimodal System involves AR, Gestures, and Tactile Feedback [21]	Gesture, Tactile	See-through based
2019	Multimodal Driver Interaction [22]	Gesture, Gaze, and Speech	See-through based
	Multimodal Referring Expressions with Mixed Reality [23]	Gesture, Speech	See-through based
	A multimodal interface for virtual information environments [24]	Eye gaze, Gesture, Speech	Pc-based
2020	Bio-Holograms in Mixed Reality [25]	Gaze, Eye, Haptic	Projectors, HMD
	A 3D model of student interest during learning using multimodal fusion [26]	Face, Gaze	Pc-based, HMD

**Table 2.** Gesture Inputs and Techniques.

	Pointing	Grabbing	Push	Pressing
Gesture Input using real hand [27]	✓	✓	✓	✗
	A pointing gesture is recognized when the user touches their finger to the ground	Grabbing is used to move objects by grabbing them and releasing them again	Push object to change the position by pushing it in a direction	Pressing gestures are a modification of pointing gestures and are used to interact with a virtual button
Fingertip using marker-based [28]	✓	✓	✗	✓
	A pointing gesture is recognized when the user touches their finger to the ground	Grabbing is used to move objects by grabbing them and releasing them again	Push object to change the position by pushing it in a direction	Pressing gestures are a modification of pointing gestures and are used to interact with a virtual button

3. Implementation

There are three phases that have been carried out to design the MR interface that covers the areas as described in the following subsections.

3.1. Configuration and User Setting

The interaction for MR-Deco includes rotation, translation, and also texture manipulation, where the user can change the texture of the object. HMD is essential in this development. There are several interaction controllers can be used with the HMD such as Oculus Touch, Oculus Remote and also Joystick. As illustrated in Figure 3 (a), the workspace to set up VR display technology, there are listed the computer desktop, Oculus Sensor, Oculus Rift and Oculus Touch. Oculus Rift is chosen as the display device for users to experience and interact with the MR environment. Leap Motion controller is used to capture the user's hands. Leap Motion is a controller to provide a workspace with appropriate gesture inputs; it was attached to HMD (as in Figure 3 (b)). HMD was attached with Zed Mini RGB depth (RGB-D) camera. A Zed Mini depth camera for images does not use infrared; it instead uses two RGB-D cameras.


(a) Hardware for workspace setting   (b) Leap Motion Controller to capture real hand
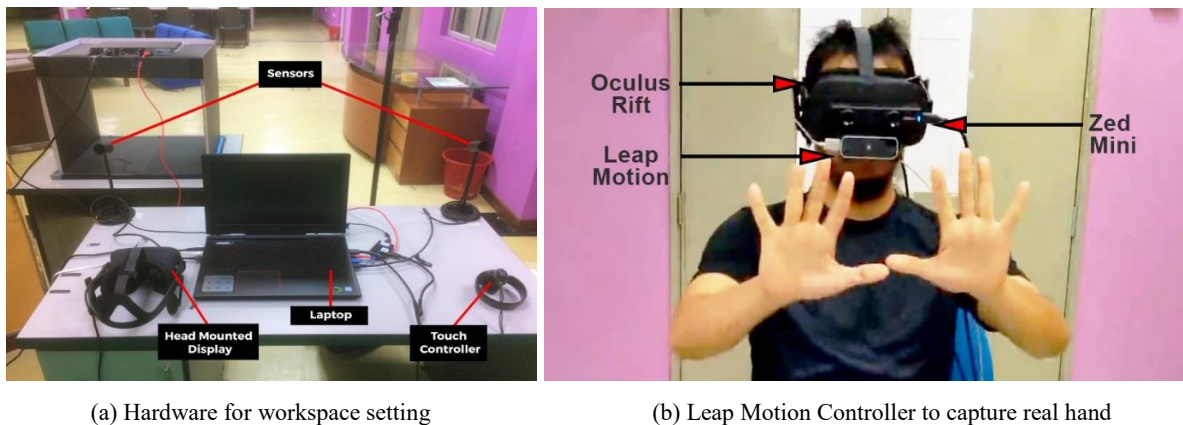
Figure 3. MR workspace setting.

3.2. Acquiring Gesture Inputs using Leap Motion

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 streamed via USB to Leap Motion library. Then the calculation takes place in the software to compute the finger positions and orientation. Then the prototype will take advantage of this information to be used as a gesture interaction. The interaction features in the prototype support two kinds of gesture – hover and grasp:

- With hover, the user can touch the 3D content, and a visual cue was projected to the user to show that the user's hands are occluded with the 3D content.
- With grasp, users are able to pick the 3D object and manipulate it. Users are able to translate and rotate the 3D object seamlessly and naturally as if the user holds a real physical object.

When the user moves the hands, the Leap Motion will track the movement in real-time. When hands are in specific space and range, the user is hovering their hands on the Leap Motion controller, and a recognition process will give the system access to its position with the orientation. As the hand inside the real-time gesture frame, the motion data is provided with a direction vector to its gesture input definitions. A gesture inputs may have numerous components such as pointing, pinch, grab, and stretch by using both hands. The pinch gesture input, for example, will imply the natural movements several times and other gestures, approximating stretch or shrink is in different axis and rotation, which can either use a single hand or both.

(a) Real environment   (b) Palm up in MR   (c) Palm down in MR
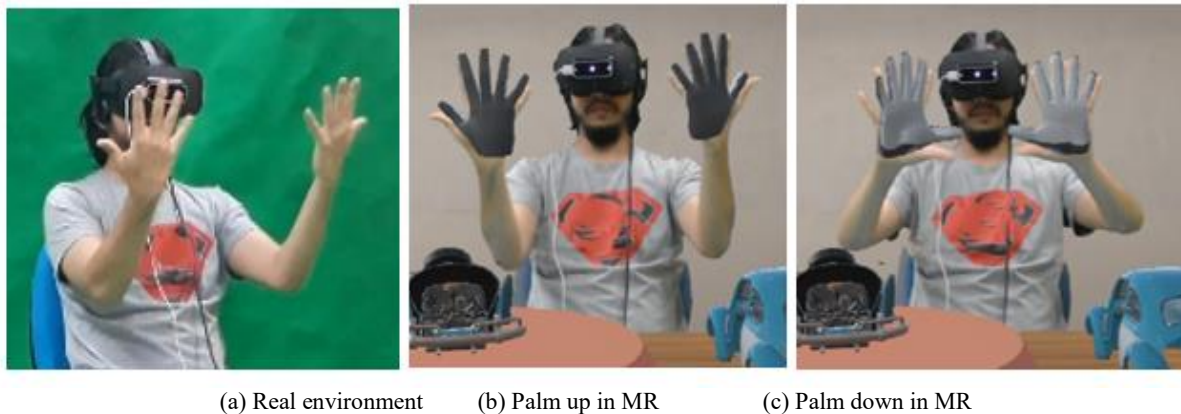
Figure 4. Both hand recognition in MR space.

Figure 4 shows the initial result to produce natural user interaction gesture interface output of the application system when user see-through the HMD. A virtual hand will be rendered in real-time according to the finger's positions and orientations. The user and his hands appear visible in MR space when gesture performs the action as shown in Figure 4, and the user was demonstrating the virtual hand palm up or palm down. Based on Figure 4 (d), the gesture is tracked by sensor-based tracking technique where hand gesture natural features are being tracked by the gesture tracking devices. Figure 4 (a) illustrates the user in the real-world environment, performing palm up gesture (as in Figure 4 (b)). With the palm up and palm down (as in Figure 4 (c)), the user can perform gesture interaction in the MR world.

3.3. Hand Gesture in MR

Figure 4 demonstrates with depth occlusion-free enabled; the virtual car still appears to occlude the real hand. Without occlusion-free, the virtual car is occluded by the user's hand. User to place their hand in front of the virtual object and user can still see their real hand, with-out being occluded by the virtual object. With depth occlusion-free enabled, the virtual pixels can be covered up by the real pixels, making placing a virtual car behind a real hand, would make the virtual car in-visible or disappear. Figure 5 (a) the real hand covers the virtual car; it still appears on occluded by it. In Figure 5 (b), the user's hand is behind the virtual car and still appear occluded by it. In this result, it enables the user to place their hand in front of the virtual object and user can still see their real hand, without being occluded by the virtual object.
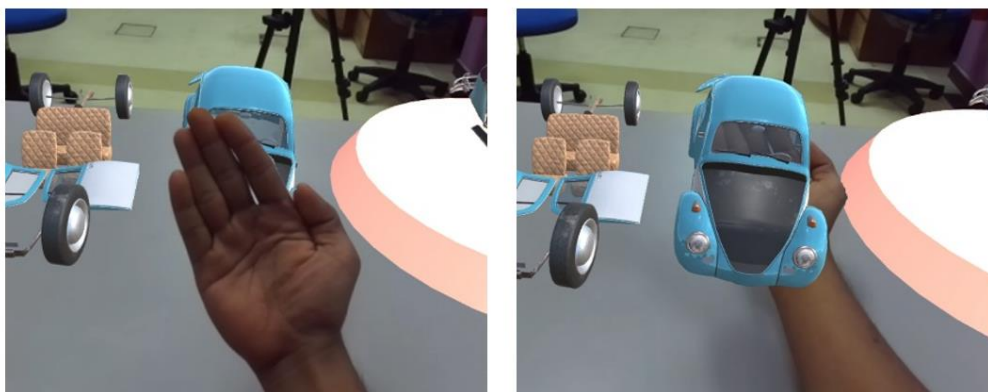
(a) Real hand covers the virtual object   (b) Virtual object covers the real hand.

Figure 5. Occlusion issue in MR for real hand gesture.

4. Application

4.1. Selection Method

The Microsoft Speech API is used to synthesize the speech input, and the process will prompt the user's input from the window application. The speech synthesizer will send the input to be recognized by the application. The process ends with producing speech output. Further research is still continuing to indicate the efficient speech output that potential to complement the gestures input for multimodal interaction. The research also concerns the cues or user's turn in order to respond to the inputs provided by more than one input modality, gesture and speech.

The research also concerns the cues or user's turn in order to respond to the inputs provided by more than one input modality, gesture, and speech. For the proposed multimodal interaction, the user needs to engage with the 3D object using gesture first and then followed with speech command. Giving the speech command first and then invoking gesture input will not trigger the multimodal interaction. Thus, Figure 6 illustrates a flowchart for multimodal interaction technique can be performed in MR.
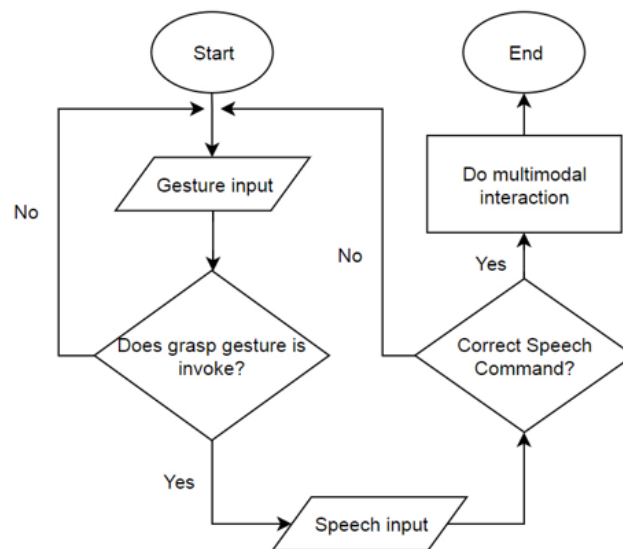
**Figure 6.** Flowchart for multimodal interaction technique.
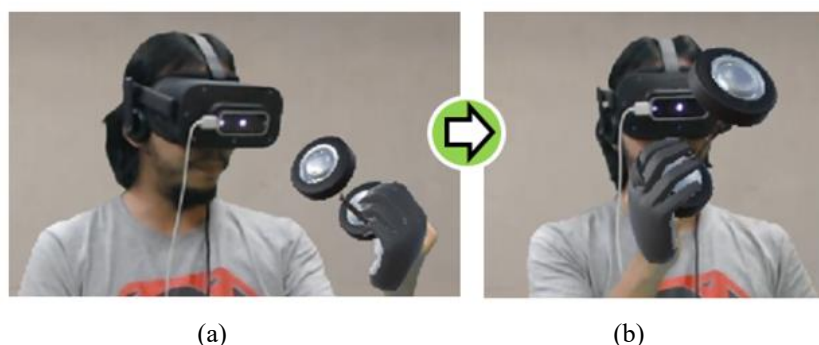
Figure 7. Multimodal interaction using gesture and speech.

Figure 7 shows the representation for multimodal interaction using gesture and speech to be executed. Firstly, the system is checked for the presence of grasp or pinch gesture, and then next it was indicating that the user is currently holding the 3D object (as in Figure 7 (a)). If such gesture input is not existent, the system continues searching for the presence of those gesture input. Until the gesture is found, only then the system allows for speech input to activate. As in Figure 7 (b)), the user performs speech input

to increase scaling factor to two times and it can be seen the object size has been increased up to be bigger. In order to perform the task, the user needs to grasp the desired car parts, and user needs to speak out the right command ("increase size by one") in order for the prototype to recognize the command. If the speech command feeds the system with the correct speech command, multimodal interaction is performed. Otherwise, the system reverts the process back to gesture checking and searching. The significant reason why the cue needs to behave like this, if the gesture is able to work without speech, it will consider gesture-only interaction, and it won't be considered as a multimodal input.

In initial result to test a fascinating and interactive virtual space that lets users virtually explore the assembly of a car. This transformative technology allows users to grasp and manipulate virtual con-tents with simple hand movements that are easy to learn. The objects tested are several parts of the car, which are windows, tires, seat and body. Based on Figure 8, the user explores the VR car model using point gaze through gesture and using speech to invoke a task to change the parts of the car, which indicates the multimodal has been used. Figure 8 (a) shows the user selecting the body of the toy car, and while the car part is selected, the user utters the speech command "change this" as in Figure 8 (a) to perform MMI to change the car part: from the roof to roofless as seen in Figure 8 (b).

(a)   (b)
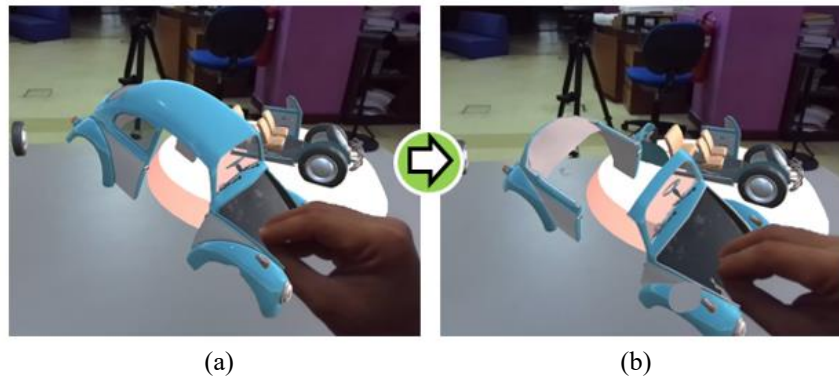
Figure 8. User perform selection using multimodal interaction in MR.

4.2. Scaling

Similar to changing the car part, 3D object scaling also performed while the user selecting the car part: tires as in Figure 9 (a), and the user utters the speech command "increase the size by one" to scale the tires as illustrates in Figure 9 (b). The speech command can be uttered more than once if the user wants to invoke the same interaction. In order to perform uniform during object manipulation, the grasping and pinching gestures are not required, but instead, the hover gesture is used for this particular interaction. The user can perform hover gestures anywhere in the scene, and while it is performed, speech command "select all" need to be uttered to perform 3D objects selection, and all 3D objects in the scene will be selected.

(a)   (b)
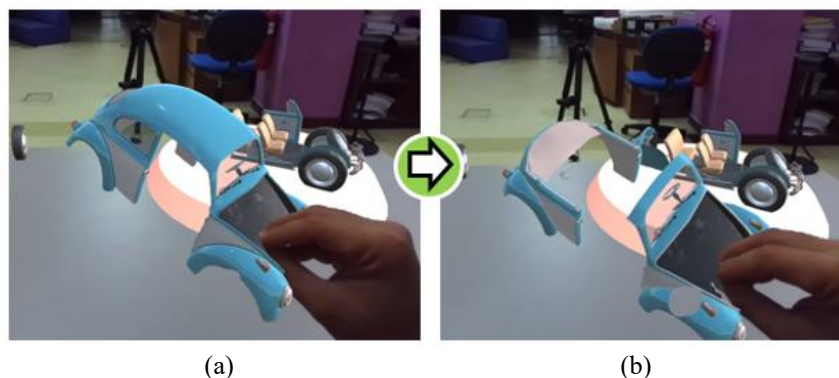
Figure 9. User perform scaling using multimodal interaction in MR.

5. Conclusion and Future Works

This paper discusses the multimodal interaction technique using gesture and speech that the experimental setup to proposed multimodal in the MR interface. MR is rapidly becoming an enjoyable experience in exploring new techniques on tracking and interaction. The intuitive multiple objects manipulation with gestures was also involved in MR. The gesture data was captured by Leap Motion has been integrated to produce a robust hand tracking technique. The gesture is combined with speech inputs in MR to perform spatial object manipulations for MR.

During the gesture recognition process, the gesture inputs were captured by the Leap Motion device that is attached to the Oculus Rift. The head movement tracking corresponds to hand tracking. Speech input is acquired from the build-in microphone in Oculus Rift, and the speech recognition system is executed right after the enable button is pressed and will remain running until the disable button is pressed or the system is close. While for the speech input, the speech grammars list has been defined and combined with gesture interaction to perform multimodal, which has been analyzed to accept a maximum of four words. The grammar panel is required to give the user a handler and feedback to determine whether to enable or disable the speech recognition process. The speech commands need to store and load asynchronously.

Further extension for the future works, the use of current technologies such as Microsoft HoloLens, or Project North Star by Leap Motion [29] can improve the user experienced by the whole as this research only adopt an opaque Oculus Rift for display with the RGB-D camera attached to it. The user can directly see their surrounding without being occluded by an opaque device, and with this increased the field of view that the user sees from their surroundings. When working collaboratively, it is bound that sometimes different user is selecting the same 3D object, so in this case, there is a need for an algorithm to give priority to whoever touches the 3D object first. Therefore, the collaborative interface such as in [30] can be considered to implement multimodal interaction to leverage conventional interfaces.

References

[1]   Ismail, A., & Noh, Z. (2008). Augmented Reality Theory and Applications. M. Shahrizal, & M. Najib, Advances in Computer Graphics and Virtual Environment, 88-105.

[2]   Ismail A.W., Sunar M.S. Intuitiveness 3D objects interaction in augmented reality using the S-PI algorithm. TELKOMNIKA Indonesian Journal of Electrical Engineering. 2013. 11(7), pp.3561-3567.

[3]   Ismail, A. W., & Sunar, M. S. (2015). Multimodal fusion: gesture and speech input in the augmented reality environment. In Computational Intelligence in Information Systems (pp. 245-254). Springer, Cham.

[4]   Ismail, A. W. (2011). User Interaction Technique With 3D Object Manipulation in Augmented Reality Environment (Doctoral dissertation, Universiti Teknologi Malaysia).

[5]   Xiao, R., Schwarz, J., Throm, N., Wilson, A. D., & Benko, H. (2018). MRTouch: adding touch input to head-mounted mixed reality. IEEE transactions on visualization and computer graphics, 24(4), 1653-1660.

[6]   Meegahapola L., Perera I. Enhanced in-store shopping experience through smart phone-based mixed reality application. In 2017 Seventeenth International Conference on Advances in ICT for Emerging Regions (ICTer). IEEE. 2017. pp. 1-8.

[7]   Marichal S., Rosales A., Perilli F.G., Pires A.C., Bakala E., Sansone G., Blat J. CETA: designing mixed-reality tangible interaction to enhance mathematical learning. In Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and

Services. ACM. 2017. p. 29.

[8]   Thiry, K. E., Wolloko, A., Kingsley, C., Flowers, A., Bird, L., & Jenkins, M. P. (2019, September). Designing Federated Architectures for Multimodal Interface Design and Human Computer Interaction in Virtual Environments. In International Conference on Human Systems Engineering and Design: Future Trends and Applications (pp. 404-410). Springer, Cham.

[9]   Wolf, E., Klüber, S., Zimmerer, C., Lugrin, J. L., & Latoschik, M. E. (2019, October). " Paint that object yellow": Multimodal Interaction to Enhance Creativity During Design Tasks in VR. In 2019 International Conference on Multimodal Interaction (pp. 195-204).

[10] Ismail A.W., Sunar M.S. Multimodal fusion: gesture and speech input in the augmented reality environment. In Computational Intelligence in Information Systems. Springer, Cham. 2015. pp. 245-254.

[11] Piumsomboon T., Altimira D., Kim H., Clark A., Lee G., Billinghurst M. Grasp-Shell vs gesture-speech: A comparison of direct and indirect natural interaction techniques in augmented reality. In2014 IEEE International Symposium on Mixed and Augmented Reality (ISMAR). IEEE. 2014. pp. 73-82.

[12] Ismail, A. W., Billinghurst, M., & Sunar, M. S. (2015, August). Vision-based technique and issues for multimodal interaction in augmented reality. In Proceedings of the 8th International Symposium on Visual Information Communication and Interaction (pp. 75-82).

[13] Cordeiro, D., Correia, N., & Jesus, R. (2015, June). ARZombie: A mobile augmented reality game with multimodal interaction. In 2015 7th International Conference on Intelligent Technologies for Interactive Entertainment (INTETAIN) (pp. 22-31). IEEE.

[14] Sankar, A., & Seitz, S. M. (2016, September). In situ CAD capture. In Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services (pp. 233-243).

[15] Fischbach, M., Wiebusch, D., & Latoschik, M. E. (2016, March). Semantics-based software techniques for maintainable multimodal input processing in real-time interactive systems. In 2016 IEEE 9th Workshop on Software Engineering and Architectures for Realtime Interactive Systems (SEARIS) (pp. 1-6). IEEE.

[16] Abidin, R. Z., Arshad, H., & Shukri, S. A. I. A. (2017). A framework of adaptive multimodal input for location-based augmented reality application. Journal of Telecommunication, Electronic and Computer Engineering (JTEC), 9(2-11), 97-103.

[17] Chen, Z., Li, J., Hua, Y., Shen, R., & Basu, A. (2017, October). Multimodal interaction in augmented reality. In 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC) (pp. 206-209). IEEE.

[18] Zimmer, C., Bertram, M., Büntig, F., Drochtert, D., & Geiger, C. (2017). Mobile Augmented Reality Illustrations that entertain and inform: Design and Implementation issues with the Hololens. In SIGGRAPH Asia 2017 Mobile Graphics & Interactive Applications (pp. 1-7).

[19] Huang, W., Billinghurst, M., Alem, L., & Kim, S. (2018, December). HandsInTouch: sharing gestures in remote collaboration. In Proceedings of the 30th Australian Conference on Computer-Human Interaction (pp. 396-400).

[20] Zimmerer, C., Fischbach, M., & Latoschik, M. E. (2018, March). Space Tentacles-Integrating Multimodal Input into a VR Adventure Game. In 2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR) (pp. 745-746). IEEE.

[21] Chan, W. P., Quintero, C., Pan, M., Sakr, M., Van der Loos, H. M., & Croft, E. (2018). A Multimodal System using Augmented Reality, Gestures, and Tactile Feedback for Robot Trajectory Programming and Execution. In ICRA Workshop on Robotics in Virtual Reality 2018.

[22] Aftab, A. R. (2019, October). Multimodal Driver Interaction with Gesture, Gaze and Speech. In 2019 International Conference on Multimodal Interaction (pp. 487-492). ACM.

[23] Sibirtseva, E., Ghadirzadeh, A., Leite, I., Björkman, M., & Kragic, D. (2019). Exploring

Temporal Dependencies in Multimodal Referring Expressions with Mixed Reality. arXiv preprint arXiv:1902.01117.

[24] Hansberger, J. T., Peng, C., Blakely, V., Meacham, S., Cao, L., & Diliberti, N. (2019, July). A multimodal interface for virtual information environments. In International Conference on Human-Computer Interaction (pp. 59-70). Springer, Cham.

[25] Romanus, T., Frish, S., Maksymenko, M., Frier, W., Corenthy, L., & Georgiou, O. (2020). Mid-Air Haptic Bio-Holograms in Mixed Reality. arXiv preprint arXiv:2001.01441.

[26] Luo, Z., Jingying, C., Guangshuai, W., & Mengyi, L. (2020). A three-dimensional model of student interest during learning using multimodal fusion with natural sensing technology. Interactive Learning Environments, 1-14.

[27] Buchmann, V., Violich, S., Billinghurst, M., & Cockburn, A. (2004, June). FingARtips: gesture based direct manipulation in Augmented Reality. In Proceedings of the 2nd international conference on Computer graphics and interactive techniques in Australasia and South East Asia (pp. 212-221).

[28] Dingler, T., Funk, M., & Alt, F. (2015, June). Interaction proxemics: Combining physical spaces for seamless gesture interaction. In Proceedings of the 4th International Symposium on Pervasive Displays (pp. 107-114).

[29] Motion, L. (2019). Project North Star is Now Open Source-Leap Motion Blog. Leap Motion Blog.

[30] Nor' a, M. N. A., & Ismail, A. W. (2019). Integrating Virtual Reality and Augmented Reality in a Collaborative User Interface. International Journal of Innovative Computing, 9(2).