Intelligent Augmented Reality System based on Speech Recognition

Juin-Ling Tseng
Minghsin University of Science and Technology
No.1, Xinxing Rd., Xinfeng Hsinchu 30401
Taiwan

Received: July 6, 2020. Revised: February 23, 2021. Accepted: March 10, 2021. Published: March 18, 2021.

Abstract—In general, most of the current augmented reality systems can combine 3D virtual scenes with live reality, and users usually interact with 3D objects of the augmented reality (AR) system through image recognition. Although the image-recognition technology has matured enough to allow users to interact with the system, the interaction process is usually limited by the number of patterns used to identify the image. It is not convenient to handle. To provide a more flexible interactive manipulation mode, this study imports the speech-recognition mechanism that allows users to operate 3D objects in an AR system simply by speech. In terms of implementation, the program uses Unity3D as the main development environment and the AR e-Desk as the main development platform. The AR e-Desk interacts through the identification mechanism of the reacTIVision and its markers. We use Unity3D to build the required 3D virtual scenes and objects in the AR e-Desk and import the Google Cloud Speech suite to the AR e-Desk system to develop the speech-interaction mechanism. Then, the intelligent AR system is developed.

Keywords—Intelligent Augmented Reality, Artificial Intelligence, Image Recognition, Speech Recognition, Augmented Reality

I. Introduction

With the rapid improvement of information technologies, 3D augmented reality (AR) technologies have been widely applied. The so-called "AR" adds 3D virtualized technologies to the senses of users to observe the environment. The technologies offer information about the real world as well as present virtual information. The complementation and stacking of both types of information can lead to a complete picture of the environment where the users are situated.

The AR technology primarily includes multimedia, 3D modeling, real-time video display control, multi-sensor fusion, real-time tracking, and scene fusion. In general, AR has three characteristics: (1) combined information of real and virtual worlds, (2) immediate interactivity, and (3) ability to add the locations of 3D virtual objects in 3D space.

The 3D AR technology has been extensively applied in different areas including military, medicine, architecture, engineering, film and television, and entertainment [1-8]. However, there are three main display devices, namely the headgear, mobile, and open-space devices. We will introduce them later in this section.

A. *Headgear:*

Headgears are well-known display devices, among which Microsoft HoloLens, as shown in Fig.1, is the most representative. Microsoft HoloLens [9] is a key device for using Windows Holographic as a pair of smart glasses operating through Windows 10. It relies on advanced sensors, high-definition 3D optical head-mounted display with all-angle transparent screens, as well as surround sound. It allows users to communicate with each other through eye contact, speech, or hand gestures in the user interface in AR.

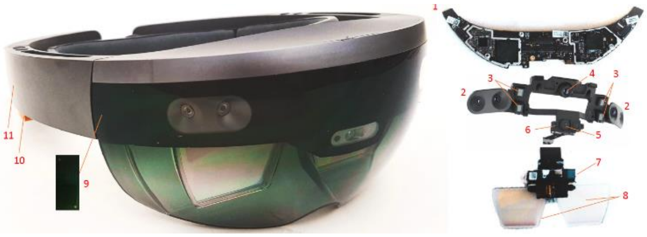


Fig. 1. Microsoft Hololens [9]

As the HoloLens headgear boosts the function of Holographic, it is not limited to applications in virtual reality (VR) or AR. In other words, it should be used more widely. In addition, what significantly distinguishes HoloLens from other headgear is that the former is in itself a computer capable of independent calculations without the need to connect to computer servers. Further, HoloLens can also deliver or receive system information in VR through wireless transmission, so that users can operate in a completely wireless and unrestricted manner.

Despite the magnificent 3D display functions of HoloLens

DOI: 10.46300/9106.2021.15.20

and other AR headgears, their interaction control is mostly through gesture operation, and it is difficult to interact more intuitively with the real objects in the scene. In addition, each user should be equipped with a set of HoloLens to access the AR system, and each set is expensive, while there is still room for improvement in the technology to simultaneously present the system to multiple people.

B. *Mobile devices:*

When you submit your final version, after your paper has been accepted, prepare it in two-column format, including figures and tables.

Presenting AR systems with mobile phones is a less expensive option, compared to using a headgear. Currently, a large number of smart phones have AR functions. In particular, the depth camera in the dual lens would monitor real information, which is used to construct the necessary real scene and realize AR combining 3D virtual objects. Currently, the mobile phones with AR features in the market include ASUS Zenfone AR, Lenovo Phab 2 Pro, and Apple iPhone X.

For example, ASUS Zenfone AR [10, 11], which is a 5.7-inch smart phone, is the first in the world to adopt the Tango system and support Daydream, as shown in Fig. 2. The Tango system is an innovative technology developed by Google for AR, which also supports the Daydream View headgear designed by Google for mobile VR.

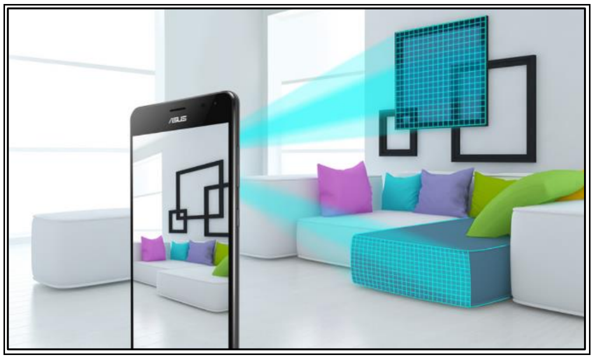


Fig. 2. AR application for interior design using ASUS Zenfone [10]

C. *Open-space devices:*

Open-space AR devices often vary in size. They can be as small as a table and as large as an entire room. Such devices usually can be shared by multiple users for watching and even interacting with one another simultaneously. In general, it is often used in open work display or other highly interactive environments. For example, as shown in Fig.3, at the 2019 Summer Travel Exhibition in Taiwan [12], the original AR rock climbing facility in the "Go Star Challenge Extreme Sports Center" in a duty-free shop was converted into an AR motion sensing game. The audience can experience the game at the scene.

AR system provides headgear, mobile devices and other related devices for image rendering. However, it still lacks the interaction friendly. In order to solve this problem, this study imported speech recognition technology, so that AR users can directly control the AR system. At present, many speech recognition technologies have been proposed. In this study, we must consider two important factors when importing speech recognition. One is the accuracy of speech recognition, and the other is the performance of speech recognition.

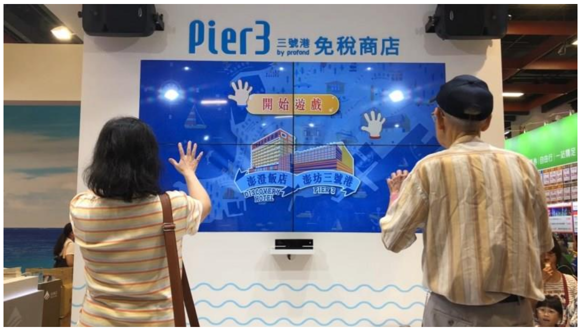


Fig. 3. The 2019 Summer Travel Exhibition in Taiwan [12]

Filippidou and Moussiades [21] compared the accuracy of three different speech recognition technologies, including IBM, Google, and Wit, in 2020. They used three Speaker for measurement analysis. It showed that Google had a lower Word Error Rate (WER) than IBM and Wit. Google also provided better speech recognition accuracy than IBM and Wit. In terms of real-time performance analysis, this study also performed the frame-per-second (FPS) monitoring through the Unity Profiler. From the experimental results, it can be found that this study can still maintain the performance of higher than 30FPS after the introduction of Google Speech System. That is, the system can also achieve the goal of real-time in our system.

II. Relevant Studies

With the development of Internet and information technologies, AR technologies have been gradually applied in many areas of daily life such as games [13], education [14], industry [15] and other related fields. These areas are introduced next:

A. *Games:*

Pokemon Go is a location-based AR action game, as shown in Fig.4. In 2016, it reached every corner of the world, with rapidly increasing popularity and considerable economic revenue. Through 1,000 qualitative surveys on positive and negative feelings of users in experiencing Pokemon Go,

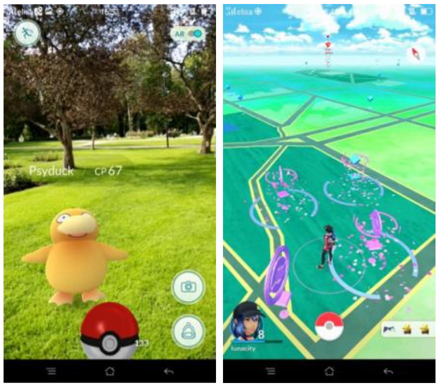


Fig. 4. Two in-game scenes from Pokémon GO. Capturing the Pokémon (left), moving in the virtual world (right). [13]

Paavilainen et al. [13] discovered that players generally think highly of the mobility and social functions, game mechanics, and branding. However, they reported negative experience because of the occurrence of technical issues, unequal game opportunities, erroneous behaviors of other players and non-players, and unsophisticated game design. Such results offer useful reference for research and design for the academia and industry professionals.

B. Education:

Barrow et al. [14] stated in their paper in 2019 that different levels of education, from public outreach activities to the academic field, have the opportunity to utilize the AR technology. The greatest appeal of AR is that the virtual and real environments can offer learners a blended learning experience that has different types including learning by teaching, learning by experience, and learning by exercise. Based on the diversity in learning experience through AR, Barrow et al. developed an AR-focused application that allows learners to switch from an ordinary learning environment to a highly interactive space, as shown in Fig.5. They also addressed the potential influence of intervening in advanced education (undergraduates or postgraduates).

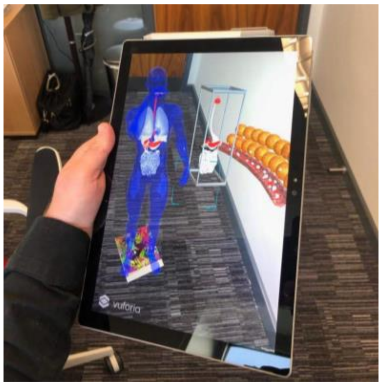


Fig. 5. Barrow et al. developed an AR-focused application in Windows tablet for education survey. [14]

C. Industry

AR is one of the essential technologies of Industry 4.0 and an important tool for producing the next generation of automated and digitalized factories. In the case of shipyards, as relevant information on ship manufacturing is essential for workers, the AR technology can assist them in acquiring the necessary information conveniently and efficiently. One of the ten biggest ship manufacturers, Navantia, has been actively studying the application of AR in the shipyard industry. Fernández-Caramés et al. [15] published a paper in 2018 on Sensors, titled "A Fog Computing and Cloudlet Based Augmented Reality System for the Industry 4.0 Shipyard." They explained the industrial AR framework of Navantia, operated based on Cloudlet and Fog Computing, as shown in Fig.6.

In this framework, AR is displayed using Microsoft HoloLens, as shown in Fig.7. The experimental results show that at a small capacity (less than 128 KB), the industrial AR is considerably fast in transmission and that when the transmitted file is relatively large, it still exhibits satisfactory performance, compared to those of other systems.

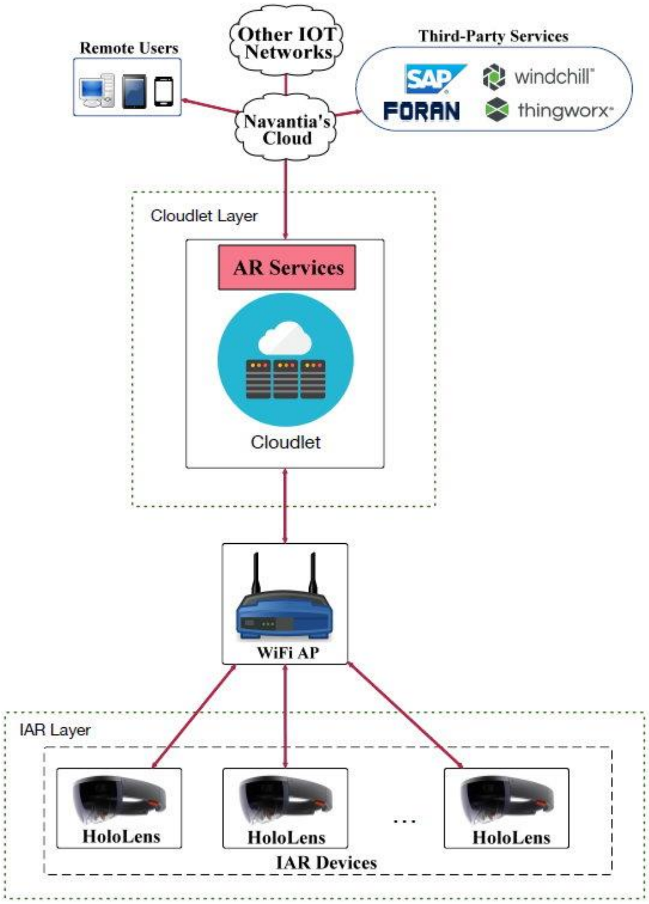


Fig. 6. The industrial AR framework of Navantia. [15]

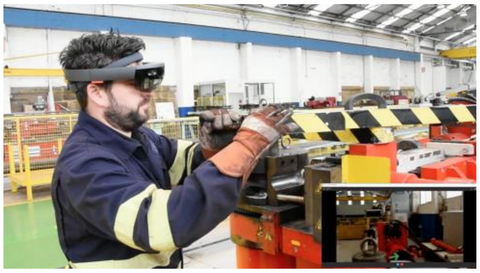


Fig. 7. In the industrial AR framework, it is displayed using Microsoft HoloLens. [15]

D. Other related fields

Besides games, education, industry, image recognition and speech recognition are often involved in facial image tracking, remote control of vehicles and other applications. Moreano and Palomino [17], for example, used the Super Vector Machine in 2020 to improve the efficiency of facial image recognition. They calculated and analyzed the facial images in the FERET and MUCT databases. Their method achieved more than 96% in the image recognition accuracy.

In addition, Pantazoglou et al. [18] developed a set of automatic speech recognition technology in the conjunction of

Greek voice recognition technology and machine learning. This automatic speech recognition technology was applied in the remote control vehicles. It effectively reduced the computing costs of remote control vehicles, and improved its accuracy, as shown in Figure 8.

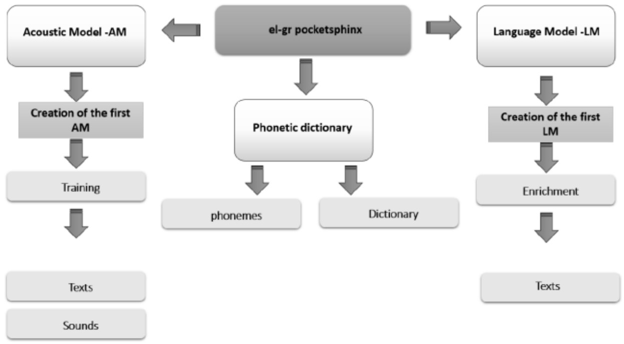


Fig. 8. The generic Greek model flow chart proposed by Pantazoglou et al. [18]

III. Design and Implementation

A. Conceptual Design

To create this intelligent AR interactive system, our research adopted Unity3D as the main software for development. Unity3D is not only a multi-platform 3D developing software; its applications can also integrate peripheral devices and systems of different kinds including the open-space AR system devices and the development environment for the Google Cloud Speech Recognition required by our research. Therefore, Unity3D is extremely suitable as a 3D integrated development software in our study. In terms of the open-space AR system, we built an AR interactive e-desk system that relies on reacTIVision [16] for operation. For the function of speech recognition in the e-desk system, we also introduced the Unity Package from Google Cloud Speech Recognition, as shown in Fig.9.

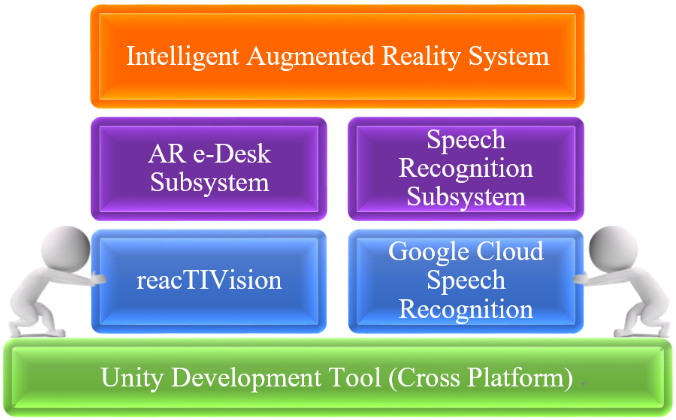

Intelligent Augmented Reality System
AR e-Desk Subsystem
Speech Recognition Subsystem
reacTIVision
Google Cloud Speech Recognition
Unity Development Tool (Cross Platform)

Fig. 9. The conceptual design proposed by this study.

B. Development of AR e-Desk system with reacTIVision

The so-called reacTIVision [16] is an open-source, multi-platform computer visual development environment. This method used TUIO client application to project interactive images through the projector into the interactive plane, where a user can locate the fiducial markers. The reacTIVision used the camera to identify the markers and determined the user's interaction instructions. That is, the user can give instructions using the reacTIVision method, as shown in Figs. 10 and 11.

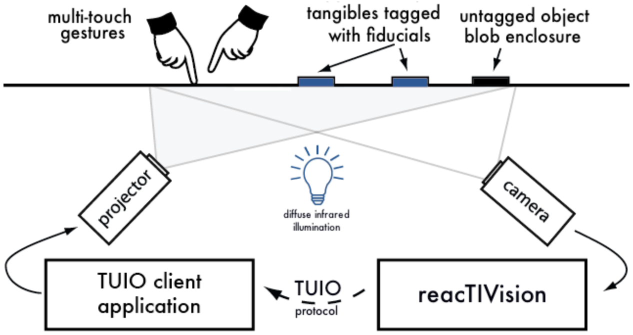

multi-touch gestures
tangibles tagged with fiducials
untagged object blob enclosure
projector
diffuse infrared illumination
camera
TUIO client application
TUIO
protocol
reacTIVision

Fig. 10. The development concept of reacTIVision.

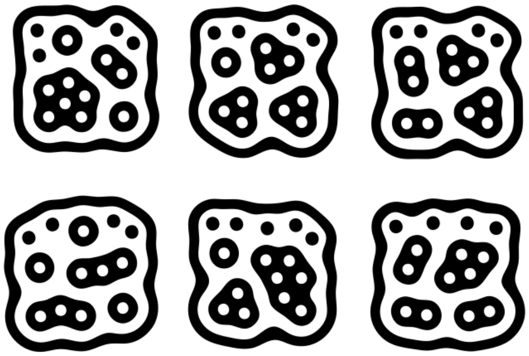


Fig. 11. The fiducial markers of reacTIVision.

It can quickly and accurately recognize images based on fiducial markers and allow for multi-touch operation. In addition, reacTIVision is a development framework for table-based tangible user interface developed by Martin Kaltenbrunner and Ross Bencina. It can be implemented on Windows, Mac OS, and Linux systems.

C. Development of required speech recognition function using Google Cloud Speech Recognition

The Google Cloud Platform is a cloud computing service leveraging Google’s core architecture, data analysis, and machine learning. Google Cloud Speech Recognition, on the other hand, is a speech-recognition service provided on the Google Cloud Platform. This technology can transform voice messages into texts through machine learning. With the Google Cloud Speech API on the Google Cloud Platform, developers can use its powerful and most advanced deep learning neural network algorithm to turn speech into words. The API can recognize more than 80 languages and dialects, thus supporting users from different regions globally.

This study adopts the length-linear function and unigram-and-bigram function [19], as shown in Equations 1 and 2, proposed by Aleksic et al. in the Google Speech Recognition to construct the contextual model. This design provides fairly accurate recognition results.

The length-linear function is defined as follows:

$$s_B(w|H) = f_1(length(Hw)) = (n-1)p_2+p_1 \qquad (1)$$

The unigram-and-bigram function is defined as follows:

$$s_B(w|H) = f_2(length(Hw)) = p_1\text{ , if } n=1 \qquad (2)$$
$$s_B(w|H) = f_2(length(Hw)) = p_2\text{ , if } n>=2$$

where $s_B(w|H)$ denotes a raw biasing model, w is the analysis word, H is a history state, Hw is the language model score of n-grams, and p_1 and p_2 indicate the control parameters for biasing model.

In general, word error rate (WER) [20, 21] is usually used to measure the accuracy of automatic speech recognition (ASR), as shown in Equation 3.

$$WER = (S+D+I) / N \qquad (3)$$

where S indicates the number of replacements, D denotes the number of deletions, I is the number of inserts, and N indicates the number of analysis words.

D. System implementation

The smart AR interaction systems that we planned to develop include the AR e-desk subsystem and speech recognition subsystem, as shown in Fig.12. In particular, the AR e-desk subsystem involves the AR virtual-scene construction module, AR virtual-scene map-construction module, AR image-recognition module, and AR interactive module, whereas the speech-recognition subsystem encompasses the speech input interface module, speech-analysis module, speech response module, etc. The main functions of each module are described as follows:

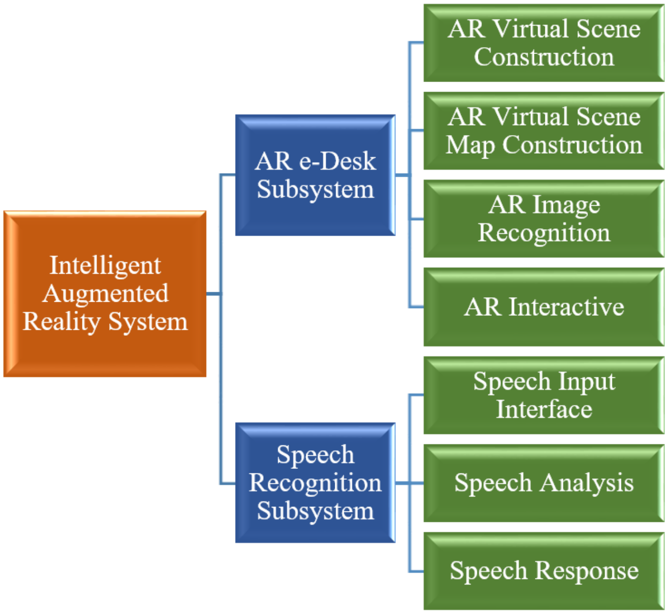

Intelligent Augmented Reality System
AR e-Desk Subsystem
Speech Recognition Subsystem
AR Virtual Scene Construction
AR Virtual Scene Map Construction
AR Image Recognition
AR Interactive
Speech Input Interface
Speech Analysis
Speech Response

Fig. 12. System structure.

➢ AR virtual-scene construction:

The main function of this module is to construct the required AR virtual scenes through Unity3D, whereas the required 3D related models within the scenes would be produced from 3D Studio Max.

➢ AR virtual-scene map construction:

The module is mainly used to build a map for the 3D virtual scenes, and the map helps outline the necessary limits for recognition in the AR image-recognition module.

➢ AR image recognition:

This module is used to recognize the required fiducial markers. To this end, we have introduced the TUIO suite to facilitate the recognition.

➢ AR interactive:

This module can receive the results from the AR image-recognition module and offer a certain response in the system based on the results including the control of rotation of 3D models.

➢ Speech input interface:

In addition to interaction using images, to enable easier interaction, we also created a speech input interface, which can only be operated in conjunction with the speech-analysis module. This current module is mostly used to read the users' speech content.

➢ Speech analysis:

This module processes the input of speech content from the speech input interface module and analyzes the content with Google Speech API. It then transcribes the speech into words and transmits them to the speech response module.

➢ Speech response:

The module receives the texts produced by the speech-analysis module and responds to the system based on the content, so that the 3D models in the system can react accordingly.

IV. Results of Implementation

Our system was developed using a CPU of Intel Core i7-6700HQ 2.6 GHz, 8 GB memory, and Windows 10 operating system. We wrote the program and integrated the system with Unity3D, C# programming language, Visual Studio, reacTIVision suite, Google Cloud Speech Recognition, and other software. Finally, we presented it in the form of an AR e-desk system.

The system is divided into the AR e-desk and speech-recognition subsystems. The AR e-desk subsystem aims to act as a mechanism that allows for AR interaction through image recognition. Thus, we adopted the reacTIVision suite and its fiducial marker image and developed four key modules, namely the AR virtual-scene construction, AR virtual-scene map construction, AR image recognition, and AR interactive. Each of these is presented in an AR e-desk, as shown in Figs. 13 and 14.

In addition, considering interactive control, we utilized the Google Cloud Speech API to operate the speech-recognition subsystem and developed three modules, namely speech input interface, speech analysis, and speech response, for speech interactive control. In particular, this subsystem features a function of speech control in the AR system. To this end, we imported the Google Cloud Speech API to develop the necessary speech modules including StartRecord(), StopRecord(), StartedRecordEventHandler(), FinishedRecord EventHandler(), and other functions related to speech recording, as shown in Fig. 15.

In addition, we built a speech-analysis module to analyze speech and to recognize the operational commands in the speech. Thus we created relevant programs for speech analysis including SetLanguage(), RecognitionSuccessEvemtHandler(), GetOperationDataSuccessEventHandler(), SetContext(), Recognize(), and ApplySpeechContextPhrase(). The speech response module produces corresponding responses based on the results from the speech-analysis module.

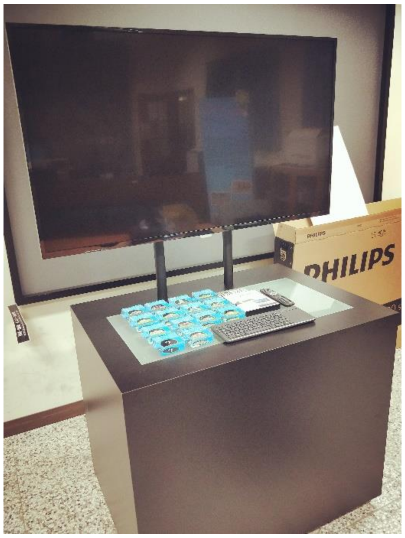


Fig. 13. AR e-desk.

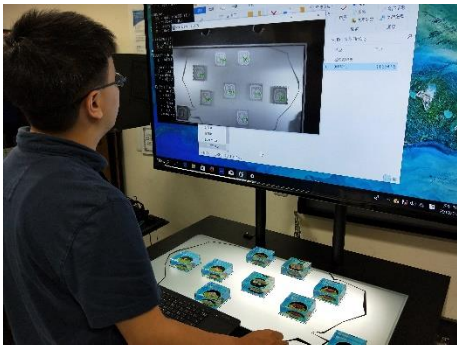
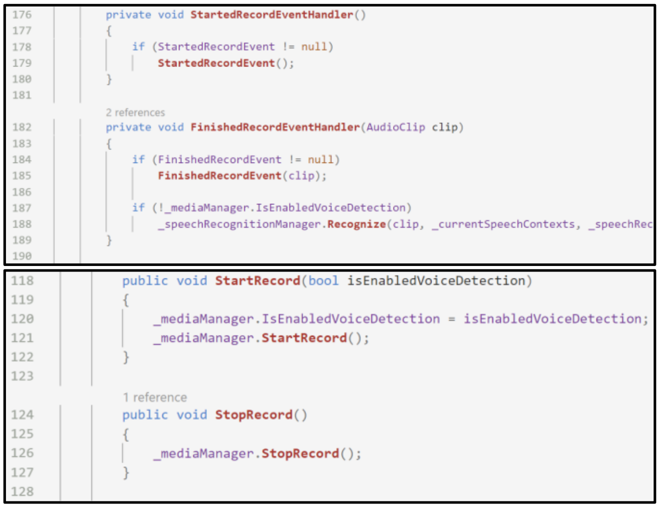


Fig. 14. This study imports fiducial markers from the reacTIVision suite



```
        private void StartedRecordEventHandler()
        {
            if (StartedRecordEvent != null)
                StartedRecordEvent();
        }

        2 references
        private void FinishedRecordEventHandler(AudioClip clip)
        {
            if (FinishedRecordEvent != null)
                FinishedRecordEvent(clip);

            if (!_mediaManager.IsEnabledVoiceDetection)
                _speechRecognitionManager.Recognize(clip, _currentSpeechContexts, _speechRec
        }

```

```
        public void StartRecord(bool isEnabledVoiceDetection)
        {
            _mediaManager.IsEnabledVoiceDetection = isEnabledVoiceDetection;
            _mediaManager.StartRecord();
        }

        1 reference
        public void StopRecord()
        {
            _mediaManager.StopRecord();
        }

```

Fig. 15. The functions of speech recording in the Speech Input Interface Module

To integrate the AR e-desk and speech-recognition subsystems into an intelligent AR system, we implemented a game system, named AR Tank War to test the interactive mechanism of AR and speech recognition. This game system is developed around the theme of tank war, where apart from the tank controlled by the player, there are three enemy tanks controlled by the computer. Each tank would patrol in its own defense area. When the player tank enters these areas, the enemy tanks would fire on the player tank until the latter leaves. Meanwhile, the player tank must defeat the enemy within the time limit, as shown in Fig. 16.

With the interactive design through reacTIVision and speech recognition, players can control their tank with operations such as forward, backward, left, right, and attack, as shown in Fig.17.

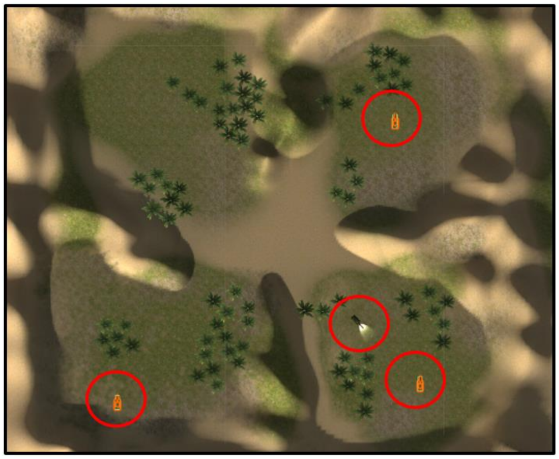


Fig. 16. The system scene

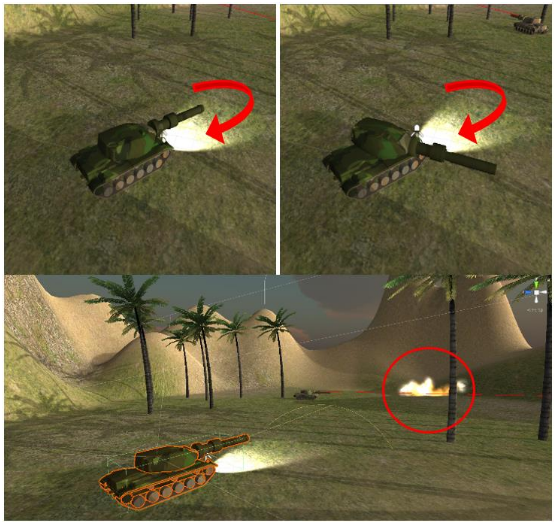


Fig. 17. Players can use the reacTIVision and speech recognition to control their tanks such as forward, backward, left, right, and attack

For example, the user can give the attack instructions by voice when playing, and steer the direction of tank through the fiducial markers at the same time.

V. Performance Analysis

In intelligent augmented reality system, in addition to image recognition, speech recognition and other interactive mechanisms, real-time is also very important. That is, its execution effectiveness must reach more than 30 frames per second (FPS). To monitor whether the system can achieve the goal of real-time, this study analyzes the computational efficiency of the developed system.

To analyze the performance of the AR e-desk subsystem and the speech-recognition system, we utilized the profile analyzer from Unity to extract 299 frames from the game operation, and listed the 100-frames execution time in the 299 frames, as shown in Table 1. The game testing involves movement control and attacks of the player tank, as well as the patrol detection and attacks of the enemy tank, including such operations as explosion of enemy tanks after the firing.

According to the experimental results in Table 1, the entire execution time of the 100 frames was 2.079 seconds, of which the speech recognition portion accounted for 41.07 percent. Of all 100 frames, only 3 frames, including frame 35, frame 39 and frame 81, have an execution time of more than 30ms, and most of the frames have an execution time of around 20ms. That is, its execution results are quite good.

Table1. The execution time of AR tank game (CPU usage in milliseconds)

Frame	SR	AR TW	Total Time	Frame	SR	AR TW	Total Time
1	7.84	11.96	19.8	51	7.42	12.06	19.48
2	11.54	11.41	22.95	52	8.17	13.46	21.63
3	8.84	12.06	20.9	53	9.88	12.51	22.39
4	8.71	11.96	20.67	54	8.07	12.3	20.37
5	7.66	10.66	18.32	55	7.48	12.45	19.93
6	9.26	11.8	21.06	56	12.9	11.51	24.41
7	7.61	12.31	19.92	57	7.47	13.34	20.81
8	7.79	14.73	22.52	58	9.64	11.56	21.2
9	7.83	17.33	25.16	59	8.47	12.68	21.15
10	7.21	14.04	21.25	60	7.34	12.5	19.84
11	10.75	12.14	22.89	61	7.54	14.13	21.67
12	8.47	12.34	20.81	62	10.16	11.51	21.67
13	8.13	12.4	20.53	63	7.89	12.39	20.28
14	7.74	11.1	18.84	64	8.32	11.7	20.02
15	9.84	12.44	22.28	65	7.38	11.95	19.33
16	8.39	10.62	19.01	66	9.3	10.67	19.97
17	7.73	11.46	19.19	67	7.88	11.1	18.98
18	7.46	11.46	18.92	68	8.01	11.7	19.71
19	9.44	11.49	20.93	69	7.61	11.04	18.65
20	8.18	11.56	19.74	70	9.75	10.64	20.39
21	8.02	11.6	19.62	71	7.9	12.1	20
22	7.63	14.22	21.85	72	8	11.25	19.25
23	7.53	13.11	20.64	73	7.74	12.81	20.55
24	10.06	12.81	22.87	74	7.69	11.75	19.44
25	7.81	14.78	22.59	75	9.55	10.39	19.94
26	7.9	10.87	18.77	76	7.87	11.98	19.85
27	7.76	11.24	19	77	7.68	10.78	18.46
28	9.97	11.26	21.23	78	7.5	11.59	19.09
29	8.01	11.23	19.24	79	9.48	11.85	21.33
30	7.71	11.41	19.12	80	8.76	10.94	19.7
31	7.62	11.02	18.64	81	7.68	24.43	32.11
32	9.63	11.87	21.5	82	7.27	11.97	19.24
33	7.88	10.96	18.84	83	9.77	11.2	20.97
34	7.61	12.41	20.02	84	7.59	11.59	19.18
35	16.79	14.4	31.19	85	7.35	12.21	19.56
36	10.61	14.4	25.01	86	7.57	10.79	18.36
37	8.11	12.27	20.38	87	9.32	11.02	20.34
38	8.27	11.58	19.85	88	7.61	11.15	18.76
39	8.1	24.49	32.59	89	7.48	11.93	19.41
40	7.95	12.26	20.21	90	7.74	10.78	18.52
41	9.63	12.43	22.06	91	7.51	12.07	19.58
42	8.23	11.12	19.35	92	9.18	12.07	21.25
43	7.69	11.31	19	93	8	11.33	19.33
44	7.66	12.42	20.08	94	15.72	11.93	27.65
45	9.76	10.4	20.16	95	8.71	10.71	19.42
46	8.32	12.54	20.86	96	10.18	13.3	23.48
47	7.86	11.9	19.76	97	8.39	12.18	20.57
48	7.74	11.23	18.97	98	7.52	13.46	20.98
49	9.58	11.39	20.97	99	7.46	13.76	21.22
50	7.69	10.45	18.14	100	9.47	12.05	21.52

Total execution time of SR: 853.92ms
Total execution time of ARTW: 1225.22ms
The complete total execution time: 2079.14ms
The Ratio of SR: 41.07%

SR: Speech Recognition
ARTW: AR Tank War (without Speech Recognition)

We analyzed the performances of speech recognition and AR tank war separately in the analysis. During the experiment on game-operation performance, the overall average frame time is 20.85 ms, in which speech recognition took 8.49 ms and AR tank war took 12.36 ms; thus, the average performance can exceed 30 FPS, as shown in Figs. 18, 19 and Table 2.

Table2. Performance analysis (CPU usage)

Execution Function	Average Frame Time (ms)	Max Frame Time (ms)	Min Frame Time (ms)
Speech Recognition	8.49 (41%)	16.79	7.04
AR Tank War (without Speech Recognition)	12.36 (59%)	26.76	10.03
Total	20.85		

In addition, the maximum frame times for speech recognition and AR tank war are 16.79 ms and 26.76 ms, respectively. Although both frame times are twice their averages, the frames that produce these two maximums are not the same and do not continue to occur. Therefore, they do not result in the immediacy of interaction in the system. According to the above performance analysis, the system can still maintain real-time interaction mechanism after importing speech recognition.

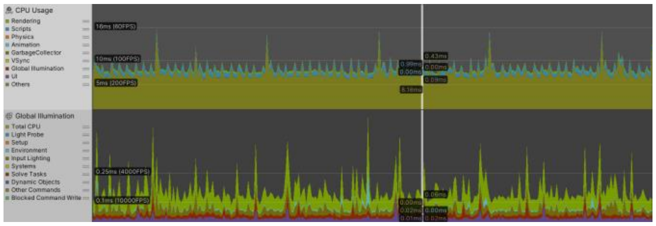


Fig. 18. Execution performance of speech recognition

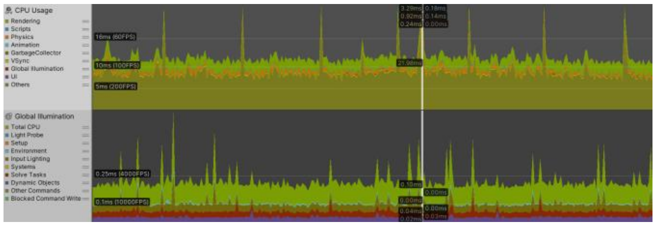


Fig. 19. Performance of game execution (AR Tank War without speech recognition)

VI. Discussion

According to the performance analysis results of Section5, the system can indeed carry out the system requirements, including real-time image processing, real-time speech recognition, high accuracy of speech instructions and so on. But in fact, not all image recognition technology and speech recognition technology can achieve these goals. This is one of the reasons why this study used the reacTIVision for image recognition interaction. From the experimental results, the high-performance advantage of reacTIVision effectively improved the performance of the system.

Although this system can achieve real-time performance after integrating the reacTIVision and Google Speech Recognition, the complexity of system content, such as the number of triangles in 3D models, is also one of the key factors affecting execution performance. Therefore, developers still have to control the complexity of system content to meet the real-time requirements.

VII. Conclusions

To develop an intelligent AR interaction system, we used Unity3D as our main operating environment and built an intelligent AR interactive desk game system. It combines the reacTIVision fiducial marker image-recognition technology and Google Cloud Platform, manipulating objects in the game using both image and speech recognition. To leverage the reacTIVision fiducial marker image-recognition technology, we introduced TUIO and the Uniducial library to facilitate the recognition of fiducial marker images and therefore control the 3D objects in the game system. In addition, on the Google Cloud Platform, we adopted Google Cloud Speech API for speech recognition, so that commands can be given through voice (such as attack, forward, and backward). After the integration, this game system no longer requires a mouse or keyboard for control and relies entirely on image and speech recognition for interaction. Image recognition and speech recognition are also the two main areas that drive AI applications at present. If these two technologies can be effectively applied to the game system under the condition of real-time execution, it will help the development of intelligent games. In addition, our system is based on Unity, an instrument software that enables multi-platform development. This means that our development content also has the feature of multi-platform porting and development. Therefore, our system offers a flexible space for multi-platform development in the future.

VIII. Future Work

In recent years, most AR systems have concentrated on interaction technologies. However, our research has incorporated image and speech recognition, which is a step forward for 3D systems towards intelligent technologies. It is expected that there would be more intelligent technologies integrated into 3D interaction systems to provide more superior human operation and interaction experience.

Acknowledgment

We would like to thank Axis 3D Technology, Inc. for industry cooperation. This work was supported by the Ministry of Science and Technology, and the Ministry of Education, Taiwan, under Contract No. MOST 109-2622-E-159-005, MOST 108-2622-E-159-001-CC3, MOST 107-2622-E-159-002-CC3 and PSK1090413.

References

[1] J. L. Tseng, "Development of a 3D Intuitive Interaction Interface for Head-Mounted Virtual Reality System", International Journal of Advanced Studies in Computer Science and Engineering, Vol.7, No.7, pp.19-25, 2018.

[2] F. Soffel, M. Zank, and A. Kunz, "Postural stability analysis in virtual reality using the HTC vive", 22nd ACM Conference on Virtual Reality Software and Technology, pp. 351-352, Nov. 2016.

[3] J. L. Tseng and Chia-Wei Chu "Interaction Design in Virtual Reality Game using Arduino Sensors", Simulation and Gaming, Chapter 7, pp.111-127, 2018.

[4] J. McGhee, B. Bailey, R. G. Parton, N. Ariotti, and A. Johnston, "Journey to the centre of the cell (JTCC): a 3D VR experience derived from migratory breast cancer cell image data", ACM SIGGRAPH ASIA 2016 VR Showcase, pp. 11, Nov. 2016.

[5] J. L. Tseng, "Raising the Learning Effects for Learners with Low Entrance Scores using Project-Based Learning in Virtual Reality Practice", IAENG International Journal of Computer Science, Vol.47, No.3, pp.516-521, 2020.

[6] E. Lemle, K. Bomkamp, M. K. Williams, and E. Cutbirth, "Virtual Reality and the Future of Entertainment", Two Bit Circus and the Future of Entertainment, Springer, pp. 25-37, 2015.

[7] L. Olcomendy, F. Santos-Cessac, Ph. Dondon, “Design of a Low Cost LIDAR Scanning System for Didactical Applications”, NAUN International Journal of Circuits, Systems and Signal Processing, Vol.13, pp.366-372, 2019.

[8] N. Stoimenov, D. Karastoyanov, “Innovative Approach for 3D Presentation of Plane Culturally-Historical Objects by Tactile Plates for Disadvantaged Users (Low-sighted or Visually Impaired)”, NAUN International Journal of Computers, Vol.13, pp. 84-88, 2019.

[9] Y. Liu, H. Dong, L. Zhang and A. E. Saddik, "Technical Evaluation of HoloLens for Multimedia: A First Look," IEEE MultiMedia, Vol.25, No.4, pp. 8-18, 2018.

[10] Zenfone AR, “ASUS Zenfone AR”, https://www.asus.com/Phone/ ZenFone-AR-ZS571KL/

[11] Y. Aoki, I. Funatsu, “Development of a Teaching Aid for Teaching Dynamic Motion Using the Tango Platform”, Vol.18, No.1, pp.7-10, 2019.

[12] Farsail, “AR Somatosensory Game-2019 Taipei International Summer Travel Show”, Farsail, 2019, https://farsail-tw.com/news/1947/.

[13] J. Paavilainen, H. Korhonen, K. Alha, J. Stenros, E. Koskinen, F. Mayra, “The Pokémon GO Experience: A Location-Based Augmented Reality Mobile Game Goes Mainstream”, 2017 CHI Conference on Human Factors in Computing Systems, pp.2493-2498, May 6–11, 2017.

[14] J. Barrow, C. Forker, A. Sands; D. O'Hare, W. Hurst, “Augmented Reality for Enhancing Life Science Education”, The Fourth International Conference on Applications and Systems of Visual Paradigms, 2019.

[15] T. M. Fernández-Caramés, P. Fraga-Lamas, M. Suárez-Albela and M. Vilar-Montesinos, “A Fog Computing and Cloudlet Based Augmented Reality System for the Industry 4.0 Shipyard”, Sensors, Vol.18, No.6, 2018.

[16] ReacTIVision, “A Toolkit for Tangible Multi-Touch Surfaces”, http://reactivision.sourceforge.net/

[17] J. A. C. Moreano, N. B. L. S. Palomino, “Efficient Technique for Facial Image Recognition With Support Vector Machines in 2D Images With Cross-validation in Matlab”, WSEAS Transactions on Systems and Control, Vol. 15, Art. #18, pp. 175-183, 2020.

[18] F. K. Pantazoglou, G. P. Kladis, N. K. Papadakis, “A Greek Voice Recognition Interface for ROV Applications, Using Machine Learning Technologies and the CMU Sphinx Platform”, WSEAS Transactions on Systems and Control, Vol. 13, Art. #63, pp. 550-560, 2018.

[19] P. Aleksic, M. Ghodsi, A. Michaely, C. Allauzen, K. Hall, B. Roark, D. Rybach, P. Moreno, “Bringing Contextual Information to Google Speech Recognition”, 16th Annual Conference of the International Speech Communication Association, pp. 468-472, 2015.

[20] B. Iancu, “Evaluating Google Speech-to-Text API's Performance for Romanian e-Learning Resources”, Informatica Economica, Vol.23, No.1, pp.17-25, 2019.

[21] F. Filippidou and L. Moussiades. “A Benchmarking of IBM, Google and Wit Automatic Speech Recognition Systems”, Artificial Intelligence Applications and Innovations: 16th IFIP WG 12.5 International Conference, pp.73-82, 2020.

Juin-Ling Tseng received his B.S. from Soochow University, Taipei City, Taiwan, in 1994 and M.S./Ph.D. from Chung Yuan Christian University, Taoyuan City, Taiwan, in 1996/2006. Currently, he is an associate professor and the director of the Department of Multimedia and Game Development at the Minghsin University of Science and Technology, Hsinchu County, Taiwan. His major research fields include 3d modeling, computer animation, game design, virtual reality, augmented reality. He currently is a member of the Institute of Electrical and Electronics Engineers (IEEE), and a senior member of the International Engineering and Technology Institute (IETI). He had the honor to get the awards of Excellent Researcher at Minghsin University of Science and Technology from 2015 to 2020.

Creative Commons Attribution License 4.0 (Attribution 4.0 International, CC BY 4.0)