Interaction with Gaze, Gesture and Speech in a Flexibly Configurable Augmented Reality System

Zhimin Wang*, Haofei Wang*, Huangyue Yu, and Feng Lu†, *Member, IEEE*

***Abstract*—Multimodal interaction has become a recent research focus since it offers better user experience in Augmented Reality (AR) systems. However, most existing works only combine two modalities at a time, e.g., gesture and speech. Multimodal interactive system integrating gaze cue has rarely been investigated. In this paper, we propose a multimodal interactive system that integrates gaze, gesture and speech in a flexibly configurable AR system. Our lightweight head-mounted device supports accurate gaze tracking, hand gesture recognition and speech recognition simultaneously. The system can be easily configured into various modality combinations, which enable us to investigate the effects of different interaction techniques. We evaluate the efficiency of these modalities using two tasks: the lamp brightness adjustment task and the cube manipulation task. We also collect subjective feedback when using such systems. The experimental results demonstrate that the *Gaze+Gesture+Speech* modality is superior in terms of efficiency, and the *Gesture+Speech* modality is more preferred by users. Our system opens the pathway towards a multimodal interactive AR system that enables flexible configuration.**

***Index Terms*—multimodal interaction, augmented reality, gaze, gesture, speech, human-computer interaction.**

I. Introduction

Augmented Reality (AR) systems aim at providing immersive experience via overlaying virtual content onto the real environment. Prior researches have extensively explored using different modalities to interact with the virtual content in AR, such as hand gesture [1], [2], [3], [4], and speech [5], [6], [7]. Each modality has its own pros and cons. The hand gesture-based system provides intuitive experience while it has to handle the occlusion problem [8], and it is likely to cause arm fatigue after long-time usage [3], [9]. The speech-based system offers better controllability, however, it requires user to remember and pronounce the verbal command correctly [5]. It also increases user's cognitive workload, especially for complex tasks.

With the development of eye tracking technology, it is possible to investigate the eye gaze behavior when using AR systems [10], [11]. Visual system is vital for us since more than 80% of the information received by the brain is from our eyes [12]. As the windows to the soul, eye contact cues increase mutual trust, collaboration and understanding [13], [14]. Eye gaze-based interaction also requires less physical effort, and provides more natural experience than gesture or speech [10], which is potentially an effective channel in wearable HMD systems [15]. However, the gaze-based system often suffers from the Midas Touch problem [16], [17], where users unintentionally trigger a target with every glance. The insufficient eye tracking accuracy also degrades the user experience [10]. Therefore, using single modality can not bring the user optimal experience.

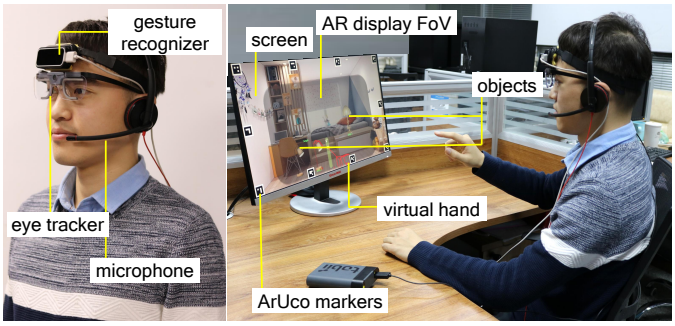

Fig. 1. Left: Gaze-Gesture-Speech AR system (GGS-AR) system setup. Right: the user is interacting with the objects using GGS-AR system.

Multimodal interactive systems seek to provide better usability by taking advantage of each modality. These systems usually combine two modalities at a time, *i.e.*, *Gesture+Speech* [18], [5], [6], *Gesture+Gaze* [10], [19], *Gaze+Speech* [20], [21] and *Gaze+Electroencephalography* [22], [23]. Each modality is assigned to an individual task, *e.g.*, the user selects an object using hand gesture and trigger an action using speech command. The modality flexibility has become a desired property of interaction systems, which meets various user preferences [24]. However, there are few works using three or more interaction modalities, and it is unclear yet which types of combination could achieve better performance. Exploring the diversity of modalities seeks not only to break the restriction of limited modalities, but also provide an intuitive and natural interaction to improve usability [25], [26]. Therefore, it is necessary to investigate the efficiency of different modality combinations.

Current AR systems viable for multimodal interaction can be categorized into two types: the head-mounted display (HMD) and the desktop setup [27], [28]. The HMDs, such as Magic Leap [29] and Microsoft HoloLens [30], provide

Manuscript received **; revised **; accepted **. This work was supported in part by China Postdoctoral Science Foundation under project #2020TQ0156, #2021M691684. (†Corresponding author: Feng Lu)

* indicates equal contribution.

Zhimin Wang, Huangyue Yu, and Feng Lu are with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing 100191, China (email: {zm.wang, yuhuangyue, lufeng}@buaa.edu.cn).

Feng Lu is also with Peng Cheng Laboratory, Shenzhen 518055, China, and Beijing Advanced Innovation Center for Big Data-Based Precision Medicine Beihang University, Beijing 100191, China.

Haofei Wang is with Peng Cheng Laboratory, Shenzhen 518055, China. (email: wanghf@pcl.ac.cn)

the user an immersive interaction experience. However, they usually rely on complex computer vision-based registration techniques to locate the real-world object [31], and the failure of localization may lead to degradation of interaction researches. In addition, the fixed hardware configuration of the above-mentioned devices cannot be easily adapted to the various requirements in different studies, such as flexible selection of different sensors and field of view (FoV). For the desktop setup, researchers have found that there is no significant difference in terms of usability as compared with HMD, and the desktop setup makes it easier to study the effectiveness of different interaction techniques [32], [33]. However, since the desktop setup lacks head tracking, the user feels less immersed during the interaction [33].

In this paper, we propose a Gaze-Gesture-Speech AR system (GGS-AR) to address the above-mentioned challenges, the system is shown in Fig. 1. A portion of this work has been introduced as an extended abstract [34], which only demonstrates the concept of the proposed system with an exemplary experiment. The GGS-AR system consists of four parts: an eye tracking glasses with a scene camera, a hand gesture recognizer, a microphone and an AR display. The system supports accurate gaze tracking, hand gesture recognition and speech recognition. It can be flexibly configured into single-modal, double-modal and triple-modal to investigate the effects of different interaction modalities. The scene camera detects the markers and tracks the user's head, thus the user's sense of immersion is increased. The primary contributions of this paper are: 1) We develop the GGS-AR system which enables the users to interact with the AR objects using gaze, gesture and speech. 2) We use the GGS-AR system to investigate the efficiency of different interaction modalities by quantitative performance measurements as well as subjective feedback. 3) The experimental results demonstrate that the *Gaze+Gesture+Speech* modality is more efficient in terms of completion time and accuracy, and the user preferred *Gesture+Speech* modality than other modalities.

II. Related Work

In this section, we review the existing works on single-modal interaction in AR, multimodal interaction in AR and the AR devices.

A. Single-modal Interaction in AR

An extensive body of research in AR explores different interaction modalities for interacting with the virtual objects. The common modalities are hand gesture, speech and gaze.

Hand gesture: Hand gesture is one of the most intuitive interaction techniques [35], [36]. Researches have been focused on creating a user-defined gesture set for the selected tasks [37], exploring bimanual and unimanual gesture for rotation and scale operations [3], and developing glove-based sensors for translation of the sign language [38].

Speech: Speech-based interaction enables the user to control the device through verbal command [6], [7]. Previous studies have found that voice command achieved comparable efficiency as gesture input [5], and multimodal voice commands are more robust to distance than embodied free-hand gestures or handheld remotes [39].

Gaze: Eye gaze is faster than manual input and requires less physical demand [11]. A few efforts have been made to take the advantage of gaze interaction in AR systems. Recent works have explored the gaze as input to select objects or buttons [15], [40], [41], [42].

To summarize, single-modal interactions have been explored in different applications while they have their own limitations [43]. Specifically, users tend to feel fatigue during gesture-based interaction [3], [9], [44], and the system performance degrades when there exists occlusion [8]. For speech input, it is difficult for the user to remember a large number of voice commands [5], especially for complex interaction scenarios. For gaze-based interaction, it usually suffers from insufficient eye tracking accuracy [10] and the Midas Touch problem [16]. Therefore, it is crucial to find an optimal strategy that combines different interaction modalities to benefit from their complementary natures.

B. Multimodal Interaction in AR

To tackle the limitations of single-modal interaction systems and enrich user experience in AR, multimodal interaction has become a recent research focus.

Gesture+Speech: The *Gesture+Speech* technique is widely used in the AR systems [18], [2], [5], [6]. For instance, using gesture to pinpoint a target and speech to take an action [2]. Lee *et al.* [5] found that integration of hand gesture and speech provides more efficient and accurate control than gesture input alone.

Gaze+Gesture: The *Gaze+Gesture* modality offers more accurate interaction experience than using gaze-only input in head-mounted AR system [10], it also outperforms the gaze-only or gesture-only in the desktop setup [19]. Pfeuffer *et al.* [42] used eye gaze to identify the objects and gesture for object manipulation.

Gaze+Speech: The *Gaze+Speech* modality is only used to perform simple operations on computer interfaces. Prior research [20] incorporated gaze with voice commands to manipulate objects in pictures of a back projection canvas. Kaur *et al.* [21] explored the gaze and speech fusion for moving objects on a computer screen.

In summary, most existing works only combine two modalities at a time. It is unclear yet that which combination is better. Here, we seek to fuse three modalities (*Gaze+Gesture+Speech*) simultaneously in a single AR system, and evaluate the effectiveness of different configurations: two single-modal techniques (*Gesture* and *Gaze*), two double-modal techniques (*Gesture+Speech* and *Gaze+Speech*) and one triple-modal technique (*Gaze+Gesture+Speech*).

C. AR Devices

We summarize the existing AR devices accessible to multimodal interaction in Table I. We compare the number of modalities, weight, and the FoV of each device.

The HMDs provide natural, head-tracked interface that engages users in an immersive environment. These systems

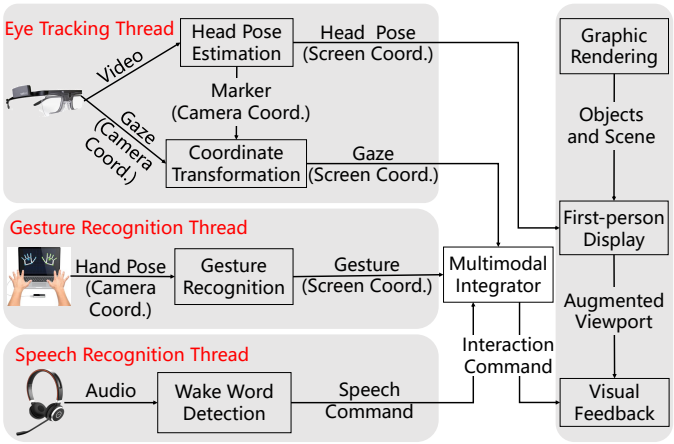

Fig. 2. The system architecture of GGS-AR. The three blocks on the left represent the eye tracking thread, gesture recognition thread, and speech detection thread. The right block represents the main workflow.

usually require complex registration techniques to locate real-world objects [31], and the registration accuracy degrades in less featured environment [45], [46]. Besides, some HMDs have the wearing comfortableness problem due to the factors such as total weight, see Table I. The hardware configuration is usually fixed, it cannot be easily adapted to the different needs in various studies, such as flexible selection of different sensors and FoVs.

The desktop setups have also been investigated in interaction research. It has been found that there is no significant difference between the desktops and the HMDs in terms of usability [32], [33]. However, the interaction study with desktop setup is highly effective, which could guide the HMD design for intuitive experiences [27].

TABLE I
EXISTING AR DEVICES.

hardware	No. of modalities	weight (g)	FoV
Magic Leap 1	3	316	40° × 30°
Microsoft HoloLens 1	2	579	30° × 17.5°
Microsoft HoloLens 2	3	566	43° × 29°
HTC Vive + ZED Mini [47]	3	875	100°×110°
AR-Rift [48]	2	790	80° × 90°

In this work, we propose a lightweight AR system that integrates a head-mounted eye tracker with scene camera, a gesture recognizer and a microphone. The system can be customized to different configurations by replacing different sensors. The GGS-AR system does not require AR registration algorithms, which impels us to focus on the study of multi-modal interaction. Since we track the user's head orientation and adjust the content on the screen accordingly, the user's sense of immersion is increased.

III. System Description

In this section, we describe the details of the GGS-AR system. The system architecture is shown in Fig. 2. The system has three threads: eye tracking thread, gesture recognition thread and speech recognition thread. Each thread outputs individual command, and the multimodal integrator combines these commands and outputs a final command. A demo video can be found at (https://youtu.be/TDFcD7CDO70).

A. Hardware Design

The hardware design features are: 1) the system should support accurate head pose estimation, gaze tracking, hand gesture recognition and speech recognition, the sensors can be easily replaced; 2) the system should be light-weight and the user feels comfortable to wear; 3) the system provides the real-time visual feedback to the user. Based on these requirements, we select the hardware as follows:

1) Hand gesture recognition: Leap Motion Controller [49].
2) Head-mounted eye tracker: Tobii Pro Glasses 2 [50].
3) Voice-input system: Plantronics Headset System for Desk Phones.
4) Head pose estimation: ArUco Markers [51].
5) Server: Intel Core i5-8500 with 3.00Ghz CPU, NVIDA GeForce RTX 2080 SUPER, and 27-inch Full HD Widescreen monitor with resolution of 1920 × 1080. Note that the screen can be replaced with a larger size, or steroscopic 3D displays.

The hardware setup of GGS-AR sytem is shown in Fig. 1. The user sits at about 50 cm in front of a screen, he wears the Tobii Pro Glasses 2, the Leap Motion Controller and the Plantronics Microphone. The eye tracker's manufacturer reports an accuracy of $0.6°$ and precision of $0.05°$ [52]. The weight of eye tracker is only 45 grams, which is comfortable for the user to wear. The Leap Motion controller is mounted on the user's head using a strap. We place the microphone near the user's mouth to reduce the environmental noise. We attach ten markers at the boundary of the screen, as shown in right part of Fig. 1. This enables us to estimate user's head pose.

The total weight of our head-mounted devices is approximately 380 grams. The FoV of GGS-AR system can be adjusted by modifying the FoV of the imaginary screen (see Fig. 3), the FoV is $50 \times 35°$ in the current settings. Each sensor of GGS-AR system can be replaced, *e.g.*, the Tobii Pro Glasses 2 can be substituted with Pupil Labs eye tracker [53].

B. Software Design

The software design features of GGS-AR are: 1) the data threads from different sensors should be synchronized; 2) the algorithm is capable of aligning virtual to real world easily without complex AR registration algorithms; 3) the system should be able to estimate the user's head pose, and provide user an immersive interaction environment. We will describe details of the software implementation as below.

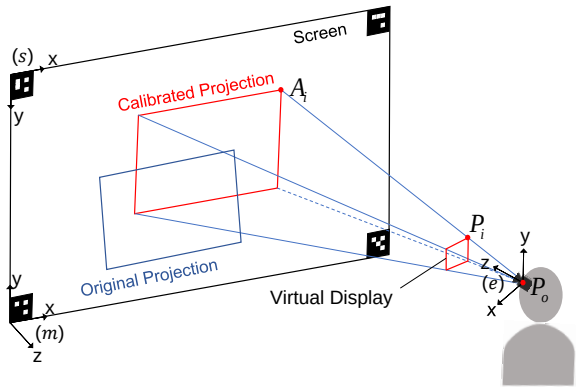

Fig. 3. This figure describes the procedure of first-person display: 1) initialize the FoV of virtual display. 2) transform the P_i or P_o from eye tracker to marker coordinates. 3) project the P_i of virtual display onto the computer screen to product a projection. 4) Calibrate the projection.

1) Overview: The key software components include Leap Motion SDK for gesture recognition, Tobii Glasses Controller for gaze tracking, Porcupine [54] for wake word detection, OpenGL for displaying computer graphics, GLSL for shader, CUDA-OpenCV [55] for accelerating image processing, and Boost [56] for multithreading.

2) Multimodal Integrator: The multimodal integrator manages the multiple threads. In the eye tracking thread, camera-based gaze coordinates are transformed into screen-based gaze coordinates using the perspective matrices obtained in marker detection. In the gesture recognition thread, screen-based gestures are extracted from original position and orientation of joints and fingertips using the following equation:

$$P_{\text{screen}} = \mathbf{M}_{\text{screen}} \cdot \mathbf{M}_{\text{projection}} \cdot \mathbf{M}_{\text{view}} \cdot \mathbf{M}_{\text{model}} \cdot P_{\text{local}}, \quad (1)$$

where P_{local} is the local coordinate of joints and fingertips relative to its local origin, and P_{screen} is the screen coordinate of these points. The rotation matrix $\mathbf{M}_{\text{model}}$ scales, rotates and translates object into scene. The rotation matrix $\mathbf{M}_{\text{view}}$ orients scene in front of camera's eye. The projection matrix $\mathbf{M}_{\text{projection}}$ applies perspective and sizes the frustum. $\mathbf{M}_{\text{screen}}$ transforms the coordinates from -1.0 and 1.0 to the pixel coordinates. [57], [58]. In the speech recognition thread, the system is triggered once the wake word is detected. All these three interactions are combined in multimodal integrator.

3) First-person Display: Fig. 3 shows the procedure of first-person display. We first imagine that there is a virtual display in front of the scene camera of eye tracker, which simulates the head-mounted display. We adjust the FoV of virtual display according to different needs. Then we project the virtual display onto the computer screen. Through the distance-size ambiguity [59], the projection can form a display in human eyes whose FoV is equal to the virtual display, thus we call it 'first-person display'. There are three steps to compute the projection: 1) coordinate transformation, 2) display projection and 3) offset calibration. We define three coordinate systems: eye tracker coordinates *e*, marker coordinates *m*, and screen coordinates *s*. The units of *e* and *m* are millimeter while the unit of *s* is pixel.

Coordinate Transformation: Let $P_i (i = 1, ..., 4)$ be one of the four control points of virtual display. P_o represents the position of scene camera. Both P_i and P_o are 3×1 vectors. We transform these vectors from eye tracker coordinates to marker coordinates using the following equation:

$$P_i^m = \mathbf{R}_m^e \cdot P_i^e + \mathbf{t}_m^e, \quad (2)$$

where $\mathbf{R}_m^e$ and $\mathbf{t}_m^e$ are the 3×3 rotation matrix and 3×1 translation matrix from eye tracker coordinates to marker coordinates. We calculate transformation matrix based on the following steps. We first detect the ArUco markers [51] using the marker detection algorithm based on traditional image processing in [60], [61]. The pixel size of these markers is scaled as small as possible for the purpose of diminishing invasion. We estimate the user's head pose following three steps: 1) we calculate the markers' physical coordinates in marker coordinate system by manual calibration of the marker's physical location before the experiment; 2) we detect the markers in the images captured from the scene camera on the eye tracking glasses; 3) we compute the rotation matrix $\mathbf{R}_m^e$ and translation matrix $\mathbf{t}_m^e$ of head relative to the marker coordinate system, by applying the Efficient Perspective-n-point (EPnP) algorithm [62].

Display Projection: Then we project P_i^m onto the computer screen. Let $A_i (i = 1, ..., 4)$ be one of the four control points of the projection, P_o^m, P_i^m and A_i^m are colinear, which can be written as:

$$\overrightarrow{(P_o A_i)}^m = \lambda_i \overrightarrow{(P_o P_i)}^m, \quad (3)$$

where $\lambda_i \in \mathbb{R}$. According to *Eq.* (2), P_o^m and P_i^m can be solved. The z value of A_i^m is equal to zero. Therefore, the x and y coordinates of A_i^m can be solved. We then transform $A_{i,(x,y)}$ from marker coordinates to screen coordinates, *e.g.*,

$$A_{i,(x,y)}^s = \text{PPI} * A_{i,(x,y)}^m, \quad (4)$$

where PPI is pixels per inch.

Offset Calibration: When user's head is perpendicular to the center of computer screen, the projection should be ideally displayed on the screen's center. The phenomenon that different people keeps distinct manner such as different sitting position and difference in wearing glasses, could invoke the deviation of the display position of projection. Therefore, we compute the distance between the projection of initial state and the ideal projection, which a compensation for subsequent states. As shown in Fig. 3, the blue quadrilateral is calibrated to the red quadrilateral.

4) AR Graphics Rendering: The GGS-AR system uses the GPU shaders to provide the realistic environment. The rendering rate is 30 fps. We exploited OpenGL and GLSL for system development, and implemented scene drawing, model loading and multimodal interaction.

IV. Experimental Setup

We conducted a 5 (modalities) × 2 (tasks) user study to compare the usability of different modalities for AR interaction. The experiment has a repeated measure within-participants design, with interaction modality as the indepen-

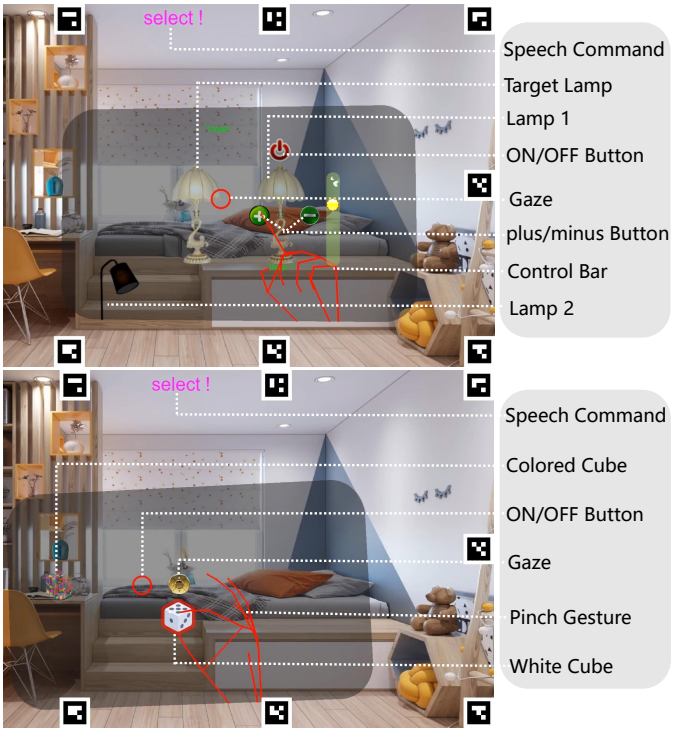

Fig. 4. Top: lamp brightness adjustment task. Down: cube manipulation task. We add additional annotations on the right of original screenshot for better understanding of the task.

dent variables. The dependent variables include objective measurements such as speed and accuracy, as well as subjective measurements such as task load and user preference.

A. Hypothesis

In this experiment, we concern about the efficiency and usability of the five modalities. We propose two hypotheses:
H1: Eye gaze-based interaction is more efficient than other modalities.

Gaze-based interaction can be faster and requires less physical demand than free-hand gesture input [42], [63]. It is also more direct than verbal command. Although gaze-based system has the Midas-touch problem, we aim to combine eye gaze with other modalities to mitigate this drawbacks and enhance the communication speed and accuracy.
H2: The Gaze+Gesture+Speech interaction modality merges the advantages of three modalities and improves user experience.

It has been reported that gaze-based interaction is more suitable to point and select objects [10], [42], descriptive voice input supports better system controllability [5], [6], [2], and free-hand gestures hold the capabilities of transforming and editing [37], [2], [3], [4]. To take advantages of different modalities, we integrate them in a single system. However, users may feel unacquainted with this new modality, especially interacting with the objects for the first time.

B. Interaction Tasks

We used two interaction tasks in our experiments: the lamp brightness adjustment task and the cube manipulation task, as shown in Fig. 4. The lamp task [64] and cube task [5], [65] are the commonly used AR interaction tasks, which represent a series of control tasks in desktop applications. These two tasks allow us to compare five modalities in different backgrounds, magnitudes and frameworks.

Task 1: The lamp brightness adjustment task requires users to adjust the brightness of a lamp to match the brightness of the target lamp. As shown in the top part of Fig. 4, there are two lamps in the work space, lamp 1 is placed on a wooden box and lamp 2 is on the floor. When a certain lamp is selected, the target lamp appears on the left of the lamp, indicating that the lamp is selected. Then the user starts to brighten or darken the lamp using different modalities: sliding up/down along the control bar, pressing "plus" or "minus" button, speaking "Brighten" or "Darken", or gazing at the button. Once the user finishes adjusting, they deselect the lamp.

Task 2: The cube manipulation task requires users to move a white cube to the position of a colored cube. As shown in the down part of Fig. 4, the white cube is placed on a wooden box and the colored cube is placed on a table. The white cube is highlighted when it is selected. Users move the cube to the target location using different modalities: pinching the cube, gazing at cubes, or using index finger to drag the cube. They repeat the manipulations until they feel the two cubes are overlapped.

To summarize, each modality involves the following two operations:

1) Brighten or darken the lamp iteratively to match the brightness of adjusted lamp with target lamp (BD: Brighten or Darken)
2) Move the white cube to the position of colored cube repeatedly until the user feels two cubes overlapped (MC: Move Cube)

C. Interaction Modalities

We define three primary elements of interacting in AR: primary pointing, confirmation and manipulation, which is similar to [10]. Whether the task is complex or simple, users have to perform these primary elements to interact in AR: 1) Primary pointing: the user searches the target using a certain modality. 2) Confirmation: the choice is confirmed and the target is selected. 3) Manipulation: the user manipulates the target, such as moving and scaling.

There are in total seven kinds of interaction modalities: three single-modal techniques, three double-modal techniques and one triple-modal technique. We chose five representative modalities out of seven modalities. We excluded the *Gaze+Gesture* and *Speech Only* techniques. For the *Gaze+Gesture* techniques, both gaze and gesture are suitable to point and select the objects [66]. Gaze-based pointing is a good alternative to hand-based pointing [67]. Therefore, we believe that gaze and gesture serve as similar roles in selecting the objects, so we excluded the *Gaze+Gesture* technique. For the *Speech Only* technique, the efficiency of speech may be limited for spatial tasks such as object translation [68]. Whitlock *et al.* [39] found that voice interaction was least efficient and least preferred. Therefore, we also excluded the

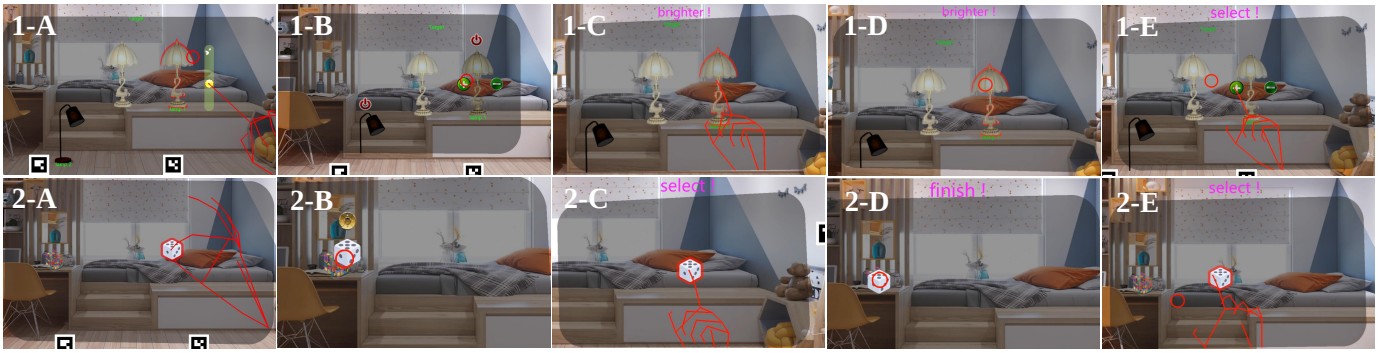

Fig. 5. **(1-A)** *Gesture Only*: sliding up/down along the control bar. **(1-B)** *Gaze Only*: hovering over the "plus" button. **(1-C)** *Gesture+Speech*: pointing the lamp and speaking "Brighter". **(1-D)** *Gaze+Speech*: hovering over the lamp and speaking "Brighter". **(1-E)** *Gaze+Gesture+Speech*: using gaze and speech to select the lamp and pressing the "plus" button. **(2-A)** *Gesture Only*: keeping pinch gesture to move the cube. **(2-B)** *Gaze Only*: moving the white cube by fixating at the colored cube. **(2-C)** *Gesture+Speech*: using speech to select the cube and index finger to drag the cube. **(2-D)** *Gaze+Speech*: looking at the colored cube and speaking "Finish" to put the cube there. **(2-E)** *Gaze+Gesture+Speech*: using gaze and speech to select the cube and index finger to drag the cube.

Speech Only technique. The reason why we assign different actions to different modalities are described below.

For single-modal interaction, we used a one-second dwell-time for *Gaze Only* interaction, and pinch for *Gesture Only* interaction. There is a trade-off between speed and accuracy: a short dwell-time may be too sensitive to select a target; a long dwell time drops the speed advantage and the interaction loses its naturalness [20]. Ware *et al.* [69] found the average gaze selection time was 950ms. Here, we empirically chose one second for the dwell time of selecting a target.

Gesture Only: The user pinches index finger and thumb for one second to select the lamp/cube. S/he uses index finger to slide up/down along the control bar to BD, see Fig. 5(1-A). S/he deselects the lamp through the same hand pinch gesture. After the cube is selected, the user keeps the pinch gesture and moves the cube until it reaches the target, see Fig. 5(2-A).

Gaze Only: The user gazes at the ON/OFF button for one second to select the lamp/cube. S/he fixates at "plus" or "minus" button to BD, see Fig. 5(1-B), and gazes at the colored cube for one second to MC, see Fig. 5(2-B).

For multimodal interaction, we mainly consider the following modality features: 1) Eye gaze is more suitable to point and select objects. Kytö *et al.* [10] used eye gaze to point the object and used secondary modalities to refine the selection. 2) Gestures hold the capability of selection and transforms [37]. Gestures are usually used to select objects [2], and also used to rotate and scale targets [3]. 3) Speech commands offer better system control and are like triggers [39]. For example, Piumsomboon *et al.* [2] used speech command such as "move it" to perform an action. Therefore, in the *Gesture+Speech* modality, we use gesture to select the lamps or the cubes, and use speech to adjust the brightness of lamp or manipulate the cubes. In the *Gaze+Speech* modality, gaze is used to select the targets, speech offers the system control. In the *Gaze+Gesture+Speech* modality, gaze is used to point primarily, speech serves as a trigger, and gesture is used to transform the target.

Gesture+Speech: The user uses his/her index finger to point the lamp/cube and uses verbal command "Select" to confirm the target. Then the user uses verbal command "Brighter" or "Darker" to BD and "Finish" to deselect the target, see Fig. 5(1-C). After the cube is selected, the user uses his/her index finger dragging white cube to MC, see Fig. 5(2-C), and then uses verbal command "Finish" to complete the task.

Gaze+Speech: The user fixates at the lamp/cube and uses verbal command "Select" to confirm the objects. Then s/he uses "Brighter", "Darker" and "Finish" to complete the target operation, see Fig. 5(1-D). For the selected cube, the user gazes at the target position and uses verbal command "Finish" to place the cube at the gazed position, see Fig. 5(2-D).

Gaze+Gesture+Speech: The user fixates at the lamp/cube and uses verbal command "Select" to confirm the targets. Then s/he presses the "plus" or "minus" button to BD using the index finger, see Fig. 5(1-E), and uses verbal command "Finish" to terminate. After the cube is selected, the user uses the index finger to drag the cube to MC, see Fig. 5(2-E), and uses verbal command "Finish" to stop.

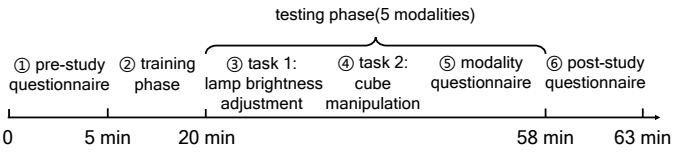

Fig. 6. The experimental procedure.

D. Experimental Procedure

The flow diagram of our experiment is shown in Fig. 6. The participants were first asked to fill in a pre-study questionnaire. Then they proceeded to a training phase where they were given instructions and practiced using the different interaction techniques. After training, they conducted the experiments including two tasks using five modalities and five questionnaires. The order of the five interaction modalities was randomized. Finally, the participants filled a post-study questionnaire to assess the system and chose which modality they preferred most. Prior to each section associated with eye gaze modality, the participants conducted a user calibration for the eye tracker. The participants were asked to balance accuracy and speed

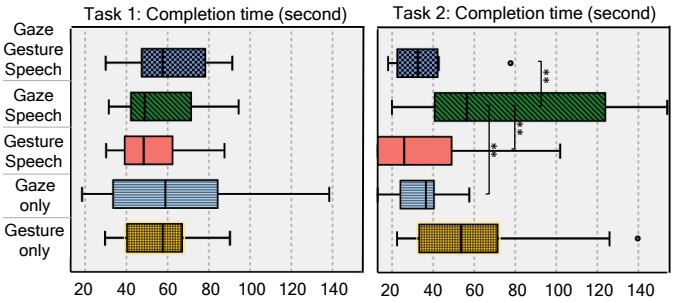

Fig. 7. Boxplots of task completion time of five modalities in two tasks. The statistical significances are labeled with ** (p $<$ 0.05). Error bars represent standard deviations. The little colored circles represent the outliers. There is no statistically significant difference for trial completion time on Task 1.

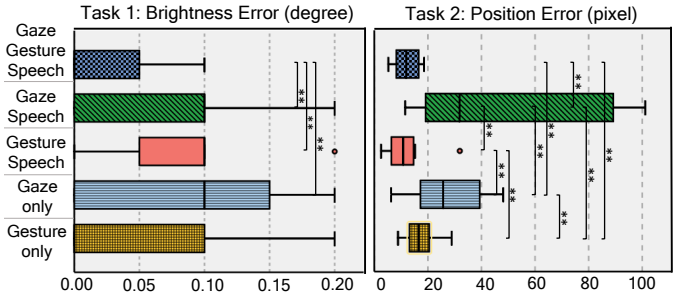

Fig. 8. Boxplots of accuracy of five modalities in two tasks. The statistical significances are labeled with ** (p $<$ 0.05). Error bars represent standard deviations. The little colored circles represent the outliers.

during the experiments. Each experiment took around 63 minutes.

E. Performance Evaluation

We evaluate the system performance by completion time and accuracy. We define the accuracy as the brightness difference between the target lamps and the adjusted lamps in Task 1, and the distance difference between the colored cube and the white cube in Task 2.

The subjective metrics report the usability and effectiveness of five modalities. The modality questionnaire included a NASA's Task Load Index [70] with 7-point Likert scales and six free-response questions to collect response on naturalness and frustration of each modality. The post-study questionnaire included a System Usability Scale (SUS) to assess the overall usability of our AR system, and a preference question, *i.e.*, "Overall, which modality do you prefer most?"

V. Experimental Results

We recruited 12 subjects on campus (8 male, 4 female), the average age is 23.8 (SD = 1.6). All participants have normal or correct-to-normal vision, and they are able to see the hint on computer screen clearly. According to results of the pre-study questionnaire with 5-point Likert scales, the participants reported low prior familiarity with AR (Mean = 2.8), the eye tracker (Mean = 2.6), and hand gesture recognition system (Mean = 2.7); medium familiarity with voice-based inputs such as Siri (Mean = 3.6). All the participants can read and speak English fluently.

A. Objective Evaluation Results

1) Completion Time: We used a repeated-measures ANOVA ($\alpha = 0.05$), in conjunction with post hoc pairwise t-tests to identify whether the task completion time is significantly different across modalities. The results are shown in Fig. 7. The statistical analysis showed that the effect of modalities on completion time for Task 2 (cube manipulation) was statistically significant (F(4, 44) = 3.89, p = 0.041, $\eta^2 = 0.26$), while it failed to reject the equality of the levels of modalities on completion time for Task 1 (lamp brightness adjustment, p = 0.404). In Task 2, we found that *Gaze Only*, *Gesture+Speech* and *Gaze+Gesture+Speech* were significantly faster than *Gaze+Speech* (p = 0.017, 0.045, 0.027, see Table II). Both *Gaze Only* and *Gaze+Gesture+Speech* slightly outperformed *Gesture Only* in terms of task time (p = 0.067, 0.091).

TABLE II
Completion Time Comparison on Task 2.

Modalities Compared	df [71]	t-statistic	p-value
GEST vs GAZE	11	2.031	0.067
GEST vs GEST-SPCH	11	1.415	0.185
GEST vs GAZE-SPCH	11	-1.015	0.322
GEST vs GAZE-GEST-SPCH	11	1.854	0.091
GAZE vs GEST-SPCH	11	-0.462	0.653
GAZE vs GAZE-SPCH *	11	-2.822	0.017
GAZE vs GAZE-GEST-SPCH	11	-0.360	0.726
GEST-SPCH vs GAZE-SPCH *	11	-2.265	0.045
GEST-SPCH vs GAZE-GEST-SPCH	11	0.121	0.906
GAZE-SPCH vs GAZE-GEST-SPCH *	11	2.544	0.027

Notes: GAZE, GEST and SPCH stand for gaze, gesture and speech.
* indicates the p-value $<$ 0.05.

There is no significant difference between *Gesture+Speech* and *Geture Only* in terms of completion time in Task 1. However, Lee *et al.* found the difference between these two modalities, that the gesture takes longer time than *Gesture+Speech* [5]. They argue that this is mainly because using the speech input in *Gesture+Speech* for changing color or shape of the objects spend less time than using gesture input. However, in our lamp brightness adjustment task, we found that there is no difference in terms of manipulation time between the speech input and gesture input for changing the brightness of a lamp.

2) Accuracy: We performed a repeated-measures ANOVA ($\alpha = 0.05$), in conjunction with post hoc pairwise t-tests to identify whether the task accuracies are significantly different across modalities. The results are shown in Fig. 8. The statistical analysis indicated that the effect of modalities on accuracy was statistically significant (brightness error, F(4, 44) = 2.949, p = 0.052, $\eta^2 = 0.21$ and position error, F(4, 44) = 10.388, p $<$ 0.001, $\eta^2 = 0.49$ respectively).

Overall, *Gaze+Gesture+Speech* appeared to be the most accurate modality according to Fig. 8. Specifically, in Task 1, we found that *Gaze+Gesture+Speech* outperformed *Gaze Only*, *Gesture+Speech* and *Gaze+Speech* in terms of accuracy (p = 0.027, 0.027, 0.026, see Table III). In Task 2, we found that *Gaze+Gesture+Speech* outperformed *Gesture Only*, *Gaze*

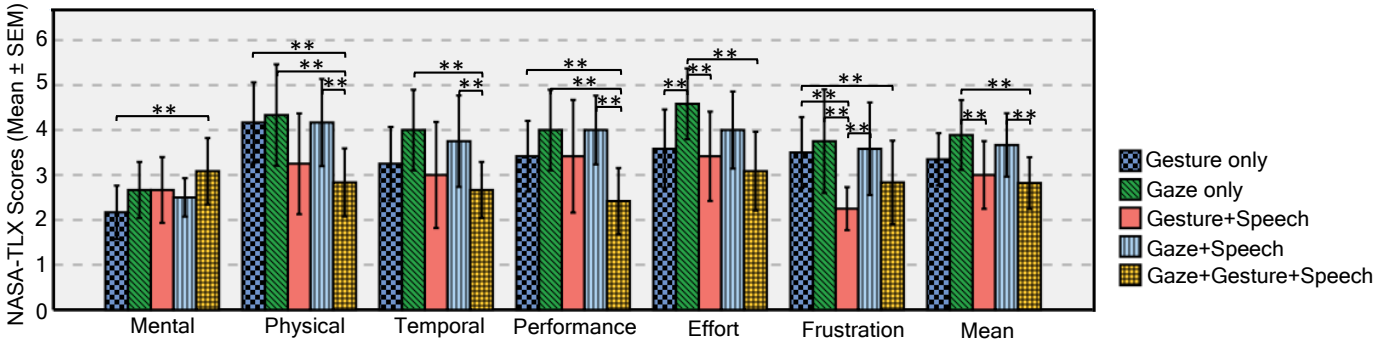

Fig. 9. Bar charts of scores on the NASA-TLX questionnaire for comparing five modalities. The statistical significances ($p < 0.05$) are labeled with **. Error bars represent standard deviations.

TABLE III
ACCURACY COMPARISON ON TASK 1.

Modalities Compared	df	t-statistic	p-value
GEST vs GAZE	11	-1.449	0.175
GEST vs GEST-SPCH	11	-1.603	0.137
GEST vs GAZE-SPCH	11	-1.173	0.266
GEST vs GAZE-GEST-SPCH	11	1.000	0.339
GAZE vs GEST-SPCH	11	0.000	1.000
GAZE vs GAZE-SPCH	11	0.266	0.795
GAZE vs GAZE-GEST-SPCH *	11	2.548	0.027
GEST-SPCH vs GAZE-SPCH	11	0.364	0.723
GEST-SPCH vs GAZE-GEST-SPCH *	11	2.548	0.027
GAZE-SPCH vs GAZE-GEST-SPCH *	11	2.569	0.026

Notes: GAZE, GEST and SPCH stand for gaze, gesture and speech.
* indicates the p-value < 0.05.

Only and *Gaze+Speech* in terms of accuracy (p = 0.016, 0.002, 0.005, see Table IV). We also found that *Gesture+Speech* outperformed *Gesture Only*, *Gaze Only* and *Gaze+Speech* (p = 0.046, 0.006, 0.004). There is a significant difference between *Gesture Only* and *Gaze Only* in move accuracy (p = 0.031). All the other modalities including *Gesture Only*, *Gaze Only*, *Gesture+Speech* and *Gaze+Gesture+Speech* outperformed *Gaze+Speech* in Task 2 (p = 0.011, 0.033, 0.004, 0.005).

TABLE IV
ACCURACY COMPARISON ON TASK 2.

Modalities Compared	df	t-statistic	p-value
GEST vs GAZE *	11	-2.467	0.031
GEST vs GEST-SPCH *	11	2.244	0.046
GEST vs GAZE-SPCH *	11	-3.035	0.011
GEST vs GAZE-GEST-SPCH *	11	2.835	0.016
GAZE vs GEST-SPCH *	11	3.407	0.006
GAZE vs GAZE-SPCH *	11	-2.443	0.033
GAZE vs GAZE-GEST-SPCH *	11	4.005	0.002
GEST-SPCH vs GAZE-SPCH *	11	-3.629	0.004
GEST-SPCH vs GAZE-GEST-SPCH	11	-0.309	0.763
GAZE-SPCH vs GAZE-GEST-SPCH *	11	3.511	0.005

Notes: GAZE, GEST and SPCH stand for gaze, gesture and speech.
* indicates the p-value < 0.05.

B. Subjective Evaluation Results

1) Task Load: Repeated-measures ANOVA analysis from the NASA TLX questionnaire indicated that different modalities resulted in different task loads. The post hoc analysis from pairwise comparisons between the modalities were shown in Fig. 9. In general, the *Gaze+Gesture+Speech* achieved lowest task load, significantly lower than *Gaze Only* and *Gaze+Speech*. However, the *Gaze+Gesture+Speech* has the highest *Mental demand*. This is reasonable since the user has to switch between the modalities. We also observed that *Gaze+Gesture+Speech* achieved the lowest *Physical/Temporal demand*, *Effort* and best *Performance* than other modalities. The *Gesture+Speech* has the lowest *Frustration*, this might because that the gesture and speech is most intuitive to use, thus caused less frustration.

2) User Preference: According to the results of user preference question, the *Gesture+Speech* modality is the most preferred by the users, which gained 58.3% votes, 33.3% users preferred *Gaze+Gesture+Speech* modality, only one participant preferred *Gaze Only* modality.

3) System Usability Scale: The average score of System Usability Scale (SUS) is 71.9 (SD = 8.7), ranging from 60 to 87.5. According to the surveys comparing SUS scores for different systems, GGS-AR achieved the level of "Good" [72].

VI. DISCUSSION

We found that there was no significant difference in Task 1 in terms of trial completion time. This might be due to the lamps are relatively large, which are easy to manipulate and require less time, while the cubes in Task 2 are small and it can clearly reflect the efficiency and usability difference between the interaction modalities. We also noticed that there was a gaze position drifting phenomenon when the user spoke too loud, which caused difficulty in selecting and manipulating the small object. It might because the *Gaze+Speech* modality takes the longest time and holds the lowest accuracy among all modalities, thus the wearing position of glasses might change after long time usage. In the experiment of *Gesture Only*, we found that the hand gesture may occlude participants' vision due to hand pinch gesture for translating the cube, which may explain the high *Effort* and *Frustration* of *Gesture Only*.

A. User Feedback

In the free questions, participants claimed that *Gaze+Gesture+Speech* is novel and efficient. "Three interactions are combined to reduce the task difficulty and I can complete it faster." (P3). P5 and P7 expressed the similar opinions. Some participants found the *Gaze Only* is fast. "The *Gaze Only* techniques was very quick and smooth." (P2). But some participants found it weak with fatigue after using for a long time. "It was also convenient to use, but I felt my eyes dry." (P6). Both our objective results and user feedback showed that the *Gaze+Speech* modality is challenging and inaccurate. "When I speak the verbal command, gaze point may drift resulting in low accuracy and the difficulty to select the target." (P12). P2 and P5 also found this phenomenon. Most of the participants found that *Gesture+Speech* enjoyable and intuitive, *e.g.*, "it facilitated the operation to a certain extent and achieved the synergistic complement." (P9). But some participants found the verbal command complex. "I need to remember the voice commands correctly." (P2) Some participants found *Gesture Only* is natural but requires high demand. "I need to keep a fixed pinch gesture all the time, which is somewhat fatiguing." (P3). P5, P8 and P9 also felt exhausted.

B. Hypothesis Validation

H1: Eye gaze-based interaction is more efficient than other modalities.

Our prediction was partially supported. In Task 1, we found that *Gaze+Gesture+Speech* achieved higher accuracy than *Gesture+Speech* (p = 0.027), see Table III, while there is no significant difference in terms of trial completion time. We argue that the size of lamps is relatively large, so that it is easy to manipulate and requires less time. In Task 2, both *Gaze Only* and *Gaze+Gesture+Speech* slightly outperformed *Gesture Only* in terms of completion time (p = 0.067, 0.091), see Table II. However, on account of the gaze drift problem, *Gaze+Speech* has lower accuracy than *Gesture Only* and *Gesture+Speech* (p = 0.011, 0.004), see Table IV. We think that small cubes require more accurate interaction modalities. Besides, user's feedback also suggested that "the gaze estimation accuracy needed to be improved" (P4). We conclude that eye gaze-based interaction improves the speed but might not guarantee the accuracy. A similar conclusion was found by the prior work [10] where *Gaze Only* achieves the fastest speed but the least accuracy among the interaction modalities. The reason for the difference could be the well-known calibration and drift problem on wearable eye trackers [73]. In our study of *Gaze Only* and *Gaze+Speech*, we found that the accuracy of gaze estimation tended to degrade over time, due to subtle shifting of eye tracker during head movement.

H2: The Gaze+Gesture+Speech interaction modality merges the advantages of three modalities and improves user experience.

Our results supported this hypothesis. The triple-modal technique tackles the limitations of single-modal interaction, and improves user experience. In Task 1, *Gaze+Gesture+Speech* achieved higher accuracy than *Gaze Only*, *Gesture+Speech* and *Gaze+Speech* (p = 0.027, 0.027, 0.026), see Table III. In Task 2, for the completion time, *Gaze+Gesture+Speech* outperformed *Gaze+Speech* (p = 0.027), see Table II. For accuracy, *Gaze+Gesture+Speech* outperformed *Gesture Only*, *Gaze Only* and *Gaze+Speech* (p = 0.016, 0.002, 0.005), see Table IV. Subjective feedback also reported that "three interactions are well integrated in *Gaze+Gesture+Speech*. Eye gaze provides rapidly primary pointing, verbal command offers better system controllability, and hand gestures are accurate." (P8). Therefore, we conclude that the *Gaze+Gesture+Speech* modality is superior in terms of efficiency.

We further investigated the characteristics of different modalities. Eye gaze is more suitable to point and select objects, which reduces physical demand and fatigue compared to hand gesture, as found in [10]. Gesture holds the capability of selection and transforms, which is more flexible for complex operations [3]. *e.g.*, Piumsomboon *et al.* [37] created 800 gestures for 40 selected tasks. Speech is like trigger [39] and does not need the dwell time to execute the command, as shown in [5]. Therefore, for the primary pointing and confirmation in the *Gaze+Gesture+Speech*, the user fixates at the lamp/cube and uses verbal command "Select" to confirm the targets, which requires the less physically demand and saves the time. For the manipulation in the *Gaze+Gesture+Speech*, the user uses his/her index finger to manipulate the targets, and uses the verbal command to terminate, which also does not need the dwell time and guarantees the accuracy. Above analysis accounts for the superior performance of *Gaze+Gesture+Speech*. However, we noticed that *Gaze+Gesture+Speech* achieved the highest *Mental demand* score in the NASA-TLX questionnaire. We considered that the novice may need some time to master how to use the triple-modal technique when interacting with objects in AR.

C. Implications and Design Recommendations

Based on aforementioned results, we bring more actionable implications for future research as follows:

1) Gaze in AR: Eye gaze-based interaction can improve the speed. However, the drift of gaze calibration cannot be negligible. Frequent recalibration is impractical in the experiment due to time-consuming and disruptive. There are three methods to alleviate the influence of gaze drift: a) Use an online offset compensation algorithm to compute the offset between the estimated fixation position and the actual fixation position [74]. b) Define proper target sizes according to the eye gaze distribution of surrounding the target [10]. c) Integrate secondary modality such as gesture, speech to refine the eye gaze-based object selection.

2) Multimodal Interaction in AR: We provide the design recommendations for multimodal interaction: a) Use eye gaze to conduct swift actions such as selecting targets; b) Use the verbal command to offer better system control and confirm operations, and c) Use the hand gesture to execute skilled actions such as panning and zooming.

D. System Limitations

In the experiment, we found that there was no statistically significant difference for trial completion time on Task 1.

We suspect that the object size had a leading impact on the interaction efficiency. The size of table lamps is $7 \times 4cm^2$, while the size of cubes is only $3 \times 3cm^2$. Therefore, it may be easier to select and manipulate the lamps. Future extensions of the GGS-AR may compare the effect of different initial scales, such as 25%, 50%, 75%, 100% of target size. Besides, the different user distances may affect the interaction efficiency. In order to capture clear image of ArUco markers, the user should sit 40 cm to 70 cm away from the screen.

Another limitation of our work is that we only examined the basic operations such as primary pointing, confirmation, moving and sliding for a small set of tasks. For more comprehensive comparison of different modalities, future explorations of the study may introduce additional interaction operations such as zoom, rotation, and more complex editing by eye gaze [75].

Finally, in the current implementation, we use a 2D computer screen to display the interactive environments. Future work could explore to display 3D environment using a stereo display. We also noticed that a limited FoV constrains how objects can be placed and search through [76]. Future explorations of GGS-AR may study the different size of FoV by setting the FoV parameters or using varisized displays. The current sample size (N = 12) is small, which may be underpowered to find large effects [77]. The lack of participants might impact the study outcome, and we will enlarge the number of participants in the future work.

VII. Conclusion

In this paper, we proposed a novel GGS-AR system to investigate the benefits of multimodal interaction in AR system. Our lightweight system integrates several sensors that support accurate gaze tracking, hand gesture and speech recognition simultaneously. We evaluated and compared various modality combinations using the proposed system. The experimental results demonstrate that the *Gaze+Gesture+Speech* modality is superior to other modalities in terms of the completion time and interaction accuracy. The *Gesture+Speech* modality is more preferred by the users. This study offers insights to design multimodal interactive AR systems in a more flexible manner.

References

[1] A. Leganchuk, S. Zhai, and W. Buxton, "Manual and cognitive benefits of two-handed input: an experimental study," *ACM Trans. on Computer-Human Interact.*, vol. 5, no. 4, pp. 326–359, 1998.

[2] T. Piumsomboon, D. Altimira, H. Kim, A. J. Clark, G. A. Lee, and M. Billinghurst, "Grasp-shell vs gesture-speech: A comparison of direct and indirect natural interaction techniques in augmented reality," in *Proc. IEEE Int. Symp. Mix. Augmented Real.*, Munich, Germany, Nov. 2014, pp. 73–82.

[3] N. Chaconas and T. Höllerer, "An evaluation of bimanual gestures on the microsoft hololens," in *Proc. IEEE Conf. Virtual Real. 3D User Interfaces*, Reutlingen, Germany, Mar. 2018, pp. 33–40.

[4] K. A. Satriadi, B. Ens, M. Cordeil, B. Jenny, T. Czauderna, and W. Willett, "Augmented reality map navigation with freehand gestures," in *Proc. IEEE Conf. Virtual Real. 3D User Interfaces*, Osaka, Japan, Mar. 2019, pp. 593–603.

[5] M. Lee, M. Billinghurst, W. Baek, R. D. Green, and W. Woo, "A usability study of multimodal input in an augmented reality environment," *Virtual Real.*, vol. 17, no. 4, pp. 293–305, 2013.

[6] S. Irawati, S. A. Green, M. Billinghurst, A. Dünser, and H. Ko, ""move the couch where?" : developing an augmented reality multimodal interface," in *Proc. IEEE Int. Symp. Mix. Augmented Real.*, California, USA, Oct. 2006, pp. 183–186.

[7] S. Goose, S. Sudarsky, Xiang Zhang, and N. Navab, "Speech-enabled augmented reality supporting mobile industrial maintenance," *IEEE Pervasive Comput.*, vol. 2, no. 1, pp. 65–70, 2003.

[8] A. O. S. Feiner, "The flexible pointer: An interaction technique for selection in augmented and virtual reality," in *Proc. ACM Symp. User Interface Softw. Technol.*, BC, Canada, Nov. 2003, pp. 81–82.

[9] J. D. Hincapié-Ramos, X. Guo, P. Moghadasian, and P. Irani, "Consumed endurance: a metric to quantify arm fatigue of mid-air interactions," in *Proc. Conf. Human Factors Comput. Syst.*, Ontario, Canada, Apr. 2014, pp. 1063–1072.

[10] M. Kytö, B. Ens, T. Piumsomboon, G. A. Lee, and M. Billinghurst, "Pinpointing: Precise head- and eye-based target selection for augmented reality," in *Proc. Conf. Human Factors Comput. Syst.*, Quebec, Canada, Apr. 2018, pp. 1–14.

[11] P. Wang, X. Bai, M. Billinghurst, S. Zhang, W. He, D. Han, Y. Wang, H. Min, W. Lan, and S. Han, "Using a head pointer or eye gaze: The effect of gaze on spatial AR remote collaboration for physical tasks," *Interact. Comput.*, vol. 32, pp. 153–169, 2020.

[12] Y. Wang, G. Zhai, S. Zhou, S. Chen, X. Min, Z. Gao, and M. Hu, "Eye fatigue assessment using unobtrusive eye tracker," *IEEE Access*, vol. 6, pp. 55 948–55 962, 2018.

[13] Y. Cheng, X. Zhang, F. Lu, and Y. Sato, "Gaze estimation by exploring two-eye asymmetry," *IEEE Trans. Image Process.*, vol. 29, pp. 5259–5272, 2020.

[14] H. Yu, M. Cai, Y. Liu, and F. Lu, "First- and third-person video co-analysis by learning spatial-temporal joint attention," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020, accepted.

[15] H. Park, S. Lee, and J. Choi, "Wearable augmented reality system using gaze interaction," in *Proc. IEEE Int. Symp. Mix. Augmented Real.*, Cambridge, United kingdom, Sep. 2008, pp. 175–176.

[16] P. Mohan, W. B. Goh, C. Fu, and S. Yeung, "DualGaze: Addressing the midas touch problem in gaze mediated vr interaction," in *Proc. IEEE Int. Symp. Mix. Augmented Real.*, Munich, Germany, Oct. 2018, pp. 79–84.

[17] B. Velichkovsky, A. Sprenger, and P. Unema, "Towards gaze-mediated interaction: Collecting solutions of the "midas touch problem"," in *Proc. Human-Comput. Interact.*, 1997, pp. 509–516.

[18] E. C. Kaiser, A. Olwal, D. McGee, H. Benko, A. Corradini, X. Li, P. R. Cohen, and S. Feiner, "Mutual disambiguation of 3d multimodal interaction in augmented and virtual reality," in *Int. Conf. Multimodal Interfaces*, BC, Canada, Nov. 2003, pp. 12–19.

[19] I. Chatterjee, R. Xiao, and C. Harrison, "Gaze+gesture: Expressive, precise and targeted free-space interactions," in *Proc. ACM Int. Conf. Multimodal Interact.*, Washington, USA, Nov. 2015, pp. 131–138.

[20] M. Elepfandt and M. Grund, "Move it there, or not?: the design of voice commands for gaze with speech," in *Proc. Workshop Eye Gaze Intelligent Hum. Mach. Interact., Gaze-In*, California, USA, Oct. 2012, pp. 1–3.

[21] M. Kaur, M. Tremaine, N. Huang, J. Wilder, Z. Gacovski, F. Flippo, and C. S. Mantravadi, "Where is it? event synchronization in gaze-speech input systems," in *Int. Conf. Multimodal Interfaces*, BC, Canada, Nov. 2003, pp. 151–158.

[22] H. Wang, X. Dong, Z. Chen, and B. E. Shi, "Hybrid gaze/eeg brain computer interface for robot arm control on a pick and place task," in *IEEE Int. Conf. IEEE Eng. Med. Biol. Soc.*, Milan, Italy, Aug. 2015, pp. 1476–1479.

[23] X. Dong, H. Wang, Z. Chen, and B. E. Shi, "Hybrid brain computer interface via bayesian integration of eeg and eye gaze," in *IEEE Conf. Neural Eng.* Montpellier, France: IEEE, Apr. 2015, pp. 150–153.

[24] M. Turk, "Multimodal interaction: A review," *Pattern Recogn.*, vol. 36, pp. 189–195, 2014.

[25] J. C. Kim, T. H. Laine, and C. Åhlund, "Multimodal interaction systems based on internet of things and augmented reality: A systematic literature review," *Appl. Sci.*, vol. 11, no. 4, p. 1738, 2021.

[26] S. Oviatt, P. Cohen, L. Wu, J. Vergo, L. Duncan, B. Suhm, J. Bers, T. Holzman, T. Winograd, J. Landay, J. Larson, and D. Ferro, "Designing the user interface for multimodal speech and pen-based gesture applications: State-of-the-art systems and future research directions," *Hum.-Comput. Interact.*, vol. 15, no. 4, p. 263–322, 2000.

[27] B. S. Santos, P. Dias, A. Pimentel, J. Baggerman, C. Ferreira, S. S. Silva, and J. Madeira, "Head-mounted display versus desktop for 3d navigation in virtual reality: a user study," *Multim. Tools Appl.*, vol. 41, no. 1, pp. 161–181, 2009.

[28] B. S. Santos, P. Dias, S. Silva, L. Capucho, N. Salgado, F. Lino, V. Carvalho, and C. Ferreira, "Usability evaluation in virtual reality: A user study comparing three different setups," in *Eurographics Symp. Virtual Environ.*, Eindhoven, Netherlands, May 2008.

[29] "Magic Leap 1," (2020). [Online]. Available: https://www.magicleap.com/en-us/magic-leap-1

[30] "Microsoft HoloLens 2," (2020). [Online]. Available: https://www.microsoft.com/en-us/hololens

[31] W. A. Hoff, K. Nguyen, and T. Lyon, "Computer-vision-based registration techniques for augmented reality," in *Proc. SPIE. Int. Soc. Opt. Eng.*, 1996, pp. 538–548.

[32] S. Sharples, S. Cobb, A. Moody, and J. R. Wilson, "Virtual reality induced symptoms and effects (VRISE): comparison of head mounted display (hmd), desktop and projection display systems," *Disp.*, vol. 29, no. 2, pp. 58–69, 2008.

[33] X. Tong, D. Gromala, D. Gupta, and P. Squire, "Usability comparisons of head-mounted vs. stereoscopic desktop displays in a virtual reality environment with pain patients," in *Stud. Health. Technol. Inform.*, California, USA, Apr. 2016, pp. 424–431.

[34] Z. Wang, H. Yu, H. Wang, Z. Wang, and F. Lu, "Comparing single-modal and multimodal interaction in an augmented reality system," in *Proc. IEEE Int. Symp. Mix. Augmented Real.*, Porto de Galinhas, Brazil, Nov. 2020, pp. 165–166, in Press.

[35] M. Cai, F. Lu, and Y. Gao, "Desktop action recognition from first-person point-of-view," *IEEE Trans. Cybern.*, vol. 49, no. 5, pp. 1616–1628, 2019.

[36] M. Cai, F. Lu, and Y. Sato, "Generalizing hand segmentation in egocentric videos with uncertainty-guided model adaptation," in *IEEE Conf. Comput. Vision Pattern Recognit.*, Washington, USA, Jun. 2020, pp. 14 380–14 389.

[37] T. Piumsomboon, A. J. Clark, M. Billinghurst, and A. Cockburn, "User-defined gestures for augmented reality," in *Proc. Conf. Human Factors Comput. Syst.*, Paris, France, Apr. 2013, pp. 955–960.

[38] K. S. Abhishek, L. C. K. Qubeley, and D. Ho, "Glove-based hand gesture recognition sign language translator using capacitive touch sensor," in *IEEE Int. Conf. Electron Devices and Solid-State Circuits*, 2016, pp. 334–337.

[39] M. Whitlock, E. Harnner, J. R. Brubaker, S. K. Kane, and D. A. Szafir, "Interacting with distant objects in augmented reality," in *Proc. IEEE Conf. Virtual Real. 3D User Interfaces*, Reutlingen, Germany, Mar. 2018, pp. 41–48.

[40] S. Nilsson, "Interaction without gesture or speech – a gaze controlled ar system," in *Proc. Int. Conf. on Artificial Reality and Telexistence*, Jylland, Denmark, Nov. 2007, pp. 280–281.

[41] M. Bâce, T. Leppänen, D. G. de Gomez, and A. R. Gomez, "ubigaze: ubiquitous augmented reality messaging using gaze gestures," in *SIGGRAPH ASIA Mob. Graph. Interact. Appl.*, Macau, China, 2016, pp. 1–5.

[42] K. Pfeuffer, B. Mayer, D. Mardanbegi, and H. Gellersen, "Gaze + pinch interaction in virtual reality," in *Proc. Symp. Spatial User Interact.*, Brighton, United kingdom, Oct. 2017, pp. 99–108.

[43] S. K. Badam, A. Srinivasan, N. Elmqvist, and J. Stasko, "Affordances of input modalities for visual data exploration in immersive environments," in *Proc. Workshop on Immersive Analytics at IEEE VIS*, 2017.

[44] K. Hinckley, R. F. Pausch, J. C. Goble, and N. F. Kassell, "A survey of design issues in spatial input," in *Proc. ACM Symp. User Interface Softw. Technol.*, California, USA, Nov. 1994, pp. 213–222.

[45] E. Ragan, C. Wilkes, D. A. Bowman, and T. Hollerer, "Simulation of augmented reality systems in purely virtual environments," in *Proc. IEEE Virtual Real.*, Louisiana, USA, Mar. 2009, pp. 287–288.

[46] C. Lee, S. Bonebrake, D. A. Bowman, and T. Höllerer, "The role of latency in the validity of ar simulation," in *Proc. IEEE Virtual Real.*, Massachusetts, USA, Mar. 2010, pp. 11–18.

[47] "HTC Vive + ZED Mini," (2020). [Online]. Available: https://www.stereolabs.com/blog/vive-pro-ar-zed-mini/

[48] W. Steotoe, "AR-RIFT," (2013). [Online]. Available: https://willsteptoe.com/post/66968953089/ar-rift

[49] "Leap Motion Controller," (2020). [Online]. Available: https://developer.leapmotion.com/

[50] "Tobii Pro Glasses 2," (2020). [Online]. Available: https://www.tobiipro.com/product-listing/tobii-pro-glasses-2/

[51] "Aruco Markers," (2020). [Online]. Available: http://www.uco.es/investiga/grupos/ava/node/26

[52] K. Holmqvist, M. Nyström, R. Andersson, R. Dewhurst, H. Jarodzka, and J. V. de Weijer, *Eye tracking: a comprehensive guide to methods and measures*. Oxford University Press, 2011.

[53] "Pupil Core," (2020). [Online]. Available: https://pupil-labs.com/products/core/

[54] "Porcupine," (2020). [Online]. Available: https://github.com/Picovoice/porcupine

[55] "OpenCV-CUDA," (2020). [Online]. Available: https://opencv.org/platforms/cuda/

[56] "Boost," (2007). [Online]. Available: https://www.boost.org/

[57] J. de Vries, "LearnOpenGL - Coordinate Systems," (2014). [Online]. Available: https://learnopengl.com/Getting-started/Coordinate-Systems

[58] J. Neider, T. Davis, and M. Woo., *OpenGL Programming Guide*. Reading, MA: Addison-Wesley, 1993.

[59] L. Swirski and N. Dodgson, "A fully-automatic, temporal approach to single camera, glint-free 3d eye model fitting," in *Proc. Pervasive Eye Tracking and Mobile Eye-Based Interact.*, 2013, pp. 1–11.

[60] S. Garrido-Jurado, R. Muñoz-Salinas, F. J. Madrid-Cuevas, and R. Medina Carnicer, "Generation of fiducial marker dictionaries using mixed integer linear programming," *Pattern Recogn.*, vol. 51, pp. 481–491, 2016.

[61] S. Garrido-Jurado, R. Muñoz-Salinas, F. J. Madrid-Cuevas, and M. J. Marín-Jiménez, "Automatic generation and detection of highly reliable fiducial markers under occlusion," *Pattern Recogn.*, vol. 47, pp. 2280–2292, 2014.

[62] V. Lepetit, F. Moreno-Noguer, and P. Fua, "Epnp: An accurate o(n) solution to the pnp problem," *Int. J. Comput. Vision*, vol. 81, no. 2, pp. 155–166, 2009.

[63] S. Zhai, C. Morimoto, and S. Ihde, "Manual and gaze input cascaded (MAGIC) pointing," in *Proc. Conf. Human Factors Comput. Syst.*, Pennsylvania, USA, May 1999, pp. 246–253.

[64] D. Mardanbegi, B. Mayer, K. Pfeuffer, S. Jalaliniya, H. Gellersen, and A. Perzl, "Eyeseethrough: Unifying tool selection and application in virtual environments," in *Proc. IEEE Conf. Virtual Real. 3D User Interfaces*, Osaka, Japan, Mar. 2019, pp. 474–483.

[65] C. S. Rosales, G. Pointon, H. Adams, J. Stefanucci, S. H. Creem-Regehr, W. B. Thompson, and B. Bodenheimer, "Distance judgments to on- and off-ground objects in augmented reality," in *Proc. IEEE Conf. Virtual Real. 3D User Interfaces*, Osaka, Japan, Mar. 2019, pp. 237–243.

[66] N. Cournia, J. D. Smith, and A. T. Duchowski, "Gaze- vs. hand-based pointing in virtual environments," in *Proc. Conf. Human Factors Comput. Syst.*, Florida, USA, Apr. 2003, pp. 772–773.

[67] C. J. Lin, S.-H. Ho, and Y.-J. Chen, "An investigation of pointing postures in a 3d stereoscopic environment," *Appl. Ergonom.*, vol. 48, pp. 154–163, 2015.

[68] S. Irawati, S. A. Green, M. Billinghurst, A. Dünser, and H. Ko, "An evaluation of an augmented reality multimodal interface using speech and paddle gestures," in *Proc. Int. Conf. on Artificial Reality and Telexistence*, Hangzhou, China, 2006, pp. 272–283.

[69] C. Ware and H. H. Mikaelian, "An evaluation of an eye tracker as a device for computer input," *Proc. Conf. Human Factors Comput. Syst. and Graphics Interface*, vol. 17, pp. 183–188, May 1986.

[70] S. G. Hart, "Nasa-Task Load Index (NASA-TLX); 20 years later," in *Proc. Hum. Factors Ergon. Soc.*, vol. 50, no. 9, California, USA, Oct. 2006, pp. 904–908.

[71] J. Frost, "Degrees of Freedom in Statistics," (2017). [Online]. Available: https://statisticsbyjim.com/hypothesis-testing/degrees-freedom-statistics/

[72] A. Bangor, P. Kortum, and J. Miller, "Determining what individual sus scores mean: Adding an adjective rating scale," *J. usability studies*, vol. 4, no. 3, pp. 114–123, 2009.

[73] Y. Sugano and A. Bulling, "Self-calibrating head-mounted eye trackers using egocentric visual saliency," in *Proc. ACM Symp. User Interface Softw. Technol.*, North Carolina, USA, Nov. 2015, pp. 363–372.

[74] S. Schenk, M. Dreiser, G. Rigoll, and M. Dorr, "Gazeeverywhere: Enabling gaze-only user interaction on an unmodified desktop PC in everyday scenarios," in *Proc. Conf. Human Factors Comput. Syst.*, Colorado, USA, May 2017, pp. 3034–3044.

[75] N. Pathmanathan, M. Becher, N. Rodrigues, G. Reina, T. Ertl, D. Weiskopf, and M. Sedlmair, "Eye vs. head: Comparing gaze methods for interaction in augmented reality," in *Proc. Symp. Eye Tracking Res. Appl.*, Stuttgart, Germany, Jun. 2020, pp. 1–5.

[76] C. Trepkowski, D. Eibich, J. Maiero, A. Marquardt, E. Kruijff, and S. Feiner, "The effect of narrow field of view and information density on visual search performance in augmented reality," in *Proc. IEEE Conf. Virtual Real. 3D User Interfaces*, Osaka, Japan, Mar. 2019, pp. 575–584.

[77] K. Caine, "Local standards for sample size at CHI," in *Proc. Conf. Human Factors Comput. Syst.*, California, USA, May 2016, pp. 981–992.