IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 16, 2021, accepted March 31, 2021, date of publication April 6, 2021, date of current version April 16, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3071364

Hands-Free Human–Robot Interaction Using Multimodal Gestures and Deep Learning in Wearable Mixed Reality

KYEONG-BEOM PARK[1], SUNG HO CHOI[1], JAE YEOL LEE[1], (Member, IEEE), YALDA GHASEMI[2], MUSTAFA MOHAMMED[2], AND HEEJIN JEONG[2], (Member, IEEE)

[1]Department of Industrial Engineering, Chonnam National University, Gwangju 61186, South Korea
[2]Department of Mechanical and Industrial Engineering, University of Illinois at Chicago, Chicago, IL 60607, USA

Corresponding authors: Jae Yeol Lee (jaeyeol@chonnam.ac.kr) and Heejin Jeong (heejinj@uic.edu)

This work was supported in part by the Republic of Korea's Ministry of Science and ICT (MSIT) through the High-Potential Individuals Global Training Program supervised by the Institute of Information and Communications Technology Planning and Evaluation (IITP) under Grant 2020-0-01532, and in part by the National Research Foundation of Korea (NRF) funded by the Ministry of Education through the Basic Science Research Program under Grant 2019R1I1A3A01059082.

ABSTRACT This study proposes a novel hands-free interaction method using multimodal gestures such as eye gazing and head gestures and deep learning for human-robot interaction (HRI) in mixed reality (MR) environments. Since human operators hold some objects for conducting tasks, there are many constrained situations where they cannot use their hands for HRI interactions. To provide more effective and intuitive task assistance, the proposed hands-free method supports coarse-to-fine interactions. Eye gazing-based interaction is used for coarse interactions such as searching and previewing of target objects, and head gesture interactions are used for fine interactions such as selection and 3D manipulation. In addition, deep learning-based object detection is applied to estimate the initial positioning of physical objects to be manipulated by the robot. The result of object detection is then combined with 3D spatial mapping in the MR environment for supporting accurate initial object positioning. Furthermore, virtual object-based indirect manipulation is proposed to support more intuitive and efficient control of the robot, compared with traditional direct manipulation (e.g., joint-based and end effector-based manipulations). In particular, a digital twin, the synchronized virtual robot of the real robot, is used to provide a preview and simulation of the real robot to manipulate it more effectively and accurately. Two case studies were conducted to confirm the originality and advantages of the proposed hands-free HRI: (1) performance evaluation of initial object positioning and (2) comparative analysis with traditional direct robot manipulations. The deep learning-based initial positioning reduces much effort for robot manipulation using eye gazing and head gestures. The object-based indirect manipulation also supports more effective HRI than previous direct interaction methods.

INDEX TERMS Deep learning, eye gazing, hands-free interaction, head gestures, human–robot interaction, mixed reality, object detection.

I. INTRODUCTION

Human-robot interaction (HRI) is attracting much attention with the advent of collaborative robots that can increase productivity and efficiency in the manufacturing industry. Therefore, HRI is considered one of the important research topics for supporting human operators to interact and collaborate with robots in shared working environments.

Because existing industrial robots are both programmed and assigned to perform repetitive tasks, it is difficult to cope with uncertainties when an unexpected situation occurs or the work environment changes. However, collaborative robots can perform intelligent tasks even in dynamic and uncertain situations by utilizing various sensors, such as RGB-D sensors and pressure sensors, enabling safe collaboration between the human operator and the robot by preventing collisions between them [1], [2]. Typically, a new interaction method and an interface tool to manipulate the

The associate editor coordinating the review of this manuscript and approving it for publication was Tao Liu.

robot are required for supporting effective HRI, as the operator mainly interacts with the robot through a 2D interface such as a teach pendant or a keyboard mouse. Recently, virtual/augmented reality (VR/AR) has been considered an essential tool to provide the user with more natural interactions with the robot or the real/virtual object [3]–[8].

Mixed reality (MR) is also considered an important interaction and visualization tool since MR is the merging of real and virtual worlds to produce new environments and visualizations, where physical and digital objects co-exist and interact in real-time [9]–[12]. MR does not exclusively take place in either the physical or virtual world but is a hybrid of reality and virtual reality, unlike AR and VR. In other words, MR is a blend of physical and digital worlds, unlocking the links between human, computer, and environment interactions [13]. Therefore, MR has been applied to various industrial fields such as HRI, manufacturing task assistance, and collaboration [5], [10], [11]. Instead of overlaying virtual objects on videos or images of the real world captured from the AR camera, the virtual environment is combined with the real world in MR. To make this possible, MR performs spatial awareness by conducting real-time scanning of the real environment [14], [15].

In order to support effective and efficient interactions in MR, hand-based interactions were widely used in previous studies, which detected the 3D joints of the hand and recognized hand gestures based on them [4], [14], [16]. However, hand-based interactions cannot be utilized by the users who hold physical objects with both hands while conducting tasks or by people with disabilities who have difficulty in using their hands. For example, for people suffering from tetraplegia, almost all activities requiring user interaction are very tedious or even impossible without the help of assistants or assistive devices. HRI provides the opportunity to create workplaces for people with tetraplegia while increasing their autonomy in activities of daily living. Hands-free HRI can even provide an opportunity to integrate disabled people into working life [9], [17]–[22].

Recently, several studies were conducted to utilize eye gazing and head gestures for hands-free interactions because eye-tracking sensors and gyro-sensors are embedded in smart devices. Sidenmark and Gellersen [23], Sidenmark *et al.* [24] leveraged the synergetic movement of eye and head with naturally combined eye-head movement and refined the cursor position with gestural head movement. However, they performed only simple tasks such as selection, and they did not perform more complicated tasks such as 3D manipulation with 3D virtual objects. Other studies focused on controlling wheelchairs and robots with head gestures for disabled people who could not use their hands [18]–[21]. It is important to note that eye gazing causes less fatigue than head gesturing because the eye gaze can move quickly without head movement, but the accuracy is lower than that of the head movement due to the jittering of the eye movement and the limitation of the performance of the eye tracker [24], [25]. For this reason, when the robot is directly manipulated using eye gazing or head gestures, it is very difficult to control its position and orientation. In addition, direct manipulation might cause the robot to move out of the specified limits, which either stops the robot or causes dangerous situations [17].

This study proposes a novel hands-free interaction method for HRI through multimodal gestures such as eye gazing and head gestures and deep learning in MR environments. The proposed approach supports coarse-to-fine interactions for providing more effective and efficient task assistance. Because eye gazing is fast but unstable due to jittering, it is used for coarse interactions such as searching and previewing objects. On the other hand, since head gestures can support stable and accurate interactions while they may cause high fatigue due to frequent head movements, they are used for fine interactions such as 3D manipulation and final positioning. The coarse interaction helps the user search for an object or user interface (UI) using eye gazing as well as navigate or preview related information. The fine interaction can help the user select the navigated or previewed objects and conduct object manipulation such as 3D translation and 3D rotation using head gestures. In particular, instead of directly manipulating the end effector of the robot, the virtual object-based indirect manipulation is proposed to assist the user in controlling the robot to perform pick-and-place tasks more effectively. The indirect manipulation is based on matching the virtual object onto the real object using hands-free HRI. Furthermore, deep learning-based object detection is applied to support fast and accurate initial object positioning, which can considerably reduce much effort for HRI. A digital twin, the synchronized virtual robot of the real robot, is used to provide a preview and simulation of the real robot to manipulate it more effectively and accurately.

In order to confirm the originality and advantage of the proposed hands-free HRI, we conduct two case studies. First, we evaluate the proposed initial object positioning using a deep learning method concerning three factors: (1) the learning performance of the deep learning-based object detection, (2) the accuracy of 2D bounding boxes, and (3) the evaluation of 3D distance errors in the 3D MR environment. Second, we conduct a comparative evaluation of the proposed virtual object-based indirect manipulation with traditional direct robot manipulations including: (1) joint-based manipulation and (2) end effector-based manipulation.

The proposed hands-free HRI has the following contributions.

1) We propose a novel hands-free HRI using the coarse-to-fine metaphor in wearable MR environments, supporting effective HRI even in situations where it is difficult for the human operator to use their hands while holding tools.

2) We propose the virtual object-based indirect robot manipulation using eye gazing and head gestures by matching the virtual object onto the physical object, which can provide more effective task assistance.

3) The initial object positioning using deep learning-based object detection can reduce much effort for hands-free HRI.

4) We confirm the originality and the excellence of the proposed hands-free HRI by conducting two case studies: (1) performance evaluation of initial object positioning using deep learning-based object detection and (2) comparative analysis with traditional direct robot manipulations.

The paper is organized as follows. In Section II, we review the related work. Section III overviews the proposed approach and presents the eye gazing- and head gesture-based hands-free interaction to search, select, and manipulate virtual objects, which is effectively used for indirect robot manipulation. Section IV proposes the accurate and effective initial object positioning using deep learning and the virtual object-based indirection manipulation of the robot through the proposed hands-free HRI. Section V describes two case studies and comparative analyses. Section VI concludes the paper and presents future studies.

II. RELATED WORK

We classify the related work into three categories and review each of them. Firstly, eye gaze and head movement-based interactions in VR/AR environments are reviewed. Secondly, we discuss previous studies on assistive HRI in constrained situations where users cannot use their hands. Finally, we review VR/AR/MR applications for HRI.

A. EYE GAZE AND HEAD MOVEMENT-BASED INTERACTIONS IN POINTING AND SELECTION

Previous research works explored effective interactions with the robot and environment using different gestures such as hand gestures, eye gazing, and head movement. Hand gestures are mainly used for main interactions but eye gazing and head gestures are used as auxiliary tools. In particular, simple interactions such as pointing and selection were supported in hands-free situations by using eye gazing and head gestures [26]–[29].

Selection by pointing has two phases. First, the user identifies an intended target by pointing at it. Then, the user confirms the target via a dwelling action or head movement. Eye movement is highly effective for the pointing phase as the user can direct his/her gaze more quickly toward a target than hands or any other pointing device [23]. A gaze shift will typically start with eye movement that will be supported by head movement, not only to reach further but also to stabilize the eyes in a comfortable position after reaching a target [30], [31]. However, eye movement is unstable due to jittering.

In order to solve the disadvantage of the eye movement, the coarse-to-fine interaction was studied in which eye gazing was designated as a coarse stage, and mouse, pen, and touch interfaces were designated as a fine stage [25], [27]. Pfeuffer *et al.* [25] proposed combining the two modes for direct-indirect input modulated by gaze and introduced gaze-shifting as a new mechanism for switching the input mode based on the alignment of the manual input and user's visual attention. They implemented direct-indirect input enabled by gaze-shifting. Stellmach and Dachselt [27] proposed gaze-supported interaction as a more natural and effective way by combining a user's gaze with touch input from a handheld device. Also, some studies were conducted to reduce selection errors by setting the dwell time for improving unstable interactions of eye gazing [32], [33].

Most of the previous studies only supported simple tasks such as pointing and selection by combining eye gazing and head gestures [23], [24], [34]–[36]. Sidenmark and Gellersen [23] proposed leveraging the synergetic movement of eye and head and identified design principles for Eye&Head gaze interaction. They suggested three eye-head coordination methods: Eye&Head pointing, Eye&Head dwell, and Eye&Head convergence. Sidenmark *et al.* [24] also proposed BimodalGaze, a technique for head-based refinement of a gaze cursor. This technique leveraged eye-head coordination, which allowed users to quickly shift their gaze to targets over larger fields of view with naturally combined eye-head movement and to refine the cursor position with gestural head movement. Mardanbegi *et al.* [34] proposed a vestibule-ocular reflex (VOR)-based gaze depth estimation method to resolve target ambiguity in 3D gaze interaction. They conducted a user study that showed the possibility of resolving the ambiguity caused by the occlusion problem when target selection was made by gaze and head gestures. Nukarinen *et al.* [35] proposed a technique, HeadTurn, that allowed a user to look at a device and to then control it by turning his or her head to the left or right. They evaluated HeadTurn using an interface that linked head-turning to increasing or decreasing a number shown on the display. Špakov *et al.* [36] suggested using a combination of eye pointing and subtle head movements to achieve accurate hands-free pointing in a conventional desktop computing environment. Based on their findings, experimental results showed that head-assisted eye pointing significantly improved the pointing accuracy without a negative impact on the pointing time. Although previous works have shown promising directions using eye gaze and head movement in conducting pointing and selection tasks, they did not perform more complicated tasks such as 3D manipulation with 3D virtual objects.

B. ASSISTIVE HRI APPLICATIONS USING HANDS-FREE INTERACTIONS

Several previous studies were conducted to help disabled people manipulate robots or physical objects. Wheelchairs were controlled by using head gestures for disabled people who were unable to use hands [18]–[20]. Ruzaij *et al.* [18] proposed an auto-calibrated head orientation controller for wheelchairs and rehabilitation robotics applications using micro-electromechanical system (MENS) sensors and embedded technologies. The system movement and speed control were dependent on the position of the user's head related to the X, Y, and Z axes. Laddi *et al.* [19] proposed an unobtrusive head gesture-based directional control system for maneuvering a patient mobility cart. The gesture detection was done by the face alignment technique developed

using a regression-based supervised learning method. Ohtsuka *et al.* [20] dealt with a simple non-contact detection method of horizontal head gesture motion using a depth sensor for the development of an intelligent wheelchair.

Other studies proposed head gesture-based assistive robot control methods for disabled people. Fall *et al.* [21] described the design of a highly intuitive wireless controller for people living with upper-body disabilities with partial or complete control of their neck and shoulders. Inertial measurement units (IMUs) were connected to a microcontroller and helped to measure the position of the user's head and shoulders using a complementary filter approach. Jackowski *et al.* [17] developed an assistive robot system through adaptive head motion control for user-friendly support (AMICUS) to increase autonomy for motion-impaired people. They conducted a usability study to validate the AMICUS interaction technology and design. However, head-only movement may cause high fatigue because all interactions are made using head gestures. Kyrarini *et al.* [22] proposed a head gesture-based interface for hands-free robot control and presented a framework for robot learning from demonstration. The head gesture-based interface was suggested using a camera mounted on a hat the user wore, and the head gesture recognition was performed using the optical flow for feature extraction and the support vector machine for gesture classification. The recognized head gestures were further mapped onto robot control commands to perform object manipulation tasks. However, the proposed approach could not support the fine-tuning of the robot manipulation for accurate interactions.

C. HRI APPLICATIONS USING AR/MR AND DEEP LEARNING

Because AR/MR can embed 3D virtual information onto the real environment, they are widely used in various fields such as HRI, manufacturing, and robot manipulation [2], [5], [6], [9], [38]. Huy *et al.* [37] proposed a new interface framework for HRI using a laser-writer instead of a projector - suitable for indoor and outdoor applications. In addition, the combination of see-through head-mounted display AR and spatial AR was suggested to enhance the security level of exchanging information. Kousi *et al.* [5] presented an AR-based software suite for supporting operators in production systems that employ mobile robots. Chadalavada *et al.* [38] proposed eye-tracking glasses as safety equipment in industrial environments shared by humans and robots. They investigated the possibility of human-to-robot implicit intention transference solely from eye gaze data and evaluated how the observed eye gaze patterns of the participants were related to their navigation decisions. Chacko and Kapila [6] proposed a mobile AR interface for HRI in a shared working environment by fusing marker-based and markerless AR technologies. The mobile AR interface enabled a smartphone to detect planar surfaces and localize a manipulator robot in the working environment. The AR interface and robot manipulator were integrated to help users to perform pick-and-place tasks effortlessly. However, it suffers from problems such as attaching markers on the objects directly as well as mismatching errors caused by the marker occlusion. On a similar note, Mohammed *et al.* [39] addressed real-time 3D object tracking of shared working environments with AR integration. They focused on sending warnings to users based on predicted human-robot collision severity. Krupke *et al.* [9] proposed the concept and implementation of an MR-based human-robot collaboration system in which a human can intuitively and naturally control a co-located industrial robot arm for pick-and-place tasks. They compared two different multimodal HRI techniques to select the pick location on a target object using head orientation or pointing, both in combination with speech. The results showed that head-based interaction techniques are more precise while requiring less time. However, they might cause high fatigue because the robot was manipulated using only head gestures. Guhl *et al.* [7] presented a system that allowed the user to interact with an industrial robot and other cyber-physical systems via AR and VR. Although their approach could visualize paths of the robot's end-effector through 3D virtual lines, it could not support robot manipulation tasks such as picking and placing.

Recently, deep learning was applied to support task assistance in HRI and assembly. Park *et al.* [10] proposed a smart and user-centric task assistance method that combined deep learning-based object detection and instance segmentation with wearable AR technology to provide more effective visual guidance with less cognitive load. They also proposed a deep learning-based mobile AR for intelligent task assistance by conducting 3D spatial mapping between the physical and virtual robots without pre-registration using AR markers [11]. Although their method enabled the manipulation of the robot directly and effectively, it could not support hands-free interactions in situations where the user's hands could not be used.

Although previous studies on hands-free interactions support various kinds of targeting and selection tasks using eye gazing and head gestures in MR/AR environments, most of them do not support complicated tasks such as 3D manipulation. HRI using eye gazing and head gestures takes a lot of time and effort since the user has to manipulate multiple objects. In addition, frequent head movements can cause physical stress during the interactions, and direct manipulations of the robot's joints or end effector makes it difficult for the user to conduct pick-and-place tasks using eye gazing and head gestures. For these reasons, there is still much room for improvement of previous studies concerning hands-free HRI and robot manipulation.

III. HANDS-FREE INTERACTIONS FOR HRI BY COMBINING EYE GAZING AND HEAD GESTURES

This study proposes a new approach to the coarse-to-fine hands-free interactions using multimodal gestures such as eye gazing and head gestures and deep learning-based initial object positioning in constrained HRI environments. There are two different types of HRI interactions: 1) search and navigation and 2) 3D manipulation. It should be noted that

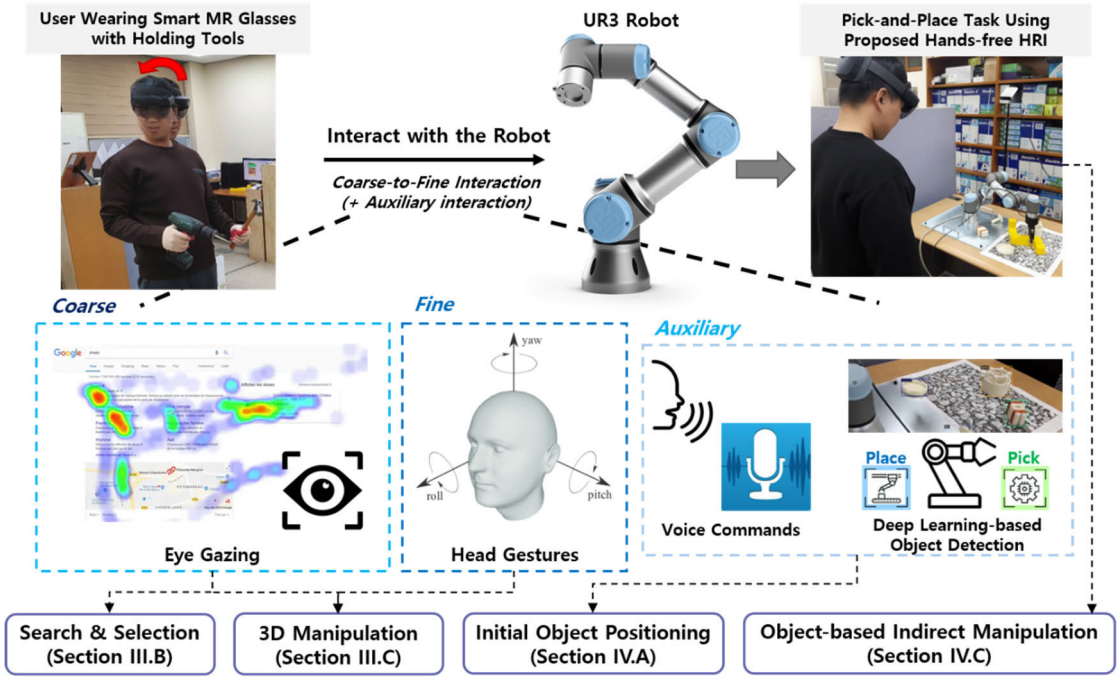

FIGURE 1. The framework of the proposed hands-free HRI.

most previous research works have not dealt with 3D object manipulation because their methods are mainly applied to 2D plane-based selection tasks, not 3D operations.

A. OVERVIEW OF THE PROPOSED APPROACH

The framework of the proposed hands-free HRI is shown in Figure 1. It consists of coarse-to-fine main interactions and auxiliary interactions. The coarse-to-fine main interactions consist of 1) eye gazing-based coarse interactions and 2) head gesture-based fine interactions. It is important to note that eye movement is faster and requires less energy, while head movement is less jittery and more controlled [23], [24], [30], [34], [35]. Therefore, the eye gazing-based coarse interaction is used for the search and preview of intended target objects or UIs by tracking the user's pupils and calculating the eye pointer's location in the MR display which the user is looking at. The head gesture-based fine manipulation is used for the final selection and 3D manipulation of objects. The fine interaction can also be effectively used for exact matching between the virtual and physical objects using pitch, yaw, and roll-based head gestures. Details on the search & preview and selection & 3D manipulation using eye gazing and head gestures will be described in Section III.B and III.C, respectively.

The proposed approach also supports auxiliary interactions, such as voice commands and object detection-based initial positioning. Voice commands are used to capture an MR image for object detection or reset the pose of the robot. In addition, they are used for changing manipulation modes. We utilized the voice commands API supported by HoloLens 2. Deep learning-based object detection is used to support initial object positioning to estimate the locations of physical objects in the MR space. We utilize RetinaNet [40] for multiple object detection. Details on initial object positioning will be described in Section IV.A. By applying deep learning-based initial object positioning, this approach can assist more effective HRI in hands-free situations.

Unlike the traditional methods that directly manipulate the robot's joints or end effector, this study proposes the virtual object-based indirect manipulation method, which matches the virtual object's pose with the real target object. Indirect manipulation proves that the user can conduct pick-and-place tasks more effectively. Details on the virtual object-based indirect manipulation will be described in Section IV.C.

The hardware and software used to implement the proposed approach are as follows. We use the Microsoft (MS) HoloLens 2 [14] as MR smart glasses and Unity 3D [41] to develop the MR system. MS HoloLens 2 has built-in gyroscope and acceleration sensors and an eye tracker that tracks the user's eye gaze in real-time. The Universal Robot UR3 [42], a well-known as a collaborative robot, is used for developing the proposed HRI. For testing the proposed HRI, we evaluate how easily and effectively the user wearing the HoloLens 2 can control the UR3 robot through the proposed hands-free HRI.

B. EYE GAZING AND HEAD GESTURES FOR SEARCH AND SELECTION

Through the proposed hands-free method, a user can perform a search and navigation task using eye gazing, and then he or she can perform a selection task using head gestures. Two pointers on MR smart glasses are defined for conducting a combined search and selection task: the eye pointer and the head pointer. The eye pointer is defined at the location on the

display of MR smart glasses where the pupils are looking, and the head pointer is defined at the center location of the MR display, which is the frontal direction of the user's face. Therefore, the eye pointer changes its location on the MR display when the eyes are moving while the head pointer is fixed at the center of the display, as shown in Figure 2. The built-in eye tracker of the HoloLens 2 is used for eye gazing and the eye pointer is shown as a green circle on its display. The head pointer is displayed as a green circle. The head gesture is calculated using the gyro-sensors of the HoloLens 2.

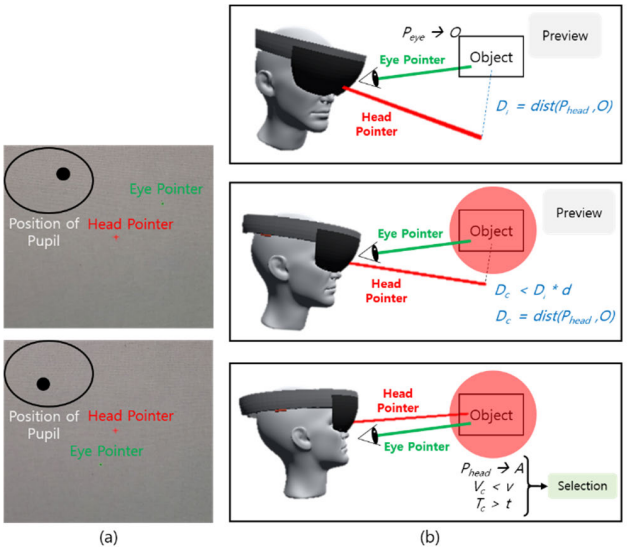

FIGURE 2. A detailed description of the proposed search and selection interaction using eye-head coordinate gesture: (a) positions of pupil, head pointer, and eye pointer, and (b) procedure of the search and selection using eye-head coordinate gesture (Object: intended target object, Preview: preview of the detailed information on the target object, Selection: final selection of the object).

The algorithm for the proposed search and selection method is shown in Figure 3. Figure 2 shows the search and selection task's process using the proposed hands-free interaction by combining eye gazing and head gestures. First, the search starts when the eye pointer touches an intended target object or the UI. Then, the preview of the detailed information of the gazed object is displayed around it. If two objects are in the direction of the eye gaze, the nearest object from the eye is selected. That is, when a ray is fired from the direction of the eye gaze, the first hit object is selected. Finally, the user can determine whether or not to finally select the gazed object while reviewing the augmented information (top in Figure 2(b)). The final selection of the gazed object through head gestures is determined as follows. The reference distance between the head pointer and the gazed object, Di, is calculated. When the head pointer moves toward the gazed object within the threshold distance $Di * d$, an activation graphical user interface (GUI) A is shown to confirm the readiness to select the gazed object (middle in Figure 2(b)). When the head pointer collides with the activation GUI, and the dwelling time is longer than t seconds at a rate less than v velocity, the object is selected [23] (bottom in Figure 2(b)). In this study, we set d= *0.8,* v= *20°/sec*, and t= *1 sec*.

C. EYE GAZING AND HEAD GESTURES FOR 3D MANIPULATION

Coarse interactions correspond to discrete interactions with objects such as search, selection, or preview through eye gazing as explained in Section III.B. On the other hand, fine interactions correspond to continuous manipulations of objects. As shown in Figure 4, the process of 3D manipulation is similar to that of the search and selection task. First, eye gazing is used in navigating to find a manipulation mode. While navigating the mode, the preview is also shown around the virtual object. Then, the head gesture is performed to select the manipulation mode. After the manipulation mode is selected, the user can conduct the direct 3D manipulation of the object using pitch, yaw, and roll-based head gestures.

Because it is difficult to manipulate the virtual object in the 3D space, the manipulation mode is defined along X, Y, and Z axes for 3D translation (Figure 4 (a)) and 3D rotation (Figure 4 (b)). In addition, the user can manipulate the object in a different mode, such as coarse translation, fine translation, coarse rotation, and fine rotation. To simplify the coarse-to-fine interaction, the user can define a plane for easily manipulating the object. Thus, the object is manipulated on the defined plane, which reduces the degrees of freedom (DOF) of the object and supports more intuitive interaction. For example, the XZ plane-based translation makes it possible for the object to move along the horizontal plane, and the XY plane-based translation makes it possible for translating the object along the vertical plane, as shown in Figure 4. In a similar manner, the rotation along the Y-axis in the world coordinate is useful when the object is located on the horizontal plane (WY, World coordinate-based Y-axis). One of the main reasons for the plane- or axis-based manipulations is that most of the robot manipulations such as pick-and-place tasks are conducted on the planar surface, such as on a horizontal desk or on a vertical wall. In addition to the world coordinate-based manipulation, the manipulation based on the camera coordinate (Camera coordinate-based Axes) is supported using pitch, yaw, and roll gestures. This is defined based on the head of the user wearing MR glasses instead of the world coordinate, as shown in Figure 4. Furthermore, auxiliary interactions such as voice commands can be used for selecting manipulation modes and reset the interaction environment. We utilized the voice commands API supported by HoloLens 2.

An example of 3D manipulation using eye gazing and head gestures for matching a virtual object onto a real object is shown in Figure 5. Figure 5(a) shows the working environment. First, a virtual object can be created in the MR environment through a user interaction. The user is then ready to match the virtual object with the physical object

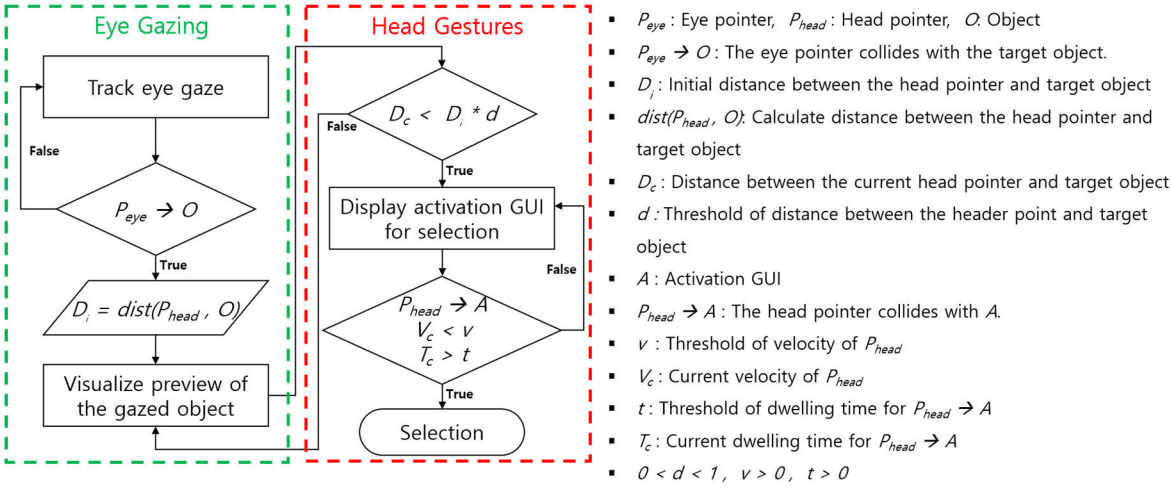

FIGURE 3. Algorithm of the search and selection using the proposed hands-free interaction for interacting with the target object in MR.

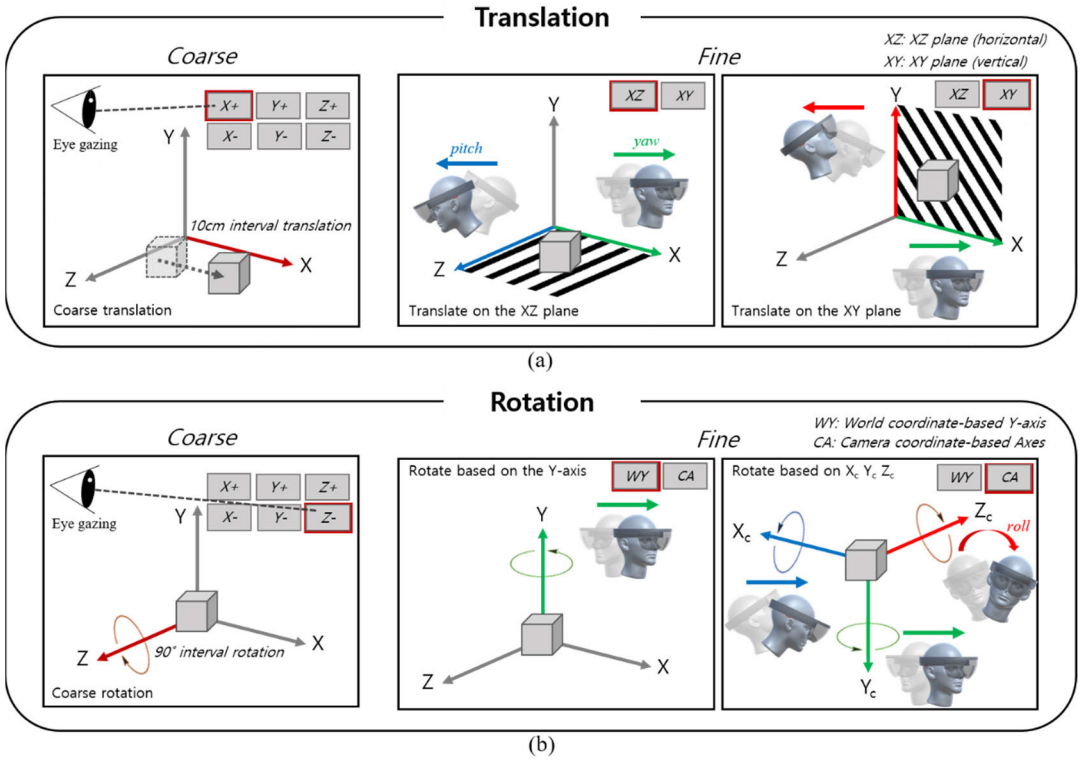

FIGURE 4. 3D manipulations using eye gazing and head gestures: (a) coarse-to-fine translation and (b) coarse-to-fine rotation.

using eye gazing and head gestures. The matching process takes coarse-to-fine manipulation steps. For coarse manipulation, the user navigates or searches for a specific menu using eye gazing and then AR smart glasses show a preview or simulation of the intermediate pose of the virtual object. In Figure 5 (b), 'FT' stands for fine translation, 'CR' for coarse rotation, 'FR' for fine rotation, 'Sel' for selection, and 'Can' for cancellation. For example, when the selection mode is chosen, the robot moves to pick or place the physical object, which will be explained in the next section. When the 'Can' mode is selected, the user cancels the manipulation of the object. Finally, the user can match the virtual object with the physical object through head gestures, as shown in Figure 5(c) and Figure 5(d). However, we have found that, when the virtual is created, its initial position in the MR space can influence the whole process of the intended manipulation using hands-free interactions. Therefore, it is very important to support accurate initial object positioning, which can reduce eye gazing- and head gesture-based interactions.

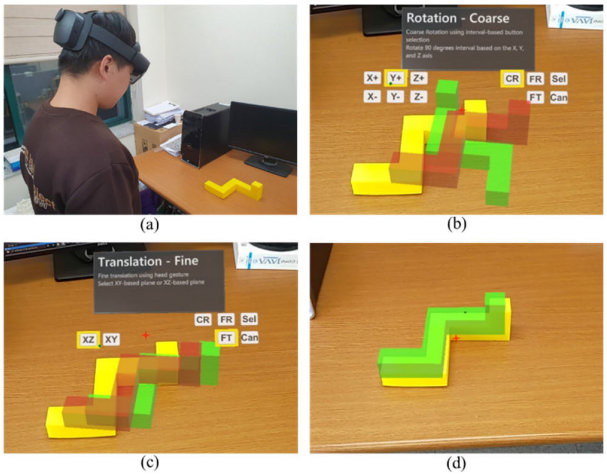

FIGURE 5. Example of manipulating 3D virtual object using hands-free interaction for matching the virtual object (green) onto the real object (yellow): (a) real working environment with hands-free interaction, (b) coarse manipulation for rotation about Y-axis and preview of the rotated virtual object (red), (c) fine manipulation for translation about the XZ-plane and preview of the translated object (red), and (d) matching between the virtual and real objects.

IV. DEEP LEARNING-BASED INITIAL OBJECT POSITIONING AND INDIRECT ROBOT MANIPULATION

One of the important key features of the proposed hands-free HRI is the application of both deep learning-based object detection and spatial mapping for initial object positioning, which helps the user conduct the pick-and-place task more effectively and intuitively while reducing eye gazing- and head gesture-based interactions. Usually, many interactions are required to control the robot before making the robot take actual pick-and-place actions. Another key feature is to support the virtual object-based indirect robot manipulation instead of its direct manipulation for the pick-and-place task. We analyzed the joint-based manipulation and the end effector-based manipulation that directly control the robot in situations where the user cannot use his or her hands. However, we found inherent problems in this such as a lot of time and effort being needed to predict the configuration of the robot's end effector while manipulating the robot using eye gazing and head gestures. Therefore, it is difficult to accurately perform pick-and-place tasks through 3D manipulations with many DOF. On the other hand, the virtual object-based indirect manipulation supports the effective and intuitive robot control. This is done by matching the physical object's pose with that of the virtual one using eye gazing and head gestures rather than directly manipulating the end effector or joints of the robot in conventional HRI. The proposed approach calculates the inverse kinematics of the robot with respect to the virtual object and communicates the calculated kinematic information to the robot, which supports more effective HRI than conventional direct manipulations. Case studies are given to prove the advantage and effectiveness of the proposed indirect manipulation for HRI in Section V.

A. INITIAL OBJECT POSITIONING USING DEEP LEARNING-BASED OBJECT DETECTION AND SPATIAL MAPPING

The initial object positioning process consists of two steps: 1) deep learning-based object detection and 2) 3D spatial mapping on the 3D reconstructed area. A deep learning-based object detection method can automatically estimate the real object's initial position, which detects object classes and 2D bounding boxes from the MR image captured from smart AR glasses. After that, 3D spatial mapping is conducted to embed the results of object detection onto the 3D MR environment.

We used RetinaNet [40] for object detection, one of the state-of-the-art research works, which solved the extreme foreground-background class imbalance encountered during the training of dense detectors. RetinaNet addressed the class imbalance by reshaping the standard cross-entropy loss such that it down-weights the loss assigned to well-classified examples. The Focal Loss focuses training on a sparse set of hard examples and prevents the vast number of easy negatives from overwhelming the detector during training [40]. The architecture of RetinaNet is shown in Figure 6. Instead of using the results on 2D images, we apply them to the 3D reconstructed MR space to find good initial object positioning before conducting coarse-to-fine manipulation.

Before applying the deep learning method to the working environment, training was performed based on five objects by fine-tuning RetinaNet. As shown in Figure 7(a), in the case study, four components of a motor in a cordless vacuum cleaner and a puzzle object used for the mental rotation test were trained. Four components of the motor consist of body, bottom case, coil, and fan. All five components are made by a 3D printer.

The training data consist of a total of 100 images for each object. The pre-trained weights learned from the COCO dataset [43] was used for transfer learning. The learning rate was 0.00025, the batch size was 4, and the epoch was 5,000. An NVIDIA GeForce 2080 Ti 11GB was used for training and inferencing. It takes about 0.066 second per image or 15.17 frame per second (FPS) for the object detection using RetinaNet. Figure 7(b) shows the result of object detection in a pick-and-place task using RetinaNet.

The object detection results are a set of bounding boxes on the MR image that must be mapped onto the 3D MR environment because of a mismatch between the 2D image space and MR-based 3D spaces. A ray-casting method is applied to spatially map the 2D image space onto the MR space, which takes the following steps [10]. The process of the proposed initial object positioning is shown in Figure 8.

1) Capturing an RGB image in HoloLens 2
2) Creating a 2D image plane in front of the user in the MR environment
3) Performing object detection and detecting 2D bounding boxes using RetinaNet
4) Calculating a direction vector between HoloLens 2 and the center of 2D bounding boxes in the 2D image plane

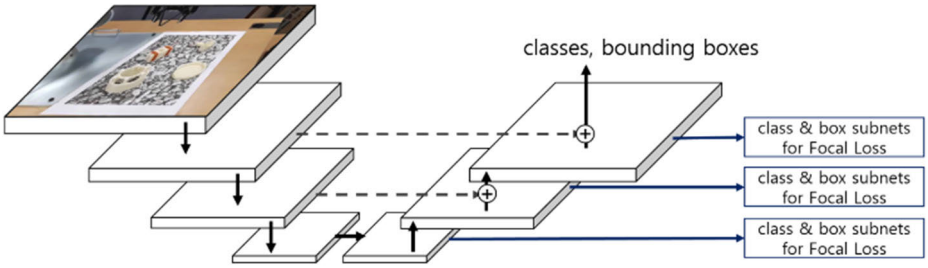

FIGURE 6. Architecture of RetinaNet [40] for object detection.

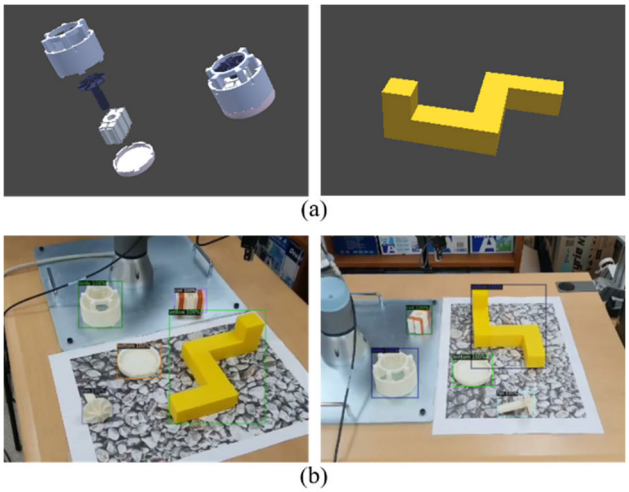

FIGURE 7. Object detection using RetinaNet to the working environment: (a) four components of a motor in a cordless vacuum cleaner and one puzzle object and (b) results of the object detection.

5) Casting a ray toward the calculated direction vector
6) Mapping the detected boxes on the reconstructed 3D MR space in HoloLens 2

B. INVERSE KINEMATICS OF UR3 FOR THE DIGITAL TWIN

It should be noted that the proposed hands-free HRI in the MR environment uses a digital twin, the synchronized virtual robot of the real (physical) robot. Therefore, it is possible to manipulate the real robot UR3 by interacting with the virtual robot through gestures. Conversely, it is possible to manipulate the virtual robot by controlling real robot. The main reason for using a digital twin is to provide a preview and simulation of the real robot to manipulate it more effectively and accurately in the MR environment. Figure 9 shows the real robot and its synchronized digital twin.

It is necessary to solve forward and inverse kinematics of the robot [44], [45] and to synchronize the solved kinematics with those of the digital twin whenever an interaction occurs, and vice versa. The robot agent plays the role to support the bi-directional communication between them. Figure 10 and Table 1 show the Denavit-Hartenberg parameters of the UR3 robot [42]. Based on these parameters, the real robot's digital twin can be virtually modeled [46]. The configuration of the digital twin is synchronized with the real robot and updated whenever an event occurs. Furthermore, when the digital twin is manipulated in the MR environment, its configuration is also sent to the real robot through the robot agent for synchronization. Details on calculating inverse kinematics are described in references 44 and 45.

TABLE 1. D-H parameter values of the UR3 robot [42].

Link	θ (radian)	a (mm)	d (mm)	α (radian)
Joint 1	θ_1	0	151.9	π / 2
Joint 2	θ_2	-243.65	0	0
Joint 3	θ_3	-213.25	0	0
Joint 4	θ_4	0	112.35	π / 2
Joint 5	θ_5	0	85.35	-π / 2
Joint 6	θ_6	0	81.9	0

C. OBJECT-BASED INDIRECT ROBOT MANIPULATION FOR HANDS-FREE HRI

One of the key contributions is that, instead of directly manipulating either the joints or the end effector of the robot, the proposed approach supports the same task more effectively and efficiently by enabling indirect manipulation of the real robot through virtual object matching-based HRI. The proposed object-based indirect robot manipulation is shown in Figure 11. After initial object positioning, the user can conduct the matching of the virtual object to the real object using eye gazing and head gestures. When the matching is completed, the proposed method calculates the inverse kinematics of the digital twin by synchronizing the coordinate of the end effector with the virtual object's pose. In addition, it verifies the kinematics by simulating the digital twin. Finally, the proposed method commands the real robot to be synchronized with the digital twin, as shown in Figure 11(c).

The proposed hands-free HRI has been implemented based on three layers: 1) interaction and visualization, 2) deep learning, and 3) digital twin, as shown in Figure 12. The interaction and visualization layer supports coarse-to-fine interactions using multimodal gestures. In addition, interaction results are superimposed on the display of the smart MR glasses that provide step-by-step instructions and previews for object manipulation. The deep learning layer supports 2D object detection that is effectively used for initial

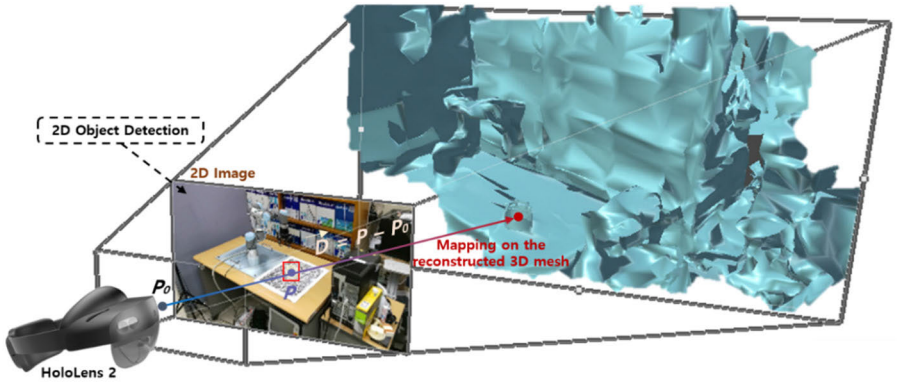

FIGURE 8. Process of the proposed initial object positioning.

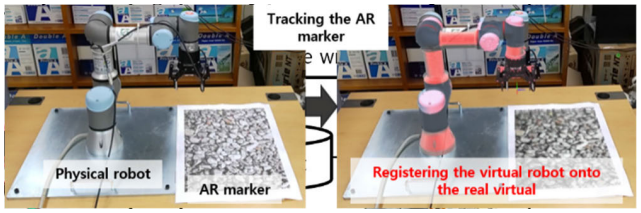

FIGURE 9. Example of registering the virtual robot onto the real robot by tracking the AR marker.

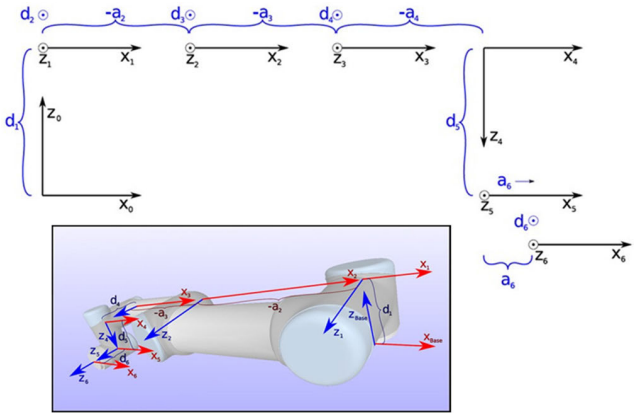

FIGURE 10. Kinematic configuration of the UR3 robot for calculating Denavit-Hartenberg parameters (D-H) [42].

object position before conducting eye gazing- and head gesture-based manipulations. The digital twin layer supports interactions between the virtual and real robots. The robot agent supports bi-directional communication between the two robots and synchronizes the kinematic parameters between them.

The process of the proposed hands-free HRI takes the following steps: (1) initial object positioning using deep learning-based object detection, (2) coarse-to-fine manipulation using eye gazing and head gestures for matching the virtual object onto the physical object needing to be manipulated, (3) previewing the robot movement and configuration through the digital twin before conducting an actual task, and (4) commanding the real robot based on the pose of the virtual object matched with the real object through calculating the inverse kinematics of the real robot. Therefore, the proposed approach makes it possible to support various tasks more effectively than previous approaches using eye gazing and head gestures.

An example of a pick-and-place task using the proposed method is shown in Figure 13. The working environment is shown in Figure 13(a). First, initial object positioning is conducted using deep learning-based object detection and 3D spatial mapping. Figure 13(b) shows an example of automatic virtual objects' initial positioning corresponding to physical objects. Without this process, it is very difficult and error-prone to conduct initial object positioning with eye gazing and head gestures. The user is then ready to conduct fine translation and rotation using head gestures after completing the menu selection using a coarse interaction, as shown in Figure 13(c). During this process, the user conducts the indirect manipulation of the virtual object to match the physical object. Then, the robot is ready to move to pick the physical object. Before controlling the real robot by sending a command, the user can preview the digital twin simulation to check whether the robot can perform the exact task the user expects. In addition, when the selection button is selected through eye gazing, the real robot moves to pick the object in the same way that the digital twin simulates (Figure 13 (e)). Moreover, as shown in Figure 13(d) and Figure 13(f), when the robot is controlled by eye gazing and head gestures, the simulation preview of the digital twin can be augmented in the MR environment before the user takes actions to control the robot. Therefore, it is possible to finally check whether the robot's control is identical to the user intent. In addition, even if the object has an arbitrary orientation, as shown in Figure 14, the virtual object can be matched through fine rotation based on the camera's coordinate system. The verification can also be performed with the previewed virtual robot to conduct a pick-and-place task accurately.

V. CASE STUDIES

We conducted two case studies to verify the proposed approach's effectiveness and advantage: 1) performance

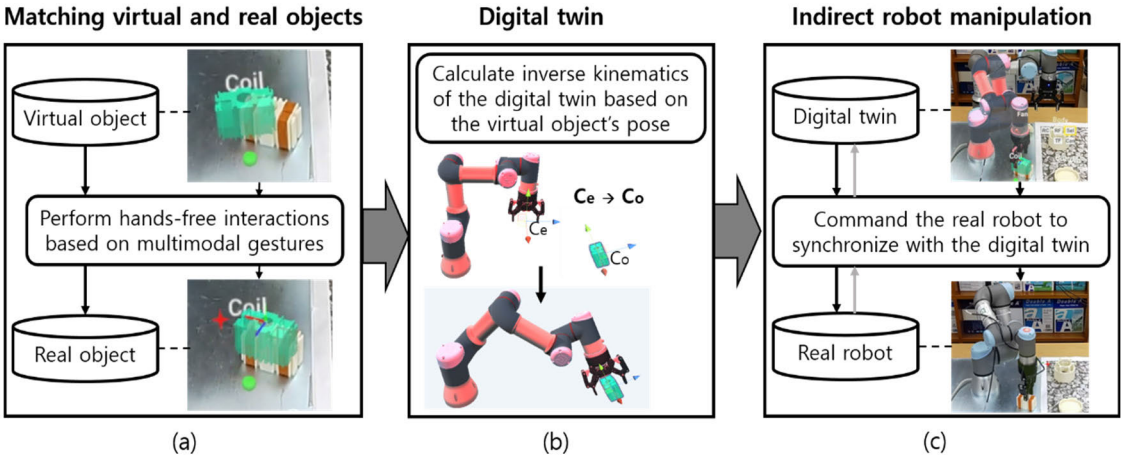

FIGURE 11. Process of virtual object-based indirect robot manipulation: (a) matching between virtual and real objects in the pick-and-place task, (b) calculating the inverse kinematics of the digital twin and simulation, (c) commanding the real robot to conduct the same task of the digital twin.

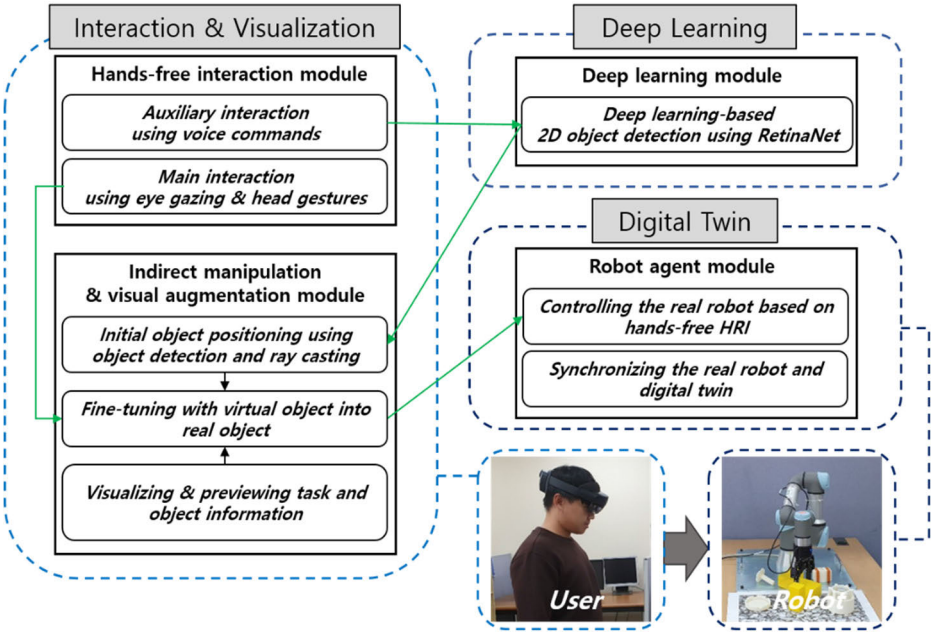

FIGURE 12. System architecture for implementing HRI in MR.

evaluation for the initial object positioning using deep learning-based object detection and 2) comparative analysis of the proposed indirect manipulation with traditional direct robot manipulation methods.

A. CASE STUDY FOR PERFORMANCE EVALUATION OF INITIAL OBJECT POSITIONING USING DEEP LEARNING-BASED OBJECT DETECTION

It takes much effort and time to estimate the target object's initial position and orientation based on the robot using eye-tracking and head gestures. Therefore, we proposed the initial object positioning through deep learning-based object detection and spatial mapping in the MR environment. We conducted a performance evaluation with respect to the accuracy of 2D object detection and the accuracy of 3D initial object positioning.

To evaluate the performance of object detection used for the pick-and-place task, we collected 150 images for the training dataset and fine-tuned RetinaNet to train the dataset. We also collected 50 images for the validation dataset. While training our dataset, we found that the losses of 150 training datasets and 50 validation datasets gradually diminished simultaneously for 5,000 iterations, which indicated that the training was performed successfully, as shown in Figure 15. As a metric for measuring the accuracy of 2D object detection, the average precision (AP) was used [47]. AP computes the average precision value for recall value over 0 to 1. The mean average precision (mAP) compares the ground-truth

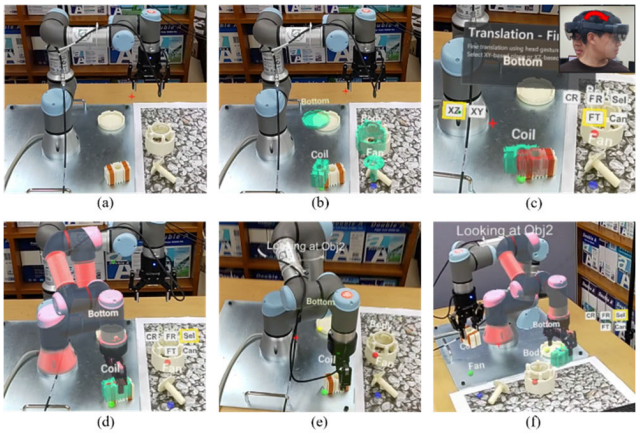

FIGURE 13. Example of the pick-and-place task using the proposed object-based indirect HRI: (a) real working environment, (b) performing initial object positioning using the deep learning-based object detection and 3D spatial mapping (green objects are virtual objects), (c) performing fine manipulation of a target object using eye gazing and head gestures, (d) previewing the simulation of the digital twin before controlling the real robot, (e) commanding the real robot toward the synchronized virtual robot and commanding the robot to pick the object, and (f) previewing the 'place' operation with the digital twin at the desired location.

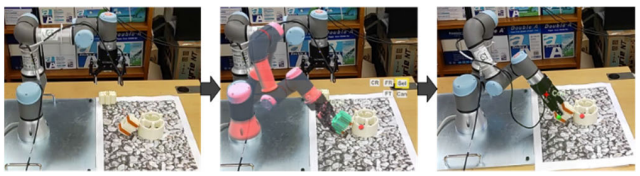

FIGURE 14. Example of picking an object rotated in an arbitrary orientation using the proposed approach.

bounding box to the detected box and returns a score. The higher the score, the more accurate the model is in its detections. Table 2 shows the performance evaluation of the 2D object detection, verifying the accurate object detection (mAP $>$ 0.9). The formulae for precision, recall, and intersection over union (IoU) are described as follows.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$IoU = \frac{area\left(Bbox_{gt} \cap Bbox_{p}\right)}{area\left(Bbox_{gt} \cup Bbox_{p}\right)} \tag{3}$$

RetinaNet was applied to the real working environment, as shown in Figure 16. Ground truth positions (blue) of the objects can be calculated based on the AR marker (Figure 16(b)). Then, the object's initial positions are estimated through deep learning-based object detection and ray-casting-based positioning in the MR environment.

We also conducted a performance evaluation of the proposed method by calculating the Euclidean distance between the ground truths (blue) and predictions (green) after applying the deep learning-based initial object positioning method, as shown in Figure 17. We tested 30 times for measuring distances of five objects within 1.5 meters (m). We have found that the distance error is less than 0.05 m on average, which shows excellent initial positioning for fine object manipulation using hands-free HRI in the next step. Based on the findings, the proposed approach has a much lower error than previous research using spatial mapping of real objects obtained with IoT sensors (the distance error was 0.2 m at a distance of 2 m in the previous study [15].).

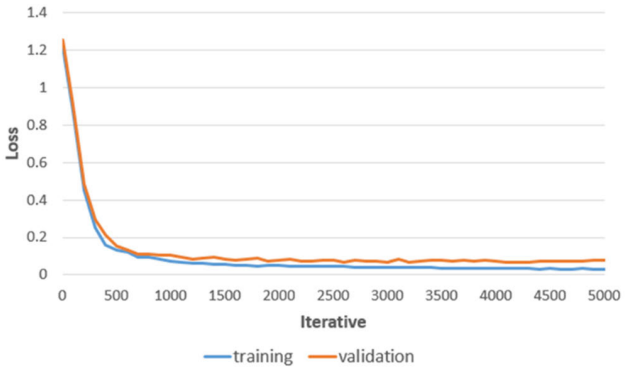

FIGURE 15. Model loss on the training of object detection for the dataset using RetinaNet.

TABLE 2. Evaluation of 2D object detection.

	Body	Coil	Fan	Bottom	Yellow	mAP
AP50	0.9505	0.9703	0.9505	0.9666	0.9741	0.9624
AP75	0.9505	0.9703	0.9505	0.9666	0.9741	0.9624
AP95	0.9062	0.8925	0.8579	0.9203	0.9662	0.9086

B. CASE STUDY FOR COMPARATIVE ANALYSIS WITH TRADITIONAL DIRECT MANIPULATIONS

We also conducted a qualitative evaluation of the proposed indirect manipulation compared with traditional direct robot manipulations including the joint-based direct robot manipulation (Figure 18) and the end effector-based direct robot manipulation (Figure 19). The joint-based direct robot manipulation enables the control of six joints of the robot directly, which requires calculating forward kinematics. The end effector-based direct robot manipulation enables the manipulation of the end effector of the robot directly, which requires the calculation of inverse kinematics. Figure 18 shows an example of the joint-based direct robot manipulation by controlling six joints. The main advantage of this method is that it does not require the calculation of the inverse kinematics of the robot. However, it takes a lot of time and effort in controlling all the joints. Furthermore, when considering the end effector movement, it is not intuitive to the user.

The end effector-based direct manipulation can support more intuitive and user-centric interactions by controlling the end effector directly, which calculates the inverse kinematics of the robot based on the end effector's pose. Figure 19 shows how to interact with the robot using hands-free HRI in the pick-and-place task. In this scenario, the user can control the robot through translation and

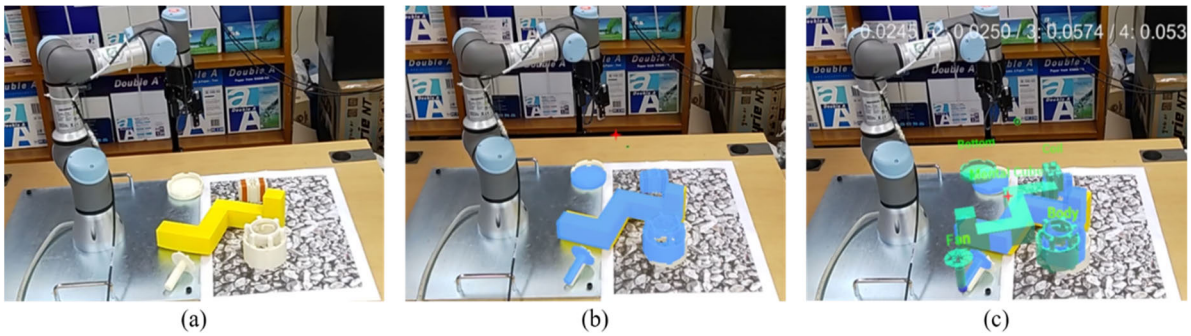

FIGURE 16. Evaluation of initial spatial mapping in MR: (a) real environment before spatial mapping, (b) ground truths (blue), and (c) calculating Euclidean distance between ground truths and predictions (green) after initial object positioning.

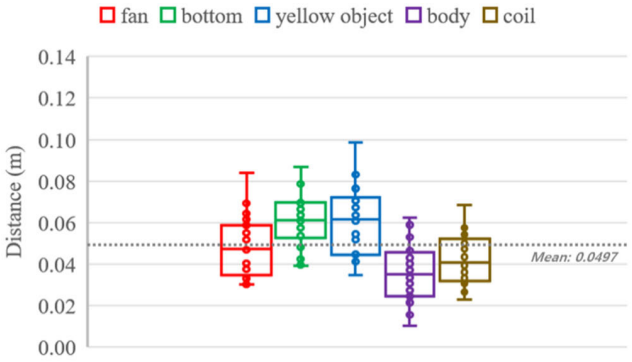

FIGURE 17. Box-and-Whisker plot for evaluating initial object positioning.

rotation operations using head gesture-based fine manipulations. However, because the inverse kinematics can generate many possible outcomes, the robot's final configuration might be unpredictable, as shown in Figure 19(b) and Figure 19(c).

The direct manipulation results might be different from the user's intent, which can cause problems such as self-collision that stops the robot. For example, direct manipulation can generate an infeasible kinematic configuration when the user tries to manipulate the end effector to have a certain configuration, as shown in Figure 20(a). On the other hand, as shown in Figure 20(b), the proposed indirect manipulation makes it possible to check the feasibility of the robot configuration in advance by calculating inverse kinematics and simulating the robot's movement using the digital twin. The proposed indirect manipulation through virtual-real object matching makes it possible to determine the infeasible configuration by simulating the robot's digital twin before controlling the real robot.

In addition, both traditional methods make it difficult for the user to conduct the pick-and-place task while directly interacting with the robot because of the lack of 3D perception, as shown in Figure 21(a) and Figure 21(b). Because traditional methods are based on direct and real-time interactions with the end effector or joints, it is very difficult to figure out the robot's 3D spatial configuration. Thus, it is not easy to control the robot using eye gazing and head gestures.

On the other hand, virtual object-based indirect manipulation can reduce the problem related to 3D spatial perception because the user does not have to directly manipulate the end effector of the robot located in the air. The object needing to be manipulated in the pick-and-place task is usually located on the floor or hung on the wall, reducing the DOF for recognizing 3D perception during the indirect manipulation. Furthermore, it is possible to figure out the feasibility of handling the object in advance. Direct manipulations make it difficult to find out whether the robot is able to pick the object before actually controlling it, as shown in Figure 22(a). However, through the virtual object-based indirect manipulation, it is possible to check the feasibility of the robot configuration for the pick-and-place task through the preview of the digital twin, as shown in Figure 22(b). In conclusion, the indirect manipulation and the preview of the synchronized digital twin can reduce inherent problems such as self-collision, 3D perception, and infeasible configuration, in comparison to previous direct manipulation methods.

C. DISCUSSION

Concerning the two case studies, the proposed method showed a very effective hands-free HRI method by combining deep learning and the digital twin. In particular, through the comparative evaluation with the existing manipulation methods, we have found that easy and fast manipulation can be supported by overcoming the shortcomings of the previous methods.

In addition to the two case studies, another test was conducted to evaluate whether the proposed approach shows better performance in task completion time. However, due to the COVID-19 pandemic, a preliminary pilot study was performed in a pick-and-place task instead of the experiment with recruited participants. Although a further study is still needed, the result shows a promising direction of the proposed hands-free interaction for HRI.

The pick-and-place task consists of picking a coil component of a motor in a cordless vacuum cleaner in a source location and placing it to a target location by manipulating the UR3 robot, as shown in Figure 23. For the joint-based direct manipulation, the user can select each joint of the robot

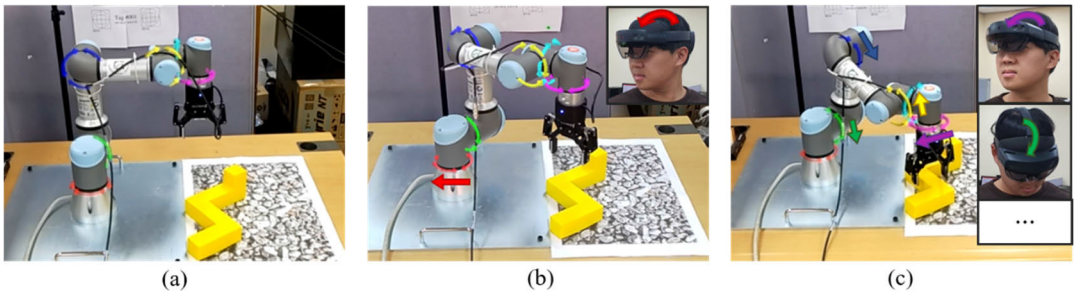

FIGURE 18. Example of the joint-based direct robot manipulation: (a) initial configuration, (b) performing joint-based control (e.g., rotation along a red axis), and (c) performing multiple axis-based rotations for picking the object.

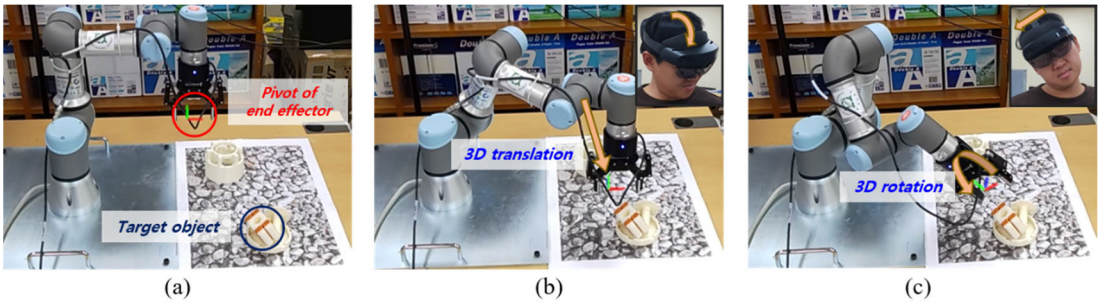

FIGURE 19. Example of the end effector-based direct robot manipulation: (a) initial configuration, (b) performing 3D translation, and (c) performing 3D rotation.

FIGURE 20. Comparative analysis of self-collision: (a) traditional direct robot manipulation, and (b) proposed indirect robot manipulation.

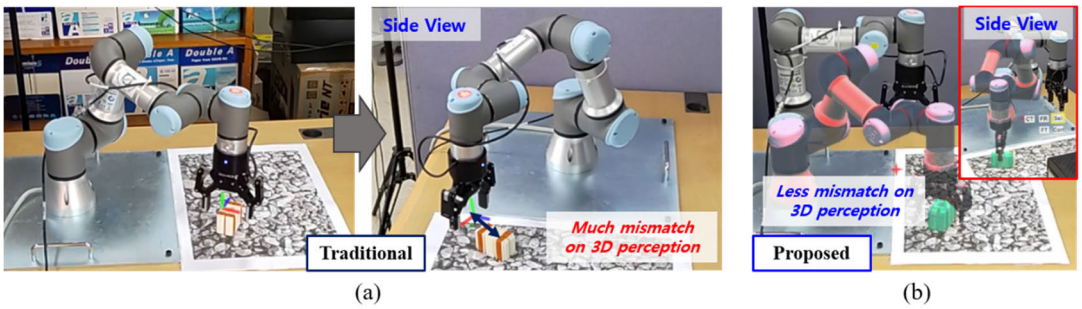

FIGURE 21. Comparative analysis of 3D perception: (a) traditional direct robot manipulation, and (b) proposed indirect robot manipulation.

through voice command and manipulate the joint through head gestures to pick and place the targeted object. For the end effector-based direct manipulation, the user can directly manipulate the end effector using head gestures. In this manipulation, the user selects a translation or rotation mode through a voice command. On the other hand, for the proposed object-based indirect manipulation, the user can indirectly conduct the manipulation by matching or positioning

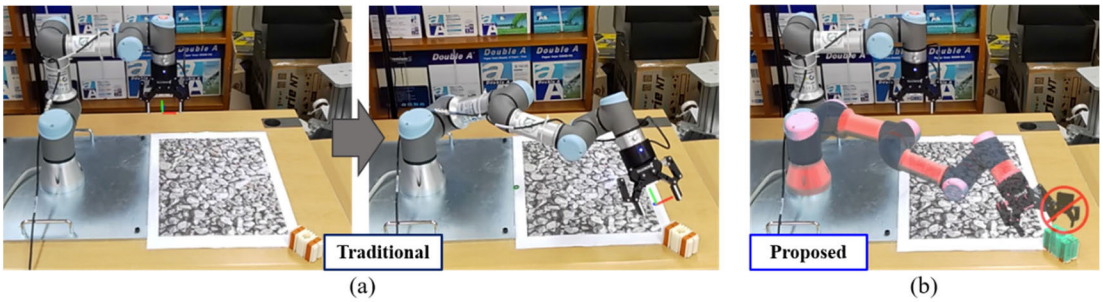

FIGURE 22. Comparative analysis of pose possibility: (a) the infeasible configuration is detected while directly controlling the robot in traditional direct robot manipulation, and (b) it is easily detected while previewing in the proposed indirect robot manipulation without controlling the robot.

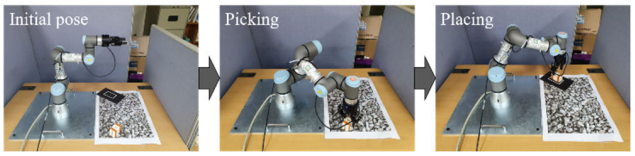

FIGURE 23. A pick-and-place task.

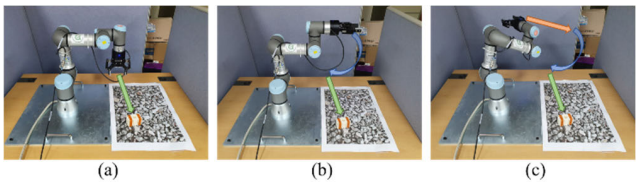

FIGURE 24. Influence of the initial pose to conduct the task: (a) easy, (b) difficult, (c) very difficult.

TABLE 3. Evaluation results of the preliminary pilot study.

Factor	Joint-based	End effector-based	Proposed
Task completion time	Long (~ 300 sec.)	Medium (~ 200 sec.)	Short (~ 110 sec.)
Number of voice commands	Many	A few	Few
Influence of initial poses	Much	Much	Little

the virtual object corresponding to the real object through deep learning and coarse-to-fine gestures.

Table 3 shows the result of the preliminary pilot study. In this study, three measurements were compared: the task completion time, the number of voice commands, and the influence of the end effector's initial pose for the task completion. The proposed method took the least time to conduct the given task, required the least number of voice commands, and was not affected by the end effector's initial pose. It is important to note that previous studies were greatly influenced by the end effector's initial pose to conduct a pick operation, as shown in Figure 24.

The joint-based direct manipulation takes much time to complete the task because it is required to separately manipulate all the robot joints. While manipulating the robot through head gestures, the end effector does not move along the user intent. Thus, it is difficult to intuitively change the pose and position of the end effector to pick the targeted object. The end effector-based direct manipulation is more intuitive than the joint-based manipulation since the user can directly manipulate the robot's end effector as a pivot. However, it is difficult to estimate the robot's configuration calculated through inverse kinematics accurately. Usually, while translation is intuitive, rotation is not intuitive. Since rotation is conducted around the end effector, even if the end effector rotates a little, the robot's overall posture may change significantly, which may cause the stopping of the robot. Accordingly, when an infeasible configuration occurs, the task may take time because the robot should restart or conduct the manipulation several times. However, if the rotation is performed properly, the user can manipulate the robot intuitively.

The proposed object-based indirect manipulation enables fast and intuitive matching because there is no need to manipulate the robot directly to determine its pose or movement range regardless of the end effector's initial pose. In the case of the existing methods, since it is difficult to recognize the depth perception, it is necessary to move the robot several times, which requires many trial-and-errors. In contrast, the proposed method can perform the task quickly with a single indirect matching. However, we have found that a mismatching might occur between the physical robot and the virtual robot due to unstable MR marker tracking when the marker was partially occluded. For more accurate task execution, further research is required for the sophisticated calibration between HoloLens, robots, and physical objects. Nevertheless, the proposed approach showed the best performance.

Meanwhile, since voice commands are affected by noises and the pronunciation and intonation of the user, voice commands do not work properly sometimes. Voice commands are frequently used in the joint-based manipulation whenever the joint to be manipulated is changed and used in the end effector-based manipulation when the manipulation mode is changed. In contrast, the proposed method rarely causes voice recognition errors because voice commands are not used when performing manipulation except for capturing an image for initial object positioning. In this respect, the proposed method is more suitable for use in noisy actual workplaces.

It is also believed that the proposed method can be effectively used as an assistive technology in various situations where both hands cannot be used. For example, the proposed approach can help the elderly or handicapped interact with the robot to conduct assistive applications.

Although the case and preliminary pilot studies verified the excellence and advantage of the proposed approach, we have found that further research is still needed to conduct more objective and subjective analyses. These analyses include the evaluation of the task completion time and manipulation accuracy in more complex tasks and the user study to assess the physical and cognitive loads through questionnaires and surveys based on the National Aeronautics and Space Administration-Task Load Index (NASA-TLX) [48].

VI. CONCLUSION

This study proposed a new hands-free HRI using multimodal gestures such as eye gazing and head gestures and deep learning in MR environments. The proposed approach provides coarse-to-fine interactions that can support more effective and intuitive HRI. Coarse interaction uses eye gazing for the search and preview of objects and UIs. The preview is beneficial before making the final decision because related information is augmented around the intended target object, and the user can also check whether the result is correct with respect to his or her intent. Fine interaction can support the final selection of the object or UI and accurate 3D manipulation of the object using head gestures. In addition, the initial object positioning using deep learning-based object detection can reduce much effort on HRI using eye gazing and head gestures. Furthermore, the virtual object-based indirect manipulation can provide more intuitive and effective control of the robot using eye gazing and head gestures. Two case studies for the initial object positioning and the virtual object-based indirect manipulation were evaluated to confirm the originality and advantage of the proposed hands-free HRI. The first case study confirms that the deep learning-based object detection and 3D spatial mapping can provide more accurate and user-centric initial positioning of the object for hands-free HRI. The second case study shows that the virtual object-based indirect robot manipulation is an excellent and effective method in terms of 3D manipulation, 3D perception, and pose estimation compared to existing direct robot manipulations.

In future studies, we will apply the proposed approach to a more diverse range of industrial tasks. Different deep learning methods will be applied to HRI for assuring safety and task assistance. In addition, it is necessary to apply a new method for matching the real and virtual robots instead of using MR marker tracking. A possible approach is to apply deep learning based 3D matching or 3D pose estimation [11]. Furthermore, we will evaluate the task completion time and manipulation accuracy in more complex tasks and conduct the user study to assess the physical and cognitive loads through questionnaires and surveys based on NASA-TLX.

REFERENCES

[1] B. Matthias, S. Kock, H. Jerregard, M. Kallman, and I. Lundberg, "Safety of collaborative industrial robots: Certification possibilities for a collaborative assembly robot concept," in *Proc. IEEE Int. Symp. Assem. Manuf. (ISAM)*, May 2011, pp. 1–6.

[2] H. Liu and L. Wang, "Collision-free human-robot collaboration based on context awareness," *Robot. Comput.-Integr. Manuf.*, vol. 67, Feb. 2021, Art. no. 101997.

[3] A. Syberfeldt, O. Danielsson, and P. Gustavsson, "Augmented reality smart glasses in the smart factory: Product evaluation guidelines and review of available products," *IEEE Access*, vol. 5, pp. 9118–9130, 2017.

[4] J. Liang, H. He, and Y. Wu, "Bare-hand depth perception used in augmented reality assembly supporting," *IEEE Access*, vol. 8, pp. 1534–1541, 2020.

[5] N. Kousi, C. Stoubos, C. Gkournelos, G. Michalos, and S. Makris, "Enabling human robot interaction in flexible robotic assembly lines: An augmented reality based software suite," *Procedia CIRP*, vol. 81, pp. 1429–1434, Jun. 2019.

[6] S. M. Chacko and V. Kapila, "An augmented reality interface for human-robot interaction in unconstrained environments," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Nov. 2019, pp. 3222–3228.

[7] J. Guhl, J. Hügle, and J. Krüger, "Enabling human-robot-interaction via virtual and augmented reality in distributed control systems," *Procedia CIRP*, vol. 76, pp. 167–170, Aug. 2018.

[8] M. Kim, S. H. Choi, K.-B. Park, and J. Y. Lee, "User interactions for augmented reality smart glasses: A comparative evaluation of visual contexts and interaction gestures," *Appl. Sci.*, vol. 9, no. 15, Aug. 2019, Art. no. 3171.

[9] D. Krupke, F. Steinicke, P. Lubos, Y. Jonetzko, M. Görner, and J. Zhang, "Comparison of multimodal heading and pointing gestures for co-located mixed reality human-robot interaction," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2018, pp. 5003–5009.

[10] K.-B. Park, M. Kim, S. H. Choi, and J. Y. Lee, "Deep learning-based smart task assistance in wearable augmented reality," *Robot. Comput.-Integr. Manuf.*, vol. 63, Jun. 2020, Art. no. 101887.

[11] K.-B. Park, S. H. Choi, M. Kim, and J. Y. Lee, "Deep learning-based mobile augmented reality for task assistance using 3D spatial mapping and snapshot-based RGB-D data," *Comput. Ind. Eng.*, vol. 146, Aug. 2020, Art. no. 106585.

[12] P. Milgram and F. Kishino, "A taxonomy of mixed reality visual displays," *IEICE Trans. Inf. Syst.*, vol. 77, no. 12, pp. 1321–1329, Dec. 1994.

[13] (2020). *Mixed Reality*. [Online]. Available: https://en.wikipedia.org/wiki/Mixed_reality/

[14] (2020). *HoloLens 2*. [Online]. Available: https://www.microsoft.com/en-us/hololens/

[15] K. Huo, Y. Cao, S. H. Yoon, Z. Xu, G. Chen, and K. Ramani, "Scenariot: Spatially mapping smart things within augmented reality scenes," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, Apr. 2018, pp. 1–13.

[16] M. R. Mine, *Virtual Environment Interaction Techniques*. Chapel Hill, NC, USA: UNC Chapel Hill, 1995.

[17] A. Jackowski, M. Gebhard, and R. Thietje, "Head motion and head gesture-based robot control: A usability study," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 1, pp. 161–170, Jan. 2018.

[18] M. F. Ruzaij, S. Neubert, N. Stoll, and K. Thurow, "Auto calibrated head orientation controller for robotic-wheelchair using MEMS sensors and embedded technologies," in *Proc. IEEE Sensors Appl. Symp. (SAS)*, Apr. 2016, pp. 1–6.

[19] A. Laddi, V. Bhardwaj, N. Kapoor, D. Pankaj, and A. Kumar, "Unobtrusive head gesture based directional control system for patient mobility cart," in *Proc. Int. Conf. Signal Process., Comput. Control (ISPCC)*, Sep. 2015, pp. 236–240.

[20] H. Ohtsuka, T. Kato, K. Shibasato, and T. Kashimoto, "Non-contact head gesture maneuvering system for electric wheelchair using a depth sensor," in *Proc. 9th Int. Conf. Sens. Technol. (ICST)*, Dec. 2015, pp. 98–103.

[21] C. L. Fall, P. Turgeon, A. Campeau-Lecours, V. Maheu, M. Boukadoum, S. Roy, D. Massicotte, C. Gosselin, and B. Gosselin, "Intuitive wireless control of a robotic arm for people living with an upper body disability," in *Proc. 37th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Aug. 2015, pp. 4399–4402.

[22] M. Kyrarini, Q. Zheng, M. A. Haseeb, and A. Gräser, "Robot learning of assistive manipulation tasks by demonstration via head gesture-based interface," in *Proc. IEEE 16th Int. Conf. Rehabil. Robot. (ICORR)*, Jun. 2019, pp. 1139–1146.

[23] L. Sidenmark and H. Gellersen, "Eye&head: Synergetic eye and head movement for gaze pointing and selection," in *Proc. UIST*, Oct. 2019, pp. 1161–1174.
[24] L. Sidenmark, D. Mardanbegi, A. R. Gomez, C. Clarke, and H. Gellersen, "BimodalGaze: Seamlessly refined pointing with gaze and filtered gestural head movement," in *Proc. ETRA*, Jun. 2020, pp. 1–9.
[25] K. Pfeuffer, J. Alexander, M. K. Chong, Y. Zhang, and H. Gellersen, "Gaze-shifting: Direct-indirect input with pen and touch modulated by gaze," in *Proc. UIST*, Nov. 2015, pp. 373–383.
[26] C. H. Morimoto and M. R. M. Mimica, "Eye gaze tracking techniques for interactive applications," *Comput. Vis. Image Understand.*, vol. 98, no. 1, pp. 4–24, Apr. 2005.
[27] S. Stellmach and R. Dachselt, "Look & touch: Gaze-supported target acquisition," in *Proc. CHI*, 2012, pp. 2981–2990.
[28] J. Turner, A. Bulling, J. Alexander, and H. Gellersen, "Cross-device gaze-supported point-to-point content transfer," in *Proc. ETRA*, Mar. 2014, pp. 19–26.
[29] E. G. Freedman, "Coordination of the eyes and head during visual orienting," *Exp. Brain Res.*, vol. 190, no. 4, pp. 369–387, Aug. 2008.
[30] D. Tweed, B. Glenn, and T. Vilis, "Eye-head coordination during large gaze shifts," *J. Neurophysiol.*, vol. 73, no. 2, pp. 766–779, Feb. 1995.
[31] Z. Zhu and Q. Ji, "Novel eye gaze tracking techniques under natural head movement," *IEEE Trans. Biomed. Eng.*, vol. 54, no. 12, pp. 2246–2260, Dec. 2007.
[32] M. Khamis, C. Oechsner, F. Alt, and A. Bulling, "VRpursuits: Interaction in virtual reality using smooth pursuit eye movements," in *Proc. AVI*, May 2018, pp. 1–8.
[33] P. Mohan, W. B. Goh, C.-W. Fu, and S.-K. Yeung, "DualGaze: Addressing the midas touch problem in gaze mediated VR interaction," in *Proc. ISMAR-Adjunct*, Oct. 2018, pp. 79–84.
[34] D. Mardanbegi, T. Langlotz, and H. Gellersen, "Resolving target ambiguity in 3D gaze interaction through VOR depth estimation," in *Proc. CHI*, May 2019, pp. 1–12.
[35] T. Nukarinen, J. Kangas, O. Špakov, P. Isokoski, D. Akkil, J. Rantala, and R. Raisamo, "Evaluation of HeadTurn: An interaction technique using the gaze and head turns," in *Proc. NordiCHI*, Oct. 2016, pp. 1–8.
[36] O. Špakov, P. Isokoski, and P. Majaranta, "Look and lean: Accurate head-assisted eye pointing," in *Proc. ETRA*, Mar. 2014, pp. 35–42.
[37] D. Q. Huy, I. Vietcheslav, and G. S. G. Lee, "See-through and spatial augmented reality—A novel framework for human-robot interaction," in *Proc. ICCAR*, Apr. 2017, pp. 719–726.
[38] R. T. Chadalavada, H. Andreasson, M. Schindler, R. Palm, and A. J. Lilienthal, "Bi-directional navigation intent communication using spatial augmented reality and eye-tracking glasses for improved safety in human–robot interaction," *Robot. Comput.-Integr. Manuf.*, vol. 61, Feb. 2020, Art. no. 101830.
[39] M. Mohammed, H. Jeong, and J. Y. Lee, "Human-robot collision avoidance scheme for industrial settings based on injury classification," in *Proc. ACM/IEEE HRI*, Mar. 2021, pp. 549–551.
[40] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.
[41] (2020). *Unity3D*. [Online]. Available: https://unity.com/
[42] (2020). *Universal Robots*. [Online]. Available: https://www.universal-robots.com/
[43] (2017). *COCO Dataset*. [Online]. Available: https://cocodataset.org/#home/
[44] G. Jiang, M. Luo, K. Bai, and S. Chen, "A precise positioning method for a puncture robot based on a PSO-optimized BP neural network algorithm," *Appl. Sci.*, vol. 7, no. 10, Sep. 2017, Art. no. 969.
[45] P. M. Kebria, S. Al-Wais, H. Abdi, and S. Nahavandi, "Kinematic and dynamic modelling of UR5 manipulator," in *Proc. SMC*, Oct. 2016, pp. 4229–4234.
[46] M. Grieves and J. Vickers, "Digital twin: Mitigating unpredictable, undesirable emergent behavior in complex systems," in *Transdisciplinary Perspectives on Complex Systems*. Cham, Switzerland: Springer, Aug. 2017, pp. 85–113.
[47] M. Zhu, "Recall, precision and average precision," Dept. Statist. Actuarial Sci., Univ. Waterloo, Waterloo, ON, Canada, Working Paper 2004-09, 2004.
[48] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (task load index): Results of empirical and theoretical research," *Adv. Psychol.*, vol. 52, pp. 139–183, Apr. 1988.

KYEONG-BEOM PARK received the B.S. and M.S. degrees in industrial engineering from Chonnam National University, South Korea. He is currently pursuing the Ph.D. degree with Chonnam National University. His current research interests include AR/MR and deep learning-based applications.

SUNG HO CHOI received the B.S. and M.S. degrees in industrial engineering from Chonnam National University, South Korea. He is currently pursuing the Ph.D. degree with Chonnam National University. His current research interests include AR-based remote collaboration and human–robot collaboration.

JAE YEOL LEE (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in industrial engineering from the Pohang University of Science and Technology (POSTECH), South Korea, in 1992, 1994, and 1998, respectively. From 1998 to 2003, he worked as a Senior Researcher with the Electronics and Telecommunications Research Institute (ETRI), South Korea. Since 2003, he has been a Faculty Member of Chonnam National University, South Korea. He is currently a Professor with the Department of Industrial Engineering, Chonnam National University. His current research interests include AR/MR, deep learning, human–robot collaboration, and UX.

YALDA GHASEMI received the B.S. degree in industrial engineering from Shomal University, Iran. She is currently pursuing the Ph.D. degree in industrial engineering and operations research with the University of Illinois at Chicago. Her current research interests include augmented reality, human–computer interaction, and human performance modeling.

MUSTAFA MOHAMMED is currently pursuing the B.S. degree in neuroscience with the University of Illinois at Chicago. His primary research interests include pertain to brain–computer interfaces and computer vision, and how they can be combined to improve rehabilitative outcomes for robotic exoskeletons.

HEEJIN JEONG (Member, IEEE) received the B.S. degree in industrial engineering from the Pohang University of Science and Technology, South Korea, and the M.S.E. and Ph.D. degrees in industrial and operations engineering from the University of Michigan, Ann Arbor, in 2015 and 2018, respectively. He is currently an Assistant Professor with the Department of Mechanical and Industrial Engineering, University of Illinois at Chicago. His current research interests include human factors engineering, cognitive ergonomics, and human performance modeling.