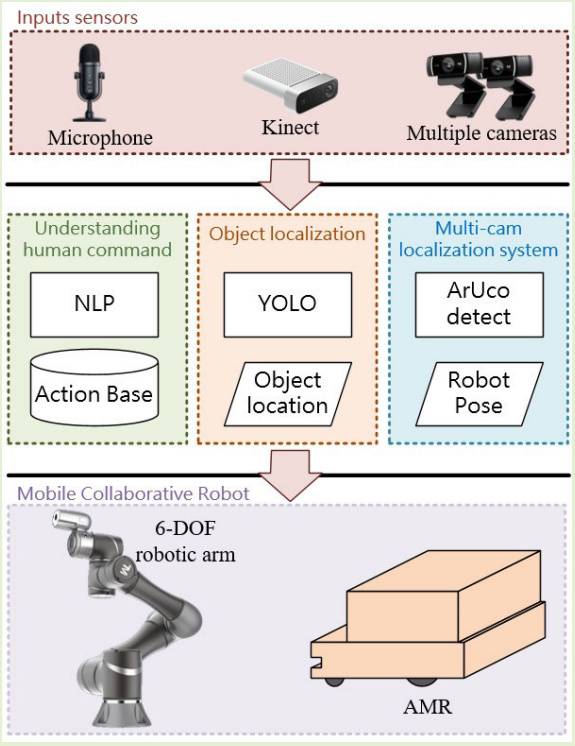
Vision-Based Mobile Collaborative Robot Incorporating a Multicamera Localization System

Chen-Chien James Hsu, *Senior Member, IEEE*, Pin-Jui Hwang, Wei-Yen Wang, *Fellow, IEEE*, Yin-Tien Wang, *Member, IEEE*, and Cheng-Kai Lu, *Senior Member, IEEE*

***Abstract*—As the Industry 4.0 landscape unfolds, collaborative robots (cobots) play an important role in intelligent manufacturing. Compared with industrial robots, cobots are more flexible and intuitive in programming, especially for industrial and home service applications; however, there are still issues to be solved, including understanding human intention in a natural way, adaptability to execute tasks, and robot mobility in a working environment. As an attempt to solve the problems aforementioned, in this article, we propose a modularized solution for mobile cobot systems, where the cobot equipped with a multicamera localization scheme for self-localization can understand the human intention via human voice commands to execute the tasks in an unseen scenario in a small-area working environment. As far as intention understanding is concerned, we devise a natural language processing approach to establish an action base to describe human commands. According to the action base, the robot can then execute the tasks by planning a trajectory with the help of an object localization module, which integrates the point cloud and the object detected by YOLOv4 to locate the object's position in 3-D space. Depending on where the cobot interacts with the object, the cobot might need to navigate around the working environment. Thus, we also establish a low-cost and high-efficiency multicamera localization system with ArUco markers to locate the mobile cobot in a larger sensing area. The experimental results show that the proposed vision-based mobile cobot can successfully interact with a human operator to assemble a wooden chair in a small workshop.**



***Index Terms*— Collaborative robot (cobot), mobile cobot, multicamera localization, natural language processing (NLP).**

Manuscript received 10 January 2023; revised 20 July 2023; accepted 26 July 2023. Date of publication 7 August 2023; date of current version 14 September 2023. This work was supported in part by the "Chinese Language and Technology Center" of National Taiwan Normal University (NTNU) from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan; and in part by the Ministry of Science and Technology, Taiwan, under the Program of AI Thematic Research Program to Cope with National Grand Challenges through Pervasive Artificial Intelligence Research (PAIR) Laboratories of the National Yang Ming Chiao Tung University, under Grant MOST 110-2221-E-003-019, Grant MOST 110-2221-E-003-020-MY2, and Grant MOST 110-2634-F-A49-004. The associate editor coordinating the review of this article and approving it for publication was Prof. Kazuaki Sawada. *(Corresponding author: Cheng-Kai Lu.)*

Please see the Acknowledgment section of this article for the author affiliations.

Digital Object Identifier 10.1109/JSEN.2023.3300301

I. Introduction

As THE Industry 4.0 landscape unfolds, collaborative robots (cobots) play a critical role in intelligent manufacturing [1], [2], [3], [4]. Compared with industrial robots, cobots are generally smaller and more nimble, enabling more flexible and intuitive programming with a teaching panel or hand guiding. As a result, better use of space, reduced downtime between tasks, and improved performance for various applications can be obtained. Now, cobots are working hand in hand with humans by balancing flexibility and productivity, taking manufacturing operations to the next level, particularly for small- and medium-sized enterprises (SMEs) [5]. With the advance of technology in recent years, cobots have already

gone beyond manufacturing environments into more complex and unstructured human working environments.

Although cobots are promising for many industrial and home service applications, there are still issues to be solved [6], including understanding human intention, adaptability to execute tasks, and robot mobility. The human intention is generally transferred to robots through haptic sensing, wearable sensors, vision, and speech. Haptic communication is the most stable way to interact with robots, in which hand guiding is one of the most intuitive approaches [7], [8], where the operator physically drags the robot around its workspace to teach points and paths. Nevertheless, the critical problem of the hand-guiding approach is that the robot cannot reproduce the task when the environment changes. Wearable sensors like robotic gloves [9], [10] and active markers [such as inertial measurement unit (IMU)] [11] are popular in the areas such as human–robot interaction (HRI) and robot learning from demonstration (LfD) because these sensors have better accuracy in locating operator's hands and fingers. However, high-performance wearable sensors are expensive and cannot economically apply to business sectors industry. Vision sensors, on the other hand, are getting increasingly popular because of their perceived ability through image processing, allowing operators to interact with the robot through gestures, actions, facial expressions, etc. However, it is still challenging to detect meticulous and high-speed movements on hands and fingers [12], [13], [14] for HRI. Unlike the social HRI [15], which focuses on social services, health care, and entertainment, the cobots for use in the industry do not have to deal with ambiguous conversations. As a result, speech [16], [17] as a natural communication mean is suitable for humans to provide explicit information about their intentions. After understanding the human's intention, the robot needs to plan a motion to execute the task. For cobots typically programed according to the perception by haptic sensing, wearable sensors, vision, and speech, particularly under the hand-guiding framework, it will be difficult for the robots to execute the task in an unseen scenario when the environment changes. As a popular solution for perceiving the environment, vision-based approaches are widely used in manufacturing industries [15], [16], making a big difference in production lines. However, there is still room for further improvement in vision-based perception to handle high-level tasks such as cognitively demanding activities in modern manufacturing. Another issue with most existing cobots is the ability to navigate. Although autonomous mobile robots (AMRs) widely used in modern manufacturing can be used as a mobile platform to combine with a robot arm to build a mobile cobot, the navigation usually relies on lidar-based localization [18], [19]. In general, the lidar-based localization framework is robust in unchanging environments. However, the framework needs to have an environment map before the robot can locate itself. Additionally, lidar-based localization requires expensive hardware and high-performance computation power. On the contrary, marker-based localization frameworks are usually used in small-area indoor localization [20], [21], [22], suitable for human–robot collaboration. If makers are allowed in the working environment, the marker-based localization framework has many advantages, such as good efficiency, accuracy, and low cost.

As an attempt to solve the problems mentioned above, we propose a modularized solution for mobile cobots systems, where the cobot equipped with a multicamera localization scheme for self-localization can understand the human intention in a natural way via voice commands to execute the tasks instructed by the human operator in an unseen scenario in a small-area working environment. As far as understanding human intention is concerned, we devise a voice-to-command approach, where the command text is first decomposed into subcommands by CKIP transformers air position indicator (API) for further analysis to recognize the action, interacted object, and destination in sequence to establish an action base to describe the human commands. According to the action base, the robot can then execute the tasks by planning a trajectory with the help of an object localization module, which integrates the point cloud and the object detected by YOLOv4 to locate the object's position in 3-D space. Depending on where the cobot interacts with the object, the cobot might need to navigate around the working environment. Thus, we also establish a low-cost and high-efficiency multicamera localization system with an easy extension to cover a larger sensing area through ArUco markers and multiple cameras to locate the mobile cobot. Existing multicamera systems generally focused on improving the accuracy of the marker location [23]. The advantage of the proposed multicamera localization system, on the other hand, lies in the capability to extend the sensing range of the localization system by deploying multiple cameras. As a result, a larger area for localization can be obtained by the easy-extension multicamera localization system.

To this end, the contributions of the proposed work are summarized as follows.

1) A modularized solution for a mobile cobot system is proposed, including voice understanding, robot localization, object localization, and HRI. Because of the modularized design, the system is easy to maintain and update if there is a better solution.
2) Based on a voice-to-command approach, the human intention expressed in a compound command context can be described by a human-readable action base, making human operators collaborate with the cobot more intuitively.
3) According to the descriptions in the action base, the cobot can interact with humans to execute tasks in an unseen scenario, greatly improving the robot's adaptability when the environment changes.
4) A novel, easy-extension multicamera localization system is proposed to locate the cobot in a 3-D space. In comparison with lidar-based localization, the proposed approach does not require predefined maps and can be executed with high efficiency at a low cost.

II. System Architecture

Fig. 1 illustrates an experimental environment in which the proposed mobile cobot works. Major components of the mobile cobot system include a Kinect RGB-D camera,

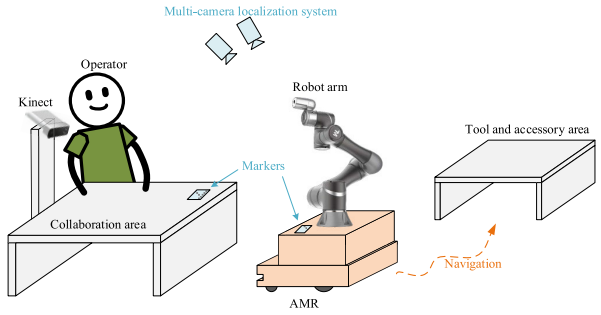


Fig. 1. Schematic illustration showing an experimental environment of the proposed mobile cobot.

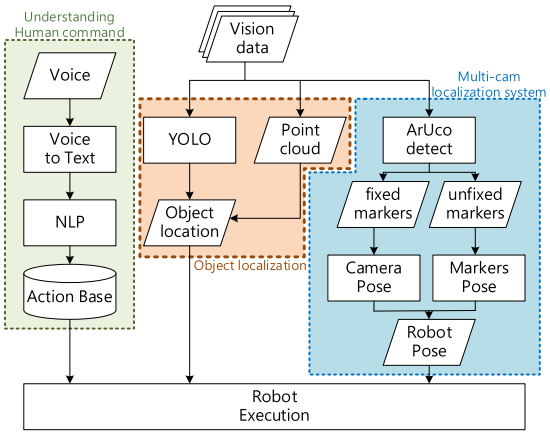


Fig. 2. Flowchart of the proposed mobile cobot.

multiple RGB cameras, and a mobile robot platform (AMR) with a 6-DOF robot arm. The flowchart of the proposed mobile cobot system is shown in Fig. 2, which includes three major modules: "understanding human commands," "object localization," and "multicam localization." The "understanding human commands" module, the voice commands will be first converted into text by Google API, followed by a proposed natural language processing (NLP) framework to create an action base describing the human intention for a given task. The action base includes all the information that the robot needs to know for execution to complete the tasks, including action, interacted object, and destination. As soon as the intention of the human is known, the robot is ready to execute the tasks according to the action base, where the interacted object is recognized through YOLOv4 to locate the object location in three dimensions by incorporating point cloud information. If the robot needs to reach the tool and accessory area, the multicamera localization system is used to locate the mobile robot for navigation to the destination through ArUco markers and multiple cameras.

The entire cobot system is based on the robot operating system (ROS), which makes the system extremely modularized, greatly helping to develop a larger system with complex interactions for every module. Fig. 3 shows the ROS node architecture of the proposed mobile cobot, including three major groups.

1) *Inputs*: Including RGB images, point cloud, and voice commands via a microphone.

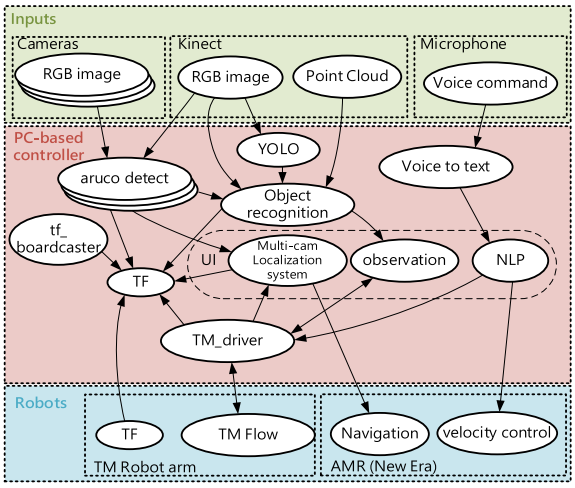


Fig. 3. ROS node architecture of the proposed mobile cobot.

2) *PC-Based Controller*: Most calculations, including understanding human commands, object localization, and multicam localization, are performed in the controller.
3) *Robots*: The robot arm and AMR platform have their respective controller for some basic motor control.

III. Understanding of Human Commands

Voice is the most natural way for humans to communicate with each other. In the proposed robot system, the robot can understand human intentions by voice commands which are subsequently analyzed and saved into an action base described in three elements: Action, Interacted object, and Destination, based on which the robot executes a given task. Fig. 4 illustrates the flowchart showing the understanding of human commands. When the voice commands are given, the voice will be first converted into command text by using the Google Speech API for further analysis to recognize Destination, Action, and Interacted objects in sequence to establish the action base according to the following steps.

Step 1: The command text is separated into several subcommands by CKIP transformers API [24], decomposing compound commands into simple subcommands. (If the command text is clear enough for analysis, the result of this step will be the same as the origin command.)

Step 2: Search for a destination from the subcommand text according to a destination dictionary which has all the destinations in the environment, including specific places, objects, and the operator's hand.

Step 3: Depending on the destination keyword found, set "Destination" in the action base by the destination keyword found. The destination keyword that is found will be removed from the subcommand text to prevent confusion in recognizing the "Interacted object" in the next steps.

Step 4: Search for an action from the subcommand text according to an action dictionary. Set "Action" in the action base by the action keyword found. If no action keyword is found in the subcommand text, set "Action" in the action base as "Place."

Step 5: Search for an interacted object from the subcommand text according to an integrated object dictionary. Set

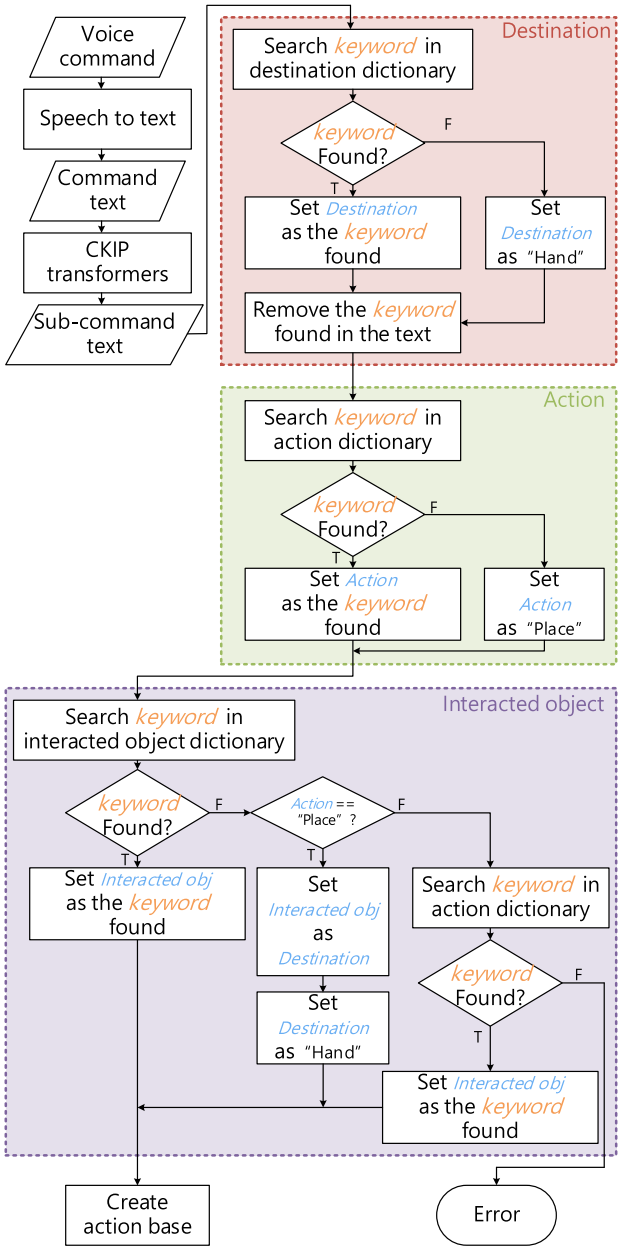


Fig. 4. Flowchart showing the understanding of human commands.

"Interacted object" in the action base as the keyword found. If no keyword is found in the subcommand, check the "Action" in the action base. If "Action" in the action base is "Place," then set "Interact object" in the action base as the content in "Destination" and set "Destination" as "Hand" in the action base. This design choice is made because, by default, the keywords in the Destination dictionary represent the "destination" of the action. Therefore, if the keywords are not found in the Destination dictionary, the appropriate "Action" to be taken is to naturally "Place" the objects on the "Hand" rather than placing the objects on a specific destination (e.g., desktop). This approach ensures that the system can provide a meaningful response even when specific keywords are not recognized in the Destination dictionary. If "Action" is not "Place," then the "Interacted object" will be set as the object associated with the keyword found in the action dictionary.

TABLE I
ILLUSTRATIVE EXAMPLE OF AN ACTION BASE

Action	Interacted Object	Destination
Place	Screw box	Hand
Sand	Sandpaper	Wooden board
Hold	Wooden board	---

TABLE II
ILLUSTRATIVE EXAMPLE OF AN ACTION BASE, WHICH IS DECOMPOSED FROM A COMPOUND COMMAND

Action	Interacted Object	Destination
Glue	Glue paste	Chair leg
Place	Chair leg	Hand

Note that "Interacted object" can sometimes be misrecognized as a "Destination" because the keyword can exist in both the interacted object dictionary and destination dictionary. Hence, we have to check the "Action" in the action base to determine the object associated with the action for the "Interacted object."

According to the steps mentioned above, an action base describing the human intention, including Action, Interacted object, and Destination, can be established. Following the descriptions of the action base, the robot can execute a given task from the human operator. As an illustrative example, Table I shows an action base, indicating the tasks to place a screw in the operator's hand, use sandpaper to sand a wooden board, and hold the wooden board for the operator in sequence. If a compound command is received, for example, "Give me a glued chair leg," it will be decomposed into subcommand as "Glue the chair leg" and "Give me the chair leg," as shown in Table II by CKIP transformers. Note that the action base is human readable to help the operator to double check or modify the details, if necessary. Furthermore, the action base only indicates the human intention, which is analyzed from the voice commands, meaning that the action base includes the information of "which object" but does not include "where the object is." The object's position will be detected in the object localization module for the robot to execute.

IV. EXECUTION OF MOBILE COBOT

To have the cobot successfully execute a given task described in the action base, we first need to locate the object. Because navigation is required in the working environment, we also establish a low-cost and efficient localization system for the proposed cobot system.

A. Object Localization

To execute a given task, the robot needs the position of the interacted object, which has been described in the action base. As a deep learning network, YOLO has been proven to be an open source, easy to train and use, stable, and efficient object detection for several years. In this article, we thus use YOLOv4 for object detection in the experimental environment. After the objects are detected, we integrate the point cloud provided by the Kinect camera to locate the object's position in 3-D space. The Kinect camera in this module is also used in the multicamera localization system.

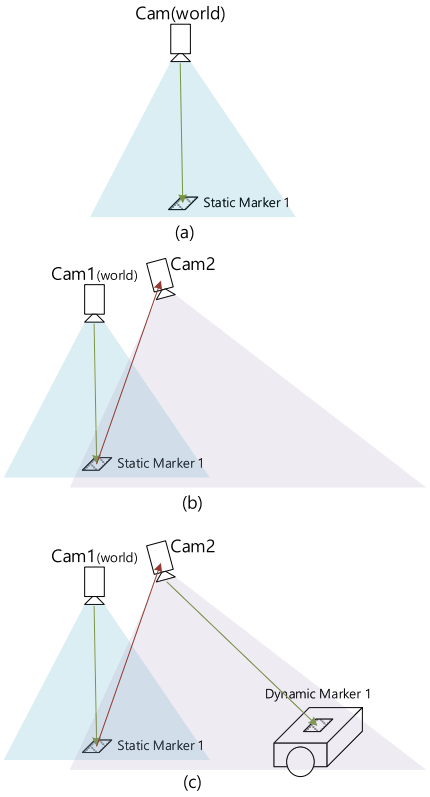


Fig. 5. Illustrative example showing the proposed multicamera localization system: (a) step 1: define a camera as a world space camera; (b) step 2: use the pose of the markers to back-calculate/derive the pose of the other cameras; and (c) step 3: locate the robot from the dynamic markers in the camera's view.

B. Mobile Robot Navigation Incorporating a Multicamera Localization System

To locate the robot, we propose a novel localization system consisting of multiple fixed-position cameras, static (fixed-position) ArUco markers, and dynamic (unfixed-position) ArUco markers attached to the AMR. Localization of the robot includes the following four steps.

Step 1: Select a camera and define it as a world space camera [Cam1 in Fig. 5(a)]. Then, locate the static markers [static marker 1 in Fig. 5(a)] in the camera's view to obtain the pose of the static markers. Notice that we do not know the poses of other cameras [Cam2 in Fig. 5(b)] now.

Step 2: Search all the other cameras' views for the static markers located in Step 1. Use the pose of the markers to back-calculate via rotation and translation to derive the pose of the other cameras [e.g., Cam2 in Fig. 5(b)].

Step 3: Locate static and dynamic markers [dynamic marker 1 in Fig. 5(c)] in the camera's view in Step 2.

Step 4: Repeat Step 2 and Step 3 for cameras deployed in the environment to extend the sensing area to locate the dynamic markers on the robot. Notice that we do not use the dynamic markers to back-calculate/derive the camera's pose because this might cause errors if the dynamic markers move.

The advantage of this system is that preparation before the localization is easy. The poses of all cameras and markers do not need to be predefined. Unlike other map-based localization methods, the proposed system does not need to create an environmental map. Therefore, the proposed system is suitable for small-area locations such as a workshop or working area for picking tasks, where cameras might be readily available for other usage. Another advantage of the system is that the processing demand is very low, so the system only needs to calculate the markers' and cameras' pose. Compared with lidar-based localization methods, the calculations required in the proposed system are almost negligible. We would like to emphasize that the advantage of the proposed multicamera localization system lies in the capability to extend the sensing range of the localization system by deploying multiple cameras. Localization accuracy is not the focus of the proposed method.

From the navigation system point of view, the navigation system is not the focus of our system. For the robot to reach a destination, we simply predefine the desired path, so that the robot follows the path according to the robot pose from the multicamera localization system. In this article, we did not propose a new path-planning algorithm. However, any path-planning algorithm, for example, the A$*$ algorithm, can be easily incorporated into the navigation system because the overall system is well-modularized.

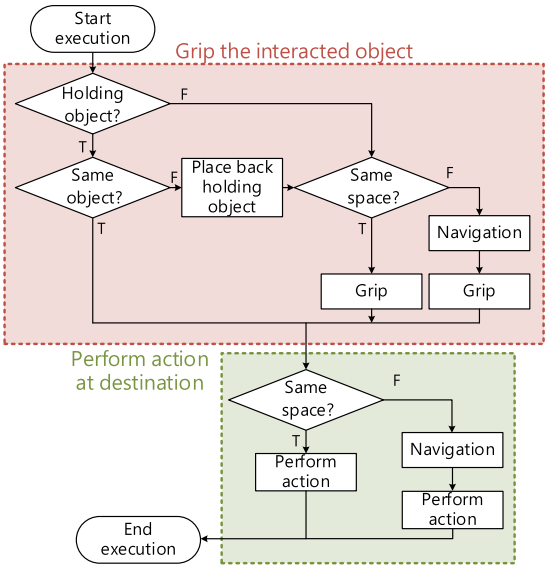


Fig. 6. Flowchart of robot execution.

C. Robot Execution/Human–Robot Collaboration

To execute the tasks provided in the action base, the robot acts according to the three modules of the proposed mobile cobot in Fig. 2. Based on the descriptions of the action base, object location and robot position can be obtained via the object localization and multicam localization modules in Fig. 2. A navigation path and a trajectory can be subsequently planned to accomplish the given task described in the action base.

For example, a human operator might want to perform the tasks such as "Place a specific object in a specific place (including operator's hand)" and "Sand a specific object" via the understanding human command module. Fig. 6 shows the flowchart of robot execution according to the tasks described in the action base. The execution flow includes two parts, where the robot first grips the interacted object and then

TABLE III
EXAMPLE OF AN ACTION BASE TO ILLUSTRATE THE ROBOT EXECUTION

Action	Interacted Object	Destination
Place	Screw box	Hand

performs the action at the destination. To grip the interacted object, the robot needs to check whether the gripper is holding any object or not. If the robot is holding an object, it will place back the holding object. Then, the robot will grip the object if it is located in the same space as the robot. If the robot cannot reach the object, the robot will need to navigate to the space to grip the object. After successfully gripping the interacted object, the robot will navigate to the destination if needed to act.

Take the action base shown in Table III as an example; the robot arm will grip the screw box using information from the object localization module. If the robot cannot reach the object, the robot will navigate to the location near the object to grip the object. Then, the robot will place the object at the destination, which is the operator's hand. The robot in the step "placing to operator's hand" will wait until the operator confirms the action "placing" with a pressing force on the robot arm to ensure the robot gripper will not release before the operator is ready.

V. EXPERIMENTAL RESULTS

To verify the proposed mobile collaboration system, several experiments are conducted. We set up two example tasks to assemble a wooden chair. In the first example task, we use a 6-DOF fixed robot arm (UR3) and Kinect camera to collaborate with the operator. In the second example, we use a TM-5 robot arm with an AMR platform to provide mobility for the cobot to work with the human operator in different working spaces. Under this circumstance, the proposed localization system will be used in example task 2.

A. Example Task 1: Human Collaborated With a Fixed Robot Arm to Assemble a Wooden Chair

This example shows the collaboration between a human and a fixed robot arm. In Fig. 7, the human tries to assemble a wooded chair with the help of the robot, in which tasks executed by the robot appearing in sequence are: (a) "pick up the screw and place it to the operator's hand"; (b) "sanding the chair seat"; (c) "give the glued chair leg to the operator"; (d) "holding the chair seat"; and (e) "taking back the screwdriver." Interested readers can refer to the following link: https://youtu.be/1KIVQKKezSA to view the full video clip.

B. Multicamera Localization System

Fig. 8 shows the experimental configuration of the localization system to validate the proposed localization system. Cameras 1 and 2 (Cam1 and Cam2) all have a resolution of 1920 × 1080, and the markers have a size of 10 × 10 cm. In Fig. 8, Cam1 is defined as a world space camera, which gives the pose of Marker1, based on which the pose of Cam2

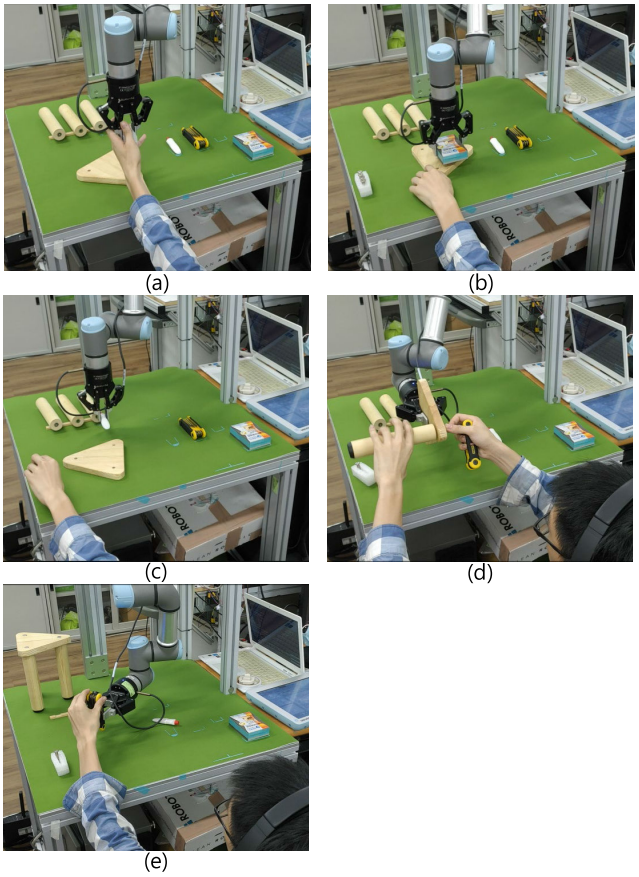


Fig. 7. Example task 1: human and fixed robot arm collaborated to assemble a wooden chair: (a) cobot executes: "pick up the screw and place it in the operator's hand"; (b) cobot executes: "sand the position of the hand"; (c) cobot executes: "give the glued chair leg to the operator's hand"; (d) cobot executes: "hold the chair seat at the position of the hand"; and (e) cobot executes: "take back the screwdriver.".

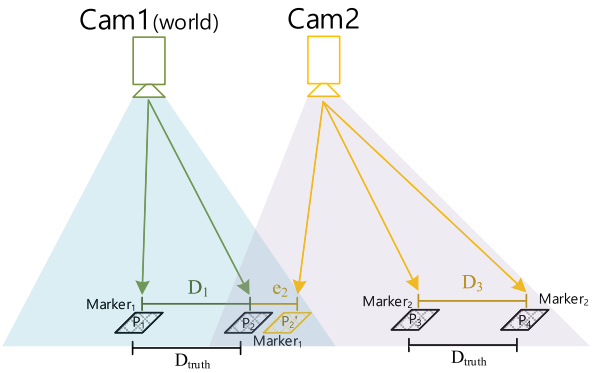


Fig. 8. Experimental environment for the multicamera localization system.

can be back-calculated. Moving Marker1 from positions P_1 to P_2, we can calculate the measured distance $D_1 = |P_1 - P_2|$ against the ground truth D_{truth}. We define $e_1 = D_1 - D_{\text{truth}}$ as the measurement error of Cam1. By analogy, $D_3 = |P_3 - P_4|$ and $e_3 = D_3 - D_{\text{truth}}$ if we move Marker2 from positions P_3 to P_4 for Cam2. P_2' is the position for Marker 1 derived from Cam2. Therefore, $e_2 = |P_2 - P_2'|$, indicating the localization error for Marker1 at position 2 due to measurement from Cam1 and Cam2 can be regarded as the error of Cam2.

TABLE IV
MEASUREMENT ERRORS OF THE MULTICAMERA LOCALIZATION SYSTEM

Distance between camera and marker	e_1 (m)	e_2 (m)	e_3 (m)
1 meter	0.0056	0.0099	0.0052
2 meters	0.0104	0.0478	0.0081

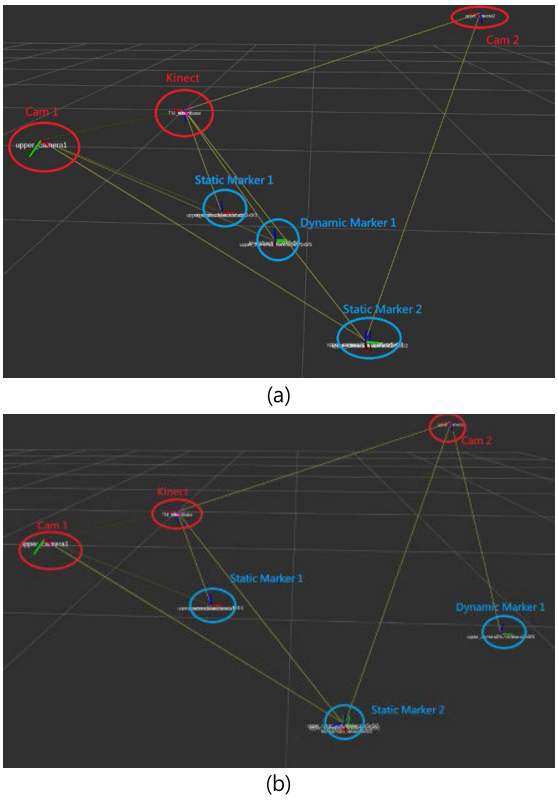


Fig. 9. Cameras and markers pose in the multicamera localization system: (a) cameras and markers pose and (b) cameras and markers pose when the robot moves.

Fig. 10. Example task 2: human operator and mobile cobot collaborated in a small workshop: (a) cobot starts to navigate to the tool area; (b) cobot navigates to the tool area; (c) cobot picks up a glue; (d) cobot gives the glue to the operator; (e) cobot picks up the screw box; (f) cobot gives the screw box to the operator; (g) cobot holds the chair seat; (h) cobot sands the chair seat; (i) cobot vacuums the place that the operator pointed to; and (j) cobot vacuums the place that the operator pointed.

To validate the performance of the proposed localization system, we establish two environments where distances between the camera and marker are 1 and 2 m, respectively. Table IV shows the measurement errors e_1, e_2, and e_3. If the distance between the camera and the marker is farther, the markers will become more blurred/unclear in the camera's view, causing higher measurement errors. To decrease measurement errors, we can use bigger markers or higher resolution cameras.

These three measurement errors contribute to the localization accuracy of the AMR. For example, if Marker 2 is placed on the target robot, the localization error is $e_1 + e_2 + e_3$. Furthermore, we can find out that e_1 is approximately equal to e_3 because the definition of e_1 and e_3 is the same except for different cameras. Suppose there is a Cam3, whose pose is back-calculated from Marker 2, and a Marker 3 is placed on the target robot. In that case, the pose of the robot can be derived from Cam3, resulting in a localization error of the target robot as $e_1 + e_2 + e_3 + e_4 + e_5$, where e_4 is the error of Cam3 and e_5 is the measurement error of Cam3. To summarize, the result of the multicamera localization system in a small area is promising for practical applications. As shown in Table IV, if the markers are 1 m far from the cameras, the localization error e_1 is smaller than 6 mm, and localization error $e_1 + e_2 + e_3$ is around 2 cm. To better illustrate the localization system, we practically build an environment with three cameras: Kinect camera, web camera Cam1, and web camera Cam2. Fig. 9(a) shows the poses of the cameras and markers, where the cameras and markers are indicated by red circles and blue circles, respectively. Note that a dynamic Marker 1 is attached to the mobile robot. When the robot moves, the system will update the pose of the dynamic marker [Fig. 9(b)].

To demonstrate the performance of the proposed method, accuracy, the number of cameras used, and vision field expandability are chosen as the evaluation indices for the comparison. Our proposed method exhibits superior accuracy compared to the state-of-the-art technique discussed in [21]. While Roos-Hoefgeest et al. [21] show an error range of approximately 1–5 cm, our method achieves an impressive accuracy of only 1 cm within a 1-m distance. Additionally, our proposed method

offers the advantage of a broader vision field by utilizing additional cameras through the use of ArUco markers. In contrast, Szalóki et al. [23] achieve the highest level of accuracy with an impressive 1-mm error. However, it requires the use of three cameras to capture a single quick response (QR) code, which incurs comparatively higher costs and restricts the vision field. This highlights a significant advantage of our proposed method, as it provides a cost-effective solution with a broader vision field, ensuring a more practical and efficient approach to marker-based localization.

C. Example Task 2: Human and Mobile Robot Collaborated to Assemble a Wooden Chair

In this example, the human tries to assemble a wooden chair with the help of a cobot in a small-area workshop. Because there is the distance from the tool and accessory area, mobility of the cobot is required. Fig. 10 shows the interaction between a human operator and mobile robot to collaborate to execute a series of commands to assemble a wooden chair, where (a) and (b) show the cobot navigates to the tool area; (c) picks up a glue; (d) gives the glue to the operator; (e) picks up the screw box on the robot platform; (f) gives the screw box to the operator; (g) holds the chair seat to help the operator assemble the chair leg; (h) sand the chair seat; and (i) and (j) vacuum the place that operator pointed to. Interested readers can refer to the following link: https://youtu.be/qMW5ihQ9grQ to view the full video clip.

VI. Conclusion

In this article, we give a modularized solution of a vision-based mobile human–robot collaboration system for assembly tasks in a workshop. The system includes three major modules: voice command understanding, object localization, and a multicamera localization. Each module is well-modularized, which can be switched or upgraded. In the voice command understanding module, we analyze the voice command to action base through our method. The action base describes the intention of the human command in three columns: Action, Interacted object, and Destination. The action base is human readable, making the operator easy to confirm or modify before the robot execution. The robot will execute the task according to the action base, and the object in the action base will locate from the object location module, which can recognize and locate the object with YOLOv4 and point cloud. If the robot cannot reach the object, the robot will navigate to the area next to the object. The navigation path and robot localization are provided by a multicamera localization module. This module comprises a few cameras and several ArUco makers. We find new camera pose and maker pose through defined camera and makers. Therefore, this makes the system easy to expand by adding a new camera, which views any defined markers in the system. To verify the system, we estimate the robot with an assembling task. The result shows that the system can understand the human voice command and collaborate with humans to assemble the wooden chair in a workshop. Although the system can execute the task accurately by human command, the system still has some limitations. First, if there is any new object, we need to retrain the learning model. Second, the positions of the objects placed and the navigation path of the mobile robot have to be in the field of view of the equipped cameras.

Acknowledgment

Chen-Chien James Hsu, Pin-Jui Hwang, Wei-Yen Wang, and Cheng-Kai Lu are with the Department of Electrical Engineering, National Taiwan Normal University, Taipei 10610, Taiwan (e-mail: Cklu@ntnu.edu.tw).

Yin-Tien Wang is with the Department of Mechanical and Electro-Mechanical Engineering, Tamkang University, New Taipei City 251301, Taiwan.

References

[1] N. Berx, L. Pintelon, and W. Decré, "Psychosocial impact of collaborating with an autonomous mobile robot: Results of an exploratory case study," in *Proc. Companion ACM/IEEE Int. Conf. Hum.-Robot Interact.*, Mar. 2021, pp. 280–282.

[2] A. Bonci, P. D. C. Cheng, M. Indri, G. Nabissi, and F. Sibona, "Human-robot perception in industrial environments: A survey," *Sensors*, vol. 21, no. 5, p. 1571, Feb. 2021.

[3] L. Liu et al., "Application, development and future opportunities of collaborative robots (cobots) in manufacturing: A literature review," *Int. J. Hum.-Comput. Interact.*, pp. 1–18, 2022.

[4] P.-J. Hwang, C.-C. Hsu, P.-Y. Chou, W.-Y. Wang, and C.-H. Lin, "Vision-based learning from demonstration system for robot arms," *Sensors*, vol. 22, no. 7, p. 2678, Mar. 2022.

[5] P.-J. Hwang, C.-C. Hsu, and W.-Y. Wang, "Development of a mimic robot—Learning from demonstration incorporating object detection and multiaction recognition," *IEEE Consum. Electron. Mag.*, vol. 9, no. 3, pp. 79–87, May 2020.

[6] A. Bauer, D. Wollherr, and M. Buss, "Human–robot collaboration: A survey," *Int. J. Hum. Robot.*, vol. 5, no. 1, pp. 47–66, 2008.

[7] S.-D. Lee, K.-H. Ahn, and J.-B. Song, "Torque control based sensorless hand guiding for direct robot teaching," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2016, pp. 745–750.

[8] S. Zhang, S. Wang, F. Jing, and M. Tan, "A sensorless hand guiding scheme based on model identification and control for industrial robot," *IEEE Trans. Ind. Informat.*, vol. 15, no. 9, pp. 5204–5213, Sep. 2019.

[9] Y. Park, J. Lee, and J. Bae, "Development of a wearable sensing glove for measuring the motion of fingers using linear potentiometers and flexible wires," *IEEE Trans. Ind. Informat.*, vol. 11, no. 1, pp. 198–206, Feb. 2015.

[10] E. Fujiwara, D. Y. Miyatake, M. F. M. D. Santos, and C. K. Suzuki, "Development of a glove-based optical fiber sensor for applications in human–robot interaction," in *Proc. 8th ACM/IEEE Int. Conf. Hum.-Robot Interact. (HRI)*, Mar. 2013, pp. 123–124.

[11] B. Sadrfaridpour and Y. Wang, "Collaborative assembly in hybrid manufacturing cells: An integrated framework for human–robot interaction," *IEEE Trans. Autom. Sci. Eng.*, vol. 15, no. 3, pp. 1178–1192, Jul. 2018.

[12] E. Seemann, K. Nickel, and R. Stiefelhagen, "Head pose estimation using stereo vision for human–robot interaction," in *Proc. 6th IEEE Int. Conf. Autom. Face Gesture Recognit.*, May 2004, pp. 626–631.

[13] M. Hasanuzzaman, V. Ampornaramveth, T. Zhang, M. A. Bhuiyan, Y. Shirai, and H. Ueno, "Real-time vision-based gesture recognition for human robot interaction," in *Proc. IEEE Int. Conf. Robot. Biomimetics*, Aug. 2004, pp. 413–418.

[14] Z. Xia et al., "Vision-based hand gesture recognition for human–robot collaboration: A survey," in *Proc. 5th Int. Conf. Control, Autom. Robot. (ICCAR)*, Apr. 2019, pp. 198–205.

[15] C. Breazeal, "Social interactions in HRI: The robot view," *IEEE Trans. Syst., Man Cybern. C, Appl. Rev.*, vol. 34, no. 2, pp. 181–186, May 2004.

[16] J. T. C. Tan and T. Arai, "Triple stereo vision system for safety monitoring of human–robot collaboration in cellular manufacturing," in *Proc. IEEE Int. Symp. Assem. Manuf. (ISAM)*, May 2011, pp. 1–6.

[17] S. Yang, W. Xu, Z. Liu, Z. Zhou, and D. T. Pham, "Multi-source vision perception for human–robot collaboration in manufacturing," in *Proc. IEEE 15th Int. Conf. Netw., Sens. Control (ICNSC)*, Mar. 2018, pp. 1–6.

[18] F. Boniardi, T. Caselitz, R. Kümmerle, and W. Burgard, "Robust LiDAR-based localization in architectural floor plans," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2017, pp. 3318–3324.

[19] C. Debeunne and D. Vivet, "A review of visual-LiDAR fusion based simultaneous localization and mapping," *Sensors*, vol. 20, no. 7, p. 2068, Apr. 2020.
[20] R. Muñoz-Salinas, M. J. Marín-Jimenez, E. Yeguas-Bolivar, and R. Medina-Carnicer, "Mapping and localization from planar markers," *Pattern Recognit.*, vol. 73, pp. 158–171, Jan. 2018.
[21] S. Roos-Hoefgeest, I. A. Garcia, and R. C. Gonzalez, "Mobile robot localization in industrial environments using a ring of cameras and ArUco markers," in *Proc. 47th Annu. Conf. IEEE Ind. Electron. Soc. (IECON)*, Oct. 2021, pp. 1–6.
[22] H. Xing et al., "A multi-sensor fusion self-localization system of a miniature underwater robot in structured and GPS-denied environments," *IEEE Sensors J.*, vol. 21, no. 23, pp. 27136–27146, Dec. 2021.
[23] D. Szalóki, N. Koszó, K. Csorba, and G. Tevesz, "Marker localization with a multi-camera system," in *Proc. Int. Conf. Syst. Sci. Eng. (ICSSE)*, Jul. 2013, pp. 135–139.
[24] M. Yang and D. Wu. *Chinese Knowledge and Information Processing Lab*. Accessed: Jul. 13, 2022. [Online]. Available: https://ckip.iis.sinica.edu.tw/

Chen-Chien James Hsu (Senior Member, IEEE) was born in Hsinchu, Taiwan. He received the B.S. degree in electronic engineering from the National Taiwan University of Science and Technology, Taipei, Taiwan, in 1987, the M.S. degree in control engineering from National Chiao Tung University, Hsinchu, in 1989, and the Ph.D. degree from the School of Microelectronic Engineering, Griffith University, Brisbane, QLD, Australia, in 1997.

He was a Systems Engineer with IBM Corporation, Taipei, for three years, where he was responsible for information systems planning and application development, before commencing his Ph.D. studies. He joined the Department of Electronic Engineering, St. John's University, Taipei, as an Assistant Professor, in 1997, and was appointed as Associate Professor in 2004. From 2006 to 2009, he was with the Department of Electrical Engineering, Tamkang University, Taipei. He is currently a Professor with the Department of Electrical Engineering, National Taiwan Normal University, Taipei. He has authored or coauthored more than 200 refereed journal articles and conference papers. His current research interests include digital control systems, evolutionary computation, vision-based measuring systems, sensor applications, and mobile robot navigation.

Dr. Hsu is a Fellow of IET.

Pin-Jui Hwang received the B.S. and M.S. degrees in mechanical and electro-mechanical engineering from Tamkang University, New Taipei City, Taiwan, in 2014 and 2016, respectively. He is currently pursuing the Ph.D. degree with the Department of Electrical Engineering, National Taiwan Normal University, Taipei, Taiwan.

His research interests include robotics and computer vision.

Wei-Yen Wang (Fellow, IEEE) received the Diploma degree in electrical engineering from the National Taipei Institute of Technology, Taipei, Taiwan, in 1984, and the M.S. and Ph.D. degrees in electrical engineering from the National Taiwan University of Science and Technology, Taipei, in 1990 and 1994, respectively.

From 1990 to 2006, he worked concurrently as a Patent Screening Member of the National Intellectual Property Office, Ministry of Economic Affairs, Taipei. In 1994, he was appointed as Associate Professor with the Department of Electronic Engineering, St. John's and St. Mary's Institute of Technology, New Taipei, Taiwan. From 1998 to 2000, he worked with the Department of Business Mathematics, Soochow University, Taipei. From 2000 to 2004, he was with the Department of Electronic Engineering, Fu Jen Catholic University, New Taipei City, Taiwan, where he became a Full Professor with the Department of Electronic Engineering, in 2004. In 2006, he was a Professor and the Director of the Computer Center, National Taipei University of Technology, Taipei. From 2007 to 2014, he was a Professor with the Department of Applied Electronics Technology, National Taiwan Normal University, Taipei, where he was the Director of the Information Technology Center, from 2011 to 2013, and is currently a Professor with the Department of Electrical Engineering. His current research interests and publications are in the areas of fuzzy logic control, robust adaptive control, neural networks, computer-aided design, digital control, and charge-coupled device (CCD) camera-based sensors. He has authored or coauthored over 200 refereed conference papers and journal articles in the above areas.

Dr. Wang is an IET Fellow. He was a recipient of Best Associate Editor Award of IEEE TRANSACTIONS ON CYBERNETICS. He is currently serving as an Associate Editor for the IEEE TRANSACTIONS ON CYBERNETICS and the *International Journal of Fuzzy Systems*. Since 2003, he has been certified as a patent attorney in Taiwan.

Yin-Tien Wang (Member, IEEE) received the B.S. degree from Tamkang University (TKU), New Taipei City, Taiwan, in 1983, the M.S. degree from the Stevens Institute of Technology, Hoboken, NJ, USA, in 1988, and the Ph.D. degree from the University of Pennsylvania, Philadelphia, PA, USA, in 1992, all in mechanical engineering.

He joined the Department of Mechanical and Electro-Mechanical Engineering, TKU, as an Associate Professor, in 1992, and was appointed as Full Professor in 2013. He served as the Chairperson of the Department of Mechanical and Electro-Mechanical Engineering, TKU, from 2016 to 2020, where he is currently a Professor and the Chairperson of the Department of Artificial Intelligence, and also in charge of robotics and machine vision courses. His current interests include computer vision research and the transference of this technology to robotic and nonrobotic application domains.

Cheng-Kai Lu (Senior Member, IEEE) received the B.S. and M.S. degrees in electronics engineering from Fu Jen Catholic University, Taipei, Taiwan, in 2001 and 2003, respectively, and the Ph.D. degree in engineering from The University of Edinburgh, Edinburgh, U.K., in 2012.

He worked as the Director of the Research and Development Division, Chyao Shiunn Electronic Industrial Company Ltd., Shanghai, China. He joined the Science and Technology Policy Research and Information Centre, National Applied Research Laboratories, Taipei. He was a Faculty Member with the Electrical and Electronic Engineering Department, Universiti Teknologi PETRONAS (UTP), Perak, Malaysia, from 2016 to 2021. He is currently a Faculty Member with the Department of Electrical Engineering, National Taiwan Normal University (NTNU), Taipei. Apart from academic experience, Dr. Lu has more than eight years of industrial work experience. He has not only published his research works on peer-reviewed articles (book chapters, journal articles, conferences, and reports), but also has filed a couple of patents. His research interests focus on medical imaging, embedded systems, artificial intelligence and their applications, and clinical decision support systems.

Dr. Lu served as an Executive Member of the IEEE EMBS Malaysia Chapter and Penang Chapter from January 2017 to February 2018 and from 2018 onward, respectively.