Speech Control Care Robot

Ti-Huai Song
Department of Electrical Engineering
National Taiwan University
Taipei, Taiwan
r09921135@ntu.edu.tw

Hsiang-Hung, Wei
Department of Electrical Engineering
National Taiwan University
Taipei, Taiwan
r10921080@ntu.edu.tw

Hsin-Yang Chang
Department of Electrical Engineering
National Taiwan University
Taipei, Taiwan
r10921013@ntu.edu.tw

Hsin-Yang Chang
Department of Electrical Engineering
National Taiwan University
Taipei, Taiwan
r10921077@ntu.edu.tw

***Abstract*—For those bedridden elderly and paralyzed patients, it would be troublesome to move their body easily and maintain a simple daily routine. Moreover, with the strive of service robot and artificial intelligence, they could be a suitable solution to these issues. Therefore, we propose a robotic system called Speech Control Care Robot, where the robot arm can be controlled by human speech commands to help with some daily activities, more specifically, grabbing things and feeding food based on the speech contents. The system is achieved by natural language processing and computer vision to be able to understand the meaning of human commands which are then transformed into corresponding actions. The experiment results also verify the effectiveness of this robot system.**

The demonstration video and source code for this project are available in the links below:

Demo video:
https://drive.google.com/file/d/1TjUMl49ecqAa1v1MMWJWH-Y2Od6cv5yC/view?usp=sharing

Source code: *https://github.com/r09921135/robotics*

Keywords—TM5-900 Robot Arm, Natural Language Processing, Computer Vision, Human-Robot Interaction

I. Introduction

The population of elderly people is growing due to rapid improvement of medical science. It is expected to increase nearly 17% of the world's population by 2050, according to National Institutes of Health(NIH) [1]. On the other hand, the population that supports elderly people is decreasing, which indicates that older people are facing difficulties dealing with their basic daily routines on their own. Likewise, the situation of physically disabled people is similar to that of the elderly. These groups of people cannot move their body easily, so they usually depend on someone else's help to do some simple daily activities, for example, grabbing things they want or having a meal. As a result, service robot technology comes in handy to address the above issues.

Service robots are getting more attention in human living environments and helping us in many ways, for example, assistants, guides, and companions. Because the labor shortages loom, the demands for robot as a replacement for manual labor continues to grow for its durability. Consequently, the Human-Robot Interaction (HRI) will become a critical problem. Interaction requires communication between humans and robots. In human-robot communication, inputs can be given through keyboard typing, joystick, remote control or blue-tooth controlled mobile applications. However, that may lead to problems such as not straightforward and indirect. In order to overcome these effects, the natural way for the human to communicate with robotic systems is speech.

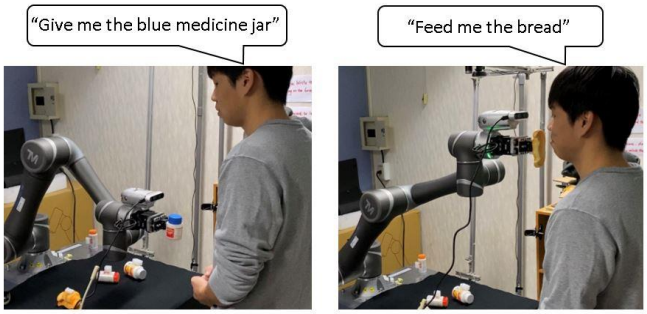

Fig. 1. Two daily activities our robotic system can perform: (a) giving things that human describe. (b) feeding the food that human want.

With the rapid development of artificial intelligence, Natural Language Processing (NLP) is becoming a more and more popular area which improves significantly the functionality of robotic system. It allows robots to process more complex control inputs. Thus, human can now communicate with robots in much easier and natural ways like via spoken dialogs. Recently, there is a surge of interest in combining NLP with Computer Vision (CV) [2, 3]. Such multi-modality applications could bring even more flexibility and diversity for the human-robot interaction.

In this paper, we focus on the scenario of deploying robot arm as a daily assistive system for those bedridden elderly people or paralyzed patients. Therefore, we propose a system called Speech Control Care Robot, where the robot can understand human speech commands and perform corresponding daily activities, more specifically, grabbing things that human describe (Fig. 1(a)) and feeding the food that human want (Fig. 1(b)). We use natural language processing to allow receiving a flexible and general speech command input, which makes it more feasible when applying in the real world. Also, computer vision is combined with NLP to be able to associate the command with the described object. Experiments are conducted to verify the effectiveness of this system.

The key contribution of this paper can be summarized as follows:

- We develop a speech control robotic system which can help grab things that people describe and feed the food that people want based on speech commands.
- We utilize natural language processing and computer vision techniques to enable the robot arm to receive a

flexible speech command input and then relate the command with the describe object.'

II. Related Work

Recently, many researches focus on the caring robot due to the increase of elderly and disabled population. Without others help, these people often have difficulties finishing some tasks independently, which are easy for normal people in daily life, for example, taking objects and eating. Hence, there are lots of researchers trying to solve the problem and create a caring robot, which aids people in such inconvenience in life.

Wen-Chang Cheng et al. [4] proposed a feeding robot system implemented in ROS system with intention detection. In their study, they use a robotic arm combined with image processing which detects the user's mouth opening and closing, to complete automatic feeding. Tapomayukh Bhattacharjee et al. [5, 6] proposed an autonomous feeding robot with assistive dexterous arm (ADA). In their study, they use a camera with two algorithms called "RetinaNet" and "SPNet" to recognize the size and shape of food, and analyze the best way to pick it up. They also proposed a force-sensing fork, which collected haptics while user biting. Using the haptics information, the robot can know whether the user finish their food. Obi Robot company [7] invent the Obi Robotic Self-Feeder. It can automatically capture the food and keep food transfer efficient. The user can simply teach it where it should bring the spoon, and it can quickly and intuitively select a food delivery location. Besides, with a click of a switch, the user can select food in different bowls.

In these works, however, most of them need the aid of user's hands. For example, user may operate a joystick to simply teach the robot arm how to move, or use a button to switch different plates. Although these robots drastically reduce the efforts from user, they are not totally automatic. Especially for those who cannot move their hands arbitrary, there still exists inconvenience. Therefore, the differences between our method and others are that our system is able to understand instructions from user's speech and find the user's position in a fully automatic fashion.

III. Experimental Setup

The experimental setup is illustrated in Fig 2., including RealSense depth camera, robot arm, microphone and workspace.

A. RealSense

RealSense is a kind of depth camera that can have RGB image and the depth information on every pixel. It will be used to calculate the distance between human and the robot arm.

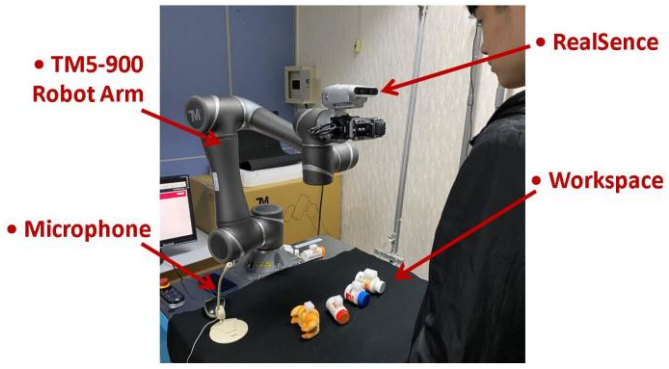

Fig. 2. The experimental setup.

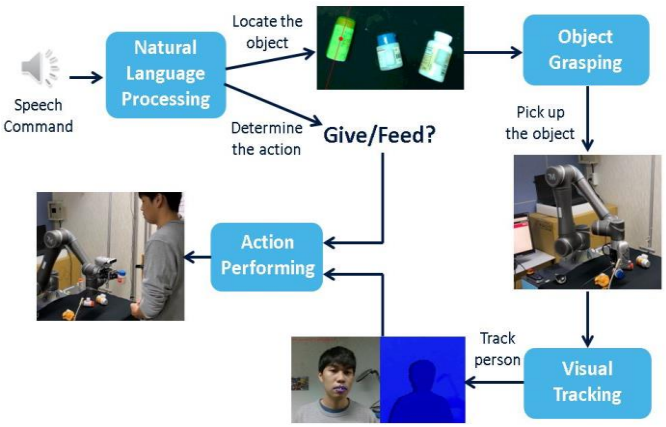

Fig. 3. The overall framework of our system.

B. Microphone

Human can give speech commands to the robot arm through microphone.

C. Robot Arm

We use the TM5-900 robot arm, its payload is 4 kilograms with a 900-millimeter reach, the typical speed is 1.4(m/s), and 0.05-millimeter repeatability (repeatable accuracy)

D. Workspace

A space where objects can be placed and robot arm can detect.

IV. System Overview

The overall framework is shown in Fig. 3. Concretely, a person will give a speech command. The speech command first goes into the *Natural Language Processing Module* to locate the target object and also determine the action category based on the semantics of the speech command. The action category is either "give" or "feed" in our project. With the location of the target object, the robot arm will pick up the target in the *Object Grasping Module*. Next, *Visual Tracking* is performed to track down the location of the person. Lastly, with the action category obtained from the earlier stages and also the location of the person, the robot arm will execute the corresponding action in the *Action Performing Module* and finish the entire procedure. In the next section, each module will be elaborated in more detail.

V. Methodology

A. Natural Language Processing

The framework of natural language processing module is shown in Fig. 4. Firstly, the speech command is transformed into text command in *Speech Recognition*. Next, in the *Object Description Extraction*, the text is divided into two subtexts: action subtext which possess the action information, and object subtext which represent the description about the referred object. For the action subtext, it is used to determine which action category it belongs to in *Action Classification*. For the object subtext, it is fed into *Referring Image Segmentation* along with an image of the workspace to

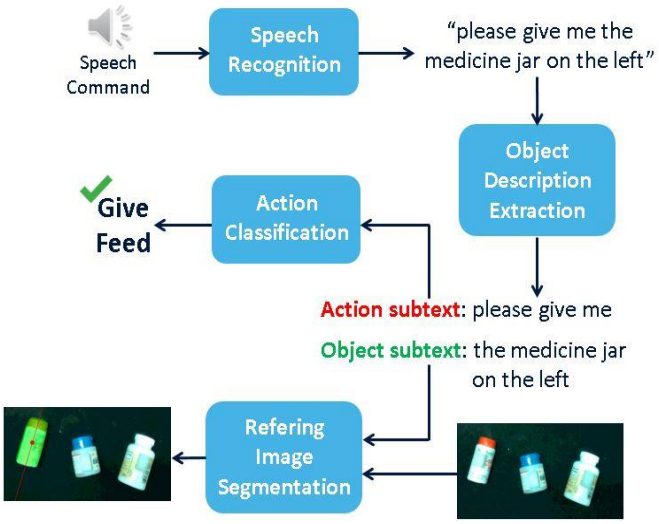

Fig. 4. The framework of the Natural Language Processing Module

Question	Sentence	Answer	start	end
What should I get for him?	Please give me medicine jar with blue cover	medicine jar with blue cover	3	7
What should I get for her?	Please give me the red medicine jar	the red medicine jar	3	6
What should I get for him?	give me the bottle on the left	the bottle on the left	2	6
What should I get for her?	Please give me bottle on the right	bottle on the right	3	6
What should I get for him?	Please give me the thermometer on the table	the thermometer on the table	3	7
What should I get for her?	give me a slice of apple	a slice of apple	2	5
What should I get for him?	Please feed me big apple	big apple	3	4
What should I get for her?	Please feed me the bread	the bread	3	4

Fig. 5. Some samples of the EQA dataset.

Sentence	label
Please give me	0
Please give me the	0
Please feed me	1
Please feed me with	1
Please give me the	0

Fig. 6. Some samples of the dataset for training action classifier.

generate the segmentation mask of the target object described in the object subtext. The mask is then used to calculate the picking location and orientation for the robot arm.

1) Speech Recognition: For the speech recognition, we directly apply the Speech-to-Text package that released by Google, which names *SpeechRecognition*. With the assistance of this package, the text command will be generated instantly after the person gives voice contents streamed from the microphone.

2) Object Description Extraction: Here, with a view to extract the object subtext from the original text command, we introduce the Extravtive Question-Answering with a fined-tune BERT [8] model to our system. BERT stands for Transformer-based Bidirectional Encoder Representations from Transformers, which is a pre-training technique for NLP proposed by Google. BERT is designed to encode deep representations of text. On the other hand, Extractive Question-Answering (EQA) is the task of extracting an answer from a text given a question, where BERT can be applied on this task. For such task, the model needs to understand relations between the question and the context, and also have the ability to locate the position of an answer phrase.

However, what EQA has to do with our system? This is the tricky part in our system. What we do is that given a text command "Please give me the medicine jar on the left.", we pre-define the question as "What should I get for him?" or "What does he want?". Then, by using BERT model, we can obtain the result "the medicine jar on the left", which is exactly the extraction of object description. This is how we reformulate our object description extraction problem as EQA.

For implementation, we use a pre-trained BERT model and then fine-tune on our EQA dataset. The contents of this dataset is shown in Fig. 5. The "Sentence" column is the input of the BERT model, which is the text command. The "Question" column is pre-defined by us and "Answer" column is the ground truth of the object subtext. "start" column and "end" column record the start position and the end position of "Answer" with respect to "Sentence". With these data, we expect BERT model to learn that for the given 'Question' and 'Sentence', it can predict the start position and end position of the object subtext in the text command. If we regard a row in the above table as a single training sample, we use 20 training samples in our dataset to fine-tuned our BERT model.

For obtaining action subtext, we simply subtract the object subtext from the entire text command and define the left subtext as the action subtext.

3) Action Classification: In our system, the robot arm can perform two kinds of actions, either "give" or "feed". Therefore, we simplify this task as a 2-way classification problem. More concretely, the action subtext first goes into the BERT model to obtain the word embedding of the subtext. This embedding will be the input to a simple multilayer perceptron (MLP) classfier to predict which action this embedding belongs to.

Fig. 6 shows some samples of our datasets that we use to train action classifier. The "Sentence" column is the input of the BERT model, which is the action subtext of the text command. The "label" column is the action ID, which is the output of action classifier. Label 0 represents the action of give and label 1 represents the action of feed. If we regard a row in the table as a single training sample, we use 15 training samples in our dataset to train the action classifier.

4) Refering Image Segmentation: In this part, we are trying to associate the object subtext with the target object on the workspace. Fortunately, there is an existing field of combining computer vision with NLP that meets our expectation called Referring Image Segmentation (RIS) [9]. RIS is the task that given a linguistic expression and an image, a binary segmentation mask will be generated for the target object which is referred to by the expression. This task meets the objective of our project where the robot needs to locate the object on the workspace based on object subtext.

The RIS model we choose is called RefVOS [10] for its competitive results in this task. RefVOS uses state of the art visual and language feature extractors, which are combined into a multi-modal embedding. This multi-modal embedding will then be decoded to generate a binary segmentation mask

Table 1. The information of the RIS dataset.

Number of images	40
Number of classes	4
Number of referring expressions per sample	2

Fig. 6. One of the samples of the RIS dataset.

for the referring target object.

For implementation, we use the RefVOS pre-trained on RefCOCO dataset [3] and then fine-tune on our RIS dataset. The information about our dataset is shown in Table 1. Specifically, for the two referring expressions of every instance, one of the expressions is made according to location, the other is made according to appearance. One sample of the RIS dataset is shown in Fig. 6. embedding belongs to. Visualizations with RefVOS are illustrated in Fig. 7. As the results show, the target object can be specified not only by its location but also by its appearance, where RefVOS is capable of correctly segmenting out the object in both ways.

B. Object Grasping

1) Transformation from camera to robot base: In order to grasping the object precisely, we have to transform the coordinate of a given point in camera frame into the position of the robot arm's end effector. This process is called frame transformation which derive the transformation matrix ${}^{base}T_{camera}$ as follows.

Let $C = [c_x, c_y, 1]^T$ be a 3-dimensional vector in which c_x and c_y represent x axis and y axis of the object in camera frame coordinate respectively. And $B = [b_x, b_y, 1]^T$ be a 3-dimensional vector in which b_x and b_y represent x axis and y axis of the object in robot base frame coordinate respectively. First, we have to calculate the relationship of B and C vector.

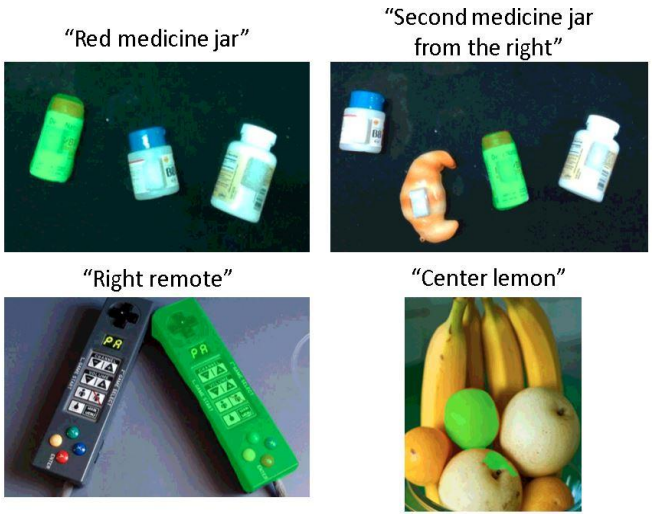

Fig. 7. Visualization results of RefVOS.

$$B = {}^{base}T_{camera} \times C \tag{1}$$

We will determine three different points $b_1 = [b_{1x}, b_{1y}, 1]^T$, $b_2 = [b_{2x}, b_{2y}, 1]^T$, $b_3 = [b_{3x}, b_{3y}, 1]^T$ in robot base frame coordinate from the workspace. And, we move robot arm to the predefine position to take a picture for workspace. After that, we can get b_1, b_2, b_3 points' pixel value in the camera frame coordinate respectively, and c_1, c_2, c_3 represent as $c_1 = [c_{1x}, c_{1y}, 1]^T$, $c_2 = [c_{2x}, c_{2y}, 1]^T$, $c_3 = [c_{3x}, c_{3y}, 1]^T$.

Due to the unit from camera and robot base coordinate are different, we have to calculate r which is the ratio of pixel to mm. Then, we take three different points with camera frame coordinate and robot base frame coordinate, it is easy to get the transformation matrix by inverse the camera frame matrix C. And the (1) can be rewritten as:

$$ {}^{base}T_{camera} = B \cdot C^{-1} \tag{2}$$

where $B = \begin{bmatrix} b_{1x} & b_{2x} & b_{3x} \\ b_{1y} & b_{2y} & b_{3y} \\ 1 & 1 & 1 \end{bmatrix}$, $C = \begin{bmatrix} c_{1x} * r & c_{2x} * r & c_{3x} * r \\ c_{1y} * r & c_{2y} * r & c_{3y} * r \\ 1 & 1 & 1 \end{bmatrix}$

2) Robot Arm Control: In our control strategy, we use ROS2 (Robot Operation System 2) as our core implementation in this project. This system allows us to transfer information among different parts: robot arm and computer. With this feature, we can use the Python code to give command to the TM5-900 robot arm. At the beginning, we use RealSense camera to get the camera frame image at predefine position. Once we get the image that can use RIS model with the user's description, it can successfully get the object's centroid pixel value. Then, use the transformation matrix ${}^{base}T_{camera}$ to let object pixel value in camera coordinate transfer to the robot base coordinate. After getting the object centroid with the robot base coordinate, we use Python to give the moving command to robot arm in order to grasp the object.

C. Visual Tracking

1) Face Recognition: For the face recognition part, we directly apply the python open package face_recognition. This package can handle plenty of tasks, including face tracking, face recognition and face landmark localization. Here, we utilize the function in tracking face landmark of human's top lip and bottom lip to localize the several points surrounding around our mouth. Then, we will calculate centroid of these points, which is the pixel for robot arm to track. The face landmark detection is shown in Fig. 8.

2) Transformation from camera to robot base: After finishing face recognition, we received a center point of lips in camera coordinate. The next thing we need to do is transforming the point from camera coordinate to robot coordinate. To finish this task, we use the similar strategy as the one applied on object grasping. Due to the complexity of calibrating a coordinate system with depth information, and considering our actual goal in this step, we compute the rotation matrix between two coordinate systems instead of the transformation matrix with displacements.

Fig. 8. The face landmark detection.

First, we define three points in the robot coordinate. Without loss of generality, we choose (1, 0, 0), (0, 1, 0) and (0, 0, 1) to be three standard point in robot coordinate and the corresponding position in camera coordinate is (0.707, 0, 0.707), (-0.707, 0, 0.707), (0, -1, 0) in camera coordinate. Since we choose the unit vector as standard points, the rotation matrix is just constructed by these three vectors in robot coordinate. Next, by (3), we can directly find out the rotation matrix R between two coordinate systems:

$$c = R \cdot r \tag{3}$$

where r is the position in robot coordinate, and c is the position in camera coordinate. By inversing the equation, we can get (4):

$$r = R^{-1} \cdot c \tag{4}$$

Using (4), when we input a point in camera coordinate (c_x, c_y, c_z), we can acquire and move to the corresponding position in robot coordinate (r_x, r_y, r_z) after rotation, which helps us to change the depth camera's view from the table to our face.

3) Robot arm control: After we get the rotation matrix and turn the depth camera facing our face, the next step is to define our tracking strategy. Since we expect our system to keep tracking before we stop moving, we create a system that allows robot arm to follow our face continuously. First, the system will determine the relative position of the center point of lips and the center of camera frame. If the center point of lips isn't in the middle of frame, the robot arm will start to move, making our lips appear in the frame's center.

We use *Move_PTP* in ROS2 system, moving the robot arm with respect to its last position. Each time, the robot arm moves 10-mm toward the center, in both x and y direction in camera coordinate; meanwhile, we set a threshold for our tracking system. If the center point of lips is less than 40 mm far from the frame's center, robot arm will stop moving. By doing so, we can make sure that our system won't be too sensitive to a small movement of our face, like breathing or winking. Last, we set a stop condition: if the robot arm didn't move for about 2 seconds, stop tracking and break the tracking system. The robot arm will be fixed at the point where the camera is directly facing our lips.

D. Action Performing

The last thing we have to do is action performing. As we mentioned, our caring robot has two different actions: give and feed. Therefore, after recognizing the users demand by speech, the robot arm needs to perform different actions according to user's instruction. For the action of give, the robot arm will move downward for about 300-mm followed by moving forward until it is 250-mm far from our body. At this position, we can easily get whatever we asked.

For the action of feed, the robot arm will go forward to the location 150-mm away from the face. This distance can be adjusted when equipped with different tableware on the robot arm. For example, if we want to use a fork to eat fruits, the distance must be longer due to the length of fork. After the robot arm stopped right in front of our face, there still needs some fine-tuning for us to eat the food more comfortably. Because we set our camera on the robot arm, the center of camera isn't really at the center of the gripper. Thus, we need to move the robot arm 100-mm up along the z direction of robot coordinate. After finishing the whole process, we can enjoy our delicious food.

After the robot arm complete the action, it will open the gripper when receiving message "open", and then move back to the origin waiting for another speech commands.

VI. Experiments

The entire experiment result is available in demo video link we provide. The experiment consists of two cases. The first case is to ask the robot arm to give the thing we describes. For example, "I want the medicine jar with blue cover". The second case is to ask the robot arm to feed us the food we want. For example, "I want to eat the bread". As the results shows, the robot arm is able to understand the meaning of the commands and not only decide which action it should perform but also what object on the workspace that is referred to in the command. After the robot pick up the target object, it can successfully track the position of the person. Lastly, the robot arm will execute the corresponding action of either give or feed. After the entire procedure is completed, the robot arm moves back to the initial location and wait for the next commands over and over again.

VII. Conclusion

In this paper, we present a robotic system called Speech Control Care Robot. The robot is capable of performing two daily activities: grabbing things and feeding food that human describe. With the NLP and CV techniques, robot can identify the target object on the workspace and determine which action to perform based on the speech commands. Moreover, an accurate transformation matrix enables the robot to reach the target. By using face recognition, robot can keep tracking the person's position from the RealSense camera. The experiment results show that this robot has high accuracy to grasp the object and finish the task.

Due to the constrained workspace of TM5-900 and the size of gripper, it can be a challenge to fulfill every task from different places and targets. For future work, we hope Speech Control Care Robot can work at everywhere by mounted on the mobile robot. Then, using a bigger gripper to adaptively grasp any size of targets.

VIII. ACKNOWLEDGMENT

In this project, we are grateful for lectures by Professor Li-Chen, Fu, and his advice and support. We have also got a lot of advice from teaching assistants. Finally, we've felt grateful for the device we used in this project, including TM5-900 robot arm, RealSense camera and etc, provided by National Taiwan University and Advanced Control Lab.

REFERENCES

[1] "World's older population grows dramatically-(NIH)", March 2016, Online Available: https://www.nih.gov/news-events/news-releases/

[2] A. Karpathy and L. Fei-Fei. Deep Visual-Semantic Alignments for Generating Image Descriptions. In CVPR, 2015.

[3] R. Hu, M. Rohrbach, and T. Darrell. Segmentation from natural language expressions. In ECCV, 2016.

[4] W. -C. Cheng, H. -C. Hsiao and C. -Y. Chung, "Implementation of ROS-based Feeding Robot System with Intention Detection," 2020 International Automatic Control Conference (CACS), 2020, pp. 1-5, doi: 10.1109/CACS50047.2020.9289772.

[5] [2] T. Bhattacharjee, G. Lee, H. Song and S. S. Srinivasa, "Towards Robotic Feeding: Role of Haptics in Fork-Based Food Manipulation," in IEEE Robotics and Automation Letters, vol. 4, no. 2, pp. 1485-1492, April 2019, doi: 10.1109/LRA.2019.2894592.

[6] How to train your robot (to feed you dinner)? University of Washington News. https://www.washington.edu/news/2019/03/11/how-to-train-your-robot-to-feed-you-dinner/

[7] Obi Independent Eating. https://meetobi.com/

[8] Kenton, Jacob Devlin Ming-Wei Chang, and Lee Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." Proceedings of NAACL-HLT. 2019.

[9] R. Hu, M. Rohrbach, and T. Darrell. Segmentation from natural language expressions. In ECCV, 2016.

[10] Bellver, Miriam, et al. "Refvos: A closer look at referring expressions for video object segmentation." arXiv preprint arXiv:2010.00263 , 2020.

WORK DIVISION

- Ti-Huai Song: Deep learning method research, BERT model research, RIS model research & training , experiment, report integration, video record & edit.
- Hsin-Yang Chang: Deep learning method research, BERT model research & training, experiment, report.
- Hsiang-Hung, Wei: robot arm control, coordinate transformation, system integration, experiment, report
- Sheng-Bang, Lin: robot arm control, coordinate transformation, system integration, experiment, report.