2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

Natural Language Instruction Understanding for Robotic Manipulation: a Multisensory Perception Approach

Weihua Wang†, Xiaofei Li†, Yanzhi Dong, Jun Xie, Di Guo*, Huaping Liu*

***Abstract*— It has always been expected that the robot can understand the natural language instruction and thus a more natural human-robot interaction is achieved. Currently, the robot usually interprets the instruction by visually grounding the textual information to its surroundings, while it may be not enough for some complex situations with only visual perception. So it is reasonable for the robot to leverage its multisensory perception ability to better understand the instruction. In this paper, we propose a multisensory perception approach to tackle the task of natural language instruction understanding for robotic manipulation, in which the robot coordinates its visual, tactile and auditory perception to fully understand the instruction and then executes the manipulation task. Extensive experiments have been conducted demonstrating the superiority of the multisensory perception compared with single sensory perception for instruction understanding. Moreover, we establish a user-friendly human-robot interaction interface where the human sends instruction to the robot via a mobile APP.**

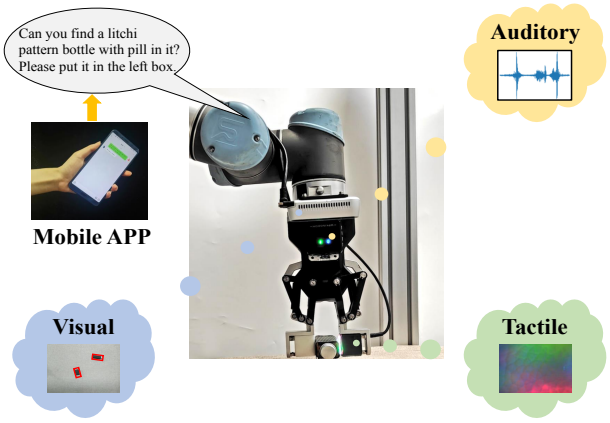

Fig. 1. The robot leverages its multisensory perception ability to understand the natural language instruction sent to it via a mobile APP.

I. INTRODUCTION

As the robots are more and more engaged in our daily life, there is a growing need to establish an approach for a natural human-robot interaction. Natural language is supposed to be a most intuitive interface for human-robot interaction. It will be helpful if the robot can understand and execute the natural language instruction given by the people. One of the great challenges in this task is how the robot can ground the mentioned objects and actions in the instruction in the real environment. To this end, we propose a multisensory perception approach, in which the robot leverages its multisensory perception ability, namely the visual, tactile and auditory perception, to understand the given natural language instruction for robotic manipulation.

To achieve a more natural human-robot interaction, multiple approaches have been investigated for years [1]. The gesture plays as a preliminary attempt for an easy-to-use human-robot interaction interface [2]. With a visual system to recognize human gesture, the robot can respond to human's intention automatically. Some robots can even recognize human's facial expressions and achieve interactions [3]. Although the gesture and facial expressions provide meaningful indications for the interaction, they are usually limited to some fixed responses given in advance and can not be generalized to new situations. Indeed, in the human-centered environment, natural language is the most intuitive interface for the human-robot interaction. It has always been our dream to seamlessly interact with the robot with natural language. And there are some works [4], [5] trying to integrate the natural language in the human-robot interaction. However, due to the great challenge of natural language understanding, the task is usually subjected to some fixed situations.

On the other hand, current robotic language instruction understanding task mainly relies on the visual information [6], [7]. The robot visually grounds the mentioned object in the scenario and then executes the manipulation task. With the great development of natural language area, recently, there are some works [8] that include more complex description in the language instruction, such as referring expression, which requests a higher semantic understanding ability of the robot. Furthermore, in some complex situations, only with visual information is not enough to thoroughly understand the natural language instruction. For example, some visually identical containers may contain different contents, while the robot can obtain different tactile [9] and auditory [10] perception information when interacting with them. It is worth noting there are some previous works trying to ground natural language to object attributes [11], [12], [13]

In this paper, we propose a multisensory perception approach to tackle the task of natural language instruction understanding for robotic manipulation (Fig. 1). Human gives language instruction to the robot via a mobile APP, and then the robot leverages its visual, tactile, and auditory perception

† denotes equal contribution.

* denotes corresponding authors: Huaping Liu (hpliu@tsinghua.edu.cn) and Di Guo (guodi.gd@gmail.com).

Weihua Wang and Yanzhi Dong are with Department of Physics and Electronic Information, Yantai University, Yantai, China.

Xiaofei Li and Jun Xie are with Department of Information and Computers, Taiyuan University of Technology, Taiyuan, China.

Di Guo is with School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China.

Huaping Liu is with Department of Computer Science and Technology, Tsinghua University, Beijing, China.

to interact with the environment so that it can understand and execute the given instruction. To enable the robot to have multisensory perception ability, we equip the robot with a wrist camera, wireless microphone and fingertip tactile sensor [14]. And a visual-auditory-tactile dataset is also collected with the robotic platform. The main contributions of the paper are summarized as the following:

- A multisensory perception approach is proposed for the robot to coordinate its visual, tactile and auditory perception to understand natural language instruction, and a multisensory dataset is collected accordingly.
- Extensive experiments are conducted demonstrating the superiority of the proposed multisensory perception approach compared with single sensory perception approach in interpreting the language instruction.
- A natural human-robot interaction interface is built in the real-world environment, where the human gives the instruction to the robot via a mobile APP and the robot executes the instruction with the proposed multisensory perception approach.

This paper is organized as follows. The related work is introduced in Section II. Section III presents the architecture of the proposed multisensory perception approach. The details of the proposed approach is described in Section IV. The collection of the multisensory dataset is introduced in Section V. Section VI provides the experimental evaluations of the proposed framework. Finally, we come to a conclusion.

II. RELATED WORK

A. Language in robotic manipulation

In a human-centered environment, it is important for the robot to execute manipulation tasks following language instructions. And many studies have been investigated to learn how the robot could interpret language instructions. In an earlier work [15], a library of verb-environment instructions is built to map instruction to robot actions. According to the language instruction, Ref. [16] propose a GAN-based method to predict the target and source objects in the scene to fetch daily objects. In [17], a gated-attention mechanism is used to fuse the visual and textual information and ground the instruction in 3D environments. Rather than just grounding the object in the visual scene, a Command Grasping Network is proposed in [18], which can directly predict robotic grasp configuration from the image and language instruction.

As the language become more complicated, a higher semantic understanding ability of the robot is required. To understand some abstract spatial concepts, [19] uses a probabilistic model to incorporate the expressive symbol space and the graphical model for the abstract concepts and concrete constituents are established for the inference. Ref. [20] proposes a pipelined architecture to perform the spatial reasoning on the textual information. It can localize all the objects in the scene and generate the pick up and place locations for the robot. Besides specifying the visual attributes of the object in the instruction, [21] develops a model which localizes the object based on the textual description of their usage and an object retrieval task is implemented. Moreover, [8] proposes to allow the robot to ask human questions in order to clarify the ambiguous expressions in the instruction. In a more complicated situation where a high level instruction for a robotic task of long execution horizons, [22] proposes a persistent spatial semantic representation method which is used to bridge the gap between the language and robot actions over the long-term task.

B. Multisensory perception

It can be seen from the above that most approaches only resort to the visual information to interpret the language instruction. However, there exist some object attributes that could only be identified by other sensory modalities. With the tactile sensor, it is convenient for the robot to capture many object attributes such as texture, weight, friction, etc [23], [24]. In [25], the robot can learn the haptic adjectives through the tactile perception. For visually identical containers, the tactile perception can be used to recognize the content in the container [9], [26]. In addition, tasks such as object hardness detection [27], sliding detection [28] [29] can also be achieved through tactile perception. It has also been demonstrated that the combination of visual and tactile perception can yield a better performance in object recognition task [30].

As an important sensory modality, sound information can also help us to better understand the environment. In [31], by the sound of shaking and pouring, the type and quantity of the content in a container can be identified. Ref. [32] builds a physical platform to predict the object by tilting the tray and collecting the sound when the object hitting the tray walls. Together with the visual information, [10] proposes a robotic audio-visual grounding operating system in which the robot actively interacts with the objects to generate the sound information, and the textual instruction is interpreted by using both the visual and sound information.

Recently, Ref. [33] proposes a multisensory dataset which includes the vision, tactile and sound information. It can provide a good multisensory learning testbed for the research community. However, to the best of our knowledge, there is still not any dataset that incorporates the natural language understanding with the multisensory perception.

III. ARCHITECTURE

The robotic platform used in this paper is shown in Fig. 2. A 6DoF UR5 robotic arm is placed in front of a table and some bottles are placed on the table. It is noted that the surfaces of the bottles are wrapped with different fabrics and different contents are in the bottles. We have a two-finger adaptive AG-95 gripper integrated at the end of the arm. We also mount the DIGIT tactile sensor [14] on the fingertip of the hand to collect the tactile data. To capture the visual information, a RealSense camera is mounted at the end of the arm. Besides, a wireless microphone is installed at the wrist of the manipulator for the sound collection.

The architecture of the proposed multisensory perception approach is demonstrated in Fig. 3. It is composed of a

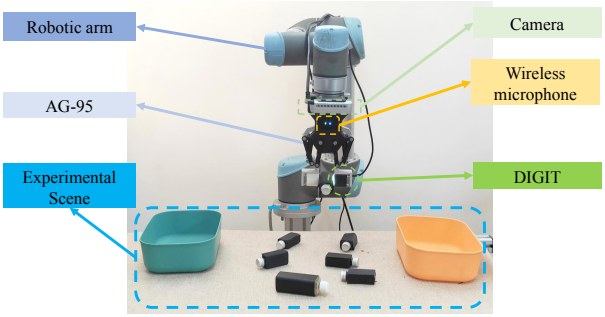

Fig. 2. Robotic platform.

human-robot interaction interface, language module, visual module, tactile module, audio module and manipulation module. The human sends the instruction to the robot via a mobile APP. It is convenient in situations where the robot and human are in different places. After the robot receives the instruction, the language module is firstly triggered to parse the instruction. And then the visual module, tactile module, and auditory module coordinate to ground the textual information in the environment. The visual module is responsible for locating bottles on the desktop. As it is difficult to identify the target bottle with only visual information, the robot uses the tactile sensor integrated in its fingertip to collect tactile information of the bottle for texture recognition. Meanwhile, the robot can manipulate with the bottle to generate sound information to recognize the content in it. With the target bottle identified, the robot picks up the it and place it in the designated location. It is noted that the robot interact with the environment with actions in the manipulation module, which contains touch, pick, whirl, shake, roll, and place.

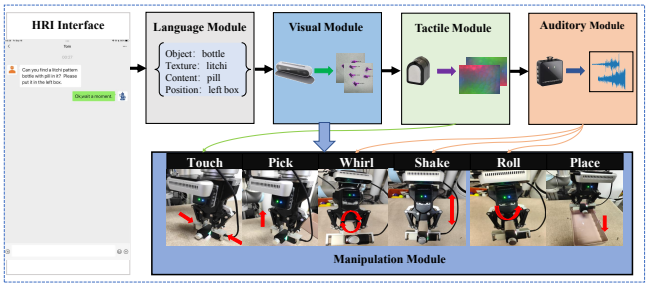

Fig. 3. System architecture.

IV. APPROACH

To fully interpret the natural language instruction, a multisensory perception approach is proposed, which mainly consists of language module, visual module, tactile module and auditory module, which are combined together for the target recognition and position detection (Fig. 4).

A. Language Module

We use the UIE model [34] to extract entities and positional relationships from the given instruction, and convert the extracted information into keywords for the multisensory perception modules. Formally, it is defined by a given text sequence x and a structural pattern indicator s, and a linearized structured extraction language y is generated:

$$y = \mathrm{UIE}(s \oplus x) \tag{1}$$

For example, given the instruction "I can't find my litchi pattern bottle, which contains pills. Please put it in the right box", it is fed into the UIE model, and then the object and position relationship vectors are obtained. The vector space is defined as $y = \{y_o, y_t, y_c, y_p\}$, among which y_o, y_t, y_c, y_p represent object, texture, content and position information in the instruction.

B. Visual Module

In terms of visual feature extraction, we use a rotating object detection framework based on the YOLO-V5 network [35]. The size of the input image is 320x320. The model can return the center coordinates and rotation angle θ of the target object for grasping. The angle θ is the angle between the long side of the detection frame and the horizontal axis. An angle prediction branch is added by converting the regression problem into a classification problem with Circular Smooth Labels (CSL).

$$\mathrm{CSL}(x) = \begin{cases} g(x), \theta - r < x < \theta + r \\ 0, \quad \text{otherwise} \end{cases} \tag{2}$$

where $g(x)$ and r represent the window function and radius. We use a Gaussian window function and r is set to 6. Finally, we transform the pixel coordinates of the center point and the rotation angle into the robotic coordinate for the manipulation.

C. Tactile Module

We use the DIGIT tactile sensor [14] in this work. It is a high-resolution optical tactile sensor, and it collects the feature map of the object through the built-in three-color camera and ordinary optical camera. It uses the opaque gel attached to the surface to feel the deformation of the object. When the gel is in contact with the object, the deformation will indirectly cause the three channels of R, G, and B to change. We can use the obtained tactile image as input for tactile recognition. Fig. 5 shows the DIGIT tactile sensor used in the experiment, and two tactile images produced by the sensor when it contacts two different objects. We feed the generated tactile images into the ResNet [36] for the classification. The input image size is 224x224.

D. Auditory Module

We predict the content in the bottle by recognizing the sound signals generated by the three predefined actions, namely whirl, shake and roll. We adopt the VGGish model [37] to extract the features from sound signals. Three auditory clips from three actions are firstly concatenated into one clip and then the clip is converted into log mel spectrogram features and the obtained audio features are fed into the VGGish model. Finally, we obtain a 128-dimensional feature vector with semantic information for the audio clip.

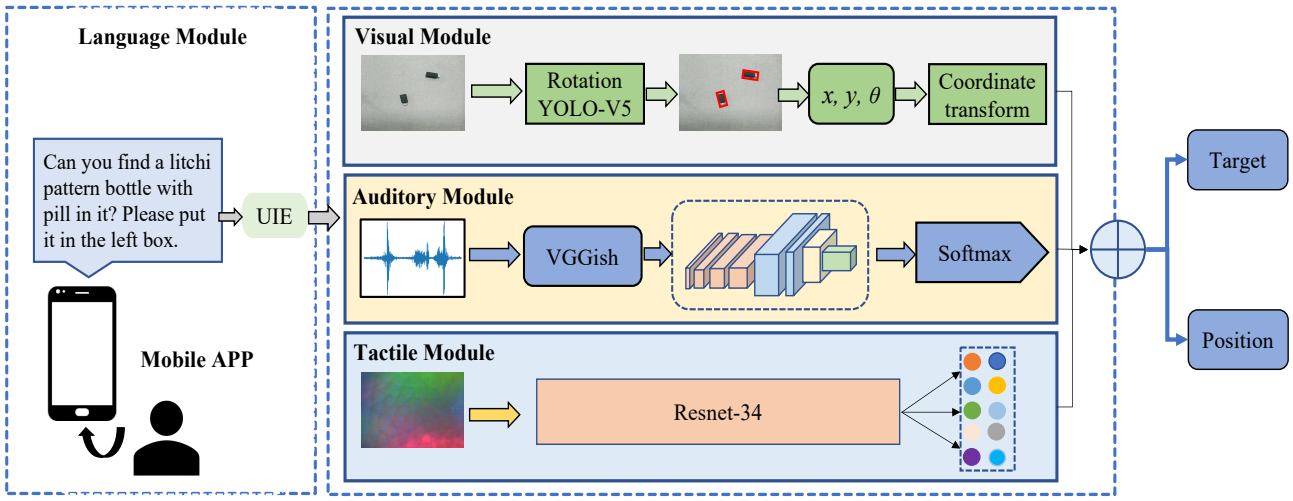

Fig. 4. An overview of the multisensory perception approach.

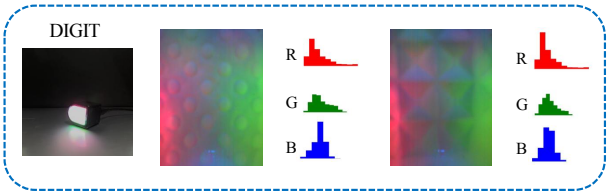

Fig. 5. The left side of is the DIGIT tactile sensor. The middle and right are two tactile images produced by two different objects and the changes in RGB channels.

V. DATASET

To collect the multisensory dataset, we consider the situation that six kinds of content (pill, grain, liquid, capsule, solid, and powder, see Fig. 6) are stored in the bottles and the bottles are the same in size but wrapped with 10 different fabrics (square pattern, grain pattern, pinstripe pattern, bamboo weave pattern, diamond pattern, widestripe pattern, cross weave pattern, litchi pattern, leather pattern, and roundhole pattern, see Fig. 8). The human sends an instruction to the robot to identify a target bottle.

A. Instruction design

For the instruction, we would like the robot to identify the specific bottle mentioned in the instruction and place it in the target location. The bottles are with different contents and different surface textures. As shown in Table I, the instructions are divided into existential instructions and logical instructions. For the existential instructions, the robot only relies on the visual perception module to identify the target object. In this situation, we have introduced five common objects such as toy, ball, banana, stapler and charger. Logical instructions require the robot to leverage multisensory perception to fulfill the task. In this situation, the robot is required to distinguish different bottles. Additionally, according to the habit of human language, the instruction expressions are divided into direct and interrogative sentences, so as to enhance the language richness.

B. Visual dataset

In order to improve the generalization ability of the model, we collect the visual data of the scene in two lighting conditions with the camera. The scene includes the bottles and some other objects in the experiment. A total of 600 images are collected and the dataset is expanded to 2400 by rotation and mirroring operations, which is divided into 2000 for training set and 400 for validation set. We manually label the object class, center coordinates, and rotating angle of the data in the scene.

C. Audio dataset

For bottles with different contents, we collect the sound signals generated when the contents collide with each other during manipulation. As shown in Fig. 6, we select 6 typical contents, namely pills, grains, liquids, capsules, solids, and powders, which are usually stored in the containers in daily life. We put the contents into the bottles and design three actions (whirl, shake, and roll) for the robotic arm to interact with bottles to generate the sound. The wireless microphone at the wrist of the arm is used to collect the audio data when the robot interacting with bottles. We collect the audio data with a duration of 2s, and 40 sets of data are collected for each action. In total, we have collected 720 sets of audio data for 6 contents. Fig. 7 shows the sonograms and spectrograms of some objects obtained by shaking them, and they demonstrate obvious different characteristics.

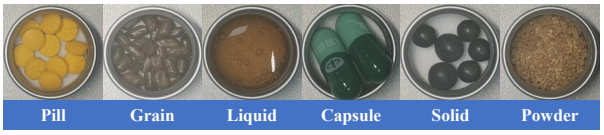

Fig. 6. Six kinds of content.

D. Tactile dataset

As shown in Fig. 8, the tactile dataset contains 10 kinds of fabrics with different textures. Different fabrics are wrapped

TABLE I

LANGUAGE INSTRUCTION MODE

Instruction type	Direct statement	Interrogative statement
Existence Instruction	Please find a <obj>, and put it in the <position>.	Is there a <obj> on the table? Please bring it to <position>.
Logical Instruction	I can't find my <tex><obj> which contains <cont>. Please put it in the <position>.	Can you find a <obj> with <tex> and <cont> in it? Please put it in the <position>.
<obj>:bottle or interfering objects; <tex>:10 textures; <content>:6 contents; <position>:left box or right box Dataset size:268 instructions,which including 28 direct & interrogative existence instructions and 240 direct & interrogative logical instructions		

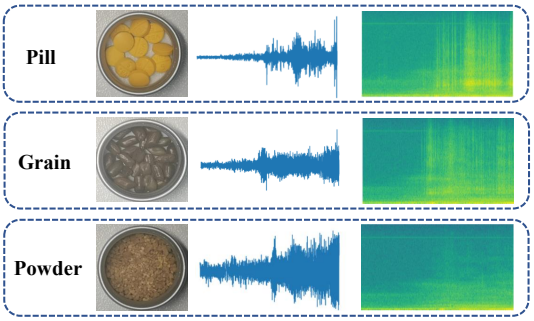

Fig. 7. Waveforms and spectrums of pill,grain and powder

over the bottles. We collect the tactile data from both the edge and surface of the bottle with different pressures. As is seen in the Fig. 8, the first two rows demonstrate the tactile images of the 10 fabrics in the experiment. And the third row indicates our collection process. We collect 120 tactile images for each type of fabric. Finally, the dataset is augmented to 2400 images by flipping operations.

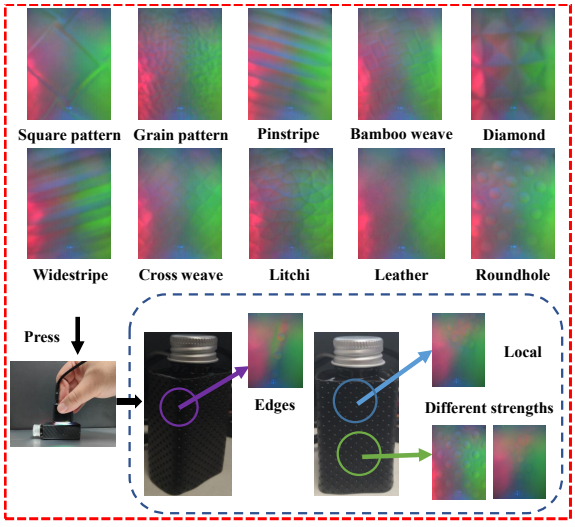

Fig. 8. The first two rows are the textures of 10 materials, and the last row is the collection process: the tactile data is collected from both the edge and surface of the bottle with different pressures.

VI. EXPERIMENT

In order to evaluate the proposed multisensory perception approach, we conduct single sensory offline evaluation experiments and the multisensory perception online evaluation experiment respectively.

A. Single sensory offline evaluation

- Auditory Model Assessment (AME)

We use the auditory dataset for the evaluation, among which 30 sets of audio data for each content. The test results are shown in Table II. As is shown in the TABEL II, an average accuracy of 77.4% is obtained for the sound recognition task. It can be seen that the recognition accuracy of different objects varies. It is observed that actually the whirl action yields the best performance. We believe that it is because that the size of the bottle used in our experiment is relatively small and there is noise when the robot is operating. The whirl action can best shake the bottles to generate the auditory data. In the future, we can reduce the environmental noise to improve the classification accuracy.

TABLE II

ACCURACY OF AUDIO OFFLINE EXPERIMENT

	Whirl	Roll	Shake	All actions
Capsule	96.7%	63.3%	56.7%	72.2%
Grain	80.0%	86.7%	60.0%	75.6%
Liquid	76.7%	73.3%	96.7%	82.2%
Pill	93.3%	90.0%	80.0%	87.8%
Powder	83.3%	86.6%	63.3%	77.7%
Solid	76.7%	63.3%	66.7%	68.9%
Average	**84.4%**	**77.2%**	**70.6%**	**77.4%**

- Tactile Model Assessment (TME)

We evaluate the tactile model with the tactile dataset to distinguish 10 fabrics used in the experiment. For each texture, 40 sets of tactile data are collected. As shown in Table III, offline experimental evaluation on 10 textured fabrics are conducted. From the test results, we can see that the overall classification accuracy reaches 90.3%, demonstrating the effectiveness of the trained tactile recognition model.

B. Mutisensory perception online evaluation

In order to evaluate our multisensory perception approach, we evaluate three fusion models online, namely language-visual-auditory model (LVA), the language-visual-tactile model (LVT), and the language-visual-auditory-tactile model (LVTA). For a easier human-robot interaction, we deployed a user-friendly human-robot interaction interface on a mobile APP, in which the human can send instruction to the robot. We interact with the robot for 200 times, among which 60 times for LVA, 60 times for LVT, and 80 times for LVTA. We calculate the success rate of the experiments with the number of successful experiments.

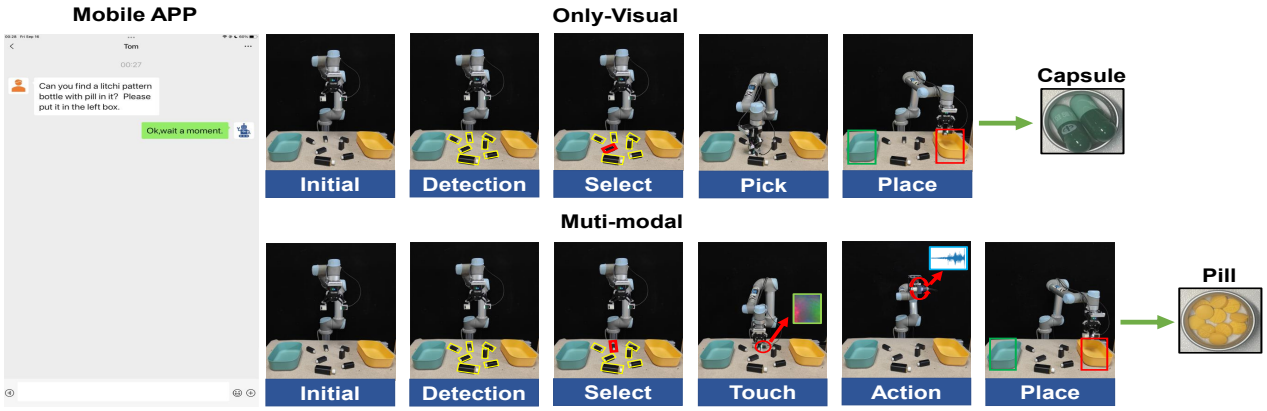

Fig. 9. Qualitative results with the proposed multisensory approach. In the only-visual setting, the robot randomly picks up a bottle, while in the multi-modal setting, the robot iteratively selects possible bottle from the detection results and conducts multisensory perception until the target bottle is found.

TABLE III
ACCURACY OF TEXTURE IMAGES

Pattern	Square	Grain	Pinstripe	Bamboo	Diamond	Widestripe	Cross	Litchi	Leather	Roundhole	**Average**
Accucary	92.5%	87.5%	75.0%	90.0%	95.0%	92.5%	95.0%	92.5%	87.5%	95.0%	90.3%

- Language-visual-auditory model (LVAM): According to the received instructions, the bottles in the scene are actively visually recognized and picked up in order, and the manipulator executes three actions to collect the sound information, and at the same time the auditory model is called to complete classification task.
- Language-vision-tactile model (LVTM): According to the received instructions, the visual model is used to recognize multiple bottles in the scene, and the tactile model is invoked to complete the active tactile perception operation task.
- Language-visual-auditory-tactile model (LVTAM): According to the received instructions, the bottles in the scene are actively visually recognized and picked up in order, and the tactile and auditory models are called respectively. The instruction understanding task is executed by the multisensory perception.

TABLE IV show our experimental results. Among them, "+" means there are other object besides the bottles in the scene, and "-" means only bottles are in the scene. The results in the TABLE IV show that the success rate of our multisensory perception approach in the three fusion modalities is greater than 70%. And an average execution time for LVTAM is 85s. The LVTM works best mainly due to the satisfying tactile recognition ability. We find that models (LVAM and LVTAM) with auditory model works not as well as expected. It is because that the bottles are small and thus the sound generated is not quite obvious. We will try better auditory model in the future.

Fig. 9 demonstrates the qualitative results of the proposed approach. The human firstly sends instruction to the robot via a mobile APP, and then the robot executes the instruction based on its understanding. We have used a visual-only approach as a baseline, in which the robot randomly picks up a bottle. It can be seen that with multisensory perception, the robot is able to correctly identify the bottle mentioned in the instruction.

TABLE IV
ACCURACY OF MUTI-MODAL LINE EXPERIMENT

Muti-modal	**+**	**-**
LVAM	70.0%	73.3%
LVTM	83.3%	86.7%
LVTAM	77.5%	80.0%

VII. CONCLUSIONS

In this work, we propose a multisensory perception approach to tackle the task of natural language instruction understanding for robotic manipulation. The robot actively coordinates its visual, tactile and auditory perception to fully understand the instruction and thus executes the manipulation task. We have also collected a language-visual-auditory-tactile dataset for the task. In the experiments, we have compared the proposed multisensory perception approach with single sensory perception model. And it is demonstrated that with multisensory perception, the robot can understand the instruction more thoroughly. We have also established a user-friendly human-robot interaction interface for the human to send instruction to the robot via a mobile APP. Physical experiments with robotic platform is also conducted. In the future work, we would like to design more complex language instructions and scenarios to improve the generalization ability of the robot.

VIII. ACKNOWLEDGEMENT

This work is supported by National Natural Science Foundation of China (62273054, 62120106005).

REFERENCES

[1] A. Bonarini, "Communication in human-robot interaction," *Current Robotics Reports*, vol. 1, no. 4, pp. 279–285, 2020.

[2] S. Waldherr, R. Romero, and S. Thrun, "A gesture based interface for human-robot interaction," *Autonomous Robots*, vol. 9, no. 2, pp. 151–173, 2000.

[3] D. Mehta, M. F. H. Siddiqui, and A. Y. Javaid, "Facial emotion recognition: A survey and real-world user experiences in mixed reality," *Sensors*, vol. 18, no. 2, p. 416, 2018.

[4] E. Bastianelli, G. Castellucci, D. Croce, R. Basili, and D. Nardi, "Effective and robust natural language understanding for human-robot interaction." in *ECAI*, 2014, pp. 57–62.

[5] H. Khayrallah, S. Trott, and J. Feldman, "Natural language for human robot interaction," in *International Conference on Human-Robot Interaction (HRI)*, 2015.

[6] M. Nazarczuk and K. Mikolajczyk, "V2a-vision to action: Learning robotic arm actions based on vision and language," in *Proceedings of the Asian Conference on Computer Vision*, 2020.

[7] R. Kartmann, D. Liu, and T. Asfour, "Semantic scene manipulation based on 3d spatial object relations and language instructions," in *2020 IEEE-RAS 20th International Conference on Humanoid Robots (Humanoids)*. IEEE, 2021, pp. 306–313.

[8] M. Shridhar, D. Mittal, and D. Hsu, "Ingress: Interactive visual grounding of referring expressions," *The International Journal of Robotics Research*, vol. 39, no. 2-3, p. 027836491989713, 2020.

[9] P. Güler, Y. Bekiroglu, X. Gratal, K. Pauwels, and D. Kragic, "What's in the container? classifying object contents from vision and touch," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2014, pp. 3961–3968.

[10] Y. Wang, K. Wang, Y. Wang, D. Guo, H. Liu, and F. Sun, "Audio-visual grounding referring expression for robotic manipulation," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 9258–9264.

[11] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *2009 IEEE conference on computer vision and pattern recognition*. IEEE, 2009, pp. 1778–1785.

[12] R. Bormann, D. Esslinger, D. Hundsdoerfer, M. Haegele, and M. Vincze, "Texture characterization with semantic attributes: Database and algorithm," in *Proceedings of ISR 2016: 47st International Symposium on Robotics*. VDE, 2016, pp. 1–8.

[13] Y. Sun, L. Bo, and D. Fox, "Attribute based object identification," in *2013 IEEE international conference on robotics and automation*. IEEE, 2013, pp. 2096–2103.

[14] M. Lambeta, P. W. Chou, S. Tian, B. Yang, B. Maloon, V. R. Most, D. Stroud, R. Santos, A. Byagowi, and G. Kammerer, "Digit: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 3838–3845, 2020.

[15] D. K. Misra, J. Sung, K. Lee, and A. Saxena, "Tell me dave: Context-sensitive grounding of natural language to manipulation instructions," *The International Journal of Robotics Research*, vol. 35, no. 1-3, pp. 281–300, 2016.

[16] A. Magassouba, K. Sugiura, A. T. Quoc, and H. Kawai, "Understanding natural language instructions for fetching daily objects using gan-based multimodal target–source classification," *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 3884–3891, 2019.

[17] D. S. Chaplot, K. M. Sathyendra, R. K. Pasumarthi, D. Rajagopal, and R. Salakhutdinov, "Gated-attention architectures for task-oriented language grounding," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[18] Y. Chen, R. Xu, Y. Lin, and P. A. Vela, "A joint network for grasp detection conditioned on natural language commands," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 4576–4582.

[19] R. Paul, J. Arkin, N. Roy, and T. M Howard, "Efficient grounding of abstract spatial concepts for natural language interaction with robot manipulators," 2016.

[20] S. G. Venkatesh, A. Biswas, R. Upadrashta, V. Srinivasan, P. Talukdar, and B. Amrutur, "Spatial reasoning from natural language instructions for robot manipulation," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 11 196–11 202.

[21] T. Nguyen, N. Gopalan, R. Patel, M. Corsaro, E. Pavlick, and S. Tellex, "Robot object retrieval with contextual natural language queries," *arXiv preprint arXiv:2006.13253*, 2020.

[22] V. Blukis, C. Paxton, D. Fox, A. Garg, and Y. Artzi, "A persistent spatial semantic representation for high-level natural language instruction execution," in *Conference on Robot Learning*. PMLR, 2022, pp. 706–717.

[23] W. Yuan, S. Wang, S. Dong, and E. Adelson, "Connecting look and feel: Associating the visual and tactile properties of physical materials," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5580–5588.

[24] A. Drimus, G. Kootstra, A. Bilberg, and D. Kragic, "Design of a flexible tactile sensor for classification of rigid and deformable objects," *Robotics & Autonomous Systems*, vol. 62, no. 1, pp. 3–15, 2014.

[25] V. Chu, I. McMahon, L. Riano, C. G. McDonald, Q. He, J. M. Perez-Tejada, M. Arrigo, T. Darrell, and K. J. Kuchenbecker, "Robotic learning of haptic adjectives through physical interaction," *Robotics and Autonomous Systems*, vol. 63, pp. 279–292, 2015.

[26] H. Liu, D. Guo, and F. Sun, "Object recognition using tactile measurements: Kernel sparse coding methods," *IEEE Transactions on Instrumentation and Measurement*, vol. 65, no. 3, pp. 656–665, 2016.

[27] W. Yuan, C. Zhu, A. Owens, M. A. Srinivasan, and E. H. Adelson, "Shape-independent hardness estimation using deep learning and a gelsight tactile sensor," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 951–958.

[28] J. Li, S. Dong, and E. Adelson, "Slip detection with combined tactile and visual information," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 7772–7777.

[29] F. Veiga, J. Peters, and T. Hermans, "Grip stabilization of novel objects using slip prediction," *IEEE transactions on haptics*, vol. 11, no. 4, pp. 531–542, 2018.

[30] H. Liu, Y. Yu, F. Sun, and J. Gu, "Visual–tactile fusion for object recognition," *IEEE Transactions on Automation Science and Engineering*, vol. 14, no. 2, pp. 996–1008, 2016.

[31] S. Donaher, A. Xompero, and A. Cavallaro, "Audio classification of the content of food containers and drinking glasses," in *2021 29th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 591–595.

[32] D. Gandhi, A. Gupta, and L. Pinto, "Swoosh! rattle! thump!–actions that sound," *arXiv preprint arXiv:2007.01851*, 2020.

[33] R. Gao, Z. Si, Y.-Y. Chang, S. Clarke, J. Bohg, L. Fei-Fei, W. Yuan, and J. Wu, "Objectfolder 2.0: A multisensory object dataset for sim2real transfer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 598–10 608.

[34] Y. Lu, Q. Liu, D. Dai, X. Xiao, H. Lin, X. Han, L. Sun, and H. Wu, "Unified structure generation for universal information extraction," *arXiv preprint arXiv:2203.12277*, 2022.

[35] H. Dong, H. Fang, and N. Zhang, "Multi-scale object detection algorithm for recycled objects based on rotating block positioning," *Journal of ZheJiang University (Engineering Science)*, vol. 56, no. 1, pp. 16–25.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[37] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.