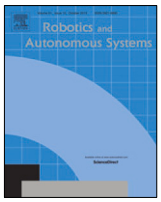
Robotics and Autonomous Systems 159 (2023) 104271

Contents lists available at ScienceDirect

Robotics and Autonomous Systems

journal homepage: www.elsevier.com/locate/robot



Audio–visual language instruction understanding for robotic sorting

Di Guo [a], Huaping Liu [b,*], Fuchun Sun [b]

[a] School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China
[b] Department of Computer Science and Technology, Tsinghua University, Beijing, China

ARTICLE INFO

Article history:
Available online 23 September 2022

MSC:
00-01
99-00

Keywords:
Audio–visual perception
Referring expression
Robotic sorting

ABSTRACT

For robot in human environment, it has always been expected that the robot can execute specified tasks following language instructions. Most current methods only rely on visual perception to understand the language instruction, while it may be not sufficient to fully interpret some language instructions when visually identical objects exist. In this paper, we propose a task of audio–visual language instruction understanding for robotic sorting, in which the robot is able to use both the visual and audio information to fully understand and execute the given instruction. To solve the proposed task, an audio–visual fusion framework is developed, which combines the visual localization and audio recognition models together for the robotic sorting task following language instruction. We have also collected a multimodal dataset for evaluation, and extensive experiments are conducted within the dataset and generalized to new scenarios in physical world demonstrating the effectiveness of the proposed framework.

1. Introduction

Language is a most intuitive interface for human–robot interaction. For robot in human environment, it has always been expected that the robot can execute specified tasks following language instructions, and a type of robotic task that combines the language, perception and action has been proposed, in which the robot should be able to understand the language by grounding the textual words with its perception information and then perform actions to fulfill the task. To assist people's daily life, there are some works investigating to introduce language instructions into robotic manipulation. In this case, it is inevitable that people may use some referring expressions to specify target objects, such as "the blue bottle in the middle", "the mug next to the laptop", etc. Therefore, it is important for the robot to reason from the visual scene and localize the target object given the language instruction for the manipulation.

There have already been a bunch of work proposed for the referring expression comprehension task, which aims at localizing the target object in the image given referring expressions [1]. A common approach is to find the region in the image that matches the referring expression most. Refs. [2,3] use the Baye's rule to rank object proposals given textual description and the one with the highest probability is selected. To handle different types of expression, Ref. [4] proposes to decompose the expression to modular components, which are then fed into different visual modules to compute individual matching scores and then an overall score can be obtained to localize the target. To further improve the semantic understanding of the textual expression and visual reasoning, some graph-based methods [5,6] are proposed for referring expression comprehension which are suitable for recognizing the mentioned object relationship in the language expression.

However, current referring expression comprehension task only depends on visual perception information and is limited to static scenarios, which is not sufficient to fully interpret some language instructions for robotic manipulation. For example, it is difficult for the robot to recognize a bottle with water among several visually identical bottle. In this situation, some other perception modalities can be used to provide complementary information in order to localize the target bottle in the scene. In [7], the robot shakes the container and recognize the content in the container by the produced sound. The sound information provides an alternative for recognizing the target object. Therefore, it is reasonable for the robot to resort to both visual and audio perception information to understand the language instruction.

In this paper, we propose a task of audio–visual language instruction understanding for robotic sorting, in which an instruction is given to the robot, and the robot is able to use both the visual and audio information to fully understand and execute the instruction. This work is inspired from our previous work [8], while a more reasonable and complex dataset for robotic sorting task is collected, and more implementation details and comparison experiments are included. As demonstrated in Fig. 1, given the instruction "Place the bottle with capsule in the left bowl", the

* Corresponding author.
E-mail address: hpliu@tsinghua.edu.cn (H. Liu).

https://doi.org/10.1016/j.robot.2022.104271
0921-8890/

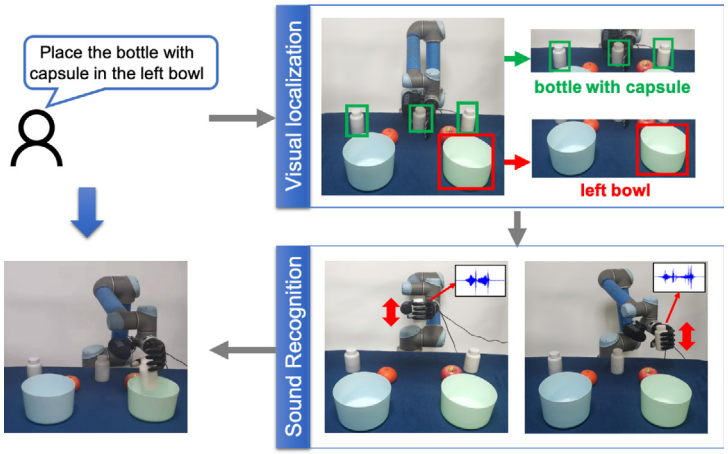


Fig. 1. An illustration of the proposed audio–visual language instruction understanding for robotic sorting. In this case, the user gives the robot a natural language instruction "Place the bottle with capsule in the left bowl". The "left bowl" and all the bottles in the scene can be firstly localized with only visual information. And then the robot manipulates with the bottle to make some sound and the content in the bottle can be recognized by the generated sound. Finally, the target bottle is identified and correctly sorted.

"left bowl" can be firstly localized with only visual information, and the robot localizes the "bottle with capsule" by firstly visually localizing all the bottles in the scene and then identifying the content in the bottle by listening to the sound when manipulating with the bottle. Finally, the target bottle is identified and correctly sorted. The main contributions of the paper are summarized as the following:

- A novel task of audio–visual language instruction understanding for robotic sorting is proposed, in which the robot is required to leverage both audio and visual information to fully understand the given instruction.
- An audio–visual fusion framework is developed, which combines the visual localization and audio recognition models together for the robotic sorting task given language instruction.
- A multimodal dataset which contains the language instructions, visual and audio information is established, and extensive experiments are conducted with the dataset and generalized to the physical world demonstrating the effectiveness of the proposed framework.

2. Related work

With the development of deep learning technologies, the area of robotic manipulation has achieved great progress recently [9,10]. To enable a more intuitive human–robot collaboration, some research begin to introduce the natural language instruction into the robotic manipulation task. To execute the given language instruction correctly, the robot has to have the ability of semantic reasoning [11] and ground the text with its perception information. So it is important for the robot to have the visual grounding ability, and some other perception modalities can also be helpful in certain situations.

A lot of research have been conducted on the visual reasoning based on the language instruction in robotic manipulation. Ref. [12] proposes a multimodal attention model which is able to conduct visual reasoning and then a series of actions are generated to complete the instruction. To extract the object relations specified in the instruction, Ref. [13] proposes a generative model to predict spatial object relations, which is further used for a pick-and-place task. Ref. [14] uses a two-stage model for spatial reasoning on the instruction, in which all the objects are firstly localized in the scene and then together with the language instruction, the pick-up and place positions of the specified objects are generated. As a popular method for robotic manipulation learning, imitation learning can also be used in language-conditioned robotic manipulation. Ref. [15] proposes to incorporate the language instruction as part of demonstration into imitation learning and the learned model is able to generalize to new human instruction. There are also circumstances where the instructions are ambiguous (e.g. the target object is not specified). In [16], a GAN based approach is proposed to predict the target object area in the scene considering the visual scenario and the robotic physical limitation. To eliminate instruction ambiguity, [17,18] further propose an interactive instruction understanding framework where the robot can ask people questions for uncertain understanding and visually ground the referring expressions in the instruction. There are also a kind of vision language navigation task in which the robot navigate by visually grounding the language instructions [19,20]. Additionally, in some embodied question answering tasks, the visual grounding and reasoning capabilities are also highly required. A manipulation question answering task is proposed in which the robot conducts manipulations to interact with the environment based on the visual scene until the answer to the given language question is obtained [21]. In [22], the robot explores and interact with the environment in order to answer the given question. Recently, the commonsense knowledge reasoning is also introduced and proved effective in the human–robot collaboration task [23].

Besides visual information, some other perception modalities can also contribute to the semantic understanding of the environment. An important one is sound modality which can reflect the characteristics of many daily events and objects. A large audio dataset has been collected which is expected to solve the audio event recognition task [24]. And it has been demonstrated that deep learning approaches such as the convolutional neural network can also achieve promising performance in sound understanding tasks [25]. Furthermore, in some circumstance where the objects are visually identical, the sound information provides an efficient alternative for content recognition. To recognize the contents in the bottle, the robot shakes the container to make some sound in order to recognize the contents [7,26]. The sound information can also work together with vision in some embodied navigation task [27–29], where the robot can navigate to a sounding object by leveraging both the sound and visual information. It is noted that the tactile information is also of great significance for semantic object understanding especially in the

robotic manipulation tasks. Ref. [30] proposes to rely on both visual and tactile information for object geometric learning, and Ref. [31] collects tactile data when interacting with the object in order to recognize the content in the container. Although there is more and more research beginning to involve multimodal perception information for semantic environment understanding. The application of multimodal perception for the robotic language understanding tasks are still at a very early stage, where the visual modality still plays a dominant role currently.

3. Problem formulation

The goal of the proposed audio–visual language instruction understanding for robotic sorting task is to enable the robot to understand the given language instruction leveraging both the visual and sound information, and then fulfill the sorting task.

Concretely speaking, the robot is given a language instruction and each target object is described specifically, among which some can be localized in the scene with only visual information, and some may need to be further identified by implementing sound recognition. For example, in the instruction "place the bottle with capsule in the left bowl", the "left bowl" can be localized directly with visual perception, while for the "bottle with capsule", the robot has to visually localize all the bottles in the scene and then manipulate with each bottle producing the sound in order to identify the target bottle. Finally, the robot can execute the sorting task accordingly by using both visual and audio information.

4. Architecture

As is demonstrated in Fig. 2, the architecture of the proposed system is composed of an audio–visual instruction understanding module and a manipulation module. Given a natural language instruction, we resort to the audio–visual instruction understanding module, which firstly tries to visually localize possible target objects in the instruction. And then for objects that are difficult to recognize with only visual information, the sound information is leveraged for recognition. To acquire sound information, the robot performs the actions in the manipulation module and the generated sound is collected. By analyzing the sound information, the robot can identify the target object specified in the instruction. Finally, all target objects in the instruction can be grounded in the scene by using both visual and audio perception. The robot performs the pick and place actions in the manipulation module and the sorting task can be implemented.

5. Method

As shown in Fig. 3, a visual localization model and a sound recognition model are used to achieve the audio–visual instruction understanding. The visual localization model is firstly used to localize target objects that can be identified with visual information, and then for other possible target objects, the robot will perform a series of actions to generate the sound, and the sound recognition model is then used to further identify these objects.

5.1. Visual localization model

Given an image-instruction pair, the visual localization model is used to interpret the specified objects in the instruction and ground the target object with a bounding box. We mainly refer to the method proposed in [32] to localize the referred object in the instruction. We firstly extract the features from both the captured image and given instruction respectively. The multimodal features are then fused semantic understanding and the best-matching region is selected in the image.

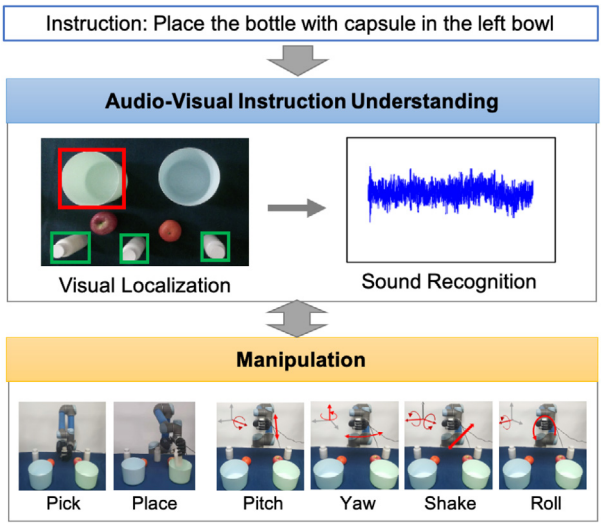


Fig. 2. The architecture of the proposed system, which is composed of an audio–visual instruction understanding module and a manipulation module.

Concretely, for the instruction, we firstly divide the instruction into two parts where one part contains the object to be sorted and the other part contains the place where the object should be placed. A bi-GRU encoder followed by a hierarchical attention module [33] is then used to extract the text feature $f_t \in \mathbb{R}^{d_t}$. In terms of visual features, we use Darknet-53 to extract the hierarchical visual features $f_{v_i}, i = 1, 2, 3$ of the input image.

A hierarchical fusion mechanism is used to obtain the multimodal feature. The text feature is firstly fused with the first level visual feature as:

$$f_{m_1} = \sigma\left(f_t \mathbf{W}_t\right) \odot \sigma\left(f_{v_1} \mathbf{W}_{v_1}\right)$$

where $\mathbf{W}_t$ and $\mathbf{W}_{v_1}$ are projection weight matrices, σ is the Leaky RelU function and $\odot$ denotes the dot-multiplication. For features of other levels, namely $i \in \{2, 3\}$, we conduct a 2×2 upsampling operation and a concatenation operation as:

$$f_{m_{i-1}} = UpSample(f_{m_{i-1}})$$

$$f_{m_i} = [\sigma(f_{m_{i-1}} W_{m_{i-1}}), \sigma(f_{v_i} W_{v_i})]$$

And then, a bottom-up path is used to fuse the above hierarchical multimodal features by conducting a 2×2 downsampling operation and a concatenation operation twice. Finally, a new multimodal feature f'_{m_1} is obtained and fed into the detection network to localize the referring expression (Fig. 3).

5.2. Sound recognition model

We integrate a microphone into the arm of the robot, which is able to collect the sound when the robot is manipulating objects. By analyzing the sound information, we can recognize the class of the object. Inspired by [7], we resort to MFCC feature and a GRU network to process the sound data (Fig. 3). We firstly preprocess the collected sound data by extracting the traditional MFCC features. To extract MFCC features, we firstly use the Hamming window with a window size of 30 ms and a step size of 15 ms to extract MFCC features from the collected sound information, and a vector of 21 Mel coefficients corresponding to the spectrum is used to represent each sound clip. The extracted MFCC features are then fed into a Bi-GRU network and a Softmax classifier is followed to recognize the class of the object. It is noted that in real experimental environment, we also set a threshold to remove inevitable noise that is generated when the robot is running.

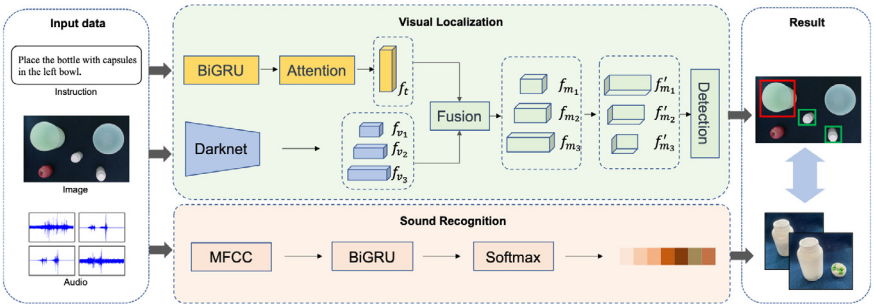


Fig. 3. The structure of the proposed audio–visual framework. The framework is mainly composed of a visual localization model and a sound recognition model.

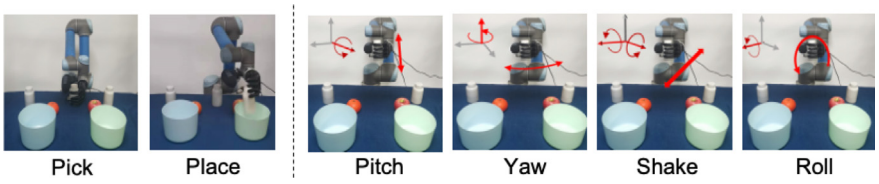


Fig. 4. The actions the robot conducts in the manipulation module.

5.3. Manipulate module

As is shown in Fig. 4, the manipulation module contains an action space that is defined as $A = \{a_{pick}, a_{place}, a_{yaw}, a_{roll}, a_{pitch}, a_{shake}\}$, among which a_{pick} and a_{place} are actions the robot use to execute the sorting task, and $a_{yaw}, a_{roll}, a_{pitch}, a_{shake}$ are four actions that the robot can use to collect sound information. For a_{pick} action, we resort to our previous work [34] to detect the grasping point when the target object is localized by the visual localization model. After the grasping point is detected, the robotic arm moves to the grasping point following the build-in motion planning algorithm, and the object is grasped in a configuration where four fingers are opposed to the thumb. For the $a_{yaw}, a_{roll}, a_{pitch}, a_{shake}$ actions, the robot follows a predefined motion (Fig. 4) with an angular velocity of 3.14 rad/s.

Specifically, when the robot wants to collect the sound information of a possible target object, the robot firstly picks up the object, and a series of actions are then conducted to generate the sound information for recognition. In this paper, the robot performs the four actions $a_{\text{yaw}}, a_{\text{roll}}, a_{\text{pitch}}, a_{\text{shake}}$ in sequence. After all the target objects are localized in the scene, the robot will picks up the target object and place it in the target location fulfilling the sorting task.

6. Dataset

6.1. Robotic platform

An overview of the robotic platform is demonstrated in Fig. 5. The platform is mainly composed of a UR5 robotic arm, a five-finger robotic hand, an operating table, a microphone and a Kinect camera.

The UR5 robotic arm is of 6 degree of freedom and a self-developed five-finger robotic hand is attached to the end of the arm. This arm-hand system is used to execute the required manipulation actions. An operating table is placed in front of the robotic arm on which there are some objects and bottles that need to be sorted in specified bowl by the robot. And a Kinect camera is placed on the top of the table top capturing the visual information of the scene. Additionally, a commercial microphone is equipped on the robotic arm besides the robotic hand, which is able to collect the generated sound when the robotic hand is manipulating the objects.

6.2. Auditory dataset

As is shown in Fig. 6, we consider 11 types of common medicine that are usually stored in the bottle, namely capsule, alcohol, red data, tablet, hawthorn, pill, seman cassiae, oyster, wax pill, cicada slough, particle, and an empty bottle without anything in it. As these bottles are visually the same, it is difficult for the robot to identify what is in the bottle visually. However, these medicines are actually with different physical characteristics, and thus different sound can be generated when manipulating with the bottles.

To collect the sound data of different contents, we integrate a microphone into the arm of the robot, and collect the generated sound when the robot is manipulating with the bottle. We employ the yaw, roll, pitch, and shake actions for the robot to manipulate with the bottle so that sufficient sound information of each content is collected. Considering that the amount of the object in the bottle may affect the sound generated, the bottles are filled with 1/4, 2/4, 3/4 of its capacity respectively for each content. For 11 different contents of different weight and the empty bottle, the bottle is firstly grasped by a five-finger robotic hand with a fixed configuration, in which the thumb and the four fingers are opposed. And then the robot performs the yaw, roll, pitch, and shake actions with an angular velocity of 3.14 rad/s in sequence. For each bottle, a six-second audio is recorded. Altogether, the bottles are manipulated 20 times for each action. The visualization of some sound waves generated by different actions are demonstrated in Fig. 7. We find that different contents reveal different sound characteristics and the sound generated by different actions are also different. We also notice that for different objects, they may generate very similar sound waved under certain actions. Additionally, for the same object, when the weight is different, the sound generated is also different. Therefore, we believe that a combination of sound information generated under all actions yields best sound characteristics.

6.3. Manipulation instruction design

As is shown in Fig. 8, some bottles with different contents and fruits are placed on the table. Two bowls used for sorting are placed in front. The items are randomly placed and the robot is required to sort the bottles and the fruits in different bowls following the instructions. Seven types of manipulation instructions

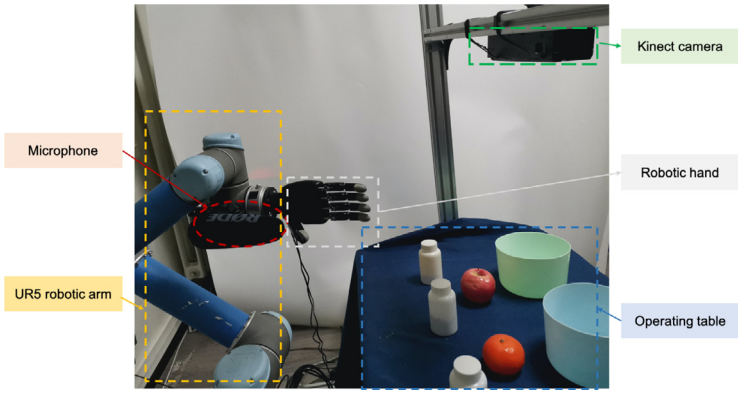


Fig. 5. The platform is mainly composed of a UR5 robotic arm, a five-finger robotic hand, an operating table, a microphone and a Kinect camera.

Fig. 6. Different types of medicines that are stored in the bottle.

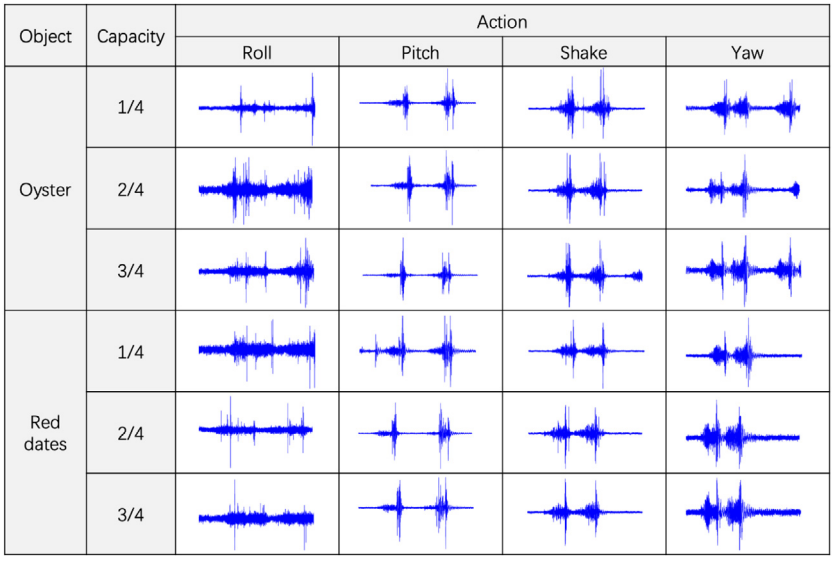


Fig. 7. Sound waves visualization.

Table 1
Templates to generate the manipulation instructions.

Template
Place the bottle with <content> in the <location> bowl.
Place the <location> bottle in the <location> bowl.
Place the bottle <relation> in the <location> bowl.
Place the bottle with <weight><content> in the <location> bowl.
Place the <location><object> in the <location> bowl.
Place the <obj><relation> in the <color> bowl.
Place the <color><object> in the <location> bowl.

are generated as shown in Table 1. And different types of referring expressions are used in the instruction. For ⟨*content*⟩, it denotes different objects in the bottle. For ⟨*object*⟩, it denotes the type of fruits on the tabletop. ⟨*location*⟩ and ⟨*relation*⟩ are used to denote the specific object in which ⟨*location*⟩ can be replaced as "right" or "left" and ⟨*relation*⟩ can be replaced by on the right of ⟨*object*⟩, on the left of ⟨*obj*⟩, or next to ⟨*object*⟩. ⟨*color*⟩ and ⟨*weight*⟩ provide the color and weight information to specify one object.

To interpret the referring expression in the instruction, different perception modalities are required. For instructions that only contain fruits, only visual expression referring is enough to interpret the instruction, while for the instructions that require to sort

Fig. 8. Sorting scenarios.

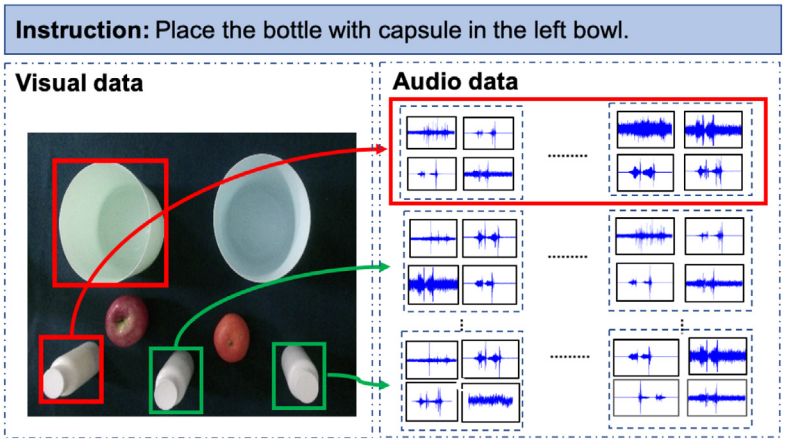


Fig. 9. Multimodal dataset. The dataset is composed of the visual information of the scene, the sound data of bottle with different contents, and corresponding manipulation instruction.

bottles, two circumstances can exist. For the one that specifies the bottle with ⟨*relation*⟩ or ⟨*location*⟩ referring expressions, only visual referring expression is employed. And for the one which specifies the ⟨*content*⟩ and ⟨*weight*⟩ of the bottle, both visual and audio referring expression are necessary to understand the instruction.

To generate the instructions, two bowls are placed on the table, and the 0–3 bottles and 0–3 fruits are also randomly placed on the table. The scenes are captured by the Kinect camera. According to the template and complexity of the scene, the instructions are generated based on the scene. In total, we have collected 180 scene images and 8998 instructions are generated.

6.4. Summary

In summary, the dataset is composed of the visual information of the scene, the sound data of the object, and corresponding manipulation instruction. Fig. 9 gives an intuitive illustration of the collected dataset. For the given scene, the bottles and bowls are placed on the table. To obtain the visual information of the scene, the scene is captured from the table with a Kinect camera. Some manipulation instructions are associated with the given scene. For each bottle, we manipulate it with all the four actions to collect the sound data. It is noted there are groups of such sound data for each bottle. All of the above information composes the multimodal dataset.

7. Experimental results

We conduct experiments both with the collected dataset and generalize the trained model to physical robot to verify the performance of the proposed framework. We firstly resort to the collected multimodal dataset to evaluate the performance of the trained sound recognition model and the visual localization model. And then, a physical robot is used to implement the sorting task in the real-world environment with the trained model.

7.1. Experiment with dataset

As is shown in Fig. 9, the collected multimodal dataset contains the visual, audio, and instruction information for each scenario. Given an instruction and a scenario, we can obtain the sound information for each bottle in the scenario from a pool of collected audio data. We also have the ground truth location information in the scene for all the target objects specified in the instruction. Therefore, it is convenient for us to evaluate the performance of the proposed visual localization and sound recognition model with the collected dataset.

7.1.1. Audio preprocessing

In the collected dataset, 11 different objects of different weight and the empty bottle have been manipulated by the four actions for 20 times, and a six-second audio is recorded for each. We first divide the data into the training and test set with a ratio of 5:1. Then to augment the data, we further divide each audio clip into 5 segments uniformly. And then, with the augmented data, for each object, we randomly select one segment from each action and concatenate segments from four actions to obtain a new combined audio. It is noted that in the combined audio, it still contains the sound generated by all the four actions.

Table 2
Sound recognition results.

	Pitch		Yaw		Roll		Shake		All Actions	
	GRU	VGGish	GRU	VGGish	GRU	VGGish	GRU	VGGish	GRU	VGGish
Alcohol	0.560	0.300	0.620	0.320	0.780	0.420	0.720	0.160	0.760	0.520
Capsule	0.700	0.340	0.620	0.440	0.820	0.640	0.780	0.460	0.900	0.580
Cicada Slough	0.620	0.420	0.740	0.360	0.760	0.520	0.740	0.380	0.700	0.580
Empty	0.400	0.620	0.880	0.300	0.760	0.620	0.700	0.420	0.700	0.740
Hawthorn	0.620	0.280	0.580	0.280	0.700	0.520	0.820	0.520	0.900	0.560
Oyster	0.580	0.420	0.760	0.240	0.700	0.420	0.800	0.440	0.860	0.740
Particle	0.660	0.320	0.560	0.340	0.800	0.540	0.860	0.460	0.800	0.740
Pill	0.640	0.720	0.820	0.620	0.800	0.720	0.720	0.640	0.940	0.860
Red Dates	0.680	0.460	0.820	0.500	0.800	0.500	0.520	0.400	0.880	0.720
Seman Cassiae	0.820	0.560	0.920	0.600	0.840	0.740	0.880	0.760	0.940	0.940
Tablet	0.800	0.660	0.820	0.580	0.800	0.700	0.840	0.380	0.940	0.780
Wax Pill	0.220	0.480	0.880	0.500	0.280	0.340	0.640	0.520	0.700	0.540
Average	**0.608**	0.465	**0.752**	0.423	**0.737**	0.530	**0.752**	0.462	**0.835**	0.692

7.1.2. Sound recognition experiment

To evaluate the performance of the proposed sound recognition model, we have conducted extensive experiments considering the collecting actions. With the collected sound data, we have trained two sound recognition models, in which one is from each action and one is from all actions. Also, the VGGish model [25] is finetuned with the collected sound data which acts as a baseline.

Table 2 demonstrates the sound recognition results in different situations. It can be seen that the average accuracy obtained from all actions achieves 0.835, which is better than any single actions. It is reasonable as multiple actions can collect more information and thus more thoroughly revealing the characteristics of the contents. We also notice that for some specific objects, certain separate action has better performance than the result obtained from all actions. For example, for the object Particle, the Shake action yields a higher accuracy, and for the object Wax Pill, the Yaw action yields a higher accuracy. However, the average accuracy of all actions is obviously better than single action. Generally, the sound information is proved effective in identifying the contents in the identical bottles. And manipulating the bottle with one action is also helpful for sound recognition though in some cases is not as good as with all actions.

We further conduct the sound recognition experiments for objects of different weight. Similarly, both the proposed GRU and VGGish models are trained with the data with a finer categories. The results are shown in Table 3. It can be seen that on average, GRU model performs better than VGGish model and the sound recognition results become worse as the capacity increases. It is reasonable as the more full the bottle is, the weaker the content generates the sound when manipulating. However, it can be noticed that an overall recognition accuracy for different weights is not as good as those for content recognition. We think that it is because that the sound difference for different weights may not be obvious and the operation noise of the robot could also affect the performance of the sound recognition model. In the future work, besides the sound information, we would like to also try the tactile perception that is more sensitive to weights to recognize the weight of the content in the bottle.

7.1.3. Visual localization verification experiment

We verify the performance of the visual localization model with the collected scene images and generated manipulations. As illustrated in Table 1, the manipulation instruction is composed of two kinds of the referring expressions. The first half of the instruction refers to the target object that needs to be sorted. And the second half of the instruction specifies where to place the target object. Either half of the instruction together with the corresponding scene image is fed into to visual localization model. The results are demonstrated in Fig. 10. It can be seen that in most cases, the referred target object can be correctly localized in the scene image. Especially, in Fig. 10(e) image, the bottle is described by ⟨*location*⟩ which could be recognized from merely visual information. However, there also exist some failure cases. For example, in Fig. 10(f), given the instruction, the robot is firstly expected to localize the bottle with visual information, while it mistakenly recognizes the left bowl as the bottle due to the imperfection of the visual detection model. We believe a stronger visual detection model could help to avoid such failure cases.

Table 3
Sound recognition for different weights.

	1/4 capacity		2/4 capacity		3/4 capacity	
	GRU	VGGish	GRU	VGGish	GRU	VGGish
Alcohol	0.250	0.100	0.275	0.300	0.363	0.125
Capsule	0.613	0.313	0.163	0.088	0.375	0.175
Cicada Slough	0.500	0.338	0.338	0.113	0.588	0.125
Empty	0.413	0.400	0.413	0.400	0.413	0.400
Hawthorn	0.513	0.163	0.388	0.213	0.338	0.263
Oyster	0.563	0.213	0.250	0.150	0.213	0.325
Particle	0.425	0.263	0.338	0.150	0.450	0.188
Pill	0.600	0.325	0.475	0.350	0.200	0.100
Red Dates	0.613	0.313	0.563	0.263	0.250	0.088
Seman Cassiae	0.600	0.363	0.750	0.350	0.788	0.350
Tablet	0.688	0.375	0.500	0.138	0.005	0.163
Wax Pill	0.300	0.750	0.038	0.138	0.125	0.005
Average	**0.507**	0.326	**0.374**	0.221	**0.342**	0.192

We test the visual localization model on all the instructions. To evaluate the referring expression comprehension task, we use the same metrics as introduced in [2]. We compute the intersection over union (IoU) between predicted box given the instruction and the ground truth box. If the IoU is larger than 0.5, we consider the model successfully localizes the referred object. Among all the 8998 instructions, referring expressions in 1726 instructions could be recognized with only visual information. And the rest instructions have to rely on both visual and audio information to localize the objects indicated in the instruction. With only visual input, a visual localization accuracy of 70.05% is achieved in circumstances where the objects referred in the instruction can be identified with only visual information, and an accuracy of 50% is obtained in circumstances where both visual and audio information are necessary to identify the objects in the instruction. Therefore, in the former situation, we could have a higher visual localization accuracy, and in the latter situation, since the model could only recognize where to place the target but not recognize any target object, a lower accuracy is obtained.

7.2. Physical experiment

To further evaluate performance of the proposed framework, we generalize the trained model to a physical robot to implement

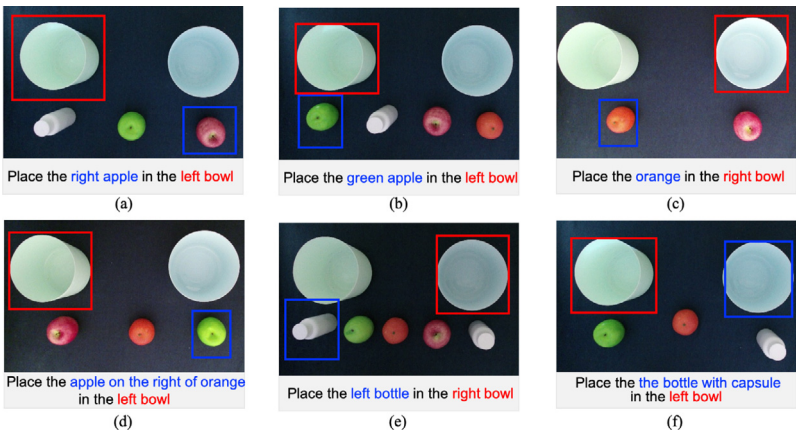


Fig. 10. The results of the visual localization model for the robotic sorting task. The blue box indicates the bounding box of the target object and the red box indicates where to place the target object. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

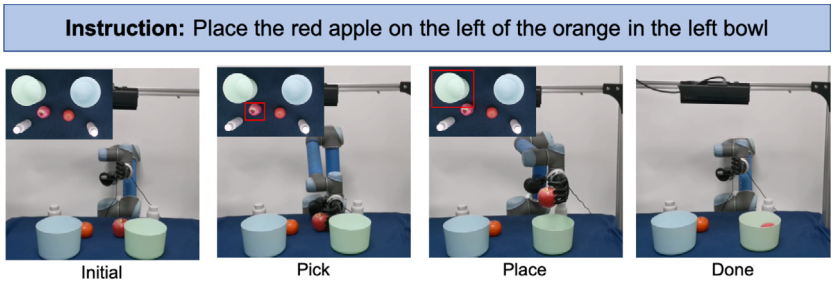


Fig. 11. With the visual information, the robot is able to pick up the target object and place it in the target bowl.

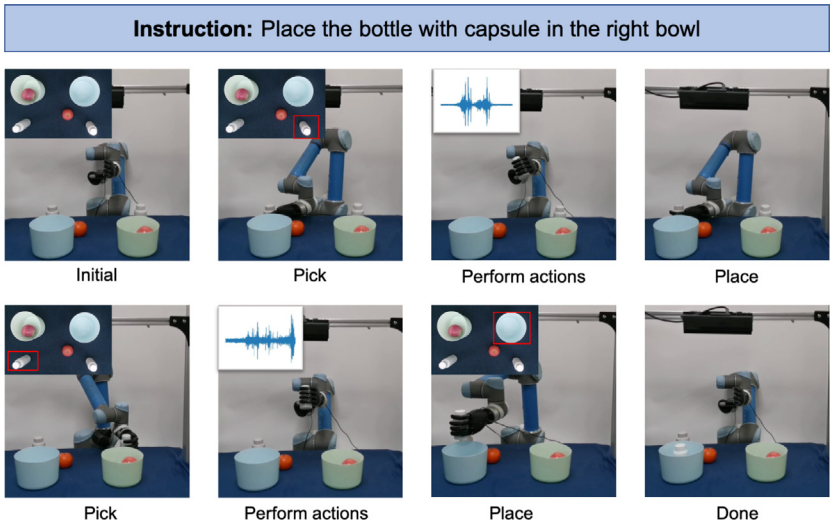


Fig. 12. With both the visual and audio information, the robot is able to firstly pick up each bottle according to the visual information. And then the robot manipulates with the bottle to generate the sound and a sound recognition module is activated. When the target object is recognized, it will be placed in the target bowl.

the sorting task in the real-world environment. We consider two circumstances in the sorting task. The one is that the system can execute the sorting task given the instruction with only the visual information. The other is the one that both visual and audio information are required in order to accomplish the given instruction.

In Fig. 11, the initial scene contains two bottles, an apple and an orange. The given instruction requires the robot to *"Place the red apple on the left of the orange in the left bowl"*. The language instruction and captured scene image are firstly fed into the proposed visual localization model. For the target object, the apple is detected and the robot picks the apple. Also, the target place is also detected from the image and the robot can place the target object in the target place as the instruction indicates.

And then, another instruction is given that requires the robot to *"Place the bottle with capsule in the right bowl"* which is shown in Fig. 12. Similarly, the language instruction and captured scene image are firstly fed into the proposed visual localization model. However, the visual localization model can only detect two bottle while it cannot decide which bottle is the target bottle. Therefore,

the robot will firstly pick up one bottle and then a series of actions are performed to generate the sound of the object. With the collected audio information, the audio recognition module will be activated to recognize the category of the object. If it is not the target object, the robot will place it back, and pick up the other bottle. The same audio recognition process is conducted again. In this example, the first bottle picked up is not the target one and the second bottle is the correct one. Finally, the robot places the correct bottle in the specified target bowl.

It can be seen from the above experiments that the trained model can be generalized to new scenarios with real robots. For the sorting task, the visual information can provide essential information to interpret the language instruction. However, for some visually indistinguishable objects, such as the bottles in our experiments, it is not enough to rely on only visual information to fully understand the instruction. And the audio information acts as an important complementary role in this situation, which validates the effectiveness of the proposed audio–visual framework.

8. Conclusion

Considering the fact that referring expressions are commonly used when people specify some particular objects in their daily conversation, and some visually indistinguishable objects actually have different sound characteristics, we develop a novel task of audio–visual language instruction understanding for robotic sorting in this paper. The robot is able to use both the visual and audio information to fully understand and execute the given instruction. And an audio–visual fusion framework is proposed for both visual localization and sound recognition. Also, a new dataset composed of the visual information, audio information, and corresponding manipulation instruction is established. Experiments are conducted with both the collected dataset and generalized to real robot demonstrating that by leveraging both the visual and audio data, the robot is able to better understand the instruction than with only the visual data in new scenarios. For the future work, we would like to design more complex manipulation instructions and scenarios, and an end-to-end framework will be further investigated.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by National Natural Science Foundation of China under Grant 62273054 and Joint Fund of Science & Technology Department of Liaoning Province and State Key Laboratory of Robotics, China (2020-KF-22-06).

References

[1] Y. Qiao, C. Deng, Q. Wu, Referring expression comprehension: A survey of methods and datasets, IEEE Trans. Multimed. 23 (2020) 4426–4440.
[2] J. Mao, J. Huang, A. Toshev, O. Camburu, A.L. Yuille, K. Murphy, Generation and comprehension of unambiguous object descriptions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 11–20.
[3] L. Yu, P. Poirson, S. Yang, A.C. Berg, T.L. Berg, Modeling context in referring expressions, in: European Conference on Computer Vision, Springer, 2016, pp. 69–85.
[4] L. Yu, Z. Lin, X. Shen, J. Yang, X. Lu, M. Bansal, T.L. Berg, Mattnet: Modular attention network for referring expression comprehension, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1307–1315.
[5] P. Wang, Q. Wu, J. Cao, C. Shen, L. Gao, A.v.d. Hengel, Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 1960–1968.
[6] S. Yang, G. Li, Y. Yu, Dynamic graph attention for referring expression comprehension, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 4644–4653.
[7] E. Strahl, M. Kerzel, M. Eppe, S. Griffiths, S. Wermter, Hear the egg-demonstrating robotic interactive auditory perception, in: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, IEEE, 2018, p. 5041.
[8] Y. Wang, K. Wang, Y. Wang, D. Guo, H. Liu, F. Sun, Audio–visual grounding referring expression for robotic manipulation, 2021, arXiv preprint arXiv:2109.10571.
[9] D. Zhang, Q. Li, Y. Zheng, L. Wei, D. Zhang, Z. Zhang, Explainable hierarchical imitation learning for robotic drink pouring, IEEE Trans. Autom. Sci. Eng. (2021) 1–17, http://dx.doi.org/10.1109/TASE.2021.3138280.
[10] Y. Laili, Z. Chen, L. Ren, X. Wang, M.J. Deen, Custom grasping: A region-based robotic grasping detection method in industrial cyber-physical systems, IEEE Trans. Autom. Sci. Eng. (2022) 1–13, http://dx.doi.org/10.1109/TASE.2021.3139610.
[11] J. Savage, D.A. Rosenblueth, M. Matamoros, M. Negrete, L. Contreras, J. Cruz, R. Martell, H. Estrada, H. Okada, Semantic reasoning in service robots using expert systems, Robot. Auton. Syst. 114 (2019) 77–92.
[12] M. Nazarczuk, 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.
[13] R. Kartmann, D. Liu, 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.
[14] S.G. Venkatesh, A. Biswas, R. Upadrashta, V. Srinivasan, P. Talukdar, B. Amrutur, Spatial reasoning from natural language instructions for robot manipulation, in: 2021 IEEE International Conference on Robotics and Automation, ICRA, IEEE, 2021, pp. 11196–11202.
[15] S. Stepputtis, J. Campbell, M. Phielipp, S. Lee, C. Baral, H. Ben Amor, Language-conditioned imitation learning for robot manipulation tasks, Adv. Neural Inf. Process. Syst. 33 (2020) 13139–13150.
[16] A. Magassouba, K. Sugiura, H. Kawai, A multimodal classifier generative adversarial network for carry and place tasks from ambiguous language instructions, IEEE Robot. Autom. Lett. 3 (4) (2018) 3113–3120.
[17] M. Shridhar, D. Mittal, D. Hsu, INGRESS: Interactive visual grounding of referring expressions, Int. J. Robot. Res. 39 (2–3) (2020) 217–232.
[18] H. Zhang, Y. Lu, C. Yu, D. Hsu, X. La, N. Zheng, INVIGORATE: Interactive visual grounding and grasping in clutter, 2021, arXiv preprint arXiv:2108.11092.
[19] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, A. Van Den Hengel, Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3674–3683.
[20] Y. Qi, Q. Wu, P. Anderson, X. Wang, W.Y. Wang, C. Shen, A.v.d. Hengel, Reverie: Remote embodied visual referring expression in real indoor environments, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9982–9991.
[21] Y. Deng, X. Guo, N. Zhang, D. Guo, H. Liu, F. Sun, MQA: Answering the question via robotic manipulation, 2020, arXiv preprint arXiv:2003.04641.
[22] D. Gordon, A. Kembhavi, M. Rastegari, J. Redmon, D. Fox, A. Farhadi, Iqa: Visual question answering in interactive environments, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4089–4098.
[23] C.J. Conti, A.S. Varde, W. Wang, Human-robot collaboration with commonsense reasoning in smart manufacturing contexts, IEEE Trans. Autom. Sci. Eng. (2022) 1–14, http://dx.doi.org/10.1109/TASE.2022.3159595.
[24] J.F. Gemmeke, D.P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R.C. Moore, M. Plakal, 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.
[25] S. Hershey, S. Chaudhuri, D.P. Ellis, J.F. Gemmeke, A. Jansen, R.C. Moore, M. Plakal, D. Platt, R.A. Saurous, B. Seybold, et al., CNN architectures for large-scale audio classification, in: 2017 Ieee International Conference on Acoustics, Speech and Signal Processing (Icassp), IEEE, 2017, pp. 131–135.
[26] S. Jin, H. Liu, B. Wang, F. Sun, Open-environment robotic acoustic perception for object recognition, Front. Neurorobot. 13 (2019) 96.
[27] C. Gan, Y. Zhang, J. Wu, B. Gong, J.B. Tenenbaum, Look, listen, and act: Towards audio–visual embodied navigation, in: 2020 IEEE International Conference on Robotics and Automation, ICRA, IEEE, 2020, pp. 9701–9707.
[28] C. Chen, U. Jain, C. Schissler, S.V.A. Gari, Z. Al-Halah, V.K. Ithapu, P. Robinson, K. Grauman, Soundspaces: Audio-visual navigation in 3d environments, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16, Springer, 2020, pp. 17–36.

[29] C. Chen, Z. Al-Halah, K. Grauman, Semantic audio–visual navigation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 15516–15525.
[30] D. Watkins-Valls, J. Varley, P. Allen, Multi-modal geometric learning for grasping and manipulation, in: 2019 International Conference on Robotics and Automation, ICRA, IEEE, 2019, pp. 7339–7345.
[31] P. Güler, Y. Bekiroglu, X. Gratal, K. Pauwels, 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.
[32] G. Luo, Y. Zhou, X. Sun, L. Cao, C. Wu, C. Deng, R. Ji, Multi-task collaborative network for joint referring expression comprehension and segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10034–10043.
[33] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, E. Hovy, Hierarchical attention networks for document classification, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016, pp. 1480–1489.
[34] D. Guo, H. Liu, B. Fang, F. Sun, W. Yang, Visual affordance guided tactile material recognition for waste recycling, IEEE Trans. Autom. Sci. Eng. (2021).

Di Guo received Bachelor degree from the Department of Instrumentation Science and Optoelectronics Engineering, Beihang University, Beijing, in 2011. She received her Ph.D degree in the Department of Computer Science and Technology, Tsinghua University, Beijing, China. Her research interests include robotic manipulation and sensor fusion.

Huaping Liu is currently an Associate Professor with the Department of Computer Science and Technology, Tsinghua University, Beijing, China. His research interests include robot perception and learning. Dr. Liu served as a Senior Program Committee Member for International Joint Conference on Artificial Intelligence 2018. He was a recipient of the Andy Chi Best Paper Award in 2017. He served as the Area Chair for Robotics Science and Systems 2018. He serves as an Associate Editor for some journals, including the IEEE TRANSACTIONS ON AUTOMATION SCIENCE AND ENGINEERING, the IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, the IEEE ROBOTICS AND AUTOMATION LETTERS, Neurocomputing, and Cognitive Computation, and some conferences, including International Conference on Robotics and Automation and IEEE/RSJ International Conference on Intelligent Robots and Systems.

Fuchun Sun is currently a Full Professor with the Department of Computer Science and Technology, Tsinghua University, Beijing, China. His research interests include intelligent control and robotics.

Prof. Sun was a recipient of the National Science Fund for Distinguished Young Scholars. He serves as the Editor-in-Chief for Cognitive Computation and Systems and an Associate Editor for a series of international journals, including the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS:SYSTEMS, the IEEE TRANSACTIONS ON FUZZY SYSTEMS, Mechatronics, and Robotics and Autonomous Systems.