Robot Assisting Water Serving to Disabilities by Voice Control

Yang Chunxin
MIST, Kyushu Institute of Technology, 680-4 Kawazu
Iizuka-shi, Fukuoka 820-8502, Japan
Sakmongkon Chumkamon
MIST, Kyushu Institute of Technology, 680-4 Kawazu
Iizuka-shi, Fukuoka 820-8502, Japan
Eiji Hayashi
MIST, Hayashi Lab, 680-4 Kawazu
Iizuka-shi, Fukuoka 820-8502, Japan
E-mail: chunxin.yang215@mail.kyutech.jp, m-san@mmcs.mse.kyutech.ac.jp
haya@mse.kyutech.ac.jp
www.kyutech.ac.jp

Abstract

ROS is an open-source robot operating system. In this paper, we use ROS to control Conbe robot arm. By introducing the YOLACT real-time instance segmentation, we trained our own model for Object Detection. Secondly, the Speech-Recognition system is established through Deep speech and Mozilla Text-To-Speech with Tacotron2 DDC model. Deep speech is an end-to-end speech system, where deep learning supersedes these processing stages. Combined with a language model, this approach achieves higher performance than traditional methods on hard speech recognition tasks while also being much simpler. In this way, we create an artificial intelligence, which accomplished a simple conversation with people. And the voice control system is established based on Speech-Recognition system. In the experiment, we successfully control the robot arm move positions and do water serving for disabilities by voice command. With this research, voice control robot arm can be apply in the life support area, it will be more convenient for disabilities in daily life.

Keywords: ROS, Water Serving, Disabilities, YOLACT, Speech-Recognition, voice control, Deepspeec

1 Introduction

1.1 Background

Robot arm has become an important role in industrial production and in people's life using the interplay of robot technology and information technology.

In this research, the robot arm is able to communicate with people accomplished by using Deepspeech and Mozilla Text-To-Speech with Tacotron2 DDC model. Communication with the arm and serving water for those who can't drink water on their own by voice control become possible.

1.2 Purpose of Research

（a）Multipurpose

Voice control robot arms can be applied to home services, commercial services and industrial production.

（b）Can be used in the life support area

The miniaturized robot arm can be used for the human body equipment, and the simple service is possible for the disabled people. By grasping objects using voice control, such as grasping cups to serve water,it will be more convenient for these people in daily life.

2 System Configuration

The system consists of four parts:

（a）Machine Vision

Using YOLACT Real-time Instance Segmentation to train dataset and detect the object that we need.

（b）Speech Recognition

By using Deepspeech and Mozilla Text-To-Speech, we create our speech dataset and trained the speech model, which successfully created an artificial intelligence accomplished a simple conversation with people and control robot arm to move.

（c）ROS And MoveIt

Using ROS and MoveIt to set robot's posture (position and angle). Using Moveit trac-ik to solve inverse-kinematics problem for robot.

(d) Conbe Robot Arm

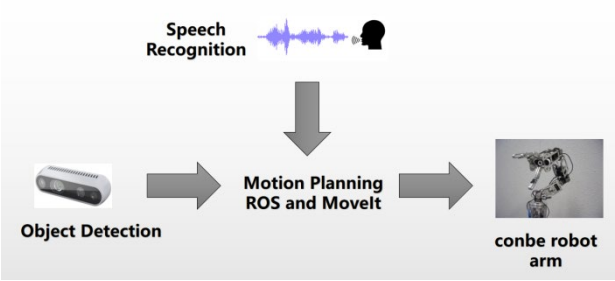

Fig.1 System configuration

3 Machine Vision

3.1 Object Detection

(a) Data collection

In order to detect an object accurately, a large amount of data are necessary. Therefore, we need to collect data by photographing each object and each object needs to take at least 100 pictures from different background, colors, angles and distance.

Then, the object will be labeled with the measurement range and the object name.

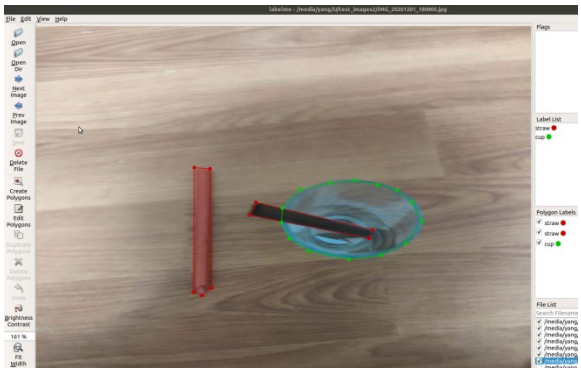

Fig.2 Label The Object

(b) Training with YOLACT

YOLACT is a simple, fully-convolutional model for real-time instance segmentation that achieves 29.8 mAP on MS COCO at 33.5 fps evaluated on a single Titan Xp, which is significantly faster than any previous competitive approach[1].

We trained date using the Resnet101-FPN model, and complete training until the loss is minimized.

(c) Program

The object detection program uses a case program on YOLACT which is eval.py with trained data[2].In order to do water serving, we need to let robot know the position of object, therefore, we add coordinates on the target using opencv centroid, calculate the coordinates of the center point from the coordinates of the bounding box. But in this way we just acquire 2-dimensional coordinates on the target. The real coordinates of the target point are obtained by recognizing the pixel coordinates and depth values of the target.

$$\frac{target\_xy\_pixel[0]-ppx}{fx}=\frac{target\_xy\_true[0]}{target\_depth}$$

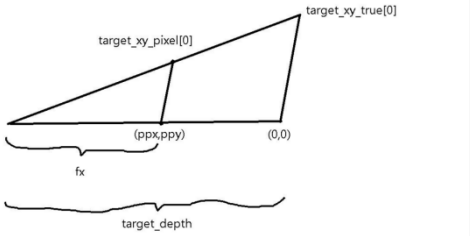

Fig.3 Convert Pixel Coordinates To Depth Value

In the figure3 ,(0,0) is the point position of object target in RGB image, and get the parameter of ppx,ppy,fx,fy from camera internal parameters. In this way, we can acquire 3-dimensional coordinates of the target. Then, we publish markers on this 3-dimensional coordinates using ROS.

3.2 Test

In the test,it can accurately identify the short distance and far distance(2 meter) for cup and straw with 30FPS.The pixel coordinates are displayed in the middle of the object.We can get the real world 3D coordinates based on this pixel coordinates.

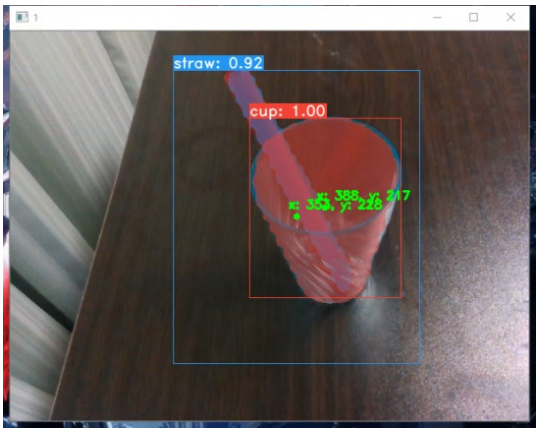

Fig.4 Object Detection Test

4 Speech Recognition

4.1 Deepspeech

Deepspeech is an end-to-end speech system, where deep learning supersedes these processing stages. Combined with a language model, this approach achieves higher performance than traditional methods on hard speech recognition tasks while also being much simpler[3].

We using Deepspeech pre-trained model and add our own database to fine-tune the model, then trained with tensor flow. Currently, we use the Deepspeech 0.9.1 pre-trained model with English alphabet.

(a) Database

For this research, we trained 20 words and sentences that we need. For audio file, we need the audio in Mono 16K with .Wav format[4].

	A	B	C	D	E
1	wav_filena	wav_filesiz	transcript		
2	01.wav	135244	move to home position		
3	02.wav	145228	pick and place		
4	03.wav	180556	please move to home position		
5	04.wav	148482	move to ready position		
6	05.wav	145410	please move to ready position		
7	06.wav	126210	return and stand by		
8	07.wav	94722	michael		
9	08.wav	129282	michael are you there		
10	09.wav	126978	please pick the object		
11	10.wav	84738	pick		
12	11.wav	108546	place		
13	12.wav	106242	please		
14	13.wav	88578	hello		
15	14.wav	84,738	water		
16	15.wav	91,650	drink		
17	16.wav	142,338	i want to drink water		
18	17.wav	110,850	water please		
19	18.wav	151,554	please give me water		
20	19.wav	121,602	put it back		
21	20.wav	96,258	put		

Fig.5 Database of Deepspeech

4.2 Mozilla TTS

Mozilla TTS is a deep learning based Text To Speech project.We use the TTS pre-trained model,Tacotron2 Double Decoder Consistency for our system to get response from our speech.Tacotron2 is a neural network architecture for speech synthesis directly from text and it faster than real-time inference performance[5].

4.3 Speech Recognition Program

To recognize people's language correctly :

(a) The language need to be record and saved as an audio file.

(b) Identify text from the audio files.

(c) Then select the corresponding answer text based on the recognized text.

(d) Convert selected answers to voice.

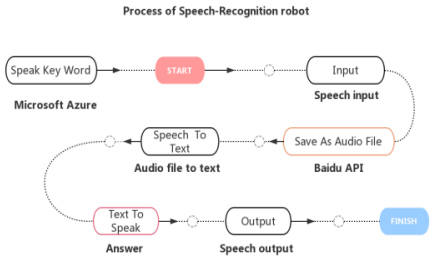

Fig.6 Construction Of Speech Conversation Program

4.4 Speech Recognition Test

In the test, we can have a simple conversation with this AI, it can recognize our words correctly and response based on our words. For example, If we say "pick and place", it will print the response and answer" right away".For trained words and sentences, the correct rate is over 90%.

4.5 Voice control

Voice control for robot use this program:

"os.system("gnome-terminal -e 'rosrun conbe_moveit_config moveit_fk_demo.py'")"

This program will run the python scripts, in this python file the pose and movement of robot is defined. When we speak the command , it will run the corresponding python scripts.

5 Experiment

5.1 Simulation With MoveIt

Moveit is a Motion Planning Framework for manipulators. The main use is calculate the path to the specified position and angle of the gripper. We use voice control such as saying"please give me water"to call Python scripts and then control robot to serve water.

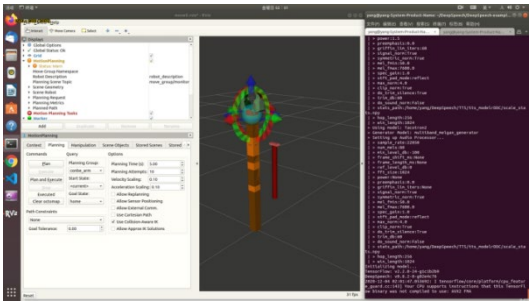

Fig.8 Setting Robot Position Using Voice Control

5.2 Water Serving With Real Robot

The Water Serving program uses MoveIt Command Interface. By setting the joint position and get target position from YOLACT, Moveit will calculate the path and inverse-kinematics for robot planning group. Robot will first pick straw into the cup and then bring cup to people.

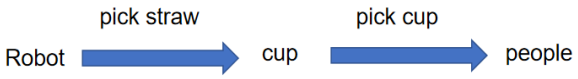

Fig.9 Water Serving

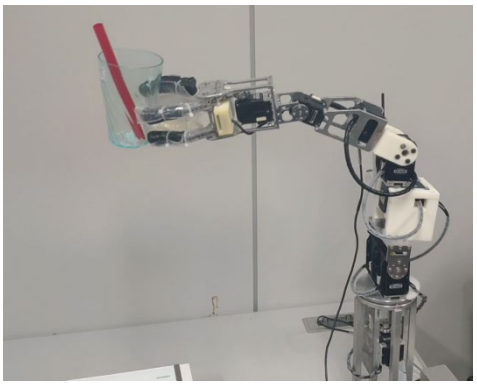

Fig.10 Pick Cup With Real Robot

6 Future Work

In the future, we would like to create more task for the robot, such as making coffee and tea. And we want to use Soft-Actor-Critic Reinforcement Learning method for our robot in order to make robot do pick and place object faster and more accurate.

For speech recognition part, we will create a interface with webserver, and a virtual character, people can communicate with robot through this interface and using voice control easier.

References

1.https://arxiv.org/abs/1904.02689

2.https://github.com/dbolya/yolact

3.https://arxiv.org/abs/1412.5567

4.https://medium.com/visionwizard/train-your-own-speech-recognition-model-in-5-simple-steps-512d5ac348a5

5.https://arxiv.org/abs/1712.05884

6.https://github.com/Eruvae/yolact_ros