Multimodal Voice-Vision Robotic Grasping System

Contribution Statement

224040237 Lai Wei – Research, coding and testing

224040219 Yu Liu – Research and coding

224040256 Haozheng Su – Research and testing

Abstract

This report presents a comprehensive multimodal robotic grasping system integrating speech recognition, real-time vision-based tracking, and precise robotic arm control. Utilizing the Whisper model for multilingual voice recognition and translation, combined with a sophisticated vision module employing HSV color-space segmentation and stability analysis, the system dynamically guides a robotic arm to locate, grasp, and transport objects accurately. The implemented approach achieves fully wireless control and demonstrates significant potential for versatile robotic applications.

1. Introduction and Background

With increasing complexity in robotic applications, single-modality systems often fail to provide sufficient adaptability in dynamic environments. Multimodal integration, particularly the combination of speech and vision inputs, has emerged as a robust solution to enable robots to interact intelligently within diverse and unpredictable settings. This project introduces an advanced robotic grasping system that leverages voice commands and visual feedback, highlighting capabilities useful in industrial automation, assistive technologies, and smart home environments.

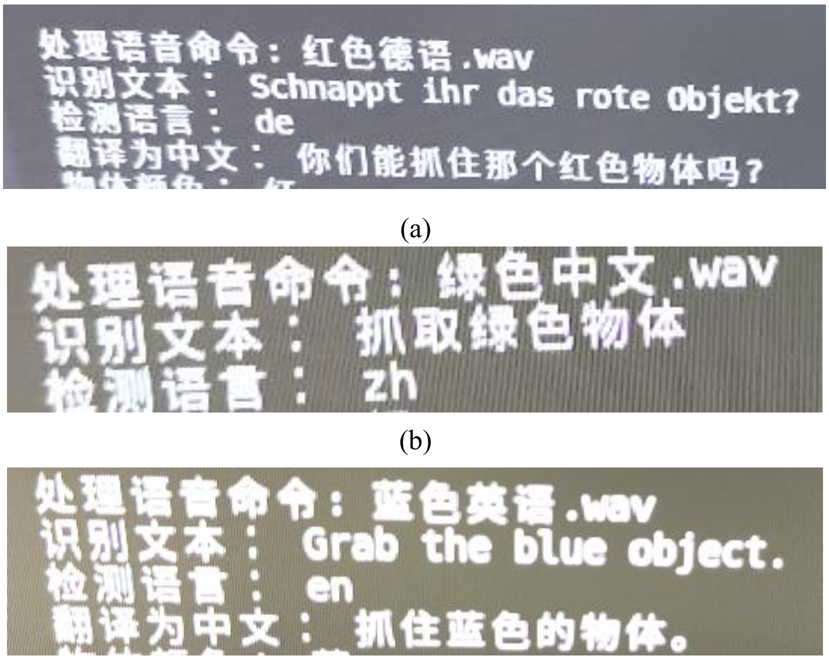
2. System Design and Methodology

2.1 Speech Recognition Module

The voice recognition subsystem employs the Whisper deep learning model (faster-whisper-large-v3) to detect multilingual voice commands, specifically in Chinese, English, and German. Commands in languages other than Chinese are automatically translated into Chinese using

MarianMT models (Helsinki-NLP) for accurate robotic interpretation.
Technical specifications:

- Model: Whisper (faster-whisper-large-v3)
- Translation: MarianMT (Helsinki-NLP series)
- Supported languages: Chinese, English, German



```
处理语音命令：红色德语.wav
识别文本： Schnappt ihr das rote Objekt?
检测语言： de
翻译为中文： 你们能抓住那个红色物体吗？
```

(a)

```
处理语音命令：绿色中文.wav
识别文本： 抓取绿色物体
检测语言： zh
```

(b)

```
处理语音命令：蓝色英语.wav
识别文本： Grab the blue object.
检测语言： en
翻译为中文： 抓住蓝色的物体。
```

(c)

Figure 1: Speech recognition workflow

2.2 Vision Tracking Module

The vision system uses a real-time camera feed processed by OpenCV. Color detection occurs through precise HSV color-space segmentation. The module includes an advanced stability-checking algorithm, which identifies stable target coordinates after analyzing multiple consecutive frames. Pixel coordinates are converted into real-world coordinates using a linear regression-based calibration technique.
Technical highlights:

- Real-time video streaming from camera
- HSV color-space segmentation for object detection
- Stability checking algorithm for robust coordinate identification
- Linear regression calibration for accurate coordinate mapping

2.3 Robotic Arm Control

The robotic arm control is executed using ROS (Robot Operating System) and MoveIt. The system plans movements dynamically based on real-time visual input rather than fixed positions. It includes Cartesian and joint-space path planning, advanced replanning capabilities, and error tolerance strategies for precise and reliable object grasping and placement.

Technical specifications:

- MoveIt planning (Cartesian and multi-stage planning)
- Precise gripper control (grip force and open-close actions)
- Replanning strategies for error recovery

3. System Integration and Operation

3.1 Multimodal Interaction Flow

Upon receiving a voice command, the system identifies the requested object's color, initiates visual tracking to dynamically detect and confirm the target's position, generates the real-time trajectory, and finally instructs the robotic arm to accurately perform grasping and placement tasks.

Operational workflow:

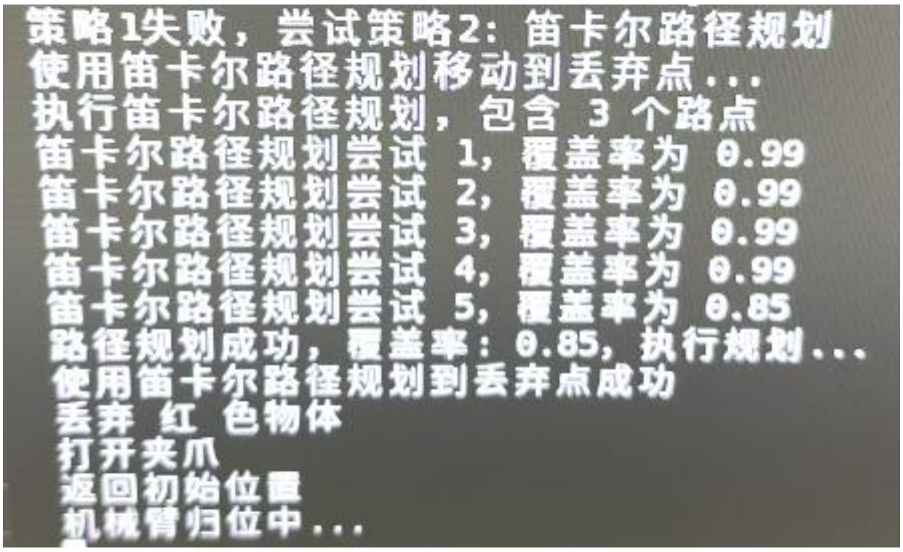
1. Speech command recognition (e.g., "grab the red object")
2. Robotic arm moves to starting position for visual scanning
3. Real-time scanning and object detection via vision module
4. Stability check and coordinate confirmation
5. Path planning and execution for grasping
6. Secure grasp and transport to placement locatio

3.2 Dynamic Vision Tracking Trajectory

Visual scanning involves a dynamic trajectory moving from the green object region to the blue object region, continuously tracking and identifying the requested object. Once the object position is stable and confirmed, the robotic arm immediately initiates the grasping sequence.

Technical highlights:

- Real-time video streaming from camera
- HSV color-space segmentation for object detection
- Stability checking algorithm for robust coordinate identification
- Linear regression calibration for accurate coordinate mapping



```
策略1失败，尝试策略2：笛卡尔路径规划
使用笛卡尔路径规划移动到丢弃点...
执行笛卡尔路径规划，包含 3 个路点
笛卡尔路径规划尝试 1, 覆盖率为 0.99
笛卡尔路径规划尝试 2, 覆盖率为 0.99
笛卡尔路径规划尝试 3, 覆盖率为 0.99
笛卡尔路径规划尝试 4, 覆盖率为 0.99
笛卡尔路径规划尝试 5, 覆盖率为 0.85
路径规划成功，覆盖率：0.85, 执行规划...
使用笛卡尔路径规划到丢弃点成功
丢弃 红 色物体
打开夹爪
返回初始位置
机械臂归位中...
```

Figure 2: Path planning and execution

3.3 Robotic Grasping and Placement Execution

The confirmed coordinates trigger the robotic grasping sequence involving:

- Positioning precisely above the object
- Descending accurately for object grasp
- Activating gripper for secure grip
- Vertical lifting and safe transportation to the destination
- Releasing object and returning arm to the home position

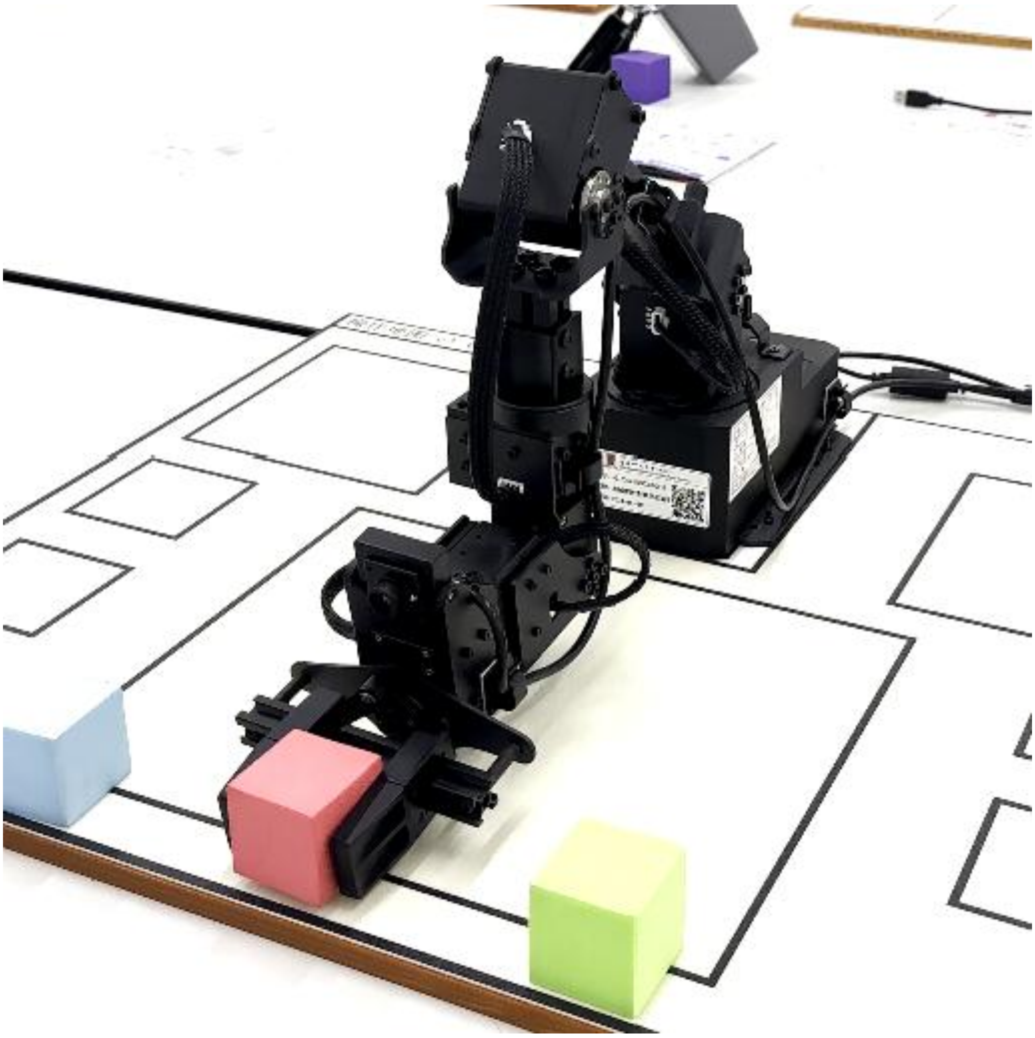


Figure 3: Grasping execution

4. Experimental Results and Performance Analysis

Due to the nature of our experiments, we conducted a qualitative evaluation rather than quantitative performance metrics. The integrated system successfully completed the overall operational workflow, demonstrating effective interaction between speech recognition, visual tracking, and robotic arm control. The grasping and placement tasks were consistently performed correctly, highlighting system reliability.

5. Strengths and Limitations Analysis

5.1 Strengths

- Multilingual voice recognition with real-time translation enhances versatility
- Dynamic vision tracking significantly improves adaptability and precision
- Fully wireless real-time control offers high flexibility and

responsiveness

5.2 Limitations

- Restricted camera field of view requires targets to be within the scanning path
- Initial loading and translation delays of the Whisper model may impact real-time response
- Imperfect path planning occasionally resulted in suboptimal object placement, as demonstrated in the provided images.

6. Future Research Directions

Future work should address the path planning limitations to improve object placement accuracy. Additionally, enhancements in real-time processing speed of the vision module and the integration of depth sensing or tactile feedback mechanisms could provide further improvements in the system's adaptability and robustness. Research into autonomous decision-making algorithms for complex task scenarios is also recommended to further the capabilities of the system.

7. Conclusion

This project has demonstrated the effective integration of voice recognition, real-time vision-based tracking, and dynamic robotic arm control within a multimodal robotic grasping system. The qualitative success of the overall operational workflow underscores the significant potential of multimodal approaches in robotic systems. Addressing identified limitations will enhance reliability and performance, paving the way for broader practical applications.