Video Description

Experiment Objective

Understand and master a basic function of a large AI model, interact with a smart car through language dialogue, and achieve the effect of the smart car describing the current real-time video.

Experiment Steps

1. Observe the IP address of the OLED screen and log in to the remote desktop via VNC.
2. According to the prerequisite configuration tutorial, both the Chinese and English versions need to complete the Tongyi Qianwen key.
3. Open a new terminal and run the following command:

```
cd /home/pi/project_demo/09.AI_Big_Model/
```

```
#Startup command for the Chinese version
python3 VideoDescription/A_video_main.py
```

```
#Startup command for the English version
python3 VideoDescription/A_video_main_en.py
```

4. Wake up the car using the wake-up phrase "Hi, Yahboom" (for international users).
5. After successfully waking up, the car will respond with a honking sound and wait for about half a second. You can then ask the car questions related to the live video.
6. After the robot recognizes your voice, wait a few seconds for relevant information to be displayed on the terminal interface and speaker. You can also terminate the conversation by using the wake-up word.
7. This concludes the conversation. To continue, repeat steps 4-6.

Experimental Results

1. Waiting for wake-up

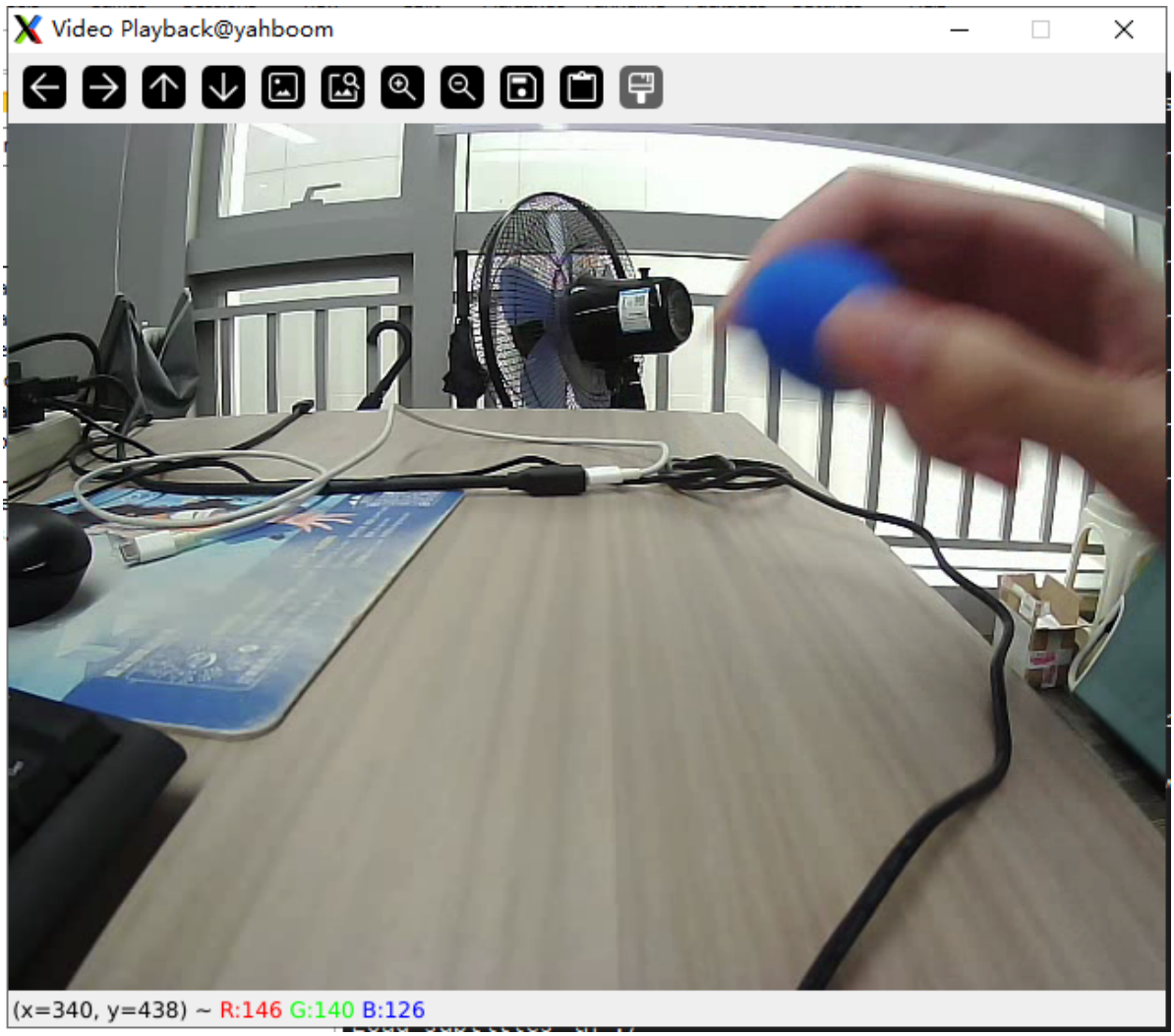
```
pi@yahboom:~/project_demo/09.AI_Big_Model $ python3 VideoDescription/A_video_main.py
serial /dev/myspeech open
Waiting for keyword...
```

2. After successful wake-up, the green-boxed message in the image appears. You can then ask your question.

```
JackShmReadWritePtr::~JackShmReadWritePtr - Init not done for -1, skipping unlock
Current volume: 111279.0, boot threshold: 3000, End threshold: 1500
start recording
3000 111279.0
Current volume: 111884.0, boot threshold: 3000, End threshold: 1500
3000 111884.0
Current volume: 80749.0, boot threshold: 3000, End threshold: 1500
3000 80749.0
Current volume: 65331.0, boot threshold: 3000, End threshold: 1500
3000 65331.0
Current volume: 68820.0, boot threshold: 3000, End threshold: 1500
```

3. After a while, the terminal will print out the question and answer.

```
Q: Has the blue ball been taken?
start recording
Recording completed! (5.01s)
A:Yes, the blue ball has been taken. In the sequence of images, the first frame
shows a hand reaching towards the ball, and the subsequent frames show the ball
being held in the hand. The ball is no longer on the desk, indicating that it ha
s been picked up.
```



Online web pages cannot be played. Only videos downloaded from Baidu Cloud will play.
Case Demonstration Video

Main Source Code Analysis

```
def Speak_Vioce():
    global response
    if TTS_IAT_Tongyi:
        tonyi_tts(response)
    else:
        Xinghou_speaktts(response)
```

```
def main():
    while True:
        if detect_keyword():
            os.system("pkill mplayer")
            time.sleep(.2)

            start_recording()
            time.sleep(1)


            if TTS_IAT_Tongyi:
                content = rec_wav_music_Tongyi()
            else:
                content = rec_wav_music()

            if content != "":
                print("Q:"+content)

                record_video()

                re =Tongyi_video_api(content)

                print("A:"+re)
                try:
                    response = re
                    tts_thread = threading.Thread(target=Speak_Vioce)
                    tts_thread.daemon = True
                    tts_thread.start()

                except:
                    pass

                play_video()
            if content == 0:
               break
        time.sleep(0.1)
```

start_recording: Records a 5-7 second video in real time.
detect_keyword: Wake-up function handler.
start_recording: Recording function handler.
Tongyi_video_api: Visual large model analysis interface.

Chinese version-specific options:
rec_wav_music_Tongyi: Tongyi Qianwen voice recognition.
rec_wav_music: iFlytek Spark voice recognition.
You can choose either voice recognition mode. You can enable or disable it in the **API_KEY.py** file. When TTS_IAT_Tongyi=True, either Tongyi Qianwen voice recognition or iFlytek Spark voice recognition is enabled.
tonyi_tts: Tongyi Qianwen speech synthesis
Xinghou_speaktts: iFlytek Xinghuo speech synthesis
You can choose between two speech synthesis options. You can enable or disable them in the **API_KEY.py** file. When TTS_IAT_Tongyi=True, either Tongyi Qianwen speech synthesis or iFlytek Xinghuo speech synthesis is enabled.

English Version
Speech recognition and synthesis use the iFlytek Spark API by default. You don't need to specify the iFlytek Spark API in API_KEY.py; simply fill in the **openAI_KEY** key.

Modifying the recording duration, start threshold, and end threshold

1. In the terminal, enter:

```
cd /home/pi/project_demo/09.AI_Big_Model/
nano audio.py
```

2. Find the source code shown below.

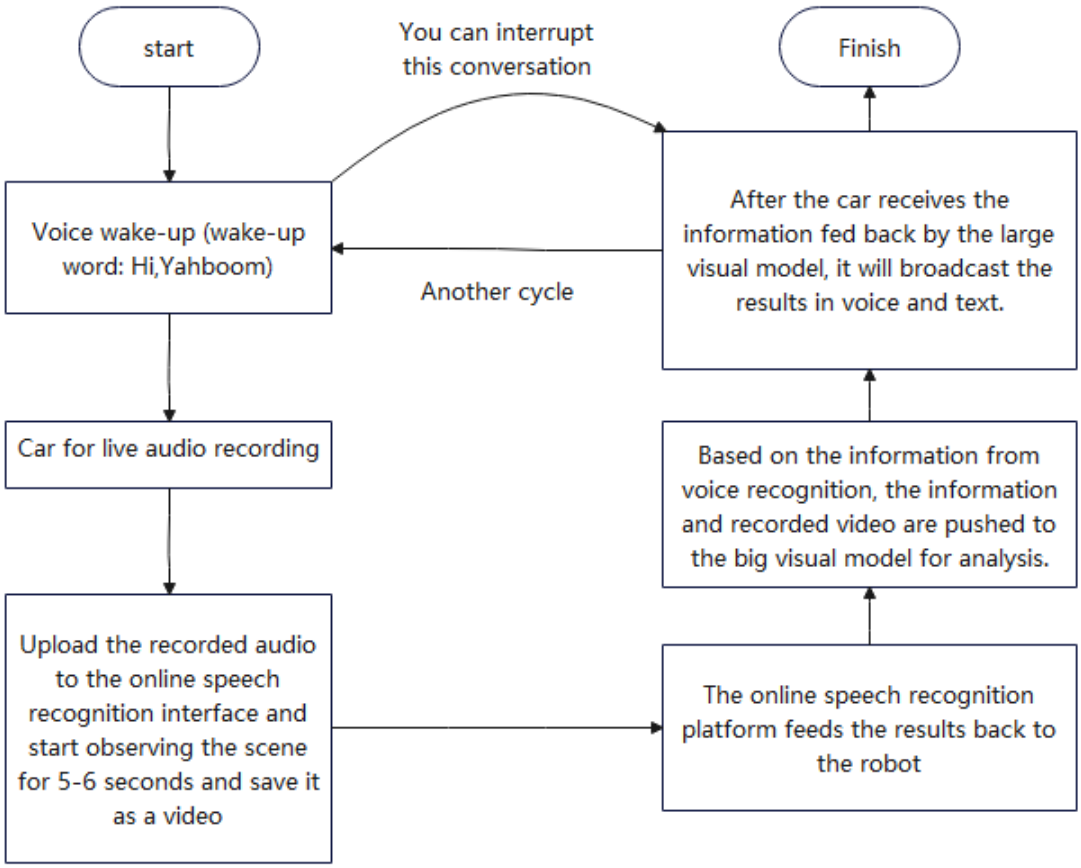
```
189 quitmark = 0
190 automark = True
191 def start_recording(timel = 3,save_file=SAVE_FILE):
192     global automark,quitmark
193     start_threshold = 3000 #30000
194     end_threshold = 1500 #20000
195     endlast = 15
196     max_record_time = 5
```

- start_threshold: The threshold for starting recording when sound is detected (reduced to 5000 in quiet environments and increased to 150000+ in noisy environments).
- end_threshold: The threshold for stopping recording when sound is detected. A recommended value is 30-50% of start_threshold.
- endlast: The number of times to stop recording. Here, 15 is used. For example, Recording will automatically terminate if 15 consecutive sound values meet the stop threshold.
- max_record_time: Recording duration, set to 5 here.
  Note: start_threshold > end_threshold. This is a required rule, and its value can be determined based on the environment.

3. Directory Structure of the Experiment's Main Files

```
├── A_video_main_en.py #English Main Program Interface
├── A_video_main.py #Chinese Main Program Interface
├── recode_video.py #Video Recording and Playback Definitions
├── tongyi_speak_iat.py #Tongyi Qianwen Speech Recognition
├── tongyi_tts.py #Tongyi Qianwen Speech Synthesis
├── tonyi_video_api.py #Tongyi Qianwen Video Analysis Model
├── xinghou_speak_iat.py #iFlytek Spark Speech Recognition
└── xinghou_tts.py #iFlytek Spark Speech Synthesis
```

Overall Flowchart


start
You can interrupt this conversation
Finish
Voice wake-up (wake-up word: Hi,Yahboom)
After the car receives the information fed back by the large visual model, it will broadcast the results in voice and text.
Another cycle
Car for live audio recording
Based on the information from voice recognition, the information and recorded video are pushed to the big visual model for analysis.
Upload the recorded audio to the online speech recognition interface and start observing the scene for 5-6 seconds and save it as a video
The online speech recognition platform feeds the results back to the robot