ANTALYA BİLİM
ÜNİVERSİTESİ

2024-2025 SPRING SEMESTER

Introduction to Robotics

-EE422 / CS421-

WEEK - 2

VISION LANGUAGE MODELS FOR ROBOTIC PERCEPTION

AGAH KUTAY ORUC

EMIN HAN CELIK

TIMUR GULMEZ

MELIH EGREK

Project Introduction and Overview

Within the scope of this project, it is aimed to develop an intelligent robot that can respond to voice commands and visual data and perform tasks in the kitchen. The main purpose of the robot is to understand and carry out the commands it receives from people. For example, when the user says “place the cups on the right side”, the robot perceives its surroundings with the camera, detects the cups and places them in the specified position. The project brings together many artificial intelligence areas such as image processing, voice command recognition and natural language processing.

Technologies Used

The project includes the integration of various artificial intelligence models:

- Visual Perception: CNN and YOLO models are used for object detection. In this way, the robot can recognize objects in its environment and determine their locations.

- Voice Command Recognition: In the system integrated with Google’s Speech API, the voice data received with the microphone is converted into text.

- Language Model (LLM): The commands translated into text are analyzed with a large language model to ensure that the robot understands them. For example, components such as object, action and location in the command are distinguished.

Visual-Language Compatible Learning with CLIP

One of the basic building blocks of the project is the CLIP model developed by OpenAI. CLIP is a Vision-Language Model (VLM) capable of matching visual and linguistic data. Thanks to this model, the robot can understand what the given command corresponds to in the image. CLIP was trained with imitation learning, a subtype of supervised learning. During training, the model was given RGB-D images and task descriptions in natural language; the correct pick-up and drop coordinates were provided as labels.

CLIP consists of two main components:

- Image Encoder (ViT or ResNet)

- Text Encoder (Transformer)

These components transform images and texts into a common embedded space. Thus, the robot can perform correct movements by matching visual elements with linguistic commands.

System Integration

The robot works in the following steps:

1. A voice command is received from the user.

2. The command is translated into text.

3. The task is interpreted with CLIP and matched with visual elements.

4. Object locations are determined with YOLO.

5. The robot performs the task by calculating the most appropriate pick-up and drop-off points.

Thanks to this multi-component system, the robot can communicate with people naturally and successfully perform complex tasks.

REFERENCES

1. **https://github.com/cliport**
2. **https://arxiv.org/pdf/2109.12098**
3. **https://openai.com/index/clip/**