Integrated Vision, Language, and Motor Control in a Humanoid Robot

Michael Ferrante '24, Andrew Steindl '25, Kenneth Livingston, Prairie Rose Goodwin

Introduction

HARPER (Humanoid Autonomous Robotic Platform for Experimental Research) was designed as a tool for human-machine interaction research. The main body and motor actuation system was constructed during the summer of 2022. This summer, our goal was to build the hardware and software infrastructure to enable human-like perception and action. As such, we designed and began the implementation of a bespoke control architecture to coordinate vision, language, and motor control processing systems.

Fig. 1 - Point Cloud of IRRL in RViz, representing the estimated depth of points in the scene via IR projection

Hardware Architecture

Human-scale perception and action abilities are computationally expensive to implement. We chose the Nvidia Jetson AGX Orin as the core of HARPER's control system for this reason. The OAK-D-Pro camera from Luxonis handles visual inputs. It incorporates infrared sensors and a stereo camera to enable on-the-fly depth-perception. A matched pair of omnidirectional electret condensers capture sound, and add-on boards for the Jetson enable analog to digital conversion and amplification. A set of Arduino Zero PWM servo control boards activate the motors and make Harper move in response.

Software Architecture

The OAK-D Pro camera can run simple object recognition models with its onboard processor, thus reducing computational demands on the Jetson and leaving greater capacity for language processing. Auditory signals are processed for speech onboard the Jetson. We utilized the SDK providede by NVIDIA for the AGX Orin platform, RIVA, that includes a neural network pre-trained to enable "automatic speech recognition" in real time. RIVA also does real-time speech synthesis to enable responses. A GPT-like LLM will be able to process responses to spoken queries. Functional integration of these disparate modalities is done by ROS2 (Robot Operating System 2), a "middleware suite" specifically designed for systems integration that also enables direct control of the Arduino Zero. A full diagram of the architecture can be seen below, centered.

Conclusion and Future Expansions

By the end of the summer the team had successfully integrated the Luxonis/DepthAI OAK-D Pro Camera within the ROS2 workspace. This enables the system to identify 80 common objects from the CoCo (Common Objects in Context) dataset within the robot's immediate spatial proximity, as well as provide an estimate for their relative location(s) within the visual field. We also were able to successfully use the RIVA SDK to facilitate ASR (Automatic Speech Recognition) using live microphone input. Further work will be required to integrate visual and auditory processing with motor control. Once complete, HARPER will be a powerful tool for experimenting with different aspects of human robot interactions.

Fig. 2 - Obverse / Wiring of HARPER 1.0

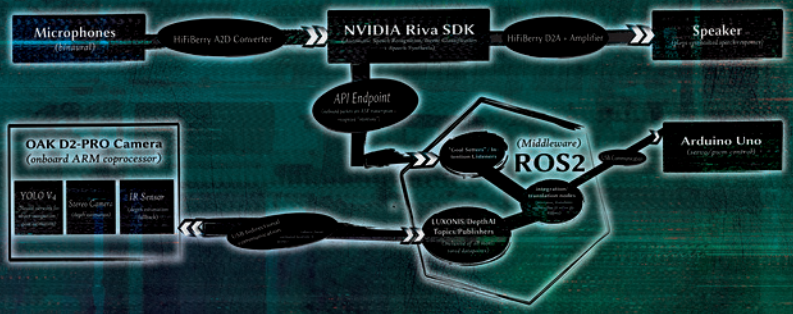

Fig. 3 - Control Architecture Diagram

Fig. 4 - Front / Face of HARPER 1.0