Abstract

In recent years, the integration of social robots into everyday environments has witnessed substantial growth, primarily driven by the increasing demand for more natural, intuitive, and effective forms of human–robot interaction (HRI). This trend reflects a broader shift toward developing robotic systems capable of understanding and responding to human social cues, thereby facilitating seamless collaboration and communication in real-world settings. In this context, this thesis addresses key challenges in social robotics by proposing a comprehensive framework that enhances HRI with a particular focus on multi-user interaction, robot's proactive behavior generation, and soft-biometric recognition through multimodal and multi-task learning strategies.

First, we introduce a hardware-agnostic architecture based on the Robot Operating System (ROS), designed to support real-time interaction in dynamic, multi-user scenarios. The system integrates multimodal perception modules, including head pose estimation, active speaker detection, and direction-of-arrival analysis, together with reasoning components built upon Behavior Trees and Finite State Machines. This design effectively addresses the Multi-Engagement Problem, enabling robots to manage simultaneous interactions with multiple users. Furthermore, the framework was extended with an interactive game to support elderly assistance and enhance user engagement through entertainment.

Second, the thesis presents a novel Vision-Language Model (VLM)-based approach for generating context-aware sentences from video input. A dual-teacher Knowledge Distillation pipeline is proposed to automatically generate training data, combining the strengths of VLMs and Large Language Models (LLMs).

The resulting lightweight model is optimized for deployment on embedded platforms, enabling proactive and contextually relevant verbal behaviors in social robots.

Third, the MAGNET architecture is introduced: a Multi-Modal Multi-Task learning framework for soft-biometric estimation, capable of jointly recognizing gender and emotion from audio and video data. By employing soft parameter sharing, MAGNET achieves state-of-the-art performance while maintaining low computational overhead, making it suitable for real-time applications.

Extensive quantitative and qualitative evaluations confirm the effectiveness of the proposed methods in improving the naturalness, responsiveness, and social acceptability of human–robot interactions. This thesis contributes a scalable and efficient solution for deploying socially intelligent robots in real-world environments.