RoboCanes-VISAGE 2024

Team Description Paper

K. Pasternak[1], C. Duarte[1], S. Luo[1], J. Ojalvo[1], U. Visser[1], and C. Lisetti[2]

[1] University of Miami, Coral Gables FL 33146, USA
robocanes@cs.miami.edu
[2] Florida International University, Miami FL 33199 USA
affect@cs.fiu.edu

Abstract. RoboCanes-VISAGE consists of faculty and PhD students from two universities in Miami, FL: the University of Miami (UM), Department of Computer Science and Florida International University (FIU), School of Computing and Information Sciences. Together, our research covers artificial intelligence, robotics, human-robot interaction, and affective social computing. This paper describes our approach for the 2023 competition.

1 Introduction

Our team has designed, developed, and implemented its autonomous agent framework and software from the ground up in both the 3D Soccer Simulation League and the Soccer Standard Platform League (NAO robots) since 2010. This framework has evolved to a flexible research platform that led to over 40 publications and shared software (cf. section 4) over the years.

For RoboCup@Home, RoboCanes has teamed up with the VISAGE lab from FIU to form the RoboCanes-VISAGE team. The VISAGE lab is led by Dr. Christine Lisetti, a faculty from the School of Computing and Information Sciences at FIU who is one of the founders of the Affective Computing research field. Dr. Lisetti's group is known for its research on social 3D Virtual Avatars.

We use Toyota's HSR to leverage the latest progress in affective social computing and socially intelligent agents, as well as using the latest technology in AI and robotics to address the RoboCup@Home challenges. It is an excellent platform to embody the integration of the UM RoboCanes agent with the FIU VIrtual Social AGEnt (VISAGE).

The RoboCanes agent is mainly responsible for managing and controlling navigation, object manipulation, grasping, etc., while the VISAGE agent handles face and facial expression recognition of the human interacting with the HSR, voice recognition and speech synthesis, and 3D-graphics for facial and gesture synthesis of the VISAGE agent.

2 Research Focus

We propose to leverage and expand the latest research on social robotics in order to enhance and personalize the capabilities of collaborative robots' (co-robots) to

communicate with humans using natural verbal and non-verbal communication techniques. We will focus on co-robots' communication and collaboration in the home environment, where natural communication is of essence. Literature reveals that a vast majority of research is focused on low-degree of freedom anthropomorphic robots. We will develop our interactive agent on a high-degree of freedom hybrid anthropomorhic robot.

The RoboCanes agent will be mainly responsible for managing and controlling navigation, object manipulation, grasping, etc., while the VISAGE agent will handle face and facial expression recognition of the human interacting with the HSR, voice recognition and synthesis, and 3D-graphics for facial and gesture synthesis of the VISAGE agent.

We will integrate and coordinate both agents toward a coherent and engaging multimodal model of communication with the human user.

3 Innovative technology

Our group has made a variety of scientific contributions in the areas of artificial intelligence, robotics, virtual agents/human-computer interaction, and with both groups (UM, FIU) together also a first human-robot interaction (cf. section 3.4).

3.1 eEVA as a Real-time Multimodal Agent Human-Robot Interface

We have developed a multimodal human-robot interface [1] for Toyota's Human Support Robot (HSR, designed to help people in homes or offices) which integrates the RoboCanes agent and the Embodied Empathetic Virtual Agent (eEVA) developed by FIU's VISAGE lab. The RoboCanes agent is responsible for managing and controlling navigation, object manipulation, and grasping, among other physical actions, while the VISAGE agent is responsible for recognizing and displaying social cues involving recognizing the user's facial expression and speech, synthesizing speech with lip-synchronization, and portraying appropriate facial expressions and gestures. Our interface also allows the RoboCanes agent to send commands to the VISAGE agent on where to direct its gaze.

We created a greeting context for the pilot study of our first social human-HSR interactions with our RoboCanes-VISAGE interface by designing a small set of greeting gestures to personalize the Toyota HSR with its users' greeting preferences, and to establish some initial rapport in preparation for more advanced studies in the future. The Toyota HSR generates greeting gestures from four different cultures: waving-hand (Western), fist-bump (informal Western), Shaka (Hawaii), and bowing (Japan) greeting gestures. The HSR's gesture greetings are performed based on (1) the user's spoken selection of one of the four greetings and (2) our pilot questionnaire aimed to assess the impact of combining the robot and the virtual agent interface on the user's experience (*e.g.*, feelings of enjoyment, boredom, annoyance, user's perception of the robot's friendliness or of competence).

In human-human interaction subtle mirroring of nonverbal cues during conversations promotes rapport building. Similarly, it could translate to human-robot interaction to improve communication. Thus, we investigated whether the ability of

a robot to mirror its user's head movements and facial expressions in real-time can improve the user's experience with it. We piloted a study [2] to assess the impact of face detection with posture mimicking and emotion mirroring on the user's sense of comfort and naturalness during their interaction with the robot. Each skill was first performed separately with either Embodied Conversational Agent (ECA), robot's head or both performing the movements. Finally, both skills were combined in last part of the experiments.

3.2 Toolbox: Low footprint reinforcement learning library RLLib

One of our main accomplishments is the creation of the RLLib tool package that we have made available for other scientists. **RLLib has been downloaded by other peer researchers more than 5,000 times to date** [3], a recognition of the significance of our contribution to the scientific community. RLLib is a C++ template library to learn behaviors and represent learnable knowledge using on/off policy RL standard, and gradient temporal-difference learning algorithms in RL. It is an optimized library for robotic applications that operates under fast duty cycles (e.g., $\leq$ 30 ms). This is a significant difference to other available packages. RLLib has been tested and evaluated on RoboCup 3D soccer simulation agents, physical NAO V4 humanoid robots, and Tiva C series launchpad microcontrollers to predict, control, learn behaviors, and represent learnable knowledge. The implementation of the RLLib library is inspired by the RLPark API, which is a library of temporal-difference learning algorithms written in Java. RLLib garnered attention when presented at RoboCup Symposium and has been used by third parties as well, e.g. on malware detection by Bidoki et al. [4]. We will make use of the library for the HSR.

3.3 Motivational interviewing with intelligent virtual agents (IVA)

We developed a virtual counseling system which can deliver brief health interventions via a **3D anthropomorphic speech-enabled interface** – a new field for spoken dialog interactions with intelligent virtual agents in the health domain. We developed our dialog system based on a Markov Decision Process (MDP) framework and optimized it by using RL algorithms with data we collected from real user interactions. The system begins to learn optimal dialog strategies for initiative selection and for the type of confirmations that it uses during the interaction. We compared the unoptimized system with the optimized system in terms of objective measures (e.g. task completion) and subjective measures (e.g. ease of use, future intention to use the system) and obtained positive results. The system is able to learn dialog strategies for initiative and confirmation selection. Our **contributions to the Spoken Dialog Systems domain** include the creation of a RL paradigm to the completely new domain of behavior change - where our dialog length is 4-5 times longer and where the nature of the dialog is less restricted than spoken dialog systems operated in the tourist information domain. We **contributed to the healthcare domain** with the first system to use speech as an input medium with a RL-based approach. Our initial evaluation showed that the dialog managers that are optimized with RL have the potential to reach optimal behavior, given enough training data [5, 6].

3.4 Human-robot interaction with a humanoid robot

We combined a spoken dialog system that we developed to deliver brief health interventions (cf. section 3.3) with a NAO robot. The dialog system is based on a framework facilitating a MDP and is optimized using RL algorithms (we used our own RLLib [3], cf. section 3.2). The spoken dialog system for the humanoid robot was a novelty at that time and exists as a proof of concept. We anticipate that the NAO robot will become a very likable and effective mode of delivery for brief interventions on target behaviors such as poor diet, overeating, or lack of exercise, among others. The appeal of the NAO to children makes it particularly suitable to become a child's favorite health coach, say, to discuss eating more fruits and vegetables on a daily basis.

3.5 Knowledge representation and reasoning

This line of research **combines** modern symbolic knowledge representation and reasoning techniques from the **Semantic Web** domain with modern **autonomous robots**. Knowledge should be represented in real-time (i.e., within ms) and deduction from knowledge should be inferred within the same time constraints. We proposed an extended assertional formalism for an expressive $\mathcal{SROIQ}(\mathcal{D})$ Description Logic to represent asserted entities in a lattice structure [7]. This structure can represent temporal-like information. Since the computational complexity of the classes of description logic increases with its expressivity, the problem demands either a restriction in the expressivity or an empirical upper bound on the maximum number of axioms in the knowledge base. We have conducted experiments in the RoboCup 3D Soccer Simulation League environment and provide justifications of the usefulness of the proposed assertional extension. We showed the feasibility of our new approach under real-time constraints and conclude that a modified FaCT++ reasoner empirically outperforms other reasoners within the given class of complexity. We intend to use our approach with incremental reasoning on HSR to model beliefs and interpret entities in uncertain environments in the near future.

3.6 Inverse Trajectory Planning (ITP)

Tracking and predicting a person's movement in three dimensional space in order to ascertain their current location and intended trajectory in the environment is a difficult task. However, this knowledge would enable an agent to learn how to navigate the environment without causing danger to a human that it is interacting with. We developed a novel probabilistic framework based on von Mises distributions for robotic systems to detect and predict a human pose. We implemented the framework on the Toyota HSR robot and showed its capabilities on the task of following a human.

3.7 Motion Planner with Geometric Heuristics

Traditionally, we used the MoveIt Motion Planning Framework. Last year, however, in order to achieve more robust manipulation performance, we constructed a new

motion planner for grasping and placing objects using geometric heuristics. Based on a known set of sampled environments, we deduced geometric heuristics that will always generate a solution for the grasp trajectory tailored to the environment assumed by the heuristics. The planner first gets a bounding box provided by our vision pipeline (cf. section 3.8). Then it computes the heuristics: (1) the decomposed angle in a desired frame and (2) the ratio of the dimensions from the position and orientation of the bounding box. From these heuristics, our planner can decide upon a combination of the base trajectory, gripper trajectory, and variations of grasp orientation of the gripper. Our experiments show that our planner using geometric heuristics outperforms the MoveIt planner with respect to overall execution duration (from the kickoff of the planning to the completion of the motion) and safer trajectories (adaptive collision avoidance).

3.8 Vision Pipeline

Our vision pipeline allows the agent to detect an object in the world seamlessly. Building on top of previous iteration of the pipeline, there is no more need for an operator to collect object data manually. Our new data generation pipeline loads meshes of the objects and automatically generates image data with various lighting and rotational conditions. The data is then automatically cropped using GrabCut [8], where the initial seed for GrabCut is the whole image as the image is without background values. In contrast to a simple bounding box segmentation, GrabCut eliminates the biases of the environment in which data is collected. The bounding box does not need to be exact in order to obtain an accurate segmentation of the object of interest. In the last step, a program augments the data by imposing the segmented images in different pose configurations into background images, (e.g., images of the environment where the object will be found). Once all the images of the segmented object are reviewed by the operator, the data is sent to our supercomputer, Triton, at the University of Miami. On the server, our instance of the YOLO convolutional neural network learns how to recognize objects from the augmented data.

4 List of externally available components

We have produced and shared a variety of our system components to support various research communities. Some of the components are part of the prior section (3) and will only be short-listed here.

HapFACS[3] is an open source software that enables (without prior knowledge of computer graphics) the animation of speaking 3D virtual human-like characters with physiologically realistic facial expressions that have been validated by experts in facial expressions [9]. Specifically, HapFACS provides the ability to manipulate the activation – in parallel or sequentially – of combinations of the smallest groups of (virtual) facial muscles capable of moving independently in the human face, for the creation of physiologically and socially believable speaking virtual agents.

[3] `http://ascl.cis.fiu.edu/hapfacs-open-source-softwareapi-download.html`

RLLib[4] (cf. section 3.2) is a C++ template library to learn behaviors and represent learnable knowledge using on/off policy RL standard, and gradient temporal-difference learning algorithms in RL.

is a software program designed to assess and develop agent behaviors in the RoboCup 3D Soccer Simulation league.

5 Tasks Approaches for RoboCup 2023

5.1 Clean Up

The main objective of this task is for the robot to perceive the objects, categorize them semantically, pick them up, and place them in the correct predefined location.

We tackled the navigation by using and adapting provided navigation tool and defining task areas to precise the locations of deposit spots to different groups of items. This is a general solution for all navigation for this task and the navigation components for all other tasks of RoboCup 2023. Once the HSR has successfully navigated to the area where different objects are initially placed, we use YOLO network trained from our vision pipeline (cf. section 3.8) to classify objects from images acquired by the HSR's Xtion RGBD camera. Then, the class of the object is parsed through feature heuristics such as color, shape, functionality, etc., and categorized into a group of similar objects. The categories of the objects are drinks, cleaning supplies, pantry items, fruit, snacks, and cutlery. Objects in each group are placed in the same, predefined by the rule book location in a way that they don't fall on each other. After categorization, the HSR extracts a 3D bounding box from RGBD images using Euclidean clustering, which can then be used for our geometric heuristic motion planner (cf. section 3.7) to plan the picking trajectory.

Once the object is picked, if it is the first object picked during the task, the HSR will generate a compact semantic mapping between feature topology and corresponding groups to which separate locations are assigned. Thus, before placing each object, HSR queries the feature topology to decide which location to place the object in.

5.2 Receptionist

The main objective of this task is to greet arriving guests, introduce them to the host and each other, and offer them a spot to sit.

We use the same navigation component as the Clean Up task (cf. section 5.1). Once HSR has reached the entrance it waits for a person's arrival. We use a previously trained on masked volunteer images, YOLO network and dlib library to recognize a person present at the door. We use our Real-time Multimodal Agent Human-Robot Interface to interact with guests to obtain information about their names and favorite drink. Then again, we use our navigation to bring the guest to the sitting area in the room and proceed with introductions where we use the head orientation to indicate the person we are referring to.

[4] `https://mloss.org/software/view/502/`

Similarly, as in the previous task (cf. section 5.1), we used a previously trained YOLO network to determine the occupancy of the couch and the chairs in the room. Once we locate a person, we compare it with the location of a predefined sitting spot. That way we can determine if the person is sitting in a chair we want to offer or not and if HSR should offer a different seat. Then we navigate to the spot across from the seat, have the robot point at it, and offer it to the new guest.

6 Conclusion

Since RoboCanes-VISAGE acquired the Toyota HSR robot four years ago, our team has kept up to speed with the latest technology of localization, navigation and manipulation, and we were able to perform at extremely high level at WRS and past RoboCup events. In addition, our team has evolved into a driving force of HRI and motion planning research with the latest eEVA integration (cf. section 3.1) and ITP (cf. section 3.6).

References

1. Pedro Peña, Christine Lisetti, Mihai Polceanu, and Ubbo Visser. eEVA: Real-time Web-based Affective Agents for Human-Robot Interface. In Dirk Holz and Katie Genter and Maarouf Saad and Oskar von Stryk, editor, *Robot World Cup XXII*. Springer Berlin / Heidelberg to appear, 2019.
2. Katarzyna Pasternak, Zishi Wu, Christine Lisetti, and Ubbo Visser. Towards Building Rapport with a Human Support Robot [Paper presentation]. RoboCup Symposium, Worldwide, 2021.
3. Saminda Abeyruwan and Ubbo Visser. Rllib: C++ library to predict, control, and represent learnable knowledge using on/off policy reinforcement learning. In *Robot Soccer World Cup*, pages 356–364. Springer, 2015.
4. Seyyed Mojtaba Bidoki, Saeed Jalili, and Asghar Tajoddin. Pbmmd: A novel policy based multi-process malware detection. *Engineering Applications of Artificial Intelligence*, 60:57–70, 2017.
5. Ugan Yasavur and Christine. Lisetti. Let's talk! speaking virtual counselor offers you a brief intervention. *Int'l Journal of Multimodal User Interfaces*, 8:281–306, 2014.
6. Christine Lisetti, Reza Amini, and Ugan Yasavur. Now all together: Overview of virtual health assistants emulating face-to-face health interview experience. *KI-Künstliche Intelligenz*, pages 1–12, 2015.
7. Saminda Abeyruwan and Ubbo Visser. A New Real-Time Algorithm to Extend DL Assertional Formalism to Represent and Deduce Entities in Robotic Soccer. In H. Akin Reinaldo, A. C. Bianchi, Subramanian Ramamoorthy, and Komei Sugiura, editors, *RoboCup 2014: Robot Soccer World Cup XVIII*, LNAI, pages 270–282. Springer Berlin / Heidelberg, 2015.
8. Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. In *ACM transactions on graphics (TOG)*, volume 23, pages 309–314. ACM, 2004.
9. Reza Amini and Christine Lisetti. Hapfacs: An open source api/software to generate facs-based expressions for ecas animation and for corpus generation. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 270–275. IEEE, 2013.

A
3rd Party Robot's Software

- YOLO
- Moveit!
- Snips NLU
- Rasa NLU
- TMC software
- PCL (Point Cloud Library)
- Google Speech API
- Amazon Polly
- Robot Operating System (ROS)

B
External Computing Devices

- Alienware 17 Gaming Laptop
  - Processor: Intel Core i7-8750H CPU @ 2.20GHz x 12
  - Graphics: GeForce GTX 1070/PCIe/SSE2
  - OS Type: 64-bit
  - Disk: 1.2 TB

Fig. 1: HSR in our lab, facing Toyota's lego block challenge