Università di Salerno

Dottorato in Ingegneria dell'Informazione

Final report

Student: Francesco Rosa

Advisor: prof. Mario Vento

Date: 15/10/2024

By the end of the 3rd year you should have got a good competence in your research area, have achieved (at least part of) your research objectives, and have results that confirm that you have been able to advance the state of the art. Hopefully, your research results should have been confirmed by peer-review as well, at least in part. You should have written an advanced draft of your PhD thesis.

Proposed title for the thesis

Video-Conditioned Multi-Task Imitation Learning for Robotic Systems: Enhancing Robustness Through Object-Centric Reasoning

Abstract of the thesis

Robot technology is one of the pillars of modern society. Advances in information, electronic, and mechanical fields enable us to build and program machines to perform tasks in very different contexts, such as industry, surgery, and space missions.

Specifically, in manufacturing, robots are mainly used to perform repetitive and unhealthy works like assembly, welding and material handling, thanks to their mechanical robustness and ability to repeatedly perform the same movements with high accuracy and precision.

While in the early day, robot systems were constrained in isolated and known environments. Over the past few decades, robots have been asked to solve tasks in dynamic and unknown/partially known environments, where they must **coexist** and **cooperate** with humans, while solving different **dynamic** tasks (e.g. pick a requested object, whose position is not known a priori).

In this scenario, the desired characteristics of such robotic systems are:

- **Adaptability to new conditions**, i.e., the system must be able to easily adapt to dynamic changes in system and environmental conditions, performing *intelligent behaviors* to handle these new scenarios and solve the desired task.
- **Adaptability to new tasks**, i.e., the system must be able to easily adapt to both new variations of a known task and completely new tasks by exploiting experience to infer actions and solve them.

These requirements can be challenging to achieve with traditional robot programming techniques based on hand-written policies and control methods. These conventional techniques often require a meticulous analysis of process dynamics, the construction of an analytical model, and the derivation of a control law that meets specific design criteria. This design process is tedious and time-consuming, particularly when high-level perception systems (e.g., cameras, microphones, motion sensors) are used to infer the state of the environment (such as the unknown position of a desired object relative to the end-effector) and the intentions of the human operator.

In contrast, significant advancements have been made by leveraging *learning techniques*, where the control policy is inferred from data. This data can be generated either by **agent experience** or by **expert demonstrations**.

In the case of agent experience, there is a trial-and-error procedure where the control policy generates actions executed by an agent, which interacts with the environment. The parameters are then tuned according to the effectiveness of the actions, based on their impact on the environment relative to the task to be solved.

In the case of expert demonstrations, the control policy parameters are directly tuned using a dataset containing examples of task execution. Here, the goal is to replicate the tasks observed in the dataset.

Given this background, the thesis is framed in the context of *Learning from Demonstration* (LfD), a learning approach based on expert demonstrations. According to the requirements of adaptability the thesis focus on a specific aspect of LfD, named **Multi-Task LfD**. In this case, the control policy is not trained to execute a single task (e.g., picking an object) with the goal of generalizing across different objects and initial conditions. Instead, it is trained to handle various variations of a specific task (e.g., picking an object from different possible locations) or even entirely different tasks (e.g., a single control policy that solves both

picking and placing tasks as well as assembly tasks). The goal is to generalize not only with respect to the objects being manipulated and the initial conditions but also with respect to the tasks themselves. This means that it is possible to achieve a system capable of solving new variations by leveraging the knowledge-sharing hypothesis.

In this scenario, the learning procedure is much more challenging because there is the need to include and define the **conditioning signal** (i.e., the signal that informs the policy about the task to execute, the object to manipulate, and the target placing location). Additionally, the environment can contain **multiple distractor objects** (e.g., objects that can potentially be manipulated but are not of interest for a given task variation).

Regarding the conditioning signal, there are at least two intuitive approaches. The first is through a natural language description of the task to be executed, and the second is through a video demonstration.

In the former, the task is described using phrases that specify the task details, such as ``Pick the red box and place it into the first bin". Given in input the phrase, the system must be able to infer the intent of the task (i.e., the pick-place operation) and the object of interest (e.g., red-box for picking and first bin for placing), and correlate this information with the environment and robot state to effectively control the robot.

In the latter, another agent (either a robot or a human operator) performs the task in a different environment configuration, records this execution, and provides the video as input to the control policy. The control policy must then infer the intent from the video (i.e., the task to be performed, the object to be manipulated, and the final state) and control the robot to complete the task according to the agent's state, the environment's state, and the commanded task.

Inspired by how humans can learn to replicate tasks by simply observing their execution, the main goal of this thesis is to develop a system capable of replicating tasks shown in a video demonstration. This involves addressing challenges related to extracting task-relevant information from the video, such as identifying the manipulated object and its final position.

Regarding the issue of distractor objects, they are typically defined as items present in the scene but never involved in manipulation operations. Modern deep architectures can handle this scenario effectively, as they can easily learn to ignore these objects since they do not participate in any manipulation tasks. However, in the context proposed in this thesis, the problem is further complicated by the fact that the semantic meaning of an object (i.e., target or distractor) is defined at run-time by the command itself. This means that if the initial configuration consists of four objects (e.g., four boxes of different colors), a specific object may or may not be of interest based on the command given to the robot.

The primary contribution of this thesis is tackling the challenge of distractor objects. A key issue identified in existing literature is **target misidentification**, where the learned control policy generates valid trajectories, enabling the robot to reach, pick, and place objects, but frequently manipulates the wrong object.

To solve this problem, two main considerations were made:

1. Architectures proposed in the current literature are predominantly end-to-end, translating high-dimensional inputs, such as images, into corresponding low-dimensional actions. As a result, the model must learn an implicit representation that encodes both the task objective and the current state of the environment, including the location of the target object.
2. The learning procedure optimizes an **action-centric** metric, meaning that it is not directly linked to task success but instead focuses on mimicking the expert's actions on average. This action-focused optimization can lead to poor encoding of critical information, such as object positions.

These two factors can result in a control policy that fails to effectively guide the robot toward the target object. It was observed that the early stages of trajectory execution are critical. Even small errors during these initial steps can cause the robot to reach and ultimately pick the wrong object.

Based on these considerations, this thesis explores the development of a **modular** architecture instead of an end-to-end approach. This architecture features modules specifically designed for reasoning about the objects of interest (e.g., target object and placement location). The outputs of these reasoning modules are then integrated to simplify the learning problem for the Control Module. This module is now informed by low-dimensional information, such as the position of the target object, which may be more effectively utilized during the learning process, especially considering the action-centric cloning loss.

To perform this explicit reasoning, a ***Conditioned Object Detector*** (COD) has been developed. This module, given the video demonstration and the current agent observation as input, predicts the category-agnostic bounding box related to the target object and the final placing location. This low-level positional information is then provided to the control module, which predicts the actions to perform.

The learning procedure is then divided into two steps. The first step involves training the *Conditioned Object Detector*} module, which focuses on explicitly solving cognitive tasks, such as detecting regions of interest represented by the object to be manipulated and its final location. The second step involves training the ***Object Conditioned Control Policy*** (OCCP), which focuses on solving the control problem using low-level positional information that can be easily mapped into the corresponding actions.

The final system has been extensively tested in simulation environments, then it was also validated on a real-world robotic platform.

Regarding the simulated environment, the system was evaluated on both **multi-variation single-task** scenarios and **multi-variation multi-tasks** scenarios, considering four simulated tasks: Pick-Place, Nut-Assembly, Stack-Block, and Button-Press. Each task had different variations based on the manipulated object and the final state. While the tasks share common properties, they also have specific characteristics. For example, the Nut-Assembly task involves contact-rich, precise manipulation, whereas Pick-Place can be solved in a much rougher manner.

Overall, the proposed methods demonstrated very promising behaviors and a general improvement over baseline methods that do not include object-related reasoning. This shows that solving manipulation tasks with an object-oriented approach can be an effective paradigm for LfD problems. Additionally, this approach provides interpretable information to the end user, as the predicted bounding boxes can be interpreted as the locations where the robot will move.

In conclusion, the proposed method was tested also in a real-world environment, where the complexity of the problem is heightened by the presence of less and noisy data collected through teleoperation. Even under these challenging conditions, the proposed method demonstrated its effectiveness in addressing both the cognitive and control problems. This confirms that the object prior can be successfully applied in real-world scenarios, enabling the development of a reliable system despite limited and noisy trajectory data.

List of publications

Related to the thesis work

Title: "Improving Learning from Visual Demonstration Methods by Target Localization"

Authors: Pasquale Foggia, Francesco Rosa*, and Mario Vento

Conference Name: IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)

Title: "Enhancing robotic demonstration-based learning method with preliminary visual target localization"

Authors: Pasquale Foggia, Francesco Rosa*, and Mario Vento

Conference Name: European Robotic Forum 2024

Not related to the thesis work

Title: Enhancing IoT Network Security with Graph Neural Networks for Node Anomaly Detection

Authors: Vincenzo Carletti*, Pasquale Foggia, Francesco Rosa, and Mario Vento

Conference Name: Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition and Structural and Syntactic Pattern Recognition S+SSPR

*: Corresponding Author

Learning activities

(list all the accomplished ones and the results you got – attach a print of your S3 career)

Mandatory learning activities:

- Name: "ADVANCED MACHINE LEARNING", Grade: Excellent
- Name: "COMPUTATIONAL PARADIGMS FOR PROBLEM SOLVING", Grade: Excellent
- Name: "ENGLISH LANGUAGE", Grade: Idoneo
- Name: "OPTIMIZATION TECHNIQUES FOR ENGINEERS", Grade: Excellent
- Name: "SCIENTIFIC WRITING AND PUBLISHING", Grade: Excellent
- Name: "EXPLOITATION OF RESEARCH RESULTS", Grade: Excellent
- Name: "FUNDING AND MANAGEMENT OF RESEARCH PROJECTS", Grade: Excellent

Additional courses:

- Name: "MOBILE ROBOTS FOR CRITICAL MISSIONS", Grade: 30L
- Name: "OTTIMIZZAZIONE", Grade: 30

Other activities you have performed in these three years

Activities outside the University of Salerno

Mandatory Period Abroad

From April 8, 2024, to July 15, 2024, I spent three months as a Visiting PhD Student under the supervision of Professor Luc Brun at the Groupe de Recherche en Informatique, Image et Instrumentation de Caen (GREYC) lab, part of the École nationale supérieure d'ingénieurs de Caen. During this time, I worked on applying Graph Neural Networks (GNNs) in the field of Robotics Learning. Specifically, the focus was on using GNNs as heuristic estimators in Task Planning. As part of this research, I conducted an evaluation of state-of-the-art GNNs to assess their effectiveness for heuristic estimation in these types of problems.

Dissemination Activity

I participated in the Application of Intelligent Systems (APPIS) conference from January 20 to 24, 2024. The primary goal of the conference was to promote interaction and collaboration in the areas of artificial intelligence, machine learning, pattern recognition, data science, and related fields. In this context, I presented the research activities I have conducted thus far.

Conference attendance

Conference attendance:

1. *European Robotic Forum 2024,* from 13 to 15 March 2024, Rimini (Italy). My participation to this conference was to present the paper titled "Enhancing robotic demonstration-based learning method with preliminary visual target localization".
2. *IEEE International Conference on Robot & Human Interactive Communication (RO-MAN),* from 26 to 30 August 2024, Pasadena (USA). My participation to this conference was to present the paper titled "Improving Learning from Visual Demonstration Methods by Target Localization".

Involvement in Research Projects

I was involved in the Research Project named "FlExible assembLy manufacturing with human-robot Collaboration and digital twin modEls" (FELICE - H2020-ICT-46-2020). Regarding this project, I worked on the problem of Command Speech Recognition for an industrial collaborative robot. The objective was to develop a module capable of recognizing spoken commands from a human operator, designed to run on an embedded device mounted on the robot.

Other activities

I was involved in the reviewing process for the following journals:

1. Pattern Recognition Letters
2. Pattern Analysis and Applications
3. Scientific Reports – Science

Moreover, I was involved in other internal research activities related to:

1. **Graph Representation Learning (GRL).** GRL, which is a subfield of Machine Learning and Deep Learning, is dedicated to obtain valuable information from data structured in the form of graphs. This approach to data representation finds relevance in diverse domains, including but not limited to Recommendation Systems, Social Network Analysis, and Biological Networks, among others. Under the supervision of Professor Pasquale Foggia and Professor Vincenzo Carletti, I directed my focus toward the topic of Anomaly Detection within the context of Internet of Things Device Networks. Our research aims to propose novel architectures and innovative learning paradigms to dynamically

analyze such networks. The ultimate objective is to discern whether a subset of devices exhibits anomalous behaviors, an indicative sign that could potentially indicate the presence of cyber-attacks in progress. This pursuit not only enriches the field of Graph Representational Learning but also holds profound implications for the cybersecurity landscape.

This activity led to the publication of a conference paper listed above, and one paper submitted to the journal named *"Pattern and Recognition Letters"* which is **under review**.

2. **Graph Matching Algorithms for GPU architectures.** Graph Matching is a well-known problem in computer science, where the objective is to identify instances of a small pattern graph within a larger target graph. The challenge lies in developing efficient algorithms for this task, as subgraph isomorphism is an NP-complete problem. Under the supervision of Professors Pasquale Foggia and Vincenzo Carletti, I worked on designing, implementing, and analyzing an algorithm to solve subgraph isomorphism by levering the high parallelism provided by GPU architectures.

3. **Command Speech Recognition.** As part of the European project FELICE, I worked on the problem of Command Speech Recognition for an industrial collaborative robot. The objective was to develop a module capable of recognizing spoken commands from a human operator, designed to run on an embedded device mounted on the robot. Key challenges included collecting a dataset representative of the task and building a system that could operate reliably despite industrial noise and varying distances between the robot and the operator. During our activity, we explicitly address these challenges adopting learning procedures that involve high-quality synthetic audio sample, data-augmentation procedures that aim to simulate the problems observed during the experimental validation performed at the *Centro Ricerche FIAT* in Melfi.