Joint Action for Social Robotics: How to Build a Robot that Works Together with Several Humans

Manuel Giuliani, Ron Petrick, Kerstin Huth, Amy Isard, Maria Pateraki, Panos Trahanias*

1 June 2011

Abstract

Joint action, the coordination of individual actions by two or more participants working on a common goal, is the basis of many everyday social interactions between humans. However, even though humans engage in such activities, seemingly with ease, how well are the mechanisms underlying such behaviours understood? Can we build a robot that is able to work together with humans?

The aim of this tutorial is to give an introduction to several key technologies that are needed to build a human-robot joint action system. In particular, we explore the topic of joint action from the viewpoint of different research fields—including robotics, computer science, electrical engineering, computational linguistics, and psycholinguistics—all of which we believe contribute to our understanding of joint action. This tutorial will give a technical introduction to the software, tools, and methods that we are using to construct a robot capable of working with humans, in the context of JAMES, a European project exploring human-robot joint action and social interaction.

The talks in this tutorial will cover a variety of topics at the heart of joint action, including the collection and analysis of empirical data from human-human joint action studies, the requirement analysis and implementation of a robot capable of joint action with a human, algorithms for visual processing of human head pose and gestures, grammar-based speech processing and output generation, and knowledge-level planning with incomplete information.

Since the implementation of a human-robot joint action system involves techniques from many diverse research areas, researchers also face the challenge of working together on common goals. Thus, in addition to the technical programme we will also share information concerning best practices for communication in a multi-disciplinary research team.

1 Tutorial Description

1.1 Motivation

Joint action is one of the key abilities allowing humans to achieve goals that cannot otherwise be achieved by individuals alone. Moreover, joint action also plays a central role in many instances of everyday activites involving social interaction, from holding open a door for another person to ordering a drink in a bar. But why are humans so good at joint action? Which mechanisms underlie human-human joint action? Most importantly, do we understand these mechanisms well enough so that we can also implement them on a robot that can work together with humans?

These questions are at the heart of the European project JAMES (Joint Action for Multimodal Embodied Social Systems), which is exploring the mechanisms underlying joint action, with an aim to implement its findings on a robot that will work together with humans. JAMES takes a *multi-disciplinary approach* to this problem, with researchers from a variety of fields such as robotics, computer science, electrical engineering, computational linguistics, and psycholinguistics. In particular, the JAMES partners believe a hybrid approach is necessary to tackle all the challenges that arise when implementing human-robot joint action. For example, a robot interacting with humans on joint tasks requires multimodal communication skills, a means of interpreting and understanding its environment, and the ability to respond to changes in the environment with multiple input and output channels. Achieving this in practice, however, requires a combination of methods from vision understanding, speech processing, planning, and robot motion.

[*] For author affiliations please refer to Section 2

At the heart of JAMES is the idea that social interaction can be viewed as an instance of joint action, and that successful social behaviour by a robot can be achieved by understanding human-human joint action. To guide this approach, JAMES makes four core assumptions about the role of joint action in social interaction:

1. Social interaction should be seen as an instance of joint action. In general, joint action can be defined as the coordination of individual actions by two or more participants in pursuit of a common goal. In the case of social interaction, joint action is inherently multimodal: behaviours such as gesture, gazing, body language, facial expressions, and even natural-language dialogue play a crucial role in interaction.

2. Successful task-based interaction relies on successful social interaction. In a task-driven context, there may be several ways to achieve a goal. However, task-based interactions will often be more successful, leading to higher levels of satisfaction among the participants, if appropriate social behaviour is used.

3. Social interaction is often multi-party, dynamic, and short-horizon. In contrast to long-term, one-on-one, companion-style relationships, many everyday interactions are much shorter and involve constantly changing group dynamics (e.g., ordering a drink at a bar). In such settings, an agent must often get the interaction right the first time, and may not have an opportunity to recover from a poor interaction.

4. Social skills should be learnt rather than preprogrammed. Creating all of the necessary rules by hand to cover every possible social situation is a difficult and time-consuming task. Instead, it makes more sense for an artificial agent to learn appropriate behaviours from the world itself, by observing human-human interactions, and by making use of its own prior experiences to robustly handle new situations.

Following these core assumptions, the JAMES project takes a multi-disciplinary approach to the problem of joint action for social interaction. By focusing on human-human joint action in social contexts, the mechanisms behind such interactions are studied in short-term, everyday activities. These findings are then used to implement components for a social robot that are tailored towards joint action, such as robot motion planning, vision processing for head pose estimation and gesture recognition, grammar-based speech processing and output generation, and knowledge-level planning.

The findings from JAMES will be tested on the robot in Figure 1, and evaluated on a scenario in which the robot acts as a bartender that takes orders from human customers and hands out drinks. This scenario is perfectly suited for illustrating the project's core assumptions, since the robot must be able to quickly recognise the needs of its human interaction partners, and respond in a socially acceptable way. In particular, the robot cannot simply follow predefined interaction patterns, but must dynamically change its actions with respect to a given situation.

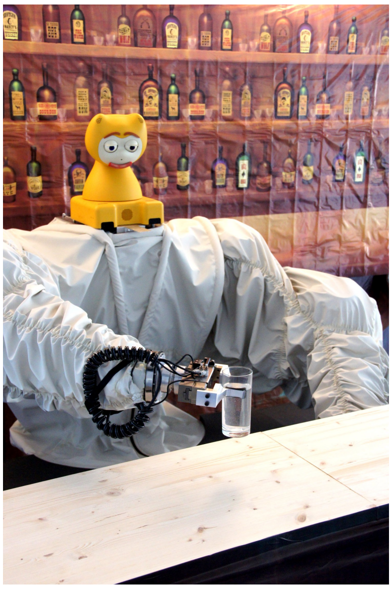

Figure 1: JAMES the social bartender

1.2 Goals and Topics

In short, attendees of this tutorial will receive

- a better understanding of the importance of joint action in the area of social robotics,
- an empirically motivated overview of the basic mechanisms of human-human joint action,
- an overview of several technologies that are needed to build a human-robot joint action system, and
- best practices for communication and collaboration in an international, multi-disciplinary research team.

In this tutorial, we will provide insights into the technologies and methods that the JAMES project is exploring to achieve the goal of implementing human-robot joint action. We will introduce each technology and the problem it addresses, together with a description of how we implement our methods, and the research findings these implementations yield. Concrete examples will be taken from the JAMES bartender scenario which will serve as a case study for each of the tutorial talks.

The tutorial will cover a wide range of multi-disciplinary topics presented by JAMES researchers, based on the core research themes of the project, including robotics, computer science, electrical engineering, computational linguistics, and psycholinguistics. Specific topics that will be presented at the tutorial include:

Empirical data acquisition and analysis This talk gives an overview of the the collection and analysis of natural interactional data in an interdisciplinary research context. The two main topics are: (1) *Collecting natural data.* Preparing and building a corpus. How to obtain maximally naturalistic data with minimal observer interference. (2) *Analysis.* How to plan the analysis in advance in order to obtain results that are useful for the interface between interactional linguistics and computer science. For that, we use an approach that combines Schank and Abelson's script theory and ethnomethodological approaches. The focus will be on the incremental construction of a behavioural script generated from the empirical data, and subsequently creating a visualisable model of the script which suits requirements from both interactional linguistics as well as computer science.
Speaker: **Kerstin Huth** (Universität Bielefeld)

Requirements for robot hardware and architecture This talk is divided into three parts which highlight important requirements for a robot that works together with a human: (1) *Safety.* The safety of the human user has to be ensured at all times. We present methods to increase the safety of the human. (2) *Robot movements.* We show how different robot motion profiles influence how humans perceive the robot and which profile is suitable so that the human can predict the robot's actions. (3) *Robot architecture.* We present our architecture, which is tailored to the robot's task of working together with several humans.
Speaker: **Manuel Giuliani** (fortiss GmbH)

Vision processing for action recognition This talk covers the visual processing components for reliably tracking location, hand gestures, and facial displays of multiple people in a constantly changing scene, which we develop for the JAMES project. Our aim is to automatically analyse and recognise hand and facial motions and facial feature changes from visual information in order to convey the intention of the person in terms of communicative signs. Highlighted topics are: (1) the use of probabilistic techniques to track humans by fusing information from various visual input channels, e.g. colour, depth, motion, (2) detection of hand activity as well as specific hand motion patterns and gestures that occur spontaneously while a person speaks and support/expand the information conveyed by words, and (3) recognition of head gestures, head pose and facial expressions.
Speakers: **Maria Pateraki, Panos Trahanias** (Foundation for Research and Technology - Hellas)

Grammar-based speech processing and output generation The language processing on the JAMES project is carried out using the OpenCCG system, which allows for bidirectional grammars. This means that the same grammar is used for parsing and generation, and the utterances of the user are parsed into logical forms of the same type as those used to create the robot's language output. The speech processing will be optimised with adaptive grammars compiled after each system output to reflect the expected user responses. The output will combine speech with robot gestures and facial expressions where appropriate.
Speaker: **Amy Isard** (University of Edinburgh)

Knowledge-level planning with incomplete information The ability to plan is essential for an intelligent agent acting in a dynamic and incompletely known world, such as the scenario we consider in JAMES. Achieving goals under such conditions often requires complex forward deliberation rather than simply reacting to a situation without considering the long term consequences of actions. This talk gives an overview of automated planning techniques and describes in detail the knowledge-level PKS (Planning with Knowledge and Sensing) planner used in JAMES. We focus on the problem of planning with incomplete information and sensing actions, and demonstrate how PKS can be used to build plans for joint action in robot domains, including those involving both physical tasks and social interaction.
Speaker: **Ron Petrick** (University of Edinburgh)

Besides the technological challenges involved in building a robot that is able to work together with humans, researchers also have to face the challenge of working together with people from different research backgrounds, to achieve a common goal. In the case of JAMES, this involves communication between researchers from five different research fields, all of which use different terminology and are spread across several European countries. Thus, following the technology presentations, we will conclude the tutorial with a panel discussion by the presenters. They will discuss best practices for teamwork and communication in a multi-disciplinary project, and effective methods for transferring knowledge and results between project partners. As part of this discussion, the audience will be asked to share their experience in multi-disciplinary collaboration.

2 Speakers

Kerstin Huth is a linguistics graduate student from the Universität Bielefeld and a research assistant in the JAMES project. Her research interests are applied pragmatics, interactional linguistics and conversational analysis.

Manuel Giuliani is a research associate at fortiss, an affiliated institute of Technische Universität München. He received a Master of Arts in computational linguistics in 2004 and a Master of Science in Computer Science in 2006. He worked on the European project JAST (Joint Action Science and Technology) and is now part of JAMES. His research interests include social robotics, human-robot interaction, natural language processing, multimodal fusion, and robot architectures.

Maria Pateraki is currently a postdoctoral researcher of the Computational Vision and Robotics Lab of ICS-FORTH. She has received her PhD in photogrammetry from the Swiss Federal Institute of Technology in Zurich (ETHZ) , Switzerland in 2005 and has been a research associate at the University of Melbourne and the Cooperative Research Center for Spatial Information (CRC-SI) in Melbourne, Australia. Her interests include topics related to robotic vision, motion tracking, registration, matching and 3D reconstruction.

Prof. Panos Trahanias is with the Dept. of Computer Science, Univ. of Crete, Greece and FORTH-ICS. In the past he has been affiliated with the Dept. of Electrical & Computer Eng., Univ. of Toronto, Canada, and was a consultant to SPAR Aerospace Ltd., Toronto. Currently, he is the Director of Graduate Studies at the Department of Computer Science, University of Crete, and the Head of the Computational Vision & Robotics Laboratory at ICS-FORTH where he is engaged in and supervises research and R & D projects in brain-based modeling, human-robot interaction, robot navigation, remote-access robotic systems and augmented reality applications. He has published over 100 papers in technical journals and conference proceedings and has contributed in two books

Amy Isard is a research fellow in the School of Informatics at the University of Edinburgh. Her research interests include natural language processing and dialogue systems. She recived a BA in French and German from the University of Cambridge and an MSc in Artificial Intelligence from the University of Edinburgh. She has worked on several EU projects including M-PIRO, Indigo, JAST and currently JAMES.

Ron Petrick is a research fellow in the School of Informatics at the University of Edinburgh. He received an MMath degree in computer science from the University of Waterloo and a PhD in computer science from the University of Toronto. His research interests include planning with incomplete information and sensing, cognitive robotics, knowledge representation and reasoning, generalised planning, and natural language dialogue. He previously worked on the European PACO-PLUS project, and is currently the Scientific Coordinator of the JAMES project.

3 Schedule

Table 1 shows the preliminary schedule for the JAMES tutorial. We are planning for a full-day tutorial.

4 Intended Audience

This tutorial is aimed at PhD students and young researchers who are interested in the technical details of joint action in a human-robot interaction system. Prior knowledge in one or more of the tutorial topics is helpful, but each talk will provide a short introduction to the covered topic as background for following the tutorial.

Time	Topic / Speaker
9:00am	Introduction / Overview
9:30am	Empirical data acquisition and analysis
	Kerstin Huth
10:30am	*coffee break*
11:00am	Requirements for robot hardware and architecture
	Manuel Giuliani
12:00pm	Vision processing for action recognition
	Maria Pateraki, Panos Trahanias
1:00pm	*lunch*
2:00pm	Grammar-based speech processing and output generation
	Amy Isard
3:00pm	Knowledge-level planning with incomplete information
	Ron Petrick
4:00pm	*coffee break*
4:30pm	Wrap-up session and panel discussion

Table 1: JAMES tutorial schedule

5 Scope of tutorial

The number of participants is not limited.