UNIVERSITÀ DEGLI STUDI DI NAPOLI
FEDERICO II

FACOLTÀ DI INGEGNERIA

Dipartimento di Ingegneria Elettrica e Tecnologie dell'Informazione

Tesi di Dottorato in Ingegneria Informatica e Automatica

Coordinatore: Prof. Francesco Garofalo

Human Gesture Recognition and Robot Attentional Regulation for Human-Robot Interaction

Salvatore Iengo
salvatore.iengo@unina.it

In Partial Fulfillment of the Requirements for the Degree of
PHILOSOPHIAE DOCTOR in
Computer Science and Automation Engineering

Supervisor
Prof. Luigi Villani

Co-Supervisor
Dr. Alberto Finzi

April 2014

«Non coronabitur nisi qui legitime certaverit»

1897

Abstract

Human-Robot Interaction (HRI) is defined as the study of interactions between humans and robots: it involves several different disciplines like computer science, engineering, social sciences and psychology. For HRI, the perceptual challenges are particularly complex, because of the need to perceive, understand, and react to human activity in real-time. The main key aspects of the perception are multimodality and attention. Multimodality allows humans to move seamlessly between different modes of interaction, from visual to voice to touch, according to changes in context or user preference, while attention is the cognitive process of selectively concentrating on one aspect of the environment while ignoring other things.

Multimodality and attention play a fundamental role in HRI also. Multimodality allows robot to interpret and react to various humans' stimuli (e.g. gesture, speech, eye gaze) while, on the other hand, implementing attentional models in robot control behavior allows robot to save computational resources and react in real time by selectively processing the *salient* perceived stimuli.

The intention of this thesis is to present novel methods for human gestures recognition including pointing gestures, that are fundamental when interacting with mobile robots, and a robot attentional regulation mechanism that is speech driven.

In the context of continuous gesture recognition the aim is to provide a system that can be trained online with few samples and can cope with intra user variability during the gesture execution. The proposed approach relies on the generation of an ad-hoc Hidden Markov Model (HMM) for each gesture exploiting a direct estimation of the parameters. Each model represents the best prototype candidate from the associated gesture training set. The generated models are then employed within a continuous recognition process that provides the probability of each gesture at each step. A pointing gesture recognition computational method is also presented, such model is based on the combined approach a geometrical solution and a machine learning solution.

Once the gesture recognition models are described, a human-robot interaction system that exploits emotion and attention to regulate and adapt the robotic interactive behavior is proposed. In particular, the system is focused on the relation between arousal, predictability, and attentional allocation considering as a case study a robotic manipulator interacting with a human operator.

Acknowledgements

First and foremost I offer my sincerest gratitude to my supervisors Prof. Luigi Villani and Dr. Alberto Finzi who have supported me throughout my thesis with their patience and knowledge whilst allowing me the room to work in my own way. I attribute the level of my PhD degree to their encouragement and effort and without them this thesis, too, would not have been completed or written. One simply could not wish for a better or friendlier supervisors.

I would like to express my special appreciation and thanks to Prof. Bruno Siciliano. His patience, encouragement, and immense knowledge were key motivations throughout my PhD.

I am most grateful to Prof. Ernesto Burattini for his valuable assistance and guidance, since the beginning of my undergraduate studies, in getting my graduate career started on the right foot and providing me with the foundation for becoming a computer scientist.

I would also like to thank Prof. Maja Matarić and my friend Dr. Ross Mead for hosting me, during the final part of my PhD program, in the "Interaction Lab" at University of Southern California Viterbi School of Engineering in Los Angeles. I consider my time there to be the one of the most valuable and significant experiences of my PhD pursuits. Working at USC with the Professor Matarić's group has been one of the most exciting and stimulating times of my life, and I have no words to thank them all for it.

Special thanks go to my good friend and colleague Dr. Haris Balta for providing me thousand of stimulating research ideas in the field. In many ways I have learnt much from him.

The Department of Computer Engineering, the PRISMA Lab, PRISCA Lab and Interaction Lab have provided the support and equipment I have needed to produce and complete my thesis and my studies.

I thank my family for supporting me throughout all my studies at University and for providing a home in which to complete my writing up.

Finally, but most importantly, I would to thank Ivana. Her support, encouragement, endless patience and unwavering love were undeniably the bedrock upon which the past four years of my life have been built.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Human-robot interaction (HRI) is the interdisciplinary study of interaction dynamics between humans and robots. Researchers and practitioners specializing in HRI come from a variety of fields, including engineering (electrical, mechanical, industrial, and design), computer science (human-computer interaction, artificial intelligence, robotics, natural language understanding, and computer vision), social sciences (psychology, cognitive science, communications, anthropology, and human factors), and humanities (ethics and philosophy).

As stated in [1], robots are poised to fill a growing number of roles in today's society, from factory automation to service applications to medical care and entertainment. While robots were initially used in repetitive tasks where all human direction is given a priori, they are becoming involved in increasingly more complex and less structured tasks and activities, including interaction with people required to complete those tasks. This complexity has prompted the entirely new endeavor of Human-Robot Interaction (HRI), the study of how humans interact with robots, and how best to design and implement robot systems capable of accomplishing interactive tasks in human environments. The fundamental goal of HRI is to develop the principles and algorithms for robot systems that make

them capable of direct, safe and effective interaction with humans. Many facets of HRI research relate to and draw from insights and principles from psychology, communication, anthropology, philosophy, and ethics, making HRI an inherently interdisciplinary endeavor.

The study of HRI contains a wide variety of challenges, some of them of basic research nature, exploring concepts general to HRI, and others of domain-specific nature, dealing with direct uses of robot systems that interact with humans in particular contexts. Real-time perception and dealing with uncertainty in sensing are some of the most enduring challenges of robotics. For HRI, the perceptual challenges are particularly complex, because of the need to perceive, understand, and react to human activity in real-time. The main key aspects of the perception are multimodality and attention. Multimodality allows humans to move seamlessly between different modes of interaction, from visual to voice to touch, according to changes in context or user preference, while attention is the cognitive process of selectively concentrating on one aspect of the environment while ignoring other things.

Multimodality and attention play a fundamental role in HRI also. Multimodality allows robot to interpret and react to various humans' stimuli (e.g. gesture, speech, eye gaze) while, on the other hand, implementing attentional models in robot control behavior allows robot to save computational resources and react in real time by selectively processing the *salient* perceived stimuli.

The main motivation for this thesis is the refinement of Human-Robot interaction mechanisms by empowering them to take into account gestures performed by the human in a flexible, fast and natural way and by the means of an attentional control architecture that enables the robot to quickly react to the users' stimuli.

1.2 The SAPHARI European Project

This thesis takes place in the context of the European project SAPHARI.

Recent progress in physical Human-Robot Interaction (pHRI) research showed in principle that human and robots can actively and safely share a common workspace. The fundamental breakthrough that enabled these results was the human-centered design of robot mechanics and control. This made it possible to limit potential injuries due to unintentional contacts. Previous projects, in particular the PHRIENDS project in which a part of the consortium has been involved, provided remarkable results in these directions, constituting the background foundation for this proposal.

Inspired by these results, SAPHARI will perform a fundamental paradigm shift in robot development in the sense that the human is placed at the centre of the entire design. The project will take a big step further along the human-centered roadmap by addressing all essential aspects of safe, intuitive physical interaction between humans and complex, human-like robotic systems in a strongly interconnected manner.

While encompassing safety issues based on biomechanical analysis, human-friendly hardware design, and interaction control strategies, the project will develop and validate key perceptive and cognitive components that enable robots to track, understand and predict human motions in a weakly structured dynamic environment in real-time.

SAPHARI will equip robots with the capabilities to react to human actions or even take the initiative to interact in a situation-dependent manner relying on sensor based decisions and background knowledge.

Apart from developing the necessary capabilities for interactive autonomy, the goal is also to tightly incorporate the human safety also at the cognitive level. This

will enable the robots to react or physically interact with humans in a safe and autonomous way. Keeping in mind the paradigm to "design for safety and control for performance", research developments will be pursued in several areas, starting with the fundamental injury mechanisms of humans cooperating with robots. The analysis will be first carried out for stiff robots and then extended to variable stiffness actuation systems in terms of safety, energy, and load sustainability. Biomechanical knowledge and biologically motivated variable compliance actuators will be used to design bimanual manipulation systems that have design characteristics and performance properties close to humans. Real-time task and motion planning of such complex systems requires new concepts including tight coupling of control and planning that lead to new reactive action generation behaviours.

Safe operation will be enforced in mobile manipulation scenarios with large workspaces by smart fusion of proprioceptive and exteroceptive sensory information, sensor-based task planning, human gestures and motion recognition and learning, and task-oriented programming, including configuration and programming of safety measures.

Finally, self explaining interaction and communication frameworks will be developed to enhance the system usability and make the multimodal communication between human and robot seamless.

The project focuses on two industrial use cases that explicitly contain deliberate physical interaction between a human and a robot co-worker, as well as on professional service scenarios in hospitals, in which medical staff and an assisting robot interact closely during daily work. These prototypical applications will pave the way towards new and emerging markets, not only in industry and professional services, but possibly also in household robots, advanced prostheses and rehabilitation devices, teleoperation, and robotic surgery. Generally, results of this project are expected to strongly impact all applications where interactive robots can assist humans and release them from dangerous or routine tasks.

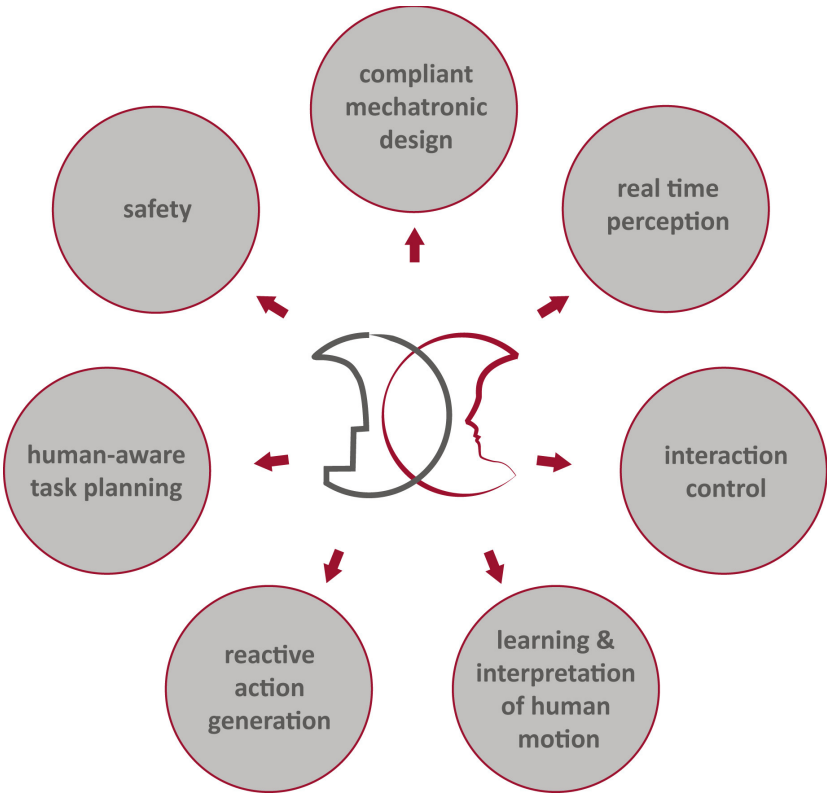


FIGURE 1.1: SAPHARI Project goals.

1.3 Thesis Scope and Objectives

The main focus of this work is on studying human communication to derive accurate models for gesture recognition, which ultimately inform the robot on how to react to human intentions. In addition, a second focus is on the robot attentional regulation through gestures and speech. Accordingly, this work affects many research areas. The original contribution of this thesis is a description of fast, simple and reliable method for human gesture recognition and interpretation through attentional regulation. Following this research program, the following challenges have been identified:

1. Fast and reliable models for human gesture recognition.

2. Accurate model for human pointing

3. Computationally light methods for attentional robot control through gestures and speech.

The four main contributions of this thesis are detailed in the following sections.

Gesture recognition

This thesis aims at developing models for fast and reliable human continuous gestures recognition that can be easily adapted to the user with a very limited training session. For this, a model for continuous dynamic gesture recognition is constructed and implemented. Dynamic gestures recognition aims to solve the problem of *what* is the user trying to describe.

Human pointing

Considering that pointing gestures recognition requires a completely different approach from classical gesture recognition, a separated model for such gestures is provided. Pointing gestures recognition deals with the problem of finding *where* the user's attention directed to.

Robot attentional control

Once that the user's intentions are recognized a method for robot's response based on attentional and adaptive behavior regulation is proposed. That method has the capability of adapting robot's behaviors to the rate of change of both the environment and its internal states reducing the computational costs of input processing.

1.4 Thesis Structure

Human-Robot interaction through gestures has been observed and modeled by several scientific disciplines and thus the **Chapter 2** summarizes and interdisciplinary review in the field and some technical aspects.

Chapter 3 addresses the problem of the gesture recognition as a classification problem. This chapter describes and extensive study on a flexible gesture recognition approach capable of providing a continuous recognition with a good accuracy and a small training set. For the analysis and recognition of gestures an ad-hoc created Hidden Markov Model is described. The results are validated through three case studies: a letter recognition case study, a case study conducted on a MRSC-12 dataset and a robotic case study.

Deictic gesture is investigated in **Chapter 4**. A computational model for deictic gesture recognition is described. Such model is based on two solutions: a geometrical solution and a regression analysis solution. In the end of the chapter an integration mechanism for multimodality in deixis is described.

Chapter 5 explores the interplay between attentional and emotional regulation in human-robot interaction. More specifically, an architecture is defined where attention allocation and emotional processes can influence the robotic interactive behavior adapting it to the human emotions, intentions, and expectations.

Chapter 2

Interdisciplinary Background

This chapter covers the main topics related to HRI as stated in the literature: the first part of this chapter provides an overview of the non-verbal communication focusing on gestures. The following parts describe linked topic like, social deixis, manual pointing, gaze pointing and speech relation and the main technologies available. The last part of the chapter concentrates on social robotics, robot behavior and attentional regulation.

2.1 Non verbal communication

According to [2] more than 65% of human communication is non-verbal. In our everyday life we respond to several non-verbal cues and behaviors like postures, eye gaze, facial expression and tone of voice.

The first scientific research on nonverbal communication and behavior starts with *The Expression of the Emotions in Man and Animals* of Charles Darwin and nowadays there are an abundance of research on the types, effects and expression of unspoken communication and behavior.

As stated in [3] research has identified several different types of nonverbal communication:

- **Facial expression**.

  Facial expressions are responsible for a huge proportion of nonverbal communication. While nonverbal communication and behavior can vary dramatically between cultures, the facial expressions for happiness, sadness, anger and fear are similar throughout the world.

- **Gestures**.

  Deliberate movements and signals are an important way to communicate meaning without words. Common gestures include waving, pointing, and using fingers to indicate numeric amounts. Other gestures are arbitrary and related to culture.

- **Paralinguistic**.

  Paralinguistics refers to vocal communication that is separate from actual language. This includes factors such as tone of voice, loudness, inflection and pitch. The tone of voice can have powerful effect on the meaning of a sentence. When said in a strong tone of voice, listeners might interpret approval and enthusiasm. The same words said in a hesitant tone of voice might convey disapproval and a lack of interest.

- **Body language and posture**.

  Posture and movement can also convey a great deal on information. Research on body language has grown significantly since the 1970's, but popular media have focused on the over-interpretation of defensive postures, arm-crossing, and leg-crossing, especially after the publication of Julius Fast's book Body Language. While these nonverbal behaviors can indicate feelings and attitudes, research suggests that body language is far more subtle and less definitive that previously believed.

- **Proxemics**.

  People often refer to their need for *personal space*, which is also an important type of nonverbal communication. The amount of distance we need and the amount of space we perceive as belonging to us is influenced by a number of factors including social norms, situational factors, personality characteristics and level of familiarity.

- **Eye gaze**.

  Looking, staring and blinking can also be important nonverbal behaviors. When people encounter people or things that they like, the rate of blinking increases and pupils dilate. Looking at another person can indicate a range of emotions, including hostility, interest and attraction.

- **Haptics**.

  Communicating through touch is another important nonverbal behavior. There has been a substantial amount of research on the importance of touch in infancy and early childhood. Harry Harlow's classic monkey study demonstrated how the deprivation of touch and contact impedes development. Touch can be used to communicate affection, familiarity, sympathy and other emotions.

- **Appearance**.

  Our choice of color, clothing, hairstyles and other factors affecting appearance are also considered a means of nonverbal communication. Research on color psychology has demonstrated that different colors can evoke different moods. Appearance can also alter physiological reactions, judgments and interpretations. Just think of all the subtle judgments you quickly make about someone based on his or her appearance. These first impressions are important, which is why experts suggest that job seekers dress appropriately for interviews with potential employers.

2.2 Gestures in Human Communication

As reported in the Oxford Concise dictionary a gesture is *a movement of a limb or the body as an expression of thought or feeling.* Nevertheless not every movement is gesture. It's important to distinguish what movements probably are gestures, and which probably are not.

In [4] Kendon conducted a study to find out "...whether or not people did consistently recognize only certain aspects of action as belong to gesture." The following summarizes his main observations:

Kendon's [4] sees the following types of actions:

- **Limb movements**: where the limb lifted sharply from the body, and subsequently returned to the same position from which it started;
- **Head movements**: such as rotations or up-down movements if rapid or repeated, and if not leading the head to be held to a new position, and if not done in coordination with eye movements;
- **Movements of the whole body**: regarded as gesture if it was seen as returning to the position from which it began, and not resulting in a sustained change in spatial location or bodily posture.
- **Movement involving the manipulation of an object**: manipulations such as changing the position of an object were never seen (by subjects) as expression.

Kendon's definition of gesture: "The word 'gesture' serves as a label for that domain of visible action that participants routinely separate out and treat as governed by an openly acknowledged communicative intent." [5].

2.3 Gestures Taxonomy

Gestures can be classified according to their function. In [6] functions are used to group gestures into three types:

- **semiotic**: those used to communicate meaningful information.
- **ergotic**: those used to manipulate the physical world and create artifacts
- **epistemic**: those used to learn from the environment through tactile or haptic exploration

In this thesis we are primarily interested in how gestures can be used to communicate with a robot so we will be mostly concerned with semiotic gestures. These can further be categorized according to their functionality. A useful taxonomy of gestural types is one offered by [7] (see Figure 2.1):

- Symbolic gestures: These are gestures that, within each culture, have come to have a single meaning. An Emblem such as the "OK" gesture is one such example, however American Sign Language gestures also fall into this category.
- Deictic gestures: These are the types of gestures most generally seen in HRI and are the gestures of pointing, or otherwise directing the listeners attention to specific events or objects in the environment. They are the gestures made when someone says "Move over there".
- Iconic gestures: As the name suggests, these gestures are used to convey information about the size, shape or orientation of the object of discourse. They are the gestures made when someone says "the two car was parked like this", while aligning their hand like the position of the car in the parking .

- Pantomimic gestures: These are the gestures typically used in showing the use of movement of some invisible tool or object in the speaker's hand.

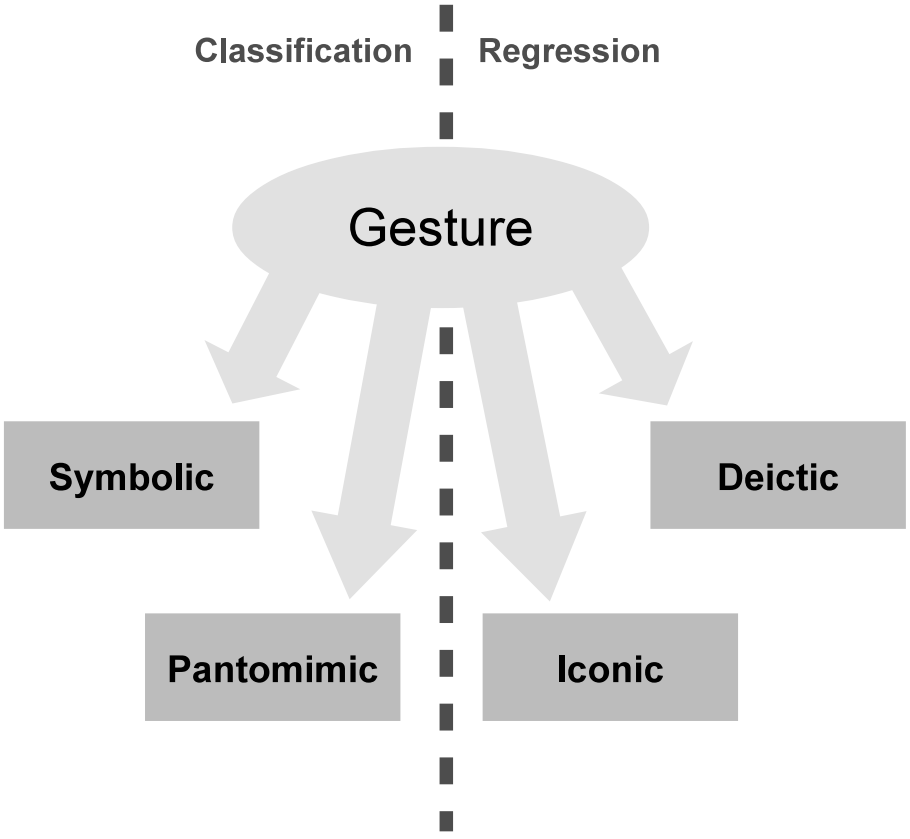


FIGURE 2.1: Human gesture taxonomy.

The very first consideration that can be made from the gesture taxonomy mentioned above is that while *symbolic* and *pantomimic* gestures can be treated, in the gesture recognition context, as a classification problem *iconic* and *deictic* gestures must be treated as a regression problem due to the fact that these gestures refer to a quantity that have to be estimated (the position of the object pointed by the gesture). In this thesis we are interested in symbolic gestures for continuous recognition (classification) and in deictic gestures for manual pointing (regression).

2.4 Social Deixis

Levinson in [8] stated that social deixis concerns with the aspects of sentences which reflect the social situation in which the speech event occurs. He points out that there are two basic kinds of social deixis information that seems to be encoded in all the languages: relational social deixis and absolute social deixis. Relational social deixis is a deictic reference to some social characteristic of referent apart from any relative ranking of referents or deictic reference to a social relationship between the speaker and addressee. Absolute social deixis is a deictic reference usually expressed in certain forms of address which will include no comparison of the ranking of the speaker and addresse.

Summarizing, social deixis can be mainly classified in:

- **Person Deixis**: Person deixis concerns with the encoding of the role of participants in the speech (e.g. "me", "you").
- **Place Deixis**: often referred as spatial deixis, where the relative location of objects is being indicated. (e.g. "that", "these", "here", "there").
- **Time Deixis**: refers to relative or absolute time in the utterances (e.g. "now", "tomorrow", "later", "at 4:00pm").

In this work we focus on *manual deixis*: a specialization of *place deixis* (e.g. manual pointing to an object or place)

2.5 Manual Deixis

The first consideration about manual deixis concerns the structure of pointing. As reported in [9] there is evidence that pointing with the whole hand serves a different

function for young humans from pointing with the index finger. Butterworth (e.g., [10]) suggested that pointing with the whole hand serves to request objects or actions on objects, whereas pointing with the index finger serves to "comment" upon something in the world. In a cross-sectional study, Franco and Butterworth [10] found that whole-handed gestures did not change in relative frequency from 12 to 18 months of age, whereas the incidence of index-finger pointing increased dramatically over the same age range. Taken together, these findings suggest that the form of pointing in humans is sensitive to contextual manipulations. Thus some humans exhibit an overwhelming reliance on the index finger for pointing and others seem to prefer to indicate distant objects with their whole hands extended [11]. In this thesis we rely on the whole arm position for estimating the pointing direction.

2.6 Gaze Pointing

Eye gaze is a very important source of communication between humans. The eyes operate as an input channel to someone who is observing, and as an output channel to others who may witness the activity of the observer's eyes.

Many researchers in human-computer interaction (HCI) regard the eyes potentially as a pointer to material on display: the user looks (gaze) at some item or area, and either by blinking or sustaining the gaze for some time-out period, the item or area is selected, the whole process not unlike clinking on a mouse.

As described in [11] the eye is, in fact, an excellent "pointer" in that one can fixate some spot, look away, and come back right on target. And, I would agree that if the only output modality available to the user is line-of-sight, as with someone severely disabled, then such use makes much sense.

An interesting consideration with respect to hand pointing is that Information coming from eye and hand may at times confict each other (e.g. the user could say "Move near that object" (looking to a specific object) while pointing to a different location with hand).

As a general rule, pointing by hand is more a significant and deliberate than looking; on that basis, a reasonable strategy is to refer to the area where the user is pointing.

In instances wherein the placement of the item is critical and the circumstances of the entire operation are of such a nature as to not tolerate errors, then the system might well insist that the user be both looking and pointing at same spot at the same time [12].

2.7 Speech and gestures

Speech and gesture relation is well summarized in [13] where is stated that use of gesture is most powerful when combined with other input modalities, especially voice. Allowing combined voice and gestural input has several tangible advantages. The first is purely practical: ease of expression. Typical computer interaction modalities are characterized by an ease versus expressiveness trade-off [14]. Ease corresponds to the efficiency with which commands can be remembered, and expressiveness the size of the command vocabulary. Common interaction devices range from the mouse that maximizes ease, to the keyboard that maximizes expressiveness. Multimodal input overcomes this trade-off; combined speech and gestural commands are easy to execute whilst retaining a large command vocabulary. Voice and gesture complement each other and when used together, creating an interface more powerful that either modality alone. In [15] Cohen shows how natural language interaction is suited for descriptive techniques, while gestural interaction is ideal for direct manipulation of objects. For example, unlike gestural

or mouse input, voice is not tied to a spatial metaphor [16]. This means that voice can interact with objects regardless of degree of visual exposure, particularly valuable in a virtual environment where objects may be hidden inside each other or occluded by other objects. Some tasks are inherently graphical, others are verbal and yet others require both vocal and gestural input to be completed [17]. So allowing both types of input maximizes the usefulness of an interface by broadening the range of tasks that can be done in an intuitive manner.

2.8 Gesture tracking technologies

A fundamental question while working on gesture recognition models is what kind of sensor use. In this section are reported the main tracking technologies for gestures.

Basically two main approaches exist:

- **Vision based approach**: involves one or more video cameras.
- **Non-vision-based approach**: device like gloves, accelerometers or joysticks.

In the following the two technologies are discussed.

2.8.1 Non-vision based sensors

There are many non-vision sensors: the most used are accelerometers, touchscreens and gloves. We can list five categories of contact sensors:

1. **Mechanical**: a typical sensor that detects movement in the space (e.g. gyroscopes).

2. **Inertial**: a sensor that measures acceleration of an object (e.g. accelerometers).

3. **Haptic**: a typical sensor that detects physical contact with an object.

4. **Magnetic**: detects changes in magnetic field. There are several health issues to be considered with these kind of sensors.

5. **Ultrasonic**: sonar like sensors useful to detect distances between objects (range finder).

FIGURE 2.2: A data-glove sensor for hand movement detection

Figure 2.2 shows a typical hand movement sensor with accelerometer, gyroscopes and contact sensors used for fingers orientation detection.

2.8.2 Vision based technologies

Vision based approaches rely on one or more cameras to detect and analyze body movement from the video sequences.

There are different camera sensors:

- **Infrared camera**: can detect movement with no light condition.
- **Monocular camera**: is the most popular and cheap camera. Typically monocular cameras are used with sensor markers. Markers can be passive if don't emit light or active otherwise.

TABLE 2.1: Vision based VS Non-vision based sensors.

Criteria	Non-vision based	Vision based
Invasive	Yes	No
Precision	Yes	No
Configuration flexibility	Yes	No
Flexible use	No	Yes
Occlusion problems	No	Yes
Health issues	Yes	No

- **Stereocameras**: stereovision can detect 3D model of the objects from the scene allowing object movement detection in three dimensions.
- **PTZ cameras**: pan-tilt-zoom camera embodies a robotic movement engine that enables the movement along three axis.

2.8.3 Advantages and disadvantages of detection technologies

Both technologies for gesture detection have advantages and disadvantages. For example, contact sensors must be necessarily worn by the user during the interaction period, while camera sensors suffer of occlusion problems.

The Table 2.1 summarizes both advantages and disadvantages.

The sensor used in the work described in this thesis is the Microsoft Kinect.

2.8.3.1 Microsoft Kinect sensor

Microsoft Kinect is a motion sensing input devices by Microsoft for Xbox 360 and Xbox One video game consoles and Windows PCs. Based around a webcam-style add-on peripheral, it enables users to control and interact with their console/-computer without the need for a game controller, through a natural user interface

using gestures. The first-generation Kinect was first introduced in November 2010. Kinect allows skeletal tracking of people and follow their actions. Using the infrared (IR) camera, Kinect can recognize up to six users in the field of view of the sensor. Of these, up to two users can be tracked in detail. An application can locate the joints of the tracked users in space and track their movements over time (see Figure 2.3.

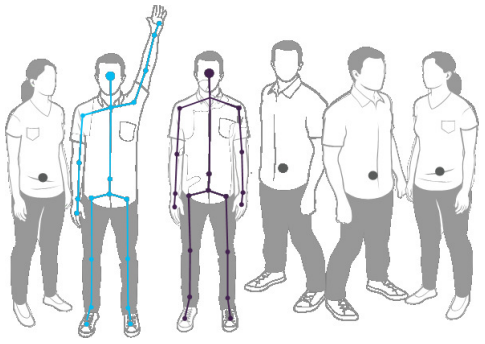


FIGURE 2.3: Skeleton tracking of multiple users performed by Kinect sensor

Microsoft has also provided a software development kit for Kinect through which the low-level data streams from the Kinect video, microphone, and depth sensors can be accessed. This SDK provided by Microsoft is capable of tracking skeletal data, too. It can track the skeleton image of one or two people moving within the Kinects field of view. We use this feature for recognizing different human postures.

2.9 Social Robotics

Socially intelligent robotics is the pursuit of creating robots capable of exhibiting natural-appearing social qualities. Beyond the basic capabilities of moving and acting autonomously, the field has focused on the use of the robot's physical embodiment to communicate and interact with users in a social and engaging manner. One of its components, socially assistive robotics, focuses on helping human users through social rather than physical interaction [18]. The study of human-robot

interaction (HRI) for socially assistive robotics applications is a new, interdisciplinary and increasingly popular research area that brings together a broad spectrum of research including robotics, medicine, social and cognitive sciences, and neuroscience, among others. Assistive robotics in general and socially assistive robotics in particular have the potential to enhance the quality of life for broad populations of users: the elderly, individuals with physical impairments and those in rehabilitation therapy, and individuals with cognitive disabilities and developmental and social disorders.

2.10 Robot behaviors

Brooks [19] stated the importance of grounding robot architectures in real-world tasks as well as stressed that a complex system (in this case, a robot) can execute correct behavior, even without an identifiable reason for that behavior. Following that philosophy, the notion of behavior-based control (BBC) and architectures for BBC [20] was conceived. The control code of a robot is divided into task-achieving modules called behaviors. These behaviors run in parallel. Behavior-based architectures form the basis for much of modern- day architecture development, including the architecture presented in this report. As reported in [21], with behavior-based architectures, the fact that many behaviors are running in parallel can be troubling in that two or more behaviors may give conflicting outputs. One way to arbitrate control is through sub-sumption [20]. Behaviors are arranged in order from lowest- to highest-level. Lower-level behavior will be suppressed by higher-level behavior. When strict ordering of the behavior hierarchy is not possible a priori, there must be a means to arbitrate conflicting instructions. In the DAMN architecture [22], behaviors cooperatively determine the robot's trajectory by voting on possible options, relying on a command arbiter to fuse the instructions. Maes and Brooks [23] developed an architecture where behaviors learn when and when not to activate through feedback. Michaud

and Audet [24] developed a system that uses artificial emotions to arbitrate conflicting goals. Plans, and the handling of plans, are of particular importance for robot control architectures. Some architectures eliminate plans altogether in favor of a completely reactive system [20], which relies on emergence to form complex behavior. Hybrid architectures have been developed that combine deliberative planning with a reactive system. These systems can be used for motion-planning [25–27], combining a reactive obstacle-avoidance system with a deliberative planner for goal-oriented navigation. Agre and Chapman [28] discusses the two uses of plans. First, plans can be used for execution, carrying out the instructions of the plan. Second, plans can be used for communication, relaying its intentions to another agent. Considering plans in this framework can be helpful for architecture design. Initially, robot architectures were concerned with movement [29], map-making [30], and path planning [31]. These systems did not emphasize perception, dialog, or human-robot interaction. Some architectures are implemented with tele-operated human control in mind. Fong et al. [32] implemented an architecture that allowed a robot to perform as a member of a team, facilitating user control through a hand-held interface. The field of HRI has been working on several fronts to better understand how humans interact with robots. To that end, several architecture innovations have been developed that facilitate better human-robot interaction. The next three sections describe such innovations relating to sensing and perception, developmental robotics, and robots designed to assist with behavior interventions.

2.11 Attentional Regulation

Attentional mechanisms applied to autonomous robotic systems have been proposed in [33, 34], mainly for vision-based robotics. In our work, we are interested in artificial attentional processes suitable for the executive control. In particular, our aim is to provide a kind of supervisory attentional system [35, 36] capable

of monitoring and regulating multiple concur- rent behaviors at different level of abstraction. The notion of divided attention [37] suggests that a limited amount of attention is allocated to tasks, with the resources involved in multi-task performances, and can be available in graded quantity. In an artificial setting, this can be obtained by introducing suitable scheduling mechanisms. In this work, we present a behavior-based control architecture endowed with attentional mechanisms which are based on periodic releasing mechanisms of activations [38]. In this context, each behavior is equipped with an adaptive internal clock that regulates the sensing rate and the resulting action activations. The process of changing the frequency of sensory readings is interpreted as an increase or decrease of attention towards relevant behaviors and particular aspects of the external environment: the higher is the frequency, the higher is the resolution at which a process is monitored and controlled. Here, we present our framework providing several case studies where we discuss the effectiveness of the approach considering its scalability and the adaptivity with respect to different environments and tasks.

Chapter 3

Continuous Gestures Recognition

3.1 Introduction

The human communication strongly relies on gestures for sharing a variety of feelings and thoughts, often together with body language in addition to speech. Consequently, a gesture can be considered to be a communicative human movement. From its beginnings, the driving vision of robotics has been to create systems which are capable of understanding human intentions without having us to learn dedicated user interfaces like a computer keyboard or a control stick. As a result of this vision the current research effort is primarily based on developing communicative systems, where the human's way of interacting governs interface design making the robots adapt to humans, not the other way round.

Today, many gesture recognition applications involve hard and hand coded strategies to recognize gestures. A more promising approach is that suggested by traditional pattern recognition techniques, where a number of example gestures are collected and subsequently summarized automatically by a process which fits a compact model to the collection of training signals. Later, when the application is running, the models are matched to the input signal. When a model matches the

signal well, the application may conclude that the gesture corresponding to the model occurred. Hidden Markov models are one such framework for the automatic learning of gestures for later recognition.

The main lack of this methods is that cannot easily be adapted to new user and new gestures for such robot interaction tasks where the human operator needs to quickly train the robot in order to face unforeseen needs.

In this chapter, we present a novel approach to real-time and continuous gesture recognition that allows a flexible, natural, and robust human-robot interaction (HRI). The proposed system should support the social interaction between the human and the robot by enabling a continuous process of evaluation and interpretation of the reciprocal movements. Furthermore, the proposed methodology should permit an incremental development of the HRI system through simple training and modular insertion of new gestures.

3.2 Related Work

In literature, we find several approaches to gesture recognition. Most of the them are based on statistical modeling, such as Principal Component Analysis (PCA), multi-dimensional Hidden Markov Models (HMM) [39–42], Kalman filtering, and condensations algorithms. On the other side, Finite State Machines (FSM) has been effectively used in modeling human gestures [43, 44]. Connectionist approaches involving neural networks have been also explored, such as time-delay neural network (TDNN) [45]. In HMM approaches the models are employed to represent the gestures and their parameters are learned from the training data. Based on the most likely performance criterion the gestures can be recognized through evaluating the trained HMMs [40],[41] and [42]. FSM methods for gesture recognition have been proposed [43]. As reported in [44], following this approach, the

structure of the model is first manually decided based on the observation of the spatial topology of the data. The model is then iteratively refined in two stages: data segmentation and model training. The recognition phase is typically accomplished using some string matching algorithm like the Knuth-Morris-Pratt [46]. As for the connectionist approaches, in [45] a time-delay neural network (TDNN) for continuous gesture recognition is used. TDNN is a multi-layer feedforward network that uses delays between all layers to represent temporal relationships between events in time. TDNN is learned in order to recognize motion patterns because gesture are spatio-temporal sequences of feature vectors defined along motion trajectories. All the methods described above have advantages and disadvantages: HMMs require the data to be temporally well aligned during the recognition phase, hence the problem of the gesture delimiter arises, TDNNs address the latter problem by exploiting temporal dependencies among the sequences, but the number of the involved parameters is typically high, while FSMs need a manual modeling of the pattern (e.g., a grammar). In addition, the connectionist approaches require a very large training set to train the corresponding gesture models (e.g., using gradient descent algorithm). In contrast with these methods, we focus on a novel method capable of quickly generalize a gesture model starting from a very small training set and perform continuous gesture recognition with a very high accuracy. This method integrates different techniques: clustering algorithm for gesture quantization, Levenshtein distance for gesture prototype election, and Hidden Markov Model for continuous gesture recognition. Ad-hoc Hidden Markov Models are then generated for each gesture exploiting a direct estimation of the parameters. Each model represents the best candidate prototype from the associated gesture training set. The generated models are then employed within a continuous recognition process that provides the probability of each gesture at each step. In particular, the proposed approach is based on the generation of an ad-hoc Hidden Markov Model (HMM) for each gesture exploiting a direct estimation of the parameters. Each model represents the best candidate prototype from the associated gesture

training set. The generated models are then employed within a continuous recognition process that provides the probability of each gesture at each step. This method integrates different techniques: clustering algorithm for gesture quantization, Levenshtein distance for gesture prototype election, and Hidden Markov Model for continuous gesture recognition. In order to assess the proposed system we tested it considering two benchmarks: a hand-performed letters recognizer and a natural gesture recognizer. The collected empirical results show the potential of the approach with respect to other methodologies in literature. Finally, we show the proposed recognition system at work in a typical human-robot interaction scenario.

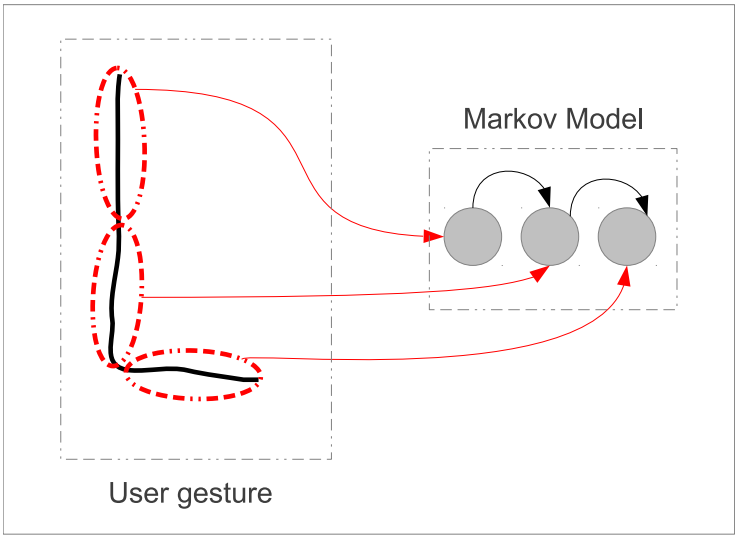


FIGURE 3.1: Building a Markov Model from a user gesture instance

3.3 System Overview

In this section, we detail the gesture recognition process. It consists of two phases: 1) a training phase, where the user shows few samples of a given set of gestures, and 2) a recognition phase, where the system recognizes the gesture performed by

the user. The gesture acquisition process consists of the following steps (see Figure 3.2): Data acquisition (from Kinect device at the sampling period of 100ms); Noise filtering (with a Monte Carlo particle filter estimator); Feature vector extraction; Vector quantization with K-means clustering; Hidden Markov Model (HMM) parameters generation, and HMM evaluation for gestures recognition.

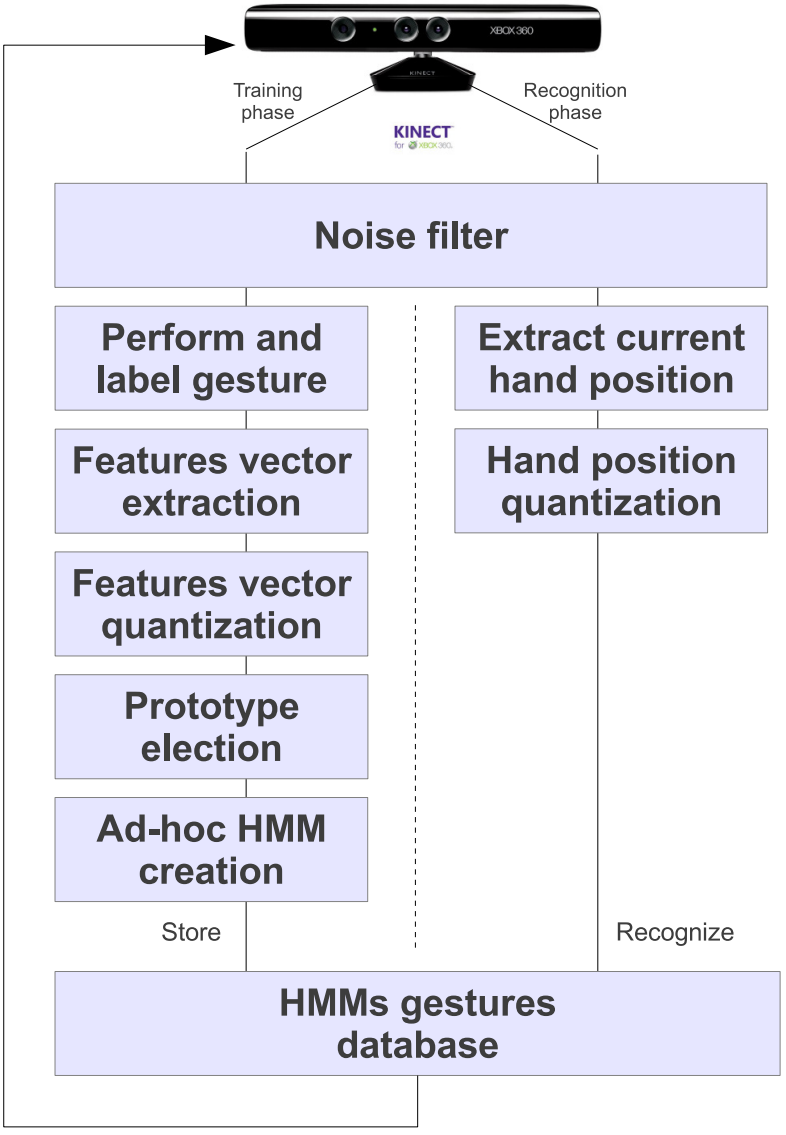


FIGURE 3.2: System architecture.

3.3.1 Gesture definition

We start defining the gesture dataset. Suppose that our gesture vocabulary consists of t gestures classes in the T gestures dataset. Our dataset is $\mathbf{T} = \{T_1, T_2, ..., T_t\}$ where every gesture class set T_j contains n_j instances so that $T_j = \{G_{1_j}, G_{2_j}, ..., G_{n_j}\}$ and n_j denotes the number of repetitions for each gesture of the given class j. Each gesture G_{c_j} is defined as $G_{c_j} = \{(x_{1_{c_j}}, y_{1_{c_j}}), (x_{2_{c_j}}, y_{2_{c_j}}), ..., (x_{m_{c_j}}, y_{m_{c_j}})\}$ where m_{c_j} denotes the number of coordinates belonging to the center of mass of the hand trajectory for the c-th repetition of the gesture belonging to the class j and $(x_{k_{c_j}}, y_{k_{c_j}})$ represents the k-th coordinate for the c-th gesture repetition for the class j.

3.3.2 Joints position estimation

Due to the noise of the perceptive system (Kinect), the hand coordinates need to be smoothed over the time for each gesture. For this purpose (and to make the tracking system robust to occlusions) we deploy an importance sampling algorithm (see Algorithm 1). The state of the hand position at the current time-step k is obtained from the initial state and all the collected measurements $Z^k = \{z_i, i = 1..k\}$ once we solve the Bayesian filtering problem. That is, we need the posterior density $p(\mathbf{x}_k|Z^k)$ of the current state conditioned on all the measurements. As usual, the computation of $p(\mathbf{x}_k|Z^k)$ requires the definition of two phases associated respectively with prediction and update. In the first phase (prediction), we evaluate $p(\mathbf{x}_k|Z^{k-1})$, where the control $\mathbf{u}_k$ vector is defined as $\mathbf{u}_k = [v_{x_k}, v_{y_k}] = [\dot{x_k}, \dot{y_k}] = [\frac{\partial x_k}{\partial k}, \frac{\partial y_k}{\partial k}]$ is the velocity model (speed vector) of the hand in terms of speed among the two axes computed as the numerical derivative of two successive spatial position. In the second phase (update) we use a *measurement model* to incorporate information from the sensor to obtain the posterior probability density function $p(\mathbf{x}_k|Z^k)$ under the assumption of conditional

independence of earlier measurements Z^{k-1} given x_k. The measurement model is given in terms of a likelihood $p(\mathbf{z}_k|\mathbf{x}_k)$ of the hand to be at location $\mathbf{x}_k$ given that $\mathbf{z}_k$ was observed. The posterior density $p(\mathbf{x}_k|Z^k)$ over $\mathbf{x}_k$ is obtained using the Bayes' Theorem.

Algorithm 1 OcclusionsAndNoiseFiltering(χ_{k-1}, $\mathbf{u_k}$, $\mathbf{z_k}$)

1: $\bar{\chi_t} \leftarrow \chi_t \leftarrow \emptyset$
2: **for** $m \leftarrow 1$ **to** M **do**
3: $\quad \mathbf{x}_k^{[m]} \leftarrow \mathcal{N}(\mathbf{x}_{k-1}^{[m]}, \alpha_0\|\mathbf{u}_k\|^2) + \beta\mathbf{u}_k$
4: $\quad w_k^{[m]} \leftarrow \eta_0(\|\mathbf{x}_k^{[m]} - \mathbf{z}_k^{[m]}\|)^{-1}$
5: $\quad \bar{\chi_t} \leftarrow \bar{\chi_k} + \langle \mathbf{x}_k^{[m]}, w_k^{[m]} \rangle$
6: **end for**
7: $m \leftarrow 1$
8: **while** $m < M$ **do**
9: $\quad q \leftarrow \eta_1 w_k^{[m]} M$
10: $\quad$ **for** $j \leftarrow 1$ **to** q **do**
11: $\quad\quad \mathbf{x}_k^{[m]} \leftarrow \mathcal{N}(\mathbf{x}_k^{[m]}, \alpha_1(1 - w_k^{[m]})^2)$
12: $\quad\quad \chi_k \leftarrow \chi_k \cup \{\mathbf{x}_k^{[m]}\}$
13: $\quad\quad m \leftarrow m + 1$
14: $\quad$ **end for**
15: **end while**
16: **return** χ_k

Given $\mathbf{x}_k^{[m]}$ and $w_k^{[m]}$ computed as in the Algorithm 1, we get the approximation of the Equation (3.1) through the expected value of the distribution as reported in Algorithm 3.1. Algorithm 1 implements a Monte Carlo particle filter for position estimation. From the line 2 to 6 new particles with the associated importance weight are generated using the control and measurement vectors $\mathbf{u}_k$ and $\mathbf{z}_k$. The lines 8-14 update the particle using the importance weight previously computed.

$$\hat{\mathbf{x}_k} = E[p(\mathbf{x_k}|z_k, u_k)] \approx \sum_{m=1}^{M} \mathbf{x}_k^{[m]} w_k^{[m]} \tag{3.1}$$

Assuming that $\hat{\mathbf{x}}_{k_{c_j}} = \hat{x}_{k_{c_j}}\, \hat{y}_{k_{c_j}}$ is the corresponding estimation of the k-th hand position coordinate for the c-th gesture repetition for the class j computed by the

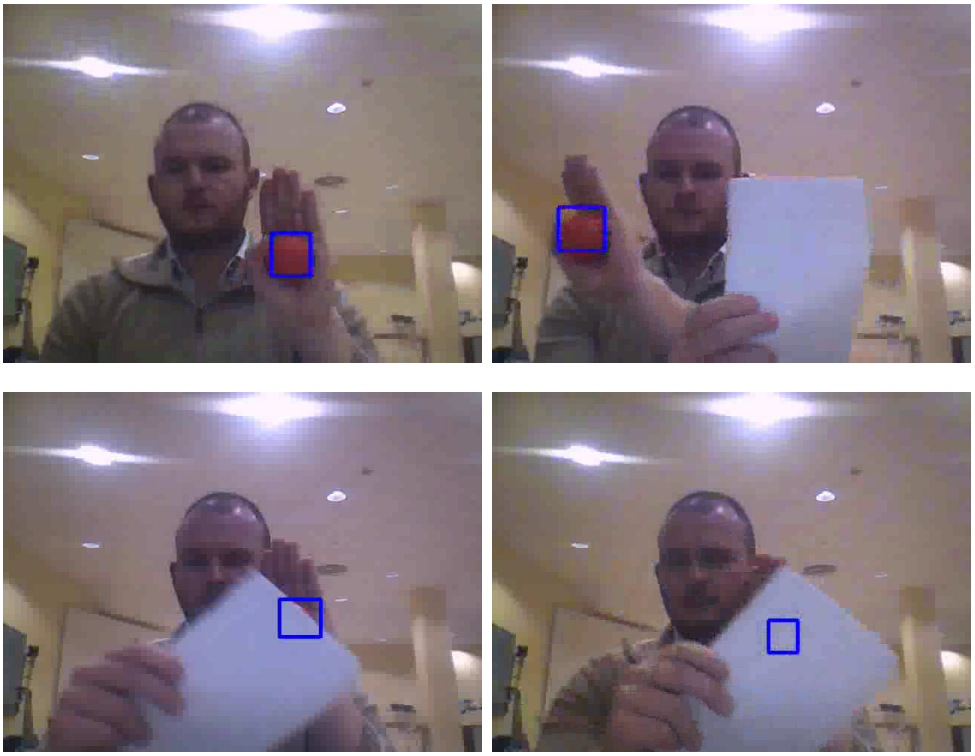


FIGURE 3.3: Example of Monte Carlo method applied to ball tracking. When the red ball goes behind the white paper (occlusion) the Monte Carlo method provides an accurate estimation of the ball position.

Equation (3.1) we get the approximation robust with respect to raft and occlusions (see Figure 3.3).$G_{c_j} = \{(\hat{x}_{1_{c_j}}, \hat{y}_{1_{c_j}}), (\hat{x}_{2_{c_j}}, \hat{y}_{2_{c_j}}), ..., (\hat{x}_{m_{c_j}}, \hat{y}_{m_{c_j}})\}$.

3.3.3 Gesture quantization

For the sake of simplicity, we temporarily replace the notation $G_{c_j} = \{(x_{1_{c_j}}, y_{1_{c_j}}), (x_{2_{c_j}}, y_{2_{c_j}}), ..., (x_{m_{c_j}}, y_{m_{c_j}})\}$ with $G = \{(x_1, y_1), (x_2, y_2), ..., (x_m, y_m)\}$ and assume $\mathbf{x}_i = [x_i\, y_i], i \in [1, m]$. In this way, a gesture instance G is quantized as $Q(G) = \{q_1, q_2, ..., q_m\} : q_i = \arg\min_{k \in [1,K]} \|\mathbf{x}_i - \mathbf{C}_k\|, i \in [1, m]$, where $q_i \in [1, K]$, $C = \{(C_{1_x}, C_{1_y}), (C_{2_x}, C_{2_y}), ..., (C_{K_x}, C_{K_y})\}$ is the set of the K centroids generated with a K-means algorithm over the whole dataset G (argmin defined with respect to the euclidean distance), and $\mathbf{C}_i = [C_{i_x}\, C_{i_y}]$, $i \in [1, K]$.

3.3.4 Generalized mean distance

For each class of the original dataset, we define the distance d of two strings on the alphabet $\mathcal{A} = \{1, 2, ..., K\}$, where d is defined as $d : \mathcal{A}^* \times \mathcal{A}^* \to \mathbb{R}$ [1]. Suppose $A = \{s_1, s_2, ...s_n\} : s_i \in \mathcal{A}^*, i \in [1, n]$ is a set of n strings defined on the alphabet $\mathcal{A}$, the problem of finding the string with minimal distance from all the others is known as the Generalized Mean distance String (GMS). The distance metric we use is the Levenshtein distance metric algorithm reported in (3.2),

$$lev_{a,b}(i,j) = \begin{cases} \max(\text{i,j}) & \text{if } min(i,j) = 0 \\ min \begin{cases} lev_{a,b}(i-1,j)+1 \\ lev_{a,b}(i,j-1)+1 \\ lev_{a,b}(i-1,j-1)+[a_i \neq b_j] \end{cases} & \text{else} \end{cases} \tag{3.2}$$

where, a and b are sequences and i and j are their indexes. Regardless the particular distance chosen, in [47] the authors demonstrated that the problem of finding the GMS is NP-Hard under Levenshtein distance for bounded and even binary alphabets. A more reasonable solution is to find the string, belonging to the set, that minimizes the sum of the distances above all the strings of a given class. This string is known as Set Median String.

$$SMS_j = \arg\min_a \sum_{i=1}^{n_j} lev_{a,Q(G_{i_j})}(|a|, m_{i_j}) : a \in Q(T_j) \tag{3.3}$$

3.3.5 Hidden Markov Model

In this section, we describe how the HMM model of the gestures is generated.

[1] $\mathcal{A}^*$ denotes the sets of strings with zero or more repetitions of the elements belonging to the set A (Kleene operator)

A hidden Markov model (HMM) is a five-tuple (S, Σ, A, B, π), where $\{S\}$ is a set of states including the initial state S_1 and a final state S_F, Σ is the alphabet of the observation symbols, A is the transition probability matrix, $A = \{a_{i,j}\}$, $a_{i,j}$ is the transition probability from state i to state j, B is the output probability matrix, $B = \{b_j(O_k)\}$ (O_k stands for a discrete observation symbol) and π is the starting probability for each state.

Let $\lambda = (A, B, \pi)$ denote the parameters for a given HMM with fixed S and Σ, the key idea of our HMM-based gesture recognition is to use multi-dimensional HMM representing each defined gestures class c as a λ_c HMM model.

Therefore, a gesture is described by a set of N distinct hidden states and r-dimensional K distinct observable symbols.

The number of states of the HMM of the j-th gesture class is chosen to be equal to the SMS_j length as defined in 3.3.

Assuming $SMS_c = Q(G_{c_i}) = \{q_{1_{c_i}}, q_{2_{c_i}}, ..., q_{m_{c_i}}\}$, we have the following HMM model $\lambda_c = (A^c, B^c, \pi^c)$ for the c gesture class:

$$a^c_{i,j} = \begin{cases} p_{trans} & \text{if j=i+1} \\ 1 - p_{trans} & \text{else} \end{cases}, i \in [1, m_{c_i}], j \in [1, m_{c_i}] \tag{3.4}$$

$$B^c = \{b^c_j(o_k)\}, j \in [1, m_{c_i}], o_k \in [1, K] \tag{3.5}$$

$$b^c_j(o_k) = \begin{cases} p_{emit} & \text{if } o_k = q_{j_{c_i}} \\ \frac{1-p_{emit}}{K-1} & \text{else} \end{cases} \tag{3.6}$$

where A^c is the m by m matrix of the transition probabilities, π^c is the starting transition probability, S^c are the model states, p_{trans} and p_{emit} are, respectively, the transition and emit probability of the HMM.

3.3.6 Gesture recognition

Given an observation sequence [2] $Q(G_{obs}) = \{q_{1_{obs_i}}, q_{2_{obs_i}}, ..., q_{m_{obs_i}}\}$ the best class is determined by $C(Q(G_{obs})) = \arg\max_c p(\lambda_c|Q(G_{obs}))$. We compute $P(\lambda_c|Q(G_{obs}))$ by applying the Bayes Theorem. Assuming $P(\lambda_c)$ and $P(Q(G_{obs}))$ constant, we can compute: $P(\lambda_c|Q(G_{obs})) \propto P(Q(G_{obs})|\lambda_c)$. $P(Q(G_{obs})|\lambda_c)$ can be computed through the Forward-Backward algorithm.

3.3.7 Continuous gesture recognition

Continuous gesture recognition is much more complex than isolated gesture recognition, this is due to the difficulty in detecting boundaries among different gestures [48]. Here, we use a temporal sliding method as illustrated in Figure 3.4.

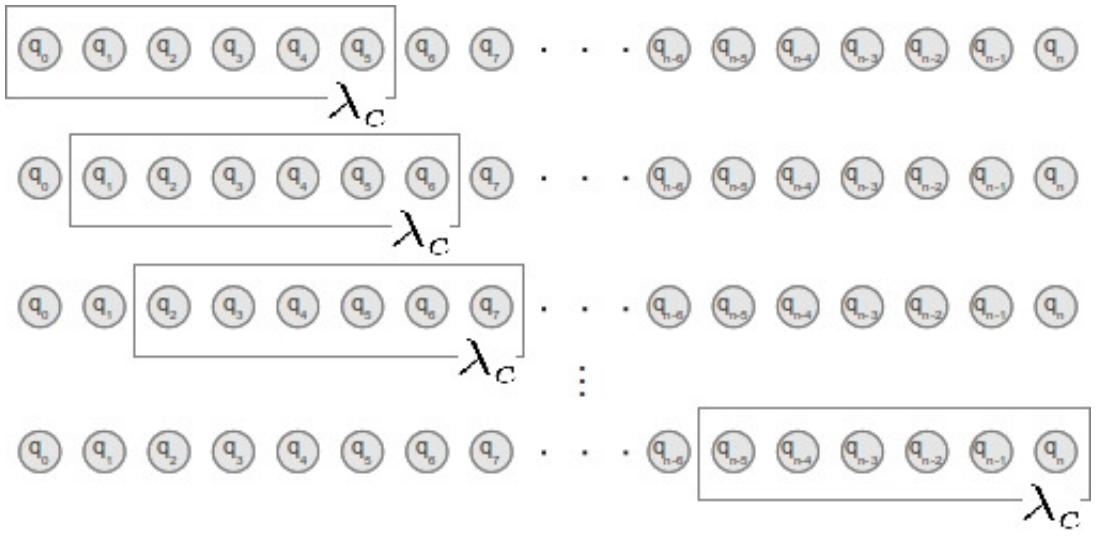


FIGURE 3.4: An example of a temporal sliding for the gesture match.

For each new observation symbol the most likely belonging class is estimated. The algorithm is reported in 2, where for each class c the observation probabilities are computed through the Viterbi algorithm and the best match is returned.

[2] For simplicity we denote sequences as sets with implicit indexing for each element. It will be clear from the context if we are referring to sets or sequences.

Algorithm 2 ContinuousRecognition($obs_{1...n}$)

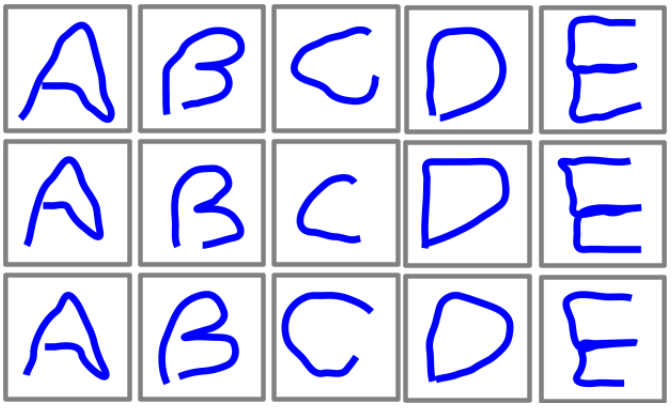
1: **for** $i \leftarrow 1$ **to** C **do**
2: Let $\mathbf{q}$ be the sequence of the last $||S_i||$ observation symbols where S_i is the set of the states of λ_i.
3: $p_i \leftarrow P(\mathbf{q}|\lambda_i)$
4: **end for**
5: **return** $argmax_i p_i$, $i \in [1, C]$

3.4 Case Study

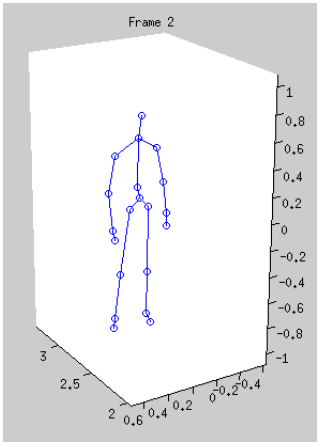
In order to assess the system performance, two standard case studies have been considered: a letter recognizer and a natural gesture recognizer. The former is based on a subset of the English alphabet (A,B,C,D,E) while the latter is based on the Microsoft Research Cambridge-12 (MSRC-12) Gesture Dataset. Furthermore, in a final case study we illustrate the system at work in a HRI scenario.

Letter case study

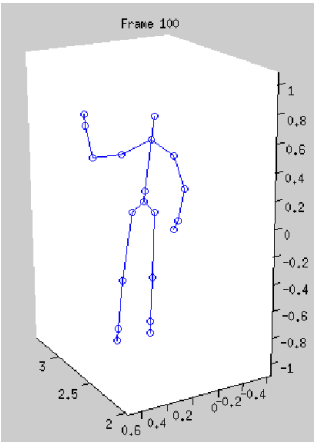
In the letter case study, we aim at validating the intra user variability robustness. We choose the first 5 upper case letter of the alphabet: A,B,C,D and E. The Figure 3.5(a) shows how the user hand trajectory describes the five letters. During the training phase the user performed only 3 samples of the 5 letters. During the recognition process the user is asked to freely move and to perform 20 continuous gestures per each of the 5 letters (for a total of 100 gestures). In this setting we assume that a gesture c is successfully recognized if the likelihood of the c-HMM letter model overcomes a given threshold (set to 65% after empirical testing), it is rejected otherwise. In the training phase the gestures are showed and labeled to the system (supervised training) while in the recognition phase no explicit segmentation is required and the recognition takes place in a continuous gesture stream.



(a) Letters Trajectories



(b) Frame 2



(c) Frame 100

FIGURE 3.5: Examples of letter trajectories performed by the user hand and Frames of human skeleton.

MSRC-12 case study

The MSRC-12 gesture dataset consists of sequences of human skeletal body part movements (represented as body part locations) and the associated tags that should be recognized by the system. The dataset comprises 594 sequences, 719359 frames collected from 30 people performing 12 gestures. In total, there are 6244 gesture instances [49]. The gestures can be categorized into two abstract categories: iconic gestures - those that imbue a correspondence between the gesture and the reference (e.g. G2 - Crouch or hide (duck)), G6 - Shoot a pistol), and metaphoric gestures - those that represent an abstract concept (e.g. G1 - Start

Music/Raise Volume, G3 - Navigate to next menu). The Figure 3.5(b)(c) shows a body skeleton taken at two successive time instants. In the experiment the 20% of the dataset was used as training set and the remaining 80% was used as test set. In particular, for each person performing a gesture the data set contains about 10 repetitions: the first 2 or 3 were used to train the model and the other 7 or 8 were used to test the results. The ground truth data is contained in separate files of the data set package - for each gesture performed there is a time stamp and a label for it. The evaluation was performed comparing the time stamp contained in the ground truth files and the time stamp provided by the proposed classification algorithm. A gesture is considered recognized if the label is the same of the ground truth and if the time stamp difference is not greater than 1 second.

Human-Robot interaction case study

In this case study, we introduce a HRI setting where the task of the robot is to interpret and to execute the intentions of the human using only gestures. We considered the following simple task: a robotic arm is posed in front of a set of objects and is to decide which one to reach (see Figure 5.5); while executing the task the robot continuously monitors the human gestures to understand whether the current operative state is adherent with the human intention or not.

Initially, the robot slowly scans the possible targets moving the end-effector in different directions waiting for some stimulus from the human, who encourages the robot to move towards one of the targets. Once the manipulator starts to move towards one of the targets, depending on the recognized gesture probability, the robot can hesitate or move with confidence in the direction of the selected object. In this context, the robotic arm speed should depend on the confidence of the recognized gesture. When the human interaction starts to become uncertain or something unexpected happens, the robot can decide to stop the motion and switch towards another target. For the task three gestures are considered: *"GO*

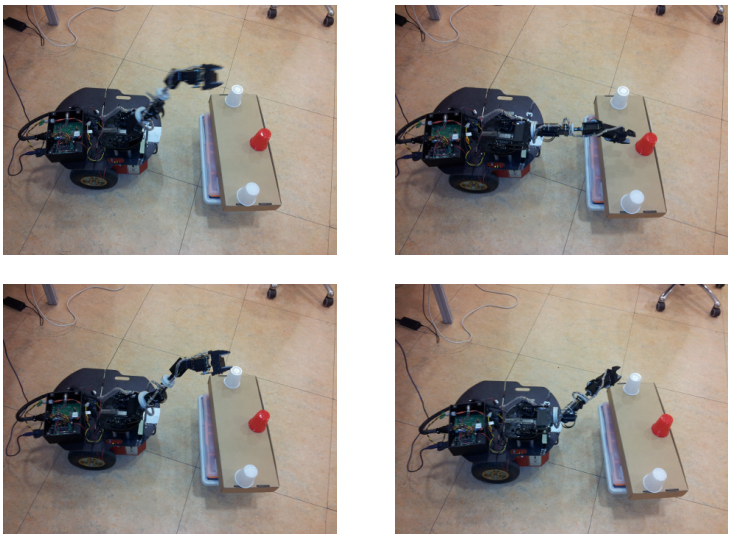


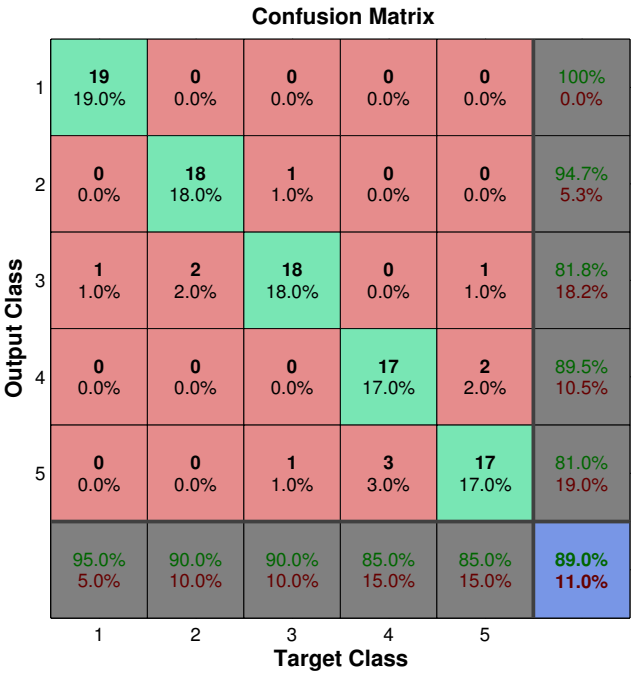
FIGURE 3.6: Human-Robot Interaction case study: interaction task.

ON", *"SLOW DOWN"* and *"SWITCH"*. The *"GO ON"* gesture is intended to suggest the robot to keep going on the current target (rotating the right arm in circle), the *"SLOW DOWN"* gesture makes the robot to decrease the approaching speed (moving right hand up and down) and *"SWITCH"* causes the robot to switch to the next target (moving right hand to left and then to right).

3.4.1 Experimental Results

Letter case study results

In the Figure 3.7(a) the confusion matrix of the letters recognition task is reported. Here the letters A,B,C,D and E are replaced by the index 1 to 5. The high values on the diagonal show that the recognition process is very effective in recognizing all the letters (successful recognition 89%). Moreover, in the Figure 3.7(b), we report the false negatives, false positives, true positives and true negatives rates for each letter. Also in this case, we can observe rare false positives and false negatives, the good performance of the classifier in this case study shows that the system is very effective on intra user variability with a very small training set.



(a) Confusion matrix.

Letter	*FN-rate*	*FP-rate*	*TP-rate*	*TN-rate*
1 *'A'*	0.0123	0.0000	1.0000	0.9877
2 *'B'*	0.0247	0.0526	0.9474	0.9753
3 *'C'*	0.0256	0.1818	0.8182	0.9744
4 *'D'*	0.0370	0.1053	0.8947	0.9630
5 *'E'*	0.0380	0.1905	0.8095	0.9620

(b) Classification results.

FIGURE 3.7: Results for letters recognition case study.

MSRC-12 case study results

As for the second case study, in Figures 3.8(a) and 3.8(b) we report the output log-probability for two gestures sequences, the *G3 - Push right* and *G4 - Googles*, respectively. The vertical (red) lines are the ground truth (e.g., when the gesture is considered performed by the user). The (blue) curve plot is the output log-probability for the given model (e.g., the confidence of a gesture to be recognized). When the probability reaches a peak and goes above a given threshold, the gesture is considered as recognized (see the green circles on top of the peaks). Here, we

TABLE 3.1: Results for the MSRC-12 case study

Gesture	*Accuracy*	*Precision*	*Recall*
G1 lift outst. arms	0.7518	0.9285	0.7506
G2 Duck	0.7800	0.9545	0.7767
G3 Push right	0.8672	0.9759	0.8664
G4 Goggles	0.8015	0.9653	0.7993
G5 Wind it up	0.8534	0.9693	0.8656
G6 Shoot	0.7582	0.9591	0.7627
G7 Bow	0.8250	0.9675	0.8332
G8 Throw	0.8739	0.9705	0.8804
G9 Had enough	0.7937	0.9491	0.7824
G10 Change weapon	0.8273	0.9831	0.8174
G11 Beat both	0.6781	0.9280	0.6398
G12 Kick	0.7893	0.9642	0.8064
Average	0.8000	0.9596	0.7984

can observe that the peek is very close to the ground truth, therefore the gesture can be considered as successfully recognized for each gesture instance.

The Table I reports the results for the MSRC-12 case study in terms of accuracy, precision, and recall.

The precision is obtained as the ratio between TP classification (true positives, i.e. correct results) and the sum of TP and FP (false positives, i.e. unexpected results). On the other way recall is obtained measuring the ratio between TP and the sum of TP and FN (false negatives, i.e. missing results) classification. Accuracy is then computed as the ratio of the sum of TP and TN over the sum of TP, TN (true negatives, i.e. correct absence of results) FP and FN.

Also in this case, the results seem to confirm a good performance of the proposed method in this case study. This result is comparable to other results obtained with standard techniques in literature [50].

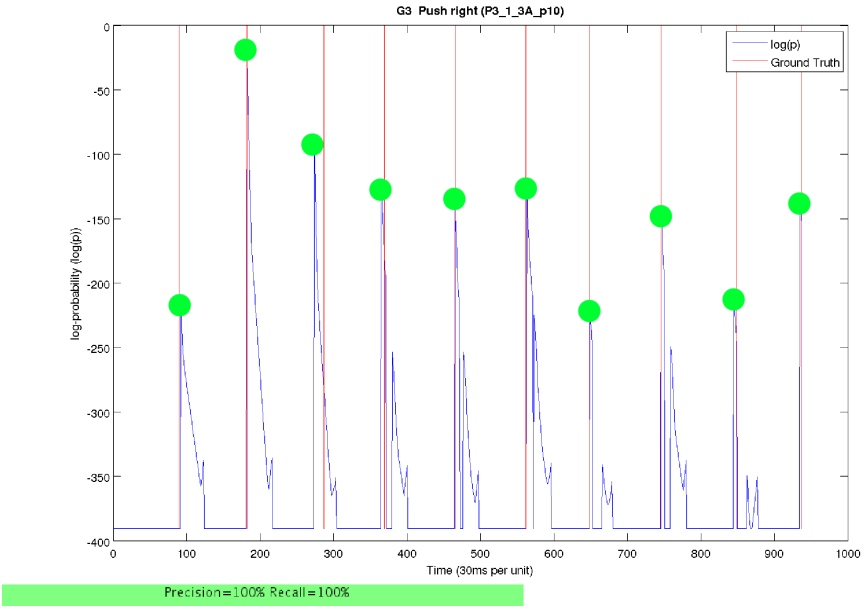
Human-Robot interaction case study results

Finally, we tested the system at work in a human-robot interaction context. Our experimental trials consist of a robotic manipulator cooperating with a human operator in simple tasks. In our setting, the robotic manipulator is close to a small table, on which three differently colored cups are placed; the red cup represents the target. The test we proposed is a game in which the participant has to drive the robotic arm towards the target object (a red cup) as many times as possible in a predefined amount of time (2 minutes). The selected testers have been explained which gestures they could use to interact with the robotic arm. 10 subjects participated in this experiment: 5 students and 5 PhD students, 6 males and 4 females. We evaluated the system considering both quantitative and qualitative performance. The quantitative measures are related to effectiveness and efficiency of the interactive system.

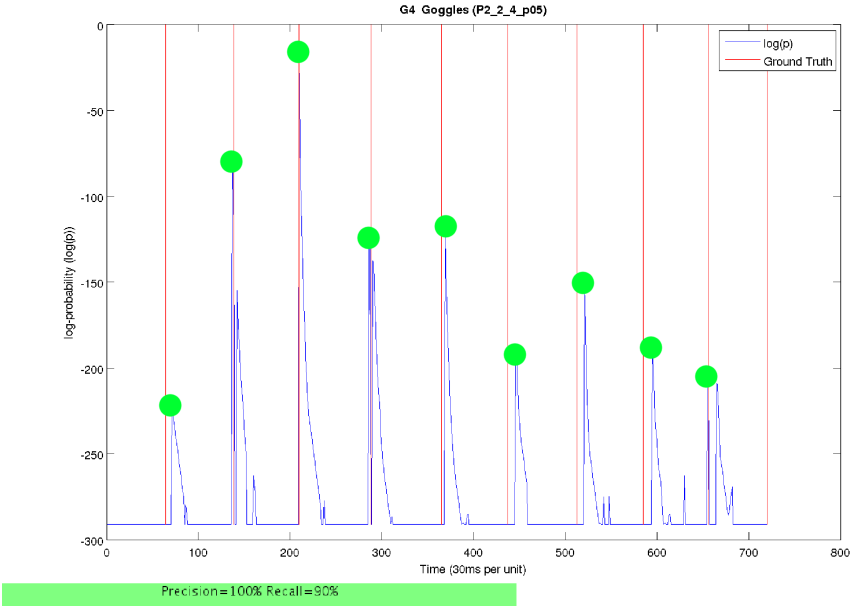
As far as qualitative performance are concerned, our aim was to evaluate the naturalness of the interaction from the operator's point of view. For this purpose, we defined a questionnaire to be filled by the tester after the overall session of tests. The questionnaire is inspired by the HRI questionnaire adopted in [51]. Its aim is to gain information about subjects perception when interacting with the robotic arm. All questions presented may be answered with a grade from 1 to 5.

In particular, we consider two main sections for the users qualitative evaluation 1) a *Specific Information* section, where questions concern respectively *a*) user competences; and the feeling of easy of use; *b*) a *General Feelings* section, asking for naturalness, satisfaction and easy of learning feelings of the interaction [52]. In Figure 3.9, we report both quantitative and qualitative results. As for quantitative results, we measured the number of successes/failures collected by each tester reporting the average values, standard deviation, min, and max. Although the majority of the subjects was not used to interact with robotic systems, they could accomplish the proposed task after few attempts (Figure 3.9(a)). This is also

confirmed by the qualitative results (Figure 3.9(b)-Q1), indeed the users considered both the naturalness (Figure 3.9(b)-Q3) and the ease of use to be satisfactory (Figure 3.9(b)-Q4). Learning how to use the system was also reported to be easy for the subjects (Figure 3.9(b)-Q5). Regarding the ease of use of the system the subjects reported a satisfaction level above the average (Figure 3.9(b)-Q4). This result is consistent with the difference the mean number of successes and the mean number of failures (Figure 3.9(a)). The realized system has been globally judged to be intuitive and satisfactory from the users' point of view.



(a) Output log-probability for a *G3 Push right*.



(b) Output log-probability for a *G4 Goggles*.

FIGURE 3.8: Log-probability for the MSRC-12 case study. On the horizontal axis is reported the time, while on the vertical axis is reported the log-probability of the recognition process. The vertical red lines represent the ground truth (when the gesture is performed by the user) while the blue line shows the log-probability of the recognition at a given time interval. Each green circle represents the local maxima of the log-probability. When the green circle is high it means that the confidence of the gesture recognition is high.

	Mean	STD	Min	Max
Failures	1.60	1.20	0	4
Successes	8.11	0.99	7	10

(a) Quantitative analysis.

	Mean	STD	Min	Max
Q1. (Competence)	1.80	0.87	1	4
Q2. (Ease of use)	3.40	0.66	2	4
Q3. (Naturalness)	3.50	0.92	2	5
Q4. (Satisfaction)	2.90	0.94	1	4
Q5. (Learning)	2.60	0.66	2	4

(b) Qualitative analysis.

FIGURE 3.9: Experimental results for the HRI case study.

Chapter 4

Deictic Gestures Recognition

4.1 Introduction

Non-verbal communication is one of the main component of human-human interaction. Humans use non-verbal cues to accentuate or substitute spoken language. In the context of HRI many approaches focus on speech recognition and well designed dialogues, although the interpretation of non-verbal cues such as body pose, facial expressions, and gestures may either help to disambiguate spoken information or further complement communication [53–55]. An important non-verbal cue in human communication is pointing. Pointing gestures are a common and intuitive way to draw somebody's attention to a certain object (joint attention). While robot gestures can be designed in a way that makes them easily understandable by humans [56], the perception and analysis of human behavior using robot sensors is more challenging. In this chapter, we propose a method for perceiving pointing gestures using a Microsoft Kinect camera. To determine the intended pointing target, frequently the line between a person's eyes and hand is assumed to be the pointing direction. However, since sometimes people tend to perform pointing gestures with line of sight free, this simple geometrical approximation is not enough

accurate. In order to improve the estimation of the pointing gesture, we combine the hand-head line with a regression model by extracting a set of body features from the Kinect skeleton and train a model of pointing directions using Kernel Recursive Least Square Regression. We evaluate the accuracy of the estimated pointing direction in a quantitative study. The results show that our combined model achieves better accuracy than simple geometrical criteria used alone.

Accordingly, the adopted approach for deixis gesture recognition is based on the fusion of two models: a geometrical model and a probabilistic model. In the geometrical model a direct estimation of the target position is computed while in the probabilistic model a machine learning approach is used.

The remainder of this chapter is organized as follows: after a review of related work, we detail our approach for recognizing pointing gestures. Finally, we evaluate the accuracy of the estimated pointing directions.

4.2 Related Work

As also reported in [11] and [70], gesture recognition has been investigated by many research groups belonging to different areas. A recent survey has been compiled by Mitra and Acharya [57]. Most existing approaches are based on video sequences (e.g., [58–60]). These approaches are sensitive to lighting conditions. In contrast, we utilize a Kinect camera which provides depth information of the human operator in the field of view. In [65] is described a task of pointing to object on a table in close range by the use of a multi-layer perceptron classifier to localize hand and finger tips in stereo images and estimate the pointing direction from the finger direction. Loper [66] recognize two gestures for commanding a robot to halt or to enter a room. They use a depth camera similar to the sensor in our approach. The supported gestures do not include any further parameters like a pointing direction that have to be estimated. In [67] neural networks on Gabor

filter responses are trained. Their approach starts from face detection and determines two regions of interest, where they extract filter responses after background subtraction. Sumioka [68] used motion cues to establish joint attention. Luber et al. [59] use a stereo vision system which is actively controlled by a behavior-based gaze controller. They track body features in proximity spaces and determine the pointing direction from the shoulder-hand line. Their system detects pointing gestures by the relative angle between forearm and upper arm. In their experimental setup, two different persons pointed to eight marked positions on the floor. In the approach proposed by Nickel et al. [60], skin color information is combined with stereo-depth for tracking 3D skin color clusters. In order to be independent of lighting conditions, the authors initialize the skin color using pixels of detected faces. The use of multiple modalities to complement and disambiguate individual communication channels has been investigated by by Fransen et al. in [69]. They integrate information from visually perceived pointing gestures, audio, and spoken language for a robot that performs an object retrieval task. The robot has to disambiguate the speaker and the desired object in the conversation. Furthermore, a Gaussian Regression Process approach is used in [70] where the body skeleton lines are used separately for evaluation purpose. In our approach, we use depth from a Kinect camera and learn the correct interpretation of the pointing direction from human observation.

4.3 Pointing Gesture Recognition

The approach used to the perception of pointing gestures is based on Microsoft Kinect device skeleton model. This allows to percieve the 3D direction in which the operator is pointing. We estimate the pointing direction with the fusion of two different models: the first is based on a geometrical solution and the second is based on regression analysis. The determined pointing direction is then fused

with other modalities like eye-gaze and speech through in order to improve the estimation.

4.3.1 Geometrical solution

The simplest approach in gesture deixis recognition consists in the analysis of the geometrical pointing direction provided by the head, shoulder, hand and elbow positions of the user skeleton. In contrast with [70] all the features of the body skeleton are used together and then combined with a machine learning approach.

In order to extract the direction intended by the user the following three directions (lines) are considered (see Figure 4.1).

- Head-Hand line
- Shoulder-Hand line
- Elbow-Hand line

In particular, let $H = (H_x, H_y, H_z)$, $s = (s_x, s_y, s_z)$, $e = (e_x, e_y, e_z)$ and $h = (h_x, h_y, h_z)$ be, respectively, the head, shoulder, elbow and hand position of the human operator in the 3D space (estimated by the Kinect sensor). The three direction cosine vectors of the head-hand ($\mathbf{v}_{Hh}$), shoulder-hand ($\mathbf{v}_{sh}$) and elbow-hand ($\mathbf{v}_{eh}$) lines are then computed as reported in Equations 4.1, 4.2 and 4.3 respectively.

$$\mathbf{v}_{Hh} = \begin{pmatrix} h_x - H_x \\ h_y - H_y \\ h_z - H_z \end{pmatrix} \tag{4.1}$$

$$\mathbf{v}_{sh} = \begin{pmatrix} h_x - s_x \\ h_y - s_y \\ h_z - s_z \end{pmatrix} \tag{4.2}$$

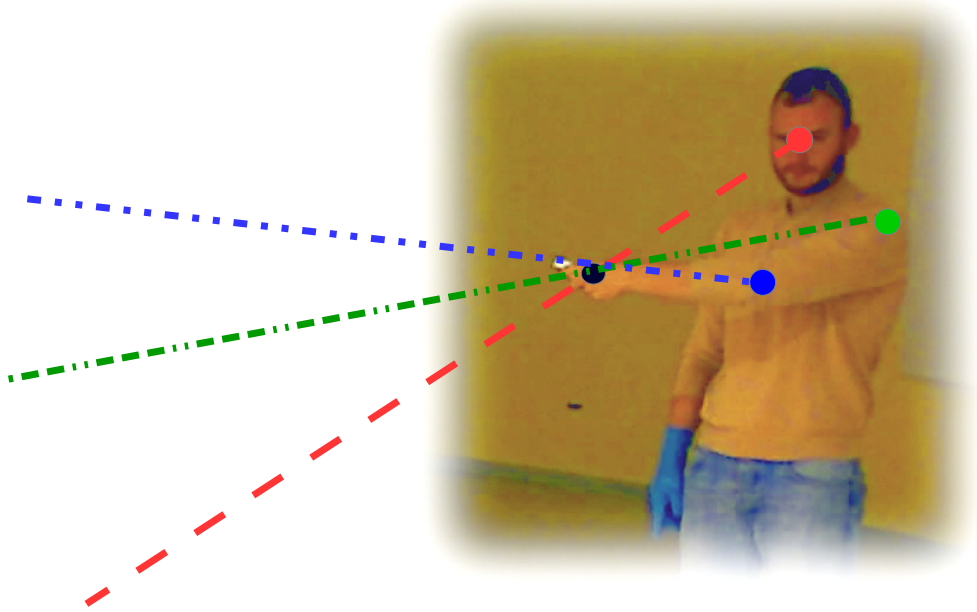


FIGURE 4.1: The three lines of interest in human pointing gestures: head-hand (red, dashed) line, shoulder-hand (green, 2 dots 3 dashes line) line and elbow-hand (blue, 2 dots 1 dash) line.

$$\mathbf{v}_{eh} = \begin{pmatrix} h_x - e_x \\ h_y - e_y \\ h_z - e_z \end{pmatrix} \tag{4.3}$$

Given the the elbow-hand ($\mathbf{v}_{eh}$) and shoulder-hand ($\mathbf{v}_{sh}$) direction cosine vectors, two conditions must be considered:

1. **Case 1**: $\mathbf{v}_{eh}$ and $\mathbf{v}_{sh}$ are different enough to be considered unaligned.
2. **Case 2**: $\mathbf{v}_{eh}$ and $\mathbf{v}_{sh}$ are close enough to be considered aligned.

In the following sections the two cases are detailed.

Case 1

If $\mathbf{v}_{eh}$ and $\mathbf{v}_{sh}$ are not aligned the estimated pointing direction $\mathbf{v}_{dir}$ is computed as the direction cosine of the line passing through the hand h and the center c of the circular base of the cone described by $\mathbf{v}_{Hh}$, $\mathbf{v}_{sh}$ and $\mathbf{v}_{eh}$ (see Figure 4.2).

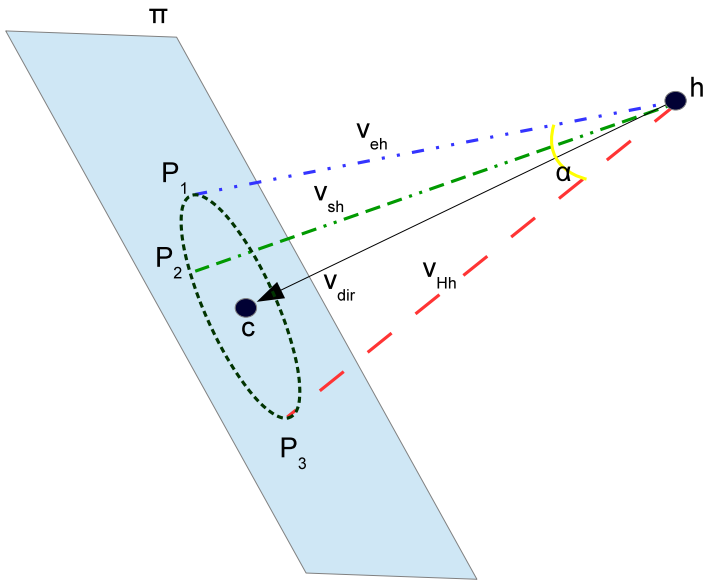


FIGURE 4.2

Let $P_1 = (P_{1_x}, P_{1_y}, P_{1_z})$, $P_2 = (P_{2_x}, P_{2_y}, P_{2_z})$, $P_3 = (P_{3_x}, P_{3_y}, P_{3_z})$ be three points chosen at the same distance from h in the directions $\mathbf{v}_{Hh}$, $\mathbf{v}_{sh}$ and $\mathbf{v}_{eh}$ respectively as Figure 4.2 shows, then the center c is computed with the following steps:

1. Find the plane $\pi : a(x - P_{x_1}) + b(y - P_{y_1}) + c(z - P_{1_z}) = 0$ passing through $P1$, $P2$ and $P3$. Where a, b and c are coefficient computed respectively as:

$$a = \begin{vmatrix} P_{2_y} - P_{1_y} & P_{2_z} - P_{1_z} \\ P_{3_y} - P_{1_y} & P_{3_z} - P_{1_z} \end{vmatrix}$$

$$b = -\begin{vmatrix} P_{2_x} - P_{1_x} & P_{2_z} - P_{1_z} \\ P_{3_x} - P_{1_x} & P_{3_z} - P_{1_z} \end{vmatrix}$$

$$c = \begin{vmatrix} P_{2_x} - P_{1_x} & P_{2_y} - P_{1_y} \\ P_{3_x} - P_{1_x} & P_{3_x} - P_{1_x} \end{vmatrix}$$

2. Find the line perpendicular to the plane π and passing through h:

$$P_\pi = \begin{cases} x = h_x + ta \\ y = h_y + tb \\ z = h_z + tc \end{cases}$$

3. Compute c as the intersection between the line P_π and the plane π as follows:
   $c = (c_x, c_y, c_z) = (h_x + ta, h_y + tb, h_z + tc)$,
   where $t = \frac{-ah_x + aP_{1_x} - bh_y + bP_{1_y} - ch_z + cP_{1_z}}{(a^2 + b^2 + c^2)}$

4. Find the angle α of the cone using the radius r of the cone circle base and the distance between c and h:
   $r = \|P_1 - c\|$
   $d = \|h - c\|$

   $\alpha_0 = \tan^{-1}(\frac{r}{d})$
   $\alpha = 2\alpha_0$

5. Find the pointing direction $\mathbf{v}_{dir}$ as the direction cosine of the line passing through c and h:

$$v = \begin{pmatrix} c_x - h_x \\ c_y - h_y \\ c_z - h_z \end{pmatrix}$$

As a result of the computations described in the above steps, the vector $\mathbf{v}_{dir}$ represents the estimated pointing direction while α is the angle of the "uncertainty"

of the measure (as α increase the size of the pointed area increases, while on the other hand, when α is small the pointed area also is small).

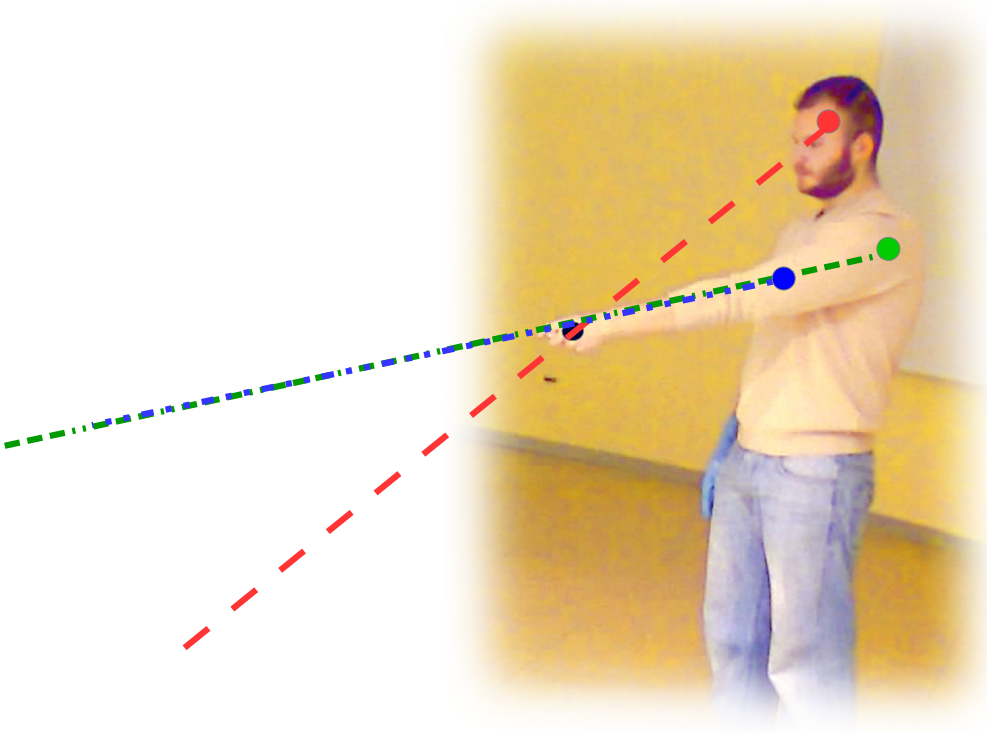


FIGURE 4.3: A pointing gesture with stretched arm.

Case 2

When the arm is completely stretched the elbow-hand line and the shoulder-hand line may overlap, in other words $\mathbf{v}_{eh}$ and $\mathbf{v}_{sh}$ are equal (see Figure 4.3). In that scenario the angle α is simply computed as the inner product of the two direction cosine vectors v_{Hh} and v_{sh}:

$\alpha = \cos^{-1}(\frac{v_{Hh} \cdot v_{sh}}{\|v_{Hh}\| \|v_{sh}\|})$

While $\mathbf{v}_{dir}$ is computed as the average of the two directions v_{Hh} and v_{sh} as follows:

$$\mathbf{v}_{dir} = \begin{pmatrix} \frac{v_{Hhx}+v_{shx}}{2} \\ \frac{v_{Hhy}+v_{shy}}{2} \\ \frac{v_{Hhy}+v_{shy}}{2} \end{pmatrix}$$

4.3.2 Regression Analysis solution

When the angle α is too wide (e.g. over 45 degrees) the geometrical solution is not able to provide an accurate estimation of the pointing direction (the area cone is too wide). In order to keep the accuracy high for such pointing gestures that cannot be handled with the first solution a supervised learning approach is used.

In this section we describe how to learn a model of pointing directions directly from the observation of humans.

4.3.2.1 Skeleton invariant representation

Before applying any supervised learning method to the problem an invariant representation of rigid body, which is invariant to translation, rotation and scaling factors is needed. Without an invariant representation the same skeleton pose could lead to different numerical representations after a change in scale (distance) or rotation. In other words the same skeleton pose, acquired from different camera orientation, could lead to different features vector and then to different estimation.

For that purpose we propose an invariant skeleton representation based on the angles measured on the skeleton.

We know that position of any vector can be uniquely identified by calculating two angles. We apply this to find out the position and orientation of the vectors built from the right arm joint points.

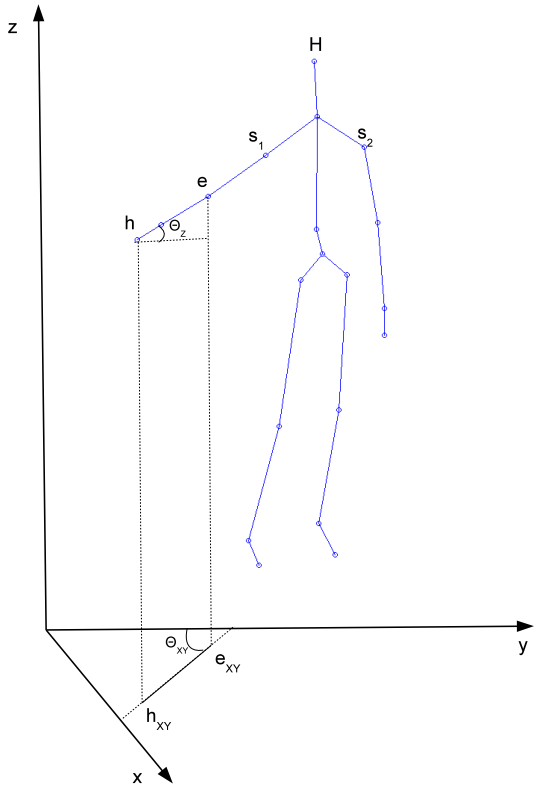
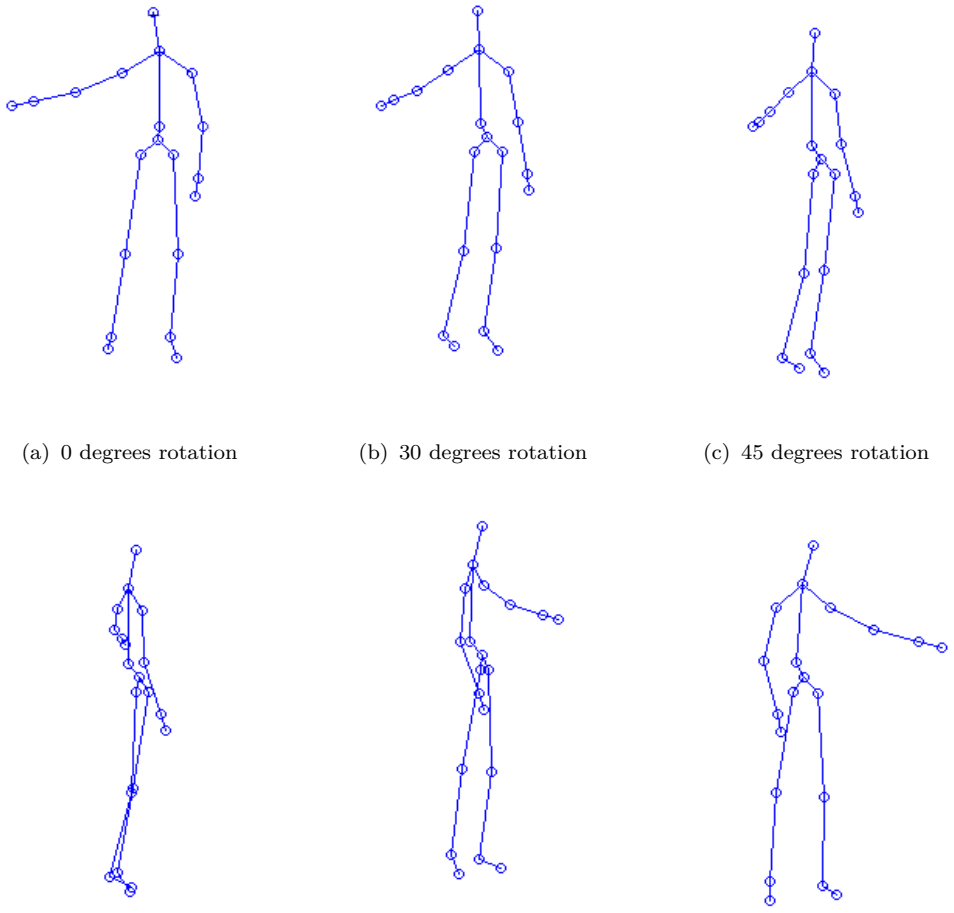


FIGURE 4.4: Scale and rotation invariant representation angles of the Kinect skeleton arm

Once set the y-axis parallel to the line passing through the shoulders, the z-axis parallel to the torso line and the x axis orthogonal to both, to uniquely identify a vector, two angles are measured for the elbow-hand vector ($\mathbf{v}_{eh}$) and the shoulder-elbow vector ($\mathbf{v}_{se}$) (see Figure 4.4):

θ_{eh}^{XY} = the angle between the projection of the vector $\mathbf{v}_{eh}$ on the XY plane.
θ_{eh}^{Z} = the angle between the $\mathbf{v}_{eh}$ vector and the positive z axis.
θ_{se}^{XY} = the angle between the projection of the vector $\mathbf{v}_{se}$ on the XY plane.
θ_{se}^{Z} = the angle between the $\mathbf{v}_{se}$ vector and the positive z axis.

The resulting invariant feature vector is $\mathbf{x} = (\theta_{eh}^{XY}, \theta_{eh}^{Z}, \theta_{se}^{XY}, \theta_{se}^{Z})$. It is trivial to prove that the above feature vector $\mathbf{x}$ is invariant with respect to rotation and scale variations (see Figure 4.5).



(a) 0 degrees rotation (b) 30 degrees rotation (c) 45 degrees rotation

(d) 90 degrees rotation (e) 120 degrees rotation (f) 180 degrees rotation

FIGURE 4.5: Various Kinect skeleton rotations around the vertical axis.

4.3.2.2 Kernel Recursive Least Square Regression

In this section we propose to learn a model of pointing directions directly from the observation of humans for such configurations where α is too wide.

We apply the Kernel Recursive Least Square algorithm (KRLS) [71] to train a function approximator that maps extracted arm features $\mathbf{x}$ to a pointing direction $\hat{v}_{dir}$.

As reported in [71], kernel methods are a relatively new class of learning algorithms utilizing Mercer kernels in order to produce nonlinear versions of conventional

linear supervised and unsupervised learning algorithms (for recent reviews, see [72] and [73]). Support vector machines (SVMs), which use kernel methods as a core ingredient, are state-of-the-art in many classification and regression tasks today [72]. The basic idea behind kernel methods is that a Mercer kernel function (see definition below), which is applied to pairs of input vectors, can be interpreted as an inner product in a high-dimensional Hilbert space (often called the feature space), thus allowing inner products in the feature space to be computed without making direct reference to feature vectors. This idea, which is commonly known as the "kernel trick," has been used extensively in recent years, most notably in classification and regression e.g. [72]-[74]. Standard approaches to the prediction problem usually assume a simple parametric form. In the standard least squares approach, one then attempts to find the value of that minimizes the squared error. Given a new sample, the number of computations performed by RLS to derive a new minimum least-squares estimate of is independent of the number of the samples. This is an essential requirement from an online algorithm - the amount of computations required per new sample must not increase (and preferably be small) as the number of samples increases.

Assuming access to a recorded sequence of input and output samples:

$Z^t = \{(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), ..., (\mathbf{x}_t, \mathbf{y}_t)\}$
where $\mathbf{x}_i$ is the i-th feature vector containing the angles extracted as shown in the previous section and $\mathbf{y}_i$ is the known corresponding pointing direction for the i-th feature vector, the resulting estimator has the form reported in Equation 4.4

$$\hat{f}(\mathbf{x}) = \sum_{i=1}^{t} \alpha_i k(\mathbf{x}_i, \mathbf{x}) \tag{4.4}$$

where α_i is the regularization parameter for the i-th training sample and k is the learned function.

The key advantage of the KRLS method is that each new training sample can be learned online, this is very important in Human-Robot interaction tasks.

4.4 Experimental Results

We evaluated the accuracy of the deixis gesture recognition system in an indoor scenario. We asked 8 people (6 male, 2 female) to perform 20 separate, natural pointing gestures to 5 fixed targets. The targets have been placed in the scene at different and fixed positions on the floor. The experiment consisted in two phases: the learning phase and the evaluation phase. In the learning phase the first half (50%) of the 800 pointing gestures performed by the users, along with the target positions, have been used as training set for the regression model. In the evaluation phase the second half of the gesture corpus have been used as test set.

For every pointing gesture, we computed the euclidean distance between the intersection of the estimated pointing line with the floor and the target position 4.6. The table 4.1 reports the mean μ_p and the standard deviation δ_p (in meters) of the distance from the desired target on the floor (error) using the geometrical solution only, the regression solution only and the combined approach.

	μ_p	δ_p
Geometrical solution only	0.34	0.21
Regression solution only	0.75	0.46
Combined approach	0.22	0.32

TABLE 4.1: Experimental results for deixis estimation: quantitative analisys

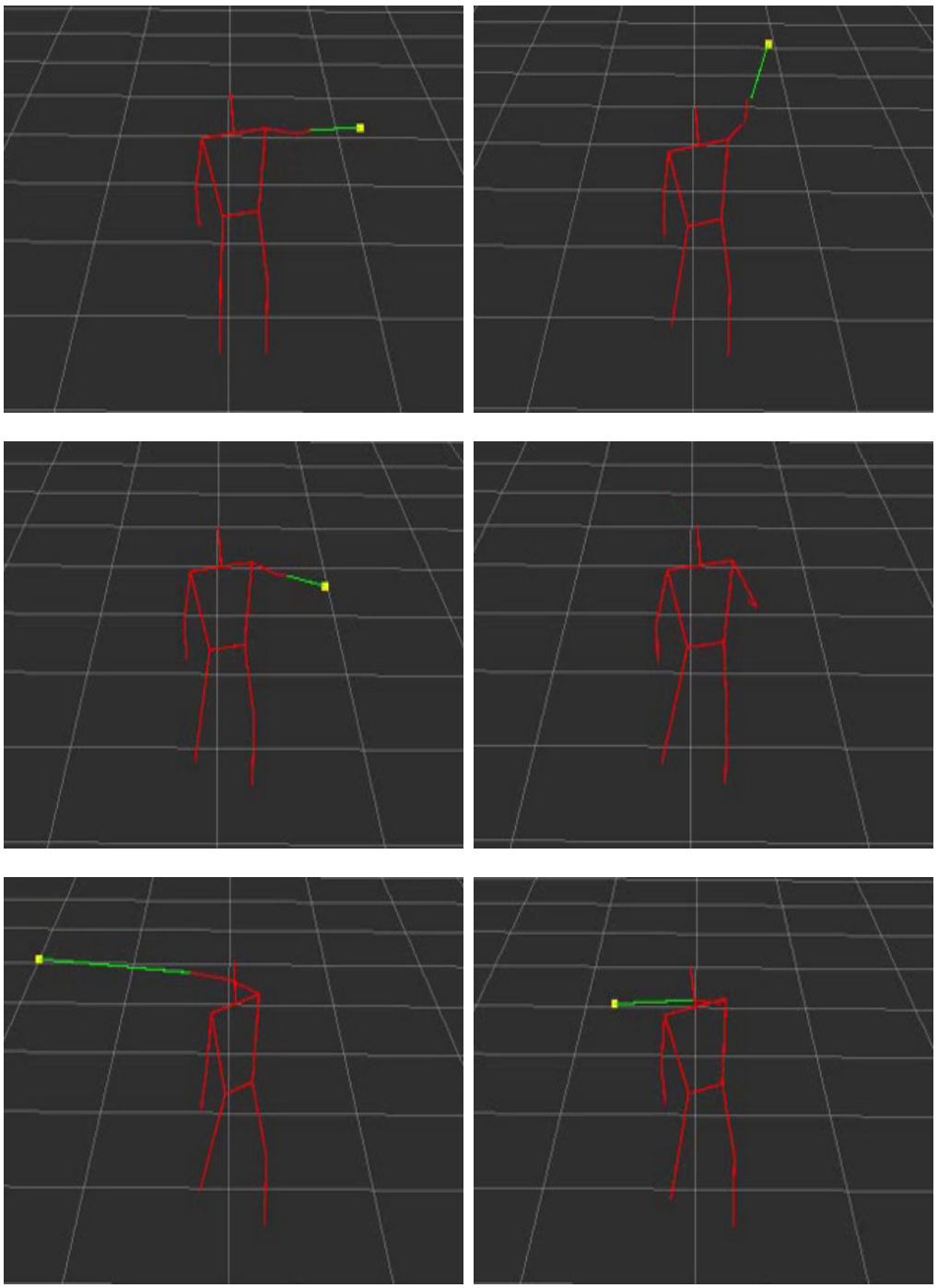


FIGURE 4.6: Deixis gesture demo snapshots

Chapter 5

Robot Attentional Regulation

5.1 Introduction

In this chapter, we explore the interplay between attentional and emotional regulation in human-robot interaction. More specifically, we aim at defining an architecture where attention allocation and emotional processes can influence the robotic interactive behavior adapting it to the human emotions, intentions, and expectations. In biological systems, attentional and emotional processes are strictly integrated and play a central role in cognitive control [75]. These processes are involved in initiation, selection, regulation, switching, and coordination of behaviors. Moreover, emotion and attention have a fundamental role in social interaction. The importance of these processes is also recognized by the social robotics and human-robot interaction literature. Emotional processes and affective computing methods are usually deployed to improve the empathic attitude of the human with respect to the robot as well as the effectiveness and naturalness of the interaction [76, 77]. On the other hand, attentional control and joint attentional mechanisms have been proposed [78, 79] to allow implicit communication and coordination during the execution of interactive tasks. Less effort has been provided in defining

a cognitive control system where affective and attentional mechanisms are tightly integrated (e.g. [76, 80]). In this work, we investigate this issue by proposing a simple human-robot interaction system endowed with a supervisory attentional system and regulated by the estimation of the human affective state. The robotic attentional system is modeled as a behavior-based system where each behavior is endowed with a simple attention allocation mechanism [81, 82] regulating sensors' sampling rates and actions' activations. The human estimated emotional state is represented in a four dimensional map [83], but we consider only the arousal and predictability values which are strictly related with attentional control. Indeed, high arousal is usually associated with cognitive effort and attentional processes [75], while high unpredictability and surprise are usually related to attentional shifts [84].

We tested our system in a minimal setting considering a simple robotic manipulator whose behavior and attentional state are influenced by the emotional content extracted from the voice of the human operator. To estimate human emotions, we rely on speech recognition techniques (which are considered as robust methods [85] when compared with humans' success rate at identifying emotions). In this settings, the prosodic features extracted from the human voice produce attentional bursts or inhibitions which can be used by the operator to guide the execution of a simple manipulation task. We assessed the system performance considering both qualitative and quantitative analysis. The collected results, show that emotional interaction combined with the attentional modulation provides an effective and natural human-robot interactive behavior.

5.2 Background and Models

In this section, we present a background on emotions and attentive systems [75] along with our proposal to connect vocal stimuli to attentive bursts by the use of

arousal and predictability.

5.2.1 Multi-dimensional model for emotions

In the literature, several attempts to define a dimensional space to map emotions can be found. This is because a dimensional, continuous model for emotions is nowadays considered more flexible and powerful, from a descriptive point of view, with respect to discrete models. Usually, dimensional models consider three axes: valence, potency, and arousal, with valence and arousal being the most investigated ones in technological applications. In [83], however, a four-dimensional model of emotions was derived from a cross-cultural study. Data presented in [83] highlighted that a better separation of emotional words in a multidimensional space could be obtained by introducing an unpredictability axis together with the three *classic* ones.

In the present chapter, we include the four dimensional model of emotions in our architecture. We believe it represents a good start point to build an interactive robotic architecture taking into account emotions. In next sections, we will concentrate on the exploration of human-robot interaction possibilities using a very low level of communication based on prosodic features extracted from human voice. While our present intent is to correlate prosodic features to the emotional axes to control a robot's behavior, it should be noted that, being emotions multimodal in nature, vocal features are differently correlated with these axes. Features coming from other channels, like facial expressions and gestures, account for emotional information that cannot be easily detected from voice only. Valence, for example, is known to be hard to evaluate from vocal features while these have been reported to be strongly correlated with dominance and arousal. Our intention, in this case, is to investigate the lowest levels of interaction in order to evaluate the impact prosodic features alone have in a task oriented scenario. At the same time, we want to retain the possibility of easily extending the framework to gradually

introduce, in the future, other levels, both parallel and higher, to evaluate the role each one covers in natural human-robot interaction. This is why it is important to integrate the most generic emotional model available, to our knowledge, in the architecture at this stage.

5.2.2 Frequency-based model for attention allocation

Our attentional system is obtained as a reactive behavior-based system where each behavior is endowed with an attentional mechanism represented by an internal adaptive clock [82, 86]. In Figure 5.1 we show a schema theory representation

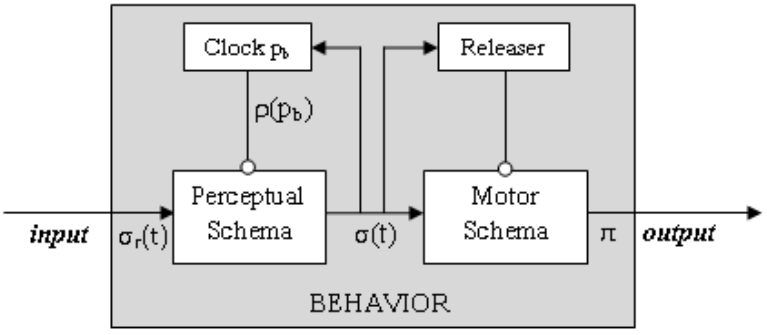


FIGURE 5.1: Schema theory representation of an attentional behavior.

of an attentional behavior. This is characterized by a Perceptual Schema (PS), which elaborates sensor data, a Motor Schema (MS), producing the pattern of motor actions, and an attentive control mechanism, called Adaptive Innate Releasing Mechanism (AIRM), based on a combination of a clock and a releaser. The releasing mechanism works as a trigger for the MS activation, while the clock regulates sensors' sampling rate and behaviors' activations. The clock regulation mechanism is our frequency-based attentional mechanism: it regulates the resolution at which a behavior is monitored and controlled, moreover, it provides a simple prioritization criteria. This attentional mechanism is characterized by:

- A period p ranging in an interval $[p_b min, p_b max]$,

- An *updating function* $f_{a,d}(\sigma(t), p_b^{t-1}) : R^n \rightarrow R$ adjusting the current clock period p_b^t, according to both internal states and environmental changes.
- A trigger function $\rho(t, p_b^t)$, which enables/disables the data flow $\sigma_r(t)$ from sensors to PS at each p^t time unit.
- Finally, a normalization function $\phi(f_{a,d}(\sigma(t), p_b^{t-1})) : R \rightarrow N$ that maps the values returned by $f_{a,d}(x)$ into the allowed range $[p_b min, p_b max]$.

The clock period at time t is regulated as follows:

$$p_b^t = \rho(t, p_b^{t-1}) \times \phi(f_{a,d}(\sigma(t), p_b^{t-1}) + (1 - \rho(t, p_b^{t-1})) \times p_b^{t-1} \tag{5.1}$$

That is, if the behavior is disabled, the clock period remains unchanged, i.e. p_b^{t-1}. Otherwise, when the trigger function is 1, the behavior is activated and, the clock period changes according to the $\phi(x)$.

5.2.3 Vocal signal, Arousal and Predictability

In an interactive framework, it has been noted that a subject's arousal level is efficiently conveyed by the portrayed vocal intensity. This is because, being influenced by subglottal pressure and vocal fold adduction, the arousal is directly linked with the tension of the phonatory apparatus [87]. In this work, we will concentrate on estimating the user's arousal level from speech energy in order to control an attentional robotic system. While the energy of the speech signal is a relatively simple feature to extract from human voice, a certain degree of refinement is necessary in order to obtain real-time reactions to vocal stimuli. In general, the main problem in the extraction of prosodic features from speech lies in the overlapping of two different components of the message: the semantic and the intonational ones. During speech activity, these two intertwined levels of communication are transmitted at the same time on the same channel. While it is

only natural for the human mind to separate first and then recombine the two in order to make sense out of the sentence with all the nuances coming with it, the task is much more difficult to describe in a technological setup. The energy profile is severely influenced by the segmental level of an utterance: energy fluctuations are, in fact, strictly related to the occurrence of basic speech units, like syllables [88].

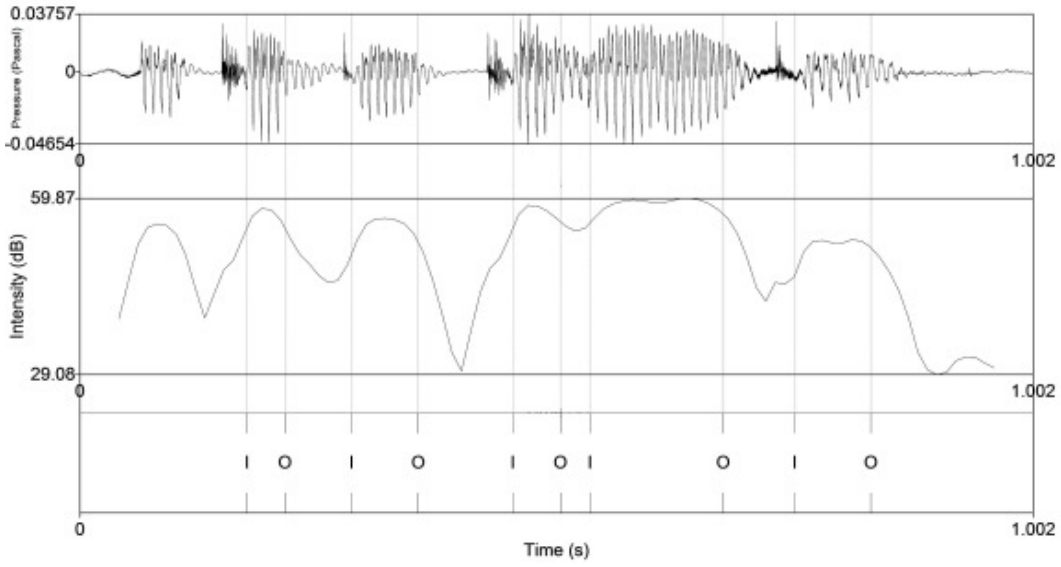


FIGURE 5.2: The waveform of a speech signal along with its energy profile. On the third tier automatically detected syllable nuclei incipits (I) and offsets (O) are reported.

Specific classes of phones, like fricatives and occlusives, cause the energy profile to drop before rising again in coincidence with the occurrence of the next vowel, usually representing the syllabic nucleus. Therefore, the control system we present here is based on the extraction of voiced local maxima to produce attentional bursts or inhibitions and to guide the execution of the task. In the system used for our tests, the sampling rate of the microphone was set at 8000 Hz and the analysis window length was set to 40 ms. First, we apply a bandpass filter ($75 - 4000$ Hz), then the energy of the frame is computed as follows:

$$E_t = 10 \times log10(\sum_{f=1}^{max_f} a_f), \tag{5.2}$$

where max_f is the maximum frequency taken from the Fourier transform of the frame and a_f is the amplitude of the $f-th$ frequency in the spectrum. The voiced/unvoiced decision is performed by considering both the autocorrelation function of the frame and the zero-crossing rate yielding pitch values between 75 Hz and 300 Hz.

If a voiced energy peak is detected, given the energy level normalized in the interval $[0, 1]$, the final effect on the arousal axis is mapped in the interval [-0.5, 0.5] as follows:

$$I_t = (-sin((3 \times x \times \pi)/2))/2 \tag{5.3}$$

The final arousal value is computed as the numerical integral of the energy profile E as described by the sequence of peaks occurred in the analysis window W, which was set to 1 second. By using this method, in absence of energy peaks, the arousal level tends to go back to zero.

$$A_t = \sum_{i=t-W}^{t} \eta E_i \tag{5.4}$$

The effect of this mapping is that low energy peaks inhibit arousal, mid-level stimuli do not alter it, as the user's intentions are less clear, while high energy peaks cause an excitation.

Other than the raw energy value, we take into account the frequency of occurrence of the energy peaks F, with reference to the analysis window, and the entropy of the signal energy H is computed as follows

$$H_t(\Delta f_i^{(j)}) = -\sum_{k=1}^{|Y^{(j)}|} p(y_k^{(j)}) ln(p(y_k^{(j)})) \tag{5.5}$$

Energy peaks frequency and signal entropy are related to the predictability axis of the emotional model. The final instantanoeus predictability value was computed as follows

$$P_t = \beta(1 - H_t) + (1 - \beta)F_t \tag{5.6}$$

With this formula, the predictability value is linked with a inversely proportional law to the H parameter, while it is related with F by means of a direct proportionality law.

5.3 Case Study

Following [89], we assume attention and arousal as multi-dimensional psychological processes closely interacting with one another and we use this assumption as the basis for designing the architecture of our interactive system.

The idea is to develop a control system endowed with attentional mechanisms influenced by emotions and suitable for human-robot interaction and communication. In this work, we are concerned with an interactive setting where the task of the robot is to interpret and execute the intentions of the human using only the emotional feedback. As a case study, we consider the following simple task: a robotic arm is posed in front of a set of objects and is to decide which one to reach; while executing the task the robot continuously monitors the human emotional state to understand whether the current operative state is adherent with the human intention or not. Initially, the robot slowly scans the possible targets moving the end-effector in different directions waiting for some stimulus from the

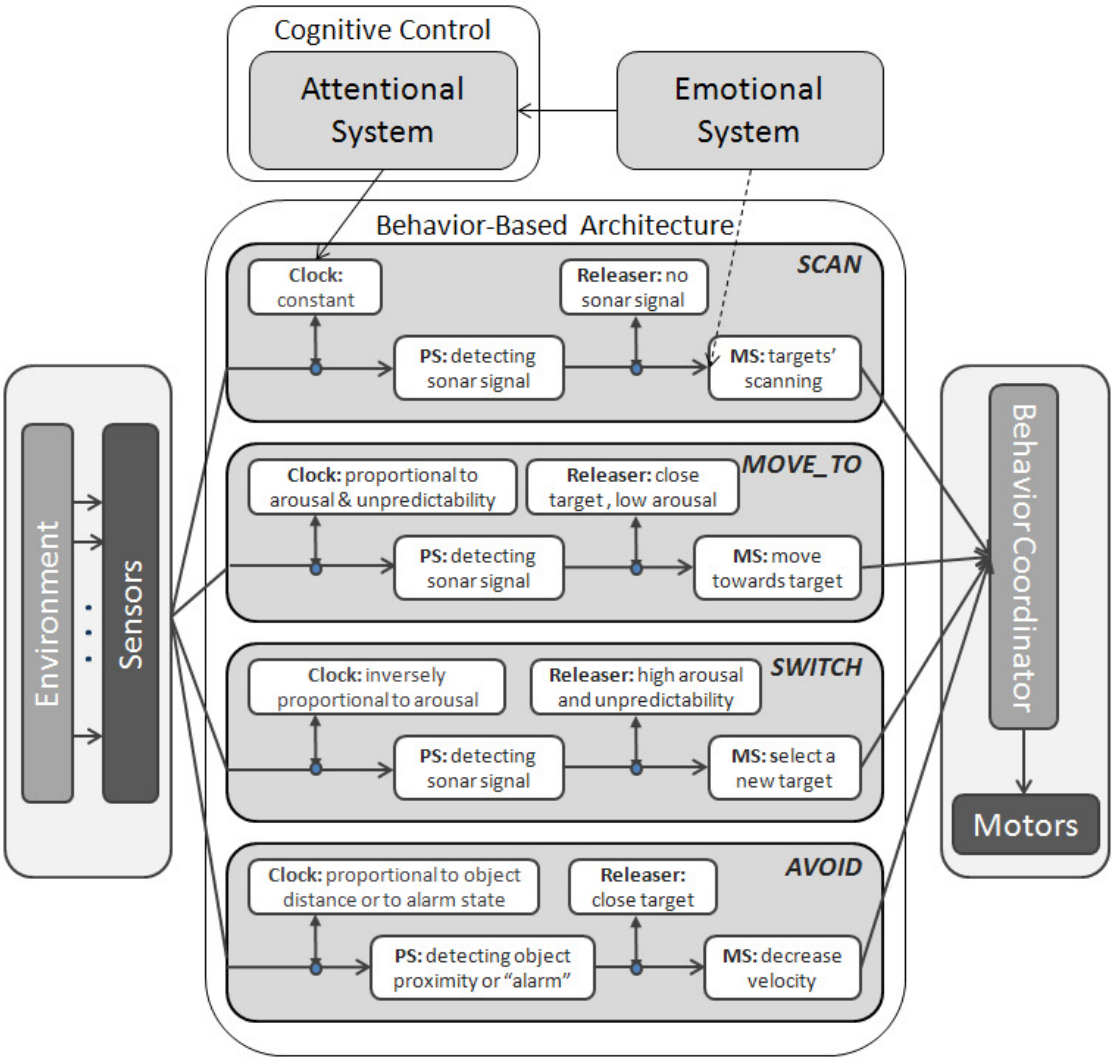


FIGURE 5.3: Attentional and Emotional Behavior-based Architecture.

human that encourages the robot to move towards one of the target. Once the manipulator starts to move towards one of the target, depending on the emotional content extracted from the human speech, the robot can hesitate or move with confidence in the direction of the selected object. When the human interaction starts to become uncertain or something unexpected happens, the robot can decide to stop the motion and switch towards another target. In this context, the robotic attentional and emotional states should depend on the emotional state estimated from the human voice: if the human arousal is high, the robotic behavior should be attentive; when the human predictability is high, this means that the robot

is doing well, hence the robotic valence is positive and the confidence is high; otherwise, when predictability is low, then the uncertainty in the human voice is interpreted as an unclear signal that affects the robotic confidence provoking hesitation and/or task switching.

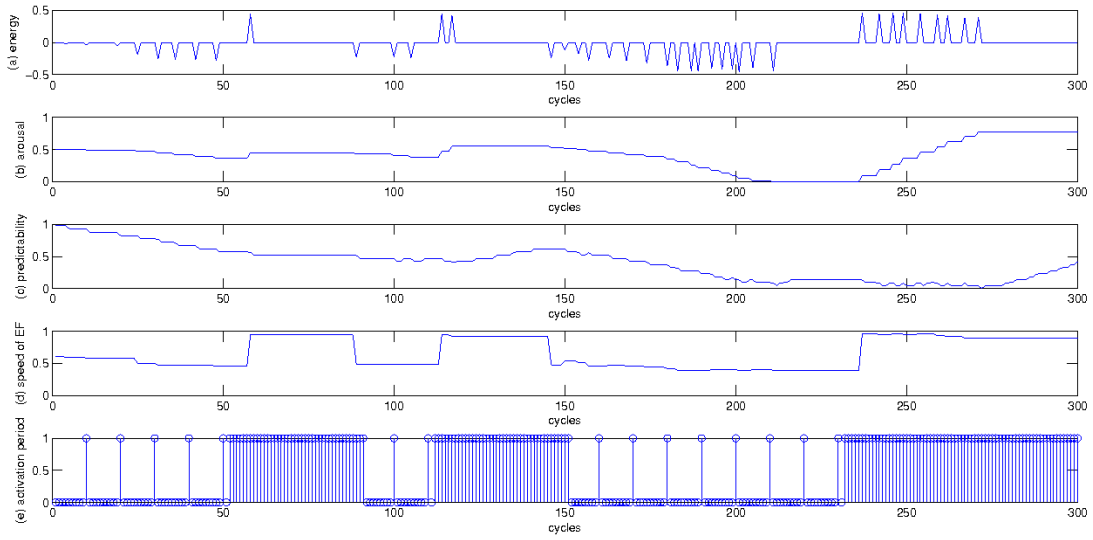


FIGURE 5.4: Speed and Behavior Activation trend as a function of Arousal and Predictability levels over time. (a) Vocal Energy, (b) Arousal Level, (c) Predictability Level, (d) end-effector speed and (e) SWITCH Behavior Activations.

In the following we illustrate how the attentional and emotional behavior is implemented as an attentional behavior-based architecture. We introduce four behaviors: (a) *SCAN*, (b) *MOVE_TO*, (c) *SWITCH*, and (d) *AVOID*, representing, respectively, (a) the robot default behavior while it is waiting for stimuli, (b) the robot motion towards a selected target, (c) the selection of an alternative target, (d) the obstacle avoidance behavior. Each behavior is endowed with an attentional mechanism influenced by the emotional state. The attentional mechanisms is represented by an adaptive clock whose frequency is regulated by the interaction with the human (emotional content of the human speech) and the environment (obstacles). The robot behaviors are influenced by the arousal and predictability levels as follows:

SCAN is the default behavior of the robots in the absence of stimuli. It continuously moves the robot end effector along a circular trajectory scanning all the possible targets keeping a slow and constant speed. Here, the clock period is not adaptive and regulated to a constant frequency, i.e. $p_s^t = k_s$.

MOVE_TO moves the end effector towards a target position (i.e. one of the target objects on the table). In this case, the adaptive clock period p_m^t depends on a linear combination of the the arousal and predictability extracted from the speech signal, i.e. $\sigma(t) = \gamma_1^m A_t + \gamma_2^m P_t$. Here, the idea is that both high arousal and high unpredictability values are associated with high attention, hence high frequency for the clock period, however, the main contribution to the period update comes from the arousal. Given the composed arousal and predictability $\sigma(t)$ value extracted from the human voice, the clock period is updated in a proportionally manner, i.e.:

$$f^m(\sigma(t), p_m^{t-1}) = \gamma_3^m \times \sigma(t).$$

The velocity is associated with the estimated arousal and predictability, but in a different manner, i.e. it is proportional to the predictability when the arousal is low, i.e. $v^m(t) = \gamma_4^m \times P_t$ for $A_t < T_1^m$, otherwise, when both predictability and arousal are low, it is associated with both, i.e. $v^m(t) = \gamma_5^m \times \sigma(t)$ for $A_t < T_2^m$ and $P_t < T_3^m$. In all the other cases, the velocity remains unchanged.

AVOID prevents the robot collision with the objects and manages the halting behavior due to abrupt changes. Indeed, it slows down the robot motions with the target proximity implementing a version of the Fitts' law, but it can also react to alarms detected in the user speech. In our setting, the robotic system detects an alarm when the arousal is high and predictability is low, given suitable thresholds, i.e. $alarm(t) = 1$ if $f(A_t, P_t) > T_1^a$ and $alarm(t) = 0$ otherwise. When the alarm is not generated, the clock period is proportional to the distance of the obstacle, i.e. $f_a^a(dist(t), p_a^{t-1}) = \gamma_1^a \times dist(t)$, otherwise, $f_a^a(dist(t), p_a^{t-1}) = p_a^{min}$. As for velocity,

it slows down proportionally the the period, i.e. $v_t^a = \gamma_2^a \times p_a^t$, and stops when the alarm is raised.

SWITCH allows the robot to change the current target depending on the emotional interaction. It is activated when both arousal and unpredictability are high. In this behavior, the clock period p_s^t is directly associated with the arousal value, i.e. the higher the arousal the smaller the period

$$p_s^t = \frac{p_{max} - \gamma_1^s \times A_t}{p_{max}}.$$

As for the velocity, it is associated with the unpredictability, that is, the higher the unpredictability the lower the velocity:

$$v_s^t = \gamma_2^s \times \frac{P_max - P_t}{P_t}.$$

Figure 5.4 illustrates the effect of the users' vocal energy (a) on arousal (b), predictability (c), end-effector speed (d) and activations of the SWITCH behavior (e) with respect to machine cycles (mc) (the duration of each cycle is around 60 ms). As we can see the arousal level trend 5.4-(b) is directly related to energy one 5.4-(a): as a matter of fact the former is an integrated version of the latter. Figure 5.4-(c) shows the predictability estimation over time accordingly to the variation of the arousal profile in the last 100 mc. As expected the predictability increases in the interval $[0, \ldots, 100]$ due to the periodicity of the energy profile in such range while it decreases around 150 mc caused by the change in the energy trend (no more positive peaks). In Figure 5.4-(d) is shown how the change in the arousal signal affects the speed of the robotic arm over time. Specifically, as arousal decreases, speed increases (i.e. the robot feels more self-confident) and vice versa. In this way the robotic arm immediately reacts to the human speech stimulus. We can observe some delays (about 500 ms) due to the system software framework.

In 5.4-(e) the clock activation signal of the SWITCH behavior is illustrated. In the first interval $[0, \ldots, 50]$ the period is larger because arousal is low and, accordingly to the underlying model, the probability of switching to another target is low too (i.e. the robot is self-confident). Conversely as the arousal increases the period decreases (the probability of a switch increases).

5.4 Experimental Results

Our experimental trials consist of a robotic manipulator cooperating with a human operator in simple tasks.

5.4.1 Platform

We adopt a 7DOF robotic arm (Cyton Arm: payload 300g, hight 60 cm, reach 48 cm, joint speed 60 rpm), endowed with a gripper (3.25 cm) as end-effector. The robot is controlled by the Player/Stage tool [90].

5.4.2 Environment

In our setting, the robotic manipulator is close to a small table, on which three differently colored cups are placed; the red cup represents the target. The test we propose is a game in which the participant has to drive the robotic arm towards the target object as many times as possible in a predefined amount of time (2 minutes). The human operator must, therefore, try to adapt the trajectory of the robotic arm in order to keep it close to the red object. The selected testers have not been explained how to interact with the robotic arm. They only knew that semantics would not have affected the robots behavior, while the tone of their voice would.

(a)  (b) 

FIGURE 5.5: (a) A snapshot of the human-robot interactive environment. (b) A snapshot of the robot field of view.

5.4.3 Experiment Trials

10 subjects participated in this experiment: 6 students and 4 PhD students, 7 males and 3 females. The experiment took an average of 10 minutes per subject, and answering a specific HRI questionnaire took an average of 1 minute per subject.

5.4.4 Results Evaluation

We evaluated the system considering both quantitative (Table 5.1) and qualitative performance (Table 5.2). The quantitative measures are related to effectiveness and efficiency of the interactive system.

As far as qualitative performance are concerned, our aim was to evaluate the naturalness of the interaction from the operator's point of view. For this purpose, we defined a questionnaire (see Table 5.2) to be filled by the tester after the overall session of tests. The questionnaire is inspired by the HRI questionnaire adopted

	Mean	STD	Max	Min
Failures	1.4	1.17	0	3
Successes	6.3	2.21	3	10
Move Act.	26%	9%	41%	12%
Switch Act.	1%	1%	2%	0%
Avoid Act.	5%	1%	8%	2%
Scan Act.	67%	10%	50%	84%

TABLE 5.1: Experimental results for the attentional regulation interaction task: quantitative analysis

in [51]. Its aim is to gain information about subjects perception when interacting with the robotic arm. All questions presented may be answered with a grade from 1 to 5.

Section	Question
Personal Information	Age? Gender? Q1. How familiarized are you with robotic applications?
Specific Information	Q2. How easy was it to perform the task? Q3. Did the robot react accordingly with your expectations?
General Feelings	Q4. How natural is this kind of interaction? Q5. How satisfying do you find the interactive system? Q6. How easy was to learn how to control the robot?

TABLE 5.2: HRI questionnaire for attentional regulation performance evaluation

Although the majority of the subjects was not used to interact with robotic systems (Table 5.3-Q1) the users considered both the naturalness (Table 5.3-Q4) and the ease of use to be satisfactory (Table 5.3-Q2). Learning to use the system was also reported to be easy for the subjects (Table 5.3-Q6). This result is validated by the hits frequency increase as a function of time (Figure 5.6). Namely, in order to evaluate the learning curve for the use of the proposed system, we discretized

the experimental time in four subintervals of 30 seconds each. Then, we evaluated the cumulative function of the number of successes obtained by each subject and considered this to be and indirect indicator of the learning curve: the higher the hits frequency, the higher the learning rate. The reported histogram shows that, after approximately 1.5 minutes of interaction, all the subjects were able to clearly increase their successes. This indicates that 1 minute and a half is the mean time needed for a user to learn how to control the robot.

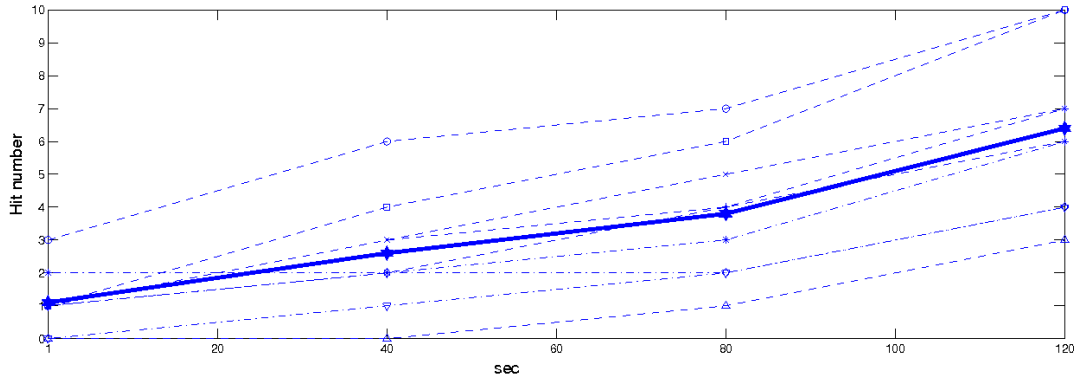


FIGURE 5.6: Cumulative histogram of the number of hits as a function of time. Dashed lines state for single subjects while the solid line shows the general trend.

Regarding the ease of use of the system the subjects reported a satisfaction level above the average (Table 5.3-Q5). This result is consistent with the difference the mean number of hits and the mean number of failures (first and second raw in Table 5.1). The realized system has been globally judged to be intuitive and satisfactory from the users' point of view. The only value under the expected average was related to the coherence of the robot behavior with respect to intentions. However this result may have been influenced by the interactors' inclination to employ semantics.

	Mean	STD	Min	Max
Q1. (Competence)	2.40	1.17	1	4
Q2. (Ease of use)	3.40	0.84	2	5
Q3. (Coherence)	2.70	1.05	1	4
Q4. (Naturalness)	3.30	1.15	1	5
Q5. (Satisfaction)	3.40	0.69	2	4
Q6. (Learning)	3.40	1.17	2	5

TABLE 5.3: Experimental results for attentional regulation interaction task: qualitative analysis

Conclusions

In this thesis, we presented a novel method for continuous gesture recognition that should support a natural and flexible human-robot social interaction. In addition we presented a human-robot interaction system based on attentional regulation modulated by low level emotional speech features.

The proposed approach for gesture recognition presents several advantages both during the training and the recognition phases. During the first phase, the available training set can be very small, hence the user can perform a very limited number of gesture samples to introduce new gestures. As for the second phase, a continuous recognition process enables the system to keep multiple hypothesis about the gesture switching from one interpretation to another depending on the context. This flexible and light process is possible because the bayesian classifier used in this work requires minimum computational resources and thus multiple gestures can be tracked in real-time. This continuous gesture recognition process allows us to face and resolve the ambiguities according to the interactive context in that enhancing the overall recognition system robustness and naturalness. The effectiveness of the recognizer has been tested in two standard case studies, while the flexibility and naturalness of the interaction has been discussed considering a simple HRI task. As a future work in the gesture recognition problem, we plan to investigate the system performance in a more sophisticated social interaction scenario considering full body gesture recognition.

Conversely in the approach for attentional regulation our aim was to evaluate the amount of information that can be extracted from speech without introducing high level analysis, like semantic interpretation. We showed that the expressiveness degree coming from basic characteristics of the human voice can be exploited to obtain a sufficiently natural interactive system. As a proof of concept, we described the approach in a minimal setting, considering the case of a simple robotic manipulator which interacts with a human considering only the energetic content of the vocal signal and mapping it on arousal and predictability axes. We evaluated our interactive system by using both qualitative and quantitative analysis. The collected results show that although the interactors' inclination to employ semantics appears to have affected their judgement, on the other hand the interaction naturalness was judged to be more than sufficient. Moreover, although the majority of the subjects reported not to be accustomed to interact with robotic systems, they reported to have been able to easily complete the task and learn how to use the system. Objective data support this subjective impression as the mean number of hits is clearly higher than the mean number of failures. The realized system clearly shows the advantage of using very simple speech characteristics, producing a reactive response and a low computational load, while providing at the same time a good level of expressiveness, at least enough to make the system intuitive and easy to use also for inexperienced users. Concluding, the presented results will be used as a baseline in future works in which we will focus on quantifying the minimum amount of information and computational effort required as the complexity degree of the tasks increases. The current architecture, including the 4-dimensional model of emotions, is designed to easily accommodate further sources of information and interpret them in this generic model. As we increase the complexity of the tasks, axes other than arousal and predictability will be included in the system's evaluations. During this process, our goal will be to investigate the boundaries at which a purely reactive system needs to be integrated by high-level content analysis and, lastly, by appraisal processes. As

future works in the attentional regulation we plan to extend the experimental setting in order to evaluate the architecture by considering other kind of actions (i.e. grasping and manipulation).

Bibliography

[1] David J. Feil-Seifer and Maja J. Matarić. Human-robot interaction. *Invited contribution to Encyclopedia of Complexity and Systems Science, Springer New York*, 2009.

[2] Albert Merhabian. *Nonverbal Communication*. Aldine, 2007.

[3] Kendra Cherry. *The Everything Psychology Book: Explore the human psyche and understand why we do the things we do*. Adams Media, 2010.

[4] A. Kendon. Some reasons for studying gesture. *Semiotica*, 1986.

[5] A. Kendon. Nonverbal communication. *Encyclopedic Dictionary of Semiotics*, 1986.

[6] C. Cadoz. Les realites virtuelles. dominos, flammarion. 1994.

[7] B. Rime and Schiaratura L. Gesture and speech. in fundamentals of nonverbal behavior. *Press Syndicate of the University of Cambridge, New York*, 1991.

[8] S. C. Levinson. Pragmatics. *Cambridge: Cambridge University Press.*, 1983.

[9] David A. Leavens. Manual deixis in apes and humans. *University of Sussex*.

[10] Butterworth G. Franco F. Pointing and social awareness: Declaring and requesting in the second year. *Journal of Child Language*, 1996.

[11] Thies Pfeiffer. Understanding multimodal deixis with gaze and gesture in conversational interfaces. *A.I. Group, Faculty of Technology,Bielefeld University*, 2010.

[12] H. P Grice. Logic and conversation. *Syntax and Semantics: Speech Acts, vol. 3. New York: Academic Press*, 1975.

[13] Bill Buxton. *Haptic Input, Gesture based Interaction*. 2011.

[14] G.L. Martin. The utility of speech input in user-computing interfaces. *Intl. J. Man-Machine Studies*, 1989.

[15] Cohen P. McGee D. Oviatt S. Pittman J. Smith I. Johnston, M. Unification-based multimodal integration. 1997.

[16] Ackerman M.S. Schmandt, C. and D. Hindus. Augmenting a window system with speech input. *IEEE Computer*, 1990.

[17] T. Ichikawa S. Chang and P. Ligomenides. Using visual concepts. *Visual Languages, New York: Plenum Press,*, 1986.

[18] IEEE Maja J Mataric Senior Member IEEE Adriana Tapus, Member and IEEE Brian Scassellati, Senior Member. The grand challenges in socially assistive robotics.

[19] R. A. Brooks. Intelligence without reason. *In Proceedings of 12th International Joint Conference on Artificial Intelligence (IJCAI-91)*, 1991.

[20] R. A. Brooks. A robust layered control system for a mobile robot. ieee transactions on robotics and automation. *IEEE Transactions on Robotics and Automation*, 1986.

[21] David J. Feil-Seifer. Socially assistive robot-based intervention for children with autism spectrum disorder. *Ph.D. Dissertation*, 2008.

[22] J. Rosenblatt. Ithe distributed architecture for mobile navigation. *Journal of Experimental and The-oretical Artificial Intelligence*, 1997.

[23] P. Maes and R. A. Brooks. Learning to coordinate behaviors. *In Proceedings, 8th National Conferenceon Artificial Intelligence (AAAI-90)*, 1990.

[24] F. Michaud and J. Audet. Using motives and artificial emotions for long-term activity of an autonomous robot. *In Proceedings of the Fifth International Conference on Autonomous Agents*, 2001.

[25] R. Arkin and T. Balch. Aura: Principles and practice in review. *Journal of Experimental and Theoretical Artificial Intelligence*, 1997.

[26] J. Connell. Sss: A hybrid architecture applied to robot navigation. *In Proceedings, IEEE International Conference on Robotics and Automation (ICRA-92)*, 1992.

[27] E. Gat. On three-layer architectures. *In R. P. B. D. Kortenkamp and R. Murphy, editors, Artificial Intelligence and Mobile Robotics*, 1998.

[28] P. Agre and D. Chapman. What are plans for. *Robotics and Autonomous Systems,*, 1990.

[29] T. Ishida M. Fujita, Y. Kuroki and T. Doi. Autonomous behaviour control architecture of entertainment humanoid robot sdr-4x. *In Proceedings of the 2003 IEEE/RSJ Intl. Conference on Intelligent Robots and Systems*, 2003.

[30] W. Burgard A. Cremers F. Dellaert D. Fox D. Hahnel C. Rosenberg N. Roy J. Schulte S. Thrun, M. Bennewitz and D. Schulz. Minerva: A second-generation museum tour-guide robot. *In Proceedings: IEEE International Conference on Robotics and Automation (ICRA '99)*, 1999.

[31] R. Sutton. Dyna, an integrated architecture for learning, planning, and reacting. *SIGART Bulletin*, 1991.

[32] C. Thorpe T. Fong and C. Baur. Collaboration, dialogue, and human-robot interaction. *In 10th International Symposium on Robotics Research (ISRR)*, 2002.

[33] N. Mitsunaga and Asada. Visual attention control for a legged mobile robot based on information criterion. *In Proc. of IROS-2002*, 2002.

[34] A.; Orlandini A.; Carbone, A.; Finzi and Pirri. Model-based control architecture for attentive robots in rescue scenarios.

[35] D. Norman and T. Shallice. Attention in action: willed and automatic control of behaviour. *Consciousness and Self-regulation: advances in research and theory*, 1986.

[36] R. Cooper and T. Shallice. Contention scheduling and the control of routine activities. *Cognitive Neuropsychology*, 2000.

[37] D. Kahneman. Attention and effort. *Englewood Cliffs, NJ: Prentice-Hall.*, 1973.

[38] E. Burattini and S. Rossi. A robotic architec- ture with innate releasing mechanism. *In 2nd International Symposium on Brain, Vision and Artificial Intelligence*, 2007.

[39] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proc. IEEE*, 77:257—-285, 1989.

[40] Starner and A. Pentland. Real-time american sign language recognition from video using hidden markov models. *Rep. TR-375*, 1995.

[41] F. Samaria and S. Young. Hmm-based architecture for face identification. *Image Vis. Comput.*, 12:537–543, 1994.

[42] J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time-sequential images using hmm. *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recogn.*, pages 379—-385, 1992.

[43] J. Davis and M. Shah. Visual gesture recognition. *Vis., Image Signal Process.*, 141:101–106, 1994.

[44] A. F. Bobick and A. D. Wilson. A state-based approach to the representation and recognition of gesture. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(12): 1235–1337, 1997.

[45] M. S. Yang and N. Ahuja. Recognizing hand gesture using motion trajectories. *Proc. IEEE CS Conf. Comput. Vis. Pattern Recogn.*, 1:466—-472, 1998.

[46] P. Hong, M. Turk, and T. S. Huang. Gesture modeling and recognition using finite state machines. pages 410–415, 2000.

[47] François Nicolas and Eric Rivals. Hardness results for the center and median string problems under the weighted and unweighted edit distances. *Discrete Algorithms*, 3(2–4):390 – 415, 2005.

[48] Tie Yang and Yangsheng Xu. Hidden markov model for gesture recognition. *CMU-RI-TR-94*, 10.

[49] Simon Fothergill, Helena Mentis, Pushmeet Kohli, and Sebastian Nowozin. Instructing people for training gestural interactive systems. *ACM Conference on Human Factors in Computing Systems*, 2012.

[50] M. Hussein, M. Torki, M. A. Gowayyed, and M. El-Saban. Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations. *IJCAI'13 Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 2466–2472, 2013.

[51] Mihai Duguleana, Florin Grigorie Barbuceanu, and Gheorghe Mogan. Evaluating human-robot interaction during a manipulation experiment conducted in immersive virtual reality. In *Proc. of International Conference on Virtual and Mixed Reality: new trends - Part I*, pages 164–173. Springer-Verlag, 2011.

[52] S. Iengo, A. Origlia, M. Staffa, and A. Finzi. Attentional and emotional regulation in human-robot interaction. *In proceedings of the 21st IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN2012*, 2012.

[53] E. Martinson S. Blisard M. Marge S. Thomas A. Schultz B. Fransen, V. Morariu and D. Perzanowski. Using vision, acoustics, and natural language for disambiguation. *In Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2007.

[54] S. H. Chernova C. V. Jones M. M. Loper, N. P. Koenig and O. C. Jenkins. Mobile human-robot teaming with environmental tolerance. *In Proceedings of the 4th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2009.

[55] E. Huber and D. Kortenkamp. A behavior-based approach to active stereo vision for mobile robots. *Engineering Applications of Artificial Intelligence*, 1998.

[56] C. Eppner A. Gorog F. Faber, M. Bennewitz. The humanoid museum tour guide robotinho. *In Proceedings of the 9th IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, 2009.

[57] Sushmita Mitra and Tinku Acharya. Gesture recognition: A survey. *IEEE Transactions on Systems, Man, and cybernetics - part C: Applications and Reviews, Vol. 37, No 3., May 2007.*

[58] L. Kopp and P. Gardenfors. Attention as a minimal criterion of intentionality in robots. *volume 89 of Cognitive Studies. Lund University.*

[59] M. Luber K. Arras, S. Grzonka and W. Burgard. Efficient people tracking in laser range data using a multi-hypothesis leg-tracker with adaptive occlusion probabilities. *In Proc. ICRA*, 2008.

[60] S. Behnke T. Axenbeck, M. Bennewitz and W. Burgard. Recognizing complex, parameterized gestures from monocular image sequences. *Proc. of IEEE Humanoids*, 2008.

[61] F. Kaplan and V. Hafner. The challenges of joint attention. *Interaction Studies*, 2006.

[62] F. Fleuret J. Berclaz and P. Fua. Robust people tracking with global trajectory optimization. *In Proc. of CVPR*, 2006.

[63] D. Fox D. Schulz, W. Burgard and A.B. Cremers. People tracking with a mobile robot using sample-based joint probabilistic data association filters. *International Journal of Robotics Research*, 2003.

[64] D.B. Reid. An algorithm for tracking multiple targets. *IEEE Transactions on Automatic Control*, 1979.

[65] N. de Freitas J.J. Little K. Okuma, A. Taleghani and D.G. Lowe. A boosted particle filter: Multitarget detection and tracking. *IIn Springer, LNCS 3021*, 2004.

[66] O. Lanz. Approximate bayesian multibody tracking. *IEEE Trans.*

[67] D.M. Gavrila and S. Munder. Multi-cue pedestrian detection and tracking from a moving vehicle. *Int. Journal of Computer Vision*, 2007.

[68] T. Martinetz M. Haker, M. Bohme and E. Barth. Self-organizing maps for pose estimation with a time-of-flight camera. *In Proc. of the DAGM 2009 Workshop on Dynamic 3D Imaging*, 2009.

[69] V. V. Hafner and F. Kaplan. Learning to interpret pointing gestures: experiments with four-legged autonomous robots. *volume 3575 of LNCS. Springer*, 2005.

[70] D. Droeschel and J. Stuckler Sven Behnke. Learning to interpret pointing gestures with a time-of-flight camera. *In Proceedings of 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2011.

[71] Member IEEE Yaakov Engel, Shie Mannor and Ron Meir. The kernel recursive least-squares algorithm. *IEEE Transactions on Signal Processing, Vol. 52, NO. 8*, 2004.

[72] B. Schoolkopf and A. Smolao. Learning with kernels. *MIT Press, Cambridge, MA*, 2002.

[73] R. Herbrich. Learning kernel classifiers. *MIT Press, Cambridge, MA*, 2002.

[74] N. Cristianini and J. Shawe-Taylor. An introduction to support vector machines. *Cambridge, U.K.: Cambridge Univ. Press*, 2000.

[75] D. Kahneman. *Attention and Effort*. Englewood Cliffs, NJ: Prentice-Hall, 1973.

[76] Cynthia Breazeal. *Designing sociable robots*. MIT Press, Cambridge, MA, USA, 2002. ISBN 0-262-02510-8.

[77] Ronald C. Arkin, Masahiro Fujita, Tsuyoshi Takagi, and Rika Hasegawa. An ethological and emotional basis for human-robot interaction. In *Robotics and Autonomous Systems*, pages 191–201, 2003.

[78] J. Gregory Trafton, Nicholas L. Cassimatis, Magdalena D. Bugajska, Derek P. Brock, Farilee E. Mintz, and Alan C. Schultz. Enabling effective human-robot interaction using perspective-taking in robots. *IEEE Transactions on Systems, Man, and Cybernetics*, 35:460–470, 2005.

[79] C. Breazeal and L. Aryananda. Recognizing affective intent in robot directed speech. *Autonomous Robots*, 12(1):83–104, 2002.

[80] Alexander Stoytchev and Ronald C. Arkin. Incorporating motivation in a hybrid robot architecture. *JACIII*, 8(3):269–274, 2004.

[81] J. Senders. The human operator as a monitor and controller of multidegree of freedom systems. pages 2–6, 1964.

[82] E. Burattini, A. Finzi, S. Rossi, and M. Staffa. Attentional human-robot interaction in simple manipulation tasks. In *Proc. of HRI-2012, Late-Breaking Reports*, 2012.

[83] J. R. Fontaine, K. R. Scherer, E. B. Roesch, and P. Ellsworth. The world of emotion is not two-dimensional. *Psychological Science*, 18:1050–1057, 2007.

[84] P. F. Baldi and L. Itti. Of bits and wows: A bayesian theory of surprise with applications to attention. *Neural Networks*, 23(5):649–666, Jun 2010.

[85] E. Hudlicka. To feel or not to feel: The role of affect in human-computer interaction. *International Journal of Human-Computer Studies*, pages 1–32, 2003.

[86] Balta H., R. Mil, Rossi S., Iengo S., and Siciliano B. Adaptive behavior-based control for robot navigation: A multi-robot case study. *Information, Communication and Automation Technologies (ICAT), 2013 XXIV International Symposium on*, 2013.

[87] S. Patel, K. R. Scherer, E. Bjorkner, and J. Sundberg. Mapping emotions into acoustic space: The role of voice production. *Biological Psychology*, pages 93–98, 2011.

[88] Jespersen. *Lehrbuch der Phonetic*. B.G. Teubner, Leipzig e Berlin, 1920.

[89] JT. Coull. Neural correlates of attention and arousal: insights from electrophysiology, functional neuroimaging and psychopharmacology. *Prog Neurobiol*, 55(4):343–61, 1998.

[90] B.Gerkey, R. Vaughan, and A. Howard. The player/stage project: Tools for multi-robot and distributed sensor systems. *In Proc. ICAR 2003*, pages 317–323, 2003.