From Vision to Actions
Towards Adaptive and Autonomous Robots

Jürgen Leitner

Abstract

This document describes my proposed PhD research and presents a review of my work done so far, as well as, an outline of the work to be done to finalise my dissertation.

The goal of this research is to overcome limitations of highly complex robotic systems, especially in terms of autonomy and adaptation. Although robotics research has seen advances over the last decades robots are still not in wide-spread use outside industrial applications. Yet a range of proposed scenarios have robots working together, helping and coexisting with humans in daily life. In all these a clear need to deal with a more unstructured, changing environment arises.

The main focus of research is to investigate the use of visual feedback for improving reaching and grasping capabilities of complex robots. A developmental, step-wise learning approach is proposed to generate adaptive and autonomous perception and behaviours for the iCub humanoid robot. To facilitate this a combined integration of computer vision, artificial intelligence and machine learning is employed on the robot.

Research Advisor
Jürgen Schmidhuber

Research Co-advisor
Alexander Förster

Internal Committee Members
Michael Bronstein, Rolf Krause

External Committee Members
Peter Corke, Stefano Nolfi
This proposal has been approved by the dissertation committee and the director of the PhD program:

__________________________
Jürgen Schmidhuber, Research Advisor and Committee Chair,
IDSIA (Dalle Molle Institute for Artificial Intelligence), Switzerland

__________________________
Alexander Förster, Research Co-Advisor,
IDSIA (Dalle Molle Institute for Artificial Intelligence), Switzerland

__________________________
Michael Bronstein, Università della Svizzera Italiana, Switzerland

__________________________
Peter Corke, Queensland University of Technology, Australia

__________________________
Rolf Krause, Università della Svizzera Italiana, Switzerland

__________________________
Stefano Nolfi, Consiglio Nazionale delle Ricerche, Italy

__________________________
Antonio Carzaniga, PhD Program Director,
Università della Svizzera Italiana, Switzerland
# Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Contents</strong></td>
<td>iii</td>
</tr>
<tr>
<td>1 <strong>Introduction</strong></td>
<td>1</td>
</tr>
<tr>
<td>2 <strong>Background</strong></td>
<td>3</td>
</tr>
<tr>
<td>2.1 Artificial Intelligence, Machine Learning and Robotics</td>
<td>3</td>
</tr>
<tr>
<td>2.2 Vision and Robot Perception</td>
<td>6</td>
</tr>
<tr>
<td>2.3 Object Manipulation and Robots</td>
<td>7</td>
</tr>
<tr>
<td>3 <strong>Towards Autonomous Object Manipulation in Humanoid Robots</strong></td>
<td>9</td>
</tr>
<tr>
<td>3.1 Perception</td>
<td>10</td>
</tr>
<tr>
<td>3.2 Motion Generation and Learning</td>
<td>16</td>
</tr>
<tr>
<td>3.3 Sensorimotor (Eye-hand) Coordination and Adaptation</td>
<td>17</td>
</tr>
<tr>
<td>4 <strong>Timeline and Schedule for Future Work</strong></td>
<td>19</td>
</tr>
<tr>
<td>4.1 Timeline of Work Performed So Far</td>
<td>19</td>
</tr>
<tr>
<td>4.2 Schedule for Completion of the Dissertation</td>
<td>20</td>
</tr>
<tr>
<td><strong>Bibliography</strong></td>
<td>21</td>
</tr>
<tr>
<td><strong>Appendix A: icVision Architecture</strong></td>
<td>31</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

In the last century robots have transitioned from science fiction to science fact. After centuries of imagining automated machines that would help humans, the last few decades brought into existence an ever growing number of these programmable appliances. The field of robotics developed from interdisciplinary research in the fields of electronics, mechanics and computer science, with the first prominent robots, such as the famous Shakey at SRI [Nilsson, 1969, 1984], appearing in academia during the 1960s. In its honour the Association for the Advancement of Artificial Intelligence (AAAI) awards trophies, named "Shakeys", to the best robot and artificial intelligence videos every year. Our work on the iCub was honoured with the best student video award recently. In the 1980s the digital revolution, yielding programmable personal computers and embedded systems, kick-started the creation of robotic systems and their use in large scale industrial production. Nowadays robotic arms are fulfilling automation tasks in factories all over the world. In recent years the field is moving towards other areas of utilisation, such as, for example, the use of robots in household settings. Proposed applications range from cleaning tasks, grocery shopping to helping in hospitals and care facilities. These involve the robots working around humans in daily life situations.

My research presented herein is aiming to overcome the limitations of current robots, with a focus on object manipulation in unstructured environments using a complex humanoid robot. A better perception and sensorimotor coordination is expected to be one of the key requirements to increase current robotic capabilities [Ambrose et al., 2012; Kragic and Vincze, 2009]. To facilitate this a combined integration of computer vision, artificial intelligence (AI) and machine learning (ML) on the robot is proposed.

To decide and act in unstructured environments, the robot needs to be able to perceive its surroundings, as it is generally not feasible or possible to provide the robot with all the information it needs a priori. This might be because of hardware limitations, such as limited storage space, or because the information is simply not available or unknown at the time (the most striking example for this is robotic space exploration). Therefore to extend robotic applications the need for more autonomous, more intelligent robots
Sensory feedback, that is the robot’s perception, is of critical importance. Creating a useful perception system is still a hard problem, but required for acting in a purposeful, some might say ‘intelligent’, way. In order to work naturally in human environments robots will need to be much more flexible and robust in the face of uncertainty and unseen observations. To create better ‘perception’, ‘motion’ and ‘coordination’ and deal with uncertainties I use various ML and AI techniques. The aim is to answer the question, whether it is possible (and if so how) for a humanoid robot to use vision and visual feedback to improve its reaching, grasping and (general) manipulation skills. Sensory information collected might be incomplete or inconsistent demanding means of managing uncertainty. Research in the fields of AI and ML has extensively explored ways of dealing with incomplete information in the last decades. A wide range of sensors have been used to build models of the environment the robot is placed in. Visual feedback, though it tends to be harder to interpret than other sensory information, is an active research area. A good motivation is that our world is built around (human) visual perception, which also means that to allow ‘natural’ interaction of robots it needs to be able to understand its environment based on the camera images it receives.

Another aspect of these ‘natural’ interactions is the human capability of adapting to changing circumstances during the action execution. Even if the environment can be perceived precisely it will not be static in most settings. For this reason the robot needs to embody a certain level of adaptation also on the motor side. This flexibility could again be provided by AI and ML techniques, leading to robots capable of reaching, grasping and manipulating a wide range of objects in arbitrary positions.

There are currently still no robotic systems available that perform autonomous dexterous manipulation (or do so only in very limited settings). At IDSIA the iCub humanoid robot, a state-of-art high degree-of-freedom (DOF) humanoid robot is available (see Figure 1.1) which was specifically designed for object manipulation research in robotic systems. The overarching goal of this research is to extend the capabilities of our experimental platform to allow for more autonomous and more adaptive behaviours.

![Figure 1.1. Our research platform, the iCub humanoid robot.](image)
Chapter 2

Background and Related Work

Object manipulation in real-world settings is a very hard problem in robotics, yet it is one of the most important skills for robots to possess [Kemp et al., 2007]. Through manipulation they are able to interact with the world and therefore become useful and helpful to humans. Yet to produce even the simplest human-like behaviours, a humanoid robot must be able to see, act, and react continuously. Even more so for object manipulation tasks, which require precise and coordinated movements of the arm and hand. The understanding of how humans and animals control these movements is a fundamental research topic in cognitive- [Posner, 1989] and neuro-sciences [Jeannerod, 1997]. Despite the interest and importance of the topic, e.g. in rehabilitation and medicine, the issues and theories behind how humans learn, adapt and perform reaching and grasping behaviours remain controversial. Although there are many experimental studies on how humans perform these actions, the development of reaching and grasping is still not fully understood and only very basic computational models exist [Oztop et al., 2004]. Vision is seen as an important factor in the development of reaching and grasping skills in humans [Berthier et al., 1996; McCarty et al., 2001]. For example, imitation of simple manipulation skills has been observed already in 14-month-old infants [Meltzoff, 1988]. Robots in contrast are only able to perform (simple) grasps in specific settings.

2.1 Artificial Intelligence, Machine Learning and Robotics

Although research has developed algorithms to play chess on a level good enough to win against (and/or tutor) the average human player [Sadikov et al., 2007], the robotic manipulation of a chess piece, in contrast, is still not feasible at a human (not even a child like) level. To produce even the simplest autonomous, adaptive, human-like behaviours, a humanoid robot must be able to, at least:

- Identify and localise objects in the environment, e.g. the chess pieces and board
- Execute purposeful motions for interaction, e.g. move a piece to a desired position
State-of-the-art humanoid robots such as Honda’s Asimo [Sakagami et al., 2002], NASA’s Robonaut and R2 [Bluethmann et al., 2003; Diftler et al., 2011], Toyota’s Partner Robots [Takagi, 2006], Justin and TORO [Ott et al., 2012] from the German Aerospace Center (DLR) and the iCub [Tsagarakis et al., 2007] are stunning features of engineering. They are capable of producing complex, repeatable motions allowing them to walk, run and manipulate delicate objects such as musical instruments [Kusuda, 2008; Solis and Takanishi, 2011]. The caveat is that every last detail of these behaviours is currently programmed by hand (often with the help of advanced tools). As a result those resourceful machines, full of capabilities, are not yet realising their full potential due to the complexity of programming them in a flexible way. Their capacity to adapt to changes in the environment or to unexpected circumstances are still very limited.

Along with the concept of adaptiveness comes the notion of acting autonomously. To exploit the full versatility of an advanced robot, a broad spectrum of actions is required. In recent years the interest of the robotics community in Artificial Intelligence and Machine Learning techniques as tools to build intelligent robots has increased. At the same time, the ML community has shown an increased interest in robots as ideal test bench and application for new algorithms and techniques [Konidaris, 2013]. It uses these techniques, and tests them, in a physical entity and real environments to interact/move in. This ‘embodiment’ is seen as an important condition for the development of cognitive abilities both in humans and robots [Brooks, 1999; Wolpert et al., 2001; Pfeifer et al., 2007]. ML algorithms, have been applied in experimental robotics to acquire new skills, however the need for carefully gathered training data, clever initialisation conditions, and/or demonstrated example behaviours limits the autonomy with which behaviours can be learned. To build robots that can perform complex manipulation skills that help users in their activities of daily living (ADL) is the aim of various research projects [Canigelosi et al., 2008; GeRT, 2012; THE, 2012; WAY, 2012].

**Robot Learning: Cognitive, Developmental and Evolutionary Robotics**

As mentioned above the programming of these highly complex robot systems is a cumbersome, difficult and time-consuming process. It generally also describes each precise movement in detail, allowing little to no flexibility or adaptation during execution. **Robot Learning** generally refers to research into ways for a robot to learn certain aspects by itself. Instead of providing all information to the robot a priori, for example, possible motions to reach a certain target position, the agent will through some process ‘learn’ which motor commands lead to what action. **Cognitive Robotics, Developmental Robotics** and **Evolutionary Robotics** are fields that recently emerged with that specific aim to investigate how robots can ‘learn’ for themselves and thereby generate more autonomous and adaptive capabilities.

---

1‘Embodiment’, i.e., the claim that having a body that mediates perception and affects behaviour plays an integral part in the emergence of human cognition and how we learn.
In *Cognitive Robotics*  [Asada et al., 2001] the aim is to provide robots with cognitive processes, similar to humans and animals. An integrated view of the body is taken, including the motor system, the perceptual system and the body’s interactions with the environment. The acquisition of knowledge, may it be through actions (e.g. motor babbling) or perception is a big part of cognitive robotics research. Another is the development of architectures for these tasks. A variety has been proposed [Burgard et al., 2005; Shanahan 2006; Vernon et al. 2007b; Chella et al., 2008; Wyatt and Hawes, 2008], but the promised improvements in robotic applications still need to be shown. This can be attributed to the varying definitions of cognition and the complex human cognitive system, whose workings are still not fully understood. To build cognitive architectures two distinct approaches have been tried. The research seems to mainly focus on top-down architectures. A bottom-up approach has been described as more suitable for the use with robots (e.g. the proposed *iCub* cognitive architecture Vernon et al. [2007a]).

*Developmental Robotics*  [Asada et al., 2001; Weng, 2004; Kuipers et al., 2006; Mee den and Blank, 2006; Asada et al., 2009] is aiming to put more emphasis on the development of skills. It is an interdisciplinary approach to developmental science. It differs from the previous approaches, as the engineer only creates the architecture and then allows the robot to explore and learn its own representation of its capabilities (sensory and motor) and the environment. As above, the body and its interactions with the environment are seen as being fundamental for the development of skills. Aims are to build adaptive robotic systems by exploration and autonomous learning, i.e. learning without a direct intervention from a designer [Lungarella et al., 2003]. Here interesting areas to explore are selected by building on previous knowledge, while seeking out novel stimuli.

*Evolutionary Robotics* [Harvey et al., 1997; Nolfi and Floreano, 2000; Doncieux et al., 2011] is another approach to add adaptiveness and developmental processes to robots. It emerged as a new approach to overcome the difficulties of designing control systems for autonomous robots: (a) coordinating the (increasing) number of DOF both in mechanics and control is hard, especially since the complexity scales with the number of possible interactions between parts [Cliff et al., 1993] (see ‘Curse of Dimensionality’) (b) the environment and how the robot interacts with it are often not known before. Its main focus is on evolve a control system based on artificial neural networks. These neuro-controllers (NC), inspired by the neuron activity in the human brain, have been shown to work in a wide range of applications [Nolfi et al., 1994; Dachwald, 2004; Leitner et al., 2010]. An important issue is that to ‘learn’ behaviours, a large number of iterations (or generations) is required. This works fine in simulation but is hard to achieve on a real robotic platform. Nolfi et al. [1994] showed that evolving a NC on hardware is, while time consuming, feasible, at least for simple mobile robots. Hybrid approaches, where NCs are trained first in simulation and then transferred to the real hardware, seem preferential. The performance of the controllers in the real world can then be used to improve the simulation [Bongard et al., 2006]. How to effectively train and apply NCs to real, high-DOF hardware is still an open research question.
Other Approaches to robot learning have been developed in the past. The area of Reinforcement Learning (RL) [Sutton and Barto, 1998] has appealed to many roboticists, especially for learning to control complex robotic systems. A general RL algorithm and the means to inform the robot whether its actions were successful (positive reward) or not (negative reward) is all that is required. RL and its applicability to humanoid robots has been investigated by [Peters et al., 2003]. Imitation Learning or Apprenticeship Learning is of importance in human skill development as it allows to transfer skills from one person to another. In robotics Robot Learning from demonstration or Programming by Demonstration is a similar paradigm for enabling robots to learn to perform novel tasks. It takes the view that an appropriate robot controller can be derived from observations of a another agent’s performance thereof [Schaal, 1999].

2.2 Vision and Robot Perception

To be useful in the above proposed scenarios a robot must be able to see, act, and react continuously. Perception is a key requirement in order to purposefully adapt robot motion to the environment, allowing for more successful, more autonomous interactions. Vision and the visual system are the focus of much research in psychology, cognitive science, neuroscience and biology. A major problem in visual perception is that what individuals ‘see’ is not just a simple translation of input stimuli (compare optical illusions). The research of Marr in the 1970s led to a theory of vision using different levels of abstraction, from a two-dimensional visual array (projected onto the retina) to a three-dimensional description of the world as output. The stages include: a 2D sketch of the scene (using feature extraction), a sketch of the scene (using textures to provide more information) and a 3D model of the world [Marr, 1982].

Research on perception has been an active component for developing artificial vision systems, in industry and robotics. Computer Vision generally describes the field of research dealing with acquiring, processing, analysing, and understanding images in order to produce decisions based on the observation. As a scientific discipline, computer vision is concerned with the theory behind artificial systems that extract information from a variety of image data, such as video sequences, views from multiple cameras, or multidimensional/spectral data, e.g. from a medical scanner. The fields of computer vision and AI have close connections, e.g. autonomous planning or decision making for robots require information about the environment, which could be provided by a computer vision system. AI and computer vision share other topics such as pattern recognition and learning techniques. Furthermore computer vision spawned a multitude of research sub-domains, such as, scene reconstruction, event detection, object recognition, motion estimation, etc.

The first thing done by the human visual system is widely understood to be the segmentation and detection of objects in the visual space (the 2D images). In vision research this would fall into the area of ‘image processing’ [Gonzalez and Woods, 2002].
The techniques generally provide ways of extracting information from the image data and can be grouped into the following categories: pre-processing (e.g. noise reduction, enhancement, scaling, etc.), feature extraction (e.g. lines, edges, interest points, etc.), segmentation (e.g. separating fore- and background), and high-level processing (e.g. recognition and decision making). Another important topic in computer vision is ‘image understanding’. With the aid of geometry, physics, statistics, and learning the goal is to mimic the abilities of the human (visual) perception system.

Research into vision for the special requirements of robotic systems is referred to as robot vision or machine vision [Horn 1986; Hornberg 2007]. For example, visual feedback has extensively been used in mobile robot applications, for obstacle avoidance, mapping and localisation. With the advancement of humanoids and the increased interest in working around humans, object detection and manipulation are more and more driving the development of robot vision systems. An important problem is that of determining whether or not the image data contains some specific object, feature, or activity. While this has been researched for quite some time already, the task seems harder than expected and no solution for the general case of detecting arbitrary objects in arbitrary situations exists. Most of the work is heavily relying on artificial landmarks and fiducial markers to simplify the detection problem. Furthermore existing methods can at best solve it for specific objects (simple geometries, faces, printed or hand-written characters, or vehicles) and in specific situations (in terms of well-defined illumination, background, and pose of the object wrt. the camera). For a detailed introduction and overview of the foundations and the current trends the reader is referred to the excellent survey by Kragic and Vincze [2009].

2.3 Object Manipulation and Robots

Only after the scene is observed and the robot has an idea about which objects are in the environment it can start interacting with these. But manipulating arbitrary objects is not a trivial thing, even for humans. The development of hand control in children, for an apparently simple, prototypical precision grasp task is not matured until the age of 8-10 years [Forssberg et al., 1991]. Moreover, complexity, as can be seen by the number of neurons comprising the control of the arm and hand, is staggeringly high. Even after manipulation skills have been learnt they are constantly adapted by an perception-action loop to yield desired results.

In recent years good progress was made with robotic grasping of objects. The various manipulators, mainly hands and grippers, and techniques clearly improved. Also novel concepts of ‘grippers’ have been designed and some are quite ingenious solutions to some of the issues. One such example is the granular gripper made by Brown et al. [2010], which is made out of grounded coffee beans which are able to ‘flow’ around the object and then fixed in position by creating a vacuum. This concept has recently been extended to a full sized arm elephant-trunk-like arm [Cheng et al., 2012]. Also in terms
of how to grasp objects with regular grippers and ‘hands’ recent results highlight the advanced state of research in grasping. For example, Maitin-Shepard et al. [2010], with their research showed that robots are able to pick up non-rigid objects, such as, towels. Their robot is able to reliably and robustly pick up a randomly dropped towel from a table by going through a sequence of vision-based re-grasps and manipulations—partially in the air, partially on the table. In the DARPA ARM project, which aims to create highly autonomous manipulators capable of serving multiple purposes across a wide variety of applications, NASA’s JPL winning team showed an end-to-end system that allows the robot to grasp diverse objects (e.g. power drill, keys, screwdrivers, ...) from a table [Hudson et al., 2012; Hebert et al., 2012]. On the other hand Saxena et al. [2008] have presented a way for a robot to learn, from only a small number of real world examples, where good grasping points are on a wide variety of previously unknown objects.

All this has lead Dr. Pratt, DARPA Manager of robotics related research projects, to his somewhat controversial statement of “Grasping is solved” at last year’s IROS conference. While this might be a bit too optimistic it seems like the research is at a good enough state to have better system integration. The direct interface between various components, which makes robotics such a hard but interesting field, clearly need to improve to allow for robust object manipulation. Only by combining sensing and control of the whole robotic platform a fully functional ‘pick-and-place’ capable system will appear. To allow for a variety of objects to be picked up from various positions the robot needs to see, act and react within such a tightly integrated control system.

A must read for roboticists is Corke’s [2011]’s ‘Robotics, Vision and Control’. It puts in the spotlight the integration of these three components and describes common pitfalls and the issues that arise with integration.

Our Experimental Platform: The iCub Humanoid at IDSIA

The iCub humanoid [Tsagarakis et al., 2007] (shown in Figure 1.1) is an open-system robotic platform developed during various European projects. It consists of two arms and a head attached to a torso roughly the size of a human child. The head and arms follow an anthropomorphic design and provide a high DOF system that is used for researching human cognitive and sensorimotor development. The iCub is an excellent experimental platform for cognitive and sensorimotor development and embodied artificial intelligence [Metta et al., 2010] and was designed to investigate human-like object manipulation. The robot’s movements need to be coordinated with feedback from visual, tactile, and acoustic perception. Of interest is also to test the scalability of machine learning methods towards such complex systems interacting with the real-world [Levinson et al., 2010; Sicard et al., 2011; Leitner et al., 2012].

[3] Legs are under development – recently first standing/balancing demos with the iCub were shown.
[4] Acoustic feedback might be used when manipulation objects to know if a grasp was successful.
Chapter 3

Towards Autonomous Object Manipulation in Humanoid Robots

One of the most important problems in robotics at the moment is to improve the robots’ abilities to understand and interact with the environment around it: it needs to be able to perceive, detect & locate objects in its surrounding and then be able to plan and execute actions to manipulate these. To enable a more autonomous object manipulation, more specifically how to enable some level of eye-hand coordination to perform actions more successfully, is of high interest to the robotics community (see e.g. NASA’s Space Technology Roadmap calls for ‘Real-time self-calibrating hand-eye System’ [Ambrose et al., 2012]). I am working on integrating learning and object manipulation in robotic systems. IDSIA is especially interested in the investigation of how rewards and motivation effect the development of complex actions and interactions between an (embodied)

![Figure 3.1. IDSIA’s research towards a functional eye-hand coordination on the iCub.](image)

9
agent and the environment. My aim is to improve adaptivity and autonomy in robot grasping based on visual feedback to close the loop and perform grasping of objects, while adapting to unknown, complex environments. Figure 3.1 gives an overview of the research on-going at IDSIA towards the goal of integrating eye-hand coordination into the iCub. On top the perception side is shown. It includes the modules for the detection and identification of objects (in the images), as well as, the localisation (in 3D Cartesian space). The details of these implementations are described in Section 3.1. The bottom have shows the action, motion side. To generate motion from learning (see Section 3.2) a crucial feature is the avoidance of collisions, both between the robot and the environment and the robot and itself. Section 3.3 aims to describe how we aim to generate a level of eye-hand coordination not previously seen on the iCub.

### 3.1 Perception

My PhD research so far has put an emphasis on investigating visual perception in robots. The term refers, in humans and animals, to the ability to interpret their environment based on the information from visible light reaching the eyes. In robots, to imitate eyes, cameras are used to capture the light of the environment. The reason for the focus on vision is twofold, firstly the sensing capabilities of robotic platforms (cameras are cheap) and secondly, vision is the most important sense for humans. Another reason is that with a humanoid robot, such as the iCub at IDSIA, a natural tendency exists to be inspired by human perception and behaviour. A lot of my work, in the last two years, focussed on building a visual perception system for the iCub humanoid, somewhat similar to the description of human perception by Marr [1982]. Various experiments performed over the course of the last two years are described briefly in the next sections.

**Framework for Robot Vision and Cognitive Perception**

During the first years of my PhD I developed the icVision framework at IDSIA [Leitner et al., 2012c, 2013b]. It allows for an easier development, testing and integration of the on-going computer vision research into the real hardware. One of the main design goals is to allow rapid prototyping and testing of vision software on the iCub and reduce development time by removing redundant and repetitive processes. icVision is implemented in C++, open-source and uses the YARP middleware and OpenCV for the underlying image processing. Its various modules provide means to detect (known) objects in the images and estimate their 3D location wrt. the robot’s reference frame (see Appendix A). It interfaces with the MoBeE environment to place a (given) geometric subject into the robot’s world model enabling the robot to, for example, avoid colliding with them during operation. An overview of the experiments done using the framework can be found in [Leitner et al., 2013a].

My research is partly funded by the European Union grant ‘IM-CLeVeR’ [Baldassare et al., 2009].
Visual Perception and Object Detection

As mentioned in Section 2.2, object detection is not yet solved in a general sense, yet it is a very important issue to enable autonomous grasping. A hard problem on the iCub in particular although there has been some recent progress, especially from the lab at IIT [Giliberto et al., 2011; Fanello et al., 2013; Gori et al., 2013]. For autonomous object manipulation with humanoid robots it is of importance to detect and identify the objects to interact with in the environment. This is still a challenging problem in robotics, especially in settings where lighting varies, viewing angles change and the environment can be described as ‘cluttered’ (i.e. lots of different objects in the scene, partly obstructing each other). Given the maturity of the field of image processing, it should be possible to construct programs that use much more complicated image operations and hence incorporate domain knowledge.

We developed a technique based on Cartesian Genetic Programming (CGP) [Miller, 1999, 2011] allowing for the automatic generation of computer programs for robot vision tasks [Harding et al., 2013]. A large subset of the functionality of the freely available OpenCV image processing library [Bradski, 2000] are used as building-blocks of these programs. The implementation provides an effective method to learn (unique) object detection modules. If the training set is chosen correctly, even different lighting conditions are no longer problematic [Leitner et al., 2012d]. The basic algorithm works as follows: Initially, a population of candidate solutions is composed of randomly generated individuals. Each of these individuals, represented by its genotype, is tested to see how well it performs the given task, i.e. segment the object of interest from the rest of the scene in a set of test images. This step, known as evaluating the ‘fitness function’, is used to assign a numeric score to each individual in the population. Generally, the lower this error, the better the individual is at performing the task. The next step generates a new population of individuals from the old population. This is done by taking pairs of the best scored genotypes and performing functions analogous to recombination and mutation. Fitness scores are then calculated for these new individuals are then tested using the fitness function. The process of test and generate is repeated until a solution is found or until a certain number of individuals have been evaluated.

While much of the previous work with machine learning in computer vision has focused on the use of grey scale images, our implementation Cartesian Genetic Prog-
Perception for Image Processing (CGP-IP) can handle colour images with multiple channels. Treatment for colour images is to separate them into RGB (red, green and blue) and HSV (hue, saturation and value) channels, and provide these as available inputs. Each available channel is presented as an input to CGP-IP, and evolution selects which inputs will be used, as all functions operate on single channel images. Our implementation generates human readable C# or C++ code based on OpenCV to be run directly on the real hardware – using our icVision framework. An example of a CGP-IP genotype is illustrated in Figure 3.2. The first three nodes obtain the components from the current test image (e.g. grey scale version, red and green channels). The fourth node adds the green and red images together. This is then dilated by the fifth node. The sixth node is not referenced by any node connected to the output (i.e. it is neutral), and is therefore ignored. The final node takes the average of the fifth node and the grey scale input.

We apply CGP-IP to a variety of problems in computer vision. The efficacy of our approach has been shown in basic image processing (noise reduction), medical imaging (mitosis detection) [Harding et al., 2013] and object detection in robotic applications.

**Visual Terrain Classification** As a first example, we apply CGP-IP to visual detection and identification of rocks and rock formations. Automatically classifying terrain such as rocks, sand and gravel from images is a challenging machine vision problem. Using data from mobile robots (on other planets) our system learns classifiers and detectors for different terrain types from hand-labelled training images. The learned program outperforms currently used techniques for classification when tested on a panorama image collected by the Mars Exploration Rover Spirit [Leitner et al., 2012b].

**Object Detection and Identification** Vision systems are expected to work in ‘real-world’, cluttered environments, with lots different objects visible. We show that the CGP-IP approach is able to work robustly in different lighting conditions, and when the target object is moved or rotated. To train, a number of images are collected from the *iCub* cameras. In each one the target (and other objects) is repositioned. Hence our training set implicitly contains multiple views, different angles, scales and lighting conditions. In the training set the target object is segmented by hand. Figure 3.3 shows the result of one evolved programme for a tea box. Nine training images were used for

![Figure 3.3](http://www.youtube.com/watch?v=xszOcj4AlE)

*Figure 3.3.* The detection of a tea box in changing lighting condition performed by a learned CGP-IP filter. Full video at: [http://www.youtube.com/watch?v=xszOcj4AlE](http://www.youtube.com/watch?v=xszOcj4AlE)
training. It can be seen that the trained individual is able to cope with variations in scale, orientation and lighting. Unique programmes, allowing for identification as well as detection, were evolved for a variety of objects of interest, e.g. different cups, tin cans, kids blocks and tea boxes.

**Robot Hand Detection** CGP-IP was then applied to detect the robot’s own hands and fingers in the images – an important prerequisite for eye-hand coordination. Previous approaches make use of external systems or markers to provide the position of the hand. In contrast we do so using only camera images. At first we verified our approach by visually detecting the hands of the Nao robot, which are of a less complex design and also simpler in appearance. The detection of the mechanically complex hands and fingers of the *iCub* seems to be a much harder computer vision task. Separate programs resulted in better detection of the fingers and the hand \cite{Leitner2013} (Figure 3.4).

**Autonomous Learning of Object Detection** CGP-IP is using a supervised learning technique. An obvious limitation is that the training images need to be provided. In our first object detection experiments with CGP-IP on the robot these images were hand-labelled in a rather tedious and time-consuming fashion. We noticed that the area specified by the labels did not need to be very precise as our technique seems to be robust to some errors in the provided bounding labels. The evolved solution was able to segment the object according to its outline rather than the label provided. A visual system that actively explores the environment is combined with CGP-IP to create a more autonomous detection of objects \cite{Leitner2012}. A neuromorphic model, based on \cite{itti98}, is applied to identify salient points in the scene. Around these points features are detected using the FAST corner detection algorithm \cite{Rosten2010}. Local descriptors are calculated using Gabor wavelets, i.e. convolutional kernels having the shape of plane waves restricted by a Gaussian envelope function \cite{Wiskott1997}. The points are matched between the two camera images, to allow a rough segmentation of the object based on depth information. Clustered matched interest points provide the rough segmentation for the CGP-IP training set. The system enables the robot to
autonomously learn to detect and identify various objects on a table and build a world model using the spatial information as described below.

It is worth mentioning that this object detection, identification, and tracking is currently implemented to use only single frames. We believe that the performance can be increased further by using information from multiple, sequential images, such as, optical flow, motion tracking, or just simply propagating information about the segmentation (e.g., centre and outline of blob) to the next frame.

Spatial Perception and Object Localisation

To localise objects in the environment the iCub has to rely solely on a visual system based on stereo vision, similarly to human perception, The iCub's cameras are mounted in the head of the robot therefore the method for localisation must be able to cope with motion to work while the robot is controlling its gaze and upper body for reaching. As part of my thesis work I am investigating how spatial representation can be learned from visual feedback [Leitner et al., 2012d,e].

Developing an approach to perform robust localisation on a real humanoid robot is important. It is a necessary input for current on-line motion planning, reaching, and object manipulation systems. Stereo vision describes the extraction of 3D information out of images and is similar to the biological process of stereopsis in humans. Its basic principle is the comparison of images taken of the same scene from different viewpoints. To obtain a distance measure the relative displacement of a pixel between the two images is used [Hartley and Zisserman, 2000]. The most logical choice of coordinate system, from a planning and control standpoint, in our case is a world reference frame originating from the robot’s hip. To transform coordinates from the eyes' frames to the world coordinates an accurate kinematic model of the robot is necessary.

While these approaches, based on projective geometry, have been proven effective under carefully controlled experimental circumstances, they are not easily transferred to robotics applications. For the iCub platform several approaches have previously been developed. One of these methods is a biologically inspired approach mimicking the human retina [Traver and Bernardino, 2010]. Another, the ‘Cartesian controller module’, provides basic 3D position estimation functionality [Pattacini, 2011] and gaze control. This module works very well on the simulated iCub, however on the hardware platform multiple sources of noise and errors exist. It therefore generates absolute errors in the 2-4 cm range, in an area where the humanoid can reach and manipulate objects.

We use a learning based approach to develop spatial perception on the iCub. A Katana robotic arm was used to teach the iCub how to perceive the location of the objects it sees. To do this, the Katana positions an object within the shared workspace, and tells the humanoid where it has placed it. While the iCub moves about, it observes the object, and the robot ‘learns’ how to relate its pose and visual inputs to the object location in

---

2Even bigger errors are seen at the edge of the frame or further away from the robot.
Cartesian coordinates. We show that satisfactory results can be obtained for localisation even in scenarios where the kinematic model is imprecise or not available. Furthermore, we demonstrate that this task can be accomplished safely. For this task MoBeE was used to prevent collisions between multiple, independently controlled, heterogeneous robots in the same workspace (see Section 3.2).

The learning is done on a feed-forward artificial neural network (ANN), consisting of three layers: one input layer, a hidden layer, and an output layer. This structure of an ANN is sometimes also referred to as a multi-layer perceptron [Fausett 1994]. Each input, in our case image coordinates and robot pose information, arrives at an input node in the input layer of our ANN. As our network is fully connected this input is then provided to every node at the hidden layer. Each connection has a weight attached to it. In addition to all inputs, each hidden node also has a connection to a bias node. Similarly the output value of every hidden node is then provided to the output node. These connections are again weighted. The output node is as well connected to a bias node. We use neurons in the hidden and output layer containing a sigmoidal activation function of the following form to calculate their respective outputs. The ANNs were trained using the standard error back-propagation algorithm [Russell and Norvig 2010] method on the dataset collected. This method is a generalisation of the delta-rule and requires the activation function of the neurons to be differentiable. Back-propagation can be described in two phases: (i) a forward propagation and (ii) a weight update phase. In phase one inputs from the training dataset are propagated through the network using the current weights for each connection. The error between the networks output and the correct output is then used to calculate a gradient at each connection, which defines how the weights will be updated.

We show that the robot can learn to estimate object locations based on visual perception. Our approach does not need explicit knowledge of a precise kinematic model or camera calibration. It learned to localise objects with comparable accuracy to currently available systems. Furthermore our approach is computationally less expensive and less dependent on other concurrently running modules (such as image rectifying techniques). As the learnt models are 'light weight' they could easily be incorporated into embedded systems and other robotic platforms. The resulting locations are easily usable with the current operational space controller on the iCub. The results using our first collected dataset show the feasibility of the method to perform full 3D position estimation. The experiments show that the approach can learn to perceive the location of objects in contrast to currently existing engineered methods.

In addition our method can learn whether an object visible in both cameras is within the reachable workspace of the robot. While learning spatial perception in Cartesian coordinates is not necessarily the best option, also humans are not good at precisely determining distances of perceived objects, it enables the use of currently existing operational space controllers on the iCub.
3.2 Motion Generation and Learning

Even with great advances in sensing technologies, the motion of the robot still needs to be controlled to perform manipulation tasks. While explicit model-based control is still the prime approach — which work very well when the world’s state is known — it has limitations when it comes to uncertainties. Robots, used in a human environment, cannot expect to estimate the state of the surroundings with certainty. It seems almost inevitable that learning will play an important role in robot manipulation. Directly programming robots by writing code can be tedious, error prone, and inaccessible to non-experts. Through learning, robots may be able to reduce this burden and continue to adapt even after being deployed, creating some level of autonomy. As our platform is a highly complex system, learning and the approaches described above are deemed to be the most suitable. For example, Hart et al. [2006] showed that a developmental approach can be used for robots to learn to reach and grasp. When attempting to synthesise behaviour on a complex robot like the iCub, the shortcomings of state-of-the-art learning and control theories can be discovered and addressed in future research.

The issue of uncertainty in real-world applications has been addressing these with robust closed-loop controllers using sensory feedback. For example, Piatt et al. [2006] and the Robonaut group at NASA/JSC and Hart et al. [2006] at UMass Amherst have explored ways to learn and compose real-time, closed-loop controllers in order to flexibly perform a variety of autonomous manipulation tasks in a robust manner. Edsinger and Kemp [2006] at MIT often use hand-coded behavior-based controllers that specify tasks in terms of visual or other feedback driven control.

We have developed a novel, Modular Behavioural Environment (MoBeE) framework for humanoids and other complex robots, which integrates elements planning, and control, facilitating the synthesis of autonomous, adaptive behaviours. MoBeE has been used in various experiments with the iCub humanoid. From adaptive roadmap planning by integrating a reactive feedback controller into a roadmap planner [Frank et al., 2012], to teaching the iCub spatial perception [Leitner et al., 2012e]. It allows the integration of suitable learning methods for controlling the robot with existing software tools and the on-going research in other fields in our lab.

Generally the approach is to start with low complexity control, such as, controlling the robot’s gaze and will slowly include more DOF until the full 7DOF arms are controlled for reaching. This has been extensively investigated previously, also on the iCub, e.g. using a combination of open and closed loop control [Natale et al., 2007; Nori et al., 2007] and motor babbling [Caligiore et al., 2008]. To advance towards the goal of general object manipulation, where a robot can autonomously manipulate any given object within its workspace, it would ideally encode and reuse knowledge in terms of task features that are invariant to the object. Confronted with a novel instance of a specific task the robot needs to establish appropriate correspondences between objects and actions in its repertoire and their counterparts in the current task.
At IDSIA we pursue this by building task-relevant roadmaps (TRMs), using a new sampling-based optimiser called Natural Gradient Inverse Kinematics (NGIK) based on previous research in our lab \cite{Stollenga et al., 2013}. To build TRMs, NGIK iteratively optimises postures covering task-spaces expressed by arbitrary task-functions, subject to constraints expressed by arbitrary cost-functions, transparently dealing with both hard and soft constraints. TRMs are grown to maximally cover the task-space while minimising costs. Unlike Jacobian methods, our algorithm does not rely on calculation of gradients, making application of the algorithm much simpler. In our experiments NGIK outperforms recent related sampling algorithms.\footnote{Recently during the course of the IM-CleVeR final demo we showed that MoBeE also enables the robot to ‘learn’ to plan motions using reinforcement learning (RL) and ‘artificial curiosity’. While these fields usually are employed in toy-scenarios, e.g. navigation in a simple maze, in our case we embody it into the real high-DOF robotic hardware. A low-level reactive controller is coupled with our RL framework to find ‘interesting’ states through interaction with the environment. We show that the robot can learn models to represent large regions of the iCub’s configuration space \cite{Frank et al., 2013}.}

3.3 Sensorimotor (Eye-hand) Coordination and Adaptation

Although there exists a rich body of literature in computer vision, path planning, and feedback control, wherein many critical subproblems are addressed individually, most demonstrable behaviours for humanoid robots do not effectively integrate elements from all three disciplines. Consequently, tasks that seem trivial to us humans, such as picking up a specific object in a cluttered environment, remain beyond the state-of-the-art in experimental robotics.

The main aim of my doctoral research is a closer integration of the sensory and motor sides and creating an closely tied action-perception loop. We created interfaces between MoBeE and icVision allowing for a preliminary visual based localisation of the detected objects into the world model. Currently we are performing very simple operational space control of the robot arm based on the reported location using the approaches described above. This, rather open-loop control, is the first step towards visual guided reaching and grasping of objects with the iCub. It enables the robot to detect an object in the visual frame then localise it in Cartesian coordinates and eventually executing a reach, with either the existing operational space controller or our TRM approach. While this approach is already a step forward of current systems, it requires a very accurate calibration of both camera and mechanical links to be successful. However some mechanical non linearities still cannot be taken into account.

Clearly vision is an important thing for a humanoid robot, but it needs to be closely integrated into the control system. Sensorimotor Development is aiming to learn a basic eye-hand coordination. \cite{Langdon and Nordin, 2001} have shown this on a simple

\footnote{The demo video won the AAAI Best Student Video Award ‘Shakey’: \url{http://youtu.be/N6x2e1Zf_yg}}
humanoid robot using GP techniques. Various methods for learning this sensorimotor mapping have been investigated \[\text{Hoffmann et al., 2005; Hülse et al., 2010}\]. These lead to biologically inspired mappings, yet applying these directly to control the robot is still an issue. Adaptation is needed for precise object manipulation, as highly accurate models of the world and also of the robotic system can not be assumed in most interesting scenarios. An estimation of the robot kinematics might help in generating more precise motions. So far no module exists to estimate the kinematics of the \textit{iCub}, this is partly due to the openly available CAD models and thorough calibration procedures that need to/should be applied regularly. For other robotic platforms machine learning has been used to estimate the kinematic model, e.g. \text{Gloye et al., 2005} used visual feedback to learn the model of a holonomic wheeled robot and \text{Bongard et al., 2006} used low-dimensional sensory feedback to learn the model of a legged robot.

Currently I am looking at ways to close the control loop on complex, high-DOF humanoid robots. \textit{Visual Servoing} is a commonly used approach to closed-loop vision based control. It refers to the use of computer vision data to control the motion of a robot \[\text{Chaumette and Hutchinson, 2006, 2007}\]. It relies on techniques from image processing, computer vision, and control theory. The visual information is usually coming from a camera mounted on a robot manipulator or a mobile robot. Various configurations of the robot and camera(s) are possible, the most common in literature being the eye-in-hand case. Visual servoing (VS) has been investigated to be used with human-style robots, but there is yet no publication (known to the author) that uses a full humanoid setup, as in our research platform the \textit{iCub}. The issue here is that it is not a eye-in-hand nor a eye-to-hand system, as the camera is on the robot, but only some of the reaching movements will imply a motion also with the head (this only happens when the hip is moved). Generally there are two separate approaches: a position-based VS (PBVS) and an image-based VS (IBVS). There is no definitive answer which ones is better, PBVS tends to be used more in setups with a 3D sensor (e.g. LIDAR, Kinect, etc.) whereas IBVS seems to be preferential when using cameras, like in our case.

In the next few months I will add a (most likely) image-based VS to our system. With our current setup we should be able to measure the error between the hand (which we can filter) and the object we want to pick up, similar to previous approaches on less complex robots or with the help of external sensors \[\text{Hager et al., 1995}\]. More experiments will be added to show the possibility of ‘forcing’ the robot arm in operational (or image) space to reduce the error, hopefully leading to pick-and-place actions for a variety of objects at random location on the table.
Chapter 4

Timeline and Schedule for Future Work

The PhD regulations at USI state that the dissertation proposal shall also describe ‘a timeline of the work performed since the beginning of the PhD, and a schedule for conducting and completing the work’. The following sections are aiming to do exactly that.

4.1 Timeline of Work Performed So Far

I started working at the Dalle Molle Institute for AI (IDSIA) Robotics Lab in Oct 2010. After applying to the PhD programme of the Faculty of Informatics at the Università della Svizzera Italiana (USI), I officially became a PhD student in Feb 2011. The research work at IDSIA is project driven. The EU funded STIFF research project aimed to enhance biomorphic agility of robot arms & hands through variable stiffness and elasticity. The focus of my work was on using ML techniques to model data from human trials, in collaboration with J. McIntyre’s lab at UPD, Paris. The tasks were mainly focussed on reaching and motions along rigid surfaces with variable ‘stiffness’ applied by the human subject. The project finished in early 2012. The next project I worked on is IM-CLeVeR. It aims to develop a new methodology for designing robot controllers that can: "(1) cumulatively learn new, efficient skills through autonomous development based on intrinsic motivations, and (2) reuse such skills for accomplishing multiple, complex, and externally-assigned tasks." [Baldassare et al. 2009] My work focusses on developing and implementing a computer vision framework, for object detection and localisation on the iCub. I started investigating how object manipulation skills can be learned and how vision can help in those situations. This project is currently finishing.

In 2013 work started on the WAY project, in which, together with Scuola Superiore Sant’Anna in Pisa, we aim to develop non-invasive methods that allow bidirectional links between a hand assistive device and a patient. I apply machine learning techniques to decode multilevel biosignals (mainly EMG) for controlling a hand prosthesis.
4.2 Schedule for Completion of the Dissertation

**PhD Requirements at USI**

**Teaching**  I had the great opportunity to be teaching assistant to Dr. Alexander Förster for his course on “Systems Programming” (in the BSc track) already three times.

**Breadth of Knowledge**  USI requires from students to pass at least 12 ECTS points of courses in multiple disciplines of informatics. I attended the following:

- Introduction to Doctoral Studies (mandatory), 2 ECTS
- Intelligent Systems (cross listed from MSc programme), 4 ECTS
- Research Policy and Grant Proposal Writing, 3 ECTS
- Innovation and Patents, 2 ECTS
- Gate-Level Hardware Design and Arithmetics, 2 ECTS

In addition I attended three summer schools: the EUCog Summer School ‘Neural dynamics approach to cognitive robotics’ (2011), the iCub Summer School ‘Veni, Vidi, Vici’ (2012) and the IEEE Summer School ‘Robotic Vision and Applications’ (2012). These together should suffice the requirement for ‘breadth of knowledge’.

**Scientific/Academic Work**

During my PhD so far, especially thanks to a fruitful collaboration with Dr. Harding, I have been able to author various papers (most of them as first author), including one book chapter [Harding et al., 2013], 3 journal publications and several papers at highly-ranked conferences in the field of robotics (IROS), AI/ML (CEC, IJCNN), and developmental robotics (ICDL/EpiRob). Further submissions to various robotic journals are planned. I was also reviewing submissions for conferences and journals.

### 4.2 Schedule for Completion of the Dissertation

A rough plan for finishing my PhD project includes the following steps:

- closer sensorimotor integration of perception and actions [1-2 months]
- implementation of selected vision-based manipulation methods (and preliminary experiments) on the iCub [2-3 months]
- finalising the implementation of a visual reaching and grasping on the iCub and gathering of the required experimental results [2-3 months]
- finalising and thesis write-up [1-2 months]
Bibliography


M.A. Diftler, J.S. Mehling, M.E. Abdallah, N.A. Radford, L.B. Bridgwater, A.M. Sanders, 
Savely, and R.O. Ambrose. Robonaut 2 - the first humanoid robot in space. In Proc. of 

S. Doncieux, J.B. Mouret, N. Bredeche, and V. Padois. Evolutionary robotics: Exploring 


and Automation (ICRA), 2013.


H. Forssberg, A.C. Eliasson, H. Kinoshita, R.S. Johansson, and G. Westling. Development 
of human precision grip i: basic coordination of force. Experimental Brain Research, 

M. Frank, A. Förster, and J. Schmidhuber. Reflexive Collision Response with Virtual 
Conference on Agents and Artificial Intelligence (ICAART), 2012.

M. Frank, J. Leitner, M. Stollenga, A. Förster, and J. Schmidhuber. Learning to plan 


A. Gloye, F. Wiesel, O. Tenchio, and M. Simon. Reinforcing the Driving Quality of Soccer 

0-201-18075-8.


G.D. Hager, W.-C. Chang, and A.S. Morse. Robot hand-eye coordination based on stereo 

S. Harding, J. Leitner, and J. Schmidhuber. Cartesian genetic programming for image 
processing. In R. Riolo, E. Vladislavleva, M. D. Ritchie, and J. H. Moore, editors,


Appendix A: icVision Architecture

Figure 1. The icVision Architecture. It consists of loosely-coupled modules to provide a simple framework for development of computer vision solutions and integration of the research directly on the iCub hardware.

icVision is a distributed framework for running robot vision modules (see Figure 1). The 3D localisation is done the following: At first the camera images are acquired from the hardware via YARP. The images are converted into grayscale, as well as, split into RGB/HSV channels and distributed to all active icVision filters. Each filter then processes the images received using OpenCV functions. The output of this is a binary image, segmenting the object to be localised. A blob detection algorithm is run on these binary images to find the (centre) location of the detected object in the image frame. The position of the object in both the right and left camera images is sent to the 3D localisation module, where together with the robot's pose, a 3D location estimation is generated. The last step places the object in the existing world model (see Figure 2).
Figure 2. The 3D localisation workflow through the icVision framework modules.